****

****

Proceedings of

National Journal of Parallel and Soft Computing

Volume 02, Issue 01

Organized by

University of Computer Studies, Yangon

Ministry of Science and Technology, Myanmar

December, 2021

EDITORIAL BOARD

REVIEWER BOARD

National Journal of Parallel and Soft Computing Volume02, Issue 01

December,2021

CONTENTS

****

Distributed and Parallel Computing

Image Processing

Natural Language Processing

Networking and Security

Cloud Computing

Consistency of Caching Dynamic Web Data of Clinic Appointment System Using MESI Protocol

Nwe Ni Khine, Khine Khine Oo
University of Computer Studies, Yangon
nwenikhine@ucsy.edu.mm, khinekhineoo@ucsy.edu.mm

Abstract

The Internet and the World Wide Web have seen tremendous growth in the last decade. This growth has made it possible for millions of users to gain access to geographically distributed web content. However, due to the magnitude of the increase in the user population and the non- uniformity of content accesses, popular objects (especially those which change frequently), create server and network overload, and thereby significantly increase the latency for content access. Caching is a commonly used technique to reduce content access latencies. In caching system, data items are retrieved from the server machines, cached and processed at the client machines, and then write back to the server. Our Proposed system controls the cache consistency by using MESI (Modified, Exclusive, Shared, Invalid) protocol. This system can get the data transparency on any data update of any user of the clinic appointment system because of the MESI protocol.

Key words: ***MESI, consistency control, clinic appointment system***

1. Introduction

In the caching system, client fetches objects from the sever machine, operate on them locally and send back any modifications to the server. Such architectures improve system performance by utilizing the processing power of client machine. The server's load is reduced by performing as much computation as possible on client machine. Although, the consistency control is a vital role for the simultaneous execution of transaction over a database, client caching introduces inconsistency into the system. If two clients simultaneously read the same file and then both modify it, several problems occur. For one, when a third process reads the file from the server, it will get the original version, not one of the two new ones. This problem can be effects of modifying a file are not supposed to be visible globally.

Another problem is that when the two files are written back to the server, the one written last will overwrite the other one. When a cache entry (file or block) is modified, the new value is kept in the cache, but is also sent immediately to the server. As a consequence, when another process reads the file, it gets the most recent value.

The proposed system can control the data consistency of clinic appointment system by the use of MESI protocol; because of the user can notify the data inconsistency by the "Invalidation" phase.

2. Related Work

The related works of concurrency controls are discussed in this session.

Consistency Control in Shared Data Using Distributed Certification Algorithm [1]: This consistency control system for sharing data is implemented by using Distributed Certification algorithm. Distributed Certification algorithm detects

the conflict at transactions commit time and operates by exchanging certification information during the commit protocol. Moreover, by using distributed certification algorithm, data consistency can be validated data locally update and data globally update to avoid the lost and buried update. This system allows not only the server but also the clients to maintain a directory for each cached page. The directory for a page is organized as a status of processing (i.e. clients that cache the same page), indicating which client has a copy of that page in its cache, together with the state information. The related directory information is tagged with the data page and sent to the requester.

Implementation of Home-based lazy release consistency system for a distributed application [2] is an implementation of proving data centric consistency model using home-based lazy release consistency approach. It is a consistency control schemes for client server system. Client acquires a lock, and then client fetches data from server machine, and then execution the update locally. When client release a lock, the system generates the copy for the update data. It then sends these update data to server. When other client fetch the same data from server machine, the server responds the most updated contents of the data. Train ticketing sales system as case study is used to generate consistent transactions for the shared database system.

A study of Causally-Consistent Lazy Model for a Data-Centric Distributed Application [3]: This system implements a distributed database replication consistency management system using causally-consistent lazy replication. By using causally-consistent, it can help to get consistent data when updating copies of databases across different sites. Replication can give high availability from creation many copies of data at many servers that reduces data access latency and replica transparency. Replication can give good performance and reliability when one copy crashes by node failures and network partitions.

3. Background Theory

Maintaining Consistency of Client-Cached Data : In the client-server database environment, the server provides shared database access for multiple client workstations and that client's workstations may cache a portion of the database. This system intends to maintain the consistency of the client cache. The application program runs as a client process and communicates with the database server through messages. This increases the cost of each data request. One solution is to reduce the number of requests by caching a portion of the database on the client.

When a client cache is used, there must be a protocol between the client and server to ensure that the client cache remains consistent with the shared database. In this sense, the client cache may be viewed as active data since updates should trigger a cache-refresh operation. Active database allow applications to be informed of changes to some portion of a shared database by other transactions. In practice, this has meant that all updates to the database must be monitored by the database management system to determine if the updates affect the active data. When active data is updated, the database system must inform the affected clients that a change has occurred. Thus, all transactions incur additional overhead to support a service that they may never use i.e, detection and notification of updates. To maintain cache consistency, the system can integrate the cache consistency algorithm with the database management system.

3.1. MESI protocol

The **MESI protocol** is an Invalidate-based cache coherence protocol, and is one of the most common protocols which support write-back caches. It is also known as the **Illinois protocol** (due to its development at the University of Illinois at Urbana-Champaign). There is always a dirty state present in write back caches which indicates that the data in the cache is different from that in main memory. This marks a significant improvement in the performance.

MESI represent four exclusive states that a cache line can be marked with:

- Modified(M)
- Exclusive(E)
- Shared(S)
- Invalid(I)

3.2. States of MESI

Modified (M): The cache line is present only in the current cache, and is *dirty* - it has been modified (M state) from the value in main memory. The cache is required to write the data back to main memory at some time in the future, before permitting any other read of the (no longer valid) main memory state. The write-back changes the line to the Shared state(S).

Exclusive (E): The cache line is present only in the current cache, but is *clean* - it matches main memory. It may be changed to the Shared state at any time, in response to a read request. Alternatively, it may be changed to the Modified state when writing to it.

Shared (S): Indicates that this cache line may be stored in other caches of the machine and is *clean* - it matches the main memory. The line may be discarded (changed to the Invalid state) at any time.

Invalid (I): Indicates that this cache line is invalid (unused).

4. MESI Pseudo Code for Proposed System

```
Let SD = Search_Doctor,
    Y/N = Confirmation;
BEGIN
SD ← Enter doctor name to search;
Search_Doctor_for_OPD (SD)
{
   If (SD is exist in doctor list)
     {
        Check for other users are currently requested
        for that doctor;
       If (other user exist)
        {

           Set_Mode_for_SD = "Shared"
         }
        Else
         {

           Set_Mode_for_SD = "Exclusive"
         }
         End If
Display: Doctor Name,
         Speciality,
         Days for OPD,
         Periods for each OPD day;
         Message "Want to continue to get
                  appointment!";
         Y/N ← Submit for confirmation;
         If (Y/N == Yes)
         {

           Get_Appointment_for_OPD (SD, Days
                for OPD, Periods for each OPD);
         }
         Else
         {
                Exit from the system;
         }
         End If
}
Else
{
         Message "Doctor name mismatch!";
}
End If

}
Get_Appointment_for_OPD (SD, Days for OPD,
Periods for each OPD)
{
Check for current status for that doctor;
         If (Mode_for_SD = ="Shared")
         {
           Check for other modified mode user;
         If (Has other modified mode user)
         {

           Set_Mode_for_SD ← "Invalid";

           Message "Try another period";
```

```
        }
        Else
        {

            Set_Mode_for_SD ← "Modified";

            Appointment submission commit;
         }
        End If
}
Else If (Mode_for_SD == "Exclusive")
{

    Set_Mode_for_SD ← "Modified";

    Appointment submission commit;
 }
End If
}
END
```

4.1. Implementation of System

Online appointment scheduling is getting popular day by day is that it helps the patient to make the appointment to their doctor, clinic or hospital in an easier way. It makes it through the computer, access a website or software and makes an appointment, than to go to the hospital, wait in a line for a number of hours, just to make an appointment with the doctor for the next week or next month. Although the clinical appointment system has those benefits, the system may have consistency problems because of multiple users from different places are simultaneously processing. Therefore, this system presents a consistence clinic appointment system by using MESI protocol for the control of data consistency between multiple users from different web browsers. The detail processing steps are as shown in figure 1. This system only allows the registered user to get the appointment. So, the new user must be registered before requesting the desire appointment.

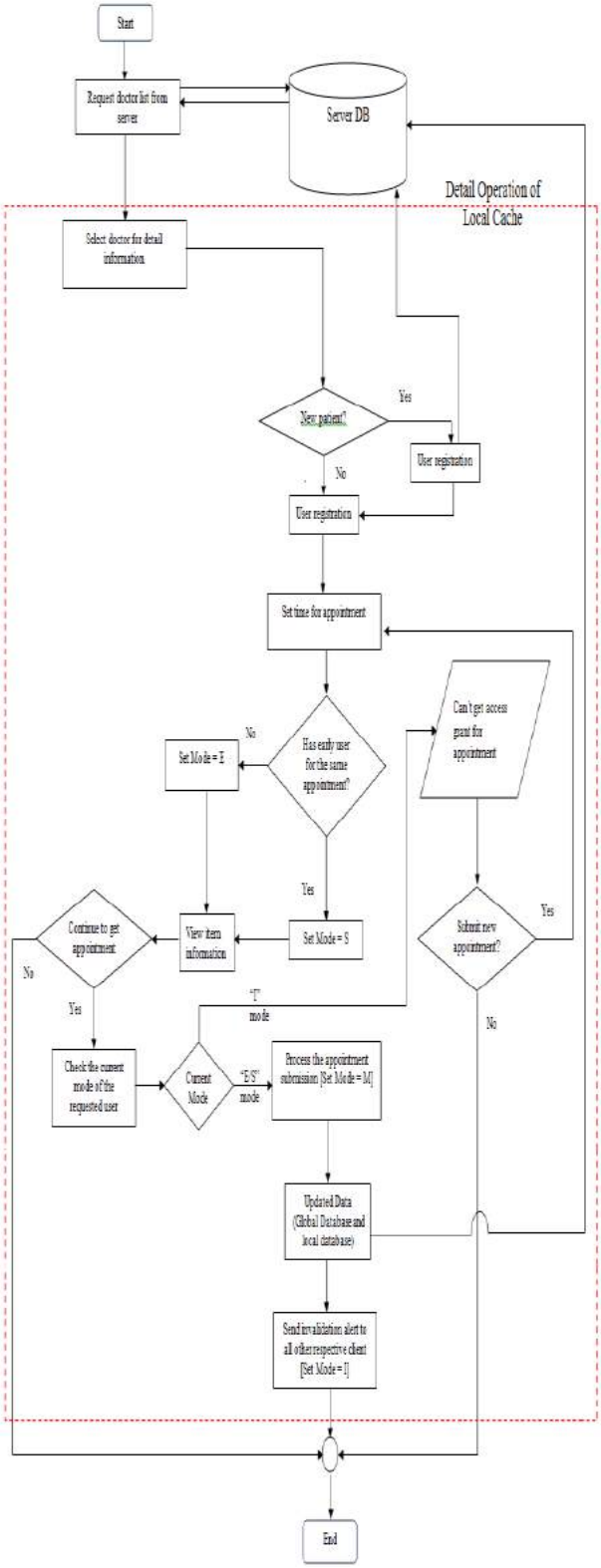


Figure1. The System Flow

The system controls the cache consistency by using MESI (Modified, Exclusive, Shared, Invalid) protocol. So, this system can get the data

transparency on any data update of any user of the clinic appointment system because of by noticing each process status (Whether "Modify" or "Exclusive" or "Share" or "Invalidation") of MESI protocol.

4.2. Database Design of the System

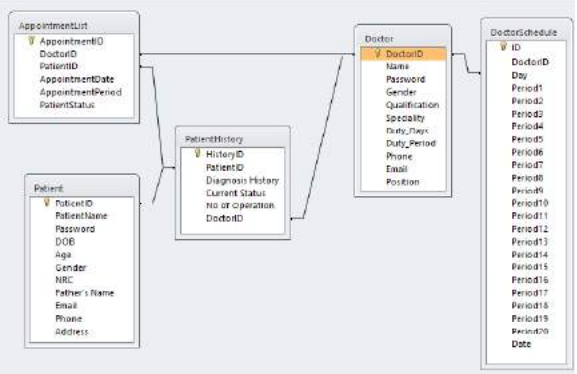


Figure2. The System Flow

This proposed system consists of five data table : "Patient" Table, "PatientHistory" Table, "Appointlist" Table, "Doctor" Table and "Doctor Schedule" Table. The interface design of the schedule list of a doctor is shown in figure 3.

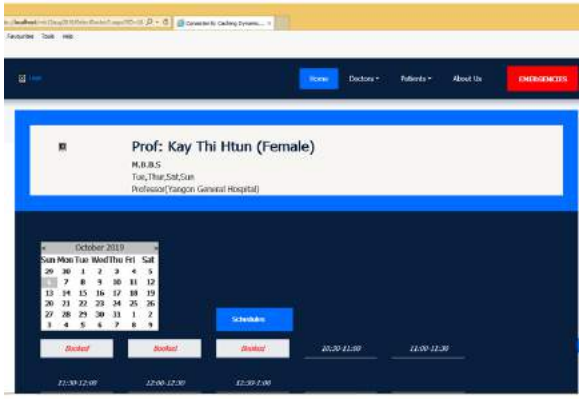


Figure2. Schedule of a Doctor

Patient information details are stored in "Patient" Table and the diagnosis history of the patients are stored in "PatientHistory" Table. Doctors' information are stored in "Doctor" Table and clinical schedule periods are stored in "Doctor Schedule" Table. "AppointmentList" Table is main control table for this proposed cache consistency control.

6. Conclusion

Appointment scheduling systems are used to manage access to service providers. The proposed clinic scheduling system allows individuals to conveniently and securely book their appointments online. Therefore, the proposed appointment system could significantly increase patient's satisfaction with registration and reduce total waiting time effectively. Then, this system also supports the data consistency by the use of MESI protocol.

REFERENCES

[1] "Consistency Control in Shared Data Using Distributed Certification Algorithm", Chan Myae Thu, M.C.Sc 2015, University of Computer Studies, Yangon.

[2] "Implementation of Home-based lazy release consistency system for a distributed application", Zar Zar Moe, M.C.Sc 2011, University of Computer Studies, Yangon.

[3] Elmasri R. and Navathe S. ―Fundamentals of Database Systems‖, Pearson Addison Wesley, 7th edition, 2015.

[5] Gray J. and Reuter A., ―Transaction Processing: Concepts and Techniques‖, San Francisco, Calif.:Kaufmann, 2011.

[6] Jose M. Faleiro and Daniel J. Abadi. ―FIT: A Distributed Database Performance Tradeoff‖, IEEE Data Engineering Bulletin, 38(1): 10-17, 2015.

[7] Kjetil Norvag, Olav Sandsta, and Kjell Bratbergsengen, ―Concurrency Control in Distributed Object-Oriented Database Systems‖,

Advances in Databases and Information Systems, 1997.

[8] Maabreh K. and Hamami A., ―Increasing database concurrency control based on attribute level locking‖, on the proceedings of International Conference on Electronic Design, ICED, IEEE, pp1-4, Issue 1-3, Malaysia, Penang. Dec. 2008.

[9] Maabreh K. and Hamami A., ―Implementing New Approach for Enhancing Performance and Throughput in a Distributed Database‖, The International Arab Journal of Information Technology, Vol. 10, No. 3, May 2013.

[10] Matthias N. and Matthias J., ―Performance Modeling of Distributed and Replicated Databases‖, IEEE transactions on knowledge data engineering, Vol.12 No.4, pp 645-672, July 2000.

[11] Muhammad Atif, ―Analysis and Verification of Two-Phase Commit &Three-Phase Commit Protocols‖, International Conference on Emerging Technologies (ICET), pp:326-331, Islamabad,19-20 Oct. 2009.

[12] Ozsu T. and Valduriez P., ―Principles of distributed database systems‖, Springer science and business, 3rd edition, New York, 2011.

[13] Silberschatz A., Korth H. and Sudarshan S. ―Database System Concepts‖, McGraw-Hill, New York, 6th edition, 2010.

[14] "MESI Cache Coherence Simulator for Teaching Purposes". Clei Electronic Journal. Volume 12, Number 1, Paper 5, April 2009.

[15] Jump up Culler, David. Parallel Computer Architecture. Morgan Kaufmann Publishers. pp. Figure 5–15 State transition diagram for the Illinois MESI protocol.

[17] “Teaching The Cache Memory Coherence With The MESI Protocol Simulator”, Electrotecnia y Electrónica. Escuela Politécnica Superior. Universidad de Córdoba. Av. Menéndez Pidal s/n. 14081. Córdoba. Spain.

[18] “Online Clinic Appointment Scheduling”, Xin Dai, Dr. Robert H. Storer Thesis Advisor , Lehigh University (2013).

Consistency Control in Group-Work Discussion Using Eager Invalidation

Khin Sandar Thein, Sabai Phyu

University of Computer Studies, Yangon

khinsandarthien@ucsy.edu.mm,sabaiphyu@ucsy.edu.mm

Abstract

Information Technology becomes essential for people for many purpose (Especially, for business and other organizations). Therefore, any type of systems are related to their works are used to get their purpose. This paper introduces a group work discussion using distributed system as a form of knowledge sharing in a private organization. User in this group can learn, share and discuss their opinion from any location. At the same time, update transactions from more than one user in the group on the same document may be occurred concurrently. So, this paper emphasizes consistency control using Eager Invalidation to get the reliable data.

Key words: ***Eager Invalidation, knowledge sharing, distributed system***

1. Introduction

Nowadays, the distributed system has become more and more important in Information Technology. It is the collection of autonomous computers and individual computers are communicated by interconnecting network over some hardware concept and they share some particular resources attempting to achieve parallel computations. Most of the real-time application areas are designed as a distributed system. Distributed system are used to enable data sharing capabilities and this can provide an abundant of benefits to the user.

In Distributed Data Sharing System, Data Consistency Control is a vital role. In this paper, distributed system consistency control for the Group-work project discussion is managed the concurrent updated transaction by using Eager Invalidation method. Eager invalidation may broadcast the invalidate message first when a data item in the group project is updated.

Therefore, users are reliable for the data consistency although they are not request for update data checking.

2. Related Work

The related works of concurrency controls are discussed in this session.

An important challenging in distributed system [1] described the consistency for replication. This means that when one copy is updated, we need to ensure that the other copies are updated as well; otherwise the replicas will no longer be the same. To achieve high performance of operations on shared data, designers of parallel computers have paid much attention to different consistency models for distributed shared memory systems. Advantages: This paper can control write consistency by using vector clock time synchronization algorithm like a global clock, it can define precisely which write operation is the last one. Disadvantages: Client-centric consistency is characterized by the lack of simultaneous updates. This may lead to increase latency in data update propagation.

The purpose of [2] is to prevent inconsistent retrievals among users who are simultaneously accessing on share database. This system will implement the Home-based lazy release consistency

control and vector timestamp synchronization by using train ticket sales system as case study. Advantage: Consistency control is performed on those databases. The goal of this system is to prevent inconsistent retrievals among users who are simultaneously accessing on share database. After communicating copy of data to the homes, they can be discarded. Disadvantage: If the home assignment to the shared pages is not well matched with an application's memory access pattern, the home-based protocol will suffer.

Nancy Lynch1 and Alex Shvartsman2, [3] mentioned to develop and analyze algorithms to solve problems of communication and data sharing in highly dynamic distributed environments. Advantages: Work on distributed services that provide useful guarantees and that make the construction of sophisticated distributed applications easier. Disadvantages: dynamic encompasses many types of changes, including changing network, changing sets of participating client processes, a wide range of types of processor and network failures, and timing variations. Constructing distributed applications for such environments is a difficult programming problem.

3. Background Theory

Data sharing - becoming increasingly important for many users - especially for businesses and organizations aiming to gain profit because of higher productivity. For example: Businesses get more work done as well as making collaboration with peers much more efficient - key to satisfying their business goals. Students - benefit when working on group projects - better able to collaborate with members - get work done more efficiently.

3.1. Distributed System

A distributed system can be defined as follows:

- A distributed system is a collection of autonomous computers linked by a computer network that appears to the users of the system as a single computer.
- A distributed system is one in which hardware or software components located at networked computers communicate and coordinate their actions only by passing messages.

3.2. Distributed Database System

Distributed Database System (DDBS) technology is the union of what appears to be two diametrically opposed approaches to data processing: database system and computer network technologies. A distributed database is a logically interrelated collection of share data that is stored on computer at several sites of a computer network and in which users can access data at any site in the network. An essential feature of a true distributed database is that users or programs work as if they have access to the whole database locally.

Distributed Database System consists of a collection of sites, connected together via some kind of communication network, in which

- Each site is a full database system site in its own right, but
- The sites have agreed to work together so that a user at any site can access data anywhere in the network exactly as if the data were all stored at user's own site.

4. The Requirements for Controlling in Distributed System

When multiple clients make concurrent updates the same record, requirements lead to a need to be consistency. To control the data consistency of the

distributed system, there are many kinds of controlling methods.

Lazy Release Consistency

- Lazy Release Consistency is a further optimization of the Release Consistency.
- It assumes that the thread executing an acquire access does not need the values written by other threads until the acquire access has completed.
- Datum and lock release are propagated together.

4.1. Types of Lazy Release Consistency

There are five types of lazy release consistency:

1. Lazy Invalidation
2. Lazy Update
3. Lazy Hybrid
4. Eger Invalidation
5. Eger Update

4.2. Lazy Invalidation

In Lazy Invalidation: only send the invalidation message when the user requests for the validation check.

System: Suitable for the periodically data update processing system (e.g. Data backup systems)

Advantages: Processing cost is lowered because of every updating is not sent to all participants as soon as the data updated.

Disadvantages: Not suitable for timely critical data update needed real time system. (Example: Banking system)

4.3. Lazy Update

Lazy Update: Never sends invalidation message. Only sends the data update when the user requests the data. Let the stale data in user site.

Advantages: Can get update when the user need.

Disadvantages: This is not suitable for Stock share trading and e-commerce site or pages. This can't support for timely updated data.

4.4. Lazy Hybrid

Lazy Hybrid: combination of Lazy Invalidation and Lazy Update. This technique sends invalidation message to all clients. When the user requests data, it sends the respective updated data lazily.

Advantage: When the user needs the updated data - update is sent.

Disadvantage: Send the invalidation message to all participants.

4.5. Eager Invalidate

Eager invalidation acquiring processor invalidates all pages in its cache for which it receives write notices. When updating participant releases the lock – sends invalidation message to all participants who own the respective updated data item.

System Example: Suitable for the Groupware systems (example: Open Source share data editing)

Advantages: No need to request to send the invalidation message to the respective participants. As soon as the updating participant released the lock, the invalidation messages are sent.

Disadvantages: When a lot of users are in the groupware system, the invalidation messages broadcasting cost will be high.

4.6. Eager Update

When updating participant releases the lock– sends data update to all participants who own the respective updated data item. (no need to request for data update)

System example

- Banking System
- Real Time System

Advantage: Suitable for banking system. The system always support for timely updated data.

Disadvantage: As the system sends the update immediately, the original data (old version) will also be lost immediately. Update processing cost will be high.

5. The System Overview

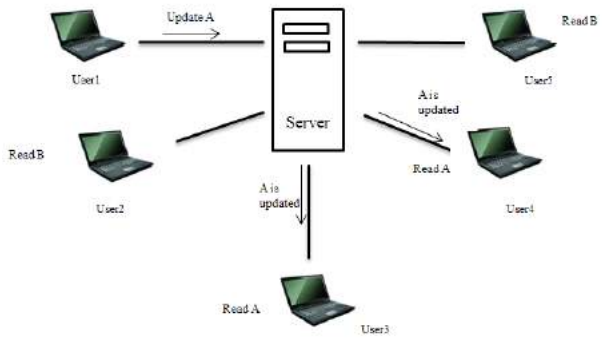


Figure 1: System Overview

Figure 1 shows overview of the proposed system. According to figure, the distributed data consistency at each site is controlled by the Eager Invalidation method. The proposed system is organized by a server and then the server data is distributed at each site. Therefore, the client at each site must be registered user at the server. By this way, the server can send the invalidation message to the respective data user when one of the members of the system is update a data item. Although the read transactions at each site cannot effect the data consistency, the write transactions may lead to data inconsistent state. In this case, this system will be used Eager Invalidation for consistency control.

5.1. Implementation of the System

Figure 2 shows the proposed system flow. As the data are shared in the proposed system, the text files will be applied. The implicit dynamism of client behavior makes the difficulties to obtain accurate results. To obtain accurate results: Firstly, the system checks whether the global access data in server and local cache data are matched or not. If these data contents are not the same, document will be get from global and then copy is stored in local cache. If these data contents are the same, the document will be shown which is selected by user.

At the start of any transaction, the system will be set the timestamp and action (read or write) for the document which is accessed. And then, the system/server checks that the status is concurrent state. If the status is concurrent, the system compares the time stamps of the concurrent transactions and sends messages to the conflict sites instead of locking. The conflict sites will be opened read mode only. The early time stamp of client's write set transaction will be allowed the write access. When the early time-stamp of client's write set transaction was performed, the time stamp of read set transaction will know this document is out of data (or) the later time stamp of write set transaction will know that this document is already opened to write by other user.

If the status is not concurrent, the current transaction of the data contents will be committed and then update the global and then multicast data update to other caches.

Some clients have an inconsistent directory relative to the server for the same page. As a result, those clients may have outdated presence flags in their directories. Such directory inconsistency causes a problem only when those clients want to update the page.

When the server receives a speculative update request for a shared page, the server compares its directory with that of the client. If the server detects

that the client directory is outdated, it grants the speculation, but at the same time it informs the client of the discrepancy. If there are some new sharing clients, the speculative client is not allowed to commit before the new sharing clients invalidate their copies.

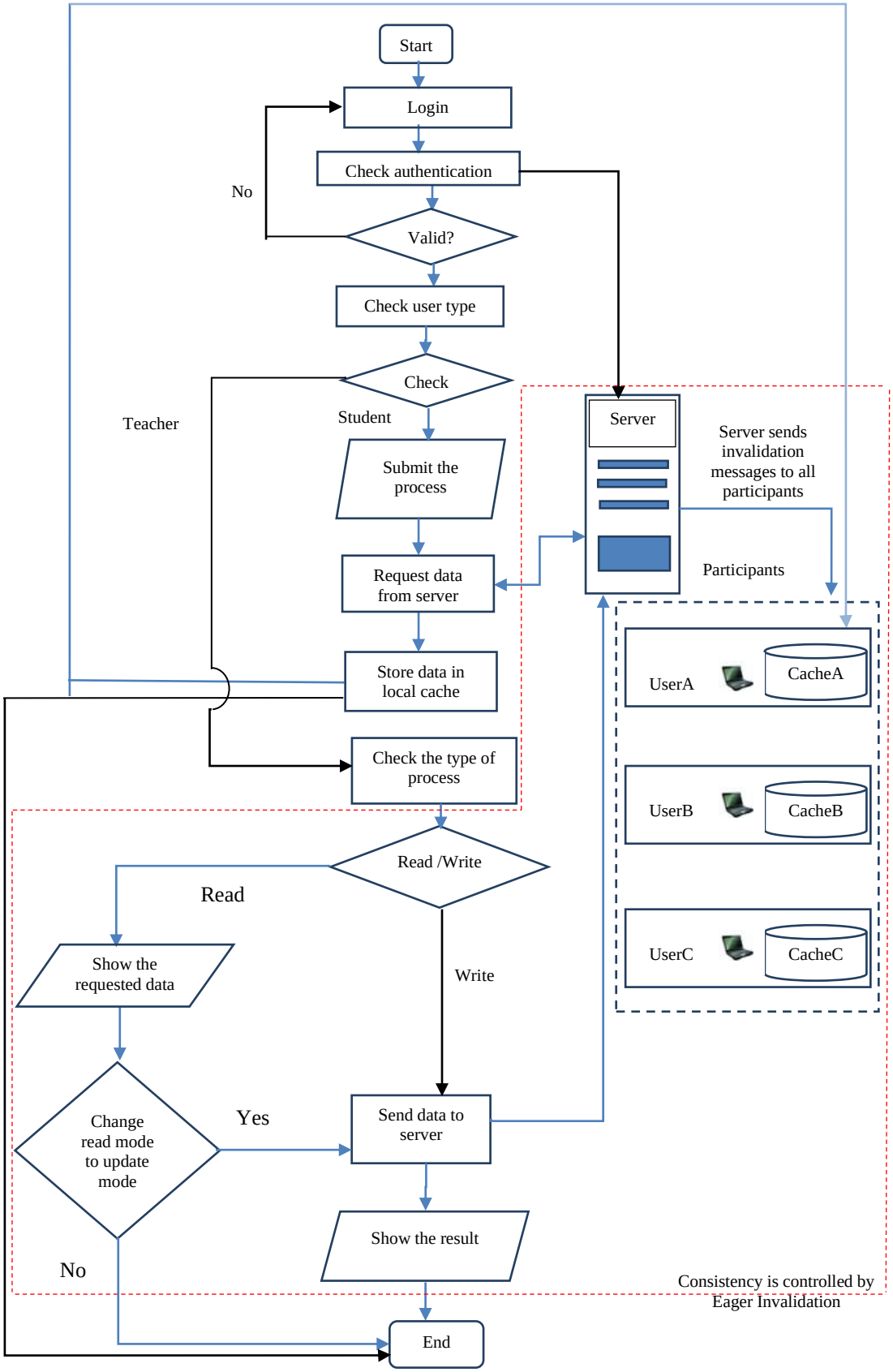


Figure 2: System Flow

6. Conclusion

This paper mentions about distributed system that used Eager Invalidation method. In the proposed system, documents can be shared from the teachers to the students and can be updated by the teachers. When teacher update a document, this proposed system allows sending updated version of that document to the students of the group via message. Data consistency can be validated data locally update and data globally update to avoid the lost and buried update by using Eager Invalidation.

This proposed system implements the data sharing system for approving concurrency and consistency control. Since it is the shared database system, it depends on the server database and client database. This proposed system can be extended priority based validation. Moreover, the system with caching data in client sides and working with those cached data while server is down will improve and solve the failures of the central database system.

REFERENCES

[1] Aye Nyein Mon, Thinn Thu Naing,"Implementation of Client-Centric Consistency Control and Synchronization System For Distributed Replication(For Mobile Clients)", University of Computer Studies, Yangon, 2010.

[2] Zar Zar Moe, Thinn Thu Naing, zarzarmoe.88@gmail.com, "Implementation of Home-based lazy release consistency system for a distributed application", University of Computer Studies, Yangon, 2010.

[3] Nancy Lynch1 and Alex Shvartsman2, "Communication and Data Sharing for Dynamic Distributed Systems", 2013.

[4] Philip S. Yeager , "A Distributed File System For Distributed Conferencing System", (University Of Florida), 2016.

[5] Spyros Voulgaris, Maarten Van Steen, Aline Baggio, And Gerco Ballintijn "Transparent Data Relocation In Highly Available Distributed Systems", 2013.

[6] Philip Homburg, Maarten van Steen, Andrew S. Tanenbaum: "An Architecture for A Wide Area Distributed System", 2006.

[7] Kjetil Norvag, Olav Sandsta, and Kjell Bratbergsengen, ―Concurrency Control in Distributed Object-Oriented Database Systems‖, Advances in Databases and Information Systems, 201 7.

[8] Maabreh K. and Hamami A., ―Increasing database concurrency control based on attribute level locking‖, on the proceedings of International Conference on Electronic Design, ICED, IEEE, pp1-4, Issue 1-3, Malaysia, Penang. Dec. 2008.

[9] Maabreh K. and Hamami A., ―Implementing New Approach for Enhancing Performance and Throughput in a Distributed Database‖, The International Arab Journal of Information Technology, Vol. 10, No. 3, May 2013.

[10] Zin Phyo Ko, "Consistency Controlling in Distributed File System by Using Wound Wait (WW) Control", University of Computer Studies, Yangon, 2017.

Data Recovery in Cloud Computing By Using Enriched Genetic Algorithm (EGA)

Phyu Phyu Thant, Yu Mon Zaw, Khine Moe Nwe
University of Computer Studies, Yangon
phyuphyuthant@ucsy.edu.mm, yumonzaw@ucsy.edu.mm, khinemoenwe@ucsy.edu.mm

Abstract

Cloud computing provides various kinds of services to its users. Storage-as-a-service is one of the services provided by cloud infrastructure in which large amount of electronic data is stored in cloud. As valuable and important data of enterprises are stored at a remote location on cloud we must be assured that our data is safe and be available at any time. In situations like Flood, Fire, earthquakes or any hardware malfunction or any accidental deletion our data may no longer remain available. To maintain the data safety there must be some data recovery technique for cloud platform to recover valuable and important data efficiently in such situations mentioned above. In this paper, a multi servers system based on Enriched Genetic Algorithm to recover the lost data by using four cloud backup servers is discussed. To achieve reliability the proposed method provides the flexible for the user to accumulate data (data restore) from the backup server at the same time when main cloud server loses its data and is unable to provide data to user.

Keywords: Cloud Computing, Data Recovery, Backup, Data Restore

1. Introduction

Nowadays cloud computing is one of the useful generation. It provides on demand resources for assets to the consumer/user. Cloud Computing becomes the delivery of computing offerings which include servers, storage, database, networking, software etc., over the internet to provide quicker innovation, flexible resources and economic scale. Clouds have many capabilities distributed over multiple locations from central servers. Cloud computing has three cloud service models. They are: *Software as a Service (SaaS):* SaaS is a way of delivering applications over the Internet—as a service. Instead of installing and maintaining software, the user can simply access it via the Internet, freeing themselves from complex software and hardware management. SaaS applications are sometimes called Web-based software, on-demand software, or hosted software. Whatever the name, SaaS applications run on a SaaS provider's servers. The provider manages access to the application, including security, availability, and performance. *Platform as a Service (PaaS):* PaaS provides customers to develop his very owns application the user of the tools and programming languages. This service is hosted in cloud and accessed by clients via the internet. Google App engine, Amazon AWS provides the platform as a service. *Infrastructure as a Service (IaaS):* IaaS provides the shared useful resource services. It affords the computing infrastructure like storage, virtual machine, network connection, bandwidth, IP address. IaaS is entire package for computing. Amazon, GoGrid provides the infrastructure as the service to the user [1]. Cloud computing has four deployment models. They are: *Public Cloud:* A public cloud is available to any user with an internet facility, is much less steady than the private cloud due to the fact it may be accesses by general public. *Private Cloud:* Private cloud is available to a particular organization in order that the user who belongs to that organization can have access the data. It is extra steady than the public cloud due to its private nature. *Hybrid Cloud:* The hybrid cloud is essentially combination of no much less than two clouds such as mixture of private, community or public cloud. *Community Cloud:* Community cloud permits the resources and system to be accessible by number of related organization. Data storage is one of the maximum great offerings provided by cloud computing technology. But, recovering the lost data is one of the challenging issues in cloud computing paradigm. A quick evaluate of facts healing in cloud computing is mentioned below.

Data saved on the datacenter is growing day by day it leads into big quantity of data storage in cloud and outcomes into issues such as data loss, data breach etc. There is a need of an efficient technique if

the data get destroyed or deleted via way of means of mistake to recover the data from any backup server. In enterprises continuity if the system crashed or any sort of natural or human made disaster occurred then there may be hazard of data loss and it may also cause the financial loss. By using some of the data recovery techniques the original data can be recovered. But, the existing recovery techniques are not efficient and reliable. Therefore, to get better misplaced original data, efficient and reliable recovery technique is needed.

2. Related Work

This segment presents summary of some of the data backup and recovery techniques in cloud computing.

In paper [1], author proposed the DR-Cloud version which is fault tolerant multi cloud storage, it makes use of DR XOR codes which affords data redundancy and uses minimum repair traffic during data transmission. DR-Cloud acts as interface between user application and multi cloud servers. The paper [2] proposed the unconventional technique to recover the data. It solves all current problems with data recovery by automatically compressing and decompressing the data earlier than the backup of the data. Dual backup system was used. The dual system affords the excessive reliability and higher bandwidth utilization of data storage. In paper [3], author proposed the Advanced Encryption Standard (AES) and Seed Block algorithm (SBA) technique to carry out the smart remote data backup in cloud computing surrounding. The proposed method makes use of the AES and seed block algorithm. If the data gets deleted via way of means of the mistake then we can get it from the remote server. This method takes much less time to recovery the data and solves the time related issues. Thus the method offers an efficient security mechanism for the data saved in the cloud environment. The paper [4] proposed the cloud mirroring method. It makes use of the mirroring algorithm. The technique affords the excessive availability, integrity of the data, recovery of the data and minimizes the data loss. This technique can be implemented to any sort of the cloud. Cost to recover the data is also less.

In paper [5], author has proposed the data backup and recovery technique. This method affords the data protection from the service failure and also decreases the cost of solution. By the use of this method the procedure of migration will become easy and also removes the cloud vendor dependency. They proposed a powerful data backup technique to recover the data from the server in case of data loss. For every enterprise it is far critical to return up the data to keep away the data loss. The paper [6] presented a way which incorporates business service procedure (BSP) and disaster recovery procedure (DRP) with an assistance of cloud environment as a way to keep away from disaster recovery problems. The work employs priority based technique towards data recovery. The proposed method guarantees that it may offer protection to entire organization datasets, which may contain log, account files. It additionally guarantees that it can reduce the time required to get better organization data within small quantity of time. In paper [7], evaluated has been carried out on distribution of data in cloud environment by construction of privacy preserving techniques and RBD. In order to carry out smart RBD the system employs encryption and compression methods. The paper aims to maintain user privacy. System demonstrated that it can overcome time associated issues and are also solved via way of mean of encryption and compress techniques.

In the proposed system, cloud computing involves networks of groups of servers with specialized connections that spread data-processing storage across them. This shared cloud infrastructure contains large pools of systems which are linked together. These backup servers are used to maximize the power of cloud computing and data is stored in the form of distributed replicated data store.

3. Proposed Data Recovery Technique in Cloud Computing

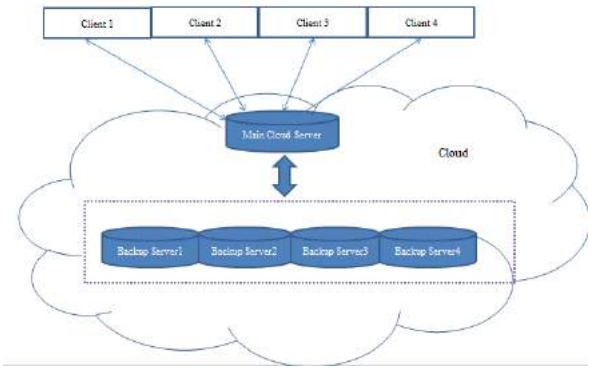


Figure 1: Architecture of remote server

The proposed architecture represents three modules inclusive of 1.Remote Backup Server 2. Main Cloud Server 3.Number of Clients/users as shown in figure1. Remote Backup server keeps the replicated copies of main server and is known as central repository it stores all the user data. The user uploads the file to main cloud server; the main cloud server stores all the data in backup server. If user wants to retrieve the file from the cloud then the file is searched in main cloud server firstly, if the data is not present in main server then it is checks in backup server to retrieve lost data.

If the loss of data or data crash takes place because of natural disasters or human made disasters in main server, to recover the lost data a recovery technique is essential. Recovery of data can be achieved through the use of proposed algorithm efficiently. To provide the reliability two or four backup data cloud storage could be used. The figure 2 shows the system architecture with four backup cloud storages.

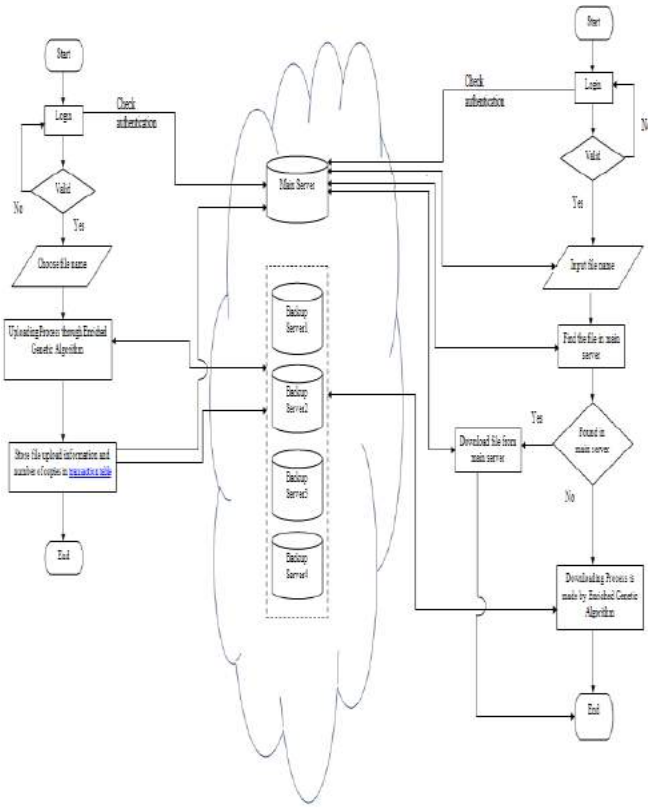


Figure 2: System architecture

The system architecture illustrates four backup servers as shown in figure2. The replicated copies of data are maintained in more than one server to recover data. When data loss occurs at one location it can be retrieved from other backup server using Enriched Genetic Algorithm (EGA).

3.1. Procedure of Enriched Genetic Algorithm (EGA)

1. User uploads the file F to the N cloud servers.
2. From the file F generates hash code H1 and stored in data base.
3. Calculate the size of file.
4. User has to select the file to be downloaded.
5. If the file is deleted then it is retrieved from the backup server
6. Select the file to be downloaded and generate the hash code H2.
7. If both the hash codes are same then we retrieved the original file.

3.2. File Uploading Algorithm

Let N be the number of copies needed.
BEGIN
User uploads the file F to the cloud server.
Let N = the number of copies needed [Server count included main server and backup server]
File F generates hash code H1 and store in database for integrity checking.
For (i=1 ; i <=N; i++)
{Select i th server and its available balance;
New Balance = Available Balance – Request Balance of Uploaded File;
If (New Balance < 0)
{Display “No Space to upload file”;
Break ;}
Else
{Upload the file to server and update the balance;
Store the file upload information and number of copies in transaction table ;}
}
End If}
End For
END

3.3. Recovering Algorithm

BEGIN
Select the file to be downloaded.
From the transaction table get the numbers of cloud storage containing the file and N server configuration details.
For (i=1; i <=N; i++)
{Select i th server status,
If (status is activated)

{Download file from i th server;
Generate hash code H2 from the file;
Fetch the hash code from Data base;
If (H1=H2)
{Display “File Integrity check is successful”;
Display “File recovered Successful”: STOP}
Else
{Display “File Integrity check is not successful” ;}
End If
}
End if
}
End for
END

3.4. MD5 Hash Generator

MD5 is a type of algorithm that is known as a cryptographic hash algorithm. MD5 is an upgraded version of MD4. MD5 produces a hash value in a hexadecimal format. This competes with other designs where hash functions take in a certain piece of data, and alter it to provide a key or value that can be used in place of the original value. It is used for the reason of data verification in transmission protocols. In many web applications, MD5 hash is used to prevent security breaches, hacking etc. by the way of enhancing security. MD5 algorithm is very much helpful as it is a bit easier for storing and comparing smaller hashes than the large text of variable length. It’s simple to develop a message digest from the original message.

4. File Uploading with Hashing

In File uploading module, user has to select the file to upload to cloud by selecting the number of copies of replication required to store. While uploading charm application will read the file size in kb. Then it will select the best cloud storage server based on the storage availability, pricing cost, predictor, size etc. For integrity Verification process it will generate the Hash Key (HK1) using MD5 algorithm and it will keep it in the user DB. Finally based on the Replication details the File will be stored in the Cloud Storage Server.

5. File recovery

When user request the file from main cloud server he has to select the file from the Data Recovery Application then the data Recovery Server will select the corresponding cloud sever details from the DB and also it will check for the cloud availability for recovering the file, if cloud server is not available then it will be recovered from another cloud server. While recovering it will generate the HASH Key (HK2), then it will check for the HK1 &HK2 for the integrity check. Finally the file is recovered from the backup server.

6. Sequence Diagram of System

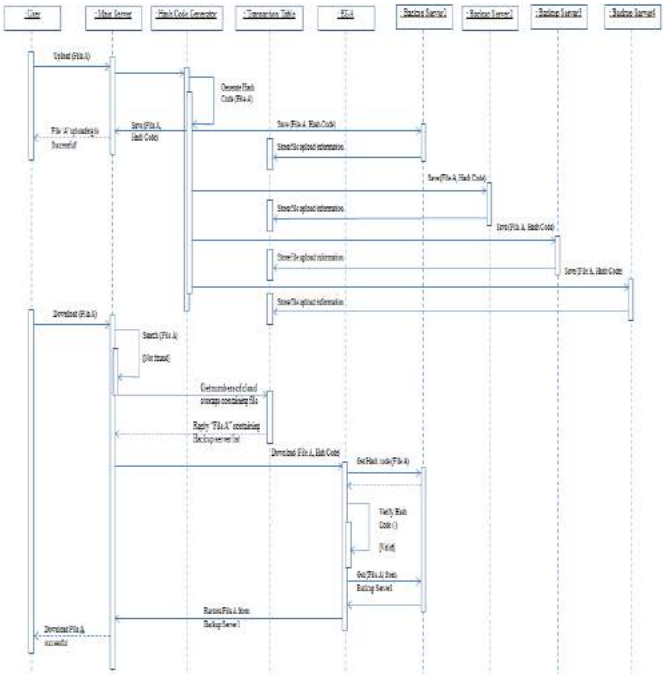


Figure 3: Sequence Diagram

At the starting of algorithm, user has to register first by entering personal information. After successful completion of registration user can login into the system with the help of username and password which are assigned. The user information is stored in the system’s servers. The next step after the login is uploading the file. After user uploads the file, the system generates the hash code H1 automatically for particular file using the MD5 algorithm and stores that file in the system’s server. If user wants to retrieve the file from cloud then the system is searches in main cloud server firstly and downloads from the main server. In the case of data loss due to any reason such as natural calamity, man-maid attacks, and server crash file on main cloud server gets deleted or crashed. If the data is not present in main server then then the system searches the file from any backup server and generates hash code H2. To download the deleted file, the user must type hash code H2 in the text box firstly. The user can find hash code H2 from

the mail box. After typing the hash code H2, the system fetches the hash code H1 from the database. If the hash code H1 and H2 are match, the user can retrieve the original file and the system restores the original file in the main server.

7. Experimental Results

The proposed model is simulated using C# (asp.net) language in window 10 PC. This system uses Google Cloud with four backup servers. And then this system simulate by using ASP.NET version 4.5.2 and Microsoft SQL version 10.0.2573.0. This system can use in all operation system (Linux, Window, Max). Your files are stored in our servers until you delete it and this system make uploading and recovering your files quick and easy. You can store in our servers for many different file types and sizes.

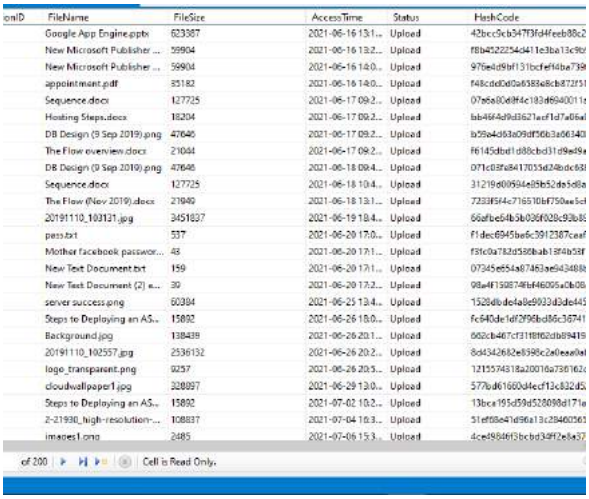


Figure 4: Different file types and sizes in our servers.

The experiment is conducted by taking different types of files and their sizes as shown in table1.

Table1: Different types of files and their sizes

Type	File Size	File size in remote servers	File size of recovered file
.txt	**250KB**	**250KB**	**250KB**
.pdf	**580KB**	**580KB**	**580KB**
.jpg	**30KB**	**30KB**	**30KB**
.png	**40KB**	**40KB**	**40KB**

Table2: The percentage of the originality of recovered files with the different numbers of files and that are different sizes.

No.	Size Range (KB)	Number of Recovery Testing Files	Percentage of recovered file size with respect to original file size	Recovery Time (Time consuming)
Testing Group 1	10 – 50	100 files	100%	5ns
Testing Group 2	100 – 1000	100 files	100%	5ns
Testing Group 3	1001 –10000	100 files	100%	6ns
Testing Group 4	10001 – 100000	100 files	100%	10ns
Testing Group 5	100001 1000000	100 files	100%	12ns

In these experiments, there are five groups of files according to the data volume ranges. This system mainly emphasized on data recovery for zero data loss. So, this system can prove zero data loss according to the experiment results of table2. The recovery time in delay point of view is an acceptable minimum time as shown in the above table.

The advantages and disadvantages of all the above discussed techniques from related works are described in Table3. And due to the high applicability and need of backup process in many companies and enterprises, the role of a remote data back –up server with an efficient technique is very important and a hot research topic.

Table3: The advantages and disadvantages of techniques from related works and proposed technique

No.	Approach	Advantages	Disadvantages
1	DR-Cloud version	-affords data redundancy and uses minimum	-Costly, -Increased -redundancy

		repair traffic during data transmission	
2	Unconventional technique	-Reliable -Privacy -Low cost	-High complexity
3	Advanced Encryption Standard (AES) and Seed Block algorithm (SBA) technique	-Takes much less time to recovery the data and solves the time related issues.	-Cost increases as data increases
4	Cloud mirroring method	-Cost to recover the data is also less.	-High bandwidth, Complete server backup at a time
5	Procedure of migration	-Data protection from the service failure and also decreases the cost of solution.	-Increased redundancy
6	Business service procedure (BSP) and disaster recovery procedure (DRP)	-Offer protection to entire organization datasets.	-High Complexity (due to contain log, account files)
7	Proposed Technique (Enriched Genetic Algorithm)	-Exact match retrieval, privacy. -Recovery time is not enormously increases as data increases	-Increased complexity

8. Conclusion

Now a day's large amount of data is stored in the cloud and becoming very important to all the organization. The four backup server's concept is used to recover the deleted data. The experimental result section show the proposed method has reliable and efficient. Because user files can recovery from any back server among four backup servers in the case of data loss and although the data size increases, the recovery time does not increase.

9. References

[1] Mahantesh N. Birje, Praveen S. Challagidad, "Cloud computing review: concepts, technology, challenges and security", *International Journal of Cloud Computing*, InderScience Publishers, vol. 6, issue 1, 2017.

[2] P. S. Challagidad, M. N. Birje, "Hierarchical Attribute-based Access Control with Delegation Approach in Cloud", Proceedings of the 11th INDIACom; INDIACom-2017; IEEE Conference ID: 40353 *2017 4th International Conference on "Computing for Sustainable Global Development*", 01st - 03rd March, 2017.

[3] Greeshma Radhakrishnan, Chenni Kumaran, "DR – Cloud: Multi- Cloud Based Disaster Recovery Service", *International Journal of Innovative Research in Science, Engineering and Technology*, Vol. 5, Issue 3, March 2016.

[4] Megha Rani Raigonda, Tahseen Fatima, "A Cloud Based Automatic Recovery and Backup System with Video Compression", *International Journal of Engineering and Computer Science*, ISSN: 2319-7242, Vol. 5, Issue 09, and September 2016.

[5] Tanay Kulkarni, Sumit Memane, "Intelligent Cloud Security Back-Up System", *International Journal of Technical Research and Applications*, Vol. 3, Issue 2, Mar-Apr 2015.

[6] Shilpi U. Vishwakarma and Praveen D. Soni, "Cloud Mirroring: A Technique of Data Recovery", *International Journal of Current Engineering and Technology*, Vol. 5, No. 2, March 2015.

[7] PS. Vijayabaskaran, "Efficient Backing up Data for Migrating Cloud to Cloud", *International Journal of Computer Science and Information Technologies*, Vol. 6, 2015.

Data Mining and Web Mining

Data Clustering using on Differential Evolution Algorithm

Phyo Ei Nyein
University of Computer Studies, Yangon
phyolaypsk@gmail.com

Abstract

Clustering (or cluster analysis) aims to organize a collection of data items into clusters, such that items within a cluster are more "similar" to each other than they are to items in the other clusters. There are many applications for clustering such as image segmentation, marketing, ecommerce, business, scientific and engineering. The k-means has served as the most widely used partitional clustering algorithm. However, in most cases it provides only locally optimal solutions. Evolutionary algorithm such as genetic algorithm and differential evolution can be used to find global optimal solution for optimization problem. Clustering can be regarded as optimization problem of finding optimal partition of data according to cluster validity measures. Differential evolution (DE) algorithm is a novel evolutionary algorithm (EA) for global optimization, where the mutation operator is based on the distribution of solutions in the population. The paper presents the differential evolution for clustering. The empirical studying is conducted on five datasets from UCI data repository.

Keywords: Clustering, Global Optimization Algorithm, Evolutionary Algorithm, Differential Evolution, K-Mean.

1. Introduction

The goal of clustering is to group similar objects together based on some notion of similarity. Over the years, many clustering algorithms have been developed, each utilizing different distance/similarity measures and/or objective functions. Applying different methods, or the same method with different parameter choices to the same data, the algorithm can obtain varying clustering results.

Clustering aims at representing large datasets by a fewer number of prototypes or clusters. It brings simplicity in modeling data and thus plays a central role in the process of knowledge discovery and data mining. Data mining tasks, in these days, require fast and accurate partitioning of huge datasets, which may come with a variety of attributes or features. This, in turn, imposes severe computational requirements on the relevant clustering techniques. A number of evolutionary algorithm has recently emerged to solve the optimization problem that can be applied to a number of real world clustering problems. Traditional clustering technique such as K-means is sensitive to the initialization step for cluster accuracy. In order to eliminate the shortcoming of K-mean, one have to place every data point in the dataset as a starting cluster centers, that is computationally infeasible. Another drawback of k-mean algorithm is it can only find local optima solution. Differential Evolution algorithm is a global optimization algorithm can be used for optimization problems. Clustering can be seemed as optimization where the objective is to find the cluster solution according to some cluster validity measure. So, Differential evolution is a good choice for partitional clustering problem due to their nature of solving effectively in real point optimization problem. Differential Evolution used population of chromosomes searching for the multidimensional problem space. Instead of using only one search solution like k-mean, DE clustering used multiple solution that are encode in chromosome to search for appropriate clusters.

2. Related Work

There have been many reports on literature on clustering based on evolution and optimization algorithm.

Sandra Paterlini et al [1] used Differential Evolution for partitional clustering problem. Medoid representation is used for the chromosome in the Differential evolution where medoid represents the cluster medoid. For the fitness meature, TRW (Trace within Measure) and MC(Marriott's criterion) is used. Authors compared clustering with Genetic Algorithm, PSO (Particle Swarm Optimization) , K-mean and Differential Evolution algorithm and experimental results stated that DE should be used for clustering rather than GA.

Swagatam Das and Ajith Abraham used DE for clustering large unlabled data set without required the number of cluster to be specified [2], algorithm select the optimal number of clusters automatically. Centroid based representation was used and DB index was used as the fitness function.

D.Zhaire et al [3] used DE for document clustering problem and can automatically discover number of clusters. Compactness, connectedness and separability are used as quality measure of their study. Bag-of-word representation is used for representing the document in a vector form.

DE was also applied to image clustering problem where images are segmented by clustering using DE algorithm [4]. Authors use pixels as the data objects for image clustering. The image is converted into its corresponding RGB values. The gray scale of these values are computed which represents the intensity of the brightness. Quantization error is used as the fitness measure, centroid representation is used.

In [5], Yanfei Zhong et al used adapitve multi-objective DE based fuzzy clustering for clustering of remote sensing images. Fuzzy Objective function and XB index were used as the fitness criteria of the algorithm.

3. Background Theory

Clustering algorithms can be categorized as either hierarchical or optimization. Partitional clustering algorithms aim to identify homogeneous groups by finding similarities between objects regarding their characterizing attributes. The algorithmic task can be stated as an optimization problem for which the objective is to maximize the similarities among objects within the same clusters while minimizing the dissimilarities between different clusters. This can be quantified by a statistical criterion, such as by defining the objective as the minimization of the trace of the within variance matrix. The most popular partitional clustering algorithm is k-mean algorithm which has advantages of ease of implementation and efficiency due to its linear time. The drawbacks of k-means are: first , they are sensitive to random initialization of cluster center, second, the clustering algorithm used the greedy approach which is a local search algorithm and can easily be trapped into the local optima solution.

Differential evolution (DE) [10] algorithm is a novel evolutionary algorithm for faster optimization, which mutation operator is based on the distribution of solutions in the population. Among DE's advantages are its simple structure, ease of use, speed and robustness. Experimental results have shown that the DE-based clustering algorithms can provide higher performance than GA-based clustering algorithms..

3.1. K-means Clustering

The K-means algorithm is simple, straightforward and is based on the firm foundation of analysis of variances. It clusters a group of data vectors into a predefined number of clusters. It starts with randomly selected initial cluster centers and keeps reassigning the data objects in the dataset to cluster center based on the similarity between the data object and the cluster center. The reassignment procedure will not stop until a stopping criterion is satisfied (e.g., the fixed iteration number, or the cluster result does not change after a certain number of iterations). The K-means algorithm can be summarized as:

(1) Randomly select cluster centroid vectors to set an initial dataset partition.
(2) Assign each data vector to the closest cluster centroids.
(3) Recalculate the cluster centroid vector c_j using the following equation

$$c_j = \frac{1}{n_j}\sum_{\forall d_j \in S_j} d_j \qquad (1)$$

where d_j denotes the data vectors that belong to cluster S_j; c_j stands for the centroid vector; n_j is the number of data vectors that belong to cluster Sj.
(4) Repeat step 2 and 3 until the stopping criteria is satisfied.

The main drawback of the K-means algorithm is that the result is sensitive to the selection of the initial cluster centroids and may converge to the local optima. Therefore, the initial selection of the cluster centroids affects the main processing of the K-means and the partition result of the dataset as well.

3.2. Differential Evolution

The DE algorithm, as proposed by Storn and Price [6], is a new floating-point encoded evolutionary algorithm for global optimization. The DE algorithm has demonstrated good convergence properties, is fundamentally easy to understand, and has been used in many different fields. The aim of DE is to find an individual which minimizes the objective function by mutation, crossover and selection operators, as shown in following algorithm.

Input:
- numbers of clusters
- population size

Output:
- the best fitness as the solution

(1) set the generation counter, t=0;
(2) Initialize the control parameters,
(3) Create and initialize the population, C(0),of n_s individuals;
(4) while stopping condition(s) not true do
(5) for each individual, $x_i(t) \in C(t)$ do
(6) Evaluate the fitness, f(xi(t));
(7) Create the trial vector, $u_i(t)$ by applying the mutation operator;
(8) Create an offspring, $x'_i(t)$, by applying the crossover operator
(9) if $f(x'_i(t))$ is better than $f(x_i(t))$ then
(10) Add $(x'_i(t))$ to C(t+1);
(11) end

(12) else
(13) Add ($x_i(t)$) to C(t+1);
(14) end
(15) end
(16) end

3.2.1 Mutation

The DE mutation operator produces a trial vector for each individual of the current population by mutating a target vector with a weighted differential. This trial vector will then be used by the crossover operator to produce offspring. For each parent, xi(t), generate the trial vector, ui(t),as follows: Select a target vector, xi1 (t),from the population, such that $i \neq i1$. Then, randomly select two individuals, x_{i2} and x_{i3} , from the population such that $i \neq i1 \neq i2 \neq i3$ Using these individuals, the trial vector is calculated by perturbing the target vector as follows:

$$u_i(t)= x_{i1}(t)+ \beta(x_{i2}(t) - x_{i3}(t)) \quad (2)$$

where $\beta \in (0,\infty)$ is the scale factor, controlling the amplication of the differential variation.

3.2.2 Cross over

The DE crossover operator implements a discrete recombination of the trial vector, $u_i(t)$, and the parent vector, $x_i(t)$, to produce offspring, $x'_i(t)$. Crossover is implemented as follows:.

$$x'_{ij}(t) = \begin{cases} u_{ij}(t) & \text{if } j \in \mathcal{J} \\ x_{ij}(t) & \text{otherwise} \end{cases} \quad (3)$$

where $x_{ij}(t)$ refers to the j-th element of the vector $x_i(t)$,and J is the set of element indices that will undergo perturbation (or in other words, the set of crossover points).

3.2.3 Selection

Original vector or chromosome to be mutate is called target vector. Donor vector is achieved by selecting three random vectors and performing mutation operation on them. Trial vector is achieved by crossover operation of target vector and donor vector.

Selection is applied to determine which individuals will take part in the mutation operation to produce a trial vector, and to determine which of the parent or the offspring will survive to the next generation. To construct the population for the next generation, deterministic selection is used: The offspring replaces the parent if the fitness of the offspring is better than its parent; otherwise the parent survives to the next generation.

3.3 Clustering with Differential Evolution Algorithm

In order to apply the DE as clustering algorithm, clustering must be viewed as optimization algorithm. Two important design factors are need to solve optimization problem using DE, chromosomes representation and fitness calculation. In the context of data clustering, a single individual represents the K cluster centroids. That is, each individual cluster centroids are encodes as chromosomes in the DE algorithm. Therefore, a population represents a number of candidate clusterings. Quantization error is used as the fitness measure

$$J_e = \frac{\sum_{j=1}^{N_c}[\sum_{\forall \mathbf{z}_p \in C_{ij}} d(\mathbf{z}_p, \mathbf{m}_j)/|C_{ij}|]}{N_c} \quad (4)$$

Where J_e is the quantization error, N_c is the number of cluster, $d(z_p,m_j)$ is the Euclidean distance between data point m_j and center z_p. $|C_{ij}|$ is the number of member in cluster j .

Purity is used to measure the quality of two clustering algorithm. Let there be k clusters(the k in k-means) of the dataset D and size of cluster Cj be |Cj |.Let $|Cj|_{class=i}$ denote number of items of class i assigned to cluster j. Purity of this cluster is given by

$$\text{Purity } (C_j) = \frac{1}{|Cj|} \max_i (|Cj|_{class=i}) \quad (5)$$

The overall purity of a clustering solution could be expressed as a weighted sum of individual cluster purities.

$$\text{Purity} = \sum_{j=1}^{k} \frac{|Cj|}{|D|} \text{ purity } (Cj) \quad (6)$$

4. System Implementation

This section describes the design and implementation of the proposed system. Two algorithm for clustering (K-Mean and DE based clustering algorithm) are implemented and cluster quality on two algorithms were tested on the dataset downloaded from the UCI dataset. Proposed system is developed in Java programming language and Microsoft Access 2007 is used to store dataset. Figure 4.1 presents the flow of the DE clustering algorithm.

At the start of the algorithm, dataset were loaded from the database and they are represented as vector object that can be easily used by both k-mean and DE clustering algorithm. To start the DE clustering algorithm, the user must supply the number of cluster and iteration. The next step initialized the DE population with random chromosome. Each chromosome of the DE algorithm encodes a cluster solutions. DE used populations of individual chromosomes and these chromosomes are initialized with random centroid picked from the dataset.

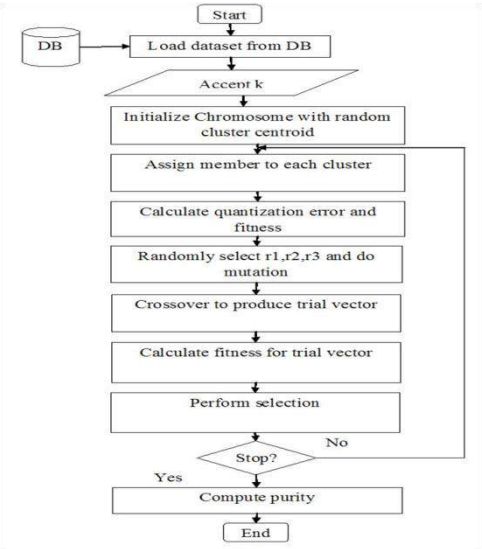


Figure 4.1 Flow of Data clustering using DE algorithm

The next step is to compute the fitness value of each chromosome. Fitness defined how well a chromosome solves the problem in hand. In this situation, for the clustering problem, fitness determines how well cluster solution which is encoded by the chromosome is. To compute the fitness, the proposed system used the quantization error, which is simply the sum of distance of each cluster data point with their centroid. So, the first step in calculating fitness is to assign the data point to centroid of each chromosome. Distance between each data point and cluster centroid in chromosome is computed using Euclidean distance. Data points are assigned to cluster cenroid with nearest distance. The second step computed the quantization error based on the cluster members and their centroid. Cluster solution is better when quantization error is smaller, so to convert to maximization problem, fitness is taken as 1/quantization error. So, the more fitness, the better cluster solutions is. Each individual chromosomes in DE algorithm is evaluated to compute its fitness.

The next step in DE is mutation which creates donar vector by selecting three random chromosomes and performing mutation operation as specified in section 3.2.1. Each chromosome undergoes mutation producing donar vector for each chromosomes. After mutation is applied to the DE algorithm, cross over is applied for each chromosomes to produce offspring trial vector as specified in section 3.2.2.
Selection step of DE compare the fitness of trial vector and original vector, chromosome with higher fitness survive to next generation. After all of these operation are completed, the next generation is formed and algorithm can now start next generation. DE algorithms typically run multiple generations to generate the solution. Throughout the multiple generations of DE algorithm, the best chromosome with best fitness is saved and it is used as the output of the algorithm. After DE is successfully finished, purity of the cluster solution output from DE algorithm is computed.

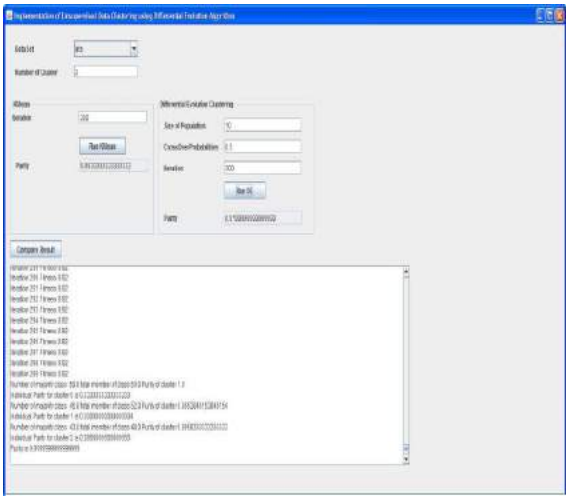


Figure 4.2 Clustering result form for both algorithms

Figure 4.2 presents the clustering result form for both algorithm, by using this form, the user can cluster data set in two algorithm and compare their result. In order to cluster, user must select the dataset from the combo box. For both algorithms, user must supply the number of cluster. For the k-mean algorithm, user must supply the number of iteration, for the DE clustering algorithm, user must supply size of population, cross over probabilities and number of generation to run DE. After the algorithm is finished the system will display the purity and cluster results .
If the user want to compare the purity using chart, it can be compare using the “Compare” button and the system will show bar chart comparing the purity of two algorithm.

For each run of k-mean algorithm and DE algorithm, the system stored related setting and their purities in the database allowing user to view the experimental result as shown in figure 4.3.

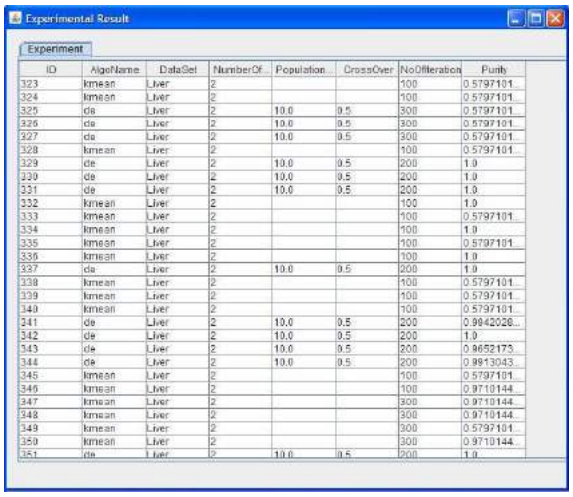


Figure 4.3 Form showing the experimental result

5. Experimental Result

This section presents the results of K-means and differential evolutional algorithms on five datasets namely Iris, wine, heart, liver and pima. These dataset are downloaded from the UCI machine learning repository. The main purpose is to compare the quality of respective clustering, where quality is measured according to purity. The datasets used for the purpose of this paper are:

Table 1. Dataset Information

Dataset	Num of records	Num of attributes	Attributes Values
Iris	150	4	Numerical
Wine	178	13	Numerical
Heart	303	14	Numerical
Liver	345	7	Numerical
Pima	768	8	Numerical

Table 2. Comparison of K-means and Cluster Ensemble

Data set	Num of clusters	Iteration	Purity of K-mean	Purity of DE
Iris	**3**	**300**	**0.820**	**0.908**
Wine	**3**	**300**	**0.8442**	**0.8478**
Heart	**2**	**300**	**0.867**	**0.9217**
Liver	**2**	**300**	**0.579**	**0.960**
Pima	**2**	**300**	**0.660**	**0.6682**

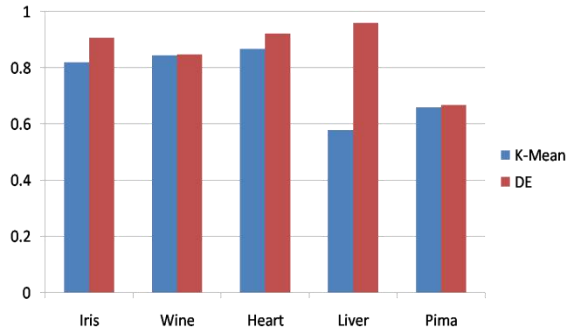


Figure 5.1 Comparing purities of two algorithms

Table1. describe the dataset information of used in testing.

Table 2. summarizes the results obtained from the two clustering algorithm. All the results reported are average value of 10 simulations. The number of cluster is not limited for two algorithms. The results of purity does not depend on number of clusters. The results show that cluster ensemble is robust and outperform than k-mean clustering algorithm on all of five dataset. For the cluster merging purpose, DE algorithm is better than K-mean algorithm.

Figure 5.1 presents the purities of two algorithms on five datasets. For Iris data set, the purity of K-mean is between 0.666 and 0.893 than the average purity is 0.820.The purity of DE is between 0.880 and 0.940, than the average purity is 0.908. So, DE can give better purity than K-means. For Wine data set, the purity of K-mean is between 0.839 and 0.859 , the average purity is 0.8442.The purity of DE is between 0.724 and 0.870, the average purity is 0.8478. For Heart data set, the purity of K-mean is 0.866 and 870 , the average purity is 0.867. The purity of DE is between 0.877 and 0.955, the average purity is 0.9217. For Liver dataset, the purity of K-mean is 0.579 . The purity of DE is between 0.924 and 1.0. The average purity is 0.960. For Pima dataset, the purity of K-mean is 0.660. The purity of DE is between 0.651 and 0.686, the average purity is 0.6682.

Among five datasets, the liver dataset is most appropriate for DE algorithm. The liver dataset has noisy data. DE algorithm can ignore the noisy data.

6. Conclusion

This paper presented the implementation of a differential evolution based clustering algorithm compares the results of algorithm with K-means algorithm on UCI datasets. According to the experimental result, DE algorithm outperforms the k-mean algorithm in all of the dataset. DE algorithm is robust then k-mean clustering algorithm and can

achieve better cluster quality but required more computational requirement than k-mean clustering algorithm. So, DE algorithm is suitable when robust clustering or better cluster solution is needed and it can also be used for distributed clustering purpose due to their parallel nature.

References

[1] Sandra Paterlini, Thiemo Krink, High Performance Clustering with Differential Evolution, Proceedings of Congress on Evolutionary Computation (CEC-2004) (vol. 2). Piscataway, NJ: IEEE PRESS. ISBN: 9780780385153

[2] Swagatam Das,Ajith Abraham, Automatic Clustering Usingan Improved Differential Evolution Algorithm , IEEE Transcations On Systems,Man, And Cybernetics—PARTA:SYSTEMS AND HUMANS,VOL.38,NO.1,JANUARY2008

[3]D.Zaharie, F.Zamfirache, V.Negru, D.Pop ,and H.Popa, A Comparison Of Quality Criteria For Unsupervised Clustering of Document Based On Differential Evolution, KNOWLED GEENGINEERING: PRINCIPLES AND TECHNIQUES . Proceedings of the International Conference on Knowledge Engineering, Principles and Techniques, KEPT 2007 .Cluj-Napoca(Romania),June6{8,2007,pp.25{32

[4] G Sudhakar, Polinati Vinod Babu, Suresh Chandra Satapathy, Gunanidhi Pradhan, Effective Image Clustering with Differential Evolution Technique, Int. J. of Computer and Communication Technology, Vol. 2, No. 1, 2010

[5] Yanfei Zhong, Shuai Zhang, Liangpei Zhang, Automatic Fuzzy Clustering Based on Adaptive Multi-Objective Differential Evolution for Remote Sensing Imagery. IEEE Journal of Selected Topic in Applied Earth Observation and Remote Sensing.

[6]R.Storn and K.Price, “Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces,” J.Global Optim.,vol.11,no.4,pp.341–359,1997

Information Retrieval System using Reciprocal Fusion with BM25 and Cosine Similarity

Nan Wint Yee Myint, Phyu Hninn Myint
University of Computer Studies, Yangon, Myanmar
nanwintyeemyint@ucsy.edu.mm, phm.ucsy@ucsy.edu.mm

Abstract

Nowadays, we have to face a variety and large number of paper information and often waste time much on searching the needs of user. They want to search effective and efficient information that they are needed. Information Retrieval (IR) is a very useful way for searching. Thesis candidates and supervisors have to spend so much time for searching large number of Thesis Titles on the shelves of the Library. This thesis intends to implement effective keyword search system for Master and Ph.D. thesis titles of UCSY Digital Library by using multiple ranking methods: Cosine similarity, Okapi (BM25), and Reciprocal fusion. The most relevant documents are ranked by using Cosine Similarity ranking, Okapi (BM25) ranking and then fusion with Reciprocal Ranking.

Keywords: digital library, cosine similarity, Okapi (BM25) ranking method, Reciprocal Fusion, Indexing, Ranking, Information retrieval

1. Introduction

Information retrieval (IR) is the process of retrieving information or documents that contain information which is relevant to the given query from data collections. It can search large information collection and return the relevant information to user's information needs. The efficient searching methods can provide for proper and relevant information to the user. In query processing, the user enters a query and the system finds the relevant document, ranked the documents and displays to user. Ranking of query results is one of the fundamental problems in information retrieval (IR), behind search engines. An efficient IR system for UCSY digital library will be developed by applying Inverted Index, Cosine Similarity ranking, Okapi(BM25) ranking and Reciprocal ranking fusion.

2. Information Retrieval

Information Retrieval (IR) is the science of information within relational databases, documents, text, multimedia files, and the World Wide Web [4]. Many users are engaged in the IR field especially reference librarians, professional researchers, political analysts, governmental investigators, and market forecasters. The applications of IR are diverse; they include but not limited to extraction of information from large documents, searching in digital libraries, information filtering, spam filtering, object extraction from images, automatic summarization, and document classification and clustering, and web searching. Information retrieval plays a vital role in this information age. The most important type of evaluation of IR system is retrieval effectiveness evaluation. Retrieval effectiveness evaluation measures how well a given system or algorithm can match, retrieve and rank documents that are relevant to user's information need. Information retrieval system is typically composed of indexing and querying system. Indexing is process of storing the term and term list in the computer for effective retrieval. The most widely used technique for indexing is the inverted index. In document organizing or indexing process, the documents are preprocessed and stored in computers suitable for the efficient query processing by using a data structure such as inverted index. In query processing, the user enters a query and the system finds the relevant document, ranked the documents and displays to user. The purpose of the ranking system is to retrieve from the collection of documents the most relevant ones in response to a query. This is done by computing, for each matching document, a score which should reflect the relevance of that document with respect to the given query. Then, the documents are sorted descending according to the computed score, and only a subset of them is returned as a result set. This system focused on retrieving master thesis titles by Cosine similarity and Okapi (BM25) ranking methods, and then fusion with reciprocal ranking method to represent the most

relevant information to the users. By using this proposed system the user can get more accurate information by searching thesis titles of UCSY.

Information Retrieval models

There are four main Information Retrieval models:

Boolean Model

The Boolean Model is the earliest and simplest information retrieval models. It uses the notion of exact matching to match documents to the user query i.e., (using Boolean Query AND, OR, NOT). Both query and retrieval are based on Boolean algebra. In the Boolean Model, documents and queries are represented as a set of terms. That is only considered present or absent in a document.

Vector Space Model

Vector Space Model is the best known and most widely used IR model. A document in IR model is represented as a weight vector, in which each component weight is computed based on some variation of term-based features. That model can rank documents by adopting an inexact strategies (documents are ranked according to the value of predefine similarity measure).

Statistical Language Model

Statistical Language Models (or simply language models) are based on probability and have foundation in statistical theory. This basis idea of this approach to retrieval is simple. Documents and queries are represented as terms in this model. It first estimate a language model for each document, and then ranks document by the likelihood of the query given the language model. Statistical Language Models are used in most works based on unigram, i.e., only individual terms (words) are considered

Probabilistic Model

In Probabilistic Model, documents and queries are represented as terms to evaluate the probability of relevance. It is used to rank documents according to their estimated probability of being relevant. The Probabilistic Model is able to cope with the uncertainty of the retrieval process**.**

3. System Overview

When the user enters a query, preprocessing steps are required prior to the information retrieval process. The preprocessing steps include tokenization, stop words removal and stemming. The documents that is stored in computers suitable for the efficient query processing by inverted indexing. The concept of processing the original data into a highly efficient cross reference lookup in order to facilitate rapid searching is called indexing [11]. To rank matching documents according to their relevance to a given search query, it is necessary to assign a numerical score to each document based on BM25 and Cosine Similarity ranking function which incorporates features of the document, the query and a document collection. BM25 and Cosine Similarity rank results are fused with Reciprocal Rank Function which yields better results than individual. Reciprocal rank is the inverse of the rank of first correct answer. The final scores of reciprocal rank results will be displayed to the user. The block diagram of the proposed system is shown in Fig1.

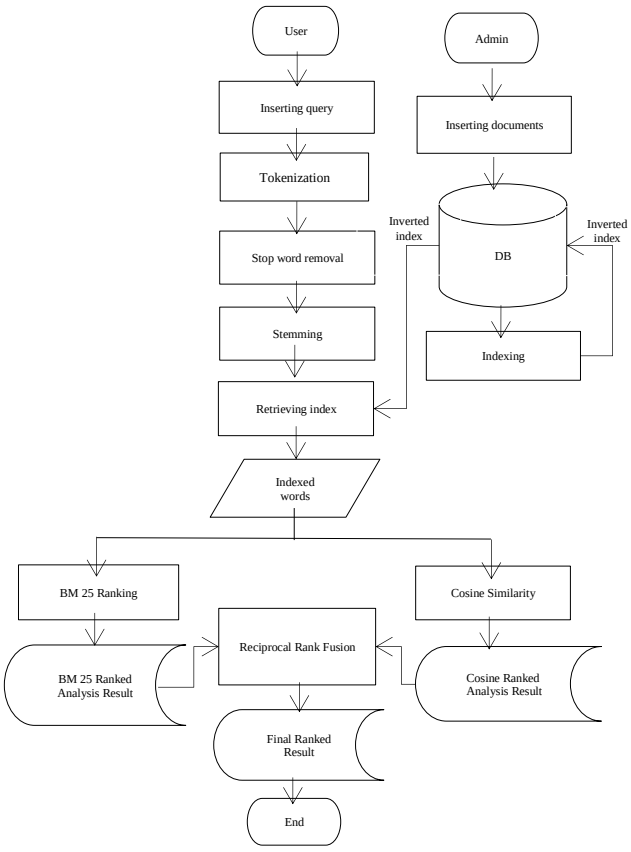


Figure.1 System Overview

4. Preprocessing

4.1 Tokenization

Tokenization is the process of breaking a stream of text into words, phrases, symbols or other meaningful elements called tokens. The list of tokens becomes

input for further processing such as parsing or text mining. In tokenizing task, each string is cut into individual word. The results of tokenizing words in the document and query information are change into lower case.

e.g., Decision support system for attending career education school

=>decision support system for attending career education school

4.2 Stop words Removal

The words that are frequently occurring but meaningless in terms of information retrieval are called stop words. By filtering those words out before running the processing part on the data the runtime will go down, there will used less space and the similarity will be more precise. The removal of stop words usually improves IR effectiveness. A stop word list contains the list of words to ignore when indexing the document collection such as preposition, articles, pronouns, some adverbs and adjectives and some frequent words.

e.g., decision support system for attending career education school

=>decision support system attending career education school

For removal of stop words we used the PubMed stop words list as shown below.

4.3 Stemming

Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form – generally a written word form. The process of stemming is often called conflation. The stemming is widely used in information retrieval tasks. Many researchers demonstrate that stemming improves the performance of information retrieval systems. Porter Stemmer is the most common algorithm for English stemming [11]. However, this stemming algorithm has several drawbacks, since its simple rules cannot fully describe English morphology. Errors made by this stemmer may affect the information retrieval performance. Thus, this system used customized Porter Stemming.

e.g., retrieval => retrieve

5. Inverted Index

The inverted index is central to the first major concept in information retrieval. In information technology, an inverted index is an index data structure storing a mapping from content. The inverted index for a document collection consists of a set of so-called inverted lists, known as posting lists. Each inverted list corresponds to a word, which stores all the IDs of documents where this word appears in ascending order [1].

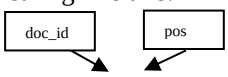
Inverted Indexing Pseudo-code

```
class MAPPER (doc_id n, doc d)
H → new ASSOCIATIVE ARRAY
for all term t ∈ doc d do
H{t} → H{t}+1
for all term t ∈ H do
EMIT (term t, posting <n, H{t}>)

class REDUCER (term t, postings [<a1,f1>, <a2,f2>…])
P → new LIST
for all postings <a,f>, ∈ postings [<a1,f1>, <a2,f2>…….] do
APPEND (p, <a,f>)
SORT(P)
EMIT (term t, posting P)
```

Inverted Index: Structure

An inverted index consists of a collection of postings lists, one associated with each unique term in the collection. Each postings list consists of a number of individual postings. Each posting holds a document identifier (doc_id) and the position of the term in that document. The complete inverted index would look something like this:



decision: {132, 0}
support: {136, 3} {214, 4} {231, 2}
attend: {40, 2}{4, 4}
career: {4, 5}
educ: {4, 6} {119, 5} {156, 7} {270, 3} {308, 6}
school: {4, 7}

6. Ranking

6.1 Cosine Measure

Cosine Similarity is used to quantify similarity values between objects. Cosine Measure ranking uses the weight of word (term) in document or query. A measure of the total weight or length of document or query is calculated in terms of the weight and number of words in query. The theory is that the more similar the query and the document are to each other, the better the document is as an answer to the query. The

answers to a ranked query are the documents with the highest values.

$$\text{Cosine (q, d)} = \frac{\sum_t w_{t,q} * \sum_t w_{t,d}}{\sqrt{\sum_t w^2_{t,q}} * \sqrt{\sum_t w^2_{t,d}}}$$

$w_{t,\,q}$= weight of term in query

$w_{t,\,q}= f_{t,\,q} * \log (N/f_t)$

$f_{t,\,q}$= number of occurrence of term in q

N= number of documents in the collection

f_t= number of documents containing term

$w_{t,\,d}= f_{t,\,d} * \log(N/f_t)$

$f_{t,\,d}$= number of occurrence of term in document

Pseudo-code for cosine similarity algorithm:

COSINESCORE (q)

1. float Scores[N]=0
2. float Length[N] =0
3. **for each** query term t
4. **do** calculate $W_{t,q}$ and fetch posting list for t
5. **for each** pair(d,$tf_{t,d}$) in posting list
6. Scores[d]+=$W_{t,d} \times W_{t,q}$
7. Read the array Length
8. **for each** d
9. **do** Scores[d] = Scores[d] / Length[d]
10. **return** top K components of Scores[]

Sample Dissertations

Title1: implement inform retrieve acean novel force using vector space model

Title2: implement inform extract system job post pattern base approach

Title3: implement inform retrieve vector space model

Title4: inform retrieve e-document genetic algorithm

Title5: compare study inform retrieve system vector space model boolean search model

Title6: implement library system vertical fragment method distribute database

Title7: text base image retrieve system latent semantic index (lsi)

Title8: develop inform system invert index

Title9: implement inform retrieve system digit library latent semantic index (lsi)

Title10: hazard inform retrieve automatic keyword extract

Query: *inform retrieve system*

Table [1] Ranking results of Cosine Similarity Measure

	inform	retrieve	system	Score
Title 1	0.535251874	0.15490196	0.000000	0.632161884343
Title2	0.22184874	0.000000	0.313294346	0.535143086208
Title3	0.434611	0.434611	0.000000	0.869222590971
Title4	0.496699	0.49669	0.000000	0.993397246824
Title5	0.376749	0.154901	0.221848	0.695660494501
Title6	0.000000	0.000000	0.316466062703	0.316466062703
Title7	0.000000	0.347958699	0.347958699	0.695917398654
Title8	0.404803785	0.000000	0.434803758	0.869607515088
Title9	0.535251874	0.15490196	0.112531	0.802685185962
Title10	0.234611	0.634611	0.000000	0.869222590971

Ranked result:

6. Title 4
7. Title 8
8. Title 10
9. Title 3
10. Title 9

1. Title 7
2. Title 5
3. Title 1
4. Title
5. Title 6

6.2 Okapi BM25

In information retrieval, Okapi BM25 (BM stands for Best Matching) is a ranking function used by search engines to rank matching documents according to their relevance to a given search query. Stephen E. Robertson, Karen Sparck Jones, and others developed this algorithm in 1970 and 1980.Okapi BM25 is one of the best-known term weighting and document ranking functions. It is a probabilistic model of information retrieval. Probability assumes that each document has a binary relevance attribute stating if this document is relevant or not. The measure itself is a probabilistic value of this binary relevance. Ordering the document by their decreasing probability is assumed to give the best result the given document collection can provide. This ranking function is actually named asBM25, but it is known as Okapi BM25 because the Okapi information retrieval system was the first one to implement this ranking function. BM25 will rank a set of documents depending upon the appearance of the query terms in each document. For the calculation of BM25, we use TF (Term Frequency) and IDF (Inverse Document Frequency) functions.BM25 use inverse document frequency to distinguish between common (low value) words and uncommon (high value) words. Both also recognize (see Term frequency) that the more often a word appears in a document, the more likely is it that the document is relevant for that word. **[**3].

The Term Frequency (TF) is calculated using following formula

$$\text{TF}_{(t,d)} = \frac{(k_1+1) * t_f(t,doc)}{k_1\left(1-b+b*\frac{len(doc)}{avdl}\right)+t_f(t,doc)}$$

The Inverse Document Frequency (IDF) is calculated by using the following formula

$$\text{IDF(t)} = \log \frac{\text{number of documents in document collection}}{\text{number of documents with term}}$$

The BM25 scoring function is defined as

$$\text{Score (q, d)} = \sum_{t \epsilon Q} \text{IDF(t)} * \text{TF(t, d)}$$

$TF_{(term, doc)}$=frequency of term in document

$t_f(t, doc)$=frequency of term t appearing in document doc

$len(doc)$=length of document

$avdl$= average length of document in collection

$avdl = \frac{Total\ length\ of\ documents}{number\ of\ documents\ in\ database}$

k_1=1.2is a parameter that tunes the scaling of term frequency,

b=0.75is a parameter that tunes the scaling of document length

Pseudo-code for Okapi BM25algorithm:

Require: Query q, and Document Collection

BEGIN

1: Score[N] = 0:0, k1=1.2,b=.75
2: for all term t in query q do
3: for all document d_i in collection do
4: w_t=TF*IDF
5: avdl=$Totallengthofdocuments/ number\ of\ \ documents\ with\ term$
6: Score[i] += (wt*(k1+1)*tf(t,d))/(k1(1-b+b*len/avdl)+tf(t,d))
7: end for
8: end for
9: SORT (Score[])
10: return Score[]

END

Table [2] Ranking results of Okapi BM25 Measure

	information	retrieval	system	Score
Title1	0.096910	0.467224	0.000000	0.564134268031
Title2	0.096910	0.000000	0.590636	0.687546155739
Title3	0.096910	0.483209	0.000000	0.638110106966
Title4	0.1549	0.48321	0.000000	0.638110106966
Title5	0.096910	0.154901	0.667267	0.919077565525
Title6	0.000000	0.000000	0.497008	0.49700767001
Title7	0.000000	0.154901	0.723061	0.877961717918
Title8	0.096910	0.000000	0.74777	0.84467960931
Title9	0.096910	0.154901	0.733203	0.985013779871
Title10	0.1549	0.48321	0.000000	0.638110106966

Ranked result:

1. Title 9
2. Title 5
3. Title 7
4. Title 8
5. Title 2
6. Title 4
7. Title 3
8. Title 10
9. Title 1
10. Title 6

6.3 Reciprocal Rank

Reciprocal Rank is evaluated for any process that retrieves a list of response to a query ordered by probability of correctness. Reciprocal rank is the inverse of the rank of first correct answer. Reciprocal Rank is the popular rank combination methods that use only rank positions of each search engine. Reciprocal Rank is one of data fusion methods [7]. A data fusion algorithm accepts two or more ranked lists and merges these lists into a single ranked list with the aim of providing better effectiveness than all systems used for data fusion. Reciprocal rank of the ranked result set of a query is defined to be the reciprocal value of the rank of the highest ranking relevant document in the result set. The reciprocal rank is set to be 0 if none of the highest ranking document is relevant to query. Assign a score 1/pos to each doc. Rank based on sum of scores. The higher the reciprocal rank of the query, the better the ranking is. RRF rule involves summing the reciprocal ranks for those p occurrences to give a fused score.

$$\text{Score(doc, Q)} = \sum_{n=1}^{p} \left(\frac{1}{rank(doc)}\right)$$

RANK (doc) = rank documents obtained when the similarity scores are sorted into descending order

p ($p \leq n$) = number of times that an individual documents occurs.

Using Reciprocal Ranking: The rank result of each document is as follows:

Score (Title1) = 1/9 + 1/8 = 0.236111111
Score (Title2) = 1/5 + 1/9 = 0.311111111
Score (Title3) = 1/7 + 1/4 = 0.392857143
Score (Title4) = 1/6 + 1/1 = 1.166666667
Score (Title5) = 1/2 + 1/7 = 0.642857143
Score (Title6) =1/10 + 1/10 =0.2
Score (Title7) = 1/3 + 1/6 = 0.5
Score (Title8) = 1/4 + 1/2 = 0.75
Score (Title9) = 1+ 1/5=1.2
Score (Title10) = 1/8 + 1/3 = 0.458333333

Ranked result:

1. Title 9
2. Title 4
3. Title 8
4. Title 5
5. Title 7
6. Title 10
7. Title 3
8. Title 2
9. Title 1
10. Title 6

7. Performance Analysis

The performance of the proposed system can be measured by computing its efficiency and its effectiveness. So to measure the effectiveness of retrieval method, we use two standard measures named as Recall & Precision.

Precision is defined as the fraction of the result set which is relevant. This measures how many documents in the result set are relevant to input query. (i.e. "correct response")

Precision is computed as:

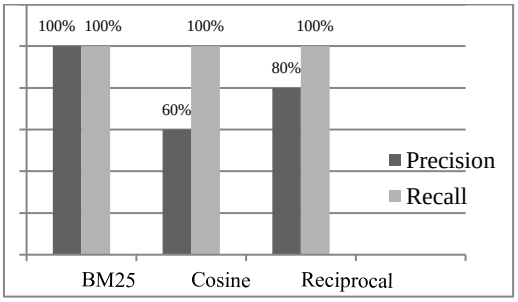
$$\text{Precision}=\frac{|\{Relevant\}\cap\{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|}$$

Recall means the fraction of relevant documents which appear in the result set. It is the measure of number of relevant documents returned by search engine from corpus.

Recall is computed as:

$$\text{Recall}=\frac{|\{Relevant\}\cap\{\text{Retrieved}\}|}{|\{\text{Relevant}\}|}$$

In order to evaluate the system, document collections with 2350 Thesis Titles are used. The evaluation of Cosine Similarity, BM25 and Reciprocal fusion results is shown in Table 3. The result of recall and precision are tested on 400 different types of input query.



Table[3] The ealuation Result

8. Conclusion

Information retrieval is the study of helping user to find information that matches their information needs. In the proposed system, multiple IR methods are applied for digital library to provide efficient searching method to users. This system will search dissertation keywords for thesis titles of UCSY digital libraries by using cosine similarity and Okapi (BM25) ranking methods, and then fusion with reciprocal ranking method. The system can be practicably useful in the digital library of our university and can effectively support thesis candidates and supervisors to get their requirement.

9. References

[1]. Ahmad T. Al-Taani, Ahmed S. Ghorab and Hazem M. Al-Najjar, "An Arabic-English Indexing System Using Inverted Index Algorithm", Department Of Computer Science, Yarmouk University.

[2]. Christian Hausler, "Methods for Determining the Similarity of Documents".

[3]. C.N.N.Kyi, "Development of Information Retrieval System using Inverted Indexing", 5th Parallel and Soft Computing Conference, Yangon, Myanmar, March 2010.

[4]. Doyle Lauren, Joseph Becker, "Information Retrieval and Processing", Melville, 1975.

[5]. Keerthiram Murugesan Lexington, "Cluster-Based Term Weighting and Document Ranking Models", University of Kentucky.

[6]. Marios Hadjieleftheriou , Amit Chandel, Nick Koudas, Divesh Srivastava, AT&T Labs–Research Florham Park, "Fast Indexes and Algorithms for Set Similarity Selection Queries", Department of Computer Science, University of Toronto.

[7]. Peter Willett, Peter Willett, "Fusing similarity rankings in ligand-based virtual screening".

[8]. Shailesh Padave, "Incorporating WordNet in an Information Retrieval System", May 2014.

[10].Thomas Roelleke and Marco Bonzanini and Miguel MartinezAlvarez, "On the Modeling of Ranking Algorithms in Probabilistic", Queen Mary University of London, Mile End Road.

[11]. Wahiba Ben Abdessalem Karaa, "A New Stemmer To Improve Information Retrieval", International Journal of Network Security & Its Applications (IJNSA), Vol.5, No.4, July 2013, University of Tunis,Higher Institute of Management, Tunisia.**[**12]. International Journal of Scientific and Research Publications, Volume 2, Issue 11, November 2012 3 ISSN 2250-3153.

Breast Cancer Classification with C4.5 Decision Tree and Weighted C4.5 Decision Tree Approach

Khin Thuzar Win, Aung Nway Oo
University of Computer Studies, Yangon
khinthuzarwin87@gmail.com, aungnwayoo78@gmail.com

Abstract

Data mining techniques is increasing becoming on medical data for discovering useful trends and patterns that are used in diagnosis and decision making. Classification is a data mining technique which addresses the problem of constructing a predictive model for a class attribute given the values of other attributes and some examples of records with known class. In this paper, we implemented the weighted C4.5 decision tree algorithms for Breast Cancer classification. Naïve Bayesian theorem was used to calculate the weight value to set the appropriate weights of training instances before trying to construct a decision tree model. In this work, the comparative analysis of weighted C4.5 decision tree algorithm and traditional C4.5 decision tree algorithm for diagnosis of Breast Cancer Datasets was also performed.

Key words: Data mining (DM), Classification, Decision Tree (DT), C4.5

1. Introduction

The term 'data mining' is devised to refer to the action of moving through large databases investigating appealing and new patterns. Data mining has become considerably important and a necessity today when data are bountiful and easily accessible. The automatic analysis of large numbers of data is possible through the methods and instruments that the field of data mining provides. Data mining is one aspect of the process of Knowledge Discovery in Databases (KDD). Some searchers think if data mining as an ambiguous expression and uses the term "Knowledge Mining" as it bears a better resemblance to gold mining. Data mining approach are mostly grounded on inductive learning i.e., constructing a mode explicitly or implicitly by forming a generalization from enough training examples. The inductive approach forms a basic assumption that the trained model is related to future unseen examples. Specifically, any form of conjecture is considered an induction on conditions that conclusions are not logically drawn from premises.

Data collection was conventionally accepted as one pivotal period in data analysis. An analyst would be able to select the variables to be collected by the application of the available domain knowledge. The number of specified variables was usually restricted and their values could be recorded by hand or using oral interviews. If computer-aided analysis was to be used, the collected data had to be entered into statistical computer package or an electronic spreadsheet. Because the process of data collection was expensive, analysts had to learn to make decisions on available information. Decision trees are regarded as well-known methods for representing classifiers. A decision tree is a classifier viewed as the repetitive subdivision of the instance space.

The decision tree is composed of nodes forming a 'rooted tree' i.e., a 'directed tree' with a node known as 'root' with no incoming edges. There is exactly one incoming edge in all other nodes. An internal node is a node with outgoing edges. All other nodes are known as leaf node. In a decision tree, it is each internal node subdivides the instance space into two or more sub-spaces by an assured discrete function of the input attributes values. Simply and most frequently, each test takes a single attribute such that the attribute's values subdivide the instance space. On the other hand, the leaf may grip a probability vector that indicates the probability of the goal attribute having a definite value. Instance, from the root of a tree to a leaf, are navigated and organized, following the outcome of the tests along the path. There have been many decision tree algorithms like ID3 [1], C4.5 [2], CART [9] etc.

Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Classification has been successfully applied to wide range application areas, medical diagnosis, weather prediction, credit approval, customer segmentation, fraud detection among the different proposals. Classification is clearly useful in many decision problems, where for a data item a decision is to be made (which depend on the class to which data item belongs).

The rest of the paper is organized as follows. Section 2 reviews the related work and section 3 presents the background theory and section 3.1 presents the overview C4.5 algorithm. Naïve Bayes theorem and weighted C4.5 algorithm were described in section 3.2 and 3.3. Overview of the system flow was illustrated in section 4. Description of dataset is presented in section 5. The experimental results are presented in section 6. Finally, conclude of this study was provided in section 7.

2. Related Work

There are many research works that proposed efficient decision tree for classification. Kohavi & John [8], who researcher of parameter settings of C4.5 decision trees made a result in optimal performance on a

particular data set. Badr Hssina et.al [9] proposed the comparative analysis of ID3 and C4.5 decision tree algorithms. Liu Yuxun and XieNiuniu [10], decision tree algorithm based on attribute importance was proposed. Gaurav & Hitesh [11], proposed C4.5 algorithm which is improved by the use of L'Hospital Rule, this simplifies the calculation process and improves the efficiency of decision making algorithms. S.VijayaRani et.al [12] the authors analyzed the performance of C4.5, RIPPER and PART algorithm. Time and Number of rules produced were provided as the measures to analyze the data for Breast cancer data and heart disease data. Dewan Md. Faraid and Chowdhury [6] proposed the method for assigning weight value to training instances to increase the classification accuracy. In this paper, comparative studies of weighted and normal C4.5 algorithms are made to approximation the breast cancer dataset.

3. Background Theory

Classification can be used as in the form of data analysis that can be used to extract models describing important data classes. Classification can be used for making intelligent decision. In this study, weighted C4.5 algorithm was used for efficient classification. Breast cancer data set was used for testing of proposed method and compares the results of normal C4.5 algorithm.

3.1. C4.5 Algorithm

The C4.5 algorithm is the modified version of ID3 algorithm and which choose splitting attributes from a dataset with the maximum information gain.

The attribute with the maximum gain ratio is selected as root node or the splitting attribute. The expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \qquad (1)$$

Let p_i is the probability that an arbitrary tuple in D belongs to class C_i and m is the quantity in class label. The log function to the base 2 is used, because the information is encoded in bits. The information is based on the proportions of tuples of each class.

Information needed (after using attribute A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \quad (2)$$

where $Info_A(D)$ is the expected information of each attribute in data D and v is types of data in that attribute. Information gained by branching on attribute A.

$$Gain(A) = Info(D) - Info_A(D) \quad (3)$$

In other words, *Gain(A)* tells how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A.

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2(\frac{|D_j|}{|D|}) (4)$$

where SplitInfo(A) is the expected split information of each attribute in data D and v is types of data in that attribute. The attribute that yields the largest Gain Ratio is chosen for the decision node. The attribute with the maximum gain ratio is selected as the splitting attribute

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \qquad (5)$$

Gain Ratio for each attribute may be computed by equation 5. The attribute that yields the largest Gain Ratio is chosen for decision node. For building decision trees of C4.5 algorithm [14]:

Algorithm: Generate Decision Tree by C4.5
Input: Dataset D, attribute_list
Output: Tree
Begin
  Check for the base cases.
  For each attribute *a* in attribute_list
    Find the normalized information gain from splitting on *a*
    a_best be the attribute with the maximum normalized information gain.
  Create a decision *node* that splits on *a_best*.
  Recur on the sublists gained by splitting on *a_best*, and add those nodes as children of *node*.
End

3.2. Naïve Bayes Theorem

Naïve Bayesian (NB) classifier is a simple probabilistic classifier based on probability model, which can be trained very competently in a supervised learning [3-4]. The naïve Bayesian classifier, or simple Bayesian classifier [5], works as follows:

1. Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, $X = (x_1, x_2, \ldots, x_n)$, illustrating n measurements made on the tuple from n attributes, respectively, $A_1, A_2, \ldots, A_n$.
2. Suppose that there are m classes, $C_1, C_2, \ldots, C_m$. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class C_i if and only if

$$P(C_i|X) > P(C_j|X) \quad for \ 1 \le j \le m, j \ne i. \qquad (6)$$

Thus we maximize $P(C_i|X)$.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad (7)$$

where $P(C_i|X)$ is the posterior probability or the probability that the value, $P(C_i)$ is the probability class based on the hypothesis, $P(X|C_i)$ is the predictor

probability based on the given class. $P(X)$ is a predictor probability.

3. As $P(X)$ is constant for all classes, only $P(X|C_i)P(C_i)$ necessity be highest. If the class prior probabilities are not known, then it is commonly presumed that the classes are equally likely, that is, $P(C_1) = P(C2) = \ldots = P(C_m)$, and we would therefore maximize $P(X|C_i)$. Otherwise, we maximize $P(X|C_i)P(C_i)$. A simplified assumption: attributes are conditionally independent (i.e., no dependence relation between attributes):

$$W_i = argmax\ P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad (8)$$

The naïve Bayesian classifier is sample to use and efficient to learn. It requires only one scan of the training data. Despite the fact that the independence assumption is often violated in practice, naïve Bayes often competes well with more sophisticated classifiers. In other words, the predicted class label is the class Ci for which P(X|Ci) P(Ci) is the maximum [4].Weight value for each attribute is calculated by equation 8 which is the maximum weight value.

3.3. Weighted C4.5 Algorithm

Weighted decision tree learning algorithm was developed by assigning appropriate weights to training instances, which improve the classification accuracy. The weights of the training instances are calculated using maximum posteriori hypothesis of Naïve Bayesian theorem. Weight of each training instance is calculated with the maximum value of the class conditional probabilities.

Weighted C4.5algorithm chooses splitting attributes from a dataset with the maximum information gain by using these weights value and constructs the decision tree model for Breast Cancer classification. Given a training dataset, the weighted C4.5 algorithm initializes the weights of each training instance, *Wi* by maximum likelihood of posterior probability by assigning weights of training dataset in D. This algorithm uses the weight value calculated from Naïve Bayes probabilistic model to initialize the weights of each training instance.

The expected information needed to classify a tuple in dataset D is calculated by applying equation (1). In this case, p_i is the relative frequency of class *i* in *D*, *where* , p_i is the probability that an arbitrary tuple in *D* belongs to class C_i and m is the quantity in class label. The log function to the base 2 is used, because the information is encoded in bits. The information is based on the proportions of tuples of each class. The sum is computed over *m* classes.

To determine the information required to classify D, we examine all the possible subsets that can be formed using known values of attribute A. When considering a split, we calculate a weighted sum of the impurity of each resulting partition. And then $Info_A(D)$ is calculated by applying equation (2). In this time, the value of equation (2) is defined as follows:

$|D_j|$ = the set of tuple with weight value in training dataset that have outcome aj of attribute,

$|D|$ = total weight value tuple

Information gain is defined as the dissimilarity between the original information requirement (i.e., based on just the proportion of classed) and the new requirement (i.e., gained after partitioning on A) by using equation (3) and gain ratio to overcome the problem by using equation (4) and equation (5). We are calculated InfoA(D), Gain(A) ,SplitInfoA(D) and GainRatio to assign weight value.

The decision tree is built established on the weights of training data which results from naïve Bayes probabilities.

4. System Flow of Proposed System

The system flow for classification of breast cancer dataset with weighted C4.5 algorithm was described in the following figure.

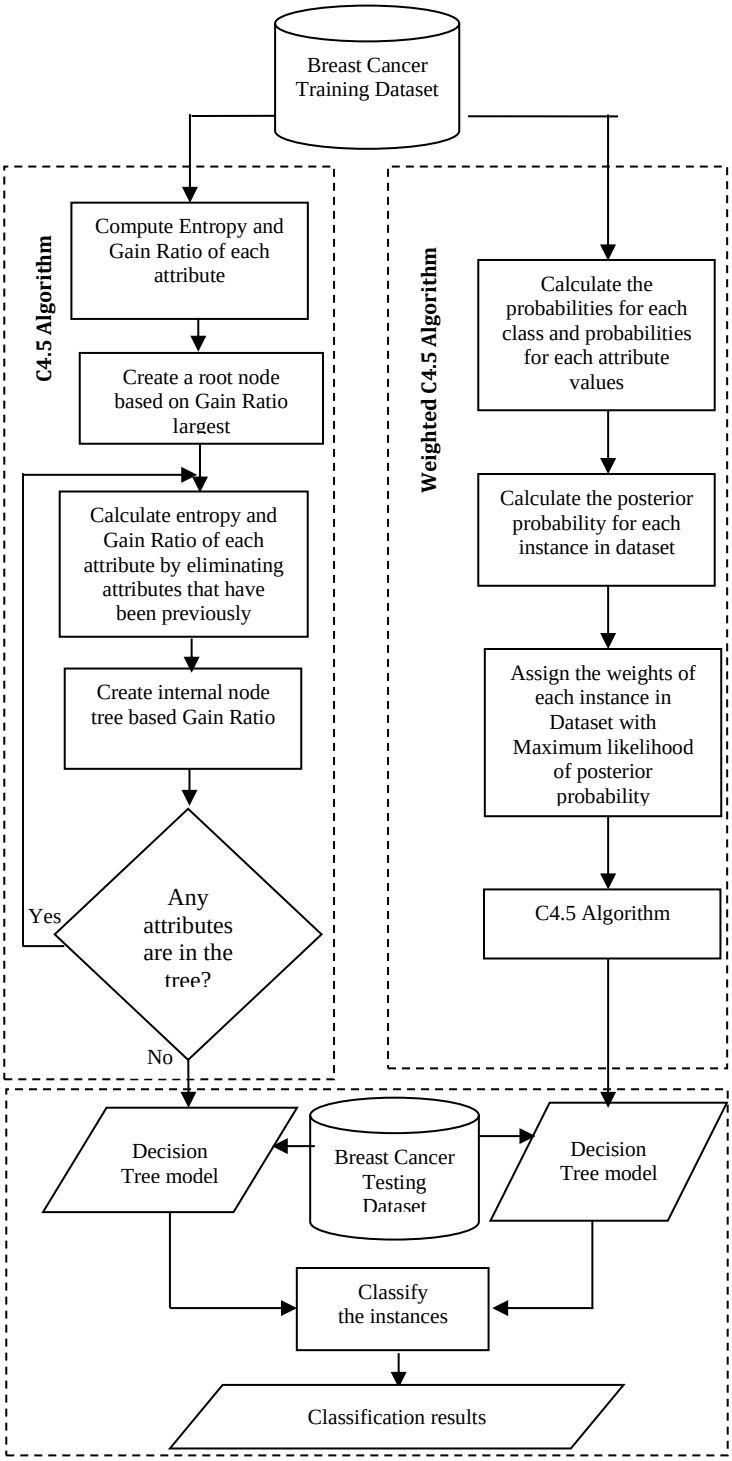


Figure 1. Overview of the Proposed System

5. Dataset Description

The breast cancer dataset contains 683 instances and 10 attributes. Each of the characteristics is assigned a value from 1 to 10 by the pathologist. The larger the value of attribute the greater the likelihood of malignancy.

The following table lists the attribute information of breast cancer dataset.

Table 1. Dataset Description

ID	Attribute Name	Value
A1	Clump Thickness	1 – 10
A2	Uniformity of Cell Size	1 – 10
A3	Uniformity of Cell Shape	1 – 10
A4	Marginal Adhesion	1 – 10
A5	Single Epithelial Cell Size	1 – 10
A6	Bare Nuclei	1 – 10
A7	Bland Chromatin	1 – 10
A8	Normal Nucleoli	1 – 10
A9	Mitoses	1 – 10
A10	Class	Benign(C1), or malignant(C2)

There are two types of classes in dataset, benign (It does not invade nearby tissue or spread to other parts of the body), or malignant (It is serious and likely to spread other parts of the body).

The following figure described the sample decision tree of Breast Cancer detection. The figure is illustrated by using attribute id.

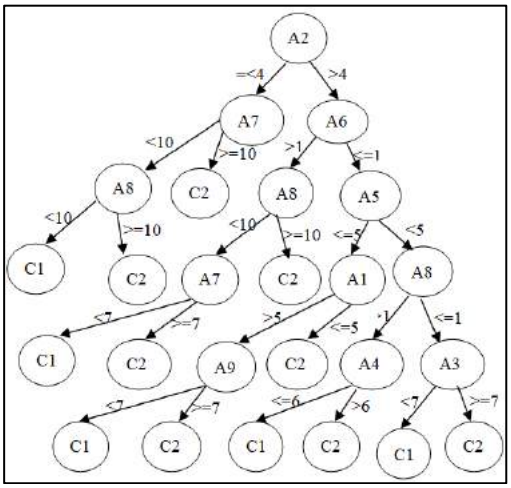


Figure 2. Sample Decision Tree for Breast Cancer Classification

6. Experimental Results

The main aim of this research is to analyze weighted C4.5 decision tree and traditional C4.5 decision tree algorithm. The breast cancer dataset from UCI [7] is used for comparative analysis. For each classifier, 2/3 of the dataset is used for training and 1/3 of datasets is used for testing.

The results of the classifiers in detecting the breast cancer were evaluated by using following equations.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

$$F-measure = \frac{2\times Recall\times Precision}{Precision+Recall} \quad (12)$$

Biometric evaluation system that assigns all authentication attempts a 'score' between closed interval [0, 1]. 0 means no match at all and 1 means a full match.

False Acceptance Rate (FAR) is calculated as follows:

$$FAR = \frac{FP}{(FP+TN)} \quad (13)$$

False Rejection Rate (FRR) is calculated as follows:

$$FRR = \frac{FN}{(TP+FN)} \quad (14)$$

Table 1. Biometric Evaluation for C4.5 and Weighted C4.5 Algorithms Classification Result

No. of Records	C4.5		Weighted C4.5	
	FAR	FRR	FAR	FRR
100	0.095	0.85	0	0
200	0.136	0.041	0	0
400	0.045	0.27	0	0
683	0.019	0.15	0	0

Table 2. Comparison of Evaluation Time Complexity (Seconds)

No. of Records	100	200	400	683
C4.5	0.467	0.827	1.453	2.905
Weighted C4.5	0.797	1.248	2.921	3.935

Table 3. Classification Result of C4.5 and Weighted C4.5 Algorithm

No. of Records	C4.5				Weighted C4.5			
	Precision	Recall	F-Measure	Accuracy	Precision	Recall	F-Measure	Accuracy
100	0.78	0.73	0.82	84.84%	1	0.93	0.96	96.96%
200	0.79	0.85	0.87	92.42%	1	1	0.98	98.49%
400	0.91	0.92	0.92	93.98%	1	1	1	99.2%
683	0.96	0.96	0.94	96.03%	1	1	1	99.6%

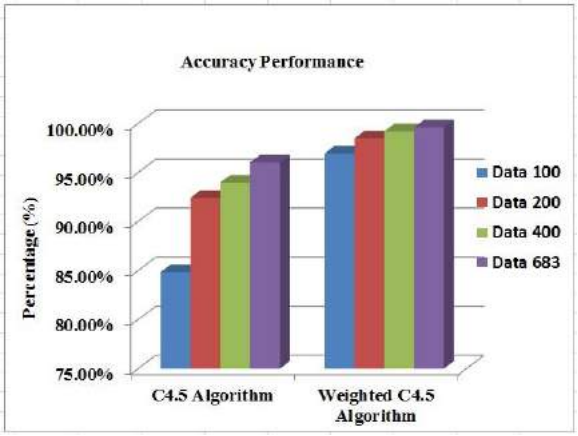


Figure 3. Comparison of Classification Accuracy

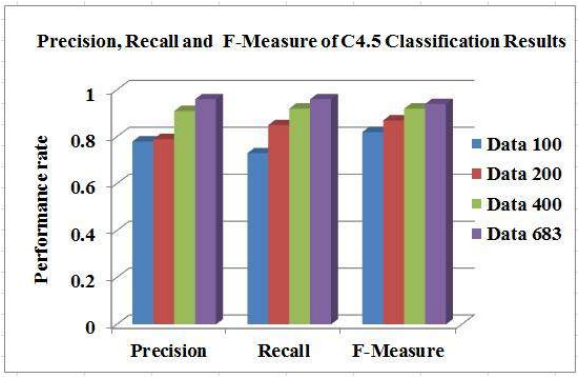


Figure 4. Confusion Matrix of C4.5 Algorithm Classification Result

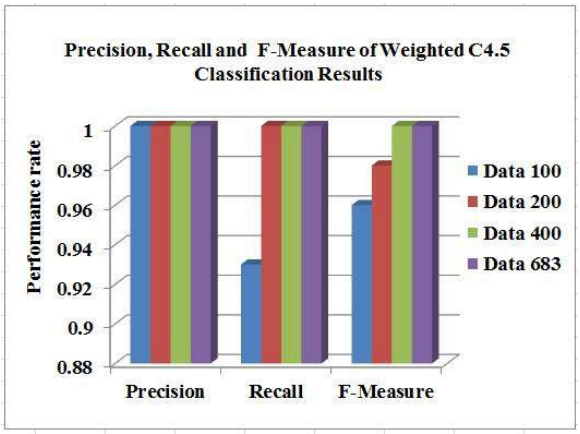


Figure 5. Confusion Matrix of Weighted C4.5 Algorithm Classification Result

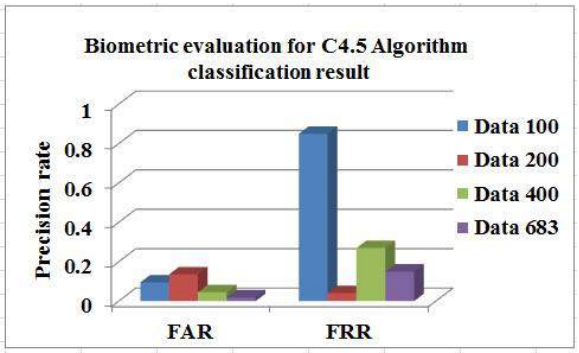


Figure 6. Biometric Evaluation for C4.5 Algorithm classification Result

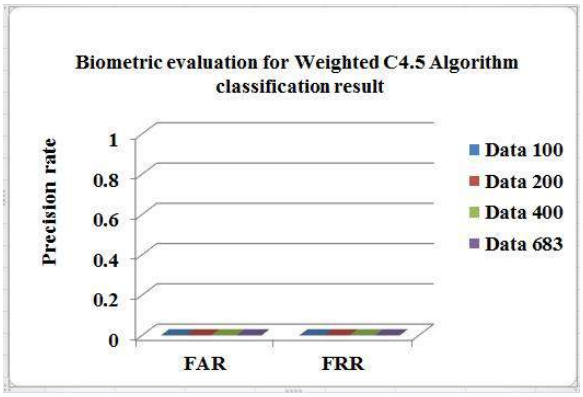


Figure 7. Biometric Evaluation for Weighted C4.5 Algorithm Classification Result

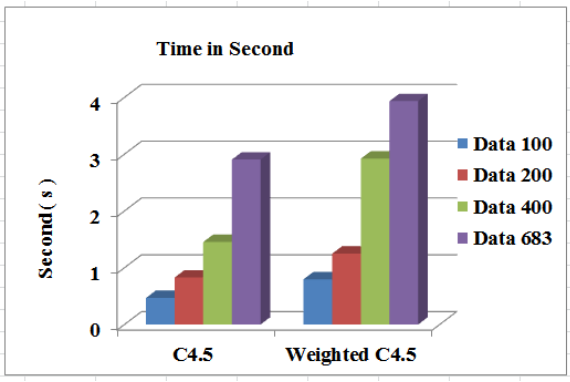


Figure 8. Comparison of Evaluation Time Complexity

7. Conclusion

In this paper, the comparative analysis of C4.5 and weighted C4.5 algorithms classification on Breast Cancer classification was presented. From this study it is found that accuracy of weighted C4.5 algorithm is better than traditional C4.5 algorithm. The time complexity of weighted C4.5 algorithm is also slower than C4.5 algorithm. The experimental results proved that the weighted C4.5 algorithm can achieve high classification rate with better performance.

Acknowledgements

I would like to express my gratitude to University of Computer Studies Yangon for allow me to do this research work. I also thanks to my supervisor for discussions that helped clarify our ideas and his support and encouragement. Finally, I also thanks to all of my colleague for their participation and contribution for this study.

References

[1]. J. R. Quinlan, "Induction of Decision Tree," Machine Learning Vol. 1, 1986, pp. 81-106.

[2]. J. R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers, San Mateo,CA, 1993.

[3]. Kononenko I, "Comparison of inductive and naïve Bayesian learning approaches to automatic knowledge acquisition," in Wieling, B. (Ed), Current trend in knowledge acquisition, Amsterdam, IOS press. 1990.

[4]. Langely, P., Iba, W., Thomas, and K., "An analysis of Bayesian classifier," in Proceedings of the 10th national Conference on Artificial Intelligence (San Matro, CA: AAAI press), 1992, pp. 223-228.

[5]. Han, Jiawei and Kamber, Micheline "Data Mining Concepts and Techniques" 2nd ed., Morgan Kaufmann Publishers, San Franciso, CA, 2007 ISBN 1-55860-901-3.

[6]. Dr. Dewan Md. Farid1 and Prof. Dr. Chowdhury Mofizur Rahman2 "ASSIGNING WEIGHTS TO TRAINING INSTANCES INCREASES CLASSIFICATION ACCURACY" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.1, January 2013.

[7]. UCI Machine Learning Repository: "Breast Cancer Wisconsin (Original) Data Set", Dr. WIlliam H. Wolberg (physician) University of Wisconsin Hospitals Madison, is consin, USA , Donor: Olvi Mangasarian (mangasarian '@' cs.wisc.edu) Received by David W. Aha (aha '@' cs.jhu.edu)

[8]. Ron Kohavi & George H. John, "Automatic Parameter Selection by Minimizing Estimated Error". In Proceedings of the Twelfth International Conference, Morgan Kaufmann Publishers, San Franciso, CA.

[9]. Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri, Mohammed Erritali, "A comparative study of decision tree ID3 and C4.5", International Journal of Advanced Computer Science and Applications, Special Issue onAdvances in Vehicular Ad Hoc Networking and Applications.

[10]. Liu Yuxun, &XieNiuniu, "Improved ID3 Algorithm", IEEE, 2010.

[11]. Gaurav L. Agrawal, & Prof. Hitesh Gupta, "Optimization of C4.5 Decision Tree Algorithm for Data Mining Application", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 3, March 2013.

[12]. Vijayarani, S. and Divya, M. (2011) "An Efficient Algorithm for Generating Classification Rules", International Journal of Computer Science and Technology,Vol.2,Issue 4,.

[13]. Rokach & Maimon "Data Mining with Decision Tree Theory and Applications" 2nd Edition ,2014.

[14]. https://en.wikipedia.org/wiki/C4.5_algorithm.

Loan Applicants Selection System for Private Banks in Myanmar Using TOPSIS

Mya Mya Aye, Thin Lai Lai Thein
University of Computer Studies, Yangon
myamyaayedku@gmail.com, tllthein@ucsy.edu.mm

Abstract

Decision support systems are achieving an increased popularity in various application area including business, engineering, the military, and medicine. They are especially useful in situations in which the amount of available information is restrictive for the intuition of human decision maker and in which precision and optimality are of importance. Bank loan plays a vital role for enterprises and the decision making for accepting or rejecting loan applicants is also important for banks. A prior selection process is needed to determine the loan applicants who will receive or not receive the loan. The propose system is expected to help the decision maker in their loan applicants selection process. Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) is used in a decision support system to search the best loan applicants because of its ability to recommend loan applicants from various variables of applicants.

Key words: *Decision support system, TOPSIS, banks, loan, loan applicants.*

1. Introduction

Decision Support Systems (DSS) is a special type of computerized information system that support business and organizational decision making activities. The application areas of DSSs are the military, management and planning in business, health care, and any area in which management will encounter complex decision situations [5].

Decision support systems are used for strategic and tactical decisions faced by upper-level management decisions [6].

One of the advantages for the banks and their customers is the existence of a bank's lending activities. For the selection of applicants, a prior selection process is needed to determine the customers who will receive or not receive the loan. Decision making process for selecting the loan applicants is important and complex tasks for decision makers. To evaluate bank loans, a checklist of bank's rules and personal judgment are used. Moreover, banking loan decisions require the use of huge and various data and substantial processing time to be able to serve a large number of variables. In order to simplify a loan officer's job, to control it and to achieve more productivity and efficiency an automated decision making system is needed for loan applicant's selection system.

One of the useful methods to implement the decision support system is TOPSIS. TOPSIS is a technique that can consider any number of measures when seeking to identify solutions close to an ideal and far from a nadir solution. TOPSIS can calculate the advantage and weakness for several criteria of customers to the decision maker [8].This system implements TOPSIS-decision support for selection of loan applicant's process.

This paper is organized as follows. Section 2 reviews previous studies on for loan applications. Section 3 describes about the overview of TOPSIS and applications of TOPSIS and presents about the loan applicants selection

system for private banks in Myanmar using TOPSIS. Finally, section 4 concludes and summarize the study.

2. Related Work

Decision support system helps in the process of selecting loan applicants. Shorouq Fathi (et.al)[7] developed a proposed model that identifies Multi-layer Feed-forward Neural Network (MLFN) with back-propagation learning algorithm as an enabling tool for evaluating credit applications to support loan decisions in the Jordanian Commercial banks. This system can be reduced the cost of loan processing and also can reduced personal judgment with high accuracy. However, the limitation of using neural network is required long training time and bank need more cases of successful loan application and bad application to enhance the accuracy. Fuzzy logic expert system is developed for approval of small business loans by applying fuzzy expert system shell [1], this consists of two stages of production rules and three levels of variables. The fuzzy approach needed interviews more experts and needed more membership functions and production rules.

3.TOPSIS in Loan Decision Support System

The acronym TOPSIS stands for Technique for Order Preference by Similarity to the Ideal Solution. TOPSIS was initially proposed by Hwang and Yoon (1981) [3], subsequently improved by many (Chu, 2004; Peng, 2000) [1]. TOPSIS finds the best alternatives by minimizing the distance to the ideal solution and maximizing the distance to the nadir or negative-ideal solution (Jahanshahloo et al., 2006) [4]. All alternative solutions can be ranked according to their closeness to the ideal solution. The processes of TOPSIS are below:

Input: Decision matrix DM

Weights for criteria W_i

Output: Sorted Lists

Method:

1. **for** each item r in DM **do**
2. calculate normalized value of r_{ij}

   $$r_{ij} = \frac{x_{ij}}{\sqrt{x_{ij}^2}}$$

   end for
3. **for** each item r in normalized matrix **do**
4. calculate weighed normalize value v_{ij}

   $v_{ij}=w_j * r_{ij}$
5. **end for**
6. determine the ideal(A^+) and negative ideal (A^-) as

   A^+=maximum value of benefit criteria

   A^-=minimum value of cost criteria
7. calculate separation measure for each altern atives

   calculate S^+ from A^+

   $$S_i^+ = \sqrt{\sum_{i=1}^{n}(V_j^+ - V_{ij})^2}$$

   calculate S^- from A^-

   $$S_i^- = \sqrt{\sum_{i=1}^{n}(V_j^- - V_{ij})^2}$$
8. calculate the relative closeness to the ideal solution as

   $C_i = S_i^- / (S_i^+ + S_i^-), 0 \leq C_i \leq 1$
9. rank the preference order.

TOPSIS method is applied for supporting the selection of loan applicants, consists of three steps. These are:

1. The applicants apply loan from online.
2. Pre-Selecting the applicants according to criteria.
3. Applying TOPSIS to analyze the alternatives and determination of the final rank.

In the first step, the applicants must provide correct information in order to make decision process more efficiently.

Thereafter, the loan officer can check applicants' information, can calculate score of applicants, can rank the applicants, can view rejected list and can delete applicants.

In this model, some criteria are checked before going to calculate with TOPSIS. Some constraints are defined according to domain expert. The pre-selection rules are shown below:

1. IF age < 20 OR age > 70 THEN reject.
2. IF nationality = "other" THEN reject.
3. IF land_types = "other" THEN reject.
4. IF business period < 3 years THEN reject.
5. IF warranty value period <3 years THEN reject.
6. IF loan_purpose = "other" THEN reje ct
7. IF business_type = "other" THEN reject
8. IF DCR_Range < 1 THEN reject.
9. IF ICR_value < 1.5 THEN reject.
10. IF WarrantyValue > 80% THEN reject.

After pre-selection, the score of applicants are calculated. In this TOPSIS based decision support system seven criteria are used to determine the applicants' selection. These are Interest Coverage Ratio (ICR), Debt Coverage Ratio (DCR), Loan from other bank exists, Loan Type, Period of Loan, Warranty Value and House Status. In order to calculate ICR and DCR, EBITDA (Earnings Before Interest, Taxes, Depreciation and Amortization) is needed. The EBITDA, ICR and DCR values are calculated according to the equation 3.1, 3.2 and 3.3 respectively.

$$EBITDA = Net\ profit + Interest + Taxes + Depreciation \quad (3.1)$$

$$ICR = \frac{\text{EBITDA}}{\text{Total Debt}} \quad (3.2)$$

$$DCR = \frac{\text{EBITDA}}{\text{Interest}} \quad (3.3)$$

For the loan type is over draft, if the loan from other bank exists total debt is calculated as equation 3.4 and if the loan from other bank does not exist total debt is calculated as equation 3.5.

[loan amount/2 *3%]*12 + [loan amount *13%]+[loan from other bank*20%]+[loan amount from other bank *13%] (3.4)

[loan amount/2 *3%]*12 + [loan amount *13%] (3.5)

For the loan type not over draft, if the loan from other bank exists total debt is calculated as equation 3.6 and if the loan from other bank does not exist total debt is calculated as equation 3.7.

[loan amount*3%]*12 + [loan amount *13%]+[loan from other bank*20%]+[loan amount from other bank *13%] (3.6)

[loan amount*3%]*12 + [loan amount *13%] (3.7)

Warranty value can get from the Forced Sale value and it can achieved from equation 3.8.

Forced Sale value=Current Price of Land* land type_percent +Current price of building* building_type_percent (3.8)

The land type percent is defined according to tan type. If land type is grand land and ancestral the percent is defined 70% and 50% is defined for others type of land. The percent of building is defined as 50% for competed buildings and otherwise only 30% is defined.

In TOPSIS weight needed to be defined as a value assigned to an evaluation criterion which indicates its importance relative to other criteria under consideration. In this model, ranking method is used for evaluating the important of weights which includes the every criterion used to consider in this system is ranked in the order of decision maker preferences. So, weights value references are assigned to each criterion according to loan officer's suggestion. The selected criteria and their associated weight value references are shown in Table (3.1). The most important criteria give the rank value 1 and so on.

Table (3.1) Weight value references.

Name	Rank Value	Weight Value
ICR	1	0.1707317
DCR	1	0.1707317
Warranty Value	1	0.1707317
House Status	2	0.1463415
Other Loan Exists	3	0.12195122
Loan Type	3	0.12195122
Loan Period	4	0.09756

The weigh value for each criteria are calculated by using rank sum (RS) method. The weights applied in the rank sum method are the individual ranks normalized by dividing by the sum of the ranks. The weights producing formula is shown in equation 3.9.

$$w_j = \frac{n - r_j + 1}{\sum_{k=1}^{n} n - r_k + 1} \qquad (3.9)$$

Where, r_j is the rank of the j^{th} criteria and n is the number to criteria.

The system testing uses applicants' data who propose loan to bank. The sample calculation steps by applying TOPSIS is shown in the following. Where C1,C2,.Cn is the applicants of loan, LT means Loan Type, P is loan period , WV is the warranty value, I is ICR , D is DCR, HS is house Status and OL is loan from other bank. The input of TOPSIS is to create the decision matrix as shown in Table (3.2).

Table (3.2). The decision matrix of applicants.

	LT	P	WV	I	D	HS	OL
C1	2	3	3	1	2	2	2
C2	2	3	4	1	2	2	2
C3	2	2	1	1	3	2	2
C4	2	1	4	4	2	2	2
C5	1	3	5	3	2	2	2
C6	2	1	3	1	2	2	2
C7	1	3	3	1	3	2	2
C8	1	2	5	4	2	2	1
C9	2	2	1	4	2	2	1

Then, the first step is to create normalized decision matrix. The resut of normalized decision matrix is shown in Table(3.3). The next step is to construct the weighted normalize decision matrix. The weigh value for each criteria is calculate according to equation(3.9). The results of weighted normalized for each customers are shown in Table(3.4). Then, determine the ideal (A+) and

negative ideal solution (A-). The results are shown in Table(3.5).

Table(3.3).The normalized decision matrix

	LT	P	WV	I	D	HS	OL
C1	0.3849	0.42426	0.28475	0.45584	0.2949	0.3849	0.365148
C2	0.3849	0.42426	0.37966	0.11396	0.2949	0.3849	0.365148
C3	0.3849	0.28284	0.09492	0.11396	0.4423	0.3849	0.365148
C4	0.3849	0.14142	0.37966	0.45584	0.2949	0.3849	0.365148
C5	0.19245	0.42426	0.47458	0.34188	0.2949	0.3849	0.365148
C6	0.3849	0.14142	0.28475	0.11396	0.2949	0.19245	0.365148
C7	0.19245	0.42426	0.28475	0.11396	0.4423	0.3849	0.365148
C8	0.3849	0.28284	0.09492	0.45584	0.2949	0.19245	0.182574
C9	0.19245	0.28284	0.47458	0.45584	0.2949	0.19245	0.182574

Table(3.4). Weighted normalized decision matrix.

	LT	P	WV	I	D	HS	OL
C1	0.04694	0.04139	0.04862	0.07783	0.0503	0.05633	0.04453
C2	0.04694	0.04139	0.06482	0.01946	0.0503	0.05633	0.04453
C3	0.04694	0.02759	0.01621	0.01946	0.0755	0.05633	0.04453
C4	0.04694	0.0138	0.06482	0.07783	0.0503	0.05633	0.04453
C5	0.02347	0.04139	0.08103	0.05837	0.0503	0.05633	0.04453
C6	0.04694	0.0138	0.04862	0.01946	0.0503	0.02816	0.04453
C7	0.02347	0.04139	0.04862	0.01946	0.0755	0.05633	0.04453
C8	0.04694	0.02759	0.01621	0.07783	0.0503	0.02816	0.022265
C9	0.02347	0.02759	0.08103	0.07783	0.0503	0.02816	0.022265

Table(3.5).The results of ideal solutions.

A+	A-
0.046939	0.02347
0.013797	0.041392
0.08103	0.01621
0.077827	0.019457
0.07552	0.05035
0.056327	0.028163
0.0222651	0.0445303

The next step is to calculate the separation measures for each alternatives. Then, relative closeness for each customeris calculated. The results of separaion measure and relative closeness are shownin Table (3.6) and Table (3.7). The final process is rank the applicants and the ranked lists are shown in Table(3.8). According to the ranked list, the decision maker of the bank can determine which applicants are more appropriate to grant the loan. The applicants who have the highest closeness value is the most appropriate for granting the loan.

Table(3.6). Separation measures values.

	S+	S-
C1	0.054234	0.076167
C2	0.074569	0.060889
C3	0.091076	0.046562
C4	0.03731	0.088747
C5	0.053106	0.080679
C6	0.079875	0.048608
C7	0.079155	0.049772
C8	0.076282	0.068147
C9	0.046562	0.091076

Table(3.7). Relative closeness values

Appicants	Score Vale
C1	0.584101
C2	0.449504
C3	0.338294
C4	0.704023
C5	0.60305
C6	0.378319
C7	0.38605
C8	0.471837
C9	0.661706

The data are collected from the Yoma Bank. There are 360 applicants. The criteria and

rules for rejection and score calculation are suggested by loan officers from Yoma bank.

Table(3.8) Ranked Lists of applicants

Applicants	Score Value
C9	0.704023
C4	0.661706
C2	0.60305
C5	0.584101
C8	0.471837
C6	0.449504
C7	0.38605
C1	0.378319
C3	0.338294

System testing is done by ranking the collected 360 applicants. The first stpe of testing is pre-processing phase and the results shown that all rules are working correctly. The next step is ranking the applicants by using TOPSIS methods. The manual and system testing is done for all applicants and the results produce similar results. Therefore TOPSIS can support the decision maker by reducing the manual analysis of applicants data. Moreover the loan applicants also save their time in loan appying process.

4.Conclusion

Decision support systems are powerful tools integrating scientific methods for supporting complex decisions with techniques developed in information science, and are gaining an increased popularity in many domains. TOPSIS is one of the numerical methods of the multi-criteria decision making. This is a broadly applicable method with a simple mathematical model. This system implements TOPSIS-decision support for selection of loan applicants. In this system, loan applicants are ranked with seven criteria (loan type, period, warranty value, ICR, DCR, house status, loan from other bank). This proposed system can be reduced the cost of loan processing and also can reduced personal judgment .By using this system, loan decision maker can compare the loan applicants by varying weight values dynamically.

References:

[1]. Chu, T., “Facility Location Selection Using Fuzzy TOPSIS Under Group Decision”, International Journal of Uncertainty, Fuzziness & Knowledge-Based Systems, Vol. 10, No. 6, pp. 687-701.2004.

[2]. Ewa ROSZKOWSKA, “RANK ORDERING CRITERIA WEIGHTING METHODS.A COMPARATIVE OVERVIEW”, OPTIMUM. STUDIA EKONOMICZNE NR 5 (65) 2013.

[3]. Hwang C.L., and Yoon K., “Multiple Attribute Decision Making: Methods and Applications” Springer-Verlag: New York. 1981.

[4]. Jahanshahloo G.R., Hosseinzadeh L. F., and Izadikhah M., “Extension of the TOPSIS Method for Decision-Making”, Applied Mathematics and Computation, Vol. 181, No. 2, pp.1544 – 1551.2006.

[5]. Marek J.Druzdzel and Roger R. Flynn. " Decision Support System", Allen Kent(ed.), New York: Marcel Dekker , Inc., 2002.

[6]. Peng, Y.,“Management Decision Analysis”, Peking: Science Publication. 2000.

[7]. Shorouq Fathi Eletter,Saad Ghaleb Yaseen, Ghaleb Awad Elrefae: "***Neuro-Based Artificial Intelligence Model for Loan Decisions***", American Journal of Economics and Business Administration2 (1) : 27-34, 2010.

[8]. Zlatko Pavic , Verdan Novoselac . “Notes on TOPSIS Method”, International Journal of Research in Engineering and Science (IJRES). June 2013.

Retrieving Semantically Relevant Documents Using Latent Semantic Indexing

Chue Wut Yee, Zon Nyein Nway
University of Computer Studies, Yangon
chuewutye@ucsy.edu.mm, zonnyeinnway@ucsy.edu.mm

Abstract

Nowadays, with the development of the internet, it is possible to store very large amounts of data and searching useful information from these data became an essential work. In this system, the most common method for information retrieval, latent semantic indexing method is used to retrieve semantically related documents. This system is able to accept a user query, search the most semantically related documents, retrieve and output the documents according to their similarity values. After the documents are collected, some pre-processing tasks are usually performed. After preprocessing the retrieval model is built by using LSI method as the training phase. In this phase almost 200 documents from biomedical data (the papers with pdf extension and approximately 2500 words) are used. With the help of this system, the users can search effectively the most semantically related documents with higher accuracy (%). Furthermore, the similarity measurement for the proposed method is presented by using cosine similarity method. And the performance value is accomplished by employing the system evaluation measures such as precision, recall, and F- measure.

Keywords: Latent Semantic Indexing (LSI), Singular Value Decomposition (SVD).

1. Introduction

Popular search Engines such as Google, Yahoo, Bing, and AltaVista give the services of the form of modern information retrieval. And the users can search for information in multimedia formats such as text, audio, still images, and moving images by looking up for keywords appeared in any documents and/or files stored in different formats, such as HTML, MS Word, PDF, and images.

Information retrieval (IR) is the process of obtaining information resources relevant to the required information from a collection of information resources. A similarity measurement is essentially performed to facilitate the users in accessing the required results effectively [6].

The vector space model has been widely used in the traditional information retrieval field. In the vector space model, text representation of objects and queries are treated as vectors in a multi-dimensional space, the dimensions of which are the words used to represent the terms. Term weighting is an important aspect of modern text retrieval systems. But that term based search cannot retrieve the semantically relevant documents. To overcome this, the proposed system implements the semantically based information retrieval system based on LSI [8].

Latent semantic indexing is an indexing and retrieval method that uses a mathematical technique called singular value decomposition to figure out patterns in the relationship between the term used and the meaning they convey. A term-document matrix is created and singular value decomposition method is performed on this matrix. This ensures that the arrangement of the matrix presents the important associative patterns in data. Points that are closer to each other refer to documents that are semantically relevant. Therefore, this system is useful for finding semantically related information from the specific collection of documents [1] [7].

The input of the proposed system is user query (phrases or lines of sentences) and the output is the related documents' titles of the collection of documents stored in database. The proposed system is able to accept a user query, search the most semantically related documents by using LSI, retrieve and rank the documents according to Cosine similarity values. Finally, the resulted documents are sorted and ranked in descending order according to the similarity values.

The rest of this paper is organized as follows. In section 2, Related Works is presented. The proposed system methodology is described in section 3 and proposes system design is explained in section 4. In section 5, the analysis and empirical results are shown. The last section 6 is about the conclusion.

2. Related Work

M.Al-Qahtani, A.Amira and N.Ramzan [3] discussed that effective information retrieval technique for e-Health systems. This system stated that data in computer systems is in coded format. However, certain data like user comments cannot be coded because it is stored in the form of free text. Extracting the valuable information from such free text is a challenge due to the complexity of the stored data. This system used the latent semantic indexing

algorithm on the Health Improvement Network. The LSI algorithm uses the computational power of multiprocessor in performing the retrieval process. The representation of the patient's data in the form of term document matrix is transformed. By using this method, processing time will be reduced and the rate of relevancy was accurate.

Nang Ei Khaing [4] presented the similarity based document retrieval system by computing the weight of each term based on its inverse document frequency (idf) in the document collection. And Jaccard Coefficient similarity method is deployed to find the similarity measurement between words. Almost 200 documents with pdf format from UCSY are trained for the retrieval model. In this paper, the author stated that although the cosine similarity takes less amount of time as compared to Jaccard coefficient method because of using mathematical formula, the Jaccard is applied in the system as this similarity measure is suitable with the tf-idf method.

R.Anita, C.N.Subalalitha and A.Dorle [5] proposed about the semantic search using latent semantic indexing and wordnet. Most of the famous search engines use the concept of semantic search. The general method (document to document) similarity search is the sequential search which involves numerous noise effects. An efficient way of improving the sequential search is latent semantic indexing (LSI) which maps the words under the conceptual space. That conceptual space depends on the queries and the document collection. And the results obtained from LSI are free of some semantic such as polysemy and synonym etc. So, the system is integrated with WordNet, a large lexical database of English language to increase the search results.

3. Proposed System Methodology

This system introduces some difficulties in the retrieval process generating a semantic gap between user needs and their requirement. The satisfaction of a specific information need on the web is supported by search engines and other tools aimed at helping users gather information from the web. The advance of internet information, search engines have a prominent role in information retrieval and web mining applications.

3.1. Information Retrieval System

Due to the vast amount of information available in these days, it cannot be effectively and efficiently search. The goal of information retrieval is to find documents that are relevant to a given user query. Today people are rare to go to the libraries and more and more searches on the Web. Traditional IR assumes that the basic information unit is a document, and a large collection of documents become the text database. A ranking of the set of documents is usually also performed according to their relevance scores to the query. IR is different from data retrieval in database using SQL queries because the data in database are highly structured and stored in relational tables, while information in text is unstructured. There is no structured query language like SQL for text retrieval [6] [8].

3.2. Vector Space Model (VSM)

The vector space model is implemented by creating the term-document matrix and a vector of query. Let the list of relevant terms be numerated from 1 to m and documents be numerated from 1 to n. The term-document matrix is an $m \times n$ matrix $A = [a_{ij}]$, where a_{ij} represents the weight of term i in document j. Each term is weighted by the number of documents in which it appears. The components of the term vector are simply term frequencies [6].

3.3. Semantically based Information Retrieval System

Semantically-based retrieval system finds semantically related documents based on a set of common concepts. The output of such retrieval should be based on the degree of relevance, where relevance is measured based on the closeness of the concepts. It is difficult to provide a precise measure of the degree of relevance between a set of terms. Therefore, the statistical method, latent semantic indexing is used as a semantically based information retrieval system.

3.3.1. Latent Semantic Indexing (LSI)

Latent Semantic Indexing (LSI) was proposed at the end of 80’s as a way to solve such problem of Vector Space Model. The basic observation is that terms are an unreliable means to assess the relevance of a document with regard to a query because of synonymy and polysemy. Thus, one should represent document using a more semantically accurate way, i.e., in terms of “concepts”. LSI is based on the terms that are used in the same contexts tend to have similar meanings. LSI retrieved documents are semantically related to the user’s search query, even if they do not share the exact same keywords. LSI extends traditional VSM by modeling the term-document relationship using reduced dimension representation computed by the matrix rank reduction technique of Singular Value Decomposition (SVD) [1] [7].

3.3.2. Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is the matrix rank reduction technique. It aims to identify the statistical association of the words. The results of

decomposition of the original matrix are descriptions of terms and documents based on the hidden correlation space derived from SVD. The method reflects the major associative patterns in the dataset. Let be the M x N term-document matrix A. If matrix A has rank k, k <= min (M, N) and then matrix A can be decomposed by singular value decomposition [2].

$$A = USV^T \quad (1)$$

Where, S is the "concept matrix"
U is the "term-concept matrix"
V is the "document-concept matrix"
k is the rank value

3.3.3. Similarity Measure

A key factor in the information retrieval system is the similarity measure between query and documents. There are two simple and well known similarity such as Jaccard and Cosine Coefficients. These measures calculate the similarity on how many words are in common [6].

Cosine Similrity considers documents and queries to be vectors in a term space. In Cosine Similarity, the lower angles presents higher similarity and higher value of an angle represent dissimilarity between query and a set of documents. The top ranked documents are regarded as more relevant to the query.

The cosine similarity is defined as

$$S_{A,B}^{Cosine} = \frac{|\{words_A\} \cap \{words_B\}|}{|\sqrt{\{words_A\}} \| \{words_B\}|} \quad (2)$$

Where, $words_A$ = word from trained dataset
$words_B$ = word from user query

4. Proposed System Design

This paper intends to implement the information retrieval model by LSI. LSI is based on vector space model. The goal of this retrieval model is to find documents that are relevant to the user query. The detail designs of the proposed system are described in figure 1 and figure 2.

The proposed system uses the combination of vector space model and dimension reduction method. The proposed system consists of two modules: building retrieval model (training phase) and finding similar documents for query terms (testing phase).

For the training module, firstly preprocessing steps will be carried out and built the retrieval model by using statistical method SVD for latent semantic indexing. The required training documents from biomedical data are used. The total trained documents are almost 200 with paper format (pdf extension).

For the similarity finding module, the input of the proposed system is user query (phrases or lines of sentences) and the output is the related documents' titles of the collection of documents stored in database together with similarity values.

To evaluate the system performance, analysis results are emphasized on training time and system accuracy results. The performance value is accomplished by employing precision, recall, and F-measure values.

4.1. Document Collection and Data Description

The collection of training documents is related with biomedical data. One document has at least one page. The collected Biomedical related document are downloaded from https://www.medicinenet.com/symptoms_and_signs/symptomchecker.htm#introView. The main topics are the popular signs and symptoms of different diseases. We used almost 200 different kinds of diseases (approximately 2500 words). According to the nature of biomedical diseases, some signs and symptoms are common for some kind of diseases. And there are many diseases that are totally different form one another. These documents are unstructured documents with pdf formats. For each document, the list of terms and the frequency of each term are required to create *m x n* term-document matrix. According to the literature, some preprocessing steps should be performed in order to improve the retrieval process and to reduce the computational time and storage requirements.

4.2. Training Phase of the Proposed System

For the training phase of the proposed system, the preprocessing step is necessary to reduce the whole document by removing words void of semantic content such as "of", "the", "and", "a" and "an", etc. The remaining words are eliminates grammatical variations of the same word by reducing it to the stem or root word form with the help of Potter Stemmer.

After preprocessing, the remaining words are the terms. Then the original term document matrix is created. Based on that matrix, LSI approach is applied by the following algorithm.

Algorithm for the Training Phase

Input	:	Set of Unstructured Documents
Output	:	Set of Document Vectors
Step 1	:	Create a list of documents
Step 2	:	For each document - Read the words from documents - Filter out the stop-words from a stop-word list For each word - Apply Stemming (using porter stemmer) - Add the stemmed word to word list - Create a document _word relation
Step 3	:	Calculate Term's count for each document.
Step 4	:	Generate the weighted term-document matrix.
Step 5	:	Compute SVD and save to database.
Step 6	:	Reduce ranked document vector for comparison with the query vector.

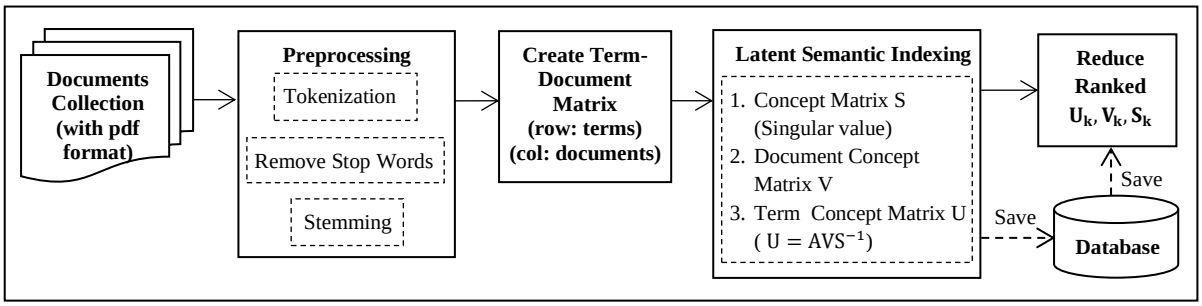


Figure 1. Process flow diagram for the Training Phase

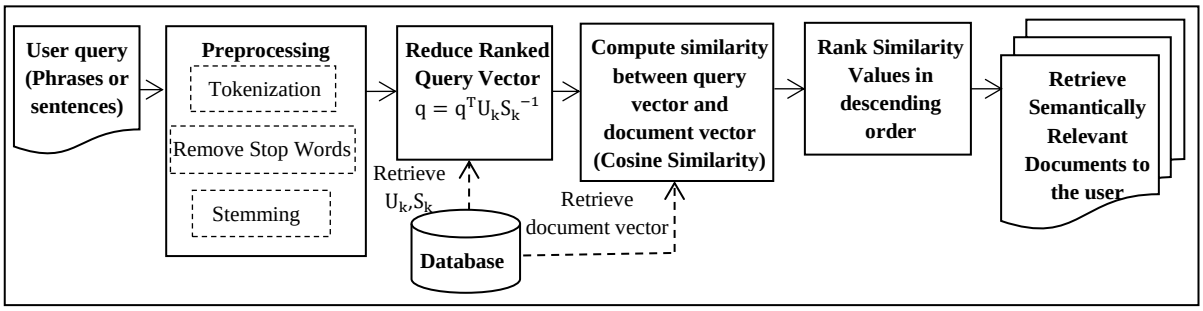


Figure 2. Process flow diagram for the Testing Phase

4.3. Testing Phase of the Proposed System

After the training phase, the retrieval model is available to use. The user input query is accepted to search the similar documents from the trained document collection. The query must be represented as a column vector whose i^{th} term is nonzero only if the i^{th} term appears in the query. Each document is scored for relevancy by computing its inner product with the input query.

The highest-scoring documents are considered the most relevant. LSI replaces the document matrix with a ranked approximation matrix generated by the truncated singular value decomposition (SVD) and computes the most similar documents by the following algorithm.

Algorithm for the Testing Phase

Input : Query Text
Output : Set of Similar Documents with the Query
Step 1 : Fetch the Documents list, Word list, S_k, U_k, A (term-document matrix) from the database.
Step 2 : Get the query text and remove stop-words and do the stemming process on query text.
Step 3 : Create a query vector using words list called q^T.
Step 4 : Compute the query vector $q = q^T U_k S_k^{-1}$
Step 5 : Compute the similarity between query and document vector.
Fetch the document vector (obtained from training phase).
Step 6 : Rank similarity values in descending order.
Step 7 : Retrieve semantically relevant documents to the user.

4.4. Sample Calculation of the Proposed System

This section explains the sample calculation for the proposed system. Suppose, there are four documents of biomedical related data. Each document describes the symptoms and signs of the specific disease.

Documents collection = {D1, D2, D3, D4}

In document D1, describes the symptoms and signs of teen depression disease, D2 describes the symptoms and signs of suicide disease, D3 describes the symptoms and signs of Schizoaffective disorder disease and D4 describes the symptoms and signs of Flu disease.

After preprocessing is done for the document collection, construct the term-document matrix A. In this matrix, which each row describes the term and each column describes the document. Each cell entry is the frequency of the term occurred in the specific document.

The original term-document matrix A is described as follow:

Matrix A

word	Teen Depression	Suicide	Schizoaffective Disorder	Flu
absence	1	0	0	0
abuse	0	3	0	0
accompan	1	0	0	0
aches	0	0	0	1
advers	0	1	0	0
age	1	1	0	0
agitation	1	0	0	0
alcohol	1	0	0	0

Based on the term-document matrix A, singular value decomposition (SVD) method is applied by $A = USV^T$. According to the SVD, the document concept matrix (V), the singular matrix (S) and the term concept matrix (U) are as follow:

Matrix V

0.3674	0.1177	-0.2913	-0.8754
0.4823	0.7578	0.4056	0.1692
0.4985	-0.0226	-0.7393	0.4522
0.6196	-0.6414	0.4518	0.0235

Matrix S

19.1825	0	0	0
0	14.4292	0	0
0	0	12.3254	0
0	0	0	9.9957

Matrix U

0.0251	0.0525	0.0329	0.0169
0.0251	0.0525	0.0329	0.0169
0.0192	0.0082	-0.0236	-0.0876
0.0251	0.0525	0.0329	0.0169
0.0503	0.105	0.0658	0.0339
0.0192	0.0082	-0.0236	-0.0876
0.0323	-0.0445	0.0367	0.0023
0.0251	0.0525	0.0329	0.0169

Therefore, the three matrices U, S and V are received as the training phase.

For the testing phase, the query term q is "depression". According to the algorithm for the testing phase, the proposed system retrieved the semantically relevant documents for the query term "depression" as follow.

D1: Teen Depression (0.8977)
D3: Schizoaffective Disorder (0.7366)
D2: Suicide (0.5415)
D4: Flu (-0.313)

In the above sample calculation, firstly D1 is retrieved, after that D3 and D2. D1 is closest to query because it contains the word "depression". D2 and D3 talks about (concept) the "depression" so it is ranked after D1. D4 does not mention anything related to query so it is lowest rank. LSI uses word-document concept to retrieve documents instead of using actual word occurrence.

5. Analysis and Empirical Results

In the proposed system, latent semantic indexing and cosine similarity method are used. This system can be used to search semantically relevant documents. The proposed system can be applied in electronic library catalogs and private search retrieval system in specific organization. Firstly, the user wants to search the relevant information for the search phrases or terms. That query terms are passed to the pre-processing stages. According to the method used, the trained collection of documents is pre-processed. In LSI,(m x n) term-document matrix must be created. Then that matrix is replaced with k-rank approximation generated by the singular value decomposition (SVD). The input query is compared with the document vectors generated by the SVD. Finally, the system output the semantically relevant documents by considering the high rank approximation.

5.1. Measuring System Accuracy

If there is no system performance evaluation, it is impossible to know how well the system is performing. Measuring system accuracy means how well the proposed system can retrieve the relevant documents from non-relevant documents for the given user query. Precision, recall and F-measure are widely used measures to determine the effectiveness of the retrieval system [8].

$$\text{Precision} = \frac{No:of..relevant..documents..retrived}{Total..no:..of..documents..retrieved} \quad (3)$$

$$\text{Recall} = \frac{No:of..relevant..documents..retrived}{Total..no:..of..relevant..documents} \quad (4)$$

$$\text{F-measure} = 2\frac{precision * recall}{precision + recall} \quad (5)$$

Precision measures the percentage (%) of documents that the system detected that are in fact as relevant. Recall measures the percentage (%) of documents actually included in the user input that were correctly identified by the system. Precision and Recall are combined into a single metric called the F-measure [6].

5.2. Performance Evaluation

We have almost 200 pdf typed documents (approximately 2500 words). Initially, we trained at most 50 documents for the training process. And then another 50 documents are added up to the 200 documents. Here, we found that the training time is increased according to the training document size by the following analysis results.

Table 1. Training Time Evaluation

Number of Documents	Training Time (min)
50	35
100	72
150	157
200	315

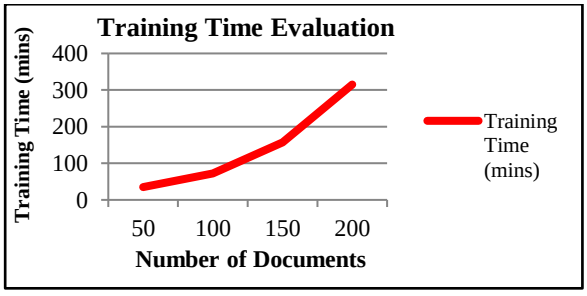


Figure 3. Line Chart for Training Time Evaluation

But only considering the time does not completely represent the powerful retrieval system. The efficiency should be considered. Therefore, the retrieval results must be analyzed for system effectiveness.

According to the Table.2, we have found out k=150 is the best value. The above table shows the system correctness at different k values. Due to the desirable results, the training data size 200 documents are chosen and the rank value k is the most suitable at 150 because the correctness of the system accuracy is higher for the later k value form 100. For the earlier k value from k=100, the correctness of the system decreases a little.

Table 2. Comparison of Different k Value

k value	System Correctness (%)
50	70
100	82
150	89
200	90

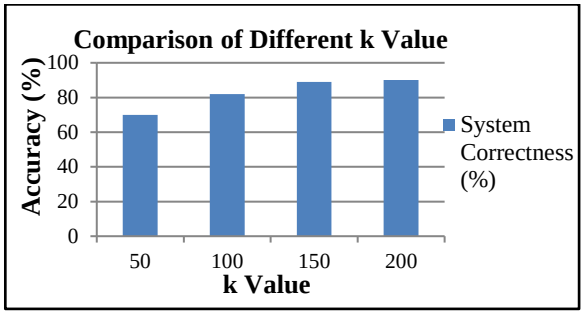


Figure 4. Column Chart for Comparison of Different k Value

Table 3. System Accuracy Evaluation

Number of Documents	Precision (%)	Recall (%)	F-measure (%)
10	80	100	88
20	80	100	88
30	100	75	85
40	100	75	86
50	100	75	86
60	80	100	89

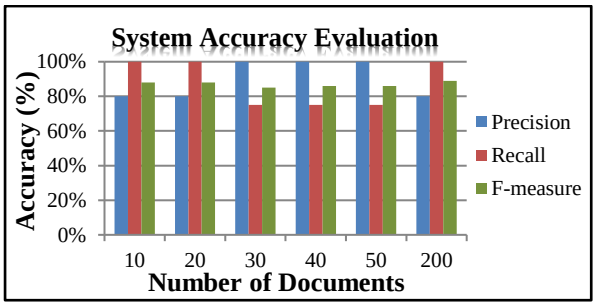


Figure 5. Column Chart for System Accuracy Evaluation

There is nothing to measure, it is impossible to know how well the system is performing. Finally, after selecting the k value (150) and the number of training documents (200), we analyzed the effectiveness of the proposed system by testing of 10, 20, 30, 40, 50 and 200 documents. Based on the number of correctly retrieved documents from total number of documents, the resulted value of precision, recall and f-measure are shown in table 3 and column chart in figure 5. Depending on the experimental result, the overall accuracy of the proposed system is 89% for the number of testing documents 200.

6. Conclusion

Latent Semantic indexing (LSI) is a useful method to implement the retrieval of documents on the basis of conceptual meaning and semantic analysis. This is a mathematical approach for the retrieval and so it makes sure that the relevancy of the obtained documents is optimized. Not all the documents using generic key word search and term matching are relevant. This method overcomes many difficulties associated with generic key word search and term matching. The main purpose of this system is to store the documents collection and then retrieve the semantically relevant documents for the user's input query. There will be very tedious for matching the incoming query with the existing documents and will not grantee to retrieve the semantically related documents without using this system.

References

[1] C.A. Kumar and S. Srinivas, "On the Performance of Latent Semantic Indexing-based Information Retrieval", Journal of Computing and Information Technology, Vol. 3, pp. 259-264, 2009.

[2] D. Kalman, "A Singular Value Decomposition: The SVD of a Matrix", The American University, Washington, DC 20016, 13 February 2002.

[3] M. Al-Qahtani, A. Amira and N. Ramzan "An Efficient Information Retrieval Technique fore-Health System", IEEE Conference on Information and Communication Technology, 2015.

[4] Nan Ei Khaing, "Similarity based Retrieval of Documents using Jaccard Coefficient Similarity Methods" Parallel and Soft Computing, 2018.

[5] R. Anita, C.N. Subalalitha, A. Dorle and K. Venkatesh, "Semantic Search Using Latent Semantic Indexing and WordNet", ARPN Journal of Engineering and Applied Sciences, 2017-2018.

[6] R.B- Yates and B.R- Nieto, "Modern Information Retrieval", 1999.

[7] S.T. Dumais, S. Deerwester, T.K. Landauer and R. Harshman, "Indexing by Latent Semantic Analysis", Journal of the American Society for Information Science, Vol. 41, pp. 391-407, 1990.

[8] W.B.Croft, D.Metzler and T.Strohman, "Information Retrieval in Practice", Pearson Education, Inc, 2015.

Cluster-based Job Matching System

Phyo Pyae Sone, Dr. Khine Moe Nwe
University of Computer Studies, Yangon
phyopyaesone@ucsy.edu.mm, khinemoenwe@ucsy.edu.mm

Abstract

Online job recruitment platform is one of the most prominent channels for both job seekers and recruiters to hunt jobs and find suitable employees respectively. In the traditional job matching process, manually scanning the resume or profile of a job seeker and matching the resume of job seekers and requirements of job recruiters takes time-consuming and makes difficulties for both seekers and recruiters. The related studies applied k-means clustering for providing the similar clusters of data but gives less relevant data. In this paper, we present a job matching system using k-means and word2vec is to output the clusters with semantically similar words. As a result, using k-means clustering and word2vec model, recruiters can get the most relevant candidates that fit employers' needs than k-means clustering only.

Key words: ***k-means, word2vec model, job matching***

1. Introduction

The advancement of IT has greatly enhanced job recruitment. Computerize job recruitment system evolved from the prototypical system that existed for a long time. A computerized job selection program has been designed and implemented in this project. This provides user-friendly interactive software environment with value-added services such as accurate outcome processing system and much more. In the design of the computerized recruitment system, the structured system analysis and design methodology was adopted.

Because of the limited talent pool and recruiters ' experience, it becomes necessary to automatically find the right candidates. This system proposed a system for job recruitment based on Kmeans clustering and word2vec model, providing a more accurate algorithm for the resume and job matching. It provides an automated process of recruitment that reduces the need for manual processes and paperwork. The job seeker dataset will be inserted as training data into the database in this system. Notifications will be sent to the job seekers as soon as the recruiters post recruitments relevant with their resumes. Besides, the system will recommend the list of the most suitable candidates ranking for recruiter when their recruitments were published.

2. Related Work

The related works of replication are discussed in this session.

In the Clustering Approach to Analyze Student Data using Kmeans Algorithm: The main challenge of clustering is effectively meaningful groups that are succinct annotated. The data clustering of students is the automated grouping of students into clusters or classes so that students within a cluster can participate in a cluster have a high resemblance compared to each other, but they are very different from other cluster students. Cluster analysis and kmeans algorithm use in the education sector in this system. The information of the participant is grouped according to their marks of examination. The system is able to analyze the data of the student and provide the results of the study. The proposed system provides for any high school of

government technology that has the relationship betw een the outcome of the students ' entry exam and their success.

In Cluster based Ranking Index for Enhancing Recruitment Process using Text Mining and Machine Learning, January 2017. Using Term Document Matrix, this method provides an efficient approach to extracting specific terms from the resumes. A tech nique for clustering was used to identify different res umes. Based on the cluster, the value of that word wa s determined which makes this paper unique. The term information matrix defines the number of words in th e frequency of all summaries. Every row represents o ne resume in the term file matrix, and each column re presents one word, and each entry represents the freq uency count of a particular word in that particular res ume. This paper will produce vector type based on the number of frequencies, but Term Document Matrix c an not assign semantic meaning.

Using the Divisive Correlation Clustering Algorithm to detect varying patterns for job search, the Divisive Correlation Clustering Algorithm (DCCA) is used to create similar job patterns. This system created user clusters dynamically. Clustering solution from work expression datasets can be obtained by DCCA. For further study, other measures with similar properties may be used. DCCA can detect clusters with jobs with similar variation in expression patterns without taking the expected number of clusters as an input. The performance comparison results make DCCA more significant than the k-means algorithm. The DCCA algorithm continues to cluster until all clusters have only positively correlated work sets.

3. Data Mining

Data mining is the process by which data patterns are extracted. The training set is called the database from which the data mining system tries to extract knowledge. The system attempts to create general rules and descriptions of patterns and relationships in the database by examining the data in this database. The goal is to acquire knowledge that is valid for other similar data as well as for the specific database considered .Using a pre-defined set of rules, a deductive system reasons for data. These rules limit how information can be used to draw conclusions and infer information from the deductive system. A correct deductive system is inferring information which is a logical consequence of the database contents.

4. Background Theory

Clustering is a process of partitioning a set of data (or objects) in a set of meaningful sub-classes, called clusters

4.1. Cluster

A collection of data objects that are “similar” to one another and can be treated collectively as one group. But as a collection, they are sufficiently different from other groups.

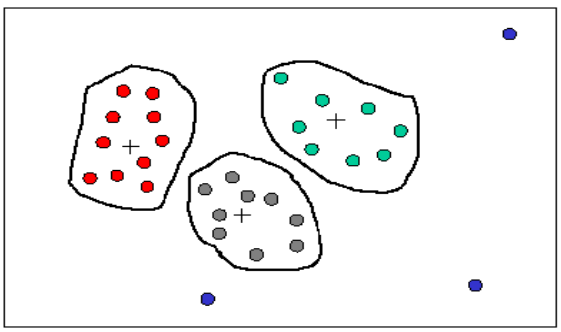


Figure1: Sample of Clustering

4.2 Clustering Methodologies

There are two general methodologies for clustering:

1. Partitioning Clustering

2. Hierarchical Clustering

Partitioning Clustering

Partitioning clustering is used to classify observations, within a data set, into multiple groups based on their similarity. The algorithms require the analyst to specify the number of clusters to be generated.

1) **K-means clustering**

Each cluster is represented by the cluster's c enter or means of the cluster's data points. T he Kmeans method is sensitive to data point s and outliers that are anomalous.

2) **K-medoid clustering**

Each cluster is represented by one of the objects in the cluster. PAM is less sensitive to outliers compared to k-means.

3) **CLARA algorithm** is an extension to PAM adapted for large data sets.

Hierarchical Clustering

Hierarchical clustering, also known as *hierarchical cluster analysis*, is an algorithm that groups similar objects into groups called *clusters*. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

Hierarchical clustering can be performed with either a *distance matrix* or *raw data*. When raw data is provided, the software will automatically compute a distance matrix in the background.

Hierarchical clustering starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps:

(1) Identify the two clusters that are closest together, and

(2) Merge the two most similar clusters. This continues until all the clusters are merged together.

5. K-Means Algorithm of Proposed System

The basic algorithm (based on reallocation method):

1. Select *K* initial clusters by (possibly) random assignment of some items to clusters and compute each of the cluster centroids.

2. Compute the similarity of each item x_i to each cluster centroid and (re-)assign each item to the cluster whose centroid is most similar to x_i.

3. Re-compute the cluster centroids based on the new assignments.

4. Repeat steps 2 and 3 until three is no change in clusters from one iteration to the next.

Cluster-centroid= $\mu_k = \frac{1}{N_k} \sum_{q=1}^{Nk} x_q$ ..... (1)

N_k= the number of instances belonging to cluster k

μ_k= the mean of the cluster k.

5.1. Word Embedding

The vector representations of words are very useful in different natural language processing tasks in order to capture the semantic meaning of words. The main idea is to find continuous **vector space representations** that carry semantic and syntactic meaning. Word embedding is the collective name for a set of language modelling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers. In very simplistic

terms, word embedding are the texts converted into numbers and there may be different numerical representations of the same text. A word embedding format generally tries to map a word using a dictionary to a vector. Many of machine learning algorithms and almost all deep learning architectures are incapable of processing strings or plain text in their raw form.

5.2. Prediction-based Embedding

Word2vec is a group of related models that are used to produce word embeddings. Such models are shallow, two-layer neural networks equipped to recreate verbs' linguistic meanings. Word2vec takes as its input a large corpus of text and produces a vector space, typically several hundred dimensions, with a corresponding vector in space being assigned to each unique word in the corpus. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located close to one another in the space.

Word2vec is a neural network architecture, developed by Mikolov et al. in 2013, to transform a word to a new vector space.

There are two different types:

1) Continuous Bag-of-Words model (CBOW) : Predicting a center word from the surrounding context

2) Skip-gram model : Predicting surrounding context words given a center word

5.2. Skip-gram model

Skip-gram model in Word2vec has a better accuracy and consistency in the linguistic meaning of words than other models. In this system, a neural network model with one hidden layer in which one-hot vector presents the input and output data for the training of the dataset. This network's purpose is to attempt to predict other neighboring words based on a single target word. Mapping between one-hot vectors to new vectors with higher dimensions through weight matrices which present the distribution of given input. The distance between words is calculated through Cosine distance.

6. Implementation of System

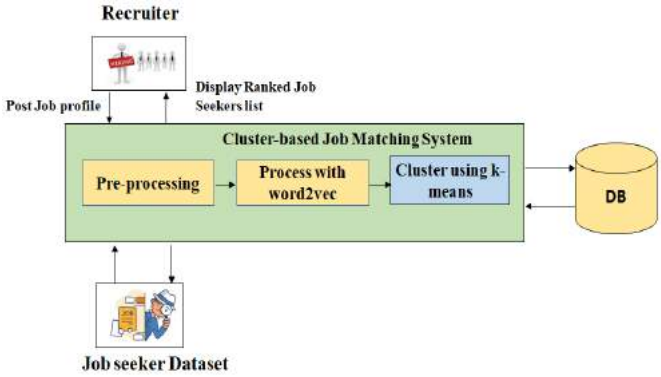


Figure2: Overall System Architecture

Inefficiencies in the labor market such as friction in matching members to jobs and the existence of skill gaps in various sectors of the economy are considered to be major problems facing economies today. The central premise is that increasing a workforce member's productivity depends crucially on identifying and recommending skills whose acquisition will yield that member's highest utility gains. To this end, this system develops work-related models [By K-Means] that can match members with other similar members as well as relevant members' jobs using shared features.

In this system, there are three users: (1) Job Recruiter (2) Job Seeker (3) Administrator.

Firstly, job seekers post their profiles and job recruiters post their job vacancies. Job seeker data is saved into database. And then, the administrator retrieve attributes of degree, job title and skill sets to convert semantic vectors using word2vec.The

administrator clusters the job seekers' profiles from their posted data as shown in fig (3).For the recruiter, Like the job seeker flow diagram, firstly new job vacancy is posted and saves into the database. Attributes that need to convert semantic vectors, degree, job title and skill set are retrieved. And then, administrator match the posted recruiters' job post with the pre-clustered job seekers as shown in fig (4). This system do evaluation for both software and hardware jobs.

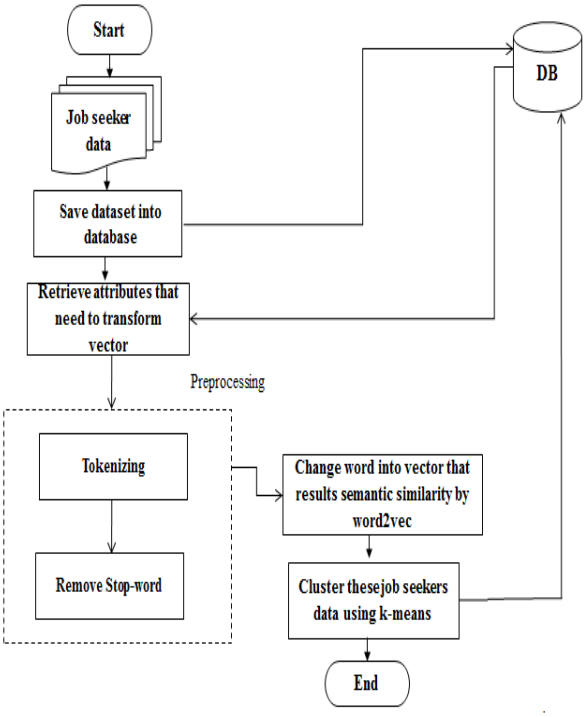


Figure3: Job Seeker Flow Diagram

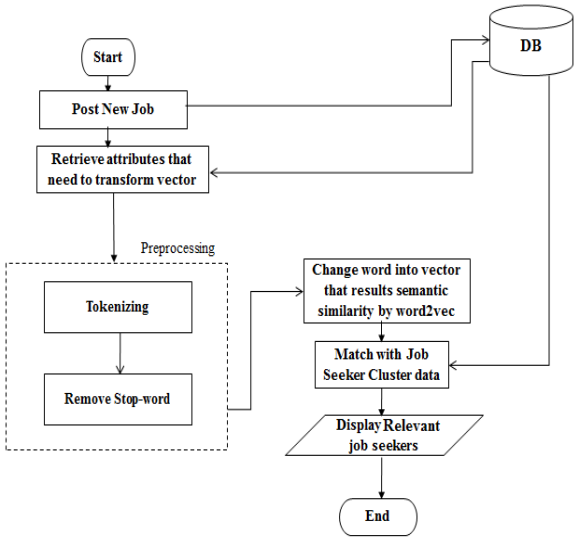


Figure4: Job Recruiter Flow Diagram

7. System Evaluation of Job Matching

Accuracy = TP+TN/TP+FP+FN+TN …….. (2)

		Predicted Class	
		Software Job Seekers	Hardware Job Seekers
Actual Class	Software Job Post	True Positive	False Negative
	Hardware Job Post	False Positive	True Negative

Table1: Accuracy Formula Table

	Relevant Job Seekers	Irrelevant Job Seekers	Accuracy %
Software Job Post	130	20	0.866666667
Hardware Job Post	121	29	0.806666667

Table2: System Evaluation for Job Matching

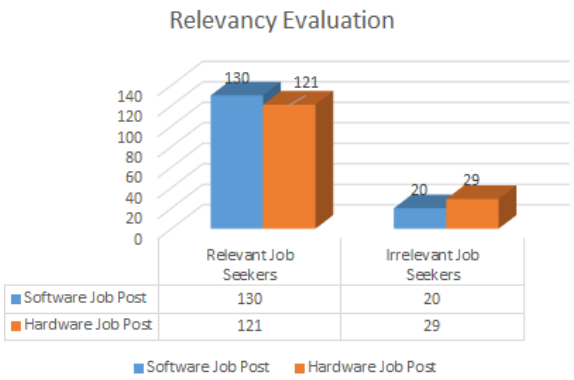


Figure5: Relevancy Evaluation Diagram

We evaluate the relevancy of matching between job seekers and job recruiters by using equation (2).We do evaluation for the relevancy of software job post and hardware job post.

During the clustering process the training data of job seekers posts "1000" and number of clusters "5" is used. According to the experimental result, true rate for relevant software job seeker posts are 130 and irrelevant posts are 20 records and true rate of relevant hardware job posts are 121 and irrelevant posts are 29 as shown in figure 5.Therefore, the accuracy rate of software job and hardware job are 87% and 81% respectively as shown in table (2).

8. Conclusion

This system proposed a cluster-based job matching process that supports effective and easy communication between recruiters and job seekers. Consequently, this system can solve the difficulties of traditional recruitment system and reduce manual work via posting in many job websites. Using k-means clustering and word2vec model, this system can give the most relevant job seekers data for recruiters.

REFERENCES

1) May Thu Naing, "Job Recommender System On Android Smart Phones By Using Hybrid APPROACH" , University of Computer Studies, Yangon,2014

2) Mayuri Verma, "Cluster based Ranking Index for Enhancing Recruitment Process using Text Mining and Machine Learning" ,India ,2017

3) A. Drigas, "An expert system for job matching of the unemployed", Department of Technological Applications, 2015

4) John von Neumann, "Skill2vec: Machine Learning Approaches for Determining the Relevant Skill from Job Description" , Vietnam, 2017

5) Anika Gupta, "Applying Data Mining Techniques in Job Recommender System for Considering Candidate Job Preferences" , India ,2014

6) Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013. p. 1301-3781.

7) InternSystems, KES2017, 6-8 September 2017, Marseille, France national Conference on Knowledge based and Intelligent Information and Engineering

8) Altszyler E, Sigman M, and Slezak DF. Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. ArXiv preprint arXiv: 1610.01520; 2016.

Total Order Based Database Replication in Banking System

Zin Phyu Phyu Phway, Sabai Phyu
University of Computer Studies, Yangon
zinphyuphway@gmail.com,sabaiphyu@ucsy.edu.mm

Abstract

Replication is the one of the good way to increase the performance of database system by separate out database in different servers. Database replication is used to provide high availability and performance for accessing data. In order to distribute database load to different database servers rather having a single server we replicate data. In this way work load on single server can be decreased. This system may lead to concern about the data consistency between replicated database. In this system, total order based database replication technique is used to ensure consistency and ease of access from any remote location of a banking system.

Key words: ***Replication, data consistency, banking system, total order.***

1. Introduction

Nowadays, commerce is larger and larger and it affects to open new branch offices in different locations. In this way, work load on single server can be decreased. Replication is the process of making a replica (a copy) of something. It is the process of copying data from a data store or file system to multiple computers that store the same data for the purpose of synchronizing the data. An account is a document issued by the branch of a Bank, which entities its holder to be one of the account holder of the Bank. An account is issued by one of the branch can be processed from the other branch. The transaction processed at one of the branch can be reflected to the other branch of the replicated data. So, the account holder no need to determine where the transaction can be made and no need to concern the physical address of the dedicated branch.

As the system used the replicated database, the user can process their account any time, everywhere. Active replication is a non-centralized replication technique. Its key concept is that all replicas receive and process the same sequence of client requests. Consistency is guaranteed by assuming that, when provided with the same input in the same order, replicas will produce the same output. Clients do not contact one particular server, but address servers as a group. In order for servers to receive the same input in the same order, client requests can be multicast to servers. The main advantage of replication is its simplicity (e.g., same code everywhere) and failure transparency.

2. Related Work

The related works of replication are discussed in this session.

In distributed system [1], the database can be replicated in multiple servers stored at different sites. Certain data items may be redundantly stored at more than one site for reliability reasons. Replication is a key to providing high availability and fault tolerance in distributed systems. A fault-tolerant service always guarantees strictly correct. There are two fault-tolerant replication systems: passive replication and active replication. This system intends to implement an online banking system by using active replication.
Advantages: The user can perform any type of banking operation (Deposit, Withdraw, Transfer) at their convenient bank and no need to concern to go their account registered bank. Although, the operations are performed at the branch of desire location, the system replicates the committed processes to all branch via the merge server for the future time of accessing to be a consistence state data.
Disadvantages: If one of the bank is disconnected from the other banks, the disconnected bank's data is only in that branch and can't access from the other branch. So, the proposed merge replicated data are only in the connected state and not suitable for offline condition.

In this system [2], if the client makes a request to the servers through the group coordinator, the group coordinator checks to all replicas are connected or not. If all replicas or more than one replicas are connected altogether, the group coordinator multicast the request to all connected replicas. Simultaneous transaction processing are

made at each replica. No coordination is necessary, as all replica process the same request in the same order.

Advantage: All replicas do the same process stages with same order for both read and write operations so that all replica have been data consistence any time. Therefore, the user can get up-to-date reliable information from this system. Disadvantage: The propose system is only developed for a bank and it maintains three replica servers for data recovery. This banking system cannot be supported for many other branches, and internal bank and external bank transaction processing control by the user of active replica servers.

3. Motivation

The nodes convey by trading messages. The messages can be lost, when the node crash happen. The system utilizes the administrations of a group communication layer which gives solid multicast informing ordering ensures (total order). The group communication framework likewise gives a participation notification administration, illuminating the replication motor about the nodes that can be come to in the present segment. The notification happens each time an availability change. The arrangement of members that can be come to by a server at a given minute in time is known as a view. The replication layer handles the node accidents utilizing the notifications gave by the group communication. The essential property gave by the group communication framework is called Virtual Synchrony and it ensures that procedures moving together starting with one view then onto the next convey the equivalent (ordered) arrangement of messages.

4. Background Theory

Replication is a typical way to deal with accomplish adaptation to non-critical failure in an appropriated framework to such an extent that replicas give repetition if there should arise an occurrence of a disappointment of a server. Two principle classes of replication are the active and passive replications. In passive replication, client manages one replication and the essential sends messages to the optional to refresh their views. A client makes an impression on the entirety of the copies in active replication and the conditions of the replicas are kept up as indistinguishable, by and large, utilizing limited state machines. To guarantee consistency of the replicas, a group communication crude called the Total Order Multicast might be utilized which ensures that the requests by the clients are gotten by all copies in a similar order.

A group membership administration deals with a group of procedures and depends on the view which is the rundown of procedures having a place with a group. View change ought to be informed to all members. There are three fundamental tasks expected to oversee group membership adequately; join, leave and avoid. Join is executed by a procedure p endless supply of it, the entirety of the procedures update their view. All the more critically, the condition of the group should be moved to the new part p. A procedure will be expelled from a group by avoidance if its accident is identified by an individual from a group and exit is an intentionally arrival of a procedure from a group without anyone else.

The group the executives module ought to likewise give the two natives; send multicast to make an impression on all individuals and get multicast to get a message sent by an individual from the group. These two natives can be acknowledged utilizing different methodologies, for example, dependable communicate, solid FIFO broadcast and total order multicast. Solid Broadcast of a message in a group guarantees that messages are conveyed by all procedures or none.

4.1. Group Communication Service

The group communication is an adequate framework for providing the multicast primitives required to implement replication. The group communication service has to manage changes in the group's membership while multicasts take place concurrently.

A group membership service has four main tasks, as follows:

- Providing an interface of group membership changes.
- Implementing a failure detector.
- Notifying members of group membership changes.
- Performing group address expansion.

4.2 Database Replication

Database replication can be utilized on numerous database management systems, for the most part with an master/slave connection between the first and the duplicates. The master logs the updates, which at that point swell through to the slaves. The slave yields a message expressing that it has gotten the update effectively, in this manner permitting the sending (and possibly re-sending until effectively applied) of resulting refreshes.

Multi-master replication, where updates can be submitted to any database node, and afterward swell through to different servers, is regularly wanted, yet presents generously expanded expenses and intricacy which may make it unfeasible in certain circumstances. The most widely recognized test that exists in multi-master replication is value-based clash anticipation or goals. Most synchronous or anxious replication arrangements do struggle counteractive action, while asynchronous arrangements need to do compromise.

For example, if a record is changed on two nodes all the while, an anxious replication framework would distinguish the contention before affirming the submit and prematurely end one of the exchanges. An apathetic replication framework would enable the two exchanges to submit and run a compromise during resynchronization. The goals of such a contention might be founded on a timestamp of the exchange, which chooses reliably on all nodes.

5. Total Order to Multicast Replication Algorithm

Let: Red Action = action within the local server by the group communication layer (not yet, determine the global order).

Green Action = action for which the server has determined the global order.

White Action = action for all server (all of the servers have already marked it as green.)

[These actions can be discarded since no other server will need them Sub-sequently]

T = user requested transaction variable.

BEGIN

Step1: List the servers in the communication layer;

Step2: Identifies a single server as a primary server. Other servers are non-primary servers.

Step3: Accept the request transaction: = T;

Step4: Make a transaction processing at local server.
Transaction (T) processing status marked as Red Action.

Step5: After the transaction (**T**) has been committed at local server, mark the transaction **(T)'s** status as **Green Action at local server.**

Step6: Processed transaction (T)'s data update sent to the primary server as a global state. Mark the transaction (T)'s status as Green Action.

Step7: Primary Server propagate the Updated data to replica server

Step8: From total order to multicast system checks all the server data received status.

If (All servers status are in **Green Action**)
{
Mark the transaction (**T**)'s status as **White Action**.
Commit (**T**) for all distributed server.
}
Else
{
Maintain the transaction (**T**)'s status of primary server in **Green Action.**

GOTO Step6;
}
End If
END

5.1. From Total Order to Database Replication

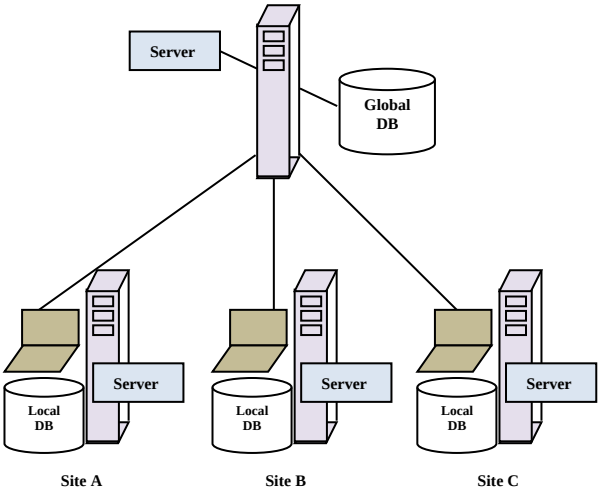


Figure1. The System Overview

In the communication network, the replication layer identifies a server as an essential segment; different parts in the system are non-essential segments as appeared in figure1. An adjustment in the membership of a part is reflected in the conveyance of a view-change notification by the group communication layer to every server in that segment. The coloring model defined the information level related with each activity. Every server denotes the activities conveyed to it with one of the accompanying colors:

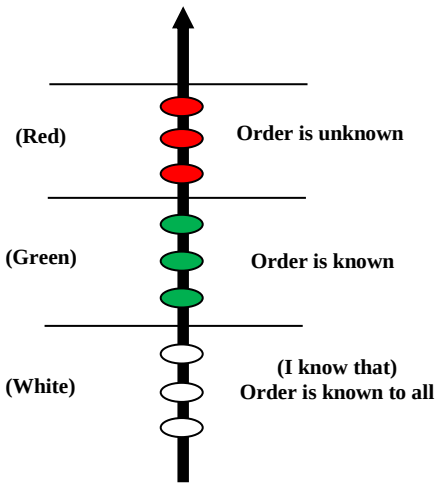
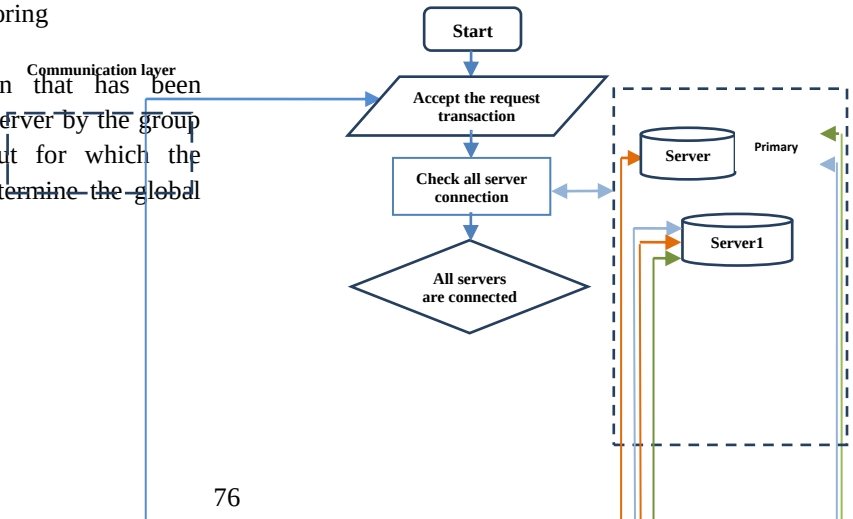


Figure2. Action Coloring

- **Red Action:** An action that has been ordered within the local server by the group communication layer, but for which the server cannot, as yet, determine the global order.
- **Green Action:** An action for which the server has determined the global order.
- **White Action:** An action for which the server knows that all of the servers have already marked it as green.

At every server, the white activities go before the green activities which, thus, go before the red ones. An activity can be stamped distinctively at various servers; in any case, no activity can be checked white by one server while it is missing or is checked red at another server. The activities conveyed to the replication layer in an primary part are stamped green. Green activities can be applied to the database promptly while keeping up the strictest consistency prerequisites as appeared in figure 3.



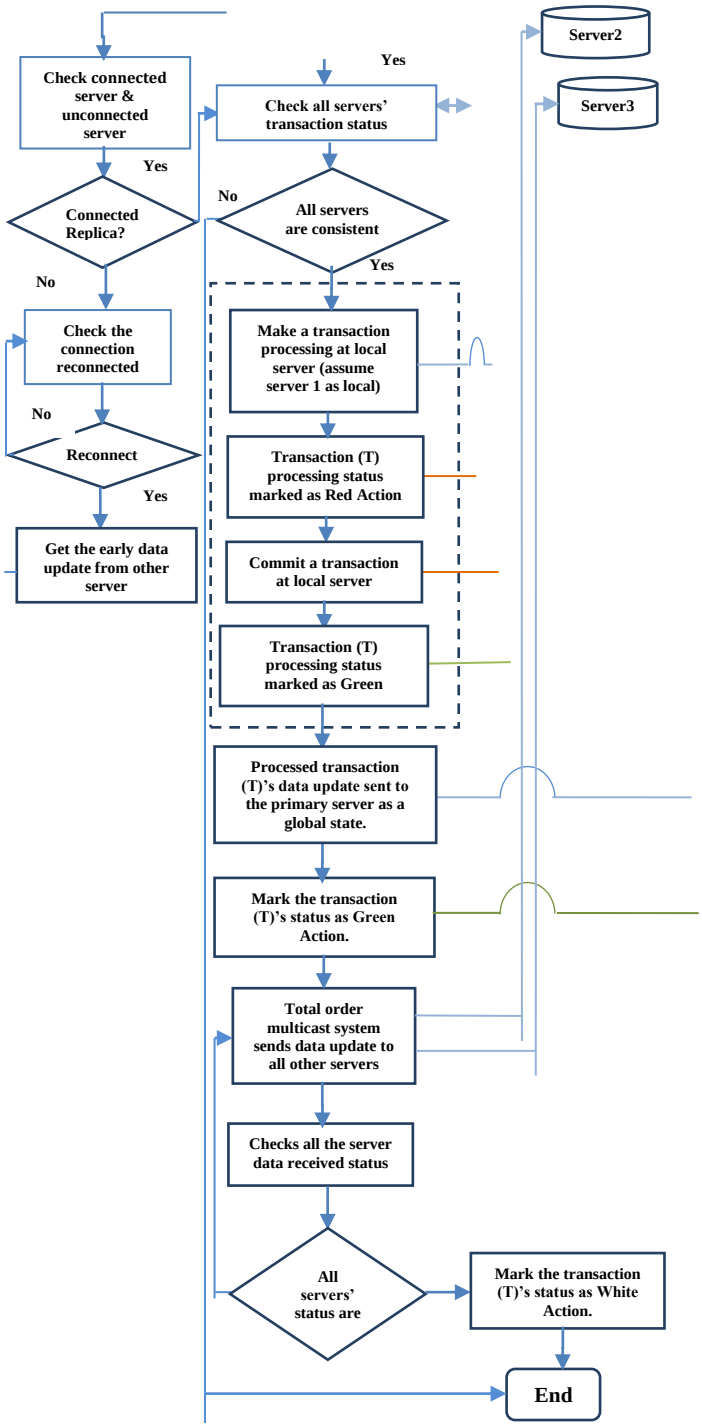


Figure3. The System Flow

5.2. Scope of the System

This system consists of three branches of bank and one control server for total order consistency control. At each branches, bank transactions (deposit, withdraw, transfer, account opening and so on) processing are take placed. But the data consistency between at the branches of bank is controlled by the server using total order multicasting algorithm.

Data consistency control can be easily noticed by the notification of color effect:

Red ➔Inconsistence,
Green ➔Consistence for locally,
White ➔Consistence for globally.

This system groups the primary server and all memberships as a primary component. In a primary component, the primary server receives the updated data from the data updated membership and multicast to all other participants in the primary component. So, the primary server is the major concern for the total order multicasting. Although this system allows the memberships' falling from the group, the primary server crashing can't support to continue the system.

6. Conclusion

Replication is valuable for improving performance and availability of information. Although the main problem of distributed database system is data consistency between all the replicas, this system can control the data consistency for replicated system. In this system, the user no need to concern about the memberships and the multicast algorithm will deliver the update data information to all. Even if one site becomes unavailable, user can continue to query or even update the remaining site.

REFERENCES

1. "Implementation of Distributed Database System using Merge Replication", Ko Ko Aung, M.C.Sc 2017.

2. "Implementation of Database Consistency by Active Replication", Khin Kaung San, University of Computer Studies, Yangon, M.C.Sc, 2016.

3. C.J.Date, "*An Introduction to Database Systems*", Seventh Edition, Addison Wesley, 2011.

4. Elmasri Navathe., "*Fundamentals of Database Systems*", fourth edition, 2004.

5. O.Amir,Y.Amir,andD.Dolev. Ahighly available application in the Transis environment. Lecture Notes in Computer Science, 774:125–139, 1993.

6. Y. Amir. Replication Using Group Communication over a Partitioned Network. PhD thesis, Hebrew University of Jerusalem, Jerusalem, Israel, 2015.

7. Y. Amir, C. Danilov, M. Miskin-Amir, J. Stanton, and C.Tutu. Practicalwide-area database replication. Technical ReportCNDS2002-1, JohnsHopkinsUniversity,Centerfor Networking and DistributedSystems, 2002.

8. Michael Kifer, Arthur Bernstein and Philip M. Lewis., "Database systems, An application oriented approach".

9. R. Soukup, K.Delaney, “Inside Microsoft SQLserver 7.0: Design, archetecture and implementation”, Microsoft Press, 2009.

Opinion Mining System of Customer Reviews by using Feature Extraction (Case Study: Tourism Review)

Nandar Moh Moh Lwin, Wai Wai Lwin
University of Computer Studies, Yangon
nandarmohmohlwinucsy@gmail.com,wai2lwin@ucsy.edu.mm

Abstract

Due to the dramatic improvement of ecommerce, web sources rapidly emerge which is important for both potential customers and service providers in prediction and decision purposes. Opinion mining techniques become popular to automatically process customer reviews by extracting features and user opinions expressed over them. To overcome the task of manual scanning through large amount of one by one review, people have interested to automatically process the various reviews and to provide the information which is useful for customers and service providers. Frequency-based feature extraction is used to extract frequent features from the review that are mentioned by the customers. By applying dependency relations, it can properly identify the semantic relationships between features and opinions of each review. It can find the numeric score of all the features using SentiWordNet. This system is intended to collect customer reviews from tourism field and then extract the related features and opinions to rate the services. Finally, it can rank each agency according to the final result of each review sentence.

Keywords: review, SentiWordNet, opinion, features, dependency relations

1. Introduction

Nowadays, there are many customers who take the services via websites. Many E-commerce enterprises often ask their customers to review the products which they have purchased and the associated services. Manufacturers can target the product features by reading the reviews from online websites. So, manufacturers can identify which elements of a product affect sales most and what are the features the customer likes or dislikes. As e-commerce is becoming more and more popular, the number of customer reviews grows rapidly. Therefore, opinion mining is also growing research area for both in natural language processing and information retrieval communities as it aims at finding subjective information, which may be more relevant to users to get useful information in many applications. A significant number of websites, blogs and forums allow customers to post reviews for various products or services (e.g., tripadvisor.com, amazon.com, and so on). These reviews are important resource for customers to help for making their purchase decisions. Based on a collection of customer reviews, the major task of opinion mining is to extract customers' opinions and predict the sentiment orientation. An opinion mining algorithm is used to track and manage customer reviews, through mining topics and sentiment orientation from online customer reviews. Certain keywords mentioned in the customer review will be mined and will be extracted with the keywords, this proposed system will rate the services provided by each trip agency. It is intended to discover the opinion by using customer reviews which are available from Web.

Today, most of the travel agency do opinion mining to analyze customers' attitudes, opinions and feelings. In these days, customers use social media to share both their positive and negative experiences with travel agency. Sentiment analysis can identify positive reviews as well as negative reviews showing strength and weakness users write about online. Sentiment analysis is based on the algorithms using natural language processing to categorize pieces of writing as positive, neutral, or negative. The algorithm is designed to identify positive and negative words, such as "wonderful", "amazing", "excellent", "terrible", "unhappy" etc. Due to language complexity, sentiment analysis has to face at least a couple of issues. The main problem of sentiment analysis tool is contrastive conjunctions which means a sentence consists of two contradictory words (they have positive and negative meaning). Moreover, sentiment analysis has to face poor grammar and incorrect spelling.

2. Related Work

Nowadays, web opinion mining is more and more popular so that there are many research papers based on the customer reviews from online website such as Tripadvisor website. Cristian Bucur [5] collected hotel reviews to generate useful information for customer. By using unsupervised method and a lexical resource, he extracted the opinion words from online. System consists of two modules a content acquisition module for collecting the reviews from website and an analysis module to preprocess the extracted data and implements opinion mining process. Firstly, each hotel review is spited into sentences. After that tokenization process split a sentence into component word. And lastly the polarity of word is evaluated using SentiWordNet to show the customer which elements of hotel are the most customer likes. Hu *et. al* [8] used frequent item sets to extract the most relevant features from a domain and pruned it to obtain a subset of features. They extract the nearby adjectives to a feature as an *opinion word* regarding that feature. Using a seed set of labeled Adjectives, which they manually develop for each domain, they further expand it using WordNet and use them to classify the extracted opinion words as positive or negative. While some researchers focus

their studies on the impact of online product reviews on sales, an important question remains unanswered, that is, can online product reviews reveal the true quality of the product? To test the validity of this hypothesis, the authors in [7] use data from Amazon to test the underlying distribution of online reviews and try to answer this question. In summary, most of the current related work focuses on problems in opinion mining, product aspect rating, review summarization etc. To the best of our knowledge, there has been no focused study regarding ranking products based on customer reviews.

3. Background Theory

Sentiment Analysis

Sentiment analysis refers to the use of natural language processing to identify and extract the subjective information from online website. Holder (source) of attitudes, Target (aspect) of attitude and Type of attitude from a set of types like, love, hate, value, desired etc. The process of sentiment extraction is automatically processed. So, it saves time and effort. It is also one of the artificial intelligence that analyses sentiment data so that human emotional is sparse. Sentiment analysis is used in broad application such as e-commerce, digital marketing, travel plans and politics [9]. In online reputation management, it can be used to analyze web and social media mentions about a product, a service, a travel agency, a marketing campaign or a brand.

Opinion mining

Opinion mining is a research subtopic of data mining which aiming to automatically obtain useful knowledge. It has been widely used in real-world applications such as ecommerce, business-intelligence, information monitoring, and public polls. Opinion mining seeks to determine the sentiment, attitude or opinion of an author expressed in texts with respect to a certain topic. On the web, there are increasing numbers of review web sites, where users can post their comments on a service of a tour agency and provide their positive or negative evaluation [4]. These are important resources providing advice to new users and helping them with their travel plans.

Opinion are central to almost all human activities and are key influencers of their behavior. This is not only true for individuals but also true for organizations. Opinion mining refers to the use of natural language processing (NLP) to identify and extract subjective information from online website.

Opinion mining is widely used to reviews and social media for a variety of applications ranging from marketing to customer services [1].

Dependency Relation for Feature-Opinion Mining

Dependency grammars mean the structures as a set of dependency relationships. A dependency relationship shows the relationship between the specified features and its related opinion words. The following diagram shows the dependency relationship of a trip review. A review sentence is spited to form a dependency tree in the POS tagging step using Stanford Parser. From the generated tree, we can capture the dependency relations between trip feature and opinion word. The syntactic structure of a sentence consists of dependencies shown in Figure. 1.

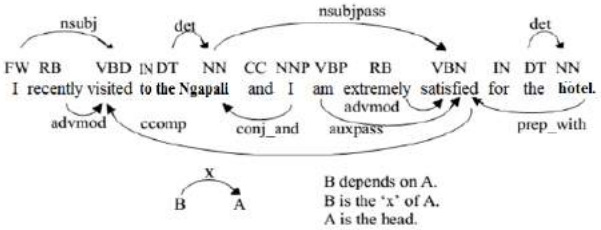


Figure 1: Sample of Dependencies relation

POS Tagging

POS tagging (or part of speech tag) reads text in some language and assigns part of speech to each other word (and other token), such as noun, verb, adjective and so on.

A set of all part of speech tags used in a corpus is called a tagset. The most common part of speech are noun (NN: singular, NNS: plural), verb (VB: base form, VBD: past tense, VBN: past participle, VBG: gerund), adverb (RB: adverb, RBR: comparative adverb, RBS: superlative adverb) and adjective (JJ: adjective, JJR: comparative adjective, JJS: superlative adjective) etc. Samples of POS tagging are as follows:

POS Tag	Description	Example
CC	coordinating conjunction	and
MD	Modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
PP	personal pronoun	I, he, it
RB	Adverb	however, usually
VB	verb be, base form	be
VBD	verb be, past tense	was, were

Extracting Features

The physical attributes of an object are called its features. Features are parsed as noun or noun phrases and are represented as _NN or _NNS. In general, the words those indicating most trip features are nouns or noun phrases. Therefore, the next step is to identify a noun phrase as a trip feature candidate. A linguistic filtering pattern is used to extract noun phrase. A process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. Feature extraction can also reduce the amount of redundant data for a given analysis [2]. There are two methods to extract such features including frequency-based feature extraction and dependency grammar-based feature extraction.

Frequency-based feature extraction

A set of nouns and noun phrases is gained per document. Determine the number of each of the nouns gained among total current lists. A new set including all words extracted is constructed and then, the frequency of each word is specified. Extract features

with a frequency lower than defined frequency threshold.

Term frequency–inverse document frequency

TF-IDF stands for term frequency–inverse document frequency. It is a formula to calculate how important a word is to a document. It works by increasing proportionally to the number of times a word appears in a document but is offset by the number of documents that contain the word. So, words that are common in every document, such as this, what, and if, rank low even though they may appear many times, since they don't mean much to that document in particular.

The **term frequency** of a word in a document is the way of calculating frequency with the simplest way. Term frequency can be calculated by taking the number of times a word appears in a document, dividing it by the total number of words in a document.

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document) Equation (1)

The **inverse document frequency** means how common or rare a word is in the entire document set. The value of a word is closer to zero, the more common a word is. This formula can be calculated by taking the total number of documents, dividing it by the number of documents that contain a word, and calculating the logarithm.

IDF(t) = log_e(Total number of documents / Number of documents with term t in it) Equation (2)

The **TF-IDF** is calculated by multiplying the results in the TF and IDF score of a word in a review document. The higher the score, that word is more relevant in that document.

TF-IDF(t)=TF(t)*IDF(t) Equation (3)

Dependency grammar based feature extraction

The polarities of consumers' sentiments on trip features are determined by analyzing dependency relationships between feature terms and sentiment terms. The Stanford parser is used to perform feature-sentiment term dependency parsing. The result of dependency parsing is dependency trees and a set of feature-sentiment term dependency pairs. Each review will generate a dependency tree. Then dependencies that possibly contain both a trip feature and a sentiment term will be identified [8]. It will search for feature terms dependencies to identify subject-predicate relationships (nsubj), verb-object relationships (dobj), adjectival modifying relations (amod), and relative clause modifying relations (rcmod). Dependency grammars mean the structures as a set of dependency relationships. A dependency relationship shows the relationship between the specified features and its related opinion words.

Extracting Opinion

Opinion words are usually feeling or attitudes of the writer. In this system, opinion words are extracted by using adjective words. Adjective words are represented as _JJ or _JJS. Extracting opinion words with relevant features are processed in this phase [2]. Extracting opinion words are very important in this step so that we need to capture the useful information in a document.

SentiWordNet

WordNet is English words which group into sets of synonyms called synsets and provides short, general definitions, and records the various semantic relations between these synonym sets. It distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules-it does not include prepositions, determiners etc. It can also use to find synonyms for all words. SentiWordNet (SWN) is an extension of WordNet, which is intended to augment the information about the sentiment of the words [3]. In SWN, each synset has a positive score, a negative score and an objectivity (neutral) score. The total values of these three scores are equal one, so they give an indication of the relative strength of the positivity, negativity and objectivity of each synset. It is made up with tens of thousands of words, there meanings, part of speech represented and the degree of positivity and negativity of the word, ranging from 0 to 1. Since same words can have different meanings with respect to the part of speech being represented, sentiwordnet was designed by ranking subjectivity of all terms /synsets according to the part of speech the term belongs to. The parts of speech represented by the sentiwordnet are adjective, noun, adverb and verb which are represented respectively as 'a', 'n', 'r', 'v'. The database has five columns, the part of speech, the offset which is a numerical ID, that when matched with a particular part of speech, identifies a synset: positive score, negative score (bottom from 0 to 1) and synset terms.

Ranking the Features

The overall weight of each review is calculated by adding the score of the opinion word with the number of sentences as following equation (4).

$$\text{TotalWeight}=\sum_{t=1}^{n}(wt\_of\_positive\_feature - wt\_of\_negative\_feature)$$

Equation (4)

where n is the number of sentence and t is the tem in this review. If the total weight of a feature is positive, then that review is termed as positive and is thought to be likely by the customer. Similarly, a negative weight indicates the feature is not liked by the user [6]. The total weight of the sentence or document is equal to the total weight of the positive features minus the total weight of the negative features.

Begin
Step 1: Read new review
Step 2: For each review, part-of-speech tagging and dependency relations are performed as preprocessing step.
Step 3: Trip feature candidates are extracted using Frequency-based feature extraction.
Step 4: Opinion words are extracted.
Step 5: The extracted opinion words are related with corresponding features by using dependency relation.
Step 6: Then, the sentiment orientation and score of the opinion words are identified with the help of SentiWordNet according to the Equation (4).

Step 7: Calculate the total weight of the document/review according to the total weight of these features.
Step 8: Find the specified features using Term Frequency-Inverse Document Frequency (TF-IDF) to show the reviewer.
8.1: Count the features in a review.
8.2: Sum the total counted times of features.
8.3: Calculate the Term Frequency using Equation (1).
8.4: Calculate the Inverse Document Frequency using Equation (2).
8.5: Calculate Term Frequency- Inverse Document Frequency (TF-IDF) using Equation (3).
End

3. Overview of the System

The system has two portions such as reviewer and viewer. From viewer portion, it can easily view all positive and negative features of review for each agency. From reviewer portion, the trip reviews are put in the review corpus as inputs. And then, the system will produce the output as ranking features by summarizing the reviews. The general steps for the ranking features are as follows:
Step 1: For each review, part-of-speech tagging and dependency relations are performed as preprocessing step.
Step 2: Feature candidates are extracted by frequency based feature extraction method.
Step 3: Opinion words are also extracted in this step.
Step 4: The extracted opinion words are related with corresponding features according to the dependency relation.
Step 5: Then, the sentiment orientation and score of the opinion words are identified with the supporting of SentiWordNet.
Step 6: Ranking the features according to the total weight of these features.
Step 7: Find the specified features using Term Frequency- Inverse Document Frequency (TF-IDF) to show the reviewer.

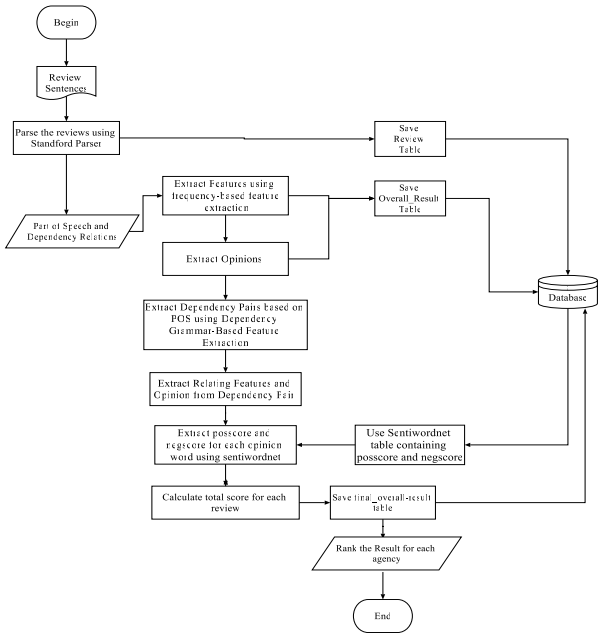


Figure 2: System Design for Reviewer

Sample of Customer Review

Most of the reviews are collected from the Tripadvisor.com and these reviews are not only structured sentences but also unstructured sentences. A structured sample review sentence is shown in below:

"I just got back from my wonderful trip to Bagan and I would like to thank Myanmar Tour Asia Agency for making sure that I got a memorable trip I have ever been. The staffs were patient and informative whenever I had some inquiries about the service or the trip or the food. All of the staffs knew exactly how to treat the guests with love and care. In fact, I love the food that they served during the trip. (My mouth is now watering just by the thought of that.) Overall, I would certainly recommend everyone to go on a trip to Bagan with Myanmar Tour Asia Agency."

Step 1: In this step, each review sentence is parsed using Stanford parser, which provides POS tagging. By using Stanford Parser, the trip review sentence is transformed into the dependency relations. The relationship between every pair of words is called a dependency relations and it is illustrated in Figure 1.

POS Tagging

I/PRP	I/PRP	sure/JJ
just/RB	would/MD	that/IsN
got/VBD	like/VB	I/PRP
back/RP	to/TO	got/VBD
from/IN	thank/VB	a/DT
my/PRP$	Myanmar/NNP	memorable/JJ
wonderful/JJ	Tour/NNP	trip/NN
trip/NN	Asia/NNP	I/PRP
to/TO	Agency/NNP	have/VBP
Bagan/NNP	for/IN	ever/RB
and/CC	making/VBG	been/VBN

Universal dependencies

nsubj(got-3, I-1) advmod(got-3, just-2) root(ROOT-0, got-3) compound: prt(got-3, back-4) case(trip-8, from-5) nmod:poss (trip-8, my-6) amod(trip-8, wonderful-7) nmod(got-3, trip-8) case(Bagan-10, to-9) nmod(got-3, Bagan-10) cc(got-3, and-11) nsubj(like-14, I-12) aux(like-14, would-13) conj(got-3, like-14) mark(thank-16, to-15) xcomp(like-14, thank-16) compound(Agency-20, Myanmar-17) compound(Agency-20, Tour-18) compound(Agency-20, Asia-19) dobj(thank-16, Agency-20) mark(making-22, for-21) acl(Agency-20, making-22) xcomp(making-22, sure-23) mark(got-26, that-24) nsubj(got-26, I-25)f ccomp(making-22, got-26) det(trip-29, a-27) amod(trip-29, memorable-28) dobj(got-26, trip-29) nsubj(been-33, I-30) aux(been-33, have-31) advmod(been-33, ever-32) acl:relcl(trip-29, been-33)

Step 2: Trip feature candidate are extracted. Then unfrequented features are removed.

Statements	Features (NN/NNS) Nouns
1	Trip
2	Staff, Inquiries, Services, Trip, Food
3	Staffs, Guests, Love, Care
4	Fact, Food, Trip
5	Mouth, Thought

6	Everyone, Trip

Step 3: Opinion words are also extracted from review.

No.	Opinion /JJ/JJR/JJS Adjectives	RB/RBR/ RBS Adverbs	VB/VBD/VB Verbs
1	Wonderful, Sure, Memorable	Just, Ever	Got, Thank, Like, Making, Got, Have, Been
2	Patient, Informative		Were, Had
3		Exactly	Knew, Treat
4			Love, Served
5		Now, Just	Are, Watering
6		Certainly	Recommend, Go

Step 4: The extracted opinion words are related with corresponding features by using dependency relation.

Extracting Dependency Pairs (nsubj/dobj/ amod/ rcmod) based on POS

No.	Dependency Pairs
1	nsubj(got-3, I-1) nsubj(like-14, I-12) nsubj(got-26, I-25) nsubj(been-33, I-30) dobj(thank-16, Agency-20)dobj(got-26, trip-29) amod(trip-8, wonderful-7) amod(trip-29, memorable-28)
2	nsubj(patient-4,staffs-2)nsubj(informative-6, staffs-2) nsubj(had-9, I-8) dobj(had-9, inquiries-11) dobj(had-9, trip-17)
3	nsubj(knew-5, All-1) dobj(treat-9, guests-11)
4	nsubj(love-5, I-4)nsubj(served-10, they-9) dobj(love-5, food-7)
5	nsubj(watering-6, mouth-3)
6	nsubj(recommend-6,I-3) dobj(recommend-6, everyone-7)

Relating Features and Opinion from Dependency Pairs

No.	Extracting Features and Opinion
1	dobj(got-26, trip-29)amod(trip-8, wonderful-7) amod(trip-29, memorable-28)
2	nsubj(patient-4, staffs-2)nsubj(informative-6, staffs-2) dobj(had-9, inquiries-11)dobj(had-9, trip-17)
3	dobj(treat-9, guests-11)
4	dobj(love-5, food-7)
5	nsubj(watering-6, mouth-3)
6	dobj(recommend-6, everyone-7)

Step 5: Then, the sentiment orientation and score of the opinion words are identified with the supporting of SentiWordNet.

POS	Offset	PosScore	NegScore	SynsetTerms
V	02359340	0	0	got#2
A	01676517	0.75	0	wonderful#1
A	00399533	0.25	0.125	memorable#1
N	10405694	0.125	0	patient#1
N	01304570	0.125	0	informative#3
V	02740745	0	0.25	had#10
V	00078760	0	0.125	treat#3
V	01465668	0.5	0	love#1
V	00882948	0.5	0	recommend#2

Step 6: Ranking the features according to the total weight of these features according to the equation (4).

Total Weight of the review =(0-0)+(0.75-0)+(0.25-0.125) + (0.125-0)+(0.125-0)+(0-0.25)+(0-0.125)+(0.5-0)+(0-0)+(0.5-0)=0+0.75+0.125+0.125+0.125-0.25-0.125+0.5+0.5 =1.75(positive)

Trip feature candidates are extracted using Frequency-based feature extraction according to the Equation (2), Equation (3) and Equation (4), respectively.

Trip=5, Staff=2, Inquiries =1, Services =1, Food=2, Guests=1, Love=1, Care=1, Fact =1, Mouth=1, Thought =1, Everyone =1

TF('Trip',review1) = 5/18, IDF('Trip')=log(3/3) = 0

TF('Staff', review1) = 2/18, IDF('Staff')=log(3/1) = 0.48

TF('Inquires',review1)= 1/18, IDF('Inquires')=log(3/1) = 0.48

TF('Services', review1) = 1/18, IDF('Services')=log(3/3) = 0

TF('Food', review1) = 2/18, IDF('Food')=log(3/2) = 0.18

TF('Guests', review1) = 1/18, IDF('Guests')=log(3/3) = 0

TF('Love', review1) = 1/18, IDF('Love')=log(3/1) = 0.48

TF('Care', review1) = 1/18, IDF('Care')=log(3/1) = 0.48

TF('Fact', review1) = 1/18, IDF('Fact')=log(3/1) = 0.48

.

TF-IDF('Trip',Document1)= 5/18*0=0

TF-IDF('Staff',Document1)= 2/18*0.48=0.05

TF-IDF('Services',Document1)= 1/18*0=0

TF-IDF('Food',Document1)= 2/18*0.18=0.02

TF-IDF('Guests',Document1)= 1/18*0=0

.

According to the result, trip, staff, services, food and guest features are extracted as specified features.

4. System Implementation

The system is implemented with java programming language and MySql database. Customer reviews are collected from internet source. The step by step processes are implemented relatively and results are shown with related screenshots as following:

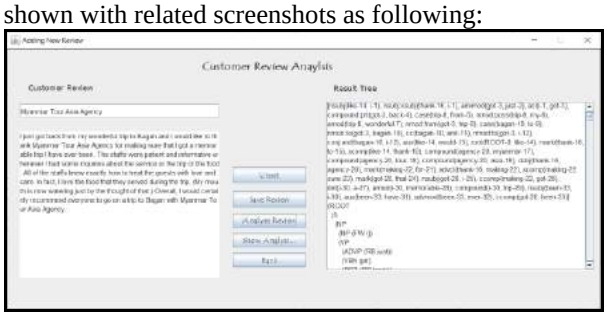


Figure 3: Customer Review Analysis Interface

Figure 4: Overall Result with Rating

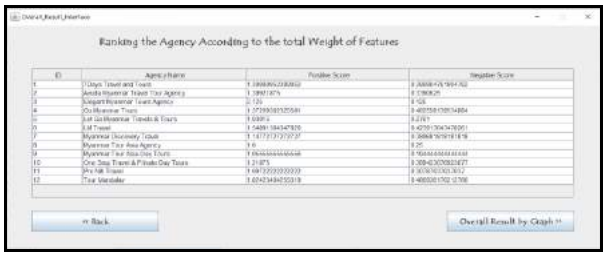


Figure 5: Agency According to the positive and negative weight of features

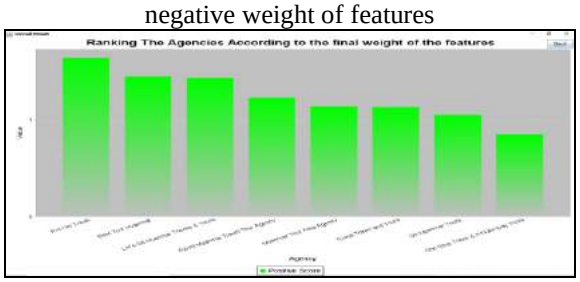


Figure 6: Overall Ranking Result for each Agency

5. Advantages of the System

This system has implemented opinion mining using SentiWordNet 3.0. It can be used in any kind of trip reviews and agency reviews and then can be analyzed the reviews. The analyzed reviews are useful to high level trip managers. Therefore, it is used in the important decision making process. The system can be used not only trip reviews but also other reviews by modifying features.

6. Performance Analysis

The performance analysis of the proposed system can be measured by computing its efficiency and its effectiveness. It measures accuracy result using F1 score. F1 score (also F1 score or F-measure) is a measure of a test's accuracy. It considers both the Precision and the Recall of the test to compute the score. Precision and recall are defined in terms of a set of retrieved documents (eg. the lists of documents produced by a search engine for a query) and a set of relevant documents (eg. the lists of all documents that are relevant for a certain topic).

This system is tested by using 60 trip reviews. First, the reviews are analyzed by using sentiment analysis module. Then, the output results are kept and checked with the manually pre-classified positive and negative reviews. After testing 30 positive pre-classified positive and negative reviews, the accuracy result can be seen in Table 1. F1 score is calculated by using precision and recall values.

Table 1 Evaluation Result of Opinion Mining

Number of Trip Reviews =60

Type	Relevant trip reviews	Retrieved trip reviews	Precision	Recall	F1 Score
Positive Reviews	30	25	0.9	0.8	0.85
Negative Reviews	30	35	0.85	0.6	0.7

7. Conclusion

In this paper it was presented an opinion mining platform for extracting and classifying hotel reviews posted by users on tourism websites. This system intended to develop a feature extraction system from tourism domain. It can parse the review sentences and identify the features efficiently. Then the weight of frequent features can be obtained and ranked these features according on their score values. As we showed the result as ranking, customers and administrators would know the trip features which are generally liked and disliked by the customer. Then, it can rank for each agency according to the final result of each agency. So, customer can get valuable facts which tour agent meet according to their desire. Moreover, each tour agency can know directly the strength and weakness of theirs so that trip agency can target in those trip features which are frequently mentioned by the customers or liked by the customers.

References

[1] Alok Choudhary, Zhang K., Narayanan R., and Choudhary A., (2009), "Mining Online Customer Reviews for Ranking Products", Technical Report, EECS department, Northwestern University.

[2] Ana-Maria Popesu and Oren Etzioni, (2005), "Extracting Product Features and Opinions from Reviews", Proceeding of Human Language Technology Conference and Conference on Empirical Methods in Natural Lagunage, ACL, Vancouver, pp. 339-346.

[3] Andrea Esuli and Fabrizio Sebastiani, (2006), "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining", In Proceedings of LREC-06, 5th Conference on Language Resources and Evaluation, Genova, IT, pp. 417-422.

[4] Cristian Bucurab,"Using Opinion Mining Techniques In Tourism",2nd Global Conference On Business, Economics, Management And Tourism, 30-31 Oct, 2014, Prague, Czech Republic.

[5] Deepanshi Sharma, Achal Kulshreshtha, Priyanka Paygude," Tourview: Sentiment Based Analysis On Tourist Domain", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3), 2015, 2318-2320.

[6] Kunpeng Zhang Ramanathan Narayanan, (2010), "Voice of the Customers: Mining Online Customer Reviews for Product Feature-based Ranking", Electrical Engineering and Computer Science Department Northwestern University.

[7] N. Hu, P. Pavlou, and J. Zhang, "Can Online Reviews Reveal a Product's True Quality? Empirical Findings and Analytical Modeling of Online Word-of-Mouth Communication", EC., 6 (2006), pp. 324-330.

[8] Qi Zhang, Yuanbin Wu, Tao Li, Mitsunori Ogihara, Joseph Johnson, Xuanjing Huang, "Mining Product Reviews Based on Shallow Dependency Parsing", SIGIR '09, Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009.

[9] Weishu Hu, Zhiguo Gong, Jingzhi Guo,(2010), "Mining Product Features from Online Reviews", Faculty of Science and Technology University of Macau , China.

Weather Prediction Using Hidden Markov Model

May Thagyan Aung, Thi Thi Soe Nyunt
University of Computer Studies, Yangon
maythagyanaung@ucsy.edu.mm, thithi@ucsy.edu.mm

Abstract

Weather prediction is an important role in meteorology and has been one of the most particularly and technologically challenging problems on all over the world [2]. Therefore, patterns on updating weather conditions are needed to observe. The aim of the proposed system is to cluster weather data and predict the forthcoming weather conditions. Thus, the proposed system includes K-means clustering algorithm and Hidden Markov Model (HMM). K-means clustering is used for clustering weather data. The clustering group of K-means clustering became the observation matrix of HMM. Hidden Markov Model (HMM) is used for prediction of next weather conditions. Three different datasets are used as case study and the performance measure involves accuracy of prediction result.

1. Introduction

Nowadays, several of the new technologies that predict the weather states have appeared. Weather forecasting is one of the most scientifically and technologically challenging problems around the world. The weather prediction has been one of the most interesting and fascinating domain since ancient times. Scientists have tried to forecast meteorological characteristics using a number of methods, some of these methods being more accurate than others. The forecasting of weather by computer is known as numerical weather prediction. To make an accurate prediction is one of the major challenges facing meteorologist all over the world. The data mining techniques are offered to analyze past data and prepare their activities. In this proposed work, the machine learning-based classification and prediction is studied in detail. These analyzed data help for decision making, business intelligence, and other technological task. The weather forecast is performed using these analyzed characteristics. The performance of any predictor is depended on the observations. Therefore, the Hidden Markov Model and K-means clustering is utilized. The K-means clustering is used to generate the observations from the input datasets. These observations are input for Hidden Markov Model. Hidden Markov Model includes two matrixes, namely, transition matrix, and observation matrix. Transition matrix is received by counting moving of current state to next state in the dataset. These two matrixes are trained in the Hidden Markov Model. And then, the prediction and performance evaluation is performed.

2. Related Work

In this section, the works related to weather prediction, are reviewed.

A.W.Robertson [1] proposed Hidden Markov Model for predicting daily rainfall occurrence over Brazil. Their experimental technique shows that its results are a great influence in optimizing the system performance and speeding the system up.

Sadegh Khanpour and Omid sojoodi [9] presented Hidden Markov Model for Clustering Web Users. Their experimental technique shows improvement attractive website for website designers, and predict future patterns of search engine users may be fruitful.

S.K Shanmuganathan [10] implemented Hidden Markov Model (HMM) for prediction model for spatio temporal trajectories. HMM is used predicting rainfall storm movement using HMMs by considering states that are related to clusters obtained by clustering points in the overall storm trajectories.

Rohit Kumar Yadav and Ravi Khatri [2] implemented weather forecasting using k-means clustering and Hidden Markov Model. K-means is used for clustering of weather dataset. Hidden Markov Model is used for prediction of weather conditions.

3. K-means

Clustering is a process of partitioning a set of data (or objects) in a set of meaningful sub-classes, called clusters. And it is a collection of data objects that are “similar” to one another and thus can be treated collectively as one group but as a collection, they are sufficiently different from other groups.

Clustering is the grouping of similar instances/objects, some sort of measure that can determine whether two objects are similar or dissimilar is required. There are two main type of measures used to estimate this relation: distance measures and similarity measures. Many clustering methods use distance measures to determine the similarity or dissimilarity between any pair of objects. The following formula is the formula to calculate the distance between each object x_i and each cluster center m_i, and to assign each object to the nearest cluster.

K-means algorithm is type in partitioning method. In case of partitioning method there is partitioning of an objects into k-clusters. It is typical clustering approach via partition datasets iteratively. In this there is division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one cluster. The concept is use K as a parameter, Divide n object into K clusters, to create relatively high similarity in the cluster, relatively low similarity between clusters.

Euclidean distance computes the root of square difference between co-ordinates of pair of objects. The Euclidean distance is using in equation (1).

$$d(x_i, m_i) = \sqrt{\sum_{j=1}^{d}\left(x_i - m_{j1}\right)^2},\ \text{i=1…N, j=1…..k} \quad (1)$$

d (x_i, m_j) is the distance between data i and cluster j.

4. Hidden Markov Model

Hidden Markov Model (HMM) is a statistical Markov Model in which the system being modeled is assumed to be a Markov process with unobservable (i.e. hidden) states. The adjective hidden refers to the state sequence through which the model passes, not to the parameters of the model; the model is still referred to as a hidden Markov model even if these parameters are known exactly.

The state transition probability matrix A = $(a)_{ij}$ is using in equation (2).

$$(a)_{ij} = a_{ij} = P(X_1 = j \mid X_0 = i) = P(X_{n+1} = j \mid X_n = i) \text{ for any n} \quad (2)$$

a_{ij} is the probability of making a transition from state i to state j in a single step.

$a_{ij} > 0$ for all i, j. This is using in equation (3).

$$\sum_{j=1}^{N} a_{ij} = 1\ ,\ 1 \le i \le N \quad (3)$$

The remark symbol probability matrix B = $\{b_j(k)\}$ is using in equation (4).

$$b_j(k) = P\left(V_k \middle| S_j\right), 1 \le j \le N\ , 1 \le k \le M \quad (4)$$

$$\sum_{k=1}^{M} b_j\,(k) = 1, 1 \le j \le N \quad (5)$$

The initial state probability vector is using in equation (6).

$$\pi_i = P(q_1 = S_i),\quad 1 \le i \le N \quad (6)$$

$$\sum_{i=1}^{N} \pi_i = 1 \quad (7)$$

For a particular hidden state sequence $Q=q_0,q_1,q_2,\ldots,q_T$ and an observation sequence $O=o_1,o_2,\ldots, o_T$, the likelihood of the observation sequence is using in equation (8).

$$P(O|Q) = \prod_{i=1}^{T} P(o_i|q_i) \quad (8)$$

The joint probability of being in a particular weather sequence Q and generating a particular sequence O is using in equation (9).

$$P(O, Q) = P(O|Q) * P(Q) = \prod_{i=1}^{T} P(o_i|q_i) * \prod_{i=1}^{T} P(q_i|q_{i-1}) \quad (9)$$

4.1 Transition matrix

Transition Probability represents the probability of moving from current state to next state. The state of the system at time t+1 depends only on the state of the system at time t. In this transition matrix, the weather conditions and the next weather conditions are organized based on the matrix [4].

4.2 Observation matrix

That is the matrix on which the clustered data is arranged in order to develop the observation matrix of the Hidden Markov Model. Additionally,

the data objects class labels are recognized here as the states of the events. These states are the natural events or the weather conditions which are required to predict.

5. Overview of the Proposed System

Figure 1 shows the overview of the proposed system. According to figure, clustering is performed to the input dataset. The proposed system can be classified by the two stages. In the first stage, clustering is performed by using K- mean clustering algorithm. In the second stage, prediction is performed by using Hidden Markov Model. The clustered data became the observation matrix of the Hidden Markov Model (HMM). Transition matrix is received from weather states in weather dataset. These two matrixes is trained in the model and then predict the next weather condition. The proposed work includes performance accuracy of prediction result.

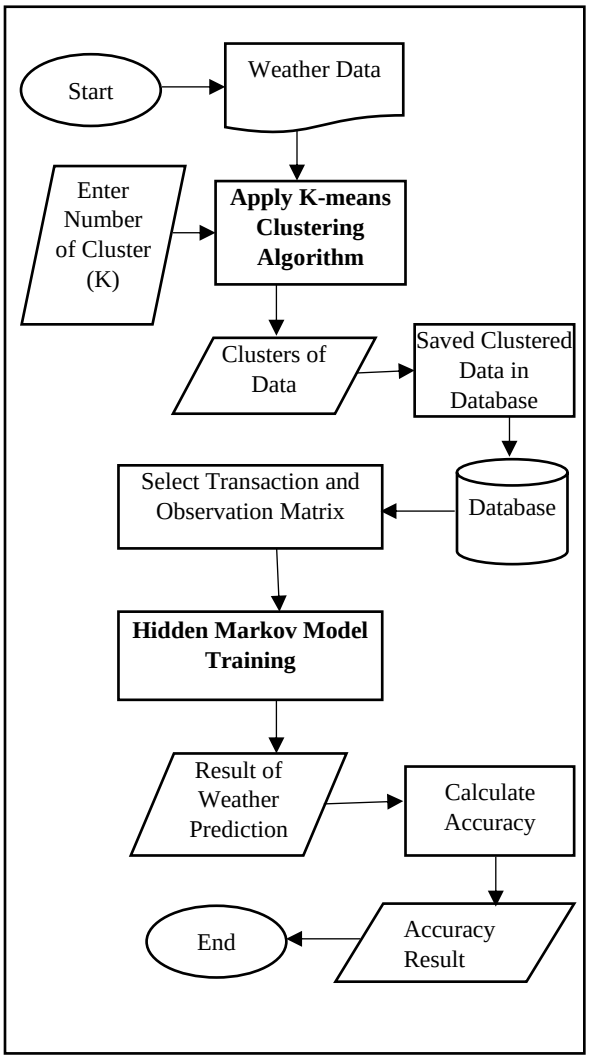


Figure-1. Overview of the proposed system

Consider the example of weather on any given day. Assume that each 'day' corresponds to a state. For simplicity, we assume that only three types of weather are possible namely, sunny, rainy and foggy. Take 1st order Markov approximation that the weather on day in decided entirely by weather on day n-1. The transition probabilities are shown in Table1. Now consider you are locked in a room for several day days and the only clue you get for the weather outside is whether the person carrying your daily meal carries an umbrella or not. The observation is shown in Table 2.

Table1. Example transition table

Current weather	Next weather		
	Foggy	Mostly Cloudy	Partly Cloudy
Foggy	0.6	0.1	0.3
Mostly Cloudy	0.1	0.7	0.2
Partly Cloudy	0.2	0.3	0.5

Table2. Example Observation table

weather	Probability of umberella	Probability of not umberella
Foggy	0.1	0.9
Mostly Cloudy	0.8	0.2
Partly Cloudy	0.3	0.7

6. Experimental Setup

In this experiment, three different data sets, WeahterHistory, Boston, weather_clean, New_York_Hourly_climate, are used as case study. Table 3 summarizes the main characteristics of these datasets. For each dataset, the number of instances, attributes and classes are shown in Table3. The source of three different datasets:

- https://www.kaggle.com/muthuj7/weather-dataset

- https://www.kaggle.com/gviso97/newyork-hourly-climate
- https://www.kaggle.com/jqpeng/boston-weather-data-jan-2013-apr-2018

The experimental environment is as follows:

- Operating System: 64-bit, CPU: Core i3@2.3 GHz, Memory: 4 GB

Table3. Data set description

Dataset Name	No. of Instances	No. of attributes	No. of classes
WeatherHistory	96453	12	5
New_York_Hourly_climate	4553	13	15
Boston weather_clean(day)	3750	24	4

7. Implementation

The implementation of the system is described in this section. When the system is started, home page will firstly appear. Figure-2 shows the welcome page or home page of the system.

Figure-2 Home Page of the Weather Prediction System

In K-means clustering form, the system clusters weather data inputted by user as shown in Figure-3. In this form, there are 4 buttons namely **"Show"**, **"Cluster Data"**, **"<<Back"**, and **"Next>>"** respectively. When the user clicks **"Cluster Data"** button, the system clusters these data. The clustering groups formed observation of Hidden Markov Model.

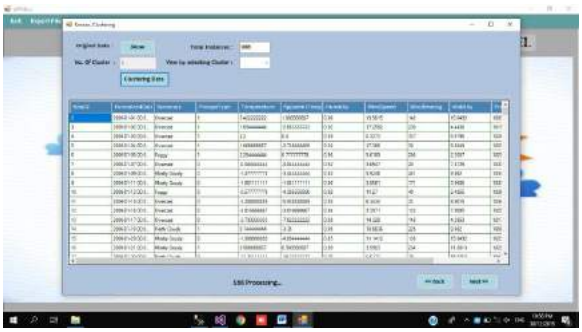


Figure-3 Clustering Data using K-means Clustering

After clicking **"Prediction Result"** menu item, the system shows **"Prediction Result"** form. In these form, user must select three fields, namely, dataset name, prediction types, and number of prediction days to see prediction result. The result depend on user selected fields is shown in grid view. The interface is shown in Figure-4. When user clicks **"<< Back"** button, the system shows **"K-means Clustering"** form is appeared. User can click **"Next >>"** button after clicking **"Apply HMM",** and **"Calculate Accuracy"** buttons. **"Accuracy Result"** form is appeared while user clicks **"Next >>"** button**.**

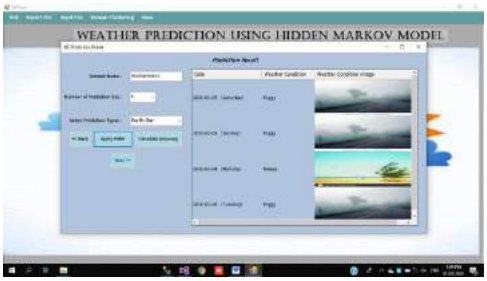


Figure-4 Prediction Result Interface

After clicking **"Accuracy Result"** menu item, **"Accuracy Result"** form is appeared as shown in Figure-5. In these form, user clicks **"Calculate"** button, the results (dataset name, number of cluster, start date, end date, accuracy) are appeared. These form is closed while user clicks **"Close"** button.

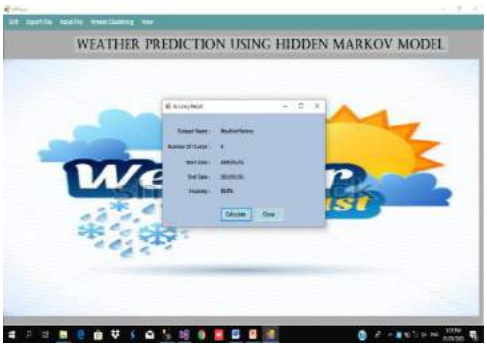


Figure-5 Accuracy Result Interface

8. Performance Evaluation

In this paper, prediction accuracy is measured. This experiment was carried out using holdout method. Two thirds of the data are allocated to the training set, and remaining one third is allocated as the test set. Prediction accuracy is an actual accumulated record of statistical difference of the predicting weather condition and the actual weather condition. As the equation,

$$\text{Accuracy} = (x/y)*100 \quad (10)$$

where,

Accuracy =accuracy rate of predicting

x= total correctly identified patterns

y=total input samples

Table 4. Accuracy comparison of Proposed Technique and Traditional Technique

	Dataset Size	Accuracy	
		Proposed Technique	Traditional Technique
WeatherHistory	1000	60%	60%
	1700	85%	79%
New_Youk_Hourly_climate	172	70%	68%
	174	75%	70%
Boston weather_clean	950	70%	65%
	1130	78%	74%

9. Conclusion

This paper focus on K-means Clustering and Hidden Markov Model (HMM). K-means Clustering algorithm is used for clustering weather datasets. The output of k-means became observation matrix of Hidden Markov Model. Transition matrix is got from weather dataset. These two matrixes are trained in the model. The model predicts next weather condition. The proposed system calculated accuracy of the prediction result. The comparison is also made with reference paper by using three different datasets. According to the obtained results the traditional algorithm namely ID3 decision tree algorithm consumes less accuracy as compared to the proposed algorithm. Accuracy of WeatherHistory, Boston Weather_climate, and New_York_Hourly_climate are 85%, 78%, and 75% respectively.

References

[1] Andrew W. Robertson, Sergey Kirshner, Padhraic Smyth, “Hidden Markov models for modeling daily rainfall occurrence over Brazil”, Technical Report UCI-ICS 03-27 Information and Computer Science, University of California, Irvine, 2003.

[2] Rohit Kumar Yadav, Ravi Khatri, “A Weather Forecasting Model using the Data Mining Technique”, International Journal of Computer Applications, April 2016.

[3] Alwis Nazir, Lia Anggraini, Lola Octavia, Fadhilah Syafria, “Hospital Patients Arrival Prediction Using Markov Chain Model Method”, 4th International Conference on Cyber and IT Service Management, 2016.

[4] Zhitang Chen, Jiayao Wen, Yanhui Geng, “Future Traffic using Hidden Markov Models” IEEE 24th International Conference on Network Protocols (ICNP), 2016.

[5]Antonello Panuccio,Manuele Bicego,Vittorio Murino, “A Hidden Markov Model-based approach to sequential data clustering”, Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Springer, 2002.

[6] Ghazaleh Khodabandelou, Charlotte Hug, Rebecca Deneckere, Camille Salinesi, “Supervised vs. Unsupervised Learning for Intentional Process Model Discovery”, Business Process Modeling, Development, and Support (BPMDS), Jun 2014.

[7] Folorunsho Olaiya, “Application of Data Mining Techniques in Weather Prediction and Climate Change Studies”, International Journal of Information Engineering and Electronic, Business, 2012.

[8] Zoubin Ghahramani, “An introduction to Hidden Markov Models and Bayesian Networks”,

International Journal of Pattern Recognition and Artificial Intelligence, 2001.

[9]Sadegh Khanpour and Omid sojoodi: **"**Improvement of a method based on Hidden Markov Model for Clustering Web Users**"**, F aculty of Electrical, Computer and IT Engineering, Qazvin Azad University, Qazvin, Iran.

[10]Sakthi Kumaran Shanmuganathan: **"**A HMM prediction model for spatio temporal trajectories**"**, The University of Texas at Arlington, 2014.

Information Retrieval System using BM25, Pivoted Normalization and CombSUM Method

Nu Yin Khaing, Ah Nge Htwe
University of Computer Studies, Yangon
Nuyinkhaing936@gmail.com,anhtwe@gmail.com

Abstract

Retrieving information is difficult and time consuming for searching a variety and large number of documents on the digital library. This paper intends to implement effective keyword search system for digital library.BM25 and Pivoted Normalization are best retrieval models for information retrieval system. The CombSUM is combining these two methods to get more relevant documents and to give better output result. The proposed system will help the user to get all relevant documents according to the given query. When the user enters the query, the most relevant documents are ranked by using BM25, Pivoted Normalization Method and CombSUM.

Keywords: ***BM25, CombSUM, Pivoted Normalization Method***

1. Introduction

Information retrieval (IR) is the process of retrieving information or documents that contain information which is relevant to the given query from data collections. An information retrieval process begins when a user enters a query into the system. Information retrieval is usually associated with document retrieval. It can search from large amount of data and return the relevant information to user's information needs. Information retrieval system retrieves the relevant information with the help of the retrieval models [6].

IR system computes a numeric score on how well each document in the database match the query, and rank the documents according to this value. In query processing, the user enters a query and system finds the relevant document, rank the documents and displays to user. The top ranking documents are matched to the query. IR is important of the success of digital library. BM25 is suitable for short document with depend term frequency. Pivoted Normalization is suitable for long document with the independent term frequency. CombSUM is suitable for short document and long document. Paper will be organized as section 2 presents related work, section 3 presents background theory, section 4 presents implementation and design of the proposed system, section 5 presents system evaluation and section 6 presents conclusion.

2. Related Work

In information retrieval system [3], despite the widespread used of BM25, there have been studies examining its effectiveness on a document description over single and multiple field combinations. This system determines the effectiveness of BM25 on various document fields. This system find that BM25 models relevance on popularity fields such as anchor text and query click information no better than a linear function of the field attributes. This system also finds query click information to be the single most important field for retrieval. In response ,they develop a machine learning approach to BM25-style retrieval that learns, using Lambda Rank, from the input attributes of BM25. The proposed model significantly improves retrieval effectiveness over BM25 and BM25F. Their data-driven approach is fast, effective, avoids the problem of parameter tuning, and can directly optimize for several common information retrieval measures. This system demonstrated the advantages of their model on a very large real-world Web data collection.

A.Singhal, C.Buckley proposed method[2], automatic information retrieval systems have to deal with documents of varying lengths in a text collection. Document length normalization is used to fairly retrieve documents of all lengths.Document length normalization is a way of penalizing the term weights for a document in accordance with its length.Various normalization techniques are used in information retrieval system.. This system presents the pivoted normalization, a technique that can be used to modify any normalization function thereby reducing the gap between the relevance and the retrieval probabilities.This paper shows that better retrieval effectiveness results when normalization strategy retrieves document with the match to their probabilities of relevance.Their system present a normalization approach for pivoted normalization.

Training of pivoted normalization on one collection can successfully use it on other (new) text collections, yielding a robust, and collection independent normalization technique.

3. Background Theory

The goal of information retrieval (IR) is to provide users with those documents that will satisfy their information need. An IR model manages how a document and a query are represented and relevance of a document to a user query is defined.

3.1. Indexing

Indexing is an important process in Information Retrieval System. It reduces the documents to the informative terms contained in them. It is collects, parses, and stores data to facilitate fast and accurate information retrieval. The process of storing the term and term list in the computer for effective retrieval. In document organization or indexing process, the documents are preprocessed and stored in database suitable for the efficient query processing by using a data structure such as inverted index. An index is used to quickly find terms in a document collection. As a digital library grows, an efficient method to do a full-text search is required. An inverted index is typically to achieve this objective.

3.1.1. Inverted Index

Documents are normally stored as lists of words, But inverted index invert this by storing for each word the list of documents that the word appears in, hence the name ” inverted index”.[1]

Storing the total frequency for each word can be useful in optimizing query execute. It is a structure used by search engines and databases to make search terms to files or documents. In information retrieval, an inverted index is an index data structure storing a mapping from document. Inverted index may contain additional information like how many times the term appears in the document,ID.Inverted index table, document divided three types of fields,title, category and abstract.There are five fundamental components of an inverted index.Each term is mapped to a list of id,field,term,dociD and frequency. Inverted index stores title, category and abstract of document The purpose of an inverted index is to allow fast full text searches, at a cost of increased processing when a document is added to the database as shown in table 1.

Table1: Example of Inverted Index

ID	Field	Term	Document	frequency
1	Title	cloud	Doc-001	1
2	Title	comput	Doc-001	1
3	Title	Study	Doc-001	1
4	Title	store	Doc-001	1
5	Title	dat	Doc-001	1
6	Title	issu	Doc-001	1
7	Category	Cloud	Doc-001	1
8	Category	comput	Doc-001	1
9	Abstract	cloud	Doc-001	1
10	Abstract	comput	Doc-001	1
11	Abstract	access	Doc-001	1
12	Abstract	correct	Doc-001	1

3.2. Okapi (BM25)

. BM25 is one of the widely used information retrieval functions because of its consistency high retrieval accuracy. It is a function of term frequencies, document frequencies, and the field length for a single field. It is a word ranking algorithm which behaves in a very similar way to TFIDF, since it discriminates terms by their numeric score of relevance [6].

BM25 is a ranking function used by search engines to rank matching documents according to their relevance to a given search query. It is one of the best-known term weighting and document ranking functions. It is a probabilistic model of information retrieval. The widely used BM25 ranking formula we use today is structured by combining the BM11 and BM15 ranking formula. [4]For the calculation of Okapi(BM25), we use term frequency and inverse document frequency.The BM25 has been one of the most efficient and widely-used information retrieval weighting models in the past three decades. BM25 is used in this system to find relevant documents to the user query.

Require: query Q and Document Collection

BEGIN

1. score[N]=0.0,k1=1.2,b=0.75

2. for all term t in query q do
3. for all document d_i in collection do
4. avdl=total length of documents/number of documents in document collection
   IDF=log(number of documents in document collection/number of documents with term)
   Score(i)+=(IDF*(k1+1)*t_f(t,doc))/(k1(1-b+b*len/avdl)+ t_f(t,doc))
5. End for
6. End for
7. Sort(Score[ ])
8. Return(Score[ ])

END

Figure 1:Pseudo-code for Okapi (BM25)

Pseudo code for Okapi(BM25) involve the words t_f (t, doc) is frequency of term t appearing in document "doc", Score (i) is score of document i, avdl is average length of documents , k1,b is constant and IDF is inverse document frequency.

K1 controls trem frequency documents.Constant k1 defines 0 and 3.The default is 1.2.Constant b controls document length influence on the scoring.Constant b defines 0 and 1.The default is 0.75.

In the BM25 algorithm, there are several steps to perform consecutively. In step1, this scheme define the constant value such as score of [N]=0, K1=1.2, and b=0.75. In step2, work the query q for each term of "t". In step3, performs each document from document collection. For each document, the step 4, the scheme calculates the values of IDF and avdl. And then calculates the values of score. Finally, the algorithm arranged the scores according to the ascending order and returned the values.

3.3. Pivoted Normalization Method

Document length normalization is used to help correctly retrieve documents of various lengths. Pivoted Normalization Method is one of the normalization techniques. In information retrieval (IR), term frequency is a fundamental and important component of a ranking model. It is one of the best performing vector space retrieval formulas.. Pivoted normalization is a technique that can be used to modify any normalization function thereby reducing the gap between the relevance and the retrieval probabilities. It uses the number of unique terms in a document as the normalization function. It used to remove the advantage the long documents have in retrieval over the short documents. Long documents usually use the same terms repeatedly. As a result, the term factors may be large for long documents. As one of the most well established IR systems, Okapi's normalization method is similar to the pivoted normalization [2].

Require: query Q and Document Collection
BEGIN

1. score[N]=0.0,s=0.02
2. for all term t in query q do
3. for all document d_i in collection do
4. avdl=total length of documents/number of documents in document collection
   IDF=log(number of documents in document collection+1/number of documents with term)
   Score(i)+=(IDF*1+log(1+log(tf(t,doc))/(1-s+s*len/avdl)
5. End for
6. End for
7. Sort(Score[ ])
8. Return(Score[ ])

END

Figure2: Pseudo code for Pivoted Normalization Method

Pseudo code for Pivoted Normalization involve the words t_f (t, doc) is frequency of term t appearing in document "doc", Score (i) is score of document i, avdl is average length of documents , s is constant and IDF is inverse document frequency.

In the pivoted normalization algorithm, there are several steps to perform consecutively. In step1, this scheme define the constant value such as score of [N]=0, and s=0.02. In step2, work the query q for

each term of "t". In step3, performs each document from document collection. For each document, the step 4, the scheme calculates the values of IDF and avdl. And then calculates the values of score. Finally, the algorithm arranged the scores according to the ascending order and returned the values.

3.4. CombSUM

Data fusion is the combination of the results of independent searches on a document collection into one single output result [7]. Fusion algorithm:

1. scored-based
2. ranked-based

CombSUM is a scored-based approach. CombSUM set the score of each document in the combination to the sum of the scores obtained by the component results. It is obtained better results in Information Retrieval (IR) by taking advantage from the combination of existing methods.

For document (i)

$$\text{CombSUM(i)}=\sum_{k=1}^{N(i)} S_k(i) \qquad \text{eq-----1}$$

$S_k(i)$ = the score of the I document on the result list(ranking)k

N(i) = the number of times a document appears on rankings.

4. Implementation and Design of the Proposed System

There are two main concepts in the proposed work: Document Scoring and Ranking. This system is retrieve relevant documents using BM25, Pivoted normalization method and CombSUM. When a user enters a query, the three steps are performed. Three steps contain tokenization, stop-word elimination and stemming.

To rank matching documents according to their relevance scores to a given search query, it is necessary to assign numerical score to each document based on a ranking function.

4.1. Pre-Processing

Preprocessing is first step and plays important role in classification techniques and applications. It is also crucial in determining the quality of the classification stage. The task is to select the significant keywords that carry the meaning and discard the words that do not contribute to distinguishing between the documents.

Tokenization: Individual word are formed from the clean lyrics. Tokenizing split the words from the training and testing lyrics. Tokenizing is done by the help of String Tokenizer.

Stop-word Removal: Words that carry no particular meaning such as "a", "and", "the" and some other common word should be eliminated. List of stop-words are maintained in the system database. The system removes the stop-words after tokenizing.

Stemming: This system used the Porter stemmer. Porter stemmer is the process for removing the commoner morphological and the inflexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems.

eg, computer-comput

Inverted index is used in this proposed system.

4.2. Proposed system for Admin and User

The proposed system is implemented to retrieve the research papers in the collection when the user submits the query.The main data type is to train document title and abstract. The proposed system has two portions:

1) Admin portion, and
2) User portion

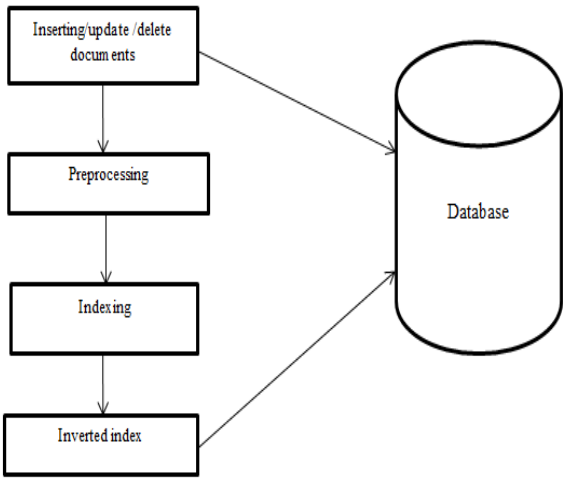


Figure 3: System Flow (For Admin)

In the admin portion, the training data set are processed for further usage. The admin loads the dataset (insert/update/delete) and then makes the preprocessing phase (tokenization, stop-word removing and stemming). After the pre-processing, the dataset contents, the indexing phase and inverted index phased are consecutively processed as shown

in figure 3.Document table stores the doc-id, category, the author the title, the year and the abstract of the research papers. Second table is Doc inverted-indexes table. It has field, term, doc-id and frequency of the term. The pre-calculated data are stored in the system database as shown in figure 4.

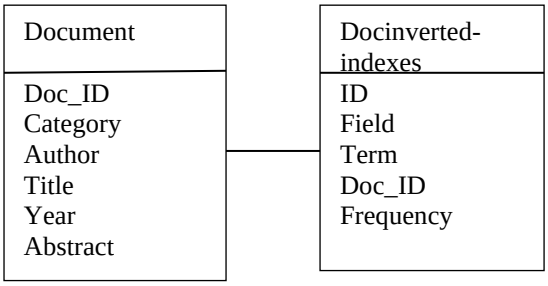


Figure 4: Database Design of the System

The next portion is the user portion and the information retrieving processes are performed by using BM25, Pivoted Normalization. Then, these resultant values are compromised by using CombSUM method. User can search document not only using proposed ranking methods but also using by author, title, year, category and abstract.

When user gives input words, the proposed system search the relevant document by using BM25 and Pivoted Normalization Method. When user inserting query, system performs preprocessing (tokenization, stopword remove, stemming).Input words is keyword that author,title,year,category and abstract of the paper. CombSUM is combine rank result of BM25 and Pivoted Normalization Method. CombSUM retrieves the most relevant document to the user. The system finally returns the rank result of CombSUM. The detail processing steps are as shown in the following figure 5.

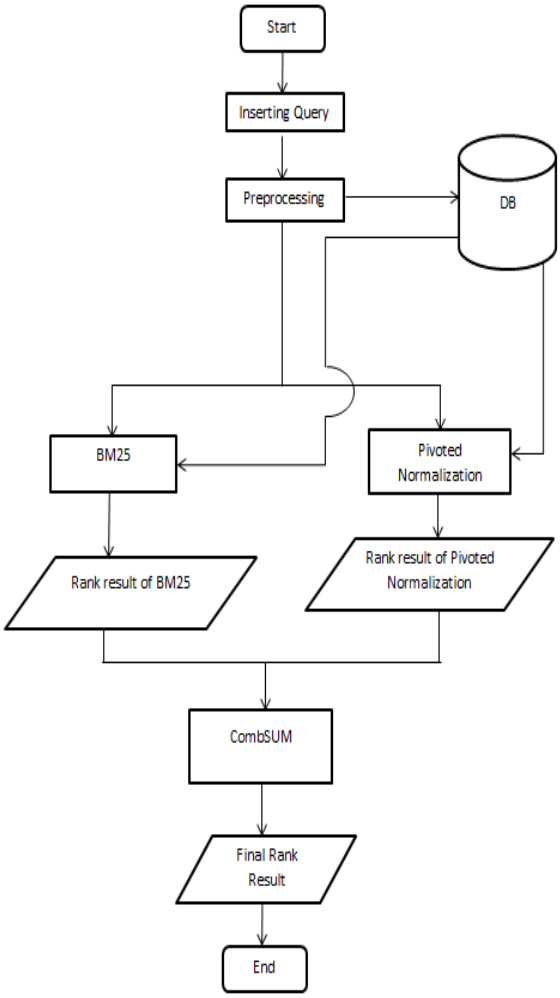


Figure 5: System Flow (For User)

5. System Evaluation

The table 2 shows the rank results of the three methods. In this testing, the 296 data records are used as training data. Input word is “Image Processing” which is used as testing data based on the top score of each result:

In BM25, the result from Doc-206 got 5.088 score which is the highest score among the results of others.For the similarity of Doc-206 which has 77 words in total, 10 matched words retrieved by BM25 method provide 13% similarity.

In pivoted normalization, the result from Doc-219 got 6.045 score which is the highest score among the results of others.For the similarity of Doc-219 which has 45 words in total, 6 matched words retrieved by pivoted normalization method provide 13% similarity.

In CombSUM, the result from Doc-233 got 10.907 score which is the highest score among the results of others.For the similarity of Doc-233 which has 63 words in total, 10 matched words retrieved by BM25 method provide 14% similarity.

So, using the combSUM on the BM25 and Pivoted Normalization can be generated the more similar and relevant result with respect to the testing data.

Table2. Rank results of BM25, Pivoted Normalization Method and CombSUM.

No.	BM25	Score	Pivoted Normalization	Score	CombSUM	Score
1	Doc-206	5.088	Doc-219	6.045	Doc-233	10.907
2	Doc-233	4.953	Doc-233	5.953	Doc-206	10.844
3	Doc-216	4.848	Doc-206	5.756	Doc-219	10.787
4	Doc-207	4.834	Doc-234	5.587	Doc-207	10.365
5	Doc-222	4.805	Doc-207	5.531	Doc-234	10.274
6	Doc-215	4.790	Doc-211	5.460	Doc-216	10.074
7	Doc-224	4.790	Doc-224	5.273	Doc-211	10.071
8	Doc-219	4.742	Doc-216	5.223	Doc-224	10.063
9	Doc-234	4.687	Doc-227	5.146	Doc-222	9.899
10	Doc-211	4.614	Doc-222	5.094	Doc-221	9.564

According to the result of the above these rankings, this system evaluates the result of each schemes based on the similarity percentages and the visual checking.

There are two types of evaluation techniques which are objective and subjective measures. Subjective measure means that measuring the match word by the visual checking of the user. This system uses subjective measure. In the result of CombSUM, The “Doc-233” provides that 9 match words from the 63 total words . Second document “Doc-206” provides that 10 match words from the 77 total words it. The third document ”Doc- 219” provides that 6 match words from the 45 total words it. This system decided the best retrieval result based on the number of result words for the user query and total words involved in document. According to the higher score of the three methods and the number of relevance words in the above query results, this system retrieved nearly all relevant documents.

6. Conclusion

Information retrieval is the study of helping user to find information that matches their information needs. In the proposed system, IR is applied for digital library to provide efficient searching method to users. In this system, user can get the most relevant results. The system can be practicably useful in the digital library.. BM25 and Pivoted Normalization Method is the best retrieval model for information retrieval system. But each method also has the drawbacks. So, using the combSUM method on BM25 and Pivoted Normalization can get the most relevant and optimal result over BM25 and Pivoted Normalization methods.

REFERENCES

[1]. A. karim,D. Enteesha ,”Enhance Invertrd Index Using in Information Retrieval”, Eng & Tech Journal,Vol 34,2016

[2]. A. Singhal, C. Buckley, M. Mitra , “Pivoted Document Length Normalization” , SIGIR 1996,p. 21-29

[3]. K. Svore, “A Machine Learning Approach for Improved BM25 Retrieval”, 18th ACM Conference on Information and Knowledge Management, 2009

[4]. M. Beaulieu, M. Gatford ,X. Huang, S.E Robertson, S. Walker, P. Wallians ,”Okapi at TREC-5”, The Fifth Text Retrieval Conference,p. 143-165,1997

[5]. N. Fuhr ,”Probabilistic Models in Information Retrieval”, The Computer Journal,vol-35,no-3,1992

[6]. R. Baeza-Yates and B. RIberia-Neto,”Modern Information Retrieval “, ACM press,ISBN -0-201-39829,2009

[7]. R. Nuray and F. Can,”Automatic ranking of information retrieval system using data fusion”,Information Processing and Management :an International Journal,v.42 n.3,p.595-614,may 2006

Duplicate Record Detection in Data Cleaning Using DCS++ Algorithm

Yin Yin Phyo, Thidar Win
University of Computer Studies, Yangon
yinyinphyo@ucsy.edu.mm, thidarwin@ucsy.edu.mm

Abstract

Duplicate Record Detection is a multiple record search process that represents the same physical entity in a dataset. It is also known as the record linkage (or) entity matching [1]. The databases contain very large datasets. Datasets contain duplicate records that do not share a common key or contain errors such as incomplete information, transcription errors and missing or differing standard formats (non-standardized abbreviations) in the detailed schemas of records from multiple databases. So, the duplicate detection needs to complete its process in a very shorter time. Duplicate detection requires an algorithm for determining whether records are duplicate records or not.

In this paper, calculate a similarity metric that is commonly used to find similar field items and use the Duplicate Count Strategy Multi-Record Increase (DCS++) Algorithm for approximately duplicate records detection over publication xml dataset.

1. Introduction

Nowadays, the amount of data within the datasets becomes more and more huge. Data errors or inconsistent data in these datasets also grow rapidly as the technology advances. In the economic world, invalid and duplicate data can be costly because it can affect the key decisions of many industry operations and the production of business organizations. Therefore, data needs to be good quality.

In order to improve the data quality, data cleaning is especially necessary when integrating disparate data sources [2]. When integrating data from different sources and implementing a data warehouse, organizations become aware of possible differences and systematic conflicts.

The problem of identifying duplicate records in the database is an important step in the data cleanup and integration process. Data reduction is the process of detecting and removing data errors, inconsistencies, and duplicate data. Duplicate detection is one of the solutions of data cleaning. It has two tasks to detect duplicate records efficiently and effectively:

(i) The representation of the data may vary slightly, so a specific similarity measure needs to be defined to compare pairs of records.

(ii) Not all records can be peer compared because the data set may be large.

To perform task (ii), a number of algorithms have been proposed that split the dataset and compare all pairs of records in each partition.

Sorted Neighborhood Method (SNM) is a known way to advance the window by classifying data based on the sorting key and comparing only the records displayed in the same window.

This paper proposes the Duplicate Count Strategy-Multi Record Increase Approach (DCS++), a variation of SNM and improvement of Duplicate Count Strategy (DCS). If a duplicate is found on the sorted dataset, it can also detect the other possible duplicates by comparing the next w - 1 record of that duplicate.

It can also reduce the comparing time by skipping windows for duplicates. Therefore, the proposed system can be faster and detects more duplicate records.

2. Related Work

Many researchers research on duplicate record detection with different efficient and effective blocking and windowing methods [3].

Ying Pei et al. [4] implemented the K-medoids clustering algorithm (IKMC) to solve the problem of detecting almost duplicate records. It is considered as one separated data object for every record in the database. It uses the Edit Distance method to get similarity values between records. Finally, clustering of these similarity values can detect duplicate records. The algorithm can automatically adjust the number of clusters by comparing the similarity value with a predefined similarity threshold. This algorithm shows good detection accuracy and high availability. Qiaoqiao Yang et al. [5] implemented the SNM algorithm based on some edit distances and variable windows to solve the shortcomings of the SNM algorithm. The algorithm proposed in this paper is based on the various edit distances and variable windows. The experiment's data set comes from the refrigeration industry management system. This proposed algorithm can efficiently recognize duplicate big data records. However, there is still the problem of improving the recall ratio and handling non-standard samples.

Jumoke Soyemi et al. [6] implemented a system for detecting duplicate records in a database using a simil matching algorithm. The Simil algorithm is based on calculating the similarity between two

strings. This proposed system can only be used to clean up data and prevent incorrect data from accessing the database.

3. Data Preparation

Data preparation is a necessary step in data cleanup before duplicate record detection process. The data preparation phase involves data parsing, data transformation, and standardization procedures. Data preparation techniques are also described in terms of ETL (extraction, transformation, loading) [7].

In the proposed system, the removing of XML tags in publication records is included in the data parsing procedure. Second, prior to the process of detecting duplicates, other data types are uniformly represented in standardization procedure. In this system, author name, date and title are standardized. Author names can be all authors participated in the publication. But only first author is extracted and formatted into first character of first name, dot (.) and last name only. The system extracts only year from date value and title must not be empty. Therefore, these preprocessed data fields can be easily used in key creation and detection.

4. Field Matching Techniques

Field Matching Technique is the inner stage of duplicate detection while the outer stage of duplicate detection is applied as the record matching technique. Duplicate detection relies on string comparison techniques to resolve typographic changes in the string data and errors in the numeric data.

Techniques for matching fields in the context of duplicate record detection include:

- Character-based similarity measurement
- Token-based similarity measurement
- Similarity measurement of pronunciation
- Numerical similarity measurement

4.1. Character-based Similarity Measurement

Character-based similarity metrics handle typographical errors well. Some similarity measures are:

- Edit distance
- Affine gap distance
- Smith-Waterman distance
- Jaro distance metric and
- Q-gram distance

In this proposed system, Edit Distance (or) Levenshtein Distance Algorithm is used to calculate field matching similarity scores.

4.1.1. Edit Distance (or) Levenshtein Distance Algorithm

Edit Distance, a.k.a. Levenshtein distance [8], is the minimum number of edit operations on a single character that is required to transform the string one into string two. Three types of edit operations are possible. They are:

(i) Insertion: insert a character into the string.
(ii) Deletion: delete a character from the string.
(iii) Substitution: replace a character with another character.

Levenshtein Distance Algorithm is described as below:

Algorithm 1: Levenshtein Distance Algorithm

```
levenshtein (source, target : STRING ) : INTEGER
-- Minimum number of edit operations to turn source into target
local
   distance                : ARRAY_2 [ INTEGER]
   i, j, del, ins, subs    : INTEGER
do
   create distance.make (source.count, target.count)
   from  i := 0  until  i > source.count  loop
      distance [ i, 0 ]  :=  i ;
      i  :=  i + 1;
end
   from  j  := 0  until   j > target.count  loop
      distance [ 0, j ]  :=  j ;
      j  :=  j + 1;
end
   from  i := 1 until  i  >  source.count  loop
     from   j  :=  1 until  j  >  target.count  invariant
     -- for  all  p :0 … i, q : 0 …j -1  ,  we can turn source  [ 1..p]
      -- into target [ 1..q]  in  distance [p,q]  operations
      loop
        if  source[i]  =  target [j]  then
           distance [i, j]  :=  distance [ i  - 1, j  - 1]
        else
          deletion :=  distance [ i - 1 , j ]
          insertion :=  distance [ i , j - 1 ]
          substitution :=  distance [ i - 1 , j - 1 ]
          distance[ i , j ]  := minimum (deletion, insertion, substitution) + 1
        end
        j  := j + 1
   end
   i  :=  i  +  1
end
Result  :=  distance (source.count , target.count)
end
```

5. Duplicate Records Detection

There are three basic techniques for detecting duplicate records. They are:

(i) Knowledge-based technology
(ii) Probabilistic technology
(iii) Empirical methods

The proposed system focuses on empirical algorithms [9] which consist of sorting, blocking and windowing methods such as Sorted Neighborhood Method, its extensions and Priority Queue Algorithm, etc. Sorted Neighborhood Method is the most representative for windowing.

5.1. Sorted Neighborhood Method (SNM)

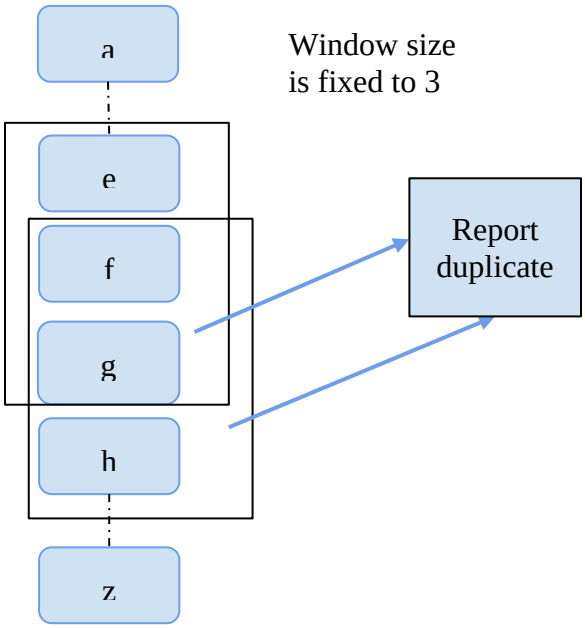


Figure 1: Sorted Neighborhood Method (SNM)

Using a sorted neighborhood approach can reduce the cost of comparing records and increase the efficiency of comparing records.

Hernáandez and Stolfo [10] describe the sorted neighborhood approach. This method involves three steps:

Step 1: Create sorting key: A key for each record in the dataset is assigned to each record. Keys are created by concatenating the values of two or more attributes.

Step 2: Sort the data: The records in the database are sorted based on the sorting key.

Step 3: Merge: A fixed size window is moved through the list of records sequentially to limit the comparison of records matching to those records in the window. Each new record that enters this window is compared to the previous record to find a "matching" record.

The Sorted Neighborhood Method uses fixed size windows. If the selection of window size is too small, some actual duplicate records may be lost and using larger window size will often result in unnecessary comparisons within the window.

The effectiveness of the sorted neighborhood approach depends greatly on the creation of sorting keys.

5.2. Duplicate Count Strategy

Duplicate Count Strategy is more than using fixed-size windows and provides windows that can be adjusted dynamically. The Duplicate Count Strategy is an improvement of the Sorted Neighborhood Method (SNM), and varies the size of the window depending on the number of identified duplicates.

5.2.1. Basic Duplicate Count Strategy (DCS)

Duplicate Count Strategy involves the following steps:

1. Sort the records by assigning a sorting key to each record.
2. Create a window with the initial window size w.
3. Compare the first record with all other records in the window.
4. Increase window size while $\frac{detected\ duplicates}{comparisons} \geq$ Ø
5. Slide the window (initial window size w)
6. Calculate transitive closure.

5.2.2. Enhancement of DCS: Duplicate Count Strategy-Multi Record Increase Approach (DCS++)

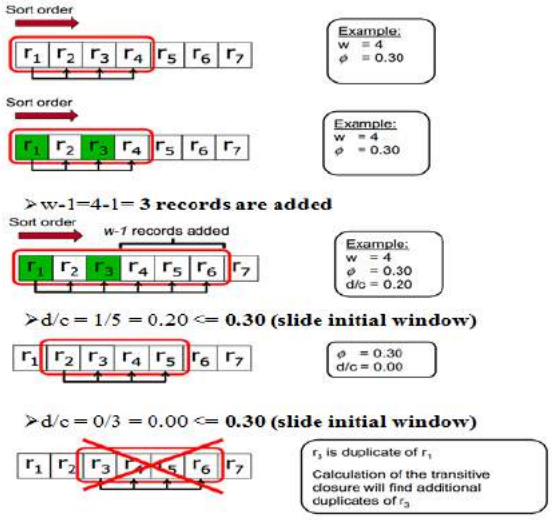


Figure 2: Duplicate Count Strategy ++(DCS++)

Idea 1: If each duplicate is found, the next w -1 records of this duplicate are added to the window.

Idea 2: Windows for duplicates have been omitted to save the comparisons.

- In **Figure 2**: Skip window for r3.
- Use the transitive closure to find additional duplicates.

- There is no loss in the window because the window for r1 covers all comparisons that r3 would have made.

6. Overview of the Proposed System

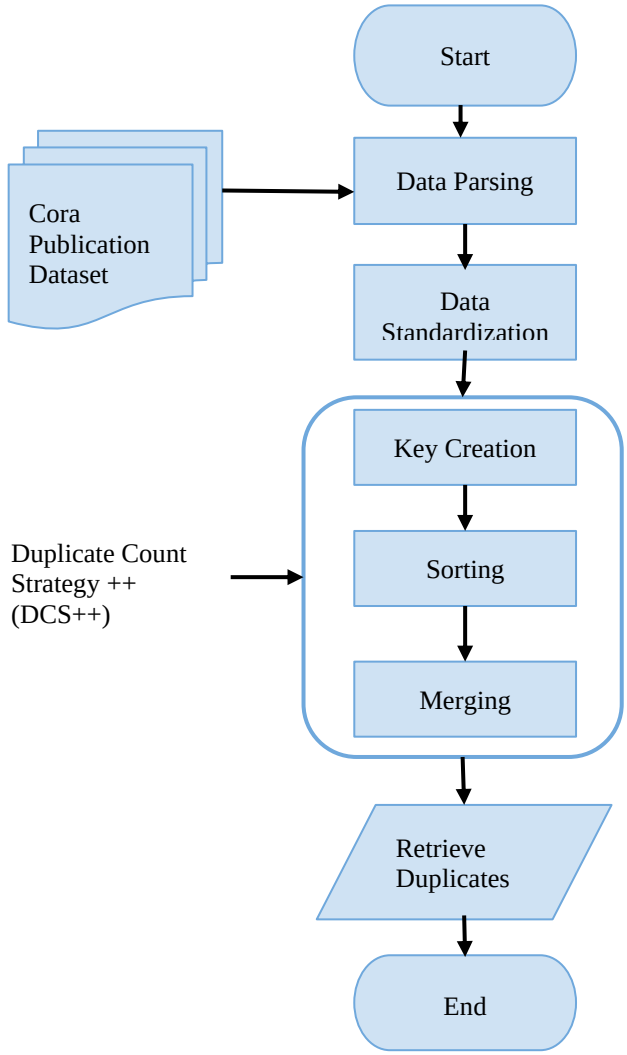


Figure 3: Overview of the proposed system

Figure 3 shows the overview of the proposed system. According to the figure, data preprocessing as data parsing and data standardization is performed to the input dataset. The proposed system can be classified by the two stages. In the first stage, field matching of each record is performed by using Levenshtein Distance Algorithm or Edit distance Algorithm. In the second stage, duplicate record detection process is performed by using Duplicate Count Strategy-Multi Record Increase (DCS++) Algorithm and then duplicate records are detected.

6.1. Duplicate Count Strategy-Multi Record Increase Algorithm

In this experiment, the initial window size (w) is provided as 20 and DCS++ threshold (Ø) is recommended as $\frac{1}{w-1}$ not to miss any duplicates. The Duplicate Count Strategy-Multi Record Increase (DCS++) algorithm is described as below:

Algorithm 2. DCS++

```
1.  w: initial window size
2.  Ø: DCS++ threshold
3.  Require: w > 1 and 0 < Ø ≤ 1
4.  sort records by sorting key
5.  populate window win with first w records of
    records
6.  skipRecords ← null
7.  for j = 1 to records.length - 1 do
8.    if win[1] NOT IN skipRecords then
9.  numDuplicates ← 0
10. numComparisons ← 0
11. k ← 2
12.     while k ≤ win.length do
13.       if isDuplicate (win[1] , win[k] ) then
14.         emit duplicate pair (win[1] , win[k])
15.      skipRecords.add ( win[k] )
16.      numDuplicates ← numDuplicates + 1
17.      while win.length < k+w-1 and j + win.length
    < records.length do
18.        win.add (records [ j + win.length + 1] )
19.      end while
20.       end if
21.       numComparisons ← numComparisons+1
22.       if k = win.length and j + k < records.length
    and (numDuplicates / numComparisons ) ≥ Ø
    then
23.         win.add (records [j + k - 1])
24.       end if
25.       k ← k + 1
26.     end while
27.   end if
28.   win.remove(1)
29.   if win.length < w and j + k < records.length
    then
30.     win.add (records [j + k - 1])
31.   else
32.     while win.length > w do
33.       win.remove (win.length)
34.     end while
35.   end if
36.   j ← j + 1
37. end for
38. calculate transitive closure.
```

7. Performance Evaluation

In this system, the performance of the algorithm is measured using: **Recall**, **False Positive Error** (FP), **False Negative Error** (FN) and **Precision**.

1. **Recall**

   The percentage of duplicate records that the system correctly identifies. Recall percentage is computed by following equation:

   Recall = $\frac{no\ of\ identified\ duplicates}{no\ of\ actual\ duplicates} \times 100\%$ **(7.1)**

2. **False Positive Error (FP)**

   The percentage of records incorrectly identified as duplicates. FP percentage is defined as the equation:

   FP = $\frac{no\ of\ wrongly\ identified\ duplicates}{total\ number\ of\ identified\ duplicates} \times 100\%$ **(7.2)**

3. **False Negative Error (FN)**

   The percentage of duplicate records that the system does not detect. FN percentage is computed by following equation:

   FN = 100% − Recall **(7.3)**

4. **Precision**

   The percentage of information reported as relevant by the system that is correct.

   Precision percentage is defined as the equation:

   Precision = 100% − FP **(7.4)**

7.1. Experimental Environment

In order to evaluate the proposed algorithms, install Xampp on this system. Apache service is started in XAMPP Control Panel as a local server.

The experimental environment is as follows:

- Operating System: Windows 8,
- CPU: Core i5@3 GHz,
- Memory: 8 GB

The software versions are as follows

- Xampp version: 7.3.0 64 bit, and
- PHP version: 7.3.0

In the next sections, the Cora dataset and performance results with different thresholds are discussed.

7.2. Dataset Used

In this experiment, a Cora Dataset is used. This dataset contains bibliographic information for scientific papers. It provides 1,879 objects.

The Cora dataset is prepared by the original Andrew McCallum and his versions of this dataset are provided on his data web page. Many publications in record linkage and entity records over the years used these various versions of the Cora dataset.

```
<CORA>
<NEWREFERENCE id="1">
ahlskog1994a
<author>
M. Ahlskog, J. Paloheimo, H. Stubb, P. Dyreklev, M. Fahlman, O.
</author>
<title>Inganas and M.R.</title>
<journal>Andersson, J Appl. Phys.,</journal>
<volume>76,</volume>
<pages>893,</pages>
<date>(1994).</date>
</NEWREFERENCE>
    ...
</CORA>
```

Figure 4: Sample XML Cora Dataset

7.3. Performance Result

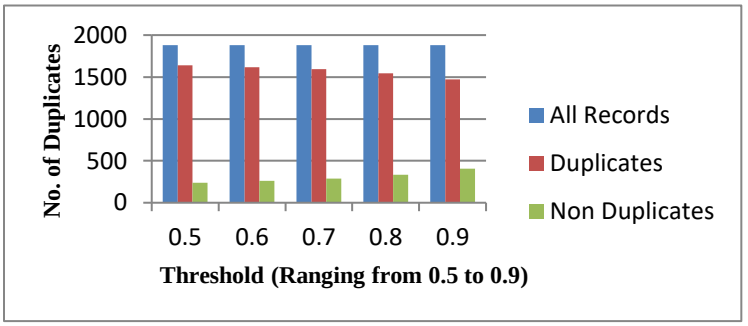


Figure 5: Execution Results of Duplicate Detection with window size 10.

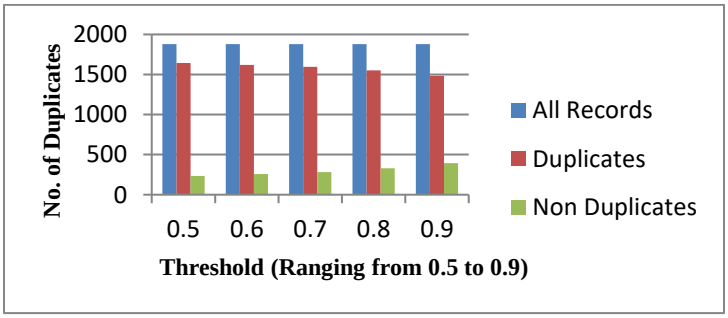


Figure 6: Execution Results of Duplicate Detection with window size 20.

The evaluation is performed by changing the threshold value from 0.5 to 0.9 at intervals of 0.1. This system is tested with different window sizes 10 and 20. **Figure 5** and **Figure 6** show the execution results of DSC++ with Edit Distance Algorithm for five threshold values.

Table 1 shows the performance evaluation results with window size 20 for five thresholds. The percentage of

recall and precision is high. Also, FP and FN is less in threshold values (0.5, 0.6, 0.7).

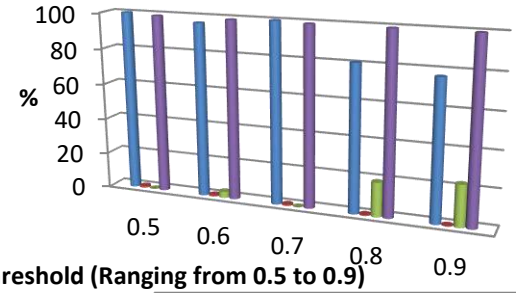
So, it determines that this system identified duplicate records being correctly in threshold values (0.5, 0.6, 0.7). Although precision and FP are good in threshold values (0.8 and 0.9), other percentage values of recall and FN are no good because the percentage of duplicate records being correctly identified by the system is less.

Table 1: Results of Performance Evaluation

DCS++ Algorithm with Edit Distance Algorithm (window size = 20)					
Sr. No	**Threshold**	**Recall%**	**FP%**	**FN%**	**Precision%**
1.	**0.5**	100	0.91	0	99.09
2.	**0.6**	96.55	0.85	3.45	99.15
3.	**0.7**	100	0.74	0	99.26
4.	**0.8**	80.95	0.58	19.05	99.42
5.	**0.9**	76.92	0.35	23.08	99.65

In **Figure 7**, the system assumes that the threshold values (0.5, 0.6 and 0.7) are better than other threshold values (0.8 and 0.9) for duplicate detection because these are less in the percentage of FP, FN and high in the percentage of precision and recall.

Among them, threshold value 0.7 is the best result for duplicate detection in this system.



	0.5	0.6	0.7	0.8	0.9
Recall %	100	96.55	100	80.95	76.92
FP%	0.91	0.85	0.74	0.58	0.35
FN%	0	3.45	0	19.05	23.08
Precision%	99.09	99.15	99.26	99.42	99.65

Figure 7: Performance Evaluation of Duplicate Detection

8. Conclusion and Future Work

The system used DCS ++ to detect duplicate records from the Cora publication xml dataset. This system has exceeded using a fixed window size.

As the result of performance evaluation threshold value 0.7 is the best result for duplicate detection of cora dataset in this system. DCS ++ detects more duplicates by adding the next w-1 records of this duplicate to the window for each detected duplicate. Time is critical in data cleaning of a large database. Usage of duplicate detection DCS++ method is to reduce the time taken on each comparison by skipping windows for duplicates. There is some limitation in this system. This system can only be used in a homogeneous source dataset such as XML format because we need the key creation for some fields of the dependent domain. The future work is to detect duplicates in domain-independent data cleaning.

References:

[1].H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James, *"Automatic linkage of vital records."* Science, vol. 130, pp. 954-99, 1959.

[2].ZM Guo, AY Zhou, *"Research on data quality and data cleaning: A survey[J]"*, Journal of Software, 2002, 13(11): 2076-2082.

[3].Draisbach, Uwe and Felix Naumann. *"A Comparison and Generalization of Blocking and Windowing Algorithms for Duplicate Detection."*, 2009.

[4] .P. Ying, X. Jungang, C. Zhiwang, and S. Jian, *"IKMC: An Improved K-Medoids Clustering Method for Near-Duplicated Records Detection"*, in Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on, Wuhan, 2009, pp. 1 -4.

[5].Q. Yang,* Z. Guo, K. Wang, *"The SNM Algorithm Based on a Variety of Edit Distance and Variable Window"*, The 7th International Conference on Computer Engineering and Networks, 2017, CENet2017.

[6].Jumoke Soyemi, James Adegboye, *"Database Record Duplicate Detection System using Simil Algorithm,"* International Journal on Computer Science and Engineering (IJCSE), 2018, vol 10.

[7].Kimball, Ralph; Joe Caserta, *"The data warehouse ETL toolkit: Practical techniques for extracting, cleaning, conforming, and delivering data"*, John Wiley & Sons, 2004.

[8].*Levenshtein V I. Binary codes capable of correcting deletions, insertions and reversals[J].* Soviet Physics Doklady, 1966, 10(1):707-710.

[9].M. Rehman and V. Esichaikul, "*Duplicate Record Detection For Database Cleansing*," in

Machine Vision, 2009. ICMV '09. Second International

Conference on , Dubai, 2009 , pp. 333 - 338.

[10].Hernández, Mauricio Antonio; Salvatore Joseph Stolfo (jan 1998). *"Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem"*.Data Mining and Knowledge Discovery 2 (1): 9--37.

Inventory Demand Forecasting using Exponential Smoothing Methods

Su Su Lin, Khin Sundee Bo
University of Computer Studies, Yangon
susulinn@ucsy.edu.mm, imksdb@gmail.com

Abstract

In today's competitive environment, forecasting the right amount of sales for the coming period is critical to ensuring product availability and improving customer satisfaction. This study uses a model that identifies the best method for prediction based on the minimum of the prediction error. The required sales data was collected from a pharmacy in Myanmar, Thiri Sandar Pharmacy. A time series chart of the medical sales data shows that demand fluctuates over time. In this system, simple exponential smoothing, double exponential smoothing (Hot's method) and triple exponential smoothing (Winter's method) were performed by statistical technique using inventory demand forecasting software. The performance of all methods was evaluated based on the accuracy of the forecasting, and analysis showed that it was the best method with predicted values and prediction periods for individual drugs. The purpose of this study is to help decision makers make better decisions as they plan their business strategy. This system is an essential need for inventory managers to manage low-stock items in the sales business. This study will guide other researchers in identifying the best prediction method.

Keywords: Inventory Demand forecasting, Time series models, Exponential smoothing methods

1. Introduction

Demand forecasting is the process in which historical sales data is used to develop an estimate of an expected forecast of customer demand. To businesses, demand forecasting provides an estimate of the amount of goods and services that its customers will purchase in the foreseeable future. Demand can be forecasted using qualitative methods or quantitative methods.

Quantitative forecasting techniques are statistical techniques for predicting the future, using numerical measurements and prior effects to predict future events. Two types of quantitative forecasting methods are time series models and associative models. Time series models examine past data patterns and predict the future based on underlying patterns derived from those data. There are many types of time series models, such as simple moving averages, weighted moving averages, seasonal indicators, trend forecasts, simple averages, and exponential smoothing. Exponential smoothing methods use weighted averages of past observations to forecast new values. Exponential smoothing is usually used to make short term forecasts, ass longer term forecasts using this technique can be quite unreliable.

This system is used primarily for medical store forecasting, such as sales data using software at healthcare stores. This system is implemented using a smoothing technique for quantitative forecasting techniques. This system has two main parts. The first is the import of medical sales data by the sales manager, and the second uses exponential smoothing methods to evaluate and determine this application. To implement the smoothing method, the system first needs to enter monthly sales data to calculate the formula for the smoothing method. Furthermore, after calculating all the equations of the smoothing method, the system will show the best method for each item. The system provides inventory reorder quantities by measuring forecast accuracy.

The structure of this paper is as follows. Section 1 introduces the forecasting application and the key exponential smoothing method for business forecasting. Section 2 introduces related

works. Section 3 describes how to use a time series model for forecasting. Section 4 describes the design of the proposed system. Section 5 presents experiments and results. The last section is section 6, with conclusions.

2. Related Work

This section describes previous work on choosing a forecasting method using time series data. Time series data can be of various types, such as power consumption, product sales / demands, and product prices.

Aye Thanda Htun, [1] have proposed "Implementation of sales forecasting system by using Linear Regression Method". It calculates the value of correlation for each item. The system can produce which items should promote or which items should not promoted in future depending on the value of correlation value of each item. This system can analyze the strong positive, negative and weak correlation between item codes by categories. This system cannot be included to identify time series patterns. Therefore, this method does not give accurate prediction results.

Eva Ostertagova, Oskar Ostertag, [2] have presented "Forecasting using simple exponential smoothing method". It used the estimated time series data using the simple exponential smoothing. This method is suitable for data with no trend, which can't give the more relevant demand forecasting for data with trend and seasonality.

3. Time Series Models

Time series models involves working on time (years, months, days, hours, minutes) based data, to derive hidden insights to make informed decision making. Time series models are very useful models when you have serially correlated data. Most companies use time-series data to analyze sales volume, website traffic, and competitive position in the next year.

Table 1. Time Series Analysis

Value	Model	Defination
1	Naive	The forecast is equal to the actual value observed during the last period and good for level patterns.
2	Simple Mean	The average of all available data and good for level patterns
3	Moving Average	The average value over a set time period. Each new forecast drops the oldest data point and adds a new observation. More responsive to a trend but still lags behind actual data.
4	Exponential Smoothing	Most frequently used time series method because of ease of use and minimal amount of data needed.

Table 1 expresses time series forecasting models for time series analysis. Table 2 also describes the definition of time series patterns.

Table 2. Time Series Patterns

Value	Pattern	Defination
1	Level or Horizontal	Data follows a horizontal pattern around the mean.
2	Trend	Data is progressively increasing or decreasing
3	Seasonal	Data exhibits a regularly repeating pattern.
4	Cycle	Data increase or decrease over time.

3.1. Exponential Smoothing

One of the most successful forecasting methods is the exponential smoothing (ES) techniques. It is also easy to adjust for past errors-easy to prepare follow-on forecasts, ideal for situations where many forecasts must be prepared, several different forms are used depending on presence of trend or cyclical variations. In short, an ES is an averaging technique that uses unequal weights; however, the weights applied to past observations decline in an exponential manner.

Table 3. Data vs Methods

Value	Data	Method
1	No trend or seasonal pattern	Single Exponential Smoothing Method
2	Linear trend and no seasonal pattern	Double Exponential Smoothing Method
3	Both trend and seasonal pattern	Triple Exponential Smoothing Method
4	Other pattern	Use Other Methods

Table 3 describes the type of exponential smoothing methods with time series patterns.

3.1.1. Simple Exponential Smoothing Method

This sophisticated method is a kind of weighted averaging method which estimates the future value based on previous forecast plus a percentage of the forecasted error. It is easy to implement and compute as it does not need maintaining the history of previous input data. It fades uniformly the effect of unusual data.

The equation of SES is as follows:

$$F_{t+1} = \alpha\, y_t + (1-\alpha)F_t$$

F_{t+1} = forecast for the next period
α = smoothing constant, which ranges from 0 to 1
y_t = observed value of series in period t (latest sales)
F_t = current forecast for period t

Notice that the smoothed value becomes the forecast for period t + 1.

3.1.2. Double Exponential Smoothing Method

Double exponential smoothing is used to forecast data having linear trend. It is an extension of simple exponential smoothing. Holt's method smooths both trend and slope in the time series using two different smoothing constants (alpha for the level and gamma for the trend).

The equation of DES is as follows:

Current level estimate

$$L_t = \alpha\, y_t + (1-\alpha)(L_{t-1} + b_{t-1})$$

The trend estimate

$$b_t = \beta\,(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$

Forecast for period m

$$F_{t+m} = L_t + mb_t$$

L_t = estimate of the level of the series at time t (current estimate of sales)
L_{t-1} = previous estimate of the level
α = smoothing constant for level ($0 \le \alpha \le 1$)
y_t = new observation or actual value of series in period t (latest sales)
β = smoothing constant for trend ($0 \le \beta \le 1$)
b_t = estimate of the slope of the series at time t
b_{t-1}= previous estimate of trend
m = periods to be forecast into the future

3.1.3. Triple Exponential Smoothing Method

When both trend and seasonality are present in data set, this procedure can be used. It is

used to smooth data employing a level component, a trend component, and a seasonal component at each period and provides short to medium range forecasting. There are two types of model: multiplicative and additive.

The equation of TES is as follows:

The trend estimate

$$L_t = \alpha \frac{y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1})$$

The seasonality estimate

$$b_t = \beta (L_t - L_{t-1}) + (1-\beta) b_{t-1}$$

$$S_t = \gamma \frac{y_t}{L_t} + (1-\gamma) S_{t-s}$$

Forecast for period m

$$F_{t+m} = (L_t + mb_t) S_{t+m-s}$$

L_t = level of series
α = smoothing constant for the data
y_t = new observation or actual value in period t
β = smoothing constant for trend estimate
b_t = trend estimate
γ = smoothing constant for seasonality estimate
S_t = seasonal component estimate
m = number of periods in the forecast lead period
s = length of seasonality
F_{t+m} = forecast for m periods into the future

4. Proposed System

The main process of this proposed system involves two main parts. The first is to enter medical sales data. The second is to use exponential smoothing to calculate sales data. Users can be categorized by position, such as Inventory Manager or Sales Manager.

Figure 1 is for sales managers. Sales managers need to log in to the system. Authorized sales managers can view, save, edit, and delete sales data information.

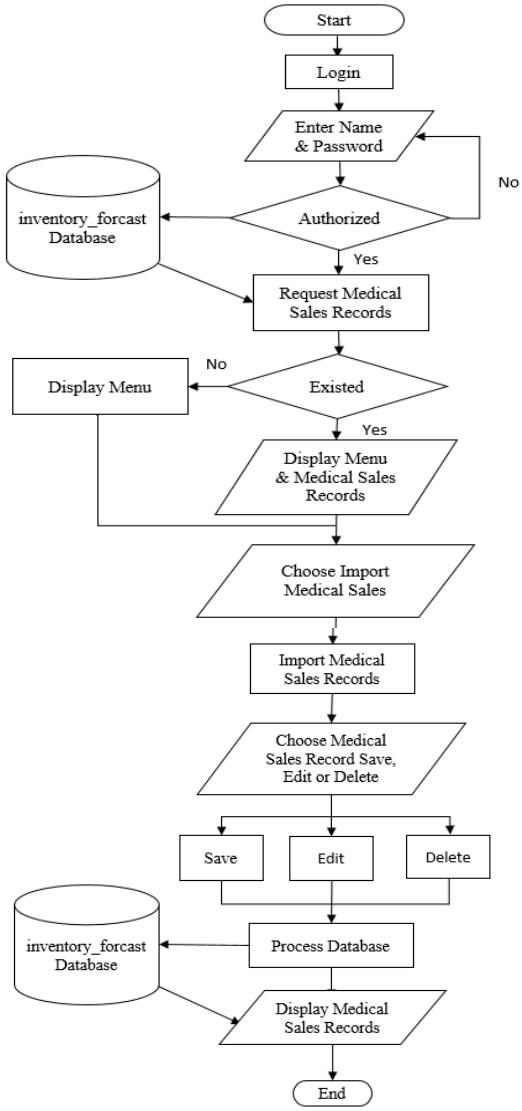


Fig.1 System Flow Diagram (Sale Manager Side)

Figure 2 shows inventory control manager side and exponential smoothing (ES) calculation process. The user who has inventory control manager position and authorized user, has to manage this system users, view charts for desired medical sales items, calculate exponential

smoothing to view sales data analysis and forecast records is needed.

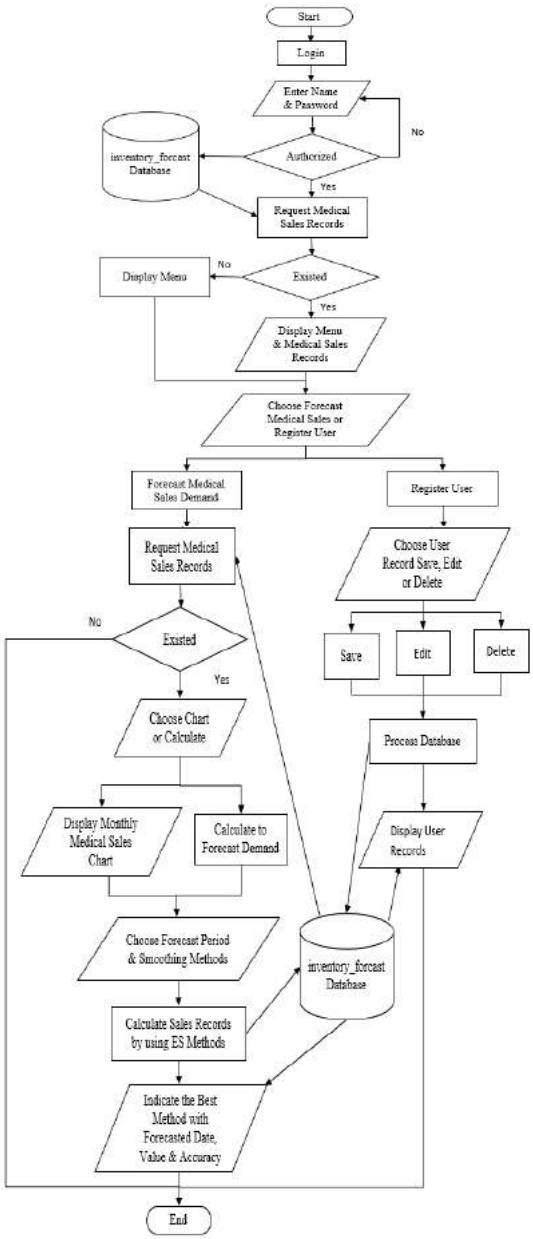


Fig.2 System Flow Diagram (Inventory Control Manager Side)

5. Experiments and Results

Implementing a quantitative forecasting technique involves dealing with each step of the quantitative methodology. The proposed system is implemented using exponential smoothing, a quantitative forecasting technique. The first implementation requires input data to calculate the formula.

Fig.1 Login with Sale Manager Position

If you are a sales manager, you must be able to log in with your existing sales manager account and the new sales manager must notify the inventory manager to register. After obtaining the password, the new sales manager can log in.

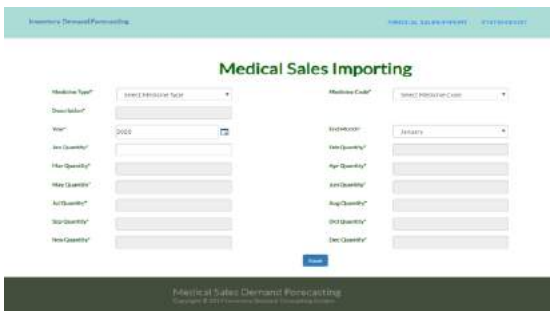


Fig.2 Input New Medical Sales Data

Users can input sales information on a monthly basis. In the inventory forecasting database, sales data is stored by category. The system automatically stores annual and quarterly sales information for each item. If the selected drug code and insertion year are already registered in the database, this drug sales information will be updated.

Fig.3 Show Inputted Medical Sales Data

The sales information will be displayed.

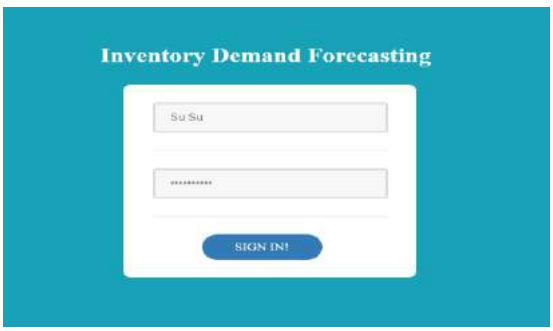


Fig.4 Login with Inventory Control Manager Position

On the inventory control manager side, the user must login to the system.

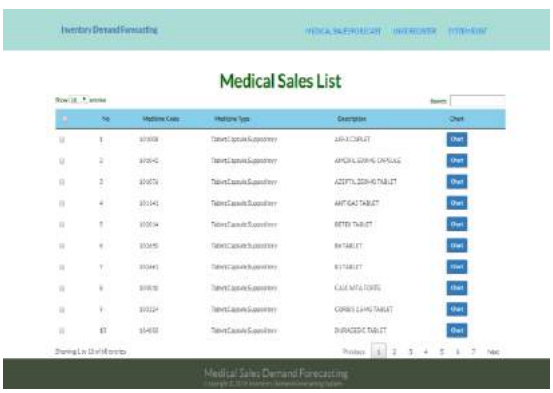


Fig.5 Show Inputted Medical Sales Data

After entering sales data from the sales manager, the entered medical sales data will be displayed to predict for future sales.

Fig.6 Time Series Chart of Monthly Sales Data (101008)

After clicking the chart button, you can see time series chart of monthly sales data for each item to know the time series pattern of the desired item.

Fig.7 Select Necessary to Calculate

To calculate using exponential smoothing methods, the user must select the desired medical sales data, forecast period such as monthly, annual, quarterly (3 months) or quarterly (4 months), forecast period, and ES methods such as SES, DES , TES were selected.

Fig.8 Report for Medical Sales Forecasting

After calculating the selected exponential smoothing method, the system displays the predicted date and value according to the prediction accuracy, and shows the best method with the least prediction error.

Fig.9 Manage Users

Authorized inventory control manager can save, edit and delete users.

6. Conclusion

In this system, exponential smoothing quantitative forecasting model was implemented to predict the number of medical sales for the next month or year or quarter. This model is based on time series analysis. The predicted medical sales numbers are accurate enough to be used by chaotic authorities to sell large quantities of drugs to customers daily. The proposed system is easy to compute and understand. This system will give the more relevant decision support for inventory reorder. In the statistical sense, prediction intervals do not really possible. Exponential smoothing produces accurate forecasts. It is best used for forecasts that are short-term and in the absence of seasonal or cyclical variations.

References

[1] Sonia Akhter, Md. Asifur Rahman, Md. Rayhan Parvez Koushik & Md. Mosharraf Hossain, “Selection of A Forecasting Technique For Beverage Production”.

[2] Ralph D. Snyder, Anne Koehler and Keith Ord, “Forecasting for Inventory Control with Exponential Smoothing”.

[3] James W. Taylor, “Multi-Item Sales Forecasting With Total And Split Exponential Smoothing.

[4] Prajakta S. Kalekar, “ Time Series Forecasting Using Holt-Winters Exponential Smoothing”.

[5] Marzena Narodzonek-Karpowska, “Smoothing Methods”.

[6] Aye Thanda Htun, “Implementation of Sales Forecasting System by using Linear Regression Method”, University of Computer Studies, Yangon, January 2014.

ONTOLOGY BASED INFORMATION RETRIEVAL SYSTEM FOR DIGITAL LIBRARY

Thet Thet Aung
University of Computer Studies, Hinthada
thetthetaung86.htd@gmail.com

[#2]Khin Lay Myint
University of Computer Studies, Hinthada
khinlaymyint.cu@gmail.com

[#3] Hlaing Htake Khaung Tin
University of Computer Studies, Hinthada
hlainghtakekhaungtin@gmail.com

Abstract

In Digital Library fields, Ontology can be used to organize bibliographic descriptions, expose the contents of documents, and share knowledge between users. IR model for digital libraries with the adaptation of Vector Space Model (VSM) and Semantic Web technologies (OWL and SPARQL) is proposed in this research. Web Ontology Language (OWL) is used to design Ontology for Digital Library using Protégé tool. In proposed IR model, metadata of resources are stored in Resource Description Framework (RDF) format and retrieved by the keywords in the user query and by the contexts defined in Domain Ontology. As the performance of IR system, 330 queries for different properties of documents are tested in 415 training documents including file types (.doc, .pdf, .txt). To evaluate the performance of IR system, precision, recall, and F-values are measured and compared. According to comparison results, Ontology-based IR system is more accurate in searching for ObjectProperty type.

Keywords: Digital Library, Ontology, OWL, Information Retrieval, Semantic Web.

1. Introduction

Nowadays, the amount of available information in both printed media and electronic/digital mediums had increased dramatically. Moreover, the number of digital documents had rapidly increased and required easy and accessed mechanized methods. In the information retrieval systems, the information is usually searched by means of a full-text search; every term in the texts of the documents can function as a search key.

Digital libraries (DLs) had become the digital counterpart of the traditional library system. In this research, Ontology-based IR system is proposed for Digital Library. Ontologies have the potential to play an important role in DL, because ontology defines a common vocabulary for researchers who need to share information in a domain.

This system intends to provide for students to retrieve the relevant information with their concept and to be able to search, read and download the textbooks, journals, thesis papers, and reference papers efficiently in the short time.

2. Related Work

In the paper with the title of Ruban S, Kedar Tendolkar, Austin Peter Rodrigues, and Niriksha Shetty presented that the domesticated plants ontology is constructed by using protégé editor in 2014. Then they suggested that the performance or result of context based search is better than the keyword based search.

In the paper with the title of Ontology Based Information Retrieval Model for Semantic Research Digital Library (SEMRDL), Shaimaa Salama, Mahmoud Abd Ellatif, and Marwa Hosny Hassan presented ontology based information retrieval model for digital library by the combination of the vector space model ranking algorithm and semantic web technologies in 2017. In this paper, the searching and retrieving results of SEMRDL model were compared with traditional digital library.

In the paper with the title of Ontology based digital library search system for enhanced information retrieval in engineering domain, Nilesh Shewale and Dr. J. Shivarama discussed about the ontologies in specific domain and semantic technologies in Digital Libraries in 2018. These ontologies and technologies can be used to access the increasing of information from the heterogeneous sources and enhance the capacity of information retrieval. And they suggested a model of ontology-based search and retrieval systems in Digital Library with exact results for the user queries and non-relevant results by eliminating.

3. Ontology for Digital Library

3.1. Digital Library

Digital libraries are a set of electronic resources and associated technical capabilities for creating, searching, and using information. They combine the structure and gathering of information, which libraries and archives have always done, with the digital representation that computers have made possible. The main purpose of a digital library is to collect, manage, and preserve in perpetuity digital content.

3.2. Ontology in Digital Library

Ontology languages allow users to design ontologies according to user needs. Ontology for digital libraries pertinent examples exist such as RDF (Resource Description Framework), XML (Extensible Markup Language), for describing data, information, and knowledge; OWL (Web Ontology Language), is becoming the standard for describing ontologies and accessing resources through the web.

3.2.1. Web Ontology Language (OWL)

Web Ontology Language (OWL) is a language for defining and instance ontologies in the Web. This includes descriptions of classes and their properties and their relationships.

OWL can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms. OWL has more facilities for expressing meaning and semantics than XML, RDF, and RDF-S. OWL ontology consists of three components: Individuals, Properties, and Classes.

3.3. SPARQL Query Language

Google SPARQL is a query language and a protocol for accessing RDF designed by the W3C RDF Data Access Working Group. RDF is a directed, labeled graph data format for representing information in the Web. This specification defines the syntax and semantics of the SPARQL query language for RDF. SPARQL contains capabilities for querying required and optional graph patterns along with their conjunctions and disjunctions. The results of SPARQL queries can be result sets or RDF graphs.

3.3.1. Context Matching with SPARQL Queries

Input Query: "Aung Myint"
Input Property: *name*-"hasAuthor", *type*-"owl:ObjectProperty", *range*-"dl:Author"
Output SPARQL:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dl: <http://www.owl-ontologies.com/library.owl#>
SELECT ?document ?i
WHERE {
        ?document dl:hasAuthor ?i.
        ?i a dl:Author.
        ?i dl:name ?label
        FILTERREGEX(?label,
        '\\baung\\b|\\bmyint\\b', 'i') }
```

3.4. Models of Information Retrieval

There are four types of models: Boolean Model, Language Model, Probabilistic model and Vector Space Model.

3.4.1. Vector Space Model

In the vector space model text is represented by a vector of terms. A vector-based information retrieval method represents both documents and queries with high-dimensional vectors while computing their similarities by the vector inner product.

3.5. Calculating TF-IDF and Similarity

The vector space model is a statistical model for representing text information for Information Retrieval. It is a simple, mathematically based approach that provides partial matching and ranked results. TF-IDF weighting, the most common term weighting approach is measured how often the term *j* occurs in document *i* (the Term Frequency) and IDF (the Inverse Document Frequency) as shown in Equation 1 and 2.

The weight equation for the term within document is as follows:

$$w_{ij} = tf_{ij} \times idf_i \quad (1)$$

where,

w_{ij} = weight of the term t_i in document d_j

tf_{ij} = the normalize term frequency (TF) of term t_i in document d_j

idf_i = the inverse document frequency (IDF) of term t_i

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \ldots, f_{|v|j}\}} \quad (2)$$

where, f_{ij} = the raw frequency count of term t_i in document d_j

$$idf_i = \log \frac{N}{df_i} \quad (3)$$

where,

df_i = number of document in which term t_i appears at least once

N = the total number of documents in the system

A query q is represented in exactly the same way as a document. The weight equation for the term within query is as follows:

$$w_{iq} = \left[0.5 + \frac{0.5 f_{iq}}{\max\{f_{1q}, f_{2q}, \ldots, f_{|v|q}\}}\right] \times \log \frac{N}{df_i} \quad (4)$$

where, w_{iq} = weight of the term t_i in query vector q

f_{iq} = the raw frequency count of term t_i in query vector q

The similarity between query q and j^{th} document retrieved by context matching process is calculated by Dice similarity method as shown in Equation 5. This is quantified as the

$$Dice(d_j, q) = \frac{2\left|\sum_{i=1}^{|v|} w_{ij} \times w_{iq}\right|}{\sum_{i=1}^{|v|} w_{ij}^2 + \sum_{i=1}^{|v|} w_{iq}^2} \quad (5)$$

4. System Design and Implementation

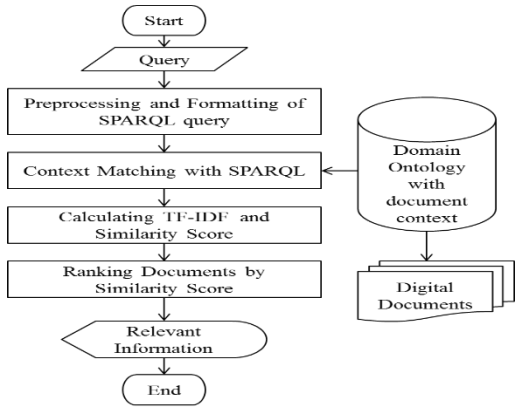


Figure 1. System Design

In Figure 1, query preprocessing, which consists of the tokenization and stopwords removal process for the user query, is performed. This system accepts the query and property selected by the user. The tokenized keywords and selected property by the user are transformed to SPARQL query format by the algorithm for the formatting of SPARQL query. The matching process with the context of documents from Domain Ontology and the

formatted SPARQL query is performed. The results of this process are relevant documents by the keywords and property of the document. And then, relevant documents retrieved are calculated for TF-IDF values and similarity scores by using the Vector Space Model (VSM) and the Dice similarity method respectively. Retrieved documents are ranked according to their similarity scores, and evaluation of the results of IR is performed in the final step by calculating its precision, recall, and f-measure values. Finally, the relevant documents retrieved by SPARQL query are ranked and displayed as the result of our Ontology-based IR system.

4.1. Implementation of Digital Library

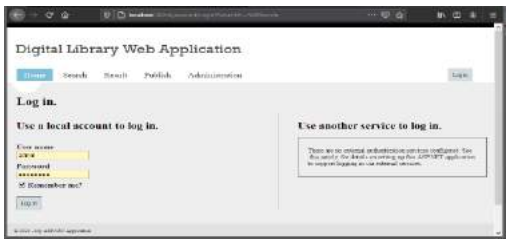


Figure 2. Login Page of Digital Library Web Application

There is no registration process in this system and user accounts is done only by the admin of our system. After the login process is performed, the webpage of the "Home" menu can be seen as shown in Figure 2.

Figure 3. Home Page of Digital Library Web Application

In Figure 3, admin and user can see Home page of the system and documents information after they log in with accounts.

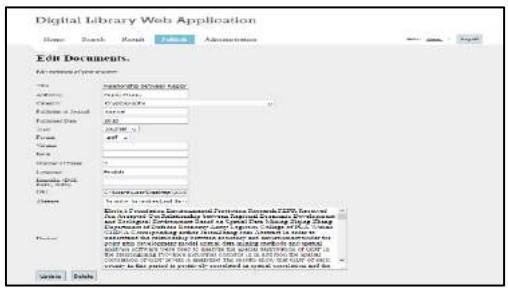


Figure 4. Edit Page of Digital Library Web Application

In Figure 4, admin not only can edit the metadata of documents but also can publish them to Ontology dataset.

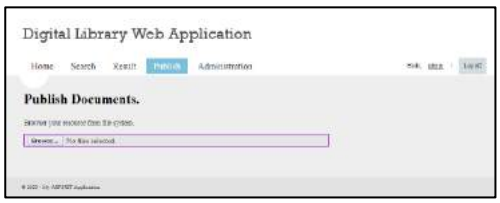


Figure 5. Publish Page for Browsing the Document

The "Publish" page which is shown in Figure 5 is used to browse a file from the computer and import it to our dataset and file storage.

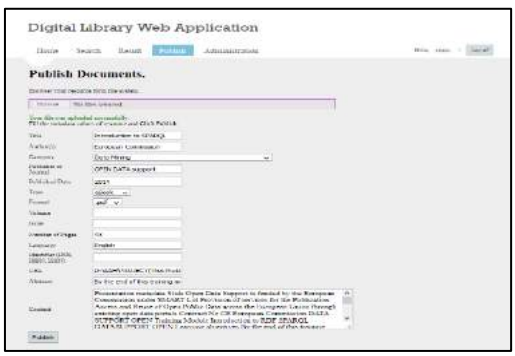


Figure 6. Publish Page with Metadata Values

Once a file is browsed, its' content, URL, and format are automatically extracted by our application. The values for all the rest of the properties of the document should be filled by the admin manually in the publication process in figure 6.

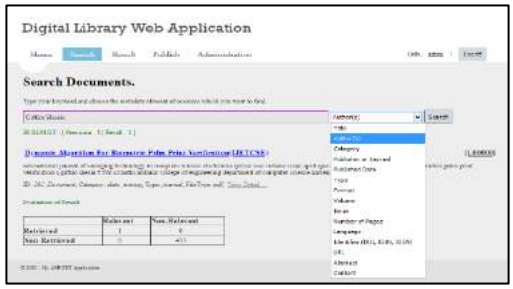


Figure 7. Search Page of Digital Library Web Application

The main functional page in our Digital Library web application is the "Search" page. On this page, both users and admin can search for different types of documents by given query with the different property of document. In Figure 7, the retrieved document is shown as a result of searching for document in Author(s) property by given keyword. In the retrieving process, the evaluation of precision and recall is performed for the current search.

Figure 8. Administration Page with User Information

The "Administration" menu of the web application is used to manage the accounts of library users. An account consists of information about username, password, email, role, and date created. Admin can define the new user by clicking the link "Add User" in the "Administration" page which is shown in Figure 8.

5. Performance Analysis

To show the performance of the system, 33 queries for different properties of documents were tested by using 415 training documents that include various file types (.doc, .pdf, .txt). These testing queries are related to Object and Datatype Properties.

To evaluate the performance of Ontology-based IR system for Digital Library, precision, recall, and F-measure methods are used as shown in Equations 6,7 and 8.

Precision (P), P = TP / (TP+FP) (6)

Recall (R), R = TP / (TP+FN) (7)

F-Measure (F), F = 2 * [(P*R) / (P+R)] (8)

Where TP and FP denote the number of relevant and non-relevant documents in retrieved documents. FP and FN denote the number of relevant and non-relevant documents in non-retrieved documents.

The precision, recall, and f-measure values of experimental results for the ObjectProperty are shown in Figure 9.

Object Property

PropertyName	Query Keywords	NoOfRetrieved	Precision	Recall	F-Measure
dl:hasAuthor	Qilin Shen	1	1	1	1
dl:hasAuthor	Aung	7	0.71	1	0.83
dl:hasAuthor	Kirti Raisinghani	1	1	1	1
dl:hasCategory	digital signal	5	1	1	1
dl:hasCategory	data mining	69	1	1	1
dl:hasCategory	artificial intelligence	23	1	1	1
dl:hasFileType	txt	1	1	1	1
dl:hasFileType	doc	30	1	1	1
dl:hasFileType	pdf	285	1	1	1
dl:hasType	ebook	124	1	1	1
dl:hasType	Journal	161	1	1	1
dl:hasType	thesis	10	1	1	1
AVERAGE			0.97	1	0.98

Figure 9. Precision, Recall and F-measure Results for ObjectProperty

In the above table, the precision (P), recall (R), and f-measure (F) values for four ObjectProperty of documents are shown. There is only one property, "has Author" which obtain the precision under 0.75 in this table. The recall, the average precision and the average F-measure value for all properties are 1, 0.97 and 0.98. According to these results, the exactness and completeness of Ontology-based IR systems in ObjectProperty is over 98%. The precision,

recall, and f-measure values of experimental results for the DatatypeProperty are shown in Figure 10.

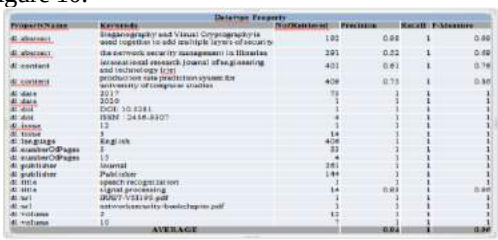


Figure 10. Precision, Recall and F-measure Results for DatatypeProperty

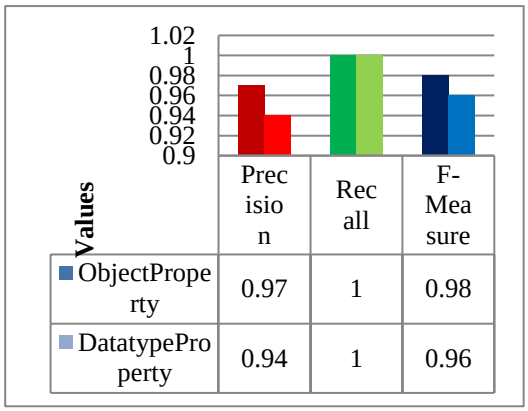


Figure 11. The Comparison of performance in precision, recall and F-measure

The average results of Ontology-based IR system for ObjectProperty and DatatypeProperty are compared and described with the bar chart in Figure 11. According to the comparison results of precision, recall, and f-measure, the Ontology-based IR system is more accurate in ObjectProperty type because the values for this property are all instances of an OWL class.

6. Conclusion

This system presents the implementation of Ontology-based information retrieval for Digital Library. This system introduces a system that users can use to retrieve digital documents from the Ontology dataset. The ontology plays a key role in the evolution of digital libraries and its method is used to represent the context model based on digital library resources. In interoperability at the semantic level, context-sensitive query processing over heterogeneous information resources requires the matching of concepts. The system presents the available heterogeneous information sources and improves the accuracy of information retrieval using semantic web technology. In addition, the system can help users to reduce the consuming time for surfing the information they wanted.

References

[1] Deborah L. McGuinness, Frank. (2004). OWL Web Ontology Language-Overview, W3C Recommendation, 10 February 2004. https://www.w3.org/TR/owl-features/access

[2] DLF (1998). A working definition of digital library. http://old.diglib.org/about/dldefinition.html

[3] An Ontology-Based Information Retrieval Model for Domesticated Plants(2014) https://www.semanticscholar.org/paper/An-Ontology-Based-Information-Retrieval-Model-for-Ruban-Tendolkar/7c000955db5bf89862a433550ece324009a8f2a7

[4] Ontology Based Information Retrieval Model for Semantic Research Digital Library (SEMRDL) (2017) https://www.academia.edu/33410886/Ontology_Based_Information_Retrieval_Model_for_Semantic_Research_Digital_Library_SEMRDL

[5] Ontology based digital library search system for enhanced information retrieval in engineering domain (2018) https://www.semanticscholar.org/paper/Ontology-based-digital-library-search-system-for-in-Shewale-Shivarama/059e32d5cc30c751160601e6c216ccbcb7aba096

GENDER CLASSIFICATION FROM MYANMAR (NRC) CARD

Soe Thiri Hlaing, Thiri Naing
Computer University (Kalay)
soethirihlaing@ucskalay.edu.mm, thuthiri@gmail.com

Abstract

Gender identification is a fundamental task of face recognition, deciding the gender according to the face image. Today, it is becoming increasingly popular for security. Gender is an important factor in social activities. In this paper, an investigation of gender classification by face from Myanmar NRC card is proposed. In this investigation, there are three parts such as face detection and extraction, facial feature extraction by PCA and classification. Initially, the facial area is detected using the Viola jones algorithm and then face region is extracted from the NRC card. In the next step, the detected face region is subjected to Principal Component Analysis (PCA) to extract facial features. For classification, these principal components are exposed to SVM classifier. This gender classification system is implemented on Matlab by using own dataset. The accuracy is obtained 89.06% at 80 number of training images. Testing error rate is about 10.94%.

Key words: Gender identification, Principal Components Analysis (PCA), Viola jones, Support Vector Machine (SVM).

1. Introduction

The face is one of the most important biometric features of the human beings and normally used as identification. Each person has their own innate face and mostly a different face. Gender classification [1,2,3] using the face is very useful for human-computer interaction and control systems. In the large population, an individual's authentication process is usually time consuming. The division of the population into two parts based on gender is the possible solution of this issue. The process of gender classification according to face image is needed to use clearly distinguished features and robust classification methods.

In many years ago, the facial images are used in gender classification has become an important role. When you see a person's face, it can be easy to identify a male or a female, but it's a difficult task for a computer. The computers require some significant information to make the classification. Gender identification can be done in different appearance such as the gait, iris, hand shape and voice etc. [3]. However, the most popular techniques for gender classification were always standing on facial features. There are different characteristics between the man and woman to classify the gender.

During the process of gender identification, firstly there are face detection and then studying the facial features is included. Therefore, the machine needs to be the suitable facial features. Principal Component Analysis (PCA) method is the most frequently used for facial classification. Gender recognition can be used in numerous applications such as Identity authentication, data collection, search engine recovery and monitoring.

In this system, the selection of principal components is used as a method for extracting facial features. The structure of this paper is as

follows: Section 1 introduces the analysis of facial features and gender classification. Section 2 reviews the gender classification that includes features based on geometrical and appearance features. An overview of the proposed system is provided in Section 3. And Section 4 descries the methods including face detection and extraction and the operation of facial features extraction by PCA. SVM classification techniques is discussed in Section 5. The implementation and experimental results of this system is presented in Section 6. In Section 7, the conclusion of the system is presented.

2. Related Works

Today's analysis of human faces is an interesting research area on gender recognition. It is necessary to be faster and more convenient system. In this study, it is tried to determine gender from Myanmar NRC card face images. The first introduction to gender classification is described in [3]. A multi-layer neural network is used to recognize gender from facial images. They had 91.9% accurate rate on 160 images.

In [4], a hybrid approach was used. It presented a hybrid approach by combining global features with local features. The Adaboost algorithm and active appearance model AAM are used to extract global features and local features individually. They point out that it receives greater accuracy by using the hybrid method. [5] proposed "Ethnicity Identification from Face Images" Ethnicity classification was done by using LDA. Different datasets separated into Asian & Non-Asian. Ensemble LDA gives more accuracy than nearest neighbor classifier. In [6], gender classification system was presented using linear discriminant classifier and Support Vector Machine. The experiment was done with different types of classification methods namely the cosine classifier, the linear discriminant classifier and SVM. The author of [7] used Continuous Wavelet Transforms to find features for each male and female face. The Wavelet Coefficient obtained was given to SVM for classification. The experiment was conducted in an ORL database containing 400 photos, both male and female. The kernel used for SVM is linear and the resulting rating is 98% compared to Radon Transform and Discrete Wavelet Transform.

3. Overview of the System

The face is one of the most important biometric aspects of humanity. Each person has their own innate facial appearance that are mostly different each other. Therefore, this system is investigated the gender classification through facial image using principal component analysis (PCA) and support vector machine (SVM).

The proposed system consists of three parts such as face detection, feature extraction from detected face by PCA and classification of male or female. The process flow diagram of this system is showed in Figure 1.

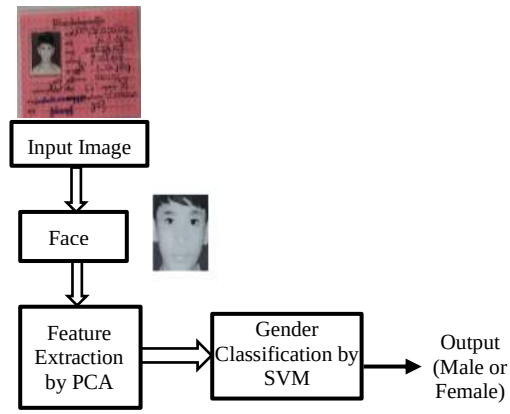


Figure 1. Block diagram of the proposed system

Initially, the facial area is detected using the Viola jones algorithm and then the detected face region is extracted from the NRC card. In the next step, the detected face region is subjected to Principal Component Analysis (PCA) to extract

facial features. The principal components are exposed to classifier to classify the gender. Lastly, Support vector machine (SVM) classification techniques will be evaluated for gender classification. For the system to be properly examined, the system needs to be trained and tested using the face from Myanmar NRC card.

4. Proposed System

Face detection and gender identification play a vital role in video investigation for monitoring area and database management system. Gender classification by face from Myanmar NRC card is proposed in this system. This system is implemented by using its own Datasets. This process involves three stages. Firstly, Face detection (Viola-Jones algorithm) is used. Then PCA is applied to the face images to extract facial features. Finally, Support vector machine (SVM) classification techniques will be evaluated for gender classification. The proposed system is implemented on the MATLAB platform.

4.1. Face Detection

It is the process of identifying one or more human faces in images or videos. The Viola-Jones algorithm is used in this system to capture faces on NRC card images. Because of its high detection rate, and its ability to run in real time. It is the most effective for frontal images of the face, and it can manage the rotation of 45 degrees for horizontal and vertical. This is due to its high intelligence and timely performance. The useful knowledge for facial features are the location and size of eyes and nose. And the darkness and brightness value are also important features.

4.1.1 Viola-jones Object Detection Framework

The Viola-Jones algorithm was the first object detection framework proposed by Paul Viola and Michael Jones in 2001 to provide real-time competitive object search rates. The algorithm looks at several sub-regions and tries to find a face by searching for specific features in each region. Viola and Jones [4,5,6] used Haar-like features to detect faces in this algorithm.

4.1.2 Haar like features

Haar like features [6] are digital images used in object recognition. It can be used to find the difference between black and light in an image. All human faces have unique properties. The eye area is deeper than its neighbors, and the nose area is brighter than the eye area. A simple way to find out which area is lighter or darker is to compile and compare the pixel values of the two areas. If one side is lighter than the other, it may be the tip of the eyebrow. Sometimes the center will be brighter than the surrounding boxes; It can be interpreted as a nose.

In Figure 2, the result of Viola Jones Face detection algorithm is shown.

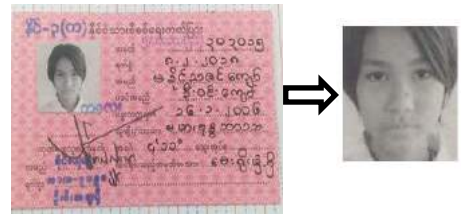


Figure 2. Face Detection Result

4.2 Features Extraction

The important thing to understand is that every face has a different pattern. The task of the face recognition system is to show that the test image belongs to a person in the database. Every image is random in nature because the lighting conditions, the orientation of eyes, facial features, hair and spectacles are different for different people. However, statistical characterization can still be carried out on this random set. They can be extracted from the original image using a mathematical tool called PCA [7,8,9].

4.2.1 Principal Component Analysis

PCA is a dimensionality reduction technique, which is used to represent each image as a feature vector in a low dimensional subspace. Although the PCA is a traditional way of representing faces and it is the most widely used method in today. In this step, principle component analysis is applied to facial images to extract facial features. Flow chart of PCA Algorithm is as shown in the Figure 3.

Some features, such as set of eyes, mouth and nose, are presented with relative distances and they can help to distinguish the face. The characteristic of these features are called the principal components or the Eigen faces. Each principal component has a different robustness according to the amount of variance in its direction. One of the key features of PCA is that reconstructing any original image from the training set by integration with the Eigen faces [8,9]. Eigen faces are represented with the certain features of the original image. The Eigen faces on the Myanmar NRC card by PCA are shown in Figure 4.

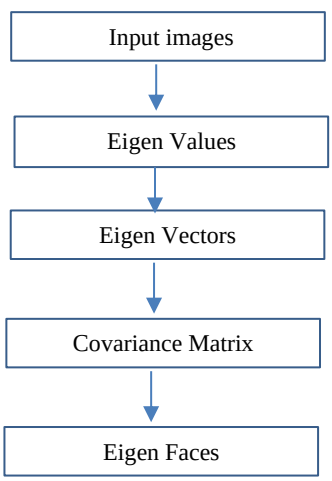


Figure 3. Flow Chart of PCA Algorithm

Step 1: The first step for the PCA is to get the data set.

Step 2: The next step of PCA is to calculate the mean of a given data set.

Step 3: The next step is to subtract the mean with each data of the data set, then, this produces a data set whose mean is zero.

Step 4: Then calculate the covariance of the matrix. If the data are two dimensional, then the covariance matrix will be of 2 × 2 and if the data is of N dimensional the covariance matrix will be of N × N.

Step 5: The next step is to calculate the eigenvalues and eigenvectors of the covariance matrix.

Step 6: The last step of PCA is to select the components and forming a feature vector.

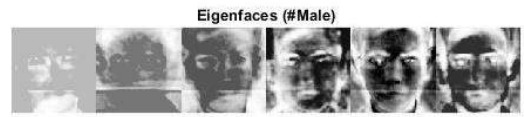


Figure 4. Eigen Faces

5. Classification

Gender classification methods can enhance the performance of many other applications human recognition and video surveillance system. The classification method, Support vector machine (SVM) is used in this system. The facial features from feature extraction step are inputted to a support vector machine (SVM) classifier to classify male or female.

5.1 Support Vector Machine (SVM)

SVM provides a very high accuracy compared to other ratings, such as logistic regression and decision trees. It is used in many applications such as genetic classification and handwriting recognition. In this system, SVM [10,11,12] is used to identify the gender. The Eigen vector W, calculated from the feature extraction stage, is used at this classification stage. The weight values calculated from the training phase are used to train

the classification algorithm and the weight values get from recognition phase are used as the input.

A Support Vector Machine performs classification by constructing an N-dimensional hyper-plane that optimally separates the data into two categories [11]. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data-points of any class. In general, if the larger the margin, the generalization error of the classifier is lower. Without any knowledge of the mapping, the SVM can find the optimal hyper-plane by using the dot product functions in original space that are called kernels. The theory of SVM [12] is based on the idea of structural risk minimization. And special property of SVM is simultaneously minimized the empirical classification error and maximize the geometric margin.

6. Experimental Setup

Gender classification by face from Myanmar NRC card is proposed in this system as in figure 5. To study the system performance, the system must be trained using the face from the Myanmar NRC card. 80 images (males and females) of own dataset is used to analyze the system performance. Data collection is captured the front view of Myanmar NRC cards by using Sony camera. The images of database are under 30 years old and the images sized are 100 ×100 pixels. To evaluate the proposed system, SVM classification method is applied. The number of training images were 80 including 40 males and 40 female images. The number of different testing images were 52 including 27 male and 25 female images. Accuracy calculates the exact gender results. It measures the percentage of face images that were classified into correct gender, which is the ratio of the accurate predictions to the total number of the ground-truth labels. Mathematically, the metric is described as follows:

$$Accuracy = \frac{No.of\ accuarate\ predicion(correctly)}{Total\ number\ of\ prediction}$$

Figure 5. Gender Classification System

The accuracy is obtained 89.06% at 80 number of training images. Moreover, the results shown on each types of gender (males and females) database separately. For the males, the number of objects is contained 67 images and for females contained 67 images individually. According to the results, the accuracy of SVM is 89% for male type and 93% for female type. Testing error rate is about 10.94%. In this experiment, the classification results of 'Male' type are also lower than the 'Female' type of gender.

Advantages: There is no need for direct contact with any like other biometric system such as finger print, voice and signature etc. **Disadvantages:** Face recognitions are not able to perform well in the variation of illumination. Face recognition systems are not always accurate.

System limitation: In poor lighting, low resolution images does not work well.

7. Conclusion

This system is proposed the human gender classification by face images from Myanmar NRC card. The experimental evaluation of proposed system confirms the good performance. According to the experimental results, SVM classifier achieves as high as 89.06% gender classification accuracy for 80 subjects. For further

experiments, the proposed system will be tested on the larger value of dataset and will be implemented the other machine learning algorithm such as KNN and Neural Network to compare with the proposed method.

References

[1] H. B. Kekre udeep, D .Thepade, C.Tejas , *"Face and Gender Recognition Using Principal Component Analysis"* International Journal on Computer Science and Engineering (IJCSE) Vol. 02, No. 04, 2010, 959-964

[2] S. Avinash, C. Mridul , K.S Deepak, C. Yogesh, "*Gender Classification using Facial Embeddings: A Novel Approach*" International Conference on computational Intelligence and Data Science (ICCIDS 2019).

[3] G.Kavitha, I.Laurence Aroquiaraj "*Face Detection and Gender Classification using Facial Features*" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019.

[4] Smriti Tikoo, Nitin Malik "*Detection of Face using Viola Jones and Recognition Using Back Propagation Neural Network*", Research gate, on 31 January 2017.

[5]. Chai, T. Y., Rizon, "*Facial Features for Template Matching Based Face Recognition*". American Journal of Applied Sciences, vol. 6, no.11, pp. 1897-1901, 2009

[6]. D. Mohammad, A. Alqudah and O. Debeir, *" Face Detection using Viola and Jones Method and Neural Networks*" IEEE International Conference on Information and Communication Technology Research , pp 40-43,2015.

[7] K.J. Diamantaras and S.Y. Kung, "*Principal Component Neural Networks: Theory and applications*", John Wiley and Sons, Inc., 1996

[8] R. Sharma and M. Patterh (2015), *"Age invariant face recognition using k-pca and k-nn on indian face age database*," International Journal of Computer applications**,** vol. 126, no. 5.

[9]. Steven M. Holand, "Principal Component Analysis(PCA)", in *Department of Geology*, University of Georgia, Athens, 2008, GA 30602-2501.

[10]. David Shaw,*'Support Vector Machines for Classifying Face Data'*, Due:12/14/2009.

[11]. Baback, Mand Ming.H.Y Yang '*Gender Classification with Support Vector Machines'* 'Proceedings of the 4th IEEE International Conference on Face and Gesture Recognition, March, 2000.

[12] Amit Jain, Jeffrey Huang, "*Integrating Independent Components and Support Vector Machines for Gender Classification"* ,ICPR , pp. 558-561, 2004.

Enhancing the Clothes Searching System using Combination of K-prototype and kNN Algorithm

Thin Thin Htwe, Hnin Pwint Phyu
University of Computer Studies, Taungoo
thinthinhtwe@ucstaungoo.edu.mm, hninpwintphyu@ucstaungoo.edu.mm

Abstract

Although k-Nearest-Neighbors algorithm is one of the most popular methods among many classification tools, it has resulted in higher computation time complexity as it used all training data to classify a new test sample. In order to solve the above mentioned problems, it is necessary to combine traditional kNN algorithm and K-prototype cluster algorithm. In proposed system, kNN is mainly used to find similar item with user preference and to produce the similar dress codes for reducing time and labor. There are two main stages in this system. The first stage is the grouping stage which is clustering the clothes group depending on their attribute values by using K-prototype algorithm. The second stage is the testing stage which is searching the most similar clothes depending on user choices by using k-Nearest- Neighbors algorithm. This system is then analyzed in terms of processing time, compared to highlight the processing time by using the combination of K-prototype and kNN algorithms.

Keywords: clothes searching system, K-prototype, kNN algorithm, normalization, clustering

1. Introduction

In digital era, every business needs to run the technology in their workplace in order to keep pace with demanding customer services and satisfaction. Today, the speed of computer software used to find the clothes that customers want is considered to be a bit slow. To resolve that problem, the computerized software can be implemented by using some data mining technique.

The supervised learning classification is one of the most well-known techniques that helps to classify the new observation in which category belongs. The most popular classification algorithm is the k-Nearest-Neighbors classifier (kNN) which is widely applied in various fields [14,5] and often used in search applications where you are looking for similar items [17]. In this system, k-Nearest-Neighbors search algorithm for searching most similar dress codes. However, the determination of classes for new data is generally accomplished by all training samples in the classification of kNN that may cause to a high level of computation complexity. For reducing the computation complexity of kNN, the hybrid of K-prototype and kNN algorithms should be used.

Among clustering algorithm, K-means algorithm is mainly used for only numeric data and K-prototype algorithm is used for the mixed data such as numeric and categorical data [12,4]. K-prototype algorithm is used to separate the different groups from a large amount of all data objects. In this system, a combine of K-prototype and kNN algorithms which is used in the grouping stage and testing stage respectively. In case, it is worth stating that kNN can select the best subset of the original dataset that are very nearly matched with user's preferences. Moreover, it is found that the processing time of using kNN with K-prototype can reduce the time of using only kNN. The rest of the paper is organized as follows: the related work is described in Section-2. The background theory is widely explained in Section-3. The next section performs the proposed system experiments and proves the efficiency of the system with tabulated data. The evaluation results

are showed in Section-5. Then, the paper is concluded in Section-6.

2. Related Work

In the area of customer assistance works, recommendation systems are implemented using data mining algorithms and user center preferences and rating.

M.M.Mon et al. [9] proposed clothes recommender system by using bisecting K-means algorithm. In this system, bisecting K-means methods are used to search user's desired clothes. This method is more efficient when data set are large. So, if data set is small, it has a few problems when separating data into two. P. W. Buana et a1. [11] proposed the news classified system. In this paper, the Indonesian news are classified by using combination of K-means and kNN based on term re-weighting. Author proposed kNN has high calculation complexity to find out the k nearest neighbor samples, all the similarities between the training samples must be calculated.

A.G.Karegowda et a1. [2] proposed the cascading model with K-means clustering and k-nearest neighbor classifier has been successfully used for markable improved classification of PIMA diabetic dataset. R.Ahuja et a1. [13] proposed the moving recommender system using K-means and k-nearest neighbor algorithm. This system uses K-means and kNN to find the movies that most closely match the user's desired movie. Author finds out as the number of clusters decreases, the value of RMSE also decreases.

In the above system, prior to the kNN classification work, only the K-means algorithm was used for preprocessing to separate clusters. K-means is suitable for clustering numeric data, but is not effective for clustering mixed data. The disadvantage of K-means is that it first converts categorical data into numerical data. In K-means, it takes more time to convert the data in the dataset to numeric before performing cluster processes. In real world, data sets usually contain both numeric and categorical attributes. For this case, K-prototype can handle mixed-attributes types without changing data type [12,4,8]. Therefore, using K-prototype algorithm saves preprocessing time. Therefore, the combination of K-prototype clustering algorithm and kNN algorithm are used in this propose system for finding the dress codes that the user wants.

3. Methodology

Data mining is the task of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses, or other information repositories. Data mining functionalities include association, classification, prediction, clustering, trend analysis, deviation analysis, and similarity analysis [6].

Clustering is supervised learning and Classification is unsupervised learning methods. Clustering is the process of grouping data into clusters, so that objects within a cluster are similar to one another and dissimilar to objects in other clusters [6,7]. Classification assigns items in a collection to target categories or classes [6]. Before starting the system, firstly the data normalization must perform preprocessing part.

3.1. Data Normalization

Normalization is used for making no gap between the values of data in training samples. If the values of the data are different, the output may not be accurate. Therefore, data normalization should be done first to get the correct output. Only then will you get the right result when splitting clusters. Min-max normalization performs a linear transformation on the original data. This method transfers data from its domain to a specific range such as between (0, 1) [3,8,10]. It is a technique that can help in improving the K-means clustering algorithm and K-prototypes algorithm [1,3,8]. The following formula in equation 1 is used for the normalization:

$$V_i' = \frac{V_i - \min_A}{\max_A - \min_A} (new\_max_A - new\_min_A) + new\_min_A \qquad (1)$$

Where, V_i means the current rating value in such that the value of new_max_A is 1 and the value of new_min_A is 0.

3.2. K-prototype Algorithm for Clustering

There are a number of clustering algorithms that has been widely used in various application domains. However, some algorithm has restricted to handle numerical data only. K-prototype is a type of clustering algorithm that is well-known of dealing with both numerical and categorical variables. The main objective of K-prototype algorithm is to partition the dataset into k clusters by reducing the cost function [12]. The K-prototypes algorithm, through the definition of a combined dissimilarity measure, integrates the K-means and K-modes algorithms to deal with the mixed data types which are more beneficial since data collected in the real world are mostly mixed type objects [4,16].

Assume that a set *n* objects, $X=\{X_1, X_2,\cdots,X_n\}$ such that $X_i=\{X_{i1},X_{i2},\cdots,X_{im}\}$ consists of *m* attributes in which m_r is numerical attributes and m_c is categorical attributes, $m=m_r+m_c$. The dissimilarity between two mixed-type objects X and Y, which are described by attributes A_1, $A_2,\ldots,..A_p$, $A_{p+1},\ldots,$ A_m, can be measured by equation 2:

$$d(X_i,C_j) = d_r(X_i, C_j) + d_c(X_i, C_j) \quad (2)$$

$$d_r(X_i,C_j) = \sqrt{\sum_{l=1}^{p}(x_{il}- c_{jl})^2} \quad \text{//For numerical data} \quad (3)$$

$$d_c(X_i,C_j) = \sum_{l=p+1}^{m} \delta(x_{il}, c_{jl}) \quad \text{//For categorical data} \quad (4)$$

Where the first part of equation (equation 3) is squared Euclidean distance on numerical attributes and the second part of the equation (equation 4) is simple matching distance on categorical attributes. The procedure of K-prototypes is described as follows [7]:

Input: X: dataset, K: the number of cluster

Output: K cluster

Step-1: Select K initial prototypes from the dataset X, one for each cluster.

Step-2: Allocate each object in X to a cluster whose prototype is the nearest to it. This allocation is done with considering the dissimilarity measure.

Step-3: After all objects have been allocated to a cluster, recalculate the centroid points of every cluster.

Step-4: Repeat step 2 and 3, until no object changes its cluster after a full circle test of X.

After clustering for each category, the cluster centres were chosen to represent the category and they become the new training sets for kNN algorithm.

3.3. k-Nearest-Neighbors Algorithm for Searching

The kNN is instance-based or lazy learners [6]. It delays the process of modeling the training data until it is needed to classify the test samples. It can be used for both classification and regression on prediction problems and often used in search applications. The advantages of kNN are very simple algorithm to implement and to justify the result of kNN. But it has some disadvantages such as high computation time since it needs to calculate distance of each test instance to all training samples [2]. The purpose of this algorithm is to find similar items. It finds the k most similar items to a particular instance based on a given distance metric like Euclidean and Manhattan distance measures.

The procedure of kNN is described as follows:

Step 1: Load the data

Step 2: Initialize the value of **k**

Step 3: Iterate from 1 to total number of training data points for getting the most similar items

3.1: Calculate the distance between new test data and each row of training data. (Using the equation-2)

3.2: Sort the test data in ascending order based on calculated distance

3.3: Get top **k** rows from the sorted data

3.4: Return the most similar item of these rows.

In this system, kNN is used for searching the most similar dress according to user input **k** number.

4. The Proposed Clothes Searching System

The clothes searching system is organized with two stages, which is called grouping similar item and searching the most similar item to show the users. The working flow of the system is as shown in Figure 1.

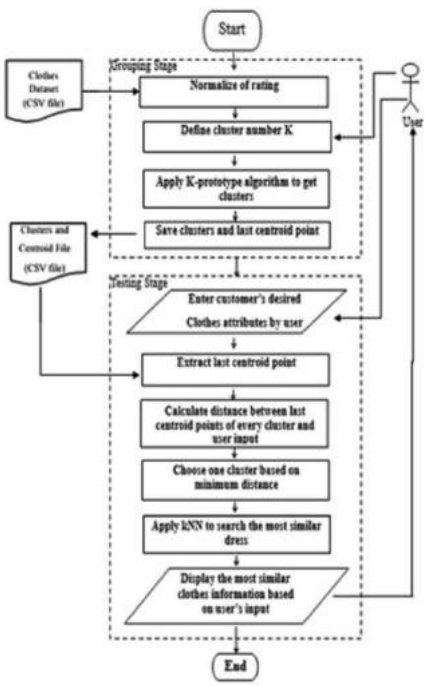


Figure 1. System flow diagram

The grouping stage firstly normalizes rating and defines the k value (i.e. number of clusters) to partition the various items group. And then, K-prototype, clustering algorithm is applied to get clusters having the similar items in dataset according to their attribute value. As a final point of this stage, the last centroid points of resultant of clusters are saved into the centroid CSV file that can be used for testing task. In the second stage (so-called testing stage), the classification is done by using kNN algorithm. Applying the kNN is to search the most similar dress. Initially, user needs to enter the attributes of desired clothes and then the system extracts the last centroid point in the centroid file. It then calculates the distance between last centroid points of every cluster and the target dress from user input and chooses one cluster based on minimum distance. When kNN makes inference about a dress, it will calculate the "distance" between the target dress and other dress from the selected cluster, then it ranks its distances and returns the top **k** nearest neighbor dresses. As a result, the system displays the most similar clothes information with their codes that is consistent with the customer's preferences. Therefore, it may reduce the manual labor as well as searching time to get the most matched clothes.

4.1. Data Collection

The datasets used in this paper is availed from UCI Machine Learning Repository [18]. To start the system, we first take a description of the dataset which contains totally 501 instances. The attributes of the dataset are price, rating, size, season, neckline, sleeve length, waist line, material, fabric type, decoration and pattern type. Among then, only rating attribute is numeric value, whilst the others are categorical values. Each instance in the dataset is represented by ID codes as shown in Figure 2.

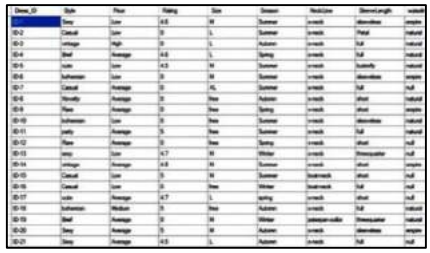


Figure 2. Original dress data

4.2. Data Normalization of Clothes Searching System

The reason why the normalization purpose is that the min-max normalization is used for rating attributes of the dataset in order to have rating values between 0 and 1 [1].

After the preprocessing task, the result data of the normalization is showed in Figure 3.

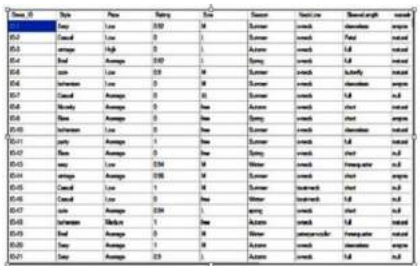


Figure 3. Result of normalization

4.3 Grouping Stage

In grouping state, the k-prototypes algorithm allocates data items into the *K* clusters with the smallest distance by calculating the distance between a new instance and each cluster center to be allocated to that cluster. For finding the most similar clothes group, the last centroid points are stored. After clustering for each group, the cluster centers were chosen to represent the group.

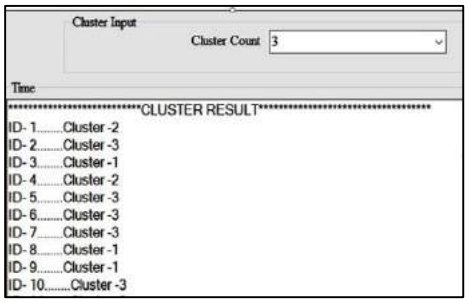


Figure 4. Clustering result form

This group becomes the new training sets for kNN algorithm. By using this approach, the number of instances for training is reduced efficiently and the time for calculating similarities in kNN algorithm is also reduced. In this section, the number of clusters is divided into three clusters. The calculation of cluster results is as shown in Figure 4.

The group split patterns are displayed with the visual image in Figure 5.

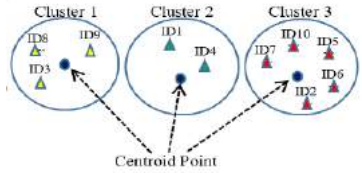


Figure 5. Sample image of resulted clusters

4.4 Testing Stage

The minimum distance group was selected by calculating the user input and distance using the centroid points obtained from the previous step. The resulting group is the closest group to the item the user wants. kNN is used for searching the most similar items according to the user input k from the group resulted from K-prototype. It finally produces the results of user-wanted items, which is the final output (dress codes) of the system as shown in Figure 6. In this section, the group closest to the user input is Cluster-3. Then, the system calculated the distance between the user input and the instances in that group by using equation 2 and searched the most suitable clothes for it. Once the user sets the desired k value is 10, the system displays the top 10 rows from the sorted array in that cluster.

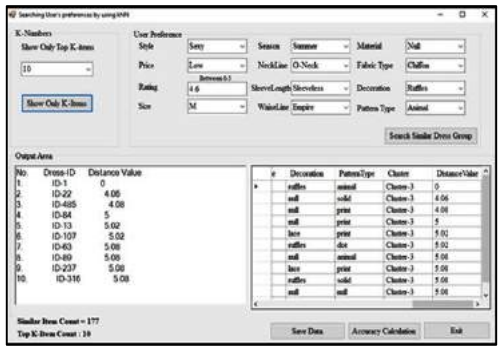


Figure 6. The most similar group with user's preference by sorting the distance

Dress_ID	Style	Price	Rating	Size
ID-1	Sexy	Low	0.92	M
ID-22	Casual	Low	0.86	M
ID-485	Casual	Low	1	M
ID-84	vintage	Low	0.92	M
ID-13	sexy	Low	0.94	M
ID-107	cute	Low	0.94	M
ID-63	Casual	Average	1	M
ID-89	Sexy	Low	1	M
ID-237	bohemian	Average	1	M
ID-316	Sexy	low	1	free

Figure 7. Showing the Dress code of the most similar items

5. Evaluation Results

The system used the two algorithms to reduce the time of classification process. In order to prove the efficiency of the proposed system, this system is performed with two versions: one is the proposed version and another is using only kNN alone for clothes searchtion process. Using these two versions, the performance is measured with processing time. The reason of measuring processing time is to prove that the processing time of using two algorithms (K-prototype and kNN) is faster than the processing time of using only kNN. The processing time differences between kNN and k-groups were tested with 10 tests, and the results are compared in Table-1. The average time for each group is also compared. On 10 tests, the average time of only kNN is 2.722 and the average time of the clusters 3, 4 and 5 are 2.372, 2.192 and 2.002.

Table 1. Comparison of processing time on traditional kNN and different K-groups of proposed method

Processing Time (s.ms)				
Test No	Only kNN	K-Prototype + kNN		
		Cluster K=3	Cluster K=4	Cluster K=5
Test-1	0.285	0.227	0.221	0.198
Test-2	0.286	0.234	0.228	0.200
Test-3	0.260	0.244	0.223	0.196
Test-4	0.254	0.236	0.224	0.185
Test-5	0.259	0.246	0.219	0.210
Test-6	0.265	0.237	0.225	0.205
Test-7	0.267	0.242	0.218	0.208
Test-8	0.3	0.23	0.199	0.201
Test-9	0.278	0.238	0.221	0.195
Test-10	0.268	0.238	0.214	0.204
Average	**2.722**	**2.372**	**2.192**	**2.002**

Different processing times are also compared with line graph.

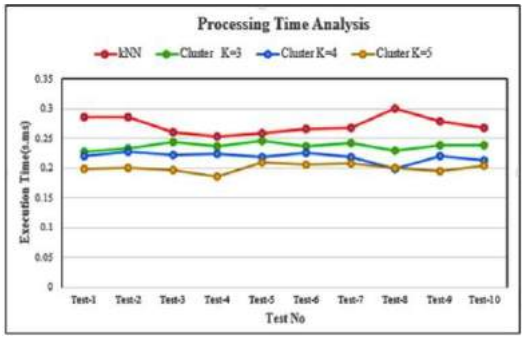


Figure 8. Comparison of different processing time based on 10 tests.

In addition, on 50 tests, the average time of only kNN is 0.288 and the average time of the clusters 3, 4 and 5 are 0.237, 0.231 and 0.229. Looking at the average times, the average time of kNN is higher and the average time decreases, the cluster groups increase. The fewest average processing time is found in the use of two algorithms with 5 clusters.

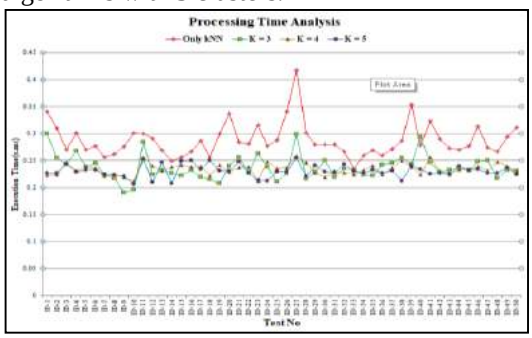


Figure 9. Comparison of different processing time based on 50 tests.

In this system, the processing time of the user searching for the desired clothes also depends on the number of clusters we have grouped. The more clusters there are, the less processing time it takes to search. However, when dividing a cluster, it is necessary to divide it based on the number of instances. The number of clusters can also be divided depending on the number of instances.

6. Conclusion

This paper has presented how data mining algorithms can help the daily clothes choosing process of customers and shop workers with their preferences from bulk of clothes selling at the counter or in the warehouse.

The proposed model with K-prototype clustering and kNN has been successfully used for mark able improved searching of clothing dataset. As a result of applying cluster based on grouping stage, kNN can be applied in the most similar cluster instead of using the whole training data. This system can save the processing time by using combination of the K-prototype and kNN algorithm rather than using traditional kNN. In addition, using K-prototype algorithm also saves preprocessing time. The limitations of the system is that it is more efficient when data set are large. So, if data set is small, it has a few issues when splitting data into clusters. In the future, a lot of further training data and testing data will be used and presented. The accuracy or error rate of the system will be calculated.

References

[1] A.Choudhary, P.Sharma and M.Singh, "Improving K-means through better initialization and normalization.", 2016 Int Conf Adv Comput Commun Informatics, ICACCI 2016;2415–9.

[2] A.G.Karegowda, M.A. Jayaram, A.S. Manjunath, "Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients", International Journal of Engineering and Advanced Technology (IJEAT); ISSN: 2249 – 8958, Volume-1, Issue-3, 2012

[3] A.S.Eesa, W.K.Arabo, "A Normalization Methods for Backpropagation: A Comparative Study.", Sci J University, Zakho. 5(4):319, 2017.

[4] B. Kim, "A Fast K-Prototype Algorithms Using Partial Distance Computation". Korea University, Seoul 02841, Korea, 2017

[5] E.Golinko, T.Sonderman, and X.Zhu, "CNFL: Categorical to Numerical Feature Learning for Clustering and Classification," in Data Science in Cyberspace (DSC), 2017 IEEE Second International Conference on, 2017.

[6] J. Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, Burlington, MA, USA, 2012

[7] K.Arun Prabha, Karthi Keyani Visalakshi, "Praticle Swarm Optimization based K-Prototype Clustering Algorithm", IOSR Journal of Computer Engineering, India; Apr.2015

[8] K.S.Ranti, K.Salim and A.S.Girsang, " Clustering steam user behavior data using KPrototypes algorithm", Bina Nusantara University, Jakarta, Indonesia 11480, doi:10.1088/1757-899X/725/1/ 012105, 2020

[9] M.M.Mon and Y.Y.Win, "Clothes Recommendation System by using Bisecting K-means", UCSM, March-2017.

[10] Patro SGK, sahu KK. "Normalization: A Preprocessing Stage.", Iarjset, 2015

[11] P. W. Buana, S. Jannet D.R.M, I. K. G.D. Putra, "Combination of K-Nearest Neighbor and K-Means based on Term re-weighting for Classify Indonesian News", Department of Information Technology Udayana University, Bali, Indonesia; 11 July 2012.

[12] R.Madhuri, M.R. Murty, J.V.R. Murty, P. Reddy and S.C. Satapathy, "Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms". ISM, Jharkhand 826004, 2014.

[13] R.Ahuja, A.Solanki, A. Nayyar, "Movie Recommender System Using K-means clustering and K-Nearest Neighbor", Gautam Buddha University, IEEE 2019

[14] W.Xindong, V. Kumar, J.R.Quinlan, "Top 10 Algorithms in Data Mining", Knowledge and Information Systems 14 (1): 1–37, 2008.

[15] Y.Zhang, X.Liu, T.Shi, Y.Guo, C .Xu, E. Zhang, J. Tang, and Z. Fang, "Fashion Evaluation Method for Clothing", *https://doi.org/ 10.1155/2017/8093057,* 2017.

[16] Z.Huang, "Clustering large data sets with mixed numeric and categorical values.", In Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, Singapore , pp. 21–34, 1997.

[17] https://www.edureka.co/blog/k-nearest-neighbors-algorithm.

[18] Dataset Source: https://archive.ics.uci. edu/ ml/ datasets/dresses_attribute_sales.

Distributed and Parallel Computing

Attribute Level Locking to Improve Data Availability in a Distributed System

Swe Zin Aung, Khaing
University of Computer Studies, Yangon
swezinaungaugust15@gmail.com, khaingaugust@ucsy.mm

Abstract

Distributed systems consist of a number of computers connected over a network in which numerous users are using such systems locally or via the Internet to meet their requirements. As the system is multiple users' environment, the lock control is an essential part to control the concurrent accessing on the same data item. This system presents a lockable unit on the attributes of a record. So, the database records increased in accessibility because of the attributes are locked instead of the entire row. This technique may allow several transactions to access the database row simultaneously by utilizing some attributes and keeping others available for other transactions. This system aims to implement employee management system with the lockable unit by using three phase locking (which contain read lock, write intent lock and write lock) to avoid dead lock and to control concurrent processing.

Key words: ***three phase locking, attribute, distributed system***

1. Introduction

Distributed system (DS) may be defined as a collection of multiple, logically interrelated computer network. Distributed system is a crucial source of information for numerous users who access the database via the network for different tasks. To meet the professional requirements of users, data must be available at all times, because data availability plays a major role in the success of information systems. During transaction execution, the lock manager locks the required database items by sending a message to the central site. If the requested lock is granted, then the lock manager sends a message to the requested site; otherwise, it waits. In case of a write operation, the lock manager must lock all copies of the requested database item in all sites where it exists, but in a read operation, the transaction is executed at any available site.

2. Related Work

The related works of concurrency controls are discussed in this session.

In distributed system [1], concurrency control is a challenge issue to distributed network. In airline reservation system, users may access data concurrently and they must get consistent data. The system embedded locking technique provides consistent data for concurrent users. It is the system that can be accessed concurrently because three sites are parallel running. This air-line reservation system uses Basic two-phase locking technique. Allow user to buy or enquiry for any tickets from any site. Because each site has its one own primary database and other two copy databases. If a transaction owns read lock on the data item, the other transaction can't get read or write lock on that data item and wait until the update operation finish.

In this system [2], the concurrency control in distributed databases is an important problem. This system is intended to build concurrency control system by using OSN (ordering by serialization number) method that uses time interval technique with short-term locks and that OSN scheduler is located on the server site. By using the OSN method, this system can control the concurrent execution of transactions as well as the consistency of database can be maintained. And also it can avoid the deadlock by using short-term locks with time interval. In this system, room selling for condominium is used as a case study.

3. Background Theory

The problem of data availability and the degree of concurrent transactions have been discussed by several researchers [2, 3, 8,] who concentrated on a strategy of dividing the database into variable size units. The size of such units is dynamically managed by the lock manager based on user needs and competition. This competition

increases more in a distributed database system than in a centralized one because of the higher number of users.

3.1. Lockable Unit

In the study of locking techniques, the size of the lockable units clearly has a major effect on the concurrency control and the availability of data, because while the database unit is locked, it will be unavailable for a time. Thus, if the locked unit is a table, then no other transaction can access that table in a conflict mode until the lock is released.

This system presents an approach for increasing the data availability by suggesting the attribute as a new lockable database unit. This technique may be implemented by increasing the database tables' attributes as lockable units instead of the entire row. The proposed approach may allow several transactions to access the same database row simultaneously, which may increase the degree of concurrency and the availability of data. (i.e. while a transaction holds the data item by read lock or write intent lock , the other transaction can get read lock on the same data item without blocking.)

3.2. Attribute Level Locking

This approach aims to include the attributes that would be the new lockable units for allowing several transactions to access the same database row concurrently. This approach may increase the database resources, which would increase the concurrency and throughput in the system and decrease deadlock occurrences. In the suggested attributes as new lockable units; the attributes may be locked individually when a transaction requires only some attributes of the database row. This locking can be performed as explained in follows.

1) The database row is locked in an intent exclusive mode (IX).
2) The key of that row is locked in a shared mode (S).
3) The required attributes can be locked by the requested transaction in read or write.

4. Three Phase Locking (3PL)

Three-phase locking protocol is a non-blocking protocol because it includes a pre-commit phase (write-intent locking phase) to prevent the blocking state. This phase is reached if all transaction participants have voted to commit. Otherwise, and if this state is not reached, the participant will be aborted and the blocked resources will be released. In the write-intent mode, the processing on the data are virtual (because of the processing is in the pre-commit phase.) Dependent on the nature of the data and the environment: changing data, many users. Types of locks used: Read, Write, Write Intent.

4.1. Benefits of 3PL

All locks are subject to timeouts, with appropriate actions (unlocks, error/warning messages to user etc) taken in the event of lock failure. This prevents deadlocks. No other protocol should be applied or considered as it may disrupt the flow of operations.

4.2. Three Phase Locking Mechanism

Num-of-sites (M) = Number of sites in the system

Num-of-DB = Number of databases in each site →1

OP-Mode = Operation mode → R, RW, W

R = Read mode → Read Lock

RW = Read / Write mode → Write-intent Lock

W = Write mode →Write Lock

Phase 1: Read mode → Read Lock

- User sign-in and request data from the database.
- Read lock is applied to the data.
- Once the data is read, the Read lock is dropped.

(There can be multiple Read locks by multiple users on the same data/tuple.)

Phase 2: Read / Write mode → Write-intent Lock

- When the user indicates that he wishes to edit the data, take out a WRITE-INTENT lock.
- UPDATE: Write Intent Lock also known as Change Lock or Protect Lock.
- Other users can still obtain a Read Lock on the data.
- Write Lock allowed only to the user who has the Write Intent Lock.
- If data locked with Write Intent Lock, then no further Write Intent Locks can be applied on it.

Phase 3: Write mode →Write Lock

- When user finishes editing the data and submits the changes, immediately Write Lock is applied on the data.
- Write Intent Lock is unlocked.
- Transaction is committed and Write Lock is unlocked.

4.3. Three Phase Locking Algorithm

```
BEGIN
Accept the lock request from user.
If (Transaction.OP-Mode = = R)
{
    Request (ReadLock);
    Release (ReadLock);
}
Else If (Transaction.OP-Mode = = RW)
{
    Check the requested item is already held by
    Write-intent Lock;

    If (RequestedItem.CurrentLock = = Write-intent
      Lock And Write Lock)
     {
       Deny (Transaction.OP-Mode);
       Message “Grant for Read Only Mode”;
       }
Else If (RequestedItem.CurrentLock = = Read
    Lock OR Null)
  {
    Grant (Transaction.OP-Mode);
  }
}
Else If (Transaction.OP-Mode = = Write)
{
    Check the requested user has Write-intent
     Lock;

    If (Has Write-intent Lock)
    {
        Grant (Transaction.OP-Mode);
        Commit (Transaction.OP-Mode);
     }
    Else
    {
        Deny (Transaction.OP-Mode);
        Message “Grant for Read Only Mode”;
     }
    End If
}
End If
END
```

Figure1. Three Phase Locking Algorithm

4.4. Control Table of Transactions By Three Phase Locking

	Read-Lock	Write-Intent-Lock	Write-Lock
Read-Lock	Grant	Grant	Deny
Write-Intent-Lock	Grant	Deny	Deny
Write-Lock	Deny	Deny	Deny

Table1: Lock Control States for Three Phases

Tuple	Transaction A	Transaction B	Lockable status
Tuple 1	Read-Lock	Read-Lock	Ok
Tuple 1	Write-Intent-Lock	Read-Lock	Ok
	Write-Intent-Lock	Write-Intent-Lock	Deny
Tuple 1	Write-Lock	Read-Lock	Deny
	Write-Lock	Write-Intent-Lock	Deny
	Write-Lock	Write-Lock	Deny

Table2: Lock Control for Each Transaction

5. Implementation of the System

This system presents an implementation of employee management system. In this system, each user level can perform the respective authorize tasks from each department of the system. So, multiple transactions may occur for the management process and may lead to concurrency problem on a specific code of data. Although there may be concurrent transaction processing, this system can control the concurrent transaction by using three phase locking instead of blocking methods with the combination of attribute level locking.

This system proposes a new approach for increasing the data availability by suggesting the attribute as a new lockable database unit. This technique may be implemented by increasing the database hierarchy tree by one more level down to include the attributes as lockable units instead of the entire row. The proposed approach may allow several transactions to access the same database row simultaneously, which may increase the degree of concurrency and the availability of data.

This system uses three-phase locking instead of two-phase locking protocol. Three-phase locking protocol has a pre-commit phase to prevent the blocking state. To simplify the implementation, a central locking approach is considered, which means there is one site that has a lock manager and must coordinate with other sites in the system. Locking can be granted on some attributes of a row, including the key of that row if no conflicts among transactions could occur. The detail step of the processing flow is shown in the following figure 2.

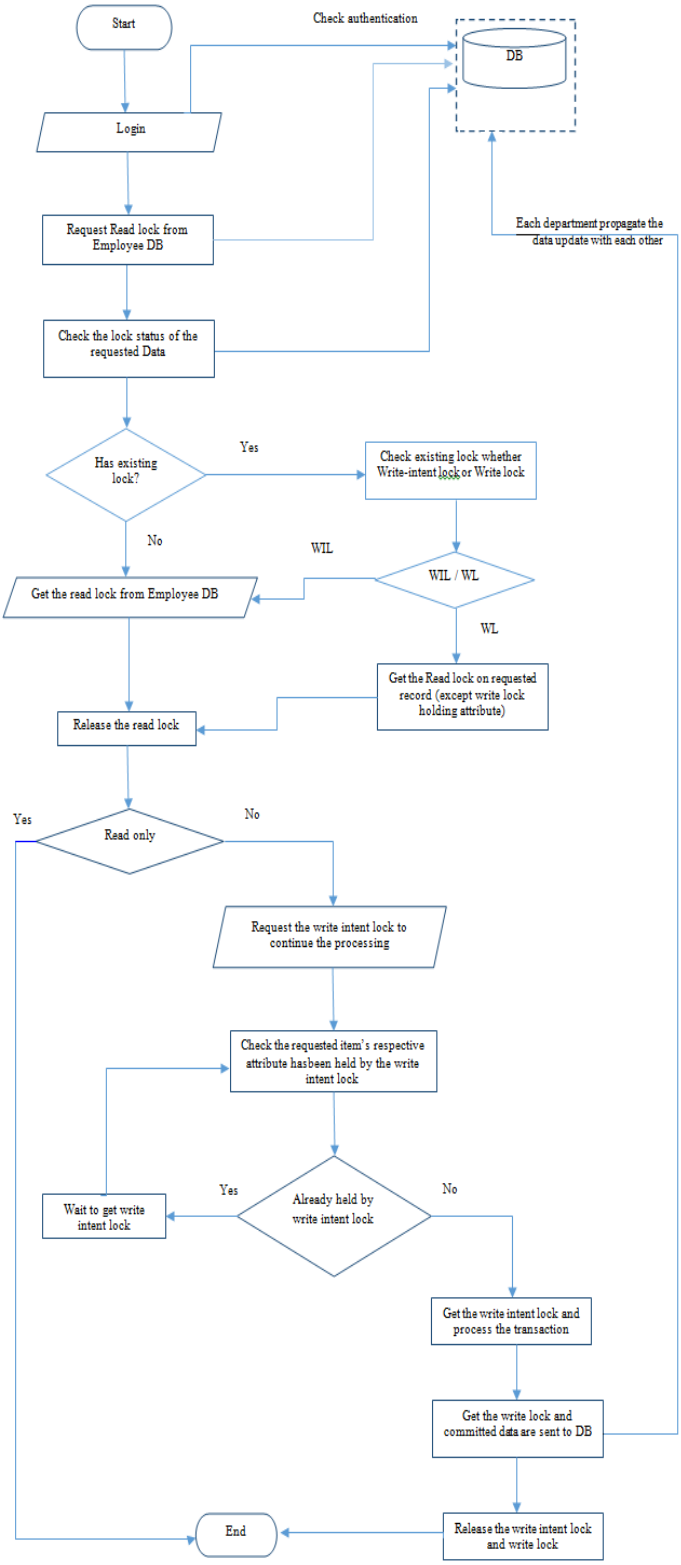


Figure 2: The System Flow

6. Conclusion

Distributed systems are considered crucial sources of information. Therefore, data contained in

such systems must be available at all times as much as possible to satisfy the user's professional needs. To increase the data availability, this paper proposes a new adaptive approach to increase the database items by reducing the size of the lockable units. This reduction can be carried out by locking the attributes instead of the database row, which remains as the other attributes become available for other transactions. The attribute-level locking increases the degree of concurrency by increasing the data availability. The overall system performance is also improved because the average waiting time is decreased. The increasing overhead is managed by returning the lock at the row level when a transaction requires many attributes of the same row. The proposed approach is suitable with short transactions of mixed read and writes operations, especially when the degree of replication is less than 50%.

REFERENCES

[1] Bernstein P. and Newcomer E., ―Principles of Transaction Processing for the Systems Professional‖, 2nd ed. Morgan Kaufmann Publisher, 2009.

[2] Chandy K., Misra J. and Hass L., ―Distributed Deadlock detection‖, ACM Transactions on Computer Systems, Vol. 1, No. 2, 1983.

[3] Croker, A. ―Improvements in Database Concurrency Control with Locking‖, Journal of Management Information Systems; Vol. 4 Issue 2, 2001.

[4] Elmasri R. and Navathe S. ―Fundamentals of Database Systems‖, Pearson Addison Wesley, 7th edition, 2015.

[5] Gray J. and Reuter A., ―Transaction Processing: Concepts and Techniques‖, San Francisco, Calif.:Kaufmann, 2011.

[6] Jose M. Faleiro and Daniel J. Abadi. ―FIT: A Distributed Database Performance Tradeoff‖, IEEE Data Engineering Bulletin, 38(1): 10-17, 2015.

[7] Kjetil Norvag, Olav Sandsta, and Kjell Bratbergsengen, ―Concurrency Control in Distributed Object-Oriented Database Systems‖, Advances in Databases and Information Systems, 1997.

[8] Maabreh K. and Hamami A., ―Increasing database concurrency control based on attribute level locking‖, on the proceedings of International Conference on Electronic Design, ICED, IEEE, pp1-4, Issue 1-3, Malaysia, Penang. Dec. 2008.

[9] Maabreh K. and Hamami A., ―Implementing New Approach for Enhancing Performance and Throughput in a Distributed Database‖, The International Arab Journal of Information Technology, Vol. 10, No. 3, May 2013.

[10] Matthias N. and Matthias J., ―Performance Modeling of Distributed and Replicated Databases‖, IEEE transactions on knowledge data engineering, Vol.12 No.4, pp 645-672, July 2000.

[11] Muhammad Atif, ―Analysis and Verification of Two-Phase Commit &Three-Phase Commit Protocols‖, International Conference on Emerging Technologies (ICET), pp:326-331, Islamabad,19-20 Oct. 2009.

[12] Ozsu T. and Valduriez P., ―Principles of distributed database systems‖, Springer science and business, 3rd edition, New York, 2011.

[13] Silberschatz A., Korth H. and Sudarshan S. ―Database System Concepts‖, McGraw-Hill, New York, 6th edition, 2010.

Mobile Learning System for Evaluating English Study Course Based On AHP Technique

Theint Wut Yi Phyo, Thin Lai Lai Thein
University of Computer Studies, Yangon
theintwutyiphyo@gmail.com, tllthein@ucsy.edu.mm

Abstract

Technology deployment in distance education has indicated it's remarkable in the transfer of knowledge for both the instructors and the learners. This is also done practicable through the use of the Internet which helps modify the traditional teaching approaches into more modern methods. Many mobile applications are developed and developers create the educational mobile applications. Most evaluation system for mobile applications is voting the level. But some users are confused to evaluate the mobile application. In this situation, Decision Support System (DSS) is a particular class of computerized information system that provides business and organizational decision making enterprises. There are two main parts in this system, the first is creating educational mobile application by student and the second part used AHP to evaluate and decision for this application. Creating mobile application is intended the secondary student who need to practice the English subject for reading, listening and grammar. The outcomes of this study can provide users, researchers, and practitioners in the education sector to allow the necessary resources and make plans to reduce the challenges and facilitate the effective use of mobile English language learning in educational practices.

Key words: mobile application, Decision Support System (DSS), AHP

1. Introduction

English language is extensive across the world and is usually used in many areas as the primary language for communication. Additionally, many learners around the world have established learning and using the English language in desire to their mother tongue. Most countries have also realized the importance of the English language in education and have founded deficiencies by making English language learning a key factor in their planning and strategy.

Computers and other tools are suitable in supporting learners studying English as second language. Instructors need to work with technology to improve learners' performance. Mobile applications are the up-to-date technical developments to support in English language learning.

Decision Support Systems (DSS) are a type of information system whose principal objective is to provide a human decision maker during the process of accessing at a decision. DSS is a particular class of computerized information system that provides business and organizational decision making enterprises. At times, it is very hard to make good decision. In preparation, decision maker commonly uses the past experiences to create a decision. These past experiences can realize a form of performing tests to derive to a correct decision. Opportunely, the developments of computer technologies and automatic learning techniques can create this easier and more efficient. Evaluation for mobile application system is the difficult task and the right decision form user's feedback is importance for product. This system is the alternatives correctness for many specific criteria, and the weights of the criteria are generally expressed in linguistic terms. The technique for AHP method is one of the familiar classical MCDM methods.

The organization of paper is as follows. Section 1 introduces the English language for mobile application and important of feedback from user by using this application. In section 2, the related works are presented. Section 3 presents the

methods for AHP model for English Study. Section 4 describes the design of the proposed system. Section 5 describes experiments and results. The last section is section 6, and there is conclusion.

2. Related Work

In this section, the previous works related to mobile English learning course and evaluation of the subjects, are reviewed.

M. Gokhale, [1] have proposed the procedure of AHP as a decision making tool for ranking the Strategic Plan's objectives and goals. It also illustrates the suitability of AHP for use in group settings where individual judgments can be aggregated. The method still depends on particular judgments for the analysis. The weights that are allocated to the decision makers can be tilted in favor such that the result will return the opinion of a select few people.

In [2], F. Colace et al. have presented "Evaluation models for e-learning platforms and the AHP approach: a case study". It used the estimated multi criteria problem as a decision hierarchy to be resolved using the Analytic Hierarchy Process (AHP). These approaches, in fact, are suitable in circumstances which require the consideration of different courses of action, which can't be estimated by the measurement of a simple, single dimension. In this way, this research can estimate an E-Learning platform allowing for both its application in the interest scenario.

N. Hlawn Moe et al. [3] have described "A Decision Support System for Supporting Academic Admission Using Analytical Hierarchical Process". This system uses the combination of AHP and Forward Chaining Method that generate better decision making with high accuracy. There are three limitations in this research. The first one is the universities and criteria are limited to be used three to five. The second one is that it should be more flexible than this for real application. The user can't save their choice for further reference or comparison for the last limitation.

3. Overview of Multi Criteria Decision Making (MCDM)

The MCDM model, by pairing theory and knowledge, supports an analytical approach to show consultation and is improved for a variety of technology and many other fields aiming at suitability assessments. The range of MCDM is simple way of selecting the adviser or a university to complex engineering application, evaluating political candidacies [8].

A decision support system (DSS) is basically a computer system that assists you make a decision by leveraging the multiple criteria decision making (MCDM) model. Decision theory methods used by high-end knowledge professionals have been effectively used for contrasting expert judgments and making educated ranges. Decision trees, artificial neural network, Bayesian learning, AHP and FAHP are also approaches of decision making techniques.

3.1. Analytical Hierarchy Process (AHP)

AHP, a tool developed by Thomas Saaty (1994) in the 1970s, uses the human ability to create sound judgments about small problems where there is relatively little aggregation of different factors to be considered. Grandzol (2005) mentions "Desirable characteristics of such an approach involve simplicity, usefulness for both individuals and groups, accommodative of intuition, compromise, and consensus building, and without prejudice toward specialized skills or knowledge" [5].

3.1.1. Basic Characteristics

AHP is a tool which integrates the quantitative and qualitative analysis. It is seizing up the problem into small sub-problems. This is realized by developing various criteria and sub-criteria which can be used to contrast the different solutions to a problem. These criteria and sub-criteria are setup in a hierarchical scheme so that they are easier to follow and compare at a lower level. The differentiations can be implemented by using significant numbers having ratio properties.

The ratios can be used to make weights or priorities that give back the relative importance of the decision criterions. The differentiations can be prepared against an absolute scale or against one another. This differentiation is managed by the expert judges or by using the available statistical data. This is where the qualitative characteristics of the process come into play.

3.1.2. Consistency Check

The key characteristic of the AHP process is consistency check of the judgments or differentiations. There is a probability that the users may be unresolved or done low judgments during the states since the evaluation process can be debilitating. These unessential checks involve computation of consistency ratios (CR).

3.1.3. Three Modes of AHP

The important effect is 'rank reversal'. To solve this effect, AHP models have three different modes: distributive and ideal modes in the relative measurement (pairwise differentiations) attitudes and an absolute measurement attitude. The distributive mode is suitable in cases where there is attention in gaining the degree of difference among the alternatives. In the distributive mode, the local priorities at any level of the hierarchy sum to one. "The ideal mode is used to obtain the single best alternative irrespective of what other alternatives there are".

3.2. Steps of AHP

Step1: Define the criteria (i=1, 2,…, m) and alternatives (j=1, 2, …., n).
Step2: Determine their relative priority P_i with respect to the objective and for each criteria i. In performing pairwise comparisons.
Step3: Differentiate the j=1, 2,…, n alternatives and determine their relative priority p_{ij} with respect to criteria i.
Step4: The system determines the final alternative's priority P_j with respect to all the criteria.
Step5: The alternatives are ranked by R_j. The most preferred alternative is the one having the largest R_j.

3.3. Applications of AHP

Table 1. Goal, Criteria and Alternatives

Goal	Criteria	Alternatives
Which session is the best and worst	**Easy to Understand Effective Convenience Flexible**	**Reading Writing Listening**

Table 1 expresses the Goal, Criteria and Alternative for this application. Table 2 also describes the definition of pairwise comparison scales for AHP model.

Table 2. Pairwise Comparison Scales for AHP Model

Value	Definition	Explanations
1	**Same important**	**Two activities assign equally to objective 1.**
3	**Slightly more important**	**Experience and judgment slightly favor one activity over another.**
5	**More important**	**Experience and judgment strongly favor one activity over another.**
7	**Lot more important**	**An activity is strongly favored, and dominance is demonstrated in practice.**
9	**Totally dominates**	**The evidence favoring one activity over another is the highest possible order of affirmation.**

4. Proposed System

The proposed system is deployed on an Android operating system which is an open source technology.

There are two main parts in this system. The first one is the creating the mobile application for English Study course and the second one is to

evaluate this mobile application using AHP model. Figure 1 shows the step by step of mobile application. There are three steps in admin side. The first one is to check the valid of the administrator. After validation, admin need to choose which session they want to update the lessons. After choosing the lesson, the administrator can insert, delete and update the lessons. The updated lessons are stored the lesson database and seeing the lessons are already updated from student side.

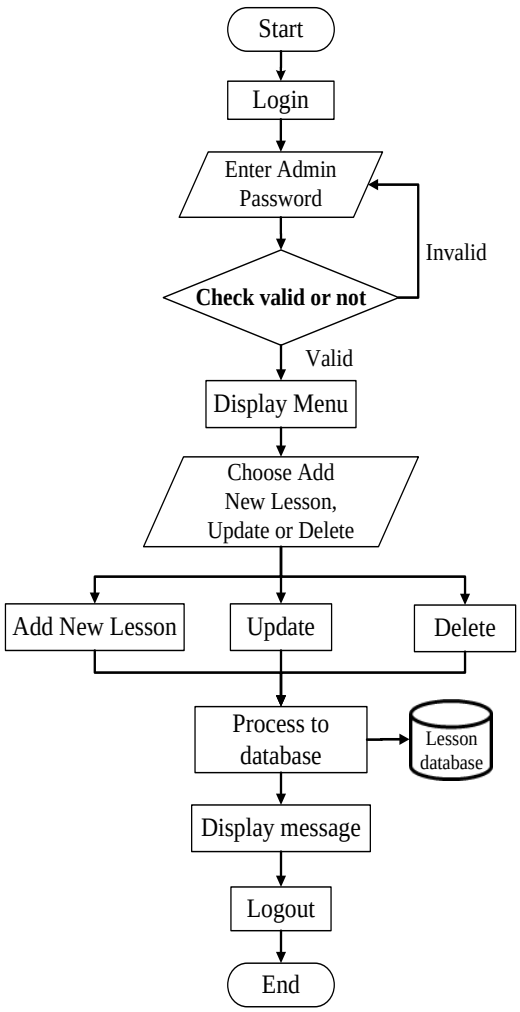


Fig.1 System Flow Diagram (Administrator Side)

Figure 2 shows the student side and AHP model calculation process**.** If the student is old user, it only needs to log in and if not it needs to sign in. After that the student chooses the session to practice and do the exercise. The last step is the important step to evaluate this mobile application. Most of the users can't decide the good decision for evaluating the application. To get the good decision, the user needs to choose their preference on the alternatives and criteria. According to the AHP model, these sessions will be displayed based on percentage**.**

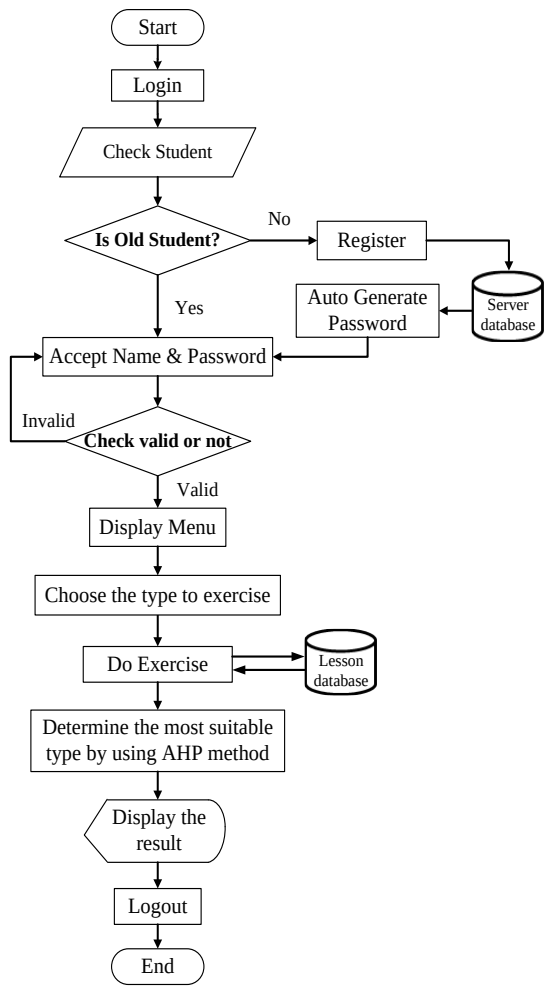


Fig.2 System Flow Diagram (Student Side)

In our proposed system, there are four criteria (Easy to Understand, Effective, Convenience and Flexible) and three alternatives (Reading, Listening and Grammar) to make which session is the best decision to evaluate the mobile application. Table 3 shows the student preference for three sessions in this English learning mobile application.

Table 3. Analysis result for three sessions

Type	Priorities
Reading	54.44
Listening	23.29
Grammar	22.16

5. Experiments and Results

To use this proposed system, users need to own android phone with the operation system version 4.0 at least and to install English Learning application. By starting the application, the user needs to choose admin or student for login. The administrator can update, delete and add lessons for reading, listening and grammar sections as shown in Figure 3. Although delete and update are simply way to access the data, add lessons can use two ways to upload the data, the first way is to upload the pdf version and the second way is to type or copy and paste the paragraph as shown in Figure 4.

Fig.3 Update Lessons from Admin Side

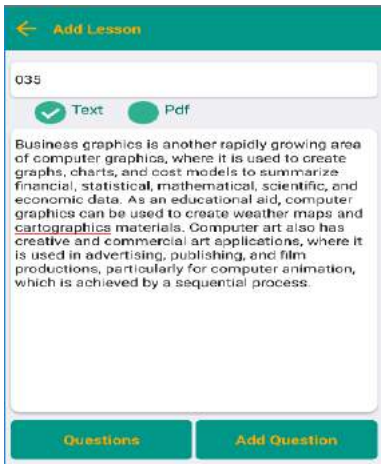


Fig.4 Two Views from Admin Side

From student side, student can login or sign in. In Figure 5, students can test their ability for their desired session and desired the lessons. After complete the test, the important part of this application is to evaluate the application. In this part, AHP model is used to support the decision. But, the users need to choose their preferences for criteria and alternatives. The final result will be displayed as % in Figure 6.

To update the application, the analysis for all the user's evaluation results will be displayed in Figure 7. So, the admin can update the weakness session.

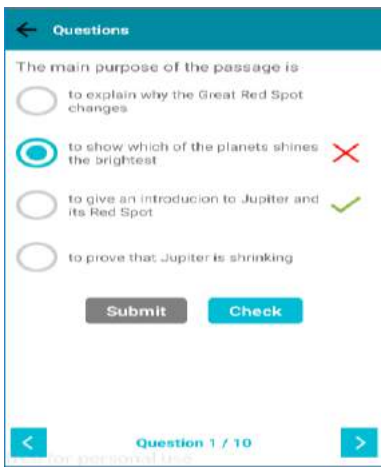


Fig.5 Answer the Reading Test

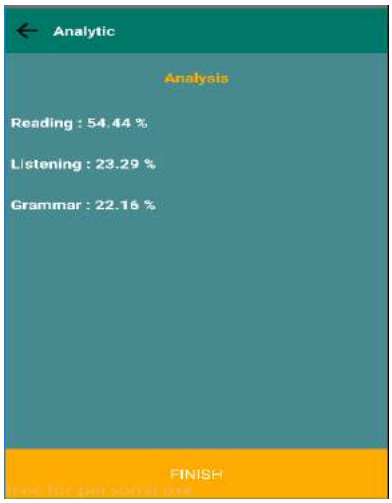


Fig.6 Decision Making

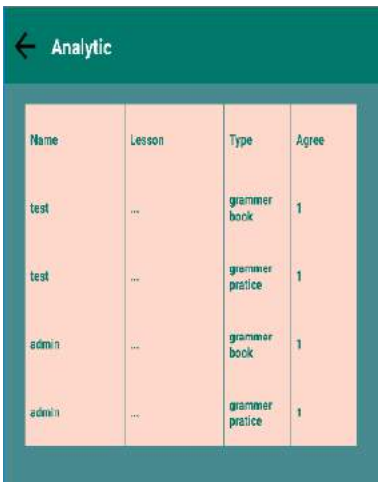


Fig.7 Analysis Result

By testing 100 students, the best session is reading and the worst is grammar. So, according to this analysis, the admin side needs to update the grammar session.

6. Conclusion

This system determined the decision support system for evaluation the English study mobile application using Analytic Hierarchy Process. The analytic hierarchy process has presented the concern of many research papers and practically useful; there are critics of the system. It's a beneficial supplement to other subjective and objective techniques that calculates product quality. The user can recognize how easy and powerful the technique is when used to display the quality. Moreover, the user can practice the technique in other fields.

References

[1] C.Ipsze Ting, "A Fuzzy Analytical Hierarchy Process Evaluation of Hotel Websites", Master Thesis, The Hong Kong Polytechnic University, Jan 2011.

[2] Ivan Pogarčić, Miro Frančić, Vlatka Davidović "The analytic hierarchy process for the decision tree with multiple criteria", Study of Information Systems, Polytechnic of Rijeka, Trpimirova.

[3] Chengjing Jounio, "Supplier Selection Based On AHP Method", Bachelor of Business Administration Helsinki Metropolia University of Applied Sciences.

[4] H. BROOVÁCzech ,"Application of AHP Method in Traffic Planning", Master thesis, University of Agriculture, Prague, Czech Republic.

[6[Lwin Htun Naing,, "Decision Support System for Phone Handset Selection Based on Analytic Hierarchy Process," University of Computer Studies, Yangon, May 2009.

http://scholarsmine.mst.edu/masters_theses/4608

Effective Music Distribution System for Online Music Industry

Su Latt Sandi and Twe Ta Oo
University of Computer Studies, Yangon, Myanmar
sulattsandi@ucsy.edu.mm

Abstract

Music industry has already been moving into online space successfully these days. The profit of this industry largely depends on the effective control of unauthorized access to music. This paper presents an audio scrambling method in wavelet domain to satisfy main requirements of online music industry: support of teaser music, access control, and acceptable security in music distribution. The proposed system deploys the different influences of audio wavelet coefficients on perceptibility of human ear. Scrambling on different audio wavelet layers produces different-quality audio files. Low- or medium-quality files can serve as teasers for potential buyers to taste the songs, whereas severely-degraded files can provide high-level access control for music distribution. Without the knowledge of scrambling keys, anyone can never recover the original song quality. Experimental results show that the proposed system is effective, fast, simple to implement, and applicable for online music industry.

Keywords: audio scrambling, wavelet, online music industry, digital rights management.

1. Introduction

Before online music stores are popular, people had to go to local music shops to buy song albums. They also had to wait for international music to be available on local shops and it might also be very expensive.

With online music stores, global users can now have easier and cheaper access to any desired music, no need to care about geographical locations. In addition, online stores allow users to choose the songs they wish instead of having to purchase an entire album in which there may be only one or two titles the buyer enjoys.

As for distributors, online distribution has allowed for potentially lower expenses such as lower coordination and distribution costs. Recently in Myanmar, music distributors are more using online not only to advertise the new songs but also to distribute songs to buyers. For example, Legacy Music Network [9] advertises their songs by sharing the chorus as teaser to potential buyers to guess the taste of the whole music.

Along with the advantages, online distribution has also brought some challenges such as loss of profits due to illegal downloads [2] [7]. If the distributor uploads the whole song as teaser, there are potentially high risk of illegal downloads. Even for sharing the chorus as sample, illegal downloading can still occur for using as ring tunes. Thus, the profit of online music industry largely depends on controlling unauthorized access to music and its future is in danger [14].

In order to avoid illegal download, some distributors are now using full audio files with degraded quality as teaser. For example, JOOX music application [5], which is a freemium service, provides most of its songs free with degraded quality. However, high-fidelity songs are only available for premium users and offered via paid subscriptions or by doing some requested tasks such as watching the advertisements.

No need to doubt, cryptographic algorithms can be used to degrade the audio quality, at the same time to control unauthorized access to music [16]. However, traditional cryptographic algorithms are not satisfactory for music distribution because of their nature of blind encryption. In this paper, an audio scrambling method which is more appropriate for online distribution than traditional encryption is proposed.

The rest of the paper is organized as follows. Section 2 discusses the key differences between audio encryption and scrambling and highlights the fact that scrambling is more appropriate for music distribution. Section 3 then presents a brief literature review on audio scrambling methods and section 4 explains the proposed system in detail. Finally, section 5 and 6 discuss the experimental results and conclusion, respectively.

2. Audio Scrambling vs. Encryption

Cryptographic algorithms, e.g. AES, have been widely used for data protection including audio. Encrypted audio will become like noise and without knowing the decryption key, anyone can never recover the full-quality song. However, this feature alone is not satisfactory for online music distribution. The primary reason is that audio media is an instance of plaintext which has specific patterns of coherence that can be judiciously exploited [18]. Based on the audio format, e.g. uncompressed (WAVE & AIFF) and compressed (MP3 & AAC) formats, each audio file must follow the predefined multimedia standard [6]. Blind encryption on audio data will break the standard and as a result, encrypted audio may not be played back by standard media players. Thus, encryption can be used to degrade the audio quality; however, that quality-degraded song cannot be used as teaser. Fortunately, audio scrambling methods can overcome that problem by obeying the media standard even after scrambling.

Audio scrambling is similar to but not a direct application of usual cryptographic techniques such as AES. It aims to minimize residual intelligibility of an audio signal with the use of a certain secret key. Unlike encryption that uses complex mathematical operations like substitution, permutation, transformation, shifting, etc., most scrambling methods try to permute only the order of audio samples without changing the values [1] [3]. Permutation order is mostly controlled by a secret key. Even if they modify the audio samples, those changes conform to the media standard.

A good scrambling method should meet the following requirements [18].

Security: Even if the algorithm is distributed in public, descrambling should still be difficult.
Efficiency: Scrambling and descrambling processes should not be too complex.
Perceptual quality: Scrambled signal should achieve very low residual intelligibility; descrambling process must also be able to recover an audio signal with nearly original quality.
Media player: Both the scrambled and descrambled audio should be playable with standard media player.

This paper proposes an audio scrambling method that meets the above requirements.

3. Related Work

This section presents a brief literature review on the previously proposed audio/speech scrambling methods.

Yan, Fu, and Kankanhalli [18] proposed a scrambling scheme for protecting the MP3 files. The basic idea was to apply multiple rounds of XOR to the Huffman codewords as per pre-defined key table. Milosevic, Delic, and Senk [17] also presented an audio scrambling algorithm that used Hadamard matrices. Both of the methods change the position and values of the audio samples. Hence, they are very effective for reducing the perceptual quality of the music.

Poblete [15] introduced the possibility of real-time use of discrete wavelet transform with pure data analysis and re-synthesis of audio signals. It was confirmed that the wavelet transform could be a powerful and interesting tool for audio processing.

Zhou and Au [12] proposed two approaches of composing the keystream to be used for audio scrambling. The first one uses dynamic password generator and the other uses pseudo-random number generator. It was said that complexity of generating the keystream using dynamic password generator is higher than the pseudo-random number generator.

This paper also presents an audio scrambling method in uncompressed domain. The proposed work is different from others in terms of low computational complexity, fast execution time in key table generation, and direct control on progressive audio quality.

4. The Proposed System

One of the main aims of the proposed system is to provide different-quality audio files, teaser or high-fidelity music, useful for online music stores. To achieve this aim, this system makes use of the effect of audio wavelet layers on human auditory system.

Discrete Wavelet Transform (DWT) can effectively be used to analyze the temporal and spectral properties of non-stationary signals like audio. When an audio signal is analyzed in DWT, it is first decomposed into detail and approximation coefficients by using an analysis filter. The approximations correspond to low-frequency components of an audio signal, while the details are high-frequency components. Normal human ear is usually most sensitive in the low-frequency range [4]. Thus, the approximation coefficients are more important for perceptibility. Our experiments also confirm that there is drastic damage on audio quality even for slight changes on approximation coefficients. Based on this nature of DWT, the proposed audio scrambling method is applied on detail coefficients to generate teaser music and applied on approximations to provide very low or zero-quality music.

Generalized process flow is shown in Fig. 1 in which raw pulse code modulation (PCM) signal is first input to the DWT decomposition process. For i-level decomposition, there are i+1 layers of coefficients $\{d_1, d_2, ..., d_i, a_i\}$, where d is details and a is approximations. For a PCM signal with n samples, the maximum number of wavelet levels that can be decomposed is $\log_2(n)$. Then, the proposed scrambling method is applied on each wavelet layers by using different keys from a pre-generated key table. After the synthesis process (IDWT) with scrambled coefficients, quality degraded audio files are obtained.

Generalized process flow of descrambling is also the same as Fig. 1, except that the audio scrambling process is replaced with descrambling process.

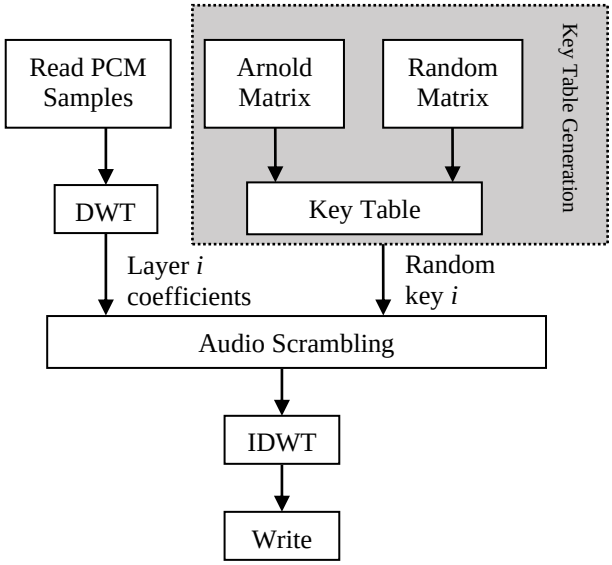


Fig. 1: Generalized process flow (scrambling process)

4.1. Key Table Generation

Security of the proposed system mainly depends on the secrecy, uniqueness, and randomness of the keys used in the scrambling/descrambling processes. In this system, a key table is pre-generated based on the Arnold and random matrices. Different keys are then randomly chosen from that key table for scrambling each wavelet layer. Thus, the larger the size of the key table, the more randomness in selection of the layer keys and thus the stronger security the system achieves.

In our experiments, we chose the key table size as 8 x 8 as an example, even though there is no restriction on key table size in our system. Firstly, an 8 x 8 Arnold matrix (A) is generated by concatenating the rows and columns of the matrix shown in eq. 1 [13] [18].

$$A(p \times q) = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \quad (1)$$

where p x q is the size of the matrix. An 8 x 8 random matrix (R) is then generated by using eq. 2 in Matlab. In

this process, we carefully set the seed used to create the repeatable arrays of random numbers by saving and restoring the generator settings [8]. Without knowing the seed, we cannot get the same random number and then we cannot derive the key needed to descramble the audio signal to recover its original quality.

$$R = randi\ (imax,\ n), \tag{2}$$

where *i*max is the maximum random number to be generated and *n* is the size of the matrix. Finally, the key table (*K*) is generated as defined in eq. 3, where *N* is the size of the key table.

$$K = (A \times R)\ mod\ N. \tag{3}$$

To recap, security offered by the proposed system is controlled by two steps: firstly without knowing the seed number and secondly without knowing the random key used to scramble each wavelet layer, anyone can never recover the original song quality.

4.2. Audio Scrambling Process

Flowchart of the proposed scrambling process is shown in Fig. 2 and detailed procedure is as follow.

Step 1: Read the DWT coefficients of the selected layer.

Step 2: Choose the key from the pre-generated key table (*K*) by randomly indexing the row and column of *K*.

$$x = randi([1, p], 1), \tag{4}$$

$$y = randi([1, q], 1), \tag{5}$$

where *p* x *q* is the size of the key table and *x* and *y* are the row and column index of the key table.

Step 3: Perform XOR on the selected layer coefficients with the selected key.

$$C' = C \otimes K(x, y), \tag{6}$$

where *C* and *C′* are the coefficients before and after scrambling, $\otimes$ is the bit-wise XOR, and *K*(*x*, *y*) is the random key selected. The above steps must be repeated for all the layers needed to be scrambled.

Step 4: Perform IDWT to construct the scrambled audio.

Step 5: After IDWT, the scrambled audio file is saved in .wav format.

In this system, scrambled audio with desired quality level can be generated by choosing the wavelet layer to be scrambled. The more the scrambled layers are, the worse the audio quality will be. This feature is really attractive for online music stores in which low-quality songs are needed as teaser and high-quality songs are needed for paid users. In addition, conforming to the media standard is also one of the main aims of the proposed system. It was achieved in this system by using simple mathematical operation like XOR. The resulting scrambled audio clips can be easily playable by standard media players.

4.3. Audio Descrambling Process

Descrambling process is exactly the same as the scrambling process shown in Fig. 2. To regain the original audio quality, the above process in Fig. 2 must be applied again on the scrambled contents by using the same keys. Only if all the keys are correct, full quality can be recovered. Moreover, quality of the descrambled audio depends on the number of layers descrambled. If *n* layers of the audio were scrambled, all *n* layers must be descrambled to retain full quality.

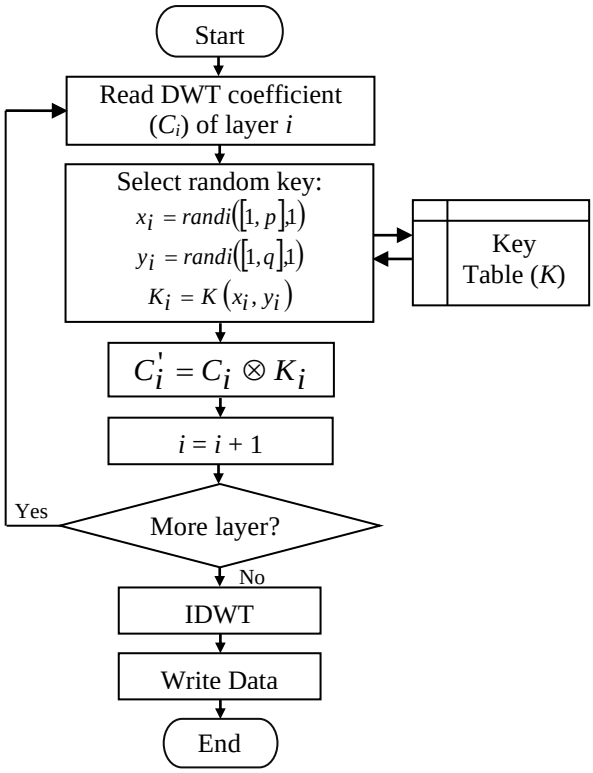


Fig. 2: Flowchart of the proposed scrambling process

5. Performance Evaluation

The proposed system is implemented in Matlab and this section evaluates the performance of the proposed system by using 25 audio files. Each file is coded in 16 bits per sample, 44.1 kHz sampling rate, and .wav format. Table 1 shows the music pieces used in the experiments, which are grouped based on genre.

For the following experiments, each audio clip is decomposed into 5 layers of coefficients {*d*1, *d*2, …, *d*4, *a*4}, i.e. 4 wavelet levels. Each layer is scrambled by using different keys. Evaluations are then done in terms of (i) the effect of scrambled wavelet layer on audio quality, (ii) progressive audio quality, (iii) execution time, and (iv) the effect of scrambling on file size.

Table 1. Music pieces for experiments

Song	Genre	Avg. Length (sec)	Avg. Size (MB)
S1-5	Classical	23	1.62
S6-10	Pop	23	1.95
S11-15	EDM	22.8	1.95
S16-20	Rock	26.2	2.22
S20-25	Jazz	25.4	2.24

5.1. Effect of Scrambled Wavelet Layer on Audio Quality

Section 4 stated that the proposed system provides different-quality audio files based on the wavelet layer scrambled. It is confirmed here by evaluating the audio

quality after scrambling each layer by means of the signal-to-noise ratio (SNR). The SNR is the objective measurement of expressing the audio quality. It is defined as the ratio of the signal power to the noise power, see eq. 7. As per the International Federation of Phonographic Industry (IFPI) [11], an audio clip with SNR ≥ 20 dB is considered to be good quality.

$$SNR = 10log\left(\frac{P_{signal}}{P_{noise}}\right)(dB), \quad (7)$$

where P_{signal} indicates the audio signal power and P_{noise} is the noise power [10].

Table 2 shows the average SNR after scrambling each wavelet layer. The results confirm that scrambling on different layer results different SNR, which means different audio quality. According to IFPI, SNR ≥ 20 dB means good audio quality. Thus, the results in Table 2 also show that scrambling on *d1* to *d3* layers produces good and acceptable quality music, whereas scrambling on *d4* and *a4* yields bad and poor quality music.

Moreover, Table 2 also shows that the SNR values are getting dropped from *d1* to *a4*. It is because *d1* is the highest in frequency among wavelet layers and *a4* is the lowest. As previously stated, human ear is more sensitive in the low frequency range. Thus, scrambling on *a4* yields the lowest audio quality.

Considering the proposed system is to be used in music distribution, the number of layers to be scrambled can also be chosen based on the system requirement. If the distributor requires high-level access control, all layers should be scrambled. Otherwise, scrambling only the approximation layer is sufficient for degrading quality. If the distributor wants to share a new song as preview, scrambling only the detail layers is sufficient.

Table 2. Average SNRs after scrambling each layer

Song	Genre	Scrambled Layer				
		d1	***d2***	***d3***	***d4***	***a4***
S1-5	Classical	28.84	24.15	20.41	17.53	7.80
S6-10	Pop	26.55	22.39	19.78	17.70	7.77
S11-15	EDM	26.38	23.35	21.69	20.09	7.82
S16-20	Rock	24.02	19.80	16.88	15.31	8.70
S21-25	Jazz	31.36	26.97	23.29	19.27	7.29
Average		**27.43**	**23.33**	**20.41**	**17.98**	**7.88**

5.2. Evaluation on Progressive Audio Quality

This section proves that the perceptible audio quality also depends on the number of wavelet layers scrambled. Table 3 and 4 show the SNR values after scrambling and descrambling, respectively. It is seen from Table 3 that the SNR values are decreasing after scrambling layer after layer. Similarly, if we descramble more layers, the recovered audio quality is getting better. It verifies that the proposed system achieves the progressive audio quality.

The last column of Table 4 also shows that the *a4* layer plays a vital role in audio reconstruction. In addition, after descrambling all layers, the average SNR values go high up to approx. 242 dB. It proves that the proposed system can perfectly recover the full audio quality after descrambling.

Table 3. Average SNR results after scrambling

Song	Genre	Scrambled Layer				
		d1	***d1, d2***	***d1 to d3***	***d1 to d4***	**All**
S1-5	Classical	28.84	22.87	18.44	14.89	6.87
S6-10	Pop	26.55	20.98	17.30	14.46	6.90
S11-15	EDM	26.38	21.47	18.49	16.11	7.10
S16-20	Rock	24.02	18.39	14.55	11.87	6.82
S21-25	Jazz	31.36	25.61	21.24	17.11	6.83
Average		**27.43**	**14.89**	**6.90**	**14.89**	**6.90**

Table 4. Average SNR results after descrambling

Song	Genre	Descrambled Layer				
		d1	***d1, d2***	***d1 to d3***	***d1 to d4***	**All**
S1-5	Classical	6.91	7.01	7.28	7.80	241.73
S6-10	Pop	6.95	7.08	7.33	7.77	242.03
S11-15	EDM	7.16	7.27	7.44	7.71	241.42
S16-20	Rock	6.92	7.18	7.75	8.70	242.31
S21-25	Jazz	6.81	6.93	7.04	7.35	242.08
Average		**6.95**	**7.10**	**7.37**	**7.87**	**241.91**

5.3. Effect on Execution Time and File Size

As per the experimental results shown in Table 5, the original and scrambled/descrambled audio clips are exactly the same in file size, no size blow-up. As for the average execution time, for a 24.1 sec long audio clip, the proposed system takes 0.64 sec for scrambling and 0.30 sec for descrambling. From these results, we can see that the proposed system is very low in computational complexity and suitable to be used for online music stores.

Table 5. Results of average execution time and file size

Song	Duration (sec)/Size (MB)	Scrambled		Descrambled	
		Duration (sec)	**Size (MB)**	**Duration (sec)**	**Size (MB)**
S1-5	23/1.62	0.82	1.62	0.29	1.62
S6-10	23/1.95	1.00	1.95	0.29	1.95
S11-15	22.8/1.95	0.49	1.95	0.29	1.95
S16-20	26.2/2.22	0.47	2.22	0.32	2.22
S21-25	25.4/2.24	0.43	2.24	0.32	2.24
Avg.	**24.1/2.00**	**0.64**	**2.00**	**0.30**	**2.00**

5.4. Proposed Application Scenario

The proposed system can be applied for online digital music services such as JOOX application that is currently popular among young consumers in Myanmar. First, an audio clip is 5-level wavelet decomposed and all layers are scrambled using different keys. That file is used for secure music distribution to subscribers. Based on the subscription type, users can get 2, 4, or 5 keys for

descrambling, as depicted in Table 6. For only 2 keys, the application will only allow users to descramble *a4* and *d4* layers and enjoy low-quality music. For VIP subscribers, they can get all 5 keys to recover high-fidelity music.

Table 6. Proposed application scenario

Quality Level	**Low**	**Medium**	**High**
Layer to be descrambled	*a4 & d4*	*a4, d4, d3 & d2*	*a4, d4 to d1*
Number of keys required	2	4	5

6. Conclusion

This paper presented an effective music distribution system for online digital music industry. The system was developed based on a DWT-based low-complexity audio scrambling method. Experimental results showed that the proposed system is fast and easy to implement. It can also provide flexible control on music quality by controlling the number of scrambled layers. In addition, the proposed method can perfectly recover the original audio quality and work well for all music genres. Furthermore, the proposed system can also support sufficient security via the control of descrambling keys.

If the security aspect of the proposed system is expected to be more enhanced, more secure key generation algorithms could be applied instead of the one based on Arnold matrix. As stated in eq. 1, the Arnold matrix used in this proposed system supports some duplicated values. Thus, randomness of the keys generated based on the Arnold matrix may be weak. However, as of now, the security provided by this proposed system is satisfactory for applications like online music distribution services.

References

[1] A. Srinivasan and P. A. Selvan, "A review of analog audio scrambling methods for residual intelligibility," Innovative Systems Design and Engineering, vol. 3, no. 7, 2012.

[2] D. Tabibzada, "How does illegally downloading music impact the music industry?," Nov, 2015.

[3] G. Dhanya and J. Jayakumari, "Permutation based speech scrambling for next generation mobile communication," International Journal of Engineering and Technology, vol. 8, no. 2, pg. 707-713, Apr-May, 2016.

[4] https://en.wikipedia.org/wiki/Hearing_range

[5] https://en.wikipedia.org/wiki/JOOX

[6] https://soundbridge.io/audio-formats-file-types

[7] https://www.marshallmusic.co.za/2017/04/05/ effects-illegal-downloading-music-industry

[8] https://www.mathworks.com/help/matlab/math/ generate-random-numbers-that-are-repeatable.html

[9] http://www.myanmarmusicstore.com/Default.aspx

[10] http://www.onmyphd.com/?p=snr.signal.noise.ratio

[11] H. Yi and C. L. Philipos, "Evaluation of objective measures for speech enhancement," Interspeech2006, pg. 1447-1450, Sept, 2006.

[12] J. Zhou and O. C. Au, "Security and efficiency analysis of progressive audio scrambling in compressed domain," Innovation and Technology Commission of the Hong Kong Special Administrative Region, China, pg. 1802-1805, 2010.

[13] M. Lin, T. Liang, and Y. He, "Arnold transform based image scrambling method," Third International Conference on Multimedia Technology, pg. 1309-1316, 2013.

[14] Priya, "Effects of piracy on music industry," Apr, 2019.

[15] R. D. Poblete, "Manipulation of audio in the wavelet domain processing a wavelet stream using PD," Institute of Electronic Music, 2006.

[16] S. Y. Salunkhe and A. R. Nigavekar, "MP3 music file protection using digital rights management and symmetric ciphering," International Journal of Engineering Research and Applications, vol. 3, issue. 2, pg. 1774-1777, Mar-Apr, 2013.

[17] V. Milosevic, V. Delic, and V. Senk, "Hadamard transform application in speech scrambling," Proceedings of 13th International Conference on Digital Signal Processing, pg. 361–364, 1997.

[18] W. Q. Yan, W. G. Fu, and M. S. Kankanhalli, "Progressive audio scrambling in compressed domain," IEEE Transactions On Multimedia, Mar, 2008.

Distributed Multi-Servers Instant Messaging System

Su Myat Hlaing, Khine Moe Nwe
University of Computer Studies, Yangon
sumyathlaing30@gmail.com, khinemoenwe@ucsy.edu.mm

Abstract

Most existing instant messaging system is built on Peer-to-peer (P2P) framework or "centralized server" architecture, which is designed by using Client/Server (C/S) structure. There is a problem about the instant messaging system due to centralization. If the local area network (LAN) network or proxy server restrict instant messaging services, or when outside the LAN connection is disconnected, the user cannot communicate with each other even in the same LAN. To resolve of this issue, this framework expects to execute a multi-server model of appropriated texting framework. This framework able to not only manage the user's important information, but also improve the quality and efficiency of communication between the users.

Keywords: ***instant messaging, P2P, Client/Server, local area network (LAN)***

1. Introduction

Instant Messaging(IM) is a constant intuitive Internet application, moment fast advancement of correspondences are significantly changing the manner in which individuals impart, team up and play. Texting framework means to sort out the trading of huge virtual networks, where clients can have a typical language, ready to share the (Friends), on-line (Online) companions can impart continuously message, voice, video and records, and so on. Next, the client can take a gander at whether their companions are on the web or not. On the off chance that the client's companions are on the web, they can send or get messages to one another, occasionally to discover online clients, pick one solicitation add to companions. On the off chance that the client's companions are disconnected, they can send to a companion mail, a companion can move records to one another. Related work is described in section 2 , Background Theory is in section3, reliable instant messaging services details provided by the proposed system is in section 4. Implementation detail is explained in section 5.

2. Related Work

The related works of concurrency controls are discussed in this session.

In Communication over Internet with Instant Messaging system [1], utilizes IM works as a book based PC gathering between at least two individuals. An IM correspondence administration makes a sort of private talk stay with another person so as to impart continuously over the web. This framework depends on customer server design. In the customer server model [5], assets and information security are controlled through the server. All information is put away in the server for sponsorship up records and looking through documents and information effectively. The customer’s server arranges model has a few restrictions. Since this model depends on a unified server, the whole framework can't work if the server goes down because of any sort of issue.

A Peer-To-Peer based chat system [2], portrays one potential approach to actualize a visit framework utilizing a Peer-To-Peer model rather than a

customer/server model. Shared correspondence instruments are utilized so as to escape from the disadvantages a customer/server model experiences. This empowers friends to discuss legitimately with one another as opposed to having a server in the center which all correspondence goes through. Gnutella is an early P2P correspondence component and utilizations flooding to discover wanted information among associated hubs [6]. At the point when information is found inside the Gnutella arrange, the information is sent straightforwardly to the mentioning hub however Inconvenient to oversee.

Architecture and Implementation of Instant Messaging in Educational Institution [3], this framework plans to deliver building structure and usage of texting (IM) which can be applied to instructive foundations as a quick, fortunate, and viable correspondence implies. This framework can encourage the correspondence from an instructive establishment to or among understudies and instructors, and assemble understudies and speakers' telephone number and email. Web Server is utilized for execution of web application and can impart by means of XMPP convention with visit server and MySQL database. The design and IM application are XMPP convention put together that can be actualized with respect to web stage with individual visit, bunch talk; communicate messages, booked messages, and record connections highlights. Difficulties to provide high quality service for all users due to centralization.

3. Background Theory

Instant messaging (IM) technology is a type of online chat that offers real-time text transmission over the Internet. A LAN messenger operates in a similar way over a local area network. Short messages are typically transmitted between two parties, when each user chooses to complete a thought and select "send".

3.1. IM Communication Process in Centralized Server Mode

Presently the correspondence mode most texting framework is basically based on "unified server." Between the client of this method of correspondence, must be first sign on to a remote server or a concentrated server ranches, build up an association, correspondence is quick. Its network model is as shown in Figure1.

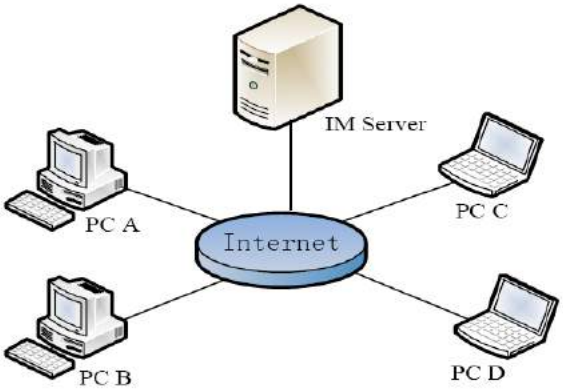


Figure1. Model of centralized server IM

Assuming a single-server IM is shown in figure 1 to establish a buddy relationship between the user A, B, C, and D. In a single-server mode communication process between users as follows:
Initially assumed that users are offline, the user A logged into the local server, and informs its IM server online, the local server record online user A's. When between user A and user C under the same server need to communicate, IM server advertised, IM buddy list server detected, notify the user A, user C's current online status. In this case, if the user is offline C, IM, server feedback to the user A, user C current offline, therefore, when the user A can only be in a wait state. If the user C already is online, then IM

server notifies the user A and user C both in-line and can communicate.

At the point when a client signs on the server, his IP address and port number of the report arranged for the future correspondence to the IM server. At the point when the new client signs on, the servers is liable for the other client status warning new login client, and inform all their online companions of IP locations and correspondence ports. Along these lines, by one another's IP delivers and correspondence ports to build up P2P association between companions, at that point you can highlight point between the two texting. In the correspondence procedure, it is expected between client A and client C can appropriately set up TCP/IP associations, texts can be transmitted through conventional Sockets interconnected between them. In any case, if there is a firewall blocking or different reasons can't build up typical TCP/IP association in both the correspondence procedure, the texts between the two still should be sent through the server.

The benefit of this method of correspondence is advantageous to the framework server upkeep and redesigns. Be that as it may, its weaknesses are self-evident: a first, the system speed isn't sufficiently quick and generally appropriated clients, this model expanded the weight on the server, it is hard to give top notch administration to all clients. Second, the "focal" server framework on the server equipment made popularity, the expense of procurement and development of such servers will in general be over the top expensive, and upkeep cost is likewise costly.

So as to facilitate the weight on the server, there are currently "center" IM server ranches. This uses various servers to finish a solitary server capacities, and group information for every client on different servers. This IM server ranch somewhat decreases the server load autonomous. In any case, the "unified" type server model pervasive an issue: If the LAN firewall or intermediary server confinements texting administrations, or when outside the LAN association is broken, even inside a similar neighborhood clients can't speak with one another.

4. Proposed Distributed Multi-server IM Model

This system aims to solve the "central" server mode problems in a multi-server distributed instant communication model. The system model is as shown in Figure 2. In this system model, the local server I provide services within users A and B, to provide services within users C, D, etc. is local server II and the server III (remote server) provide services within different LANs.

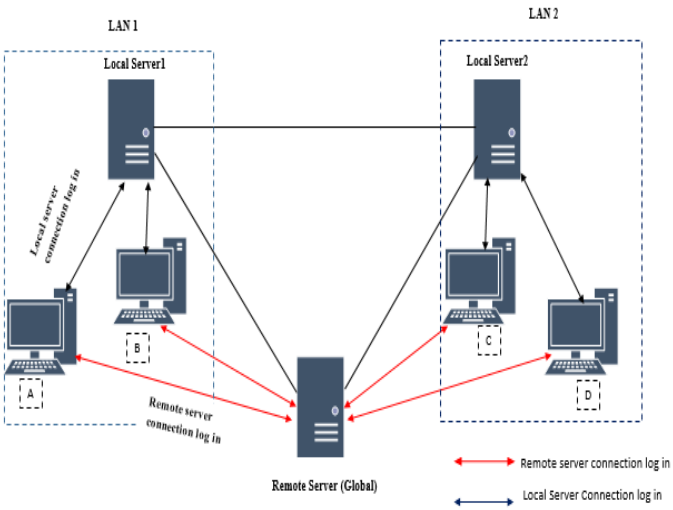


Figure2: Model of distributed multi-sever IM

Scenario I:

Assume that User A want to communicate User B within same LAN as shown in Figure 3. Firstly User A need to login to local server I and notify its online status. Local server I record its status and check LAN1 user list and their status. As User B is also in LAN1, User A need to search User B in LAN1 User List. If user B is also online, they can communicate each other without passing through

external network. But if user B is Offline, User A can send messages to User B as waiting state. Similarity, if pair of user in LAN2 wants to communicate, same scenario as above.

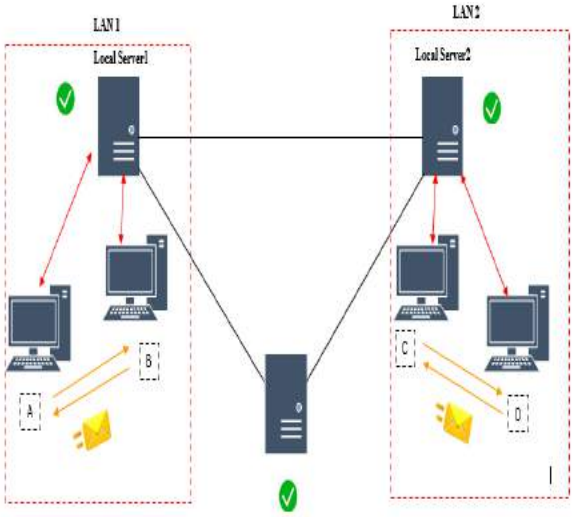


Figure3. Communication in Same LAN

Scenario II:

Assume that User A in LAN1 want to communicate User C in LAN2. User A must exit from LAN I login and re-login to remote (global) server as shown in figure 4. Remote (global) server records User A's Online status and check LAN1 and LAN2 user lists and their status. As User C is in LAN2, User A needs to search User C in LAN2 User List. If user C is also online, remote (global) server notifies the user A and user C both in-line and they can communicate each other by using remote (global) server. Similarity, if user from LAN2 wants to communicate user from LAN1, same scenario as above.

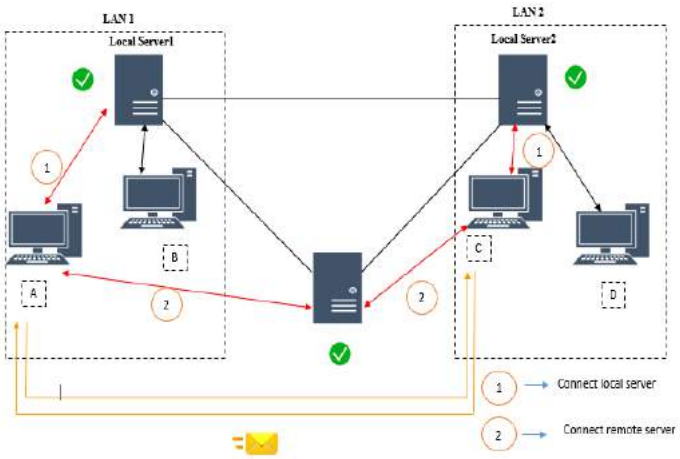


Figure4: Communication in Different LAN

Scenario III

If same LAN users want to communicate each other but local server connection is failure, can connect to remote (global) server and communicate. When local server works again, remote (global) server synchronizes data to local server. So, users can see message history at next time login by using local server.

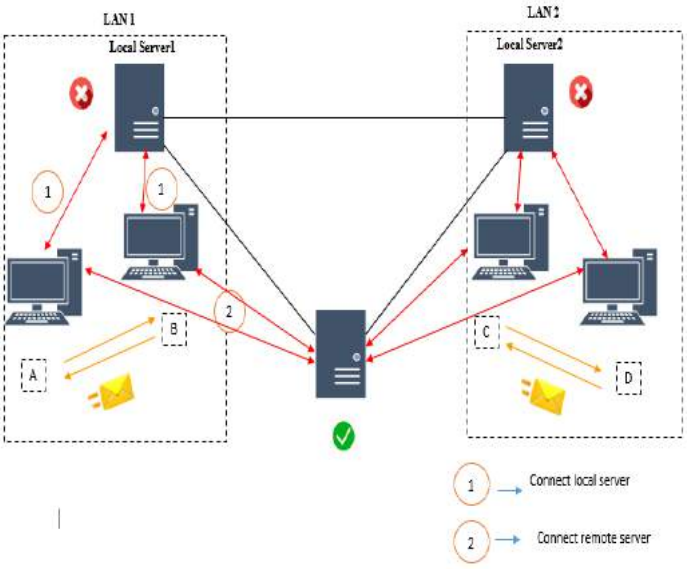


Figure5: Communication Same LAN by Remote (Global) Server

5. Implementation of the System

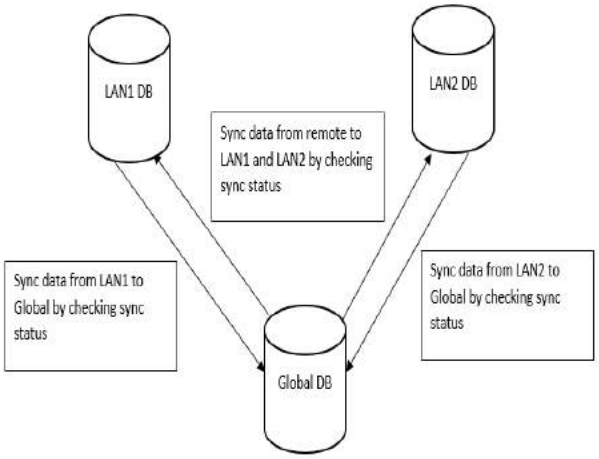


Figure6. Synchronization Process

This system aims to implement a distributed instant messaging system for a private organization. In this organization, assume Head Office (LAN1) is in location I. Branch Office (LAN2) is in Location II. Remote server is in Location III. Users in Branch Office use LAN2 server to connect each other user within same Office. Users in Head Office use LAN1 server to connect each other under Head Office. To connect users in different offices, they must use Remote (Global) server. User connects to server and use UserID to login instant messaging server and also use UserID to send or receive message. Detail processing steps are shown in the following figure 7.

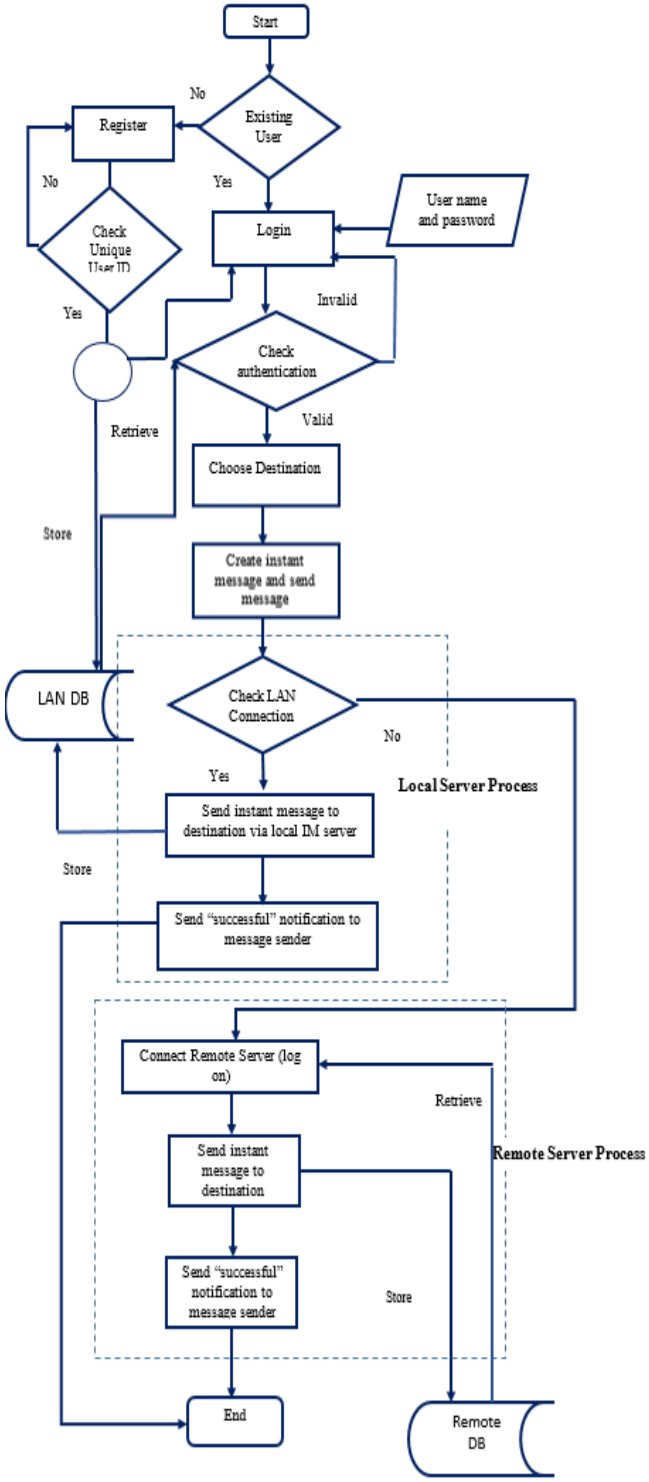


Figure7. The System Flow

6. Conclusion

This framework proposes texting framework model on circulated multi-server as a quick and viable correspondence implies. This model gives dependable texting than the unified engineering. This framework structures an appropriated multi-server texting framework. This system designs a distributed multi-server instant messaging system, which ensures reliability between different LAN users because there is global server to continue instant messaging even if local servers are disconnected and bidirectional synchronization for backup purposes in the event of failure or maintenance of an active LAN databases and for making faster access to the system from different locations.

REFERENCES

[1] Mohammad Milon, “Communication over Internet with Instant Messaging”, Helsinki Metropolia University of Applied Sciences Bachelor of Engineering Information Technology Thesis, 25 April 2012.
[2] Tommy Mattsson “A Peer-To-Peer based chat system”, 2012.
[3] Beverly Yang, Hector Garcia-MolinaBudi Yulianto, Eileen Heriyanni, Lusiana Citra Dewi, and Timothy Yudi Adinugroho, “Architecture and Implementation of Instant Messaging in Educational Institution”, Bina Nusantara University, Jl. K.H. Syahdan No. 9, Jakarta 11480, Indonesia
[4] Hao Zhang, Zhongkui Sun, He Li, “A Distributed Instant Messaging System Model ”, 2015
[5] Raymond B. Jennings III, Erich M. Nahum, David P. Olshefski, Debanjan Saha, Zon-Yin Shae, and Chris Waters, “A Study of Internet Instant Messaging and Chat Protocols”, IBM T.J. Watson Research Center, 2006.
[6] Shen Zhiwei ,Ma Shaowu, “Analysis and Design of Instant Message System Based on P2P”, China Netcom Group Labs Beijing 100032

Implementation of Travel Scheduler System by using Genetic Algorithm

Ms.Su Thitsar Hlwan Moe Thu, Daw Myint Myint Yee
University of Computer Studies, Yangon
suthitsarhlwanmoethu@ucsy.edu.mm, myintmyintyee@ucsy.edu.mm

Abstract

Traveling can expand people's views and thoughts. People believe it possible to relieve stress and purify soul by traveling. The independent trip can choose both travel date and travel route. Preparation, however, is not easy and tedious as there are certain considerations that need to be concerned such as plan, travel path, transportation and accommodation. People had always had trouble deciding where they should go on vacation and they would spend a lot of bow during the vacation. Therefore, an effective system of travel advisory is required. A good travel scheduling system can meet the time and budget constraints of the traveler for the entire trip. This paper's main purpose is to set up a travel advisory system that will allow the traveler to plan the entire trip so that they can visit many destinations in less time. The idea would use the Artificial Intelligence Genetics Algorithm to find the nearest optimal travel direction.

Key words: *Trip Advisory, Genetic Algorithm, Scheduling, Optimal Solution, Artificial Intelligent*

1. Introduction

Due to significant changes in information technology, the Internet has greatly influenced the travel services industry. Recently, the Internet has made many services available online, and in many sectors, products have appeared. Internet popularization has resulted in extensive travel information, enabling travelers to use the Internet to easily access reliable and accurate travel information and arrange travel schedules or itineraries within a limited period of time [1]. To help travelers plan their travels, many information such as timetables, routes, accommodations and restaurants are easily available.

How to prepare the most appropriate travel schedule, though, taking into account many factors, such as visiting tourist attractions, local hotels selecting, and travel budget calculation is a challenge. Well planning ahead is the main task to do in order to have an enjoyable trip. A good travel plan not only enables a visitor to gain maximum enjoyment during the trip, but also to meet their needs within time and budget constraints. One significant issue is that they only recommended the most popular travel routes or projects and are unable to prepare the schedule of travel. In addition, existing travel planning systems have limits in their ability to adapt to changes based on the needs of users and the results of planning. This study applies genetic algorithms to plan travel routes.

This paper is organized as follows. Section 2 reviews previous studies on trip scheduling. Section 3 presents the outline of the Genetic Algorithm and Genetic Algorithm implementations and focuses in section 4 on the trip scheduling system using the Genetic Algorithm. Finally, section 5 concludes and summarize the study.

2. Related Work

There are several unique or individual conditions in the scheduling of travel schedules. Travel recommendation systems (TRSs) are useful in providing adequate travel schedules for travelers [2]. Nonetheless, most TRSs are structured to suggest popular attractions or ideal travel routes without taking into account the needs of individual travelers, such as attractions, hotels and restaurants. Nonetheless, previous

TRS attractions and route suggestions were limited to recommendations based on rules and conditions, resulting in specific recommendations which minimize recommendations 'effectiveness [6].

Shivendra Goel et al [8] presents a model for trip distribution in Delhi Urban Area using Genetic Algorithm. In this study, GA is used for the distribution of travel in all areas of Delhi Urban Area and is applied to the actual collection of generated passenger travel data and attracted passenger travel in all areas of Delhi Urban Area, which in turn provides satisfactory results that can be applied in current and future scenarios. This research analyzes and compares the results of this model with Linear programming model for trip distribution. Md. Lutful Islam et al [3] developed a Heuristic Approach For Optimizing Travel Planning Using Genetics Algorithm. The main purpose is to create a Travel Planner which will allow the customer to plan the entire tour so that in less time he can visit several locations. The idea would be implemented using Artificial Intelligence's Genetics Algorithm that would be used as a search algorithm to find the nearest optimal travel direction. Moreover, In order to reduce the running time of GA, Parallelization of Genetics Algorithm would be demonstrated using Hadoop Framework. The system can automatically plan the travel itinerary that most fits the visitor's requirements and utilize Google maps to show the planned routes on visitors' screen. ," WEB-Based Tour Planning Support System Using Genetic and Ant Colony Algorithms is developed by Sheng-Yuan Tseng et al[7].This study uses genetic algorithms and ant colony algorithms to plan travel itineraries and compares the quality of the routes that the two methods have found. In addition, a travel itinerary planning website has been set up for this study to demonstrate the practicality of the proposed algorithms.

3. Background Theory

Genetic Algorithms (GAs) are an adaptive heuristic search algorithm that focuses on evolutionary concepts of natural selection and genetics [5]. As such, they reflect a wise use of a random search that was used to solve optimization problems. Although randomized, GAs are by no means random, instead they use historical information to direct the search in the search space into the better-perfect performance. The basic techniques of the GA were designed to simulate processes that are important for evolution in natural systems, particularly those that follow the principles of "the best survival" first laid down by Charles Darwin. Since in nature, competition for scarce resources among individuals results in the most suitable individuals overwhelming the weaker ones. Like older AI systems, these do not break easily, although the inputs have changed slightly or in the presence of sufficient noise. Furthermore, in searching for a large state-space, multi-modal state-space or n-dimensional surface, a genetic algorithm can offer significant advantages over more typical optimization techniques search (Linear programming, heuristic, depth-first, breath-first, and praxis.)

GAs simulate the survival of the most fit person over consecutive generation to solve a problem. Growing generation consists of a string population of character that is identical to the chromosome we see in our DNA. Every entity in a search space represents a point and a possible solution. Humans are then forced to undergo an evolutionary process in the population.

GAs are based on an analogy with the genetic structure and behavior of the chromosome within a human population using the following foundations:

Asset and partners compete for individuals in a group.

Those most successful individuals will produce more offspring in every competition than those poorly performing individuals.

Genes from healthy individuals move throughout the population so that two good parents frequently produce better offspring than either parent. This will make each successive generation more appropriate for their climate.

3.1 Advantages of GAs

GAs have some advantages that have made them very successful. These include

• Does not include any derived information (which may not be usable for many real-world issues).

- It is quicker and more efficient than traditional approaches.
- Has outstanding parallel capabilities.
- Optimizes functions that are both continuous and discrete and multi-objective.
- Optimizes both continuous and discrete functions and multi-objective problems.
- A list of "healthy" solutions is given, not just one.
- The question that gets better over time still gets a response.
- Useful when there is a very wide search space, requiring several parameters.

3.2. Limitations of GAs

There are a few restrictions on GAs as well as any strategy. Among these are:

- GAs that are not suitable for all issues, particularly simple issues that are available for derivative information.
- The fitness value is determined twice for some problems, which can be computationally expensive.
- When stochastic, there is no guarantee of the optimality or value of the solution.
- If properly implemented initial population, the GA may not converge to the optimal solution.

3.3 Genetic Algorithms

The fundamental process of the Genetic Algorithm is discussed in this section. GA's basic processes are:

1. Fitness function
2. Selection
3. Crossover
4. Mutation

An individual is characterized by a set of parameters (variables) known as Genes. Genes are joined into a string to form a Chromosome (solution).Permutation Encoding (Real number Coding) encode every chromosome is a string of integers/real values, which represents numbers in sequence. Genetic algorithms are working with the set of potential solutions, which is called population. Each solution item (individual) is measured by fitness function. The fitness value reflects an individual's performance metric, so the algorithm can pick individuals with better genetic material to produce new populations and generations to come. The simulation of evolution allows survival of better individuals and extinction of inferior ones. The aim of Evolution is to find better people in each generation. Selection, crossover and mutation maintain the process of evolution. Such systems are called genetic operators in terms of genetic algorithms. In each generation, the selection selects superior individuals and ensures that inferior individuals become extinct. The crossover operator selects two individuals (parents) from the current population and creates a new person (child) based on genetic material from the parents. Selection and crossover operators can extend superior individuals ' good characteristics throughout the entire population. The search process will also be led to a local optimum. The mutation operator adjusts an individual's value of some genes and helps search for other parts of problem space.

4.Implementing trip advisory system using Genetic Algoritm

This system implements the advisory of trip in Myanmar famous region by using genetic algorithm. The main objective is to find the optimal paths that meet the constraints of traveller provided to the system. This system find the optimal path based of travel time between places and estimate budget of whole trip.

4.1.1 Chromosome Encoding

A set of parameters (variables) known as Genes defines an individual location. Genes are linked in a chromosome (solution) sequence. The permutation encoding (real number encoding) is used in this system. -chromosome is a string of real / integer values, representing numbers in sequence. Figure (4.1) depicts the sample chromosome.

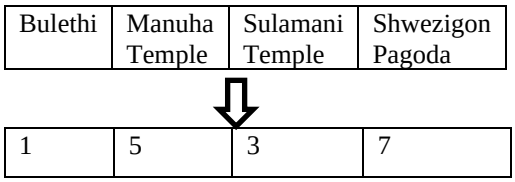


Figure (4.1) Chromosome Encoding

4.1.2 Fitness Function

The fitness value reflects a chromosome's efficiency. A chromosome with a greater fitness quality is much more likely to be selected to produce offspring. Nevertheless, the offspring produced may be an illegal remedy. Penalty methods are the most common technique used in evolutionary algorithms to prevent solutions from falling into an unfeasible area. Since the proposed chromosome encoding can create a path that violates total travel time or travel budget constraints, the following fitness function evaluates a built-in path P using the penalty function.

For each trip generated, the fitness function for entire trip, travel time and travel costs are defined in equation 4.1, equation 4.2 and equation 4.3 respectively.

$$F(i) = (C(P) - B_{max}) + (T_{max} - T(R)) \quad (4.1)$$

where,
B_{max}= maximum budget provided by visitors
T_{max}= maximum trip duration provided by visitors
C(P) = total cost
T(P) = total travelling time

$$T(P) = t_s + t_l + \sum_{i=1}^{n} t_i \quad (4.2)$$

where,

t = travel time from start point

t_l=lunch time

t_i=travel time for each point

T_{max}=maximum trip duration provided by visitors

n=number of places to visit

$$C(P) = \sum_{i=1}^{n} c_i + t_i$$

$$where C(P) \le B_{max} \quad (4.3)$$

c_i=accommodation cost per day
t_i =transportation cost per day
n=number of days to visit
B_{max}=maximum budget provided by visitors

The main objective of this system is to find the optimal travel path P that satisfied the following constraints describe in equation 4.3, equation 4.4, equation 4.5 and equation 4.6.

$$F(j) = \begin{cases} 1 \text{ if the plan contains famous places} \\ 0 \text{ otherwise} \end{cases}$$

(4.4)

$$C(P) - B_{max} = \begin{cases} 1 if C(P) - B_{max} \le 0 \\ 0 otherwise \end{cases} \quad (4.5)$$

$$T(P) - T_{max} = \begin{cases} 1 if T(P) - T_{max} \le 0 \\ 0 otherwise \end{cases}$$

$$\sum_{i=1}^{n} p_i \le n$$

$$\text{where } p_i = \begin{cases} 1 \text{ if the place is not visited} \\ 0 \text{ otherwise} \end{cases} \quad (4.6)$$

4.1.3 Reproduction

There are three stages in the reproduction process: selection, crossover, and mutation. A pair of chromosomes are chosen as the parents that will produce offspring based on roulette wheel selection. The one-point crossover will be performed on each parent trip after the selection A likelihood (the crossover rate) governs the crossover process. Nevertheless, the results obtained may be illegitimate, i.e. a produced offspring may have two identical

numbers of nodes. To overcome the illegitimacy created by the one-point crossover, the Partial-Mapped Crossover (PMX) [4] is used. Crossover points are randomly selected between 1 and max schedule length. Figure 4.2 indicates the process of the crossover. A random crossover point is chosen, copying each parent's first part to the corresponding offspring, and copying the second parts after swapping.

To change all parents, the PMX is added.

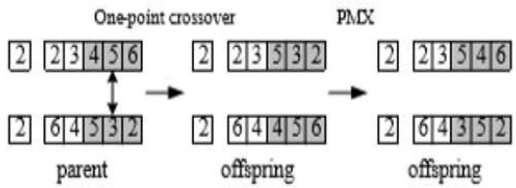


Figure (4.2) Cross over operation.

The next step is to mutate the newly created offspring by changing a portion of the new population with some low probability of P_m (the frequency of mutation). The mutation activity is shown in Figure 4.3. Two genes from the offspring are randomly selected and exchanged

Figure (4.3) Mutation Operation

4.1.4 Evolutionary Strategy

Parents are completely replaced by their offspring in a process called generational replacement in the original genetic algorithm. The technique of removing the parents completely means that in the evolutionary process some fitter individuals may be lost. In order to overcome this problem, several replacement strategies are investigated. One such strategy is (μ + λ) selection [4], in which μ parents and λ offspring compete for survival, and the μ best offspring and parents are selected as parents for the next generation. This strategy preserves the best chromosome generated so far. This study simply set λ = μ , leading to a (μ + μ) selection strategy. Let ξi represent the i-th individual of the current population and let ξ' i represent the i-th individual generated from the old population. This replacement strategy shown below retains the best individual generated to date, while retaining the diversity of the chromosome pool.

4.1.5 Experimental Result

Tests carried out on the parameters that are used in genetic algorithm, which consists of testing population size. Testing population size is used to determine the number of chromosomes in order to produce the best optimal solution in this problem. The number of population size to be tested are 10, 20, 30, 40, and 50. The average fitness values based on population size is displayed in the following figure (4.4).

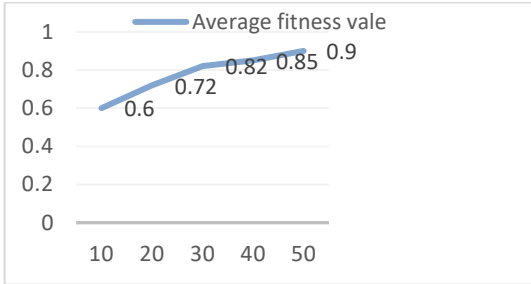


Figure (4.4) Average fitness values bsased on population size.

5.Conclusion

Travelers need advance planning in order to achieve fair tours within their expectations and constraints. This research used the Genetic Algorithm to help the traveler prepare the optimum trip. Users can easily get their suggested trip schedule after specifying their expectations and constraints. This system implements the scheduling of trips to some of Myanmar's sites. The scheduling system for the genetic base trip will help the traveler prepare their travel schedule further forward.

References:

[1] Fang-Cheng Hsu, Jiah-Shing Chen and Poren Chen, Interactive Genetic Algorithms for a Travel Itinerary Planning Problem, Proceedings of the 2nd Asia Pacific Conference on GAs and Applications, Hong Kong, May, 2000, pp.267-275.

[2] L. Console, I. Lombardi, S. Gioria, Personalized and adaptive services on board a car: an application for tourist information, J. Intell. Syst. 21 (2003) 249–284.

[3]Md. Lutful Islam, Danish Pandhare, Arshad Makhthedar, Nadeem Shaikh, "A Heuristic Approach For Optimizing Travel Planning Using Genetics Algorithm ", IJRET: International Journal of Research in Engineering and Technology , Volume: 03 Issue: 01, Jan-2014.

[4] Mehrdad Parsa, Qing Zhu, Jose Joaquin and GarciaLuna-Aceves, An Iterative Algorithm for Delay Constrained Minimum-Cost Multicasting, IEEE/ ACM Transactions on Networking, Vol.6, No.4, 1998, pp.461-474.

[5] O. Darshana,"Application of Fuzzy Logic and Genetic Algorithm in Trip Distribution ", International Journal Of Engineering Development And Research , IJEDR 2011.

[6] Y. Liu, X Huang, A. An, Personalized recommendation with adaptive mixture of Markov models, J. Am. Soc. Inf. Sci. Technol. 58 (2007) 1851–1870.

[7]Sheng-Yuan Tseng, Jen-Wen Ding, Ren-Chang Chen," WEB-Based Tour Planning Support System Using Genetic and Ant Colony Algorithms", Journal of Internet Technology Volume 11 (2010) No.7.

[8]Shivendra Goel, J.B. Singh, Ashok K. Sinha ," Trip Distribution Model for Delhi Urban Area Using Genetic Algorithm ". International Journal of Computer Engineering Science (IJCES) .Volume 2 Issue 3 ,March 2012.

Low Latency Fault Tolerance System for Distributed Application

Chu Sandy Kyaw, Sabai Phyu
University of Computer Studies, Yangon
chusandykyaw15@gmail.com, sabaiphyu72@gmail.com

Abstract

The Low Latency Fault Tolerance (LLFT) system provides fault tolerance for distributed applications within a wide-area network, using a leader-follower replication strategy. LLFT provides application-transparent replication, with strong replica consistency, for applications that involve multiple interacting processes or threads. The LLFT Messaging Protocol provides reliable, totally-ordered message delivery by employing a group multicast, where the message ordering is determined by the primary replica in the destination group. The Leader-Determined Membership Protocol provides reconfiguration and recovery when a replica becomes faulty and when a replica joins or leaves a group, where the membership of the group is determined by the primary replica. LLFT can operate in the common industrial case where there is a primary replica and one or more backup replicas. The LLFT system achieves low latency message delivery during normal operation and low latency reconfiguration and recovery when a fault occurs.

The proposed LLFT system is emphasized on the Furniture Ordering Management System. This system is implemented by using ASP.Net language and Microsoft SQL server for database engine.

Key words: ***LLFT, primary replica, backup replicas***

1. Introduction

Nowadays, commerce is larger and larger and it affects to open new branch offices in different locations. Therefore it is important to handle the order of each distributed branch. A distributed database system (DDBS) is a collection of several logically related databases which are physically distributed in different computers (otherwise called sites) over a computer network.

The Low Latency Fault Tolerance (LLFT) system provides fault tolerance for distributed applications, using a highly optimized leader-follower replication strategy, to achieve substantially lower latency and more rapid responses than existing group communication systems. LLFT provides fault tolerance for distributed applications over a wide-area network. One replica in the group is the primary, and the other replicas are the backups.

The primary multicasts messages to a destination group over a connection. The primary in the destination group orders the messages, performs the operations, produces ordering information, and supplies that ordering information to its backups. Thus, the backups can perform the same operations in the same order and obtain the same results as the primary. If the primary fails, a new primary is chosen deterministically and the new primary determines the membership of the group.

In LLFT, the processing and communication are asynchronous, but the fault detectors impose timing bounds. The assumptions of eventual reliable communication and sufficient replication enable LLFT to maintain a single consistent infinite computation, despite crash. LLFT uses the leader-follower strategy to establish a total order of messages, to establish a consistent group membership.

2. Related Work

The related works of replication are discussed in this session.

Andre Brito and Pascal Felber focused on active replication as an approach to provide fault-tolerance to Event Stream Processing (ESP) operators. More precisely, they addressed the performance costs of active replication for operators in distributed ESP applications. They used a speculation mechanism based on Software Transactional Memory (STM) to achieve the following goals: (i) enable replicas to make progress using optimistic delivery; (ii) enable early forwarding of speculative computation results; (iii) enable active replication of multi-threaded operators using transactional executions. Experimental evaluation showed that, using this combination of mechanisms, one can implement highly efficient fault-tolerant ESP operators. They have proposed new techniques to efficiently support active replication in fault-tolerant distributed event processing systems. By using an STM-based speculation mechanism, they allowed nodes to optimistically start processing events before their

final delivery (before their respective order is know with certainty) [2].

Yair Amir and Ciprian Tutu presented a complete algorithm for database replication over partitionable networks sophistically utilizing group communication and proved its correctness. Their avoidance of the need for end-to-end acknowledgment per action contributed to superior performance. They showed how to incorporate online instantiation of new replicas and permanent removal of existing replicas. They also demonstrated how to efficiently support various types of applications that required different semantics [1].

M.Wisemann, F.Pedone and A.Schiper provided an abstract and "neutral" framework to compare replication techniques from both communities (distributed systems and databases). The framework has been designed to emphasize the role played by different mechanisms and to facilitate comparisons. Their paper described the replication techniques used in both communities, compared them, and pointed out ways in which they can be integrated to arrive to better, more robust replication protocols [6].

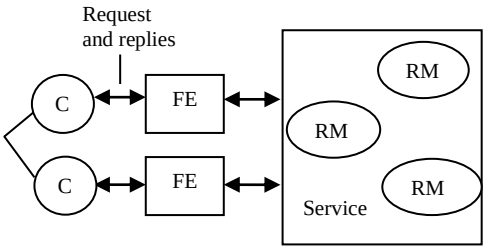
3. Background Theory

Replication is a technique for enhancing services. The motivations for replication are to improve a service's performance, to increase its availability, or to make it fault tolerant. Replication consists of two or more replicas. Replicas are physical objects, each stored at a single computer. Replicas are held by distinct replica managers. Replica managers are components that contain the replicas on a given computer and perform operations up on them directly.

There are two approaches for fault-tolerant service:

1. Passive (primary-backup) replication and
2. Active replication.

In this system, active replication will be used for fault tolerant. General model of replica management is shown in figure 1.



RM= Replica Managers

Client

Front ends

Figure1. General Model of Replica Management

3.1. Passive (primary-backup) replication

The passive model for fault tolerant is shown in figure 2.

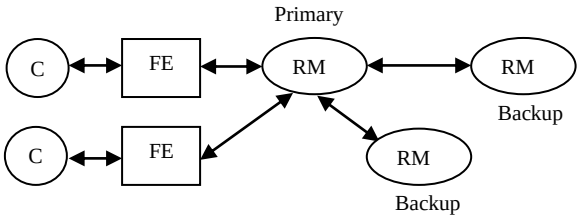


Figure2. Passive Replication

In this model, there is at any one time, a single primary replica manager and one or more secondary replica managers – 'backups' or 'slaves'. In the pure form of the model, front ends communicate only with the primary replica manager to obtain the service. The primary replica manager executes the operations and sends copies of the updated data to the backups. If the primary fails, one of the backups is promoted to act as the primary [3].

3.2. Active Replication

The model for active replication is shown in figure 3. In this model, the replica managers play as a group. Front ends multicast their requests to the group of replica managers and all the replica managers process the request and reply. If any replica manager crashes, then the remaining replica managers continue to respond in the normal way. This configuration can reduce network load dramatically.

1. *Request*: The front end attaches a unique identifier to the request and multicasts it to the group of replica managers. It does not issue the next request until it has received a response.
2. *Coordination*: The group communication system delivers the request to every correct replica manager in the same (total) order.
3. *Execution*: Every replica manager executes the request identically. The response contains the client's unique request identifier.
4. *Agreement:* No agreement phase is needed.

5. *Response:* Each replica manager sends its response to the front end. The front end passes the first response to arrive back to the client and discards the rest [3].

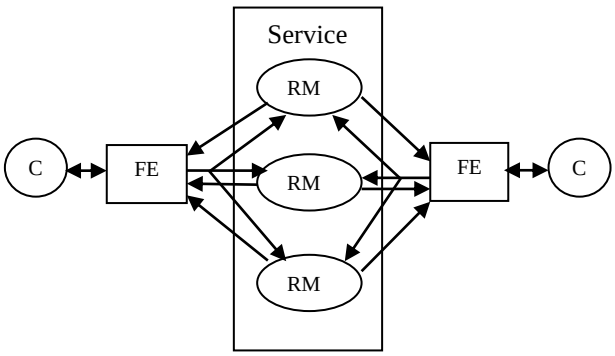


Figure 3. Active Replication

The active replication technique requires that the invocations of client processes be received by the non-faulty replicas in the same order. This requires an adequate communication primitive, ensuring the order and atomicity property. This primitive is called total order multicast or atomic multicast [4].

4. Architectural Components of LLFT

The LLFT system comprises three main architectural components, each of which employs novel techniques. These three components, which are integrated into the complete LLFT system, are described briefly below.

Low Latency Messaging Protocol: The Low Latency Messaging Protocol provides reliable, totally-ordered message delivery by communicating message ordering information from the primary to the backup replicas in a group. It ensures that, in the event of a fault, a backup has the messages and the ordering information that it needs to reproduce the actions of the primary. The replicated applications interact with each other directly, via a group multicast.

Leader-Determined Membership Protocol: The Leader-Determined Membership Protocol ensures that the members of a group have a consistent view of the membership set and of the primary replica in the group. It effects a membership change and a consistent view more quickly than other membership protocols, by selecting a new primary, based on the precedence and ranks of the replicas in the group.

The Virtual Determinizer Framework: The Virtual Determinizer Framework records the order and results of each operation at the primary, and by guaranteeing that the backups obtain the same results in the same order as the primary.

The LLFT system supports two types of leader-follower replication, namely: (1) Semi-active replication and (2) Semi-passive replication

4.1. Semi-active Replication

The primary orders the messages it receives, performs the operations, and provides ordering information for nondeterministic operations to the backups. A backup receives and logs incoming messages, performs the operations according to the ordering information supplied by the primary, and logs outgoing messages, but does not send outgoing messages.

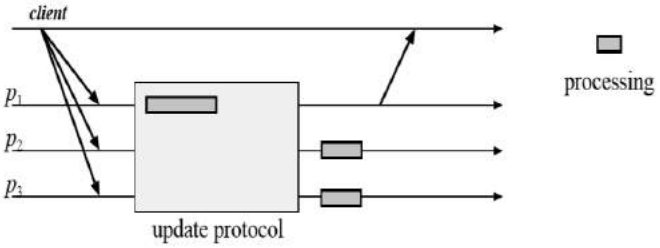


Figure4. Semi-active Replication

4.2. Semi-passive Replication

The primary orders the messages it receives, performs the operations, and provides ordering information for nondeterministic operations to the backups. In addition, the primary communicates state updates to the backups. A backup receives and logs incoming messages, and updates its state, but does not perform the operations and does not produce outgoing messages. Semi-passive replication uses fewer processing resources than does semi-active replication; however, it incurs greater latency for reconfiguration and recovery, if the primary becomes faulty. To maintain strong replica consistency, it is necessary to sanitize (mask) non-deterministic operations not only for semi-active replication but also for semi-passive replication.

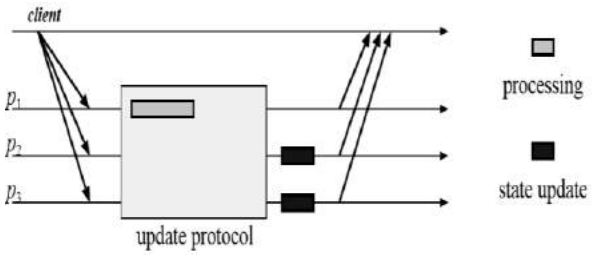


Figure5. Semi-passive Replication

4.3. Algorithm of Proposed System

BEGIN
Step1: Accept the request from clients;
Step2: Requests processing are made at primary;
If (The request processing completes)
{
GOTO: Step3;
}
Else
{
Repeat: Step2;
}
End If
Step3: The primary sends replies to the clients;
Step4: Check the primary status;
If (Status = = OK)
{
Check the replicas status;
For (int r=1; r <= number of replica; r++)
{
If (replica [r] status = = OK)
{
Primary sends its updates to the backup/replica [r].
}
Else If (replica [r] status = = Busy)
{
Primary passively sends its updates to the backup/replica [r].
}
}
}
Else If (The primary then fails before it sends its updates to the backups)
{
Revoke the new primary from the replicas / backups;
Repeat: Step2;
}
End If
Step 5: Check the group of replicas;
If (a group of replicas becomes network disconnected)
{
LLFT ensures that only one component of the partition, referred to as the primary component;
If (Partitioned component = = primary component)
{
LLFT allows continued operation and avoid blocking during partitioning.
}
Else If (network disconnected component = = other components)
{
Might terminate operations and must reapply for admission to the membership;
}
End If
}
End If
END

4.4. Implementation of the Furniture Ordering System

This system implemented a reliable furniture ordering system by the used of backup replica servers. Primarily, the furniture ordering web system used the main server and two backup servers. Every transactions (such as ordering transaction from the clients or data entry transaction of the admin) are made at the primary server of the system. Similarly, the backup servers of the system must be execute the processing transaction to maintain the data consistency between primary server and backup replica servers. When the backup replicas are not idle to execute the processing transaction to maintain the data consistency between primary server and backup replica servers, only the primary server process the requested transaction and then the committed results are propagated to replicas and applied to them.

So, this system not only controls the data consistency and reliability but also control the avoidance for the waiting to busy replicas by the used of LLFT as shown in figure 6.

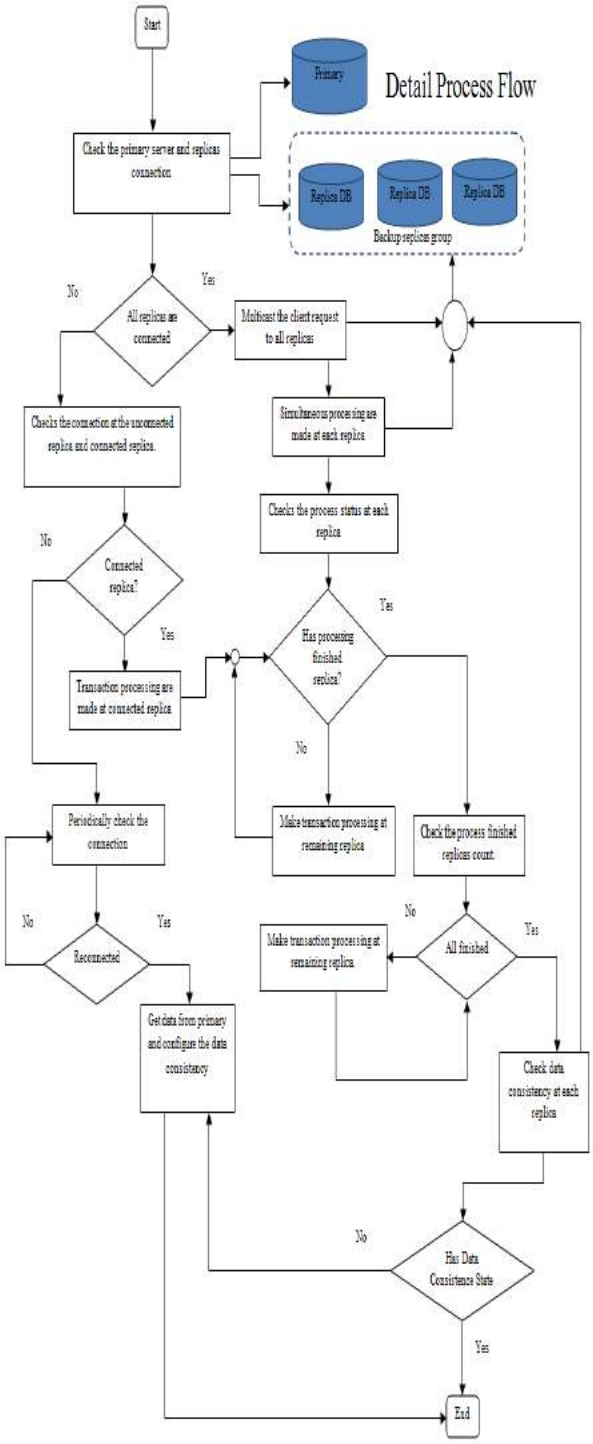


Figure6. The System Flow

Processing the request from the first client updates a data item. Processing the request from the second client updates the same data item, where the interaction between the processing of the two requests is non-deterministic. The request processing completes, and the primary sends replies to the clients. The primary then fails before it sends its updates to the backups. The processing of the requests from the two clients is repeated at the new primary.

However, the non-deterministic interactions between the processing of the two requests are sent to the clients. The processing of the requests at the new primary must repeat the same non-deterministic interactions, if the correct results are to be obtained. If a group of replicas becomes partitioned, LLFT ensures that only one component of the partition, referred to as the primary component. Within the primary component, LLFT maintains virtual synchrony, i.e., if the primary fails, the new primary must advance to the state of the old primary, and the state known to the remote groups of its connections, before the old primary failed. The processes of the other components might terminate operations and must reapply for admission to the membership. Care must be taken to recover those operations and to restore consistency. LLFT ensures to continue the operation and avoid blocking during partitioning.

5. Conclusion

The Low Latency Fault Tolerance (LLFT) system provides fault tolerance for distributed applications deployed over a wide-area network. This system can be replicated with strong replica consistency using LLFT, without any modifications to the applications. LLFT achieves low latency message delivery under normal conditions and low latency reconfiguration and recovery when a fault occurs. The application transparency, and low latency of LLFT make it appropriate for a wide variety of distributed applications, particularly for latency-sensitive applications.

References

[1]. Chandra, T. D. and Toueg, S. (1996) Unreliable failure detectors for reliable distributed systems. J. of ACM, 43(2), 225-267.

[2]. Birman, K. P. and van Rennesse, R. (1994) Reliable Distributed Computing Using the Isis Toolkit. IEEE Computer Society Press, Los Alamitos, CA.

[3]. M.Wisemann, F.Pedone and A.Schiper, “Understanding Replication in Databases and Distributed Systems”, EPFL-ETHZ DRAGON project.

[4]. R.Guerraoui and A.Schipe, “Fault-Tolerance by Replication in Distributed Systems”, Proc. Reliable Software Technologies, 1996.

[5]. R.Gueraoui and A.Schiper,“Software-Based Replication for Fault Tolerance”, Swiss Federal Institute of Technology, IEEE, 1997.

[6]. Y.Amir and C.Tutu, “From Total Order to Database Replication”, Johns Hoplins University, Department of Computer Science.

[7]. L. Lingxia , X.Jingbo , M.Zhiqiang and L.Ruixin “Rapid-Response Replication: A Fault Tolerant Algorithm Based on Active Replication”.

[8]. Y.Xinfeng, “Providing Reliable Web Services through Active Replication”,Department of Computer Science, Auckland University, New Zealand.

An Efficient DCT-Based Video Watermarking Method for Copyright Control

May Tharaphy Htun, Twe Ta Oo
University of Computer Studies, Yangon, Myanmar
mthphtun@gmail.com

Abstract

With the aim of copyright control in transmission of digital video, this paper presents a DCT-based video watermarking method. Its essence is to embed the copyright related information on digital videos in such a way that it can later be extracted in case copyright violation is detected. The embedded information needs to be invisible and robust against malicious attacks. In this system, watermark is embedded by only changing the luminance of video frames. As the human visual system cannot easily detect light intensity changes in images, the method presented in this paper well preserves the visual quality of watermarked videos by keeping the PSNR of more than 50dB. Moreover, this system solves the frame drop problem by repetitively embedding the watermark in all video frames. It also provides remarkable robustness against compression attack as it is based on the DCT, which is a proven method used in JPEG compression.

Keywords: video watermarking, DCT, copyright control, DRM.

1. Introduction

The advancement of technologies is leading the world from analog age to digital age. With the help of technologies, digital media such as video can be copied, edited, transmitted, and shared faster and easier than before. However unfortunately, it also brings illegal acts such as piracy, stealing, and infringement. It not only damages the intellectual property rights of digital works' owners but also affects the market order of electronic publications. In addition, pirated works may pose a great threat to the safety of users as they are of poor quality and may carry and spread computer viruses. Thus, the issue of piracy and copyright infringement is undoubtedly a huge obstacle to the healthy development of the digital video industry [1]. It is also the motivating factor in developing watermarking techniques.

In digital video watermarking, copyright information, e.g. logo, signature, etc., is hidden in the host video in such a way that it is difficult to be distorted by a variety of attacks. When copyright violation is detected, copyrighted owner can be proved by detecting the embedded information.

Digital video watermarking can be used in a wide variety of applications where the issues of security and ownership are vital; e.g. owner identification, copy and usage control, broadcast monitoring, theater identification, and military applications. Requirements for good watermarking methods differ based on the applied area. For copyright control, a good watermarking algorithm should be robust against both intended and unintended signal processing operations like lossy video compression. The embedded watermark must be invisible, undetectable, and undeletable as well.

The rest of the paper is organized as follows. Section 2 discusses the video watermarking literature. Section 3 explains the proposed luminance-based watermarking system in detail, including both embedding and extraction processes. Section 4 and 5 present the experimental results and conclusion, respectively.

2. Related Work

In the literature, a lot of digital video watermarking methods used in copyright control applications have been proposed. Those methods satisfy some or all of the above mentioned requirements. This section introduces some of them.

Kumar and Shukla [2] implemented a Discrete Cosine Transform (DCT) based watermarking technique for AVI videos using Matlab Simulink. In that paper, the watermark image is firstly segmented and then embedded in separate frames. It means that, for identifying the copyright owner, watermark must be extracted from multiple frames and reconstructed. In the case of frame loss, watermark detection cannot be successful and copyright protection fails.

Yadav and Anand [3] also proposed a DCT based digital video watermarking technique using Matlab Simulink. The main aim of that method was to detect modifications on the host video by using visible watermark.

Kalra [4] also proposed a DCT and thresholding based digital video watermarking method that repetitively embeds the same watermark in every frame of the video. Even if some frames were lost, the watermark could be successfully detected in other frames.

As stated, most of the video watermarking methods are developed in the DCT domain. It is because the DCT has been proved as a good choice for compressing the images; it is used in the JPEG image compression standard [5]. As a video is a sequence of image frames, DCT also works well for videos. Watermarking methods developed in the DCT domain are mostly robust against compression attacks.

The method presented in this paper also achieves good robustness against compression attack as it is developed in the DCT domain. Moreover, it also deploys the fact that the human visual system is not sensitive to light intensity change [6]. It modifies only the luminance of video frames to hide the watermark and thus achieves good invisibility as well.

3. The Luminance-based Watermarking System

Properties of the DCT such as being real, orthogonal, and separable with fast computational algorithms have made it as a great relevance for data compression [7]. To exploit the effectiveness of DCT in image compression, the method presented in this paper is developed in the DCT domain.

The DCT divides an image into spectral sub-bands of differing importance with respect to the image's visual feature

[8]. It has the property that, for a typical image, most of the visually important information about the image is concentrated in just small coefficients of the DCT [9].

In this system, the two dimensional DCT (2D-DCT) is applied on each image frames of the host video to extract visually important feature that will carry the watermark information. Detailed procedure is discussed below.

3.1 Watermark Embedding Process

Figure 1 shows the block diagram of watermark embedding process. Detailed explanation is given below.

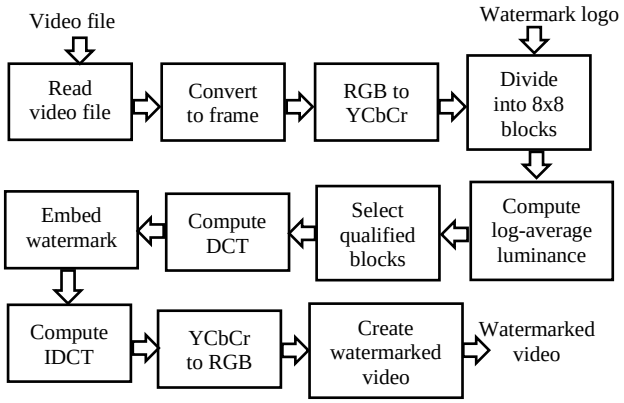


Fig. 1: Watermark embedding process

***Step 1*:** Firstly, the host video is read as an input. Video can be any type, e.g. AVI, MP4, etc.

Step 2: The host video is then converted into image frames; the number of frames will be different depending on the frame rate. As an example, a 16-sec long video with a frame rate of 30 fps consists of 480 frames. The method presented in this paper has no restriction on the frame size. For the experiments discussed in this paper, 640×480 size is used.

Step 3: The RGB frame is then converted to YCbCr frame as stated in eq. 1 to eq. 3.

$$Y = 16 + \frac{65.74}{256}.R + \frac{129.06}{256}.G + \frac{25.06}{256}.B, \quad (1)$$

$$Cb = 128 - \frac{37.95}{256}.R - \frac{74.49}{256}.G + \frac{112.44}{256}.B, \quad (2)$$

$$Cr = 128 + \frac{112.44}{256}.R - \frac{94.15}{256}.G - \frac{18.29}{256}.B, \quad (3)$$

where *R*, *G*, and *B* are the red, green, and blue components of the RGB color space, respectively.

Step 4: From YCbCr, only the *Y* component (i.e. luminance) is extracted for watermark embedding. As mentioned above, human eyes cannot easily detect the changes in luminance values. Thus, embedding the watermark in the *Y* component will keep the invisibility of the watermarks and thus preserve the video quality.

The *Y* component of each frame is then divided into 8×8 blocks. In order to keep the best possible invisibility, only the selected blocks (i.e. not the whole frame) will be used for watermark embedding.

The watermark image to be embedded is first converted to binary value and also divided into 8×8 blocks [10].

Step 5: To embed the watermark image, only 64 host image blocks are needed. Among the 4,800 blocks, the 64 blocks whose log-average luminance is the closet to that of the entire image are selected. The log-average luminance is calculated as described in eq. 4.

$$L_{avg} = \exp(\sum Log(\delta + Y_{x,y}) / N), \quad (4)$$

where $Y_{x,y}$ is the *Y* component of the pixel at location [*x*,*y*] of the host image/block, δ is a small value (0.1 in this paper) to avoid taking the log of a black pixel, L_{avg} is the log-average luminance, and *N* is the number of pixels in the image/block.

Step 6: The best 64 blocks with the log-average luminance closet to the log-average luminance of the entire image are chosen based on the criteria [$L_{avg} - \beta$, $L_{avg} + \beta$] where β is the minimum floating-point value that is enough to determine the adequate number of qualified blocks.

Step 7: The 2D-DCT defined in eq. 5 is applied to the *Y* component of each selected block.

$$D_{u,v} = \frac{1}{\sqrt{2N}} C_u C_v \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} Y_{x,y} Cos\left[\frac{(2y+1)v\Pi}{2N}\right] Cos\left[\frac{(2x+1)u\Pi}{2N}\right],$$

$$C_u = \begin{cases} \frac{1}{\sqrt{2}}, \text{for } u = 0 \\ 1, \text{ otherwise} \end{cases}, \quad C_v = \begin{cases} \frac{1}{\sqrt{2}}, \text{for } v = 0 \\ 1, \text{ otherwise} \end{cases}, \quad (5)$$

where $D_{u,v}$ is the DCT coefficient in row *u* and column *v*, and *N* = 8 for 8×8 blocks.

Step 8: The pixels in each 8×8 block of the watermark image (already converted to binary) are embedded in the DCT coefficients of each selected block as defined in eq. 6. If the watermark's pixel is white (i.e. 1) then an additional factor α is added to the DCT coefficient. If the pixel is black (i.e. 0) then α is subtracted from the DCT coefficient.

$$D'_{u,v} = \begin{cases} D_{u,v} + \alpha, & \text{where } W_{x,y} = 1 \\ D_{u,v} - \alpha, & \text{where } W_{x,y} = 0 \end{cases}, \quad (6)$$

where $D_{u,v}$ and $D'_{u,v}$ are the DCT coefficients before and after embedding, respectively.

Step 9: The watermarked DCT coefficients are converted back to the *Y* component by using the inverse DCT (eq. 7).

$$Y'_{x,y} = \frac{1}{\sqrt{2N}} \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} C_u C_v D'_{u,v} Cos\left[\frac{(2y+1)v\Pi}{2N}\right] Cos\left[\frac{(2x+1)u\Pi}{2N}\right], \quad (7)$$

where $D'_{u,v}$ is the watermarked DCT coefficient and $Y'_{x,y}$ is the watermarked *Y* component.

***Step 10*:** The watermarked frame is then constructed by converting the YCbCr to the RGB by using eq. 8 to eq. 10.

$$R = \frac{298.08}{256}.Y + \frac{408.58}{256}.C_r - 222.92, \quad (8)$$

$$G = \frac{298.08}{256}.Y - \frac{100.29}{256}.C_b - \frac{208.12}{256}.C_r - 135.57, \quad (9)$$

$$B = \frac{298.08}{256}.Y - \frac{516.41}{256}.C_b - 276.84, \quad (10)$$

All the above steps 3 to 10 are repeated for all frames. The aim of the proposed system is to identify the owner of the

media in case copyright violation is detected. Thus, the watermark image is repetitively embedded in all frames so that some frames can still be used for owner identification even if some are manipulated by malicious users.

Step 11: Finally, all the watermarked frames are combined to construct the watermarked video.

3.2 Watermark Extraction Process

Block diagram of the watermark extraction process is shown in Figure 2. Except the extraction process (the last block of Fig. 2), the first seven blocks are the same as Fig. 1.

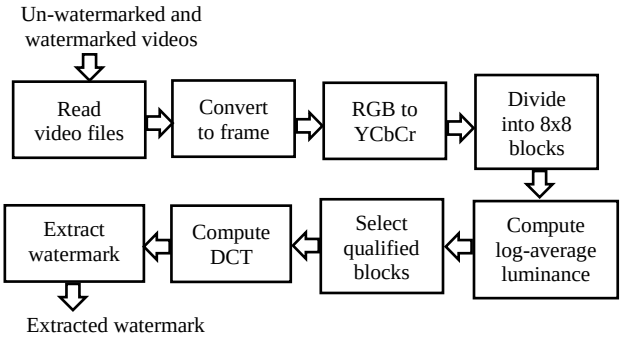


Fig. 2: Watermark extraction process

Step 8: The method presented in this paper needs the original un-watermarked video for the extraction process. As stated in eq. 11, the difference between the watermarked DCT coefficient and the un-watermarked coefficient is calculated. If the result is greater than or equal to zero, then the watermark color is assumed as white; otherwise, it is assumed as black.

$$W'_{x,y} = \begin{cases} 1, & \text{where } D'_{u,v} - D_{u,v} \geq 0 \\ 0, & \text{where } D'_{u,v} - D_{u,v} < 0 \end{cases}, \qquad (11)$$

where $W'_{x,y}$ is the extracted watermark pixel, $D_{u,v}$ and $D'_{u,v}$ are the DCT coefficients before and after embedding, respectively.

The above process is performed for all frames. Then, all the extracted watermarks are compared with the original watermark, and the one with the highest similarity is chosen as the final result.

4. Experimental Results

This section presents the experimental results, and analyzes and evaluates the performance of the proposed system.

For the experiments discussed below, the watermarking method presented in this paper is simulated in Matlab and applied on a total of 30 videos (.mp4). Among them, 10 are "Nature" videos with views of the nature backgrounds, 10 are "Cartoon" videos with the movement of people, and 10 are "Music" videos with dance and singers.

Figure 3 shows an image that will be used as the copyright logo. It can be any format and "JPEG" format with 64×64 size is used in this system. That image is converted to binary format before the embedding process.

Fig. 3: Watermark image (logo)

For embedding the watermark logo, the host MP4 video (frame size of 640x480) must be firstly converted into image frames. Then, the color format of each frame is changed to YCbCr as this system embeds the watermark in the luminance value (*Y*) of the pixels. Figure 4 shows how an image frame with RGB, YCbCr, and *Y* color format looks like.

Fig. 4: Image frames with RGB, YCbCr, and *Y* value color formats (left to right)

Then, as previously discussed in section 3.1, both the watermark logo and the *Y* component of each frame is divided into 8×8 blocks. The result is 64 blocks for the watermark logo and 4,800 blocks for *Y* component per frame. Then, the 64 *Y* component blocks whose log-average luminance are closet to the log-average luminance of the entire image are chosen for watermark embedding. Then, each selected block is converted into DCT domain and each watermark block is embedded on the DCT coefficients of each selected *Y* component block. The embedding rule is defined in eq. 6. Finally, the watermarked blocks are converted back to time domain and to RGB color space. After embedding the watermark logo in all frames, we get the watermarked video file. For checking the visibility of the watermark, Fig. 5 shows an example of the original and watermarked video frames obtained from our experiments.

Fig. 5: Original (left) and watermarked (right) video frames

4.1 Evaluation on Watermark Invisibility

In this system, invisibility of the embedded watermark is analyzed in terms of the peak signal to noise ratio (PSNR).

The PSNR is widely used to measure the invisibility of the watermark – in other words, visual quality level of the watermarked image/video. Typical PSNR values for lossy image and video compression is 30-50 dB, where the higher the better [11]. Optimal values for JPEG image transmission

over wireless channels are considered to be about 20-25 dB [12]. PSNR is calculated as defined in eq. 12.

$$PSNR(dB) = 10 \log 10 \left(\frac{(MAX_i)^2}{MSE} \right), \quad (12)$$

where MAX_i is the maximum possible pixel value of the image; *MSE* is the mean square error defined in eq. 13 and it is one of the simplest methods to measure the distortion.

$$MSE = \frac{1}{N \times M} \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \left(I(x,y) - I_W(x,y) \right)^2, \quad (13)$$

where $N \times M$ is the image size, $I(x,y)$ and $I_w(x,y)$ are the pixel values at location $[x,y]$ of the original and watermarked images, respectively.

In this system, visual quality of the watermarked video depends on the additional factor α used in the embedding process. Table 1 shows the PSNR results after watermark embedding with different α values for each video type. It can be seen that the smaller the α, the higher the PSNR which means the better the visual quality of the watermarked video.

Unfortunately, there is a tradeoff for the choice of α for good robustness and good invisibility. The larger the α, the the more robust to malicious manipulation. Thus in this system, we choose $\alpha = 5$ for keeping acceptable robustness and invisibility.

Table 1: PSNR after watermark embedding

Video	PSNR (dB)	α				
		10	*5*	*3*	*1*	*0.1*
Nature1.mp4		54.85	58.04	60.41	66.28	69.83
Music1.mp4		54.65	57.89	60.36	67.15	73.13
Cartoon1.mp4		54.62	56.98	58.15	66.35	69.75

4.2 Evaluation on Watermark Similarity

The main aim of the proposed system is to identify the media owner in case copyright violation is detected. Thus, this system should be able to extract the embedded copyright logo no matter how severe the host video is manipulated by malicious users. This section analyzes the robustness of the proposed system to malicious attacks.

Robustness is measured in terms of the similarity between the embedded and extracted watermarks [13], as mentioned in eq. 14.

$$S_{x,y} = \begin{cases} 1, & \text{if } W'_{x,y} = W_{x,y} \\ 0, & \text{if } W'_{x,y} \neq W_{x,y} \end{cases}, \quad (14)$$

where $S_{x,y}$ is the similarity of each pixel, $W'_{x,y}$ and $W_{x,y}$ are the extracted and embedded watermark pixels, respectively. Then, similarity for the whole image (σ) is calculated as shown in eq. 15. The 100% is the maximum similarity.

$$\sigma(\%) = 100 \times \frac{1}{N \times M} \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} S_{x,y}, \quad (15)$$

where $N \times M$ is the size of the watermark image (64×64).

Figure 6 shows the watermark logo extracted from a video file with no attack. The extraction process has already been discussed in section 3.2. As we can see, the original and extracted watermarks are exactly the same, which yields 100% similarity.

Fig. 6: Original (left) and extracted (right) logos for no attack

The following subsections discuss the signal processing attacks that are applied on the watermarked video for testing the robustness of the proposed system.

4.2.1 Attacks on Watermarking Methods

Watermarked videos may face a variety of unintended or intended attacks trying to destroy the copyright information. A good watermarking method should resist those attacks. In this system, the following attacks are simulated in Matlab for robustness testing.

1. Compression schemes
2. Geometric operations such as cropping and rotation
3. Signal processing operations such as quantization, filtering, and noise addition

For all those attacks, parameters are chosen guessing on attacker perspective.

4.2.1.1 Compression Attack

Compression attack may occur when users intentionally compress to distort the watermark or unintentionally compress to reduce the video file size.

This system is implemented on the DCT, which is a basis of JPEG image compression, and thus it well resists the compression attack. For simulation of this attack, the parameter "compression ratio" is defined as the storage size of the original image divided by the storage size of the compressed image. For example, if the video is originally 1000 bytes and needed to be 500 bytes then the ratio would be 2[1]. Figure 7 shows the extracted watermark after compressing the "Nature1.mp4" with compression ratio of 2. It achieves the similarity result of 100%.

Fig. 7: Extracted logo after compression attack

4.2.1.2 Cropping Attack

Cropping attack occurs when the video is unintentionally or intentionally cropped to remove some portion. The attack parameter is the desired frame size. As an example, Fig. 8 shows the extracted watermark after cropping the 640×360 frames of the "Nature1.mp4" to 620×340. It achieves the 62.01% similarity.

[1] https://www.mathworks.com/matlabcentral

Fig. 8: Extracted logo after cropping attack

4.2.1.3 Rotation Attack

Rotation attack rarely occurs when the users rotate the video for some reason like video editing. Attack parameter can be any angle value such as 90˚ and 180˚; positive value for counterclockwise and negative value for clockwise directions. Figure 9 shows the extracted watermark from the 180˚ clockwise rotated video of "Nature1.mp4", which is upside down after attack. It achieves the 62.11% similarity.

Fig. 9: Extracted logo after rotation attack

4.2.1.4 Quantization Attack

Quantization attack occurs when reducing the number of colors required to represent a digital image to reduce its file size. Attack parameter is the quantization levels specified in an N element vector. The quantized image contains $N + 1$ discrete integer values in the range 1 to $N + 1$[2]. For example, to get 4 discrete levels, N needs to be 3. Figure 10 shows the extracted result from a 4-level quantized video of "Nature1.mp4" and achieves 79.43% similarity. The original video was 256-level quantized.

Fig. 10: Extracted logo after quantization attack

4.2.1.5 Filtering Attack

Filtering attack occurs when users perform video editing for quality enhancement or intentionally filter the video to distort watermark. Motion filter is used in this attack, which results the blurring artifacts in the filtered video[3]. Two attack parameters, lens and theta, specify the length of the motion and the angle of motion in degrees in a counter-clockwise direction. The default lens is 9 and the default theta is 0, which corresponds to a horizontal motion of pixels[4]. Figure 11 shows the result from the filtered video "Nature1.mp4" with attack parameter of 10 for lens and 10 for theta. Similarity result is 59.06%.

Fig. 11: Extracted logo after filtering attack

4.2.1.6 Adding Noise

Noise addition attack is one of the most commonly occurred attacks in digital watermarking as communication channel is noisy in nature. It can also occur when users perform video editing. Additive white Gaussian noise (AWGN) is a basic noise model used in information system[5]. The noise density is important attack parameter. Figure 12 shows the result after AWGN attack on "Nature1.mp4" with noise density of 0.01. Similarity result is 69.65%.

Fig. 12: Extracted logo after adding noise attack

Table 2 summarizes the average similarity scores of all 30 videos for all kinds of attack. Each attack is simulated with two attack parameters and PSNR values show how severe the attack is. As we can see in Table 2, similarity is 100% when there is no attack and drops to 58% as the lowest after attacks. As the nature of attacks are different, we can also see that the extracted logo from "rotation" attack with 62.11% similarity is worse to see than the logo from "cropping" attack with 58.57% similarity. In general, we can see that 58% similarity is acceptable and extracted logos are still recognizable with human eyes, except for rotation attack. Thus, it can be concluded that the method presented in this paper achieves good robustness to most signal processing attacks to some acceptable extent.

Table 2: Average similarity scores for various attacks

Attack Type	Attack Parameter	PSNR (after attack)	Average Similarity	Extracted Logo
No attack	-	57.01	100%	
Compression	2	56.18	100%	
Compression	4	56.18	100%	
Cropping	620 x 340	31.10	62.01%	
Cropping	600 x 300	30.79	58.57%	
Rotation	90	26.41	58.84%	
Rotation	180	27.46	62.11%	
Quantization	3	34.50	79.43%	
Quantization	7	39.45	80.59%	
Filtering	5,5	38.98	59.08%	
Filtering	10,10	36.62	59.06%	
Noise	0.01	30.44	69.65%	
Noise	0.02	31.14	69.46%	

5. Conclusion

This paper presented a digital video watermarking method in the DCT domain. A binary logo image was embedded as copyright information in the host video, which could be any

[2]https://www.mathworks.com/help/images/ref/ imquantize.html

[3]https://en.wikipedia.org/wiki/Motion_blur

[4]https://www.mathworks.com/help/images/ref/ imfilter.html

[5]https://www.mathworks.com/help/images/ref/ imnoise.html

type such as .avi or .mp4. The proposed embedding process only changed the luminance value of the host video frames and thus it could well preserve good video quality. In addition, as the method presented in this paper was developed in the DCT domain, it achieved good robustness to a certain extent for both intended and unintended attacks. Thus, it can be concluded that the proposed system is appropriate to be used in copyright protection applications.

References

[1] X. Yu, C. Wang, and X. Zhou, "A survey on robust video watermarking algorithms for copyright protection," in Multidisciplinary Digital Publishing Institute Scientific Conference, vol. 3, Oct 2018.

[2] M. Kumar and D. Shukla, "DCT domain video watermarking technique for AVI video," vol. 3, Apr 2015.

[3] S. Yadav and P. Anand, "DCT based digital video watermarking using MATLAB/Simulink," in Proc. IEEE International Conference on Image Processing, vol.3, Feb 2015.

[4] G. Kalra, "DCT and thresholding based digital video watermarking," in Proc. IEEE International Conference on Image Processing, May 2012.

[5] A. Watson, "Image compression using the Discrete Cosine Transform," NASA Ames Research Center, Mathematica Journal, pg. 81-88, Jan 1994.

[6] M. Balsky, "Luminance values extraction from digital images," IEEE Lighting Conference of the Visegrad Countries, Sep 2016.

[7] K. Kavitha, B. Shan, "Video watermarking using DCT and DWT: a comparison," European Journal of Advances in Engineering and Technology Conference, pg. 83-87, May 2015.

[8] H. Huang, "An adaptive video watermarking technique based on DCT domain," in Multidisciplinary Digital Publishing Institute Scientific Conference, Jun 2014.

[9] S. Sharma, V. Kumar, "A survey of blind and non-blind watermarking technique," in Proc. IEEE International Conference on Image Processing, vol. 7, Dec 2016.

[10] J. Hussein, "Luminance-based embedding approach for color image watermarking," International Journal of Image, Graphics, and Signal Processing, pg. 49-55, Apr 2012.

[11] M. Yeung and F. Mintzer, "An invisible watermarking technique for image verification," in Proc. IEEE International Conference on Image Processing, vol. 2, pg. 680–683, Oct 1997.

[12] N. Thomos, N. Boulgouris, and V. Strintzis, "Optimized transmission of JPEG2000 streams over wireless channels," IEEE Transactions on Image Processing, Jan 2006.

[13] C. Rey and J. Dugelay, "A survey of watermarking algorithms for image authentication," EURASIP Journal on Applied Signal Processing Conference, 2002.

Enhancing Parallel Algorithms for Generating Combinations with Scheduling

Nanda Thant Sin, May Aye Khine
University of Computer Studies, Yangon
nandathantsin@ucsy.edu.mm, mayayekhine@ucsy.edu.mm

Abstract

In this paper we propose the scheduling methods to improve the load balancing of Sergey NOVIKOV's two parallel algorithms for generating combinations without repetitions of m out of n objects. The workloads of the tasks are not equaled. So, when the number of available processors is less than maximum possible parallelization, scheduling will be needed to improve the balancing of the load. Simple static and dynamic scheduling are used in this paper.

Keywords: ***parallel algorithm, combination without repetitions, Boolean vector, scheduling***

1. Introduction

The combinatorial analysis has an incredible number of applications in Computer Science [1]. For Example, graph theory, the study of objects and connections between them, is a part of it. Combinatorial techniques are also used in algorithm and data structures for analysis and design extensively. The enumeration of combinations without repetitions is one of the basic tasks in computer science. One of its application areas is in genetic applications [2]. To enhance the performance of computations, devices and processors that operate at the fastest speed possible are invented every year. A different approach to improve the performance is using the parallel computer which operates on several processors.

Related work is described in section 2, and Background theory and proposed approach is in section 3. Implementation detail is explained in section 4. The performance evaluation is in section 5. Some concluding remarks are given in Section 6.

2. Related Work

The related works of parallel algorithms for generating combinations are discussed in this session.

A lot of parallel algorithms have been proposed for generating combinations without repetitions [2,3,4,5]. All of them have strengths and weaknesses. Some algorithms need a constant number of processors and some require arbitrary number of independent processors. Some parallel algorithms require a special algorithm. For example, a numbering system for combinations [6] is required in Selim G. Akl's algorithm [4]. In some parallel algorithms [2,5], task sizes for each processor are different. Scheduling techniques are used to make the better balancing of the load. In Sergey NOVIKOV's paper [5], two parallel algorithms for generating combinations was proposed. The combinations are represented as Boolean vectors. Although these two parallel algorithms allow to parallelize the combination generation on the arbitrary numbers of processors, the tasks that are split are not equaled.

3. Background Theory and Proposed Approach

A combination is a mathematical technique to calculate all possible selections of the subset objects from the set of objects where the order of the selections doesn't matter. Two combinations are

considered different if they contain at least one different object. Mathematical formula for the combination without repetition is

$$C(n,r) = \binom{n}{r} = \frac{n!}{r!\,(n-r)!}$$

Where:

- n is the total number of objects
- r is the selected distinct objects

3.1. Sequential Algorithm SAGC

Sergey NOVIKOV presented a sequential algorithm for generating combination (SAGC) which uses the Boolean vector representation to represent the combinations. The algorithm allows to generate the next combination for each considered combination (excluding the last one) as well as the previous combination (excluding the primary one).

3.2. Modified Parallel Algorithm PAGC1

PAGC1 has three main steps. Firstly, the main processor splits the tasks, maximum possible parallelization is (n-m) * (m-1). For each task, the starting combination is calculated and generate the remaining combinations until it reaches to the starting combination of the next task. After all the combinations for each task is generated, they are sent back to the main processor to combine.

Scheduling can be added after the calculation of starting and ending combinations and before the combination sequences are generated in each processor. The whole algorithm is similar to previous algorithms [2,4]. Like in the paper [2], the load is not balanced.

3.3. Modified Parallel Algorithm PAGC2

PAGC1 can be think of as a horizontal splitting of the tasks while PAGC2 is a vertical splitting of the tasks. The algorithm's idea is to generate short Boolean vectors first and then combine them to make combination sequences. To combine short vectors, a sequential algorithm for generating connecting vectors SAGCV is used. The task of generating connecting vectors is similar to "balls and urns" problem.

Scheduling can be added before the generation of short Boolean vectors, connecting vectors and connection of short Boolean vectors. PAGC2 allows to parallelize the maximum processors C (s+m-1, m), where s = ⌊n/m⌋. Again, the loads are not balanced just like PAGC1.

3.4. Scheduling

When the tasks split is not balanced, scheduling is used in order to balance the load [2]. For these algorithms, four different types of scheduling, static, dynamic, guided and LPT are used.

Static scheduling distributes the tasks to the processers repeatedly until every task is distributed. It works well when the tasks have the same load balance which means the same computational cost.

In the dynamic scheduling, the task distribution is not decided beforehand. Firstly, each processor gets a single task to work. When a processor finishes the current task and there are still remaining tasks, one of the remaining tasks is assigned to that processor. This type of scheduling works better when the tasks have unbalanced computational costs.

Guided scheduling is similar to dynamic scheduling and the difference is the size of the tasks given to the processor at a time. The size of the assigned tasks equals to the number of remaining tasks divided by the number of processors. When a processor finishes the assigned tasks, it will be assigned again in the same pattern. This type of scheduling works best when the poor load balancing occurs toward the end of the computation.

If the computational costs of the tasks are known and the tasks are independent of each other, LPT scheduling can be used. In this method, the tasks are

scheduled one by one in decreasing order of processing time and each task is scheduled on the processor on which it finishes earliest until every task is assigned.

4. Implementation of the System

In PAGC1, the computational costs of the split tasks are not known until actual combination generation. Therefore, LPT scheduling cannot be used. On the other hand, in the PAGC2, the computational costs are known ahead. LPT scheduling is possible to use. The algorithms are implemented using C++ language and compiled with GCC. OpenMp, which a simple C++ compiler extension that allows to add parallelism into existing source code, is used for parallelization.

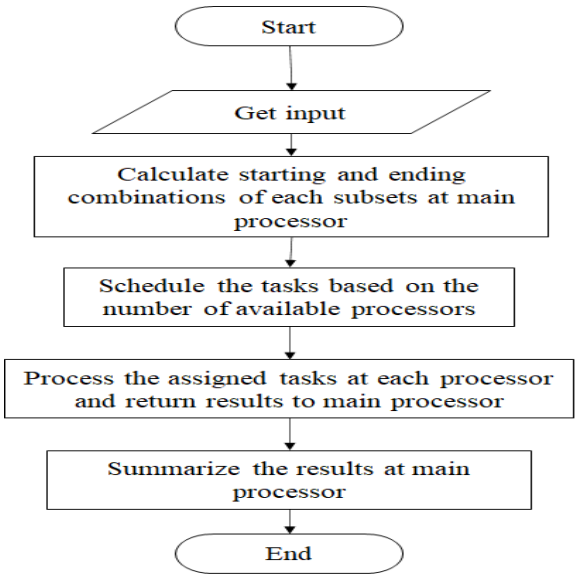


Figure 1: The System Flow of modified PAGC1

In modified PAGC1, it first calculates the starting combination and ending combination for each task. Then, the processors generate the combinations from the starting combination to ending combination for each task using the previous SAGC algorithm. Finally, the main processor sums the results into a single set. For the tasks split, maximum possible is (n-m) *(m-1) and scheduling is done before the combination generation.

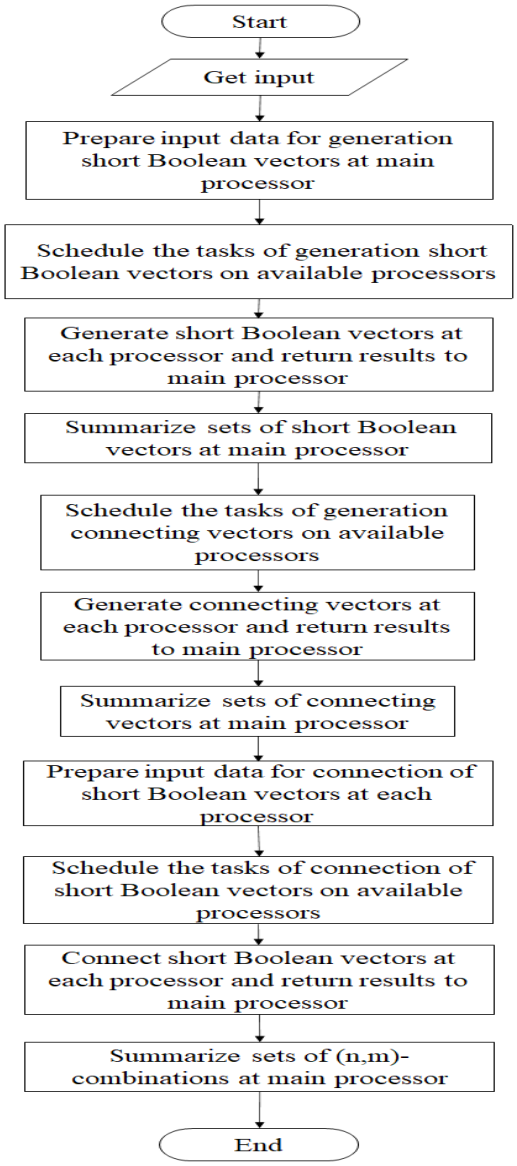


Figure 2: The System Flow of modified PAGC2

Modified PAGC2 has nine main steps. Step 1 prepare the input data to generate the short Boolean vector. The generation of short Boolean vector is done in step 2 and the results are summed together in the main processor in the step 3. In the step 4, SAGCV is used for the connecting vectors generation and the results are summed together in the main processor in the step 5. Step 6 is about preparing input data for connection of short Boolean vectors and the connecting happens in step 7. Step 8 and 9 summarize and finalize the combination generation. Scheduling is done before step 2, step 4 and step 7.

5. Performance Evaluation

Firstly, the execution times based on the scheduling are compared. PAGC1 has more task splits than PAGC2 when the value of m becomes closer to the $n/2$ and PAGC2 has more task splits than PAGC1 when the value of m becomes closer to 1 or n. To see the results clearly, we compared the execution time for enumerating C(24,12), C(25,13), C(30,9) and C(100,3) combinations, using up to 32 processors. The first two combination enumerating will show the results where m equals to n/2 and the rest will show the results of m becomes closer to 1.

In PAGC1, C(24,12) and C(25,13) allow to split the tasks in 132 and 144 respectively whereas PAGC2 allows only 13 task splits in both cases. For C(30,9), PAGC1 allows 168 task splits and PAGC2 allows 55. For C(100,3), PAGC1 allows to split only 194 tasks but PAGC2 allows up to 6545 tasks.

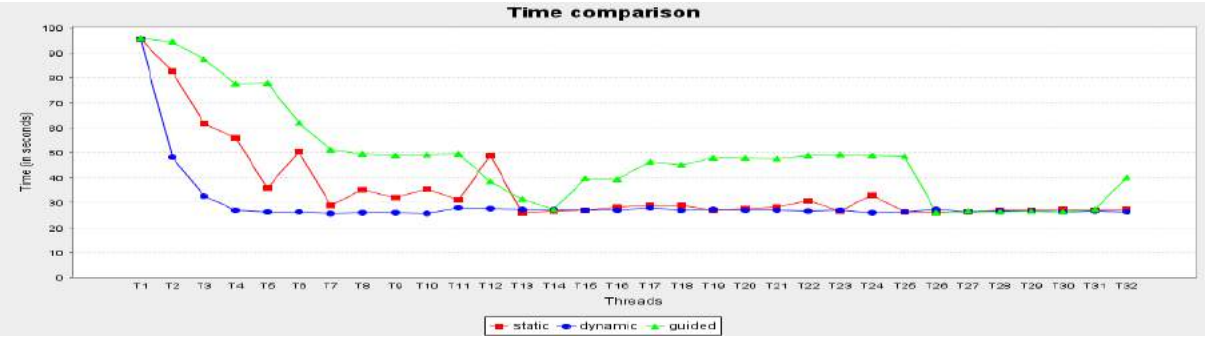


Figure 3: Comparison of total execution for enumerating C(24,12) combinations with modified PAGC1 using static, dynamic and guided scheduling

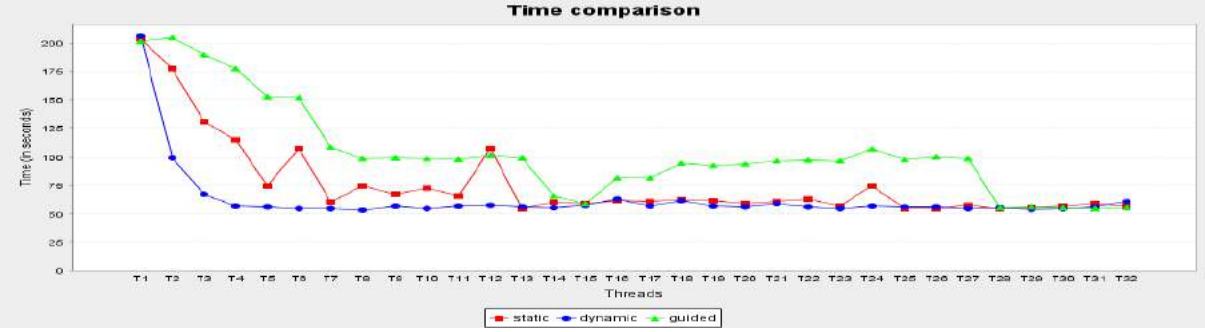


Figure 4: Comparison of total execution for enumerating C(25,13) combinations with modified PAGC1 using static, dynamic and guided scheduling

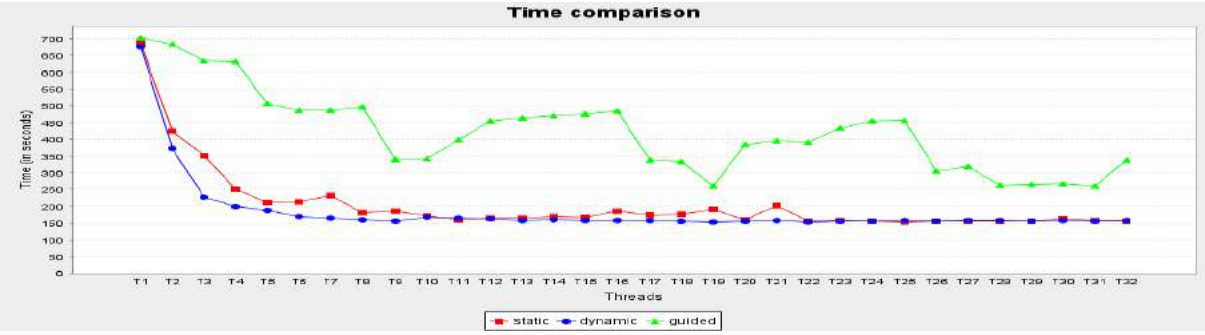


Figure 5: Comparison of total execution for enumerating C(30,9) combinations with modified PAGC1 using static, dynamic and guided scheduling

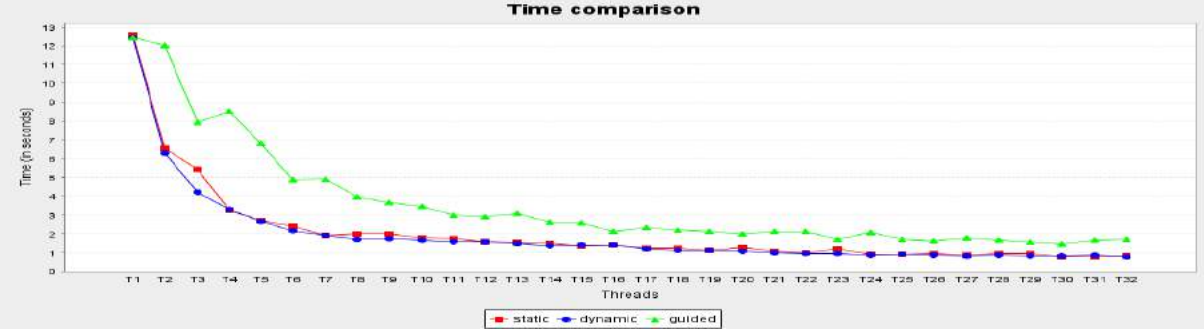


Figure 6: Comparison of total execution for enumerating C(100,3) combinations with modified PAGC1 using static, dynamic and guided scheduling

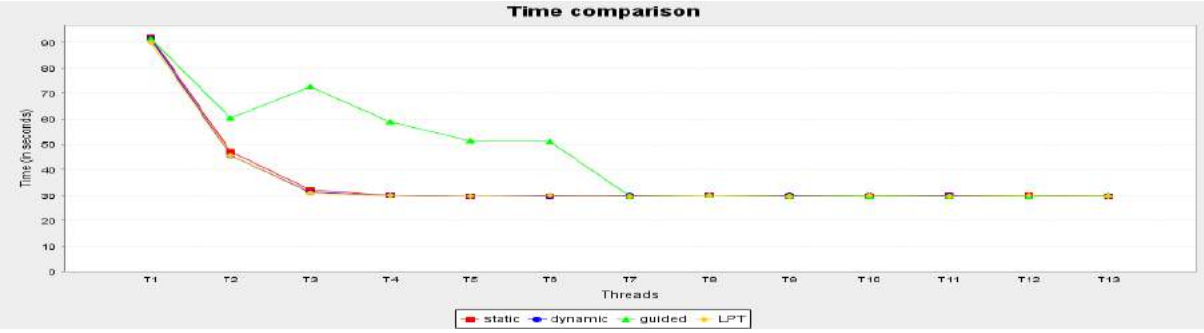


Figure 7: Comparison of total execution for enumerating C(24,12) combinations with modified PAGC2 using static, dynamic, guided and LPT scheduling

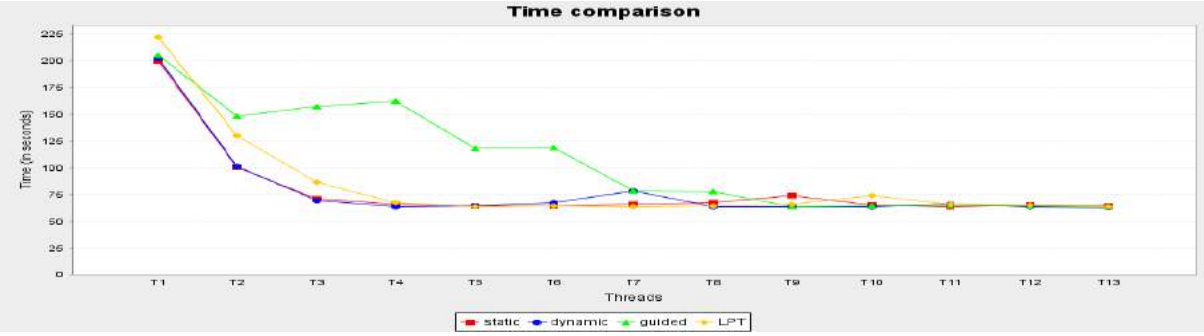


Figure 8: Comparison of total execution for enumerating C(25,13) combinations with modified PAGC2 using static, dynamic, guided and LPT scheduling

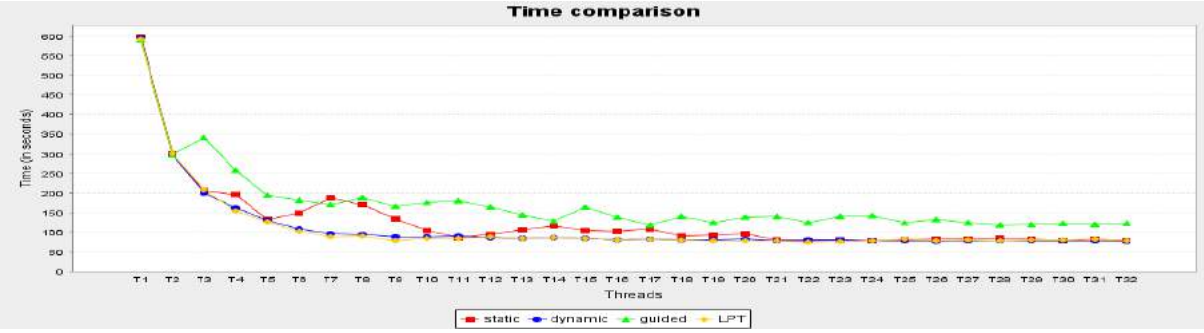


Figure 9: Comparison of total execution for enumerating C(30,9) combinations with modified PAGC2 using static, dynamic, guided and LPT scheduling

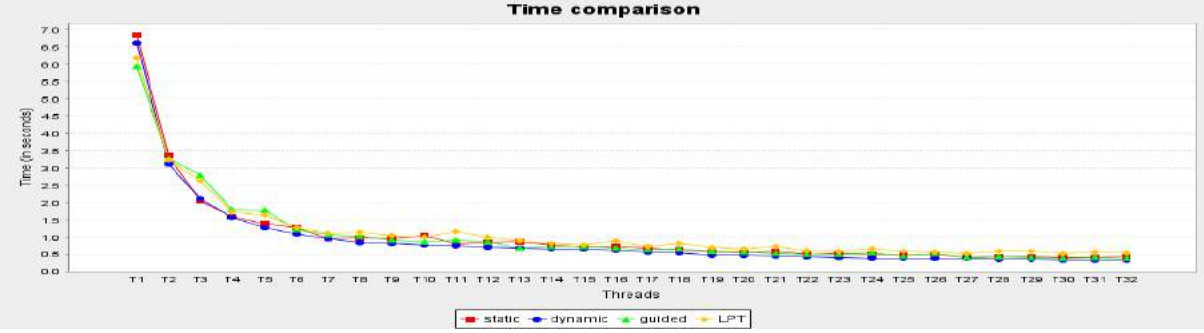


Figure 10: Comparison of total execution for enumerating C(100,3) combinations with modified PAGC2 using static, dynamic, guided and LPT scheduling

Figures 3-6 shows the execution time comparison between different scheduling using modified PAGC1. In modified PAGC1, dynamic scheduling is more efficient than other scheduling in all cases. Dynamic scheduling shows the steadiest results as the number of the processors increased. On the other hand, guided scheduling shows different results based on the number of processors. Static scheduling works better because guided scheduling works well only if the poor load balancing occurs toward the end of the computation.

Figures 7-10 shows the execution time comparison between different scheduling using modified PAGC2. LPT scheduling shows the best result in modified PAGC2. Although dynamic scheduling shows similar results, since LPT scheduling is static scheduling type, it will produce the same result but dynamic scheduling will show slightly different results for each run. Since the sizes of the tasks are known in PAGC2 before the combination generations, it is possible to estimate using how many processors can give the best results.

6. Conclusion

Although there are a lot of combination generation algorithms, all of them have strong and weak points. This paper proposes 3 different types of scheduling to improve the performance of the PAGC1 and PAGC2. Performance results show that dynamic scheduling is the best for modified PAGC1 but LPT scheduling is the best for modified PAGC2. The advantage of modified PAGC2 is that with the help of LPT scheduling, there is the possibility to estimate the minimum number of processors to get the best result.

REFERENCES

[1] Selim G. Akl., — The Design and Analysis of Parallel Algorithms, ISBN 0-33-230056-3, 2014.

[2] Torres M., Goldman A., Barrera J., — A parallel algorithm for enumerating combinations; Proceeding of the 2003 International Conference on parallel processing.

[3] Becky Chan and Selim G. Akl., — Generating combinations in parallel, BIT; Vol. 26, No. 1, 1986, pp. 2-6.

[4] Selim G. Akl., — Adaptive and optimal parallel algorithms for enumerating permutations and combinations; The Computer Journal, Vol. 30, No. 5, 1987, pp. 433-436.

[5] Sergey NOVIKOV., — Parallelization of computations for generating combinations; STUDIA INFORMATICA, Systems and information technology, Volume 1-2 (21), Wyd. UPH, Siedlce, 2017.

[6] Knott, G. D., — A numbering system for combinations; Communications of the ACM, Vol. 17, No. 1, January 1974, pp. 45-46.

Implementation of Push-based Log-transfer Replication System

Aung Chan Myint, Khaing

University of Computer Studies, Yangon

aungchanmyint18@gmail.com, khaing@ucsy.edu.mm

Abstract

Replication is important in distributed environment. Replication gives clients their own neighborhood duplicates of information. These nearby, updatable information duplicates can uphold expanded limited preparing, diminished organization traffic and simple versatility. However, the significant detriment of replication is that when a given reproduced object is refreshed, all duplicates of that article should be refreshed: the update spread issue. The push-based log-transfer replication system does not allow replica contents to become stale. This system will become more productive and accessible than customary replication frameworks that keep all the copies reliable, particularly when the organization and PCs are problematic. Therefore, this system makes to reduce the update propagation data amount in the push-based multi master optimistic replication system by using the log transfer approach. This system is emphasized on the Furniture Ordering Management System. This system is implemented by using C#.Net language and Microsoft SQL server for database engine.

Key words: ***push-based log-transfer, replication***

1. Introduction

Distributed database implies that a solitary application should have the option to work straightforwardly on information that is spread across a wide range of data sets, overseen by a wide range of DBMS, running on a wide range of machines, upheld by a wide range of working frameworks and associated together by a wide range of correspondence network where the term straightforwardly implies that the application works from a sensible perspective as though the information were totally overseen by a solitary DBMS running on a solitary machine. A conveyed information base framework comprises of an assortment of locales, associated together through correspondence organization. *Nowadays, most businesses are organized with distributed pattern. For instance, main office as server site and branch offices as slave sites. So, data about the organization becomes necessary to store redundantly. Moreover, information about the organization needs for autonomy to all offices. It spends many times for informing changes in main office to all branches and vice versa. So, this proposed system supports for organization to take the most appropriate action informing the changes in main office to all branches and actions in the branch offices to main office within a minimum of time[1].*

Informing formally is often made based on human resource and telephone rather than on the update propagation methods in replication. This practice leads to save time and human resources. The integration of organization actions with computer-based update propagation methods would improve performance, increase availability, reliability and to make it fault tolerance.

The objectives of the proposed system are to implement replication system by using push-based log-transfer and multi-master method.

2. Related Work

The related works of replication are discussed in this session.

Replication [4] is normal way to deal with accomplish adaptation to non-critical failure in an appropriated framework to such an extent that imitations give excess if there should be an occurrence of a disappointment of a worker. In this system, banking system is implemented by using active replication strategy. All replicas do the same process stages with same order for both read and write operations so that all replica have been data consistence any time. Therefore, the user can get up-to-date reliable information from this system.

Concert ticket selling system [5] is implemented by passive replication protocol. Three passive based replicas are used to build this system as a reliable system. All replicas (Site1, Site2 and Site3) do the same read and write operations so that all replicas have been updated. This information can be browsed by the registered clients using Web browsers and search engines. Therefore, the user can get up-to-date item prices from this system. They can also make selling and buying via this reliable system.

In the active replication technique [6], also called the state-machine approach, each replica handles the solicitations got from the customer, and sends an answer. The dynamic replication method necessitates that the summons of customer measures be gotten by the non-flawed copies in a similar request. This requires a sufficient correspondence crude, guaranteeing the request and atomicity property. This crude is called all out request multicast or nuclear multicast. As such, the copies act freely and the method comprises in guaranteeing that all reproductions get the solicitations in a similar request. This framework is executed as web based financial framework by utilizing dynamic replication.

3. Background Theory

3.1 Single-master and Multi-master System

Single-master frameworks assign one reproduction as the expert that stores the definitive duplicate of the item. All updates are acknowledged first on the expert and are then engendered to different reproductions, or slaves. Multi-ace frameworks let any copy issue and update whenever, and they trade and union updates among copies.

The principle favorable position of single expert framework is the straightforwardness. Since refreshes are acknowledged distinctly at one spot, single-ace frameworks can identify and report update clashes to clients promptly, making the framework less befuddling to clients. They are easier algorithmically too, on the grounds that updates stream only single direction, from the expert to the slaves.

The multi-master framework are moved accessible than single-ace frameworks. Notwithstanding, the impediment of multi-ace framework is lost update issue: they may lose the impacts of certain updates, since update clashes are recognized after the updates are acknowledged by imitations and the clients who gave them have since quite a while ago logged out from the framework [6].

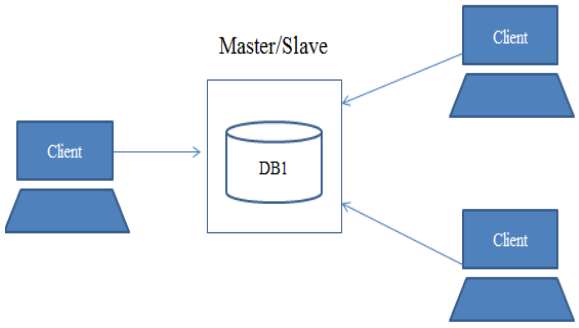


Figure1: Single Master Architecture

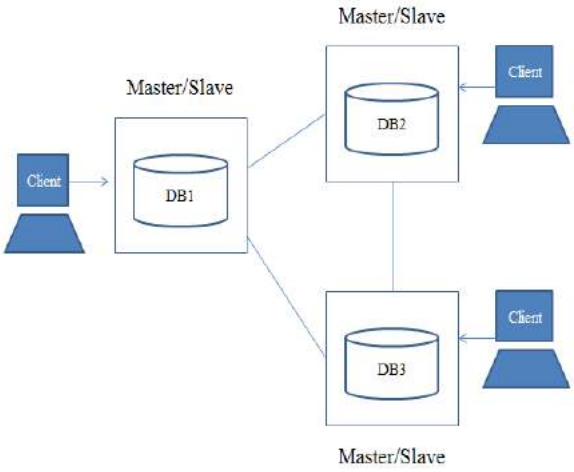


Figure2: Multi Master Architecture

3.2 Pull-based and Push-based approach

Replication is significant in conveyed climate. Replication gives clients their own nearby duplicates of information. These nearby updatable information duplicates can uphold increment confine preparing, diminished organization traffic, simple versatility and less expensive methodologies for circulated, relentless handling. In spite of the fact that, the neighborhood updatable information may prompts tackle the update proliferation issue in replication framework and to limit the quantity of update traded between reproductions. There are two kinds of update proliferation methods: Pull based and Push based update engendering.

Pull-based methodology makes every imitation answerable for surveying different copies and downloading new updates, while push-based methodology makes a copy liable for conveying the update to different reproductions. Pull-based frameworks never send a similar update to same imitation twice, in light of the fact that the arrangement of updates to be gotten is resolved definitely through surveying. Every reproduction just necessities to keep track just of its own state in draw based frameworks, in light of the fact that the condition of other copy is acquired through surveying.

3.3 Log-transfer and Contents-transfer System

A change to an item is communicated either by the new article substance or by its (regularly semantic) depiction ("log"). Framework that trades substance is called substance move frameworks, though frameworks that trade log are called log-move frameworks. Substance move framework moves the whole information base substance to different copies, though a log-move just a depiction of the update. At the point when an article is refreshed at a copy, a substance move framework would move the whole information base substance to different copies, though a log-move framework would move just a portrayal of the update. Log-move claims a few points of interest over substance move. To begin with, log-move can deal with update clashes all the more deftly, particularly in multi-ace frameworks. Second, log-move decreases both the computational and the systems administration overhead, particularly when the item size is enormous and update is little.

4. Proposed Push-based Log-transfer Replication System

A gathering participation administration deals with a gathering of cycles and depends on the view which is the rundown of cycles having a place with a gathering. View change should be advised to all individuals. There are three fundamental tasks expected to oversee bunch enrollment successfully; join, leave and bar. Join is executed by a cycle p endless supply of it, the entirety of the cycles update their view. All the more critically, the condition of the gathering should be moved to the new part p. A cycle will be eliminated from a gathering by avoidance if its accident is recognized by an

individual from a gathering and exit is a willfully arrival of a cycle from a gathering without anyone else.

The gathering the executives module ought to likewise give the two natives; send multicast to make an impression on all individuals and get multicast to get a message sent by an individual from the gathering.

The traditional replication systems that keep all the replicas consistent, therefore update propagation problem arise. Push-based Log-transfer replication system does not allow replica contents to become stale. They can propagate update among replicas for timely access from any replica. In this replication system, multi master system is used to accept and apply changes. For sending the updated data to replicas, push-based system is used. This system uses the log-transfer method to inform the changes to all replicas.

4.1 The System Overview

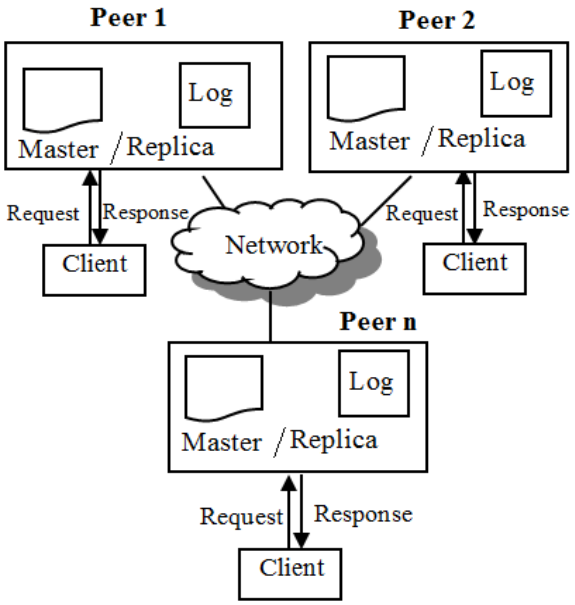


Figure 3: The System Overview

In this system, each replica stores a timestamp (ts) show the last time its contents were modified. When the user request arrived to one of the replica, this replica performs the user request. After performed the user request, update the replica's timestamp and updated request are sent to the master site and then propagate these update to all other replicas. In this propagation phase, the updated data is filtered by the log transfer approach. So, this system no needs to replicate the whole content as transfer.

4.1 Proposed Push-based Log-transfer Algorithm

The data consistency and multiple transactions controls are mainly managed by Push-based Log-transfer Algorithm. The algorithm processing steps are shown in following:

```
Let   L = log contents storage;
R = user request (Read/Write);
BEGIN
        R ← Accept the user request;
        Check R in L;
        If (R  is conflict in L's content)
        {
                Wait ( );
        }
        Else
        {
          L ← store information of R;
          Process(R);
          Commit(R);
          Check the other sites in L;
        If (L has connected sites for committed R)
        {
           Send data updates to respective sites;
           Message "Success for R";
         }

        Else
```

```
        {
          Message “Success for R”;
        }
      End If
  }
End If
```

4.3. Implementation of the System (Furniture Sale System)

In this system, the purchasing products can only be carried by the registered user. So, the user must be registered firstly and then the ordering processing can be carried out. When the user submit the order items to perform the ordering process, the respective server accepts the request, and then the server checks the order item list to query that the item ordered (requested) has sufficient amount in this database. If any value returns from the server sufficient amount for this order request, the server creates update. After this update is successfully completed, the server returns “Transaction is commit” message to the client.

Then, the updated data information are sent to all remaining replica server via the update log content transfer. So, these update approach can avoid copying the whole record in every update process. Only, the new writes are exchanged.

The control log records (shown in figure 4) the transaction status of each transaction to control the concurrency on the same data item processing. By using the logs records in the push base data update propagating processing, each branch in the proposed system can get timely update data. So, this system can maintain the data transparency and reliable to all branches of users as shown in figure 5.

4.4. Benefits of the Proposed System

Every branch can get timely update freshness data, by maintaining the logs records in the push base data update propagating processing. This system can also reduce the transaction abort rate cause for order quantity because the server checks whether the order items are sufficient amount in this database before making transaction processing. As contribution, this system also maintain the transaction expire time to eliminate the dead lock condition. Because the permission accepted transaction can cause delay or block for many reason can lead to deadlock. So, the delay or block transaction must remove from the transaction processing queue.

5. Conclusion

This system implements the push-based log-transfer and multi master approach for the propagating the updated data and update propagation on time. In this system, to maintain the consistency among replicas, use the log-transfer method. If any changes have occurred at the master site, send the log information to all replicas instead of content information. Thus the amount of information to be sent on the network traffic will reduce. In order to reduce the update propagation, use the push-based method. This framework is helpful when the quantity of reproductions is enormous and when the repeated item is huge and update is little.

References

[1] Dr. George Schussel, "Replication, the Next Generation of Distributed Database Technology", *George Schussel research paper*.

[2] Mourad Amad LaMOS, “A Log Based Update of Replicated Profiles in Decentralized Social Networks”, Research Unit of Bejaia University Bouira University, Algeria, 2018.

[3] F. Pedone, “Understanding Replication in Databases and Distributed Systems”, Operating

Systems Laboratory, Swiss Federal Institute of Technology (EPFL), 2014.

[4] Khin Kaung San, "Implementation of Database Consistency by Active Replication", M.C.Sc 2016, *University of Computer Studies, Yangon.*

[5] May Phuu Ko, "Concert Ticket Selling System by Passive Replication", M.C.Sc University of Computer Studies, Yangon, 2017.

[6] F. Schneider, **"Replication management using the state machine approach,"** in *Distributed Systems (S. Mullender,* ed.), ch. 7, pp. 169–198, Addison-Wesley.

[7] Marius Cristian, "**Database Replication",** MAZILU Academy of Economic Studies, Bucharest, Romania mariuscristian.mazilu@gmail.com, mazilix@yahoo.com, 2015

[8] Kuljeet Kaur, Poonam Sood, Simarpreet Kaur, **"Database Replication Using Eager Replication" ,** International Journal of Advanced Research in Computer Science and Software Engineering Research Paper, 2016.

[9] Thet Maung Maung Zaw, Yin Ko Latt, **"Replication with Materialized views using Distributed Database",** Computer University, Magway, 2017.

[10] Chan Mya Aye, "**Online Stock Tracking System Using Active Replication**", December, 2010, Computer University (Taung-Ngu)**.**

[11] "Correctness of Concurrency Control and Implications in Distributed Databases," Proc. COMPSAC '04 Conf. Chicago, IL.

[12] "Optimistic Methods for Concurrency Control in Distributed Database Systems," Proc. 7th VLDB Conf. 2017 Cannes, France.

Log_ID	Type of Transaction	SaleID	ItemID	Price	QTY	Transaction Start Time	Transaction End Time	Processing Time	Customer Name	Action Perform Site
1	Insert	123	488	549000	1	11-Feb-20 7:59:05 AM	11-Feb-20 7:59:13 AM	00:00:07.826818	Daw Hla Hla	BySite 3
2	Read	-	251	100000	-	11-Feb-20 7:59:18 AM	-	-	-	BySite 1
3	Insert	124	223	300000	5	11-Feb-20 8:01:47 AM	11-Feb-20 8:01:50 AM	00:00:03.752146	Daw Mya	BySite 1
4	Insert	125	210	410000	3	11-Feb-20 8:02:50 AM	11-Feb-20 8:02:55 AM	00:00:05.563214	U Khin Myint	BySite 2
5	Insert	126	159	321000	2	11-Feb-20 8:03:55 AM	11-Feb-20 8:04:01 AM	00:00:06.748214	U Win Aung	BySite 2

Figure4. Log File Structure (Control Log)

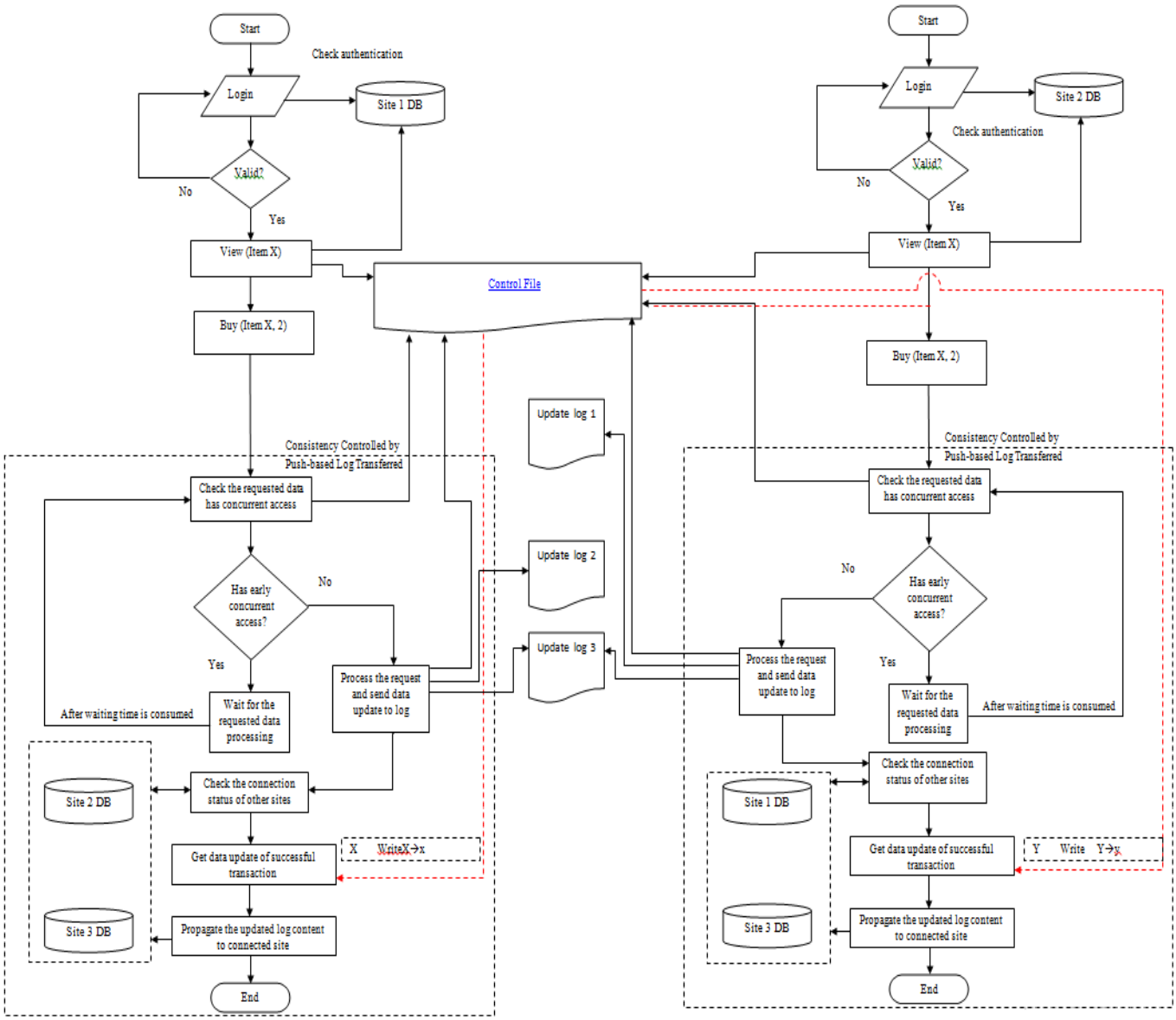


Figure5. The System Flow

Image Processing

Flower Recognition System using Chain Code Method

Wint Sandi Soe, Zin May Aye
University of Computer Studies, Yangon
wintsandisoe.ucsy@gmail.com, zinmay110@gmail.com

Abstract

Nowadays, automatic recognition of flowers using computer technology is of great social benefits. Recognition of flowers has various applications such as floriculture, flower searching for patent analysis and so on. Floriculture industry includes flower trade, seed and bulb production, micro propagation and extraction of essential oil from flowers. For such cases, flower recognition is very necessary step. The proposed system in this paper uses Freeman's Chain Code method by tracking contour information of the flower image to recognize the type of flower. This system consists of image preprocessing, chain code implementation and flower recognition process. In image preprocessing, Grabcut segmentation method, resizing, converting to grayscale image, binarization, filling holes method and normalization are applied. Chain code serves as compact representation of a binary object by a connected sequence of line segments. The chain code sequence of flowers is searched by using eight-connectivity neighborhood method. For the recognition process, predetermined and normalized chain code sequence numbers are compared with the chain code sequence of the input image by computing minimum mean square error. The minimum value of mean square errors is used to determine the recognition result of the system.

Keywords: ***Grabcut Segmentation, Filling Holes Method, Freeman's Chain Code method, Minimum Mean Square Error***

1. Introduction

Nowadays, there are many different types of image recognition systems in the world such as face recognition, finger print recognition, object recognition, character recognition which can actually be developed with the acceptable excellent results. But even the professional programmers attempt to get the more accurate solutions from the recognition systems with complex data.

Object recognition is a difficult problem to handle for the computer scientists due to the numerous challenges. The image of any object taken from different view appears in a different way. For the natural object such as flower, various species of flowers exists in the world. Some of the categories are Anthurium, Daffodil, Frangipani, Orchid, Sunflower, etc. The categorization of flower images is challenging due to variances in geometry, illumination and occlusions. In this system, flower recognition system is implemented by using Grabcut method, Filling Holes method, Freeman's Chain Code method and Minimum mean square error.

The outline of this paper is organized as follows: Section 2 describes Related Works for flower recognition. The proposed system is explained in Section 3. Section 4 presents the experimental result of the system. In Section 5, the conclusion of the system is described.

2. Related Works

Image processing and computer vision can be applied in object recognition and classification problems. Today, we can automatically classify food, car, plant and other items. Research in [1] proposed a flower recognition system basing on color and edge characteristics of the image. Histogram is used to derive hue, green, blue, red and saturation features. Classification is done by using K-nearest neighboring algorithm. Research in [2] was developed by using Random Forest Classifier method. As a preprocessing step of flower identification system, Grabcut method was used to segment the background and the foreground from the input image. The identification of flower name from the input image is done based on RGB Histogram data. Also, research in [3] shows the application of plant species identification technique to identify local fruit trees through leaf structures. The system initially converts RGB to Grayscale image using thresholding algorithm before removing noises. Chain code method is used to obtain the shape of an object in feature extraction process. Finally, the leaf's feature was recognized using Linear Comparison technique.

3. Proposed System

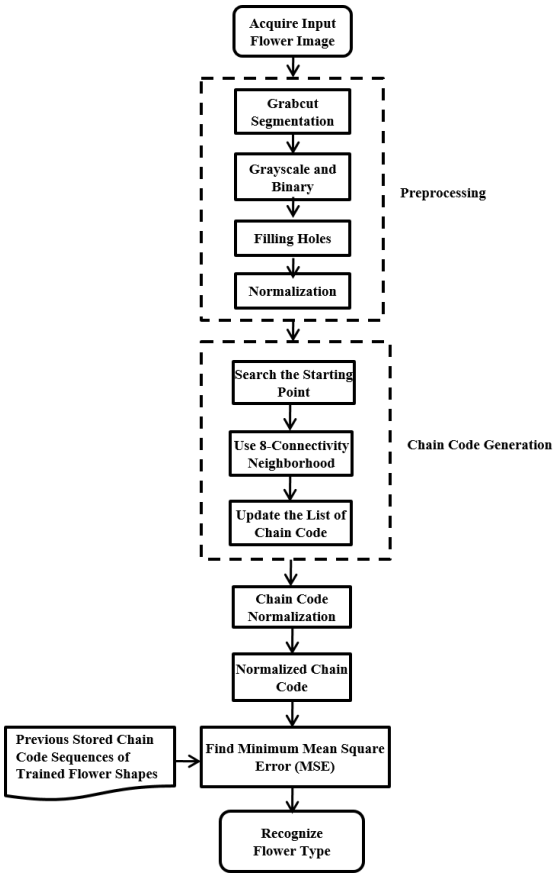


Figure1. System Overview

An overview of the Flower Recognition System is shown in Figure 1. First, the user input an image into the system for recognition. After that, the input image is cut by dragging with rectangle on it including the desired flower object. The Grabcut segmentation method is applied to get the desired foreground flower image. The segmented image is resized as preprocessing for convenience of later processes. The resized colored image is converted to grayscale image. The grayscale image is converted to binary image. Filling holes method is applied to the binary image to fill black holes in white object. The filled image is normalized for standard uniform size. The normalized image is input to chain code implementation. The chain code sequence is generated for the normalized binary level image. After chain code generation, chain code sequence is normalized for fair comparison with predefined chain code sequences. After chain code normalization, in recognition step, minimum mean square error (MSE) is computed from the chain code sequence of the input flower image to the predefined chain code sequence stored in the system. Finally, the recognized flower's name is displayed.

3.1. Grabcut Segmentation Method

The system uses Grabcut segmentation method [4] to extract foreground flower object. Grabcut is an innovative 2D image segmentation technique developed by Rother et al. [2004]. Grabcut is an iterative image segmentation technique based upon the Graph Cut algorithm and consists of statistical models of the foreground and background structure in the color space. In the system, the input image is uploaded, and then the user has to drag a rectangle around a desired flower object and Grabcut segmentation algorithm is applied and finally an extracted flower image appears as an output.

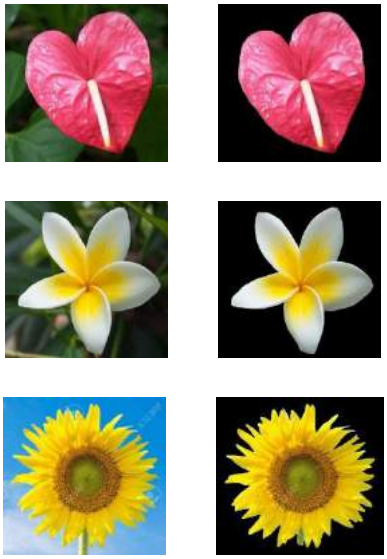


Figure 2: Three examples of Grabcut segmentation

3.2. Filling Holes Method

The filling holes method is used to fill black holes in white object in a binary image. The processing steps of filling holes method are as follows.

1. Read the image.
2. Threshold the input image to obtain a binary image.
3. Flood fill from pixel (0, 0). The difference between the outputs of 2 and 3 is that the background in 3 is now white.
4. Invert the flood filled image (i.e. black

becomes white and white becomes black).

5. Combine the thresholded image with the inverted flood filled image by using bitwise OR operation to obtain the final foreground mask with holes filled in. The image in 4 has some black areas inside the boundary. By design the image in 2 has those holes filled in, combine the two to get the mask.

Figure 3: Filling Holes Method

3.3. Chain Code Method

Chain codes are one of the shape representations which are used to represent a boundary by a connected sequence of straight line segments of specified length and direction. Chain code representation has many advantages. First of all, it is compact. Secondly, the chain code is translation invariant. Thirdly, the chain code can be used to compute many shape features, such as the perimeter and area. However, chain codes are invariant to boundary size and orientation.

Chain code technique is widely used, because chain codes allow considerable data reduction in the description of shapes. Many applications using chain code representation have been reported up to date. The first work on representing digital curves using chain code representation was introduced by Freeman in 1961. It remains the most widely used coding technique. This code moves along a digital curve or a sequence of border pixels based on eight-connectivity. The direction of each movement is encoded by using a numbering scheme $\{i \mid i = 0, 1, 2, 3, \ldots, 7\}$ denoting an angle of $45^\circ \times i$ counter-clockwise from the positive x-axis, as shown in Figure 4 (b). The chain codes can be viewed as a connected sequence of straight-line segments with specified lengths and directions. A four-directional version of the Freeman chain code is sometimes used, too. It moves in four directions with a numbering scheme $\{i \mid i = 0, 1, 2, 3\}$ denoting an angle of $90^\circ \times i$ counter-clockwise from the positive x-axis, as shown in Figure 4 (a). The directions and assigning the number of the shape using four and eight connectivity chain code methods for an image are shown in Figure 5 (c) and (d) as an examples.

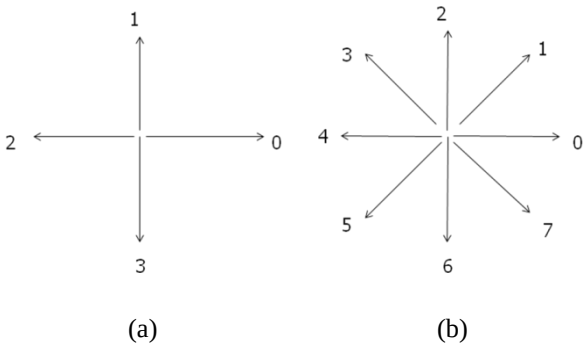


Figure 4: (a) Freeman's Chain Code in Four Directions and
(b) in Eight Directions

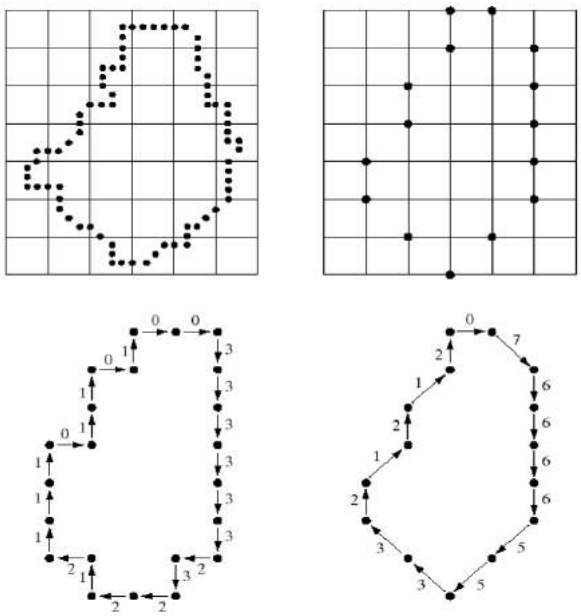


Figure 5: Examples of Chain Codes
(a)(Top-Left) Object Boundary
(b)(Top-Right) Boundary Vertices
(c)(Down-Left) 4-Directional Chain Code
(d)(Down-Right) 8-Directional Chain Code

3.4. Border Tracing Algorithm

The border tracing algorithm is used in chain code method to extract the contours of the objects from an image. When applying this algorithm, it is

assumed that the image with regions is either binary or those regions have been previously labeled. Algorithm's steps are as follows:

1. Search the image from top left until a pixel of a new region is found; this pixel P_0 is the starting pixel of the region border. Define a variable dir which stores the direction of the previous move along the border from the previous border element to the current border element. Assign (a) dir = 0 if the border is detected in 4-connectivity (Figure. 6 a) (b) dir = 7 if the border is detected in 8-connectivity (Figure. 6 b)
2. Search the 3x3 neighborhood of the current pixel in an anti-clockwise direction, beginning the neighborhood search at the pixel positioned in the direction (a) (dir + 3) mod 4 (Figure. 6 c) (b) (dir + 7) mod 8 if dir is even (Figure. 6 d) (dir + 6) mod 8 if dir is odd (Figure. 6 e) The first pixel found with the same value as the current pixel is a new boundary element P_n. Update the dir value.
3. If the current boundary element P_n is equal to the second border element P_1 and if the previous border element P_{n-1} is equal to P_0, stop. Otherwise, repeat step (2).
4. The detected border is represented by pixels $P_0 \ldots P_{n-2}$.

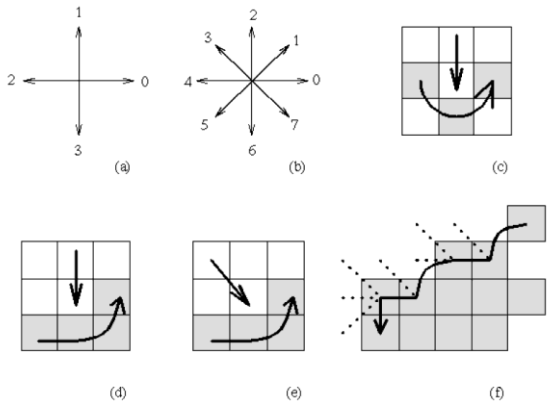


Figure 6: Border Tracing Algorithm
(a) Direction notation, 4-connectivity
(b) 8-connectivity
(c)Pixel neighborhood search sequence in 4-connectivity
(d)(e) Search sequence in 8-connectivity
(f) Boundary tracing in 8-connectivity
(dashed lines show pixels tested during the border tracing)

3.5. Chain Code Implementation

More and more images have been generated in digital form in today's world. In order to find an image, the image has to be represented or described by certain features. Shape is an important visual feature of an image. In this paper, shape representation is described by using chain code technique. The generation of chain code applied in this system is shown in Figure 7.

After image preprocessing, the binary image is applied to generate chain code sequence by using chain code algorithm with eight-connectivity. The direction and assigning the number of the shape of chain code are shown in Figure 4 and 5. The algorithm for the chain code generation can be considered as the flow diagram shown in the figure 7.

Chain code generation steps are as follows:

Step 1: All pixels in the image are found as binary level image. For the image, white color represents 1 and black for 0.

Step 2: The starting point was searched starting from the minimum 0 pixel value's element in matrix form of image at the column and the row. This coordinate was stored as variable of starting point in the chain code programming.

Step 3: A variable of direction was set to 7 because of 8 directional neighborhoods which are 0 to 7.

Step 4: Transverse the 3*3 neighborhood of the current pixel in a counter-clockwise direction. This will set the current direction counter-clockwise from the new direction.

Step 5: The direction transverse will update in variable chain code.

Step 6: This process will loop until the direction point of boundary get the same coordinate with starting point.

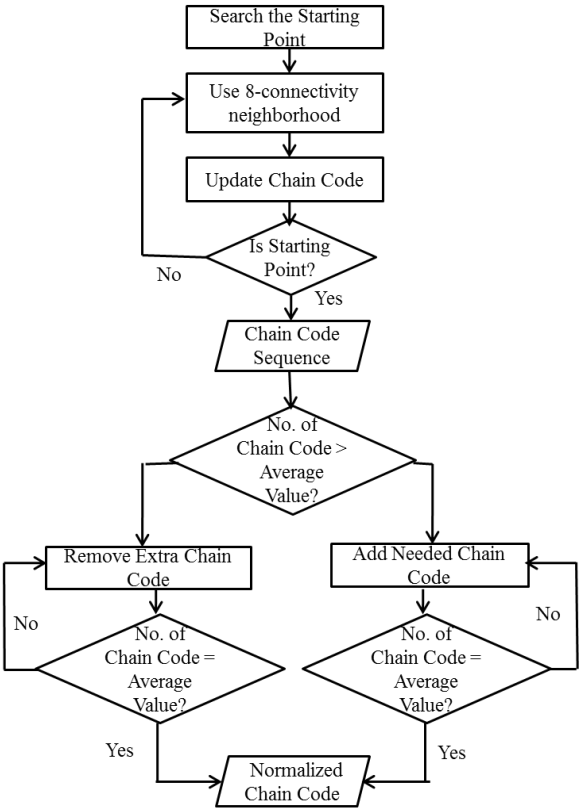


Figure 7: Generation of Chain Code

As an example chain code sequence for the input flower is

000000000000670066766666766666666665 66666666666766.

In this sequence, 50 chain codes are included. The predefined average for that flower is set to 55. The number of chain code sequence is less than the average value and it is needed to add to get the same number as the average value. The position of the added code is between the repeated code and the resulted sequence will be

000000000000067000666766666676666666666566 666666666766.

If the number of chain code is greater than the predefined value, the repeated code in the sequence can be removed from the original chain code sequence. In this system, the predefined number of chain code sequence is set to 450 as an average value.

3.6 Mean Square Error

This system will be implemented to recognize flowers by using chain code technique with minimum Mean Square Error (MSE). The mean square error (MSE) of an estimator is one of many ways to measure the difference between values implied by an estimator and the true values of the quantity being estimated. The chain code from the input flower and also the text file are used to compute mean square error (MSE). The absolute different value is stored and the minimum value at each element is determined as a result to the recognition which flower. The MSE equation is as follows.

$$\text{MSE} = \frac{1}{N}\sum_{i=0}^{N}(x_i - y_i)^2 \qquad (1)$$

where MSE is mean square error, x is the chain code sequence from the input flower, y is the chain code sequence from the text file that is previously trained and N is the number counts of chain code sequence.

5. Experimental Result and Discussion

The final result form of the proposed flower recognition system is as shown in figure 8.

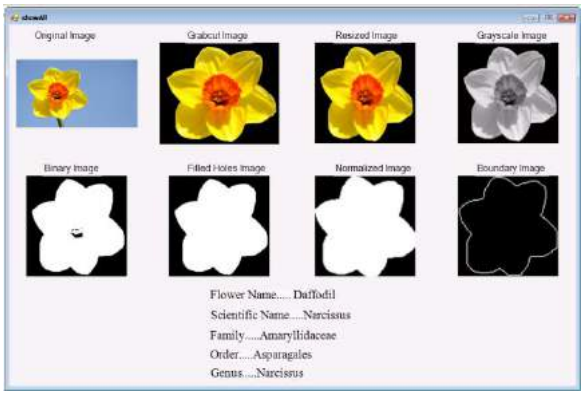


Figure 8: Final result form of flower recognition system

For training process, 100 flowers pictures for each type are used to be trained in the first time. The experiment is conducted on 5 different species of flowers as shown in Table 1. For testing, flower images for each type are tested in three conditions: complete flower image, 80% cropping flower image and 60% cropping flower image. According to experimental results, the accuracy rate of recognition result is decreasing when the input flower image does not include its whole complete parts. For complete flowers, the average recognition accuracy rate reaches to 90%.

There are some limitations in using Grabcut method. When the user inputs the rectangle on the

flower image, everything outside this rectangle will be taken as sure background. That is the reason that the rectangle should include all the foreground objects. And, Grabcut segmentation method is more suitable for the image in which foreground object and background are not complex.

Therefore, this system is intended to be effective in flower recognition for non-noisy flower images, front-view flower images and non-camouflage flower images. As further extension, the flower recognition systems with flower images which includes noises and in different directions and with more various types of flower can be implemented.

Table 1: Experimental Accuracy Result

Input flower image	80% crop image	60% crop image	Accuracy for Testing		
			For Com-plete image	For 80% crop image	For 60% crop image
			92%	54%	24%
			94%	80%	50%
			90%	70%	28%
			88%	66%	26%
			86%	50%	22%
Average			90%	64%	30%

6. Conclusion

This paper presents a system that recognizes flower from the input image. Grabcut is used in the segmentation phase to segment foreground (flower) from the background. Filling holes method is used to fill black holes in white object in the flower binary image. Chain code method is used to describe contour information as number sequence. For the recognition, minimum mean square error method is used in the system. It is a simple method to compare the two chain code sequences, one from the trained flowers and other from the test flower to achieve minimum mean square error used to determine the flower recognition. With 5 types of input flowers, Anthurium, Daffodil, Frangipani, Orchid and Sunflower, this system was able to recognize the flowers with average 90% accuracy.

According to testing experiences, for better performance in Grabcut segmentation, the user should draw a rectangular frame to fit with the flower. The user should try not to overlap the frame with bigger part of background because the algorithm would consider the background as the same type with the foreground object (flower). Also, the input flower should be clear and good quality, with different foreground and background color.

References

[1] Tiay T., Benyaphaichit P., and Riyamongkol P. (2014) Flower Recognition System Based on Image Processing. The 2014 Third ICT International Student Project Conference

[2] Warisara Pardee, Prawaran Yusungnern, Perayaripian, Flower Identification System by Image Processing (Using Grabcut method and RGB histograms), Media Technology King Mongkut's University of Technology Thonburi Bangkok, Thailand, August 2015.

[3] V., P. A., and V.K., S. Digital Image Processing Approach for Fruit and Flower Leaf Identification and Recognition. International Journal of Engineering And Computer Science 2(2013).

[4] Carsten Rother, Vladimir Kolmogorov, Andrew Blake Grabcut Method - Interactive Foreground Extraction using Iterated Graph Cuts. Microsoft Research Cambridge, UK,2004

[5] Pulipati Annapurna, Sriraman Kothuri, Digit Recognition Using Freeman's Chain Code Method, Srikanth Lukka, August 2013.

[6] Hussein El Saadi, C# Project in Optical Character Recognition (OCR) Using Chain Code, Feb 2011.

Human Action Recognition based on Motion Detection

Aye Aye Aung, Thiri Naing
Computer University (Kalay)
ayeayeaung@ucskalay.edu.mm, thuthiri@gmail.com

Abstract

System surveillance is the process of monitoring the behavior of people in the system. Recognizing human meaning is automatically realizing what actions a person makes on a video. In this paper, human action recognition based on extracted motion features by computer vision techniques is proposed. The main objective of this work is to recognize three human actions walking, running and bending. The proposed system can be classified into three main processes: human-motion detection, human-silhouettes extraction and classification. First, the sequence of the image is tracked then human silhouettes is extracted by the frame differencing method from the walking figures. In the experiment, sparse representation-based classification (SRC) is applied for classification. The result of the proposed system gives good enough performance for human-action identification.

Key words: human-silhouettes, sparse-representation classification (SRC), human-action recognition.

1. Introduction

The visual surveillance, video-based content recovery, human-computer interaction and sports annotation is the potential applications for human action recognition system. It is also popular research area. By successfully recognizing human activity, the surveillance system in a large public area can automatically generates advanced semantic information from surveillance video. And it can alert the public if it detected the suspicious action at a predetermined distance. The motion object detection is a significant task undertaking the image sequences of the surveillance area.

Detection of moving objects is the first step for object recognition system [1,2,3]. The purpose of object detection is to extract moving objects that are interest to make decision in input video sequences. The motion detection can be divided into three categories. It includes optical flow method, frame differencing method and background subtraction method. Each has its own strengths and weakness. In this system, the frame differencing method is used to detect the moving objects. Over the last few years, sparse representation has received much attention and it can be found in different fields.

This paper's structure is as follow: Section 2 is some related works of human-action recognition. The overview of the proposed system is described in Section 3. And Section 4 presents the details of the system including motion detection, motion features extraction and motion analysis. The experimental result of human action recognition system based on SRC classification techniques is discussed in Section 5. The conclusion part is in Section 6.

2. Related Works

Monitoring applications often find it difficult to obtain facial or visual information because it requires a high resolution to be recognized. In many things, human play an important play in the everyday activities of everyday life. Therefore, recognition of basic human actions is an essential component of many important applications. Finding unusual activities, such as jumping, running can provide alarms in time for enhanced security. Recognition of automatic human action is described in [4]. It illustrates the way in which view-invariant method for automatic recognition of human actions by the use of motion activities.

In study [5] it showed the behavioral walking characteristic can use to detect suspicious persons when they entering the monitoring center without permission. The video database is used side by side view in this system to evaluate. In [6], it shows the approach that combines information of the face and gait. Facial and gait features were obtained separately. The principal component analysis (PCA) and gait energy image (GEI) are used. The result is shown that the synthetic features of encoding both side of face and gait information.

Object detection from the video sequence is the task for all visual analysis based on visual monitoring system. It separates the motion objects from the background objects. From the statistical point of view, background is extracted using the Gaussian model [7] and adapted to handle the changing visual environment. In [8], it controls the point for studying the applied background extraction algorithm using Gaussian Mixture Model. The method is adapted the rate adjustments in background of the video event. The result shows that the comparison of well-known methods over the proposed system's methods. A high order spectral analysis was conducted in [9] to detect

people by recognizing such movements as running or walking.

The approach [10] provides a theoretical framework for distinguishing signals with sparse representation. It can robust to signal corruptions, noise and missing data. These experimental results have been shown to be a potential achievement for SR in signaling. In the field of pattern recognition, Sparse representation-based classification (SRC) has received much of attention. In [11], it used gene expression data to propose a new classification for tumor classification based on sparse representation (SR).

3. Overview of the System

This system can identify suspicious and unauthorized person entered to a surveillance area. It is the main part of monitoring things. Firstly, extracts the moving objects from the input video sequences. From the moving region, the corresponding silhouette of the walking figures is successively tracked through the simple correspondence method. In this system, simple frame difference background subtraction method [12] is used.

The next step of the system is motion pattern extraction on the sequence of silhouette figures for different action recognition. Different types of action are Walking, Running and Bending. The human silhouettes and static motion features (height and width) are extracted as the motion parameters.

Finally, the sparse representation-based classifier [13, 14] is applied to the extracted motion features as classification features to evaluate the discriminant ability. SR has been used successfully in human identification fields. Overview process of the proposed system is shown in Figure 1.

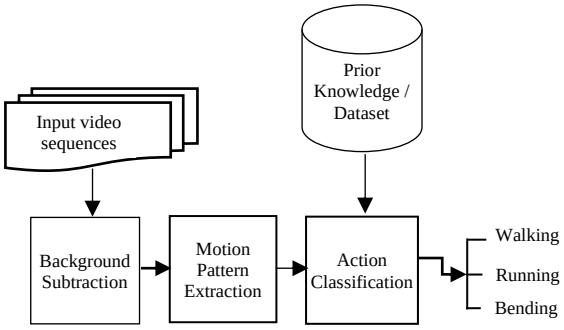


Figure 1. Block Diagram of the proposed system

4. Proposed System

In this system, Human action recognition is proposed on the basis of automated features of computer vision techniques. Recognizing a human being is defined as automatically computing what a person does in a video. The proposed approach captures human silhouettes and static movement characteristics (height and width). Firstly, the walking figure in an image sequence is detected and tracked by background subtraction method. The silhouettes are extracted by a simple motion detection method. Input video is captured with a static camera.

Human action classification is tested on the threshold level value 15 in moving objection detection. Width and height and velocity of the moving object is used to identify the three different actions in classification phase. To evaluate the proposed classification method, the SRC classification scheme is implemented based on the sparse representation.

4.1 Motion Detection and Human Silhouettes Extraction

No advance information is needed to detect moving objects from video sequences. However, only the multiple consecutives frames of the video are required. Today, video surveillance systems play an important role in the field of safety and security. It is useful for finding and examining objects that are really moving at all places, such as airports, shopping mall, hospitals and residence. The estimation of pixel properties of background and foreground are fragments from the background of each frame.

4.1.1 Frame difference background subtraction

The principle of this method is to use a model of the background and compare the current image with a reference. It has been used for years in many vision systems such as video surveillance, teleconferencing, video editing, and human-computer interfaces. It calculates pixel gray scale difference between adjacent two frames in a continuous image sequences and determines foreground by setting threshold. The input signal is changed into a grayscale with the rgb2gray function. Therefore, it can reduce computer time and memory storage.

Frame differencing technique does not require any knowledge about background and is very adaptive to dynamic environments, but suffers from the problem of foreground aperture due to homogeneous color of moving object. It is depended on the threshold value so it must be set accordingly. At the very low and very high threshold, background subtraction results are poor thus achieving low correct classification rate. Threshold value can be varying 10,15,25 and 50 etc. for various background.

In threshold level 10, it is very low and the results of background subtraction is poorest for all subjects as the background subtraction algorithm estimated the background value incorrectly. In threshold level 40,

background subtraction result is poor as it may be loss of the foreground information. Some important foreground information is removed together with the background values from the input frame because the background subtraction algorithm assumes this foreground information as the background value. Moreover, due to the intensity quality of foreground is lost, the identification results are decreased in the experiment.

In this system, the threshold value 15 is used because it can be the good accuracy for the classification. SRC classifier is applied in order to test the identification results. For the moving detection, the two images (adjacent frames in a video sequence) are compared and the differences in pixel values are determined.

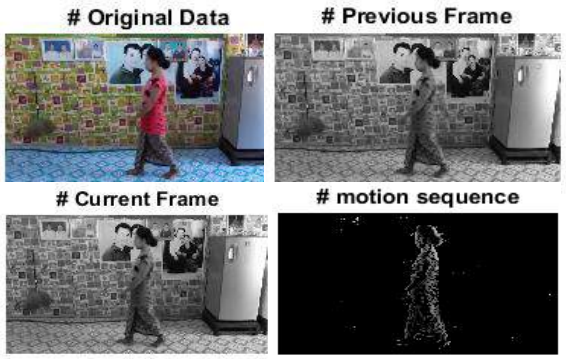


Figure 2. Background Subtraction Results

If the difference between the two is greater than the specified value of T_h, the pixel is part of the foreground. If not, take it to background mode.

$$|Img(t) - Img(t-1)| > T_h \qquad (1)$$

where $Img(t)$ is the current frame image, $Img(t-1)$ is image of the previous frame, and T_h is the threshold value. The detection foreground of the different signal frames is shown in Figure 2.

4.2 Motion Pattern Extraction

In this step, calculate the width and height of the human silhouette extracted from background subtraction. To identify the **Bending**, it is calculated the width and height of the human silhouette from each frame during the walking sequences. And also, Bending can determine by the center coordinate. The center coordinate of each human silhouette is calculated from two border points as [15]:

$$x_{center} = x_s + (x_e - x_s)/2 \qquad (2)$$

where x_s represent first pixel and x_e represent the end pixels on the vertical line. If the height of the silhouette is reduce to the under of x_{center} , it may be bending. The width can be defined by (x_e - x_s) on the horizontal line of the human silhouettes. And, the height value is also (x_e - x_s) on the vertical line of the human silhouettes.

To identify the ***Running***, the gait velocity V_i at frame *i* is calculated by:

$$V_i = (x_i - x_{i-1})/\Delta t \qquad (3)$$

Where, Δt = 1/25 (frame rate), the video is captured and played back at 25 frame per second, that means each second of video shows 25 distinct still images.

Walking: the height value of the silhouette is same and the velocity V_i value is normal.

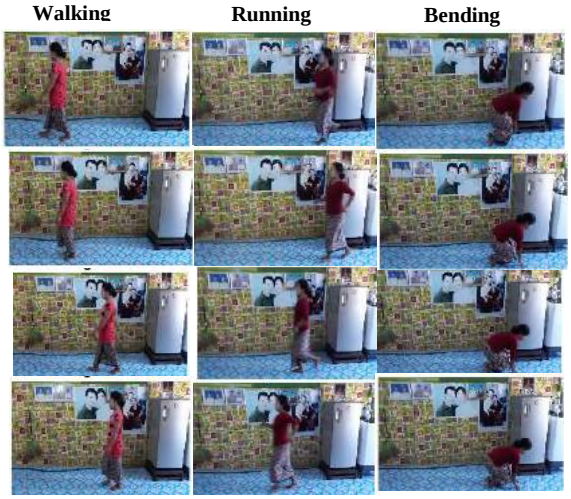


Figure 3. Three Types of Human Action

4.3 Sparse Representation Classifier (SRC)

Sparse Representation-based Classification (SRC) has received an increasing amount of attention, and it has been successfully used in the classification of various visual signals including facial expressions, hand written digits, and general images There are two key points, first, the query signal/image is collaboratively coded over a dictionary of atoms with some sparsity constraint, and then classification is performed based on the coding coefficients and the dictionary. The dictionary for sparse coding could be predefined and it directly used the training samples of all classes.

The SRC is a new and powerful classifier especially for noisy or corrupted samples data. Thus, the aim of the system is to develop a robust and

reliable human action recognition system which can work well with noisy background.

The basic idea of SRC [12] is that:

- Compute the sparse decomposition of all test samples.
- The residual error for reconstruction of the test sample is calculated by sparse decomposition coefficient.
- The test sample is classified into the class which remains the minimum reconstruction residual error.
- It can be brief as following:
  1. Input: matrix of training samples for c classes; testing sample
  2. Normalize the training matrix D to have unit l_2 norm
  3. Compute the coding vector c of y over D:
     $\delta = argmin\|y - Dc\|$ (4)
  4. Compute the approximate residuals:
     $$e_k(y) = \|y - D\delta_k(x)\|_2 \quad for\ k = 1,...,c \tag{5}$$
  5. Output: the class of the given test sample y is determined by
     $$identity(y) = \arg\min_k \{e_k\} \tag{6}$$

In SRC algorithm, there are two parts. The first one is that the coding vector of the sample y needs to be sparse. The second is the coding of y, the co-operation of the entire set of X values of each group.

5. Experimental Setup

In order to test the system properly, it is desired to train and test the system using video of actual people walking in front of a camera. Data collection is captured the gait videos from the side view of a person using a static camera. All of the input video data are converted to the '.avi' format and then these video clips are equally split into each of size about 5 secs to analyze the system performance.

Figure 4. Human Action Recognition System

The motion was filmed by means of a video camera at 25 frames/sec of frame rate and the video resolution is 320 by 240 pixels. The experimental database, 150 videos of 20 people (10 males and 10 females) is used to calculate the performance of the system. The video clips of 5 people are used to test the system end result.

In human action recognition, the recognition rate of the proposed system is presented with the threshold value 15 in background subtraction as shown in Figure 2. The sparse representation-based classification methods of SRC is tested on this system. The results shown on each types of action (walking, running and bending) database separately.

Table 1. The accuracy of the human action recognition on 15 persons (threshold = 15)

Person #	Walking	Running	Bending
P1	0.8251	0.6626	0.7527
P2	0.7752	0.6785	0.7238
P3	0.7935	0.7074	0.6289
P4	0.7458	0.7182	0.7051
P5	0.7684	0.6784	0.8352
P6	0.7924	0.7224	0.7935
P7	0.7457	0.7457	0.7458
P8	0.7927	0.7527	0.7684
P9	0.8238	0.6238	0.7985
P10	0.7289	0.7859	0.8253
P11	0.7977	0.7051	0.7960
P12	0.8626	0.6752	0.8226
P13	0.7385	0.6935	0.7853
P14	0.7774	0.6458	0.8047
P15	0.7682	0.6684	0.7826

In the database, the number of objects is contained 50 videos for walking, running and bending collectively. Using cross-validation, the proposed system is tested on three. Table 1 show the experimental results of the system. According to the result in table, the proposed classification method SRC achieved the best performance with 86% in walking. The best classification rate for running is 78% and 83% for bending. Among three types of actions, the best recognition rate is testing on the walking for all people. Walking is the better accuracy than the running and bending. The classification rate of running is the poorest among the others in this system but it is still reasonable results. The experimental evaluation of the system confirms the good enough for the performance of all types of action.

6. Conclusion

Human action recognition at a distance has gained more interest in visual surveillance systems.

Therefore, the proposed system architecture is simulated on three types of actions by static camera. This system can identify a person's activity when entering a monitoring area, which is an important part of visual surveillance systems. The experimental results of the system are good enough for the performance of all types of action. According to the experimental results, walking is the best as high as 86% classification accuracy for all types of action. As a limitation, the system only made video clips with one person. The database used does not include any external images that are in a complicated background state. For the further works, the system will be calculated the performance on the large dataset with other types of actions (fighting, clapping and falling, etc.) and using the other algorithms, KNN and SVM to compare with the proposed method.

7. References

[1] Jay. P. G, Pushkar. D, Nishant. S, Vijay. B. A, '*Analysis of Gait Pattern to Recognize the Human Activities*', International Journal of Artificial Intelligence and Interactive Multimedia, Vol. 2.

[2] Rajkumari. B. D, Yambem. J .C and Khumanthem.M. S, '*A Survey on Different Background Subtraction Method for Moving Object Detection'*, International Journal for Research in Emerging Science and Technology, Volume-3, Issue-10, Oct-2016.

[3] Ong.C. A, Lau. B.T, '*Human Activity Recognition: A Review'*, IEEE International Conference on Control System, Computing and Engineering, 28 - 30 November 2014.

[4] Zan .G,An-an Liu, Hua. Z, Guang. X, Yan. X, 'Human action recognition based on sparse representation induced by L1/L2 regulations', Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012).

[5] Lakany, H.M., Birbilis, A. and Hayes, G.M.: Recognising Walkers Using Moving Light Displays.

[6] Little, J.J. and Boyed, J.E.: Recognizing people by their Gait: the shape of motion, submitted to Videre, December1996

[7] Lakany, H.M and Hayes, G.M: An Algorithm for Recognizing Walkers.

[8] Cedras.C and Shah, M.: A Survey of Motion Analysis from Moving Light Displays, IEEE CVPR-94, Seattle, Washington, June 20-24, 1994, pages 214-221.

[9] Claudette Cedras and Mubrak shah, Motion-Based Recognition: A Survey.

[10] Johansson, G.: Visual perception of biological motion and a model for its analysis, Perception and Psychophysics, 14(2):201{211, 1973}.

[11] Huang, K. and aviyente, S.: Sparse Representation for Signal Classification.

[12] Chun-Guang Li, Jun Guo and Hong-Gang Zhang: Local Sparse Representation based Classification, 2010 International Conference on Pattern Recognition.

[13] Jort F. Gemmeke, Efficient sparse representation-based classification using hierarchically structured dictionaries.

[14] Rui Min and Jean-Luc Dugelay,: Improved Combination of LBP and Sparse Representation based Classification (SRC) for Face Recognition.

[15] Jiwen Lu, Gang Wang, and Thomas S. Huang, Gait-Based Gender Classification in Unconstrained Environments.

Natural Language Processing

Opposite Emotion Word Identification in Building Myanmar Word-Emotion Lexicon

Phyu Hninn Myint, Thiri Marlar Swe
University of Computer Studies, Yangon, University of Computer Studies, Mandalay
phm.ucsy@ucsy.edu.mm, thirimarlarswe@ucsm.edu.mm

Abstract

In English language, a word named "not" is noted as "Negation of a word". It is a preceding word of any positive word and these two words combination can be observed as one negative word. In Myanmar Language, a negative particle "ဓ-"(ma-) is noted as a negation syllable of the most words if it is a prefix of these words. But irregularly, it can be an infix syllable of some words and it is hard to identify which word is negative. Moreover, it is impossible to store all possible negations of positive words (antonyms) in a lexicon and even if it is possible, its size can be extremely large. Thus, the point of this paper is to present the lexical guidelines so as to recognize and distinguish a negative word (antonym), annotate it with opposite emotion of its positive word and update the opposite emotion lexicon. The system is built to support building a Myanmar Word-Emotion Lexicon. Evaluation has been conducted on an existed Myanmar Word-Emotion Lexicon [1] which contains 1402 word tokens and the outcomes are fully satisfied for all of the words in the lexicon.

Keywords

Natural Language Processing, Myanmar Emotion Lexicon, Opposite Word Identification

1. Introduction

Myanmar is a nation which is located in South East Asia and it is a member of ASEAN. Previously, Myanmar was otherwise called Burma and its language was called Burmese. These days, in our nation, our local language is authoritatively called Myanmar Language and it is utilized as an official language. Myanmar language is a tonal, pitch-register, and syllable-planned language, to a great extent monosyllabic and investigative language, with a subject–object–action word request. Myanmar contents are embraced from Mon content (one of the Myanmar principle national races) that is gotten from India Brahmi content. Myanmar language is the local language of the Bamar (principle nationality of Myanmar) and related sub-ethnic gatherings of the Bamar, just as that of some ethnic minorities in Myanmar like the Mon. It is spoken by 32 million as a first language and as a second language by 10 million, especially ethnic minorities in Myanmar and those in neighboring nations. It is an individual from the Tibeto-Burman language family, which is a subfamily of the Sino-Tibetan group of dialects. The language utilizes the Myanmar content, got from the Old Mon content and at last from the Brāhmī content. [6][10][11]

Myanmar Language is monosyllabic (i.e., each word is a root to which a molecule however not another word might be prefixed). Sentence structure decides linguistic relations and action words are not conjugated. Rather they have particles suffixed to them. The extra Part-of-Speech class in Myanmar is called Particles class. Since the words in Particles class are futile words, one Particles word has no significance and it cannot remain without anyone else as an important word in Myanmar language. Particles can be utilized as an attached one to other Part-of-Speech classes to be framed an important word. Also, it is conceivable to change the Part-of-Speech class of an important word by appending at least one Particles connected with this word. Some Part-of-Speech classes can likewise be consolidated to shape another Part-of-Speech class. These can be turned out in authoritative with Noun, Verb, Adjective and Adverb in the content. Moreover, the sort of POS tag can be changed over, that is, Noun appended with certain particles can get Verb or Adjective. Notwithstanding, a refutation particle "ဓ-" (ma-) is bound with Verb or Adjective or Adverb and it turns into a negative word. In Myanmar Language, the particle "ဓ-" (ma-) can likewise be a prefix or infix of a positive word. To recognize the negative or opposite word (antonym) is essential in the greater part of the NLP applications.

This paper is given seven sections. Related works are portrayed in Section 2. Section 3 presents a brief of Myanmar language. Segment 4 examines about Myanmar Emotion Lexicon. Opposite Emotion Word Identification System is demonstrated in Section 5 and Section 6 introduces the evaluation of the framework. Finally, conclusion is drawn in Section 7.

2. Related Work

In the paper [2], the syntactic structure of Myanmar linguistic classes has been dissected to be prepared to use in labeling Myanmar content with standard Part-of-Speech (POS) tags. It portrayed as during a Myanmar dictionary, all words are commented on with essential labels and these words are regularly called as stem words or root words. For ordinary POS labeling,

standardization step is required to make progressively significant words and comment on certain words with increasingly fitting better POS labels and classes.

The paper [2] presented the lexical principles for normalizing linguistic classes in Myanmar language. To standardize a few words and fundamental labels, the lexical principles must be applied so as to create increasingly precise and better tags called standard tags. In this manner, these guidelines can be utilized to build up the proposed standard POS labeling. The standard POS labels can be straightforwardly related with English POS labels and they are extremely helpful to be utilized in Myanmar to English Machine Translation System. Besides, they need to dissect casual structure of the sentence and attempt to make more guidelines that can explain a wide range of word blends in the casual sentence later on.

In this way, the paper [2] built up the special lexical rules so as to derive better or standard POS tag from essential POS labels mixes. By investigating Myanmar syntactic classes, 27 guidelines are characterized to standardize them. Right now, it has been made on an essential POS labeled corpus which contains 1000 fundamental POS labeled sentences and its normal sentence length is around 10 words. It depicted that the outcome is full fulfillment for all words in these sentences. They have tried numerous examinations utilizing their methodology on various sorts of sentences till they get the best exactness. The presentation of the lexical guidelines is assessed as far as the issues that can be experienced in Myanmar sentences in light of some sporadic word mix designs. The sentences that have sporadic examples have been tried with the framework and the precision of their principles has been noted.

As an assessment result, for certain words particularly for negative words, some mistakes have been discovered in light of the fact that unordinary example of verbal refutation is found in such examples where the second action word of a compound is set apart with the negative prefix, as in စိတ်-မ-ကောင်း <'sad'>, ဝမ်း-မ-သာ <'unhappy'>, and so on. Although the vast majority of the negative action words can be worked by consolidating a negative particle "မ-" (ma-) with an action word, some negative action words can be framed from a mix design which has a negative particle "မ-" (ma-) situated between two expressions of action word. To mitigate these mistakes, these words must be labeled with their fundamental labels, for example, verb1 + negative particle + verb2 and afterward, these three words can be consolidated together and labeled as a negative action word. If not really, just two words, that is, negative particle + action word, can be joined and labeled as a negative action word.

In this manner, in the paper [2], the greater part of the unpredictable negative words cannot be recognized as a particular negative word. However they can be labeled as isolated words with fundamental labels. It tackled the negative word distinguishing proof after POS tagging and its mix rules can be applied on just POS labeled words.

3. Myanmar Language

In Myanmar language, there are 34 essential consonants, 8 fundamental vowels, 4 average consonants or ward consonant signs, subordinate different signs, 2 punctuation marks and 10 digits. Vowels can be delegated free and ward vowels. Autonomous vowels can remain solitary and ward vowels are composed with a consonant [6].

There are nine Part-of-Speech classes for all Myanmar words since it is characterized by Myanmar Language Commission [4][5]. These are called as Noun, Pronoun, Verb, Adjective, Adverb, Conjunction, Postpositional Marker, Particles and Interjection. In English, just eighth Part-of-Speech classes are nominated. Preposition word class in English is for the most part the equivalent with Postpositional Marker in Myanmar. The extra class in Myanmar is Particles class.

In Particles class, there are negligible words which have nonbearing data, so one Particle word has no importance and it cannot remain without anyone else's input as a significant word. However, it very well may be connected with other Part-of-Speech classes so as to turn into an important word. In addition, it is conceivable to change the Part-of-Speech class of an important word by attaching at least one Particle word with this word. Additionally, some Part-of-Speech classes can be joined to form another Part-of-Speech class. [2]

In Myanmar Language, there is no particular word for negation like "not" in English. A negative particle "မ-" (ma-) is noted as a negation syllable of the most words if it is a prefix of these words. But irregularly, it can be an infix syllable of some words and sometimes, it is hard to identify which word is negative or antonym of a word.

4. Word-Emotion Lexicon

When social media comes out in recent years, it turns into a most generally utilized one around the globe. Users can show their feelings by posting status and imparting to companions in an online life. Examination of feelings gets a mainstream to apply in numerous application zones. Along these lines emotion detection systems have been proposed

by utilizing Lexicon based methodology. The term vocabulary is utilized in Natural Language Processing to represent a lexicon that is viewed as one of the parts of a Natural Language Processing framework. Emotion lexicons are made in their own dialects to apply in passionate framework. To recognize the users' emotions in Myanmar online media, lexicon was not accessible, so another word-emotion vocabulary particularly dependent on Myanmar language was important to create. We have introduced the making of Myanmar word-emotion vocabulary, M-Lexicon, which contains six fundamental feelings: happiness, sadness, fear, anger, surprise, and disgust described in the paper [1]. Facebook status written in Myanmar content are gathered and segmented. Words in M-Lexicon are at long last increased in the wake of applying stop-words removal process. To create M-Lexicon, Matrices, Term-Frequency Inversed Document Frequency (TF-IDF), and solidarity based standardization steps have been applied. Test shows that the M-Lexicon creation contains over 70% of accurately connected with six essential feelings. It is at first worked by utilizing 485 Facebook status with 3890 Myanmar words. In the wake of passing the pre-handling stage, there are 2147 words with 1743 stop words. The 2147 words are investigated into relating feelings. At last, 1947 feeling words have been characterized for every six emotions.

In the paper [1], Myanmar Emotion Lexicon which to be applied in recognizing emotion words from users' status in social media written in Myanmar content has been proposed. Language interpretation instruments and existing diverse language feeling vocabulary can be utilized for Myanmar language, however it isn't adequate in down to earth. Furthermore, an emotion lexicon can be physically fabricated, yet the time has come devouring. Consequently, another word-emotion lexicon for Myanmar language, in particular M-Lexicon, has been automatically made. This lexicon contains six fundamental emotions, for example, emotions such as happiness (joy), sadness, fear, anger, surprise, and disgust as defined by Ekman [7][8][9]. Myanmar word-emotion lexicon, M-Lexicon, has been assembled utilizing the words from Facebook status.

M-Lexicon is right now in introductory creation and the intention is to be future utilized in feeling recognition framework for Social Media (Facebook) clients in Myanmar language. To gather better and more feeling words, we will ceaselessly manufacture emotion lexicon with recently refreshed a few status and extra stop words. Table 1 shows the example expressions of M-Lexicon.

Table 1. Sample Words of M-Lexicon

Words	Emotions
စိတ်ညစ် <disappointed>	Sad
သာသာယာယာ <agreeablely>	Happy
သဘောကျ <enjoy>	Happy
ဝမ်းနည်း <sad>	Sad

Since most of the words in this M-Lexicon are positive words with respective emotions, it needs to be detected opposite words, i.e., to detect the antonym of the existing words with opposite emotions, that can be assigned with opposite emotions of the positive words. Because it is impossible to manually collect all antonym and their emotions can be opposite with existing words, opposite word identification process is introduced by deducing from the existing emotion words in this paper.

5. Opposite Emotion Word Identification

Among the particles of Myanmar Language, the prefix particle "မ-" (ma-) is a prompt constituent of the verb, which is the head of the word development as in: မ-ကြိုက် ('dislike'); မ-ပျော် ('unappy'). It can convert the positive sense into negative feeling of the word. The extent of verbal negation reaches out to the entire compound of a compound action word, as in မ-ပျော်ရွှင် ('unhappy'). Another example of negation is conceivable with action word mixes or action word states by individualized refutation of each part of the compound, as in: သဘော-မ-ကျ ('not appreciate'); မ-ပျော်-မ-ရွှင် ('miserably').

A negative particle "မ-" (ma-) is typically found as a prefix for most negative words. Notwithstanding, in certain words, it very well may be put as an infix between syllables. For instance, "ကောင်း" ('great') is a positive descriptive word and "ပျော်" ('happy') is a positive action word and when "မ-" (ma-) is prefixed to them, negative descriptor "မ-ကောင်း" ('not great') ("မ-ကောင်း" is an antonym of "ကောင်း") and negative action word "မ-ပျော်" ('unhappy') (antonym of "ပျော်") are formed. Additionally, "ဝမ်းနည်း" ('sad') is a positive action word and it has an excellent case for refutation. The negative particle "မ-" (ma-) must be included the center of the word to shape a negative action word like that "ဝမ်း" + "မ" (Negative Particle) + "နည်း" ("ဝမ်းမနည်း"('not sad')) (antonym of "ဝမ်းနည်း"). Another models are: "အံ့သြ" ('astonish') is a negative expression of "မ-အံ့သြ " ('not surprise') (antonym of "

အံ့သြ") and "စိတ်-မ-ဆိုး" ('not anger') is a negative expression of "စိတ်ဆိုး" ('anger') (antonym of "စိတ်-မ-ဆိုး").

It is impossible to manually insert all possible antonyms to the lexicon and important to correctly identify the antonym of the existed words to detect the emotion of the text. The opposite word identification system can be applied to detect all opposite words from respective existing words. Now, the system has been conducted with emotion words and aims to build opposite emotion lexicon. To detect the antonym of a word can apply all existing words with six emotions in M-Lexicon. In assigning opposite emotion for "Happiness" or "Sadness" emotions, the opposite emotion of "Happiness" is "Sadness" emotion and vice versa. For “Fear”, “Anger”, “Surprise” or “Disgust” emotions, the opposite emotions can be assigned as “Not Fear”, “Not Anger”, “Not Surprise” or “Not Disgust”. Therefore, the system built an opposite emotion lexicon for antonym words with respective emotions.

In the system, "Syllabification" is necessary to identify which syllable is negation particle "မ-"(ma-). "Syllabification Rules" which has been proposed in [3] is applied for this step. For example, given a word "စိတ်မချမ်းသာ", after syllabification, "စိတ်+မ+ချမ်း+သာ" is produced. Then, remove "မ-"(ma-) and emerge an updated word "စိတ်+ချမ်း+သာ". "စိတ်ချမ်းသာ" is an existed word in the lexicon and its emotion is "happiness", so" စိတ်မချမ်းသာ" is negated word of "စိတ်ချမ်းသာ" and it can be identified as "not happiness" (sadness) emotion. Identified opposite word with update emotions are collected in an opposite emotion lexicon which is helpful for the emotion detection process.

The following steps are developed to identify an opposite emotion word from an existed emotion lexicon which stored emotion words.

Step 1: Getting all syllables from an input word using syllabification rules.
eg: " စိတ်မချမ်းသာ" => " စိတ်+မ+ချမ်း+သာ"

Step 2: Removing an infix or prefix syllable "မ" <-ma> from the syllable sequence.
eg: " စိတ်+မ+ချမ်း+သာ" => " စိတ်+ချမ်း+သာ"

Step 3: Finding the updated word without a syllable "မ" <-ma> in Emotion Lexicon.
eg: " စိတ်+ချမ်း+သာ" is existed in the lexicon.

Step 4: Extracting the emotion of the existed word.
eg: "Happiness" is the emotion of "စိတ်ချမ်းသာ"

Step 5: Annotating the input word with the opposite emotion of the existed word.
eg: Emotion of " စိတ်မချမ်းသာ " is "Sadness".
(Not Happiness)

The architecture of the opposite emotion word identification system is shown in Figure 1.

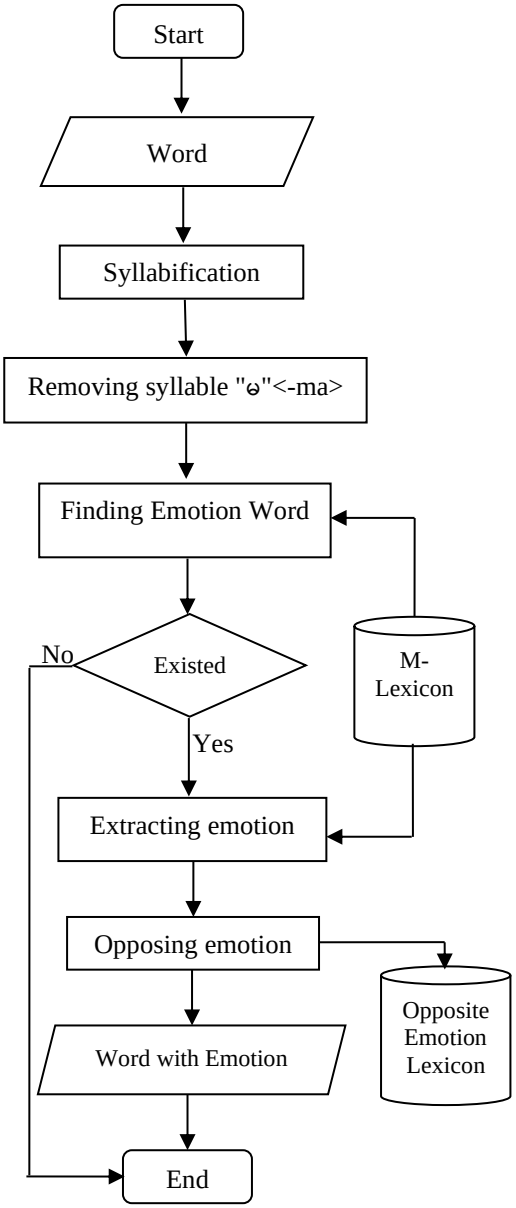


Figure 1. System Architecture

Table 2 describes some sample emotion words and their opposites, marking as syllable level.

Table 2. Sample Emotion Words and Their Opposite Emotion Words

Existed Words	Opposite Words
စိတ်+ချမ်း+သာ (H)	စိတ်+မ+ချမ်း+သာ (S)
ပျော်+ရွှင် (H)	မ+ပျော်+ရွှင် (S)
ပျော်+ရွှင် (H)	မ+ပျော်+မ+ရွှင် (S)
စိတ်+ချမ်း+မြေ့ (H)	စိတ်+မ+ချမ်း+မြေ့ (S)
စိတ်+ညစ် (S)	စိတ်+မ+ညစ် (H)
သာ+ယာ (H)	မ+သာ+မ+ယာ (S)
ဝမ်း+နည်း (S)	ဝမ်း+မ+နည်း (H)
စိတ်+ပျက် (S)	စိတ်+မ+ပျက် (H)
သ+ဘော+ကျ (H)	သ+ဘော+မ+ကျ (S)
စိတ်+ကြည်+လင် (H)	စိတ်+မ+ကြည်+မ+လင် (S)
အံ့+သြ (Sup)	မ+အံ့+သြ (Not Sup)
ေကြာက်+ရွံ့(F)	မ+ေကြာက်+မ+ရွံ့ (Not F)

Remark: (S): Sadness, (H): Happiness, (F): Fear, (A): Anger, (Sup): Surprise, (D): Disgust

6. Evaluation

The system can solve all basic emotion words in M-Lexicon and annotate them with the opposite emotion of the existed emotion words. For all existing words in M-Lexicon, the system can absolutely identify all opposite words and produce respective emotions. The system has also been tested using 100 Facebook status with opposite emotion words and it can detect and identify all opposite words from them. However, some text have unknown words if they are non-existed in M-Lexicon. The cause of the unknown word error is data sparseness. All possible emotion word collection is necessary to completely created in the lexicon. Sample Emotion Word pairs, that are correctly identified and annotated with the system, are depicted in the above table named Table 2.

Moreover, 100 irregular Myanmar Verbs and Adjectives have been tested with the system and it can identify their antonyms. Negation particles "မ-" (ma-) are infixed in most of the irregular words

Table 3. Sample Irregular Words and Their Antonyms

Words	Antonyms
ပျက်+စီး (ruin)	မ+ပျက်+မ+စီး (not ruin)
ညီ+ညာ (regular)	မ+ညီ+မ+ညာ (irregular)
နား+လည် (understand)	နား+မ+လည် (not understand)

In this case, there is no need to assign emotion and just detect the opposite words. Hence, instead of using emotion lexicon (M-Lexicon), Myanmar Lexicon [3] that contains over 22,000 word tokens is applied in this evaluation. Sample detected words are depicted in Table 3.

7. Conclusion

This paper introduces the identification of antonyms which also have irregular position of negation affixes. Myanmar positive emotion words are tested and converted to respective opposite emotion words. And then, opposite emotion lexicon has been built by using the system. Also, it can identify antonyms of positive Myanmar words without emotions.

It is a useful tool in indicating opposite emotion, sense or meaning of a word in many Myanmar Natural Language Processing systems. For Myanmar text analysis, it can be applied to identify antonyms for some applications such as Word Sense Disambiguation and Sentiment Analysis.

References

[1] Thiri Marlar Swe, Phyu Hninn Myint, "Word-Emotion Lexicon for Myanmar Language", 3rd IEEE/ACIS International Conference on Big Data, Cloud Computing, and Data Science Engineering, Springer, Japan, July 31, 2019, pp. 157-171.

[2] Phyu Hninn Myint, Tin Myat Htwe and Ni Lar Thein, "Normalization of Myanmar Grammatical Categories for Part-of-Speech Tagging", International Journal of Computer Applications, USA, December 2011, Volume 36-No.1, pp. 10-17.

[3] Phyu Hninn Myint, Tin Myat Htwe and Ni Lar Thein, "Basic Word Identification of Part-of-Speech Tagging of Myanmar Language", Proceedings of the Eleventh International Conference on Computer Application (ICCA 2013), Yangon, Myanmar, February 2013, pp 184-189.

[4] Myanmar Language Commission, "Myanmar Dictionary", 2nd ed., University Press, 2008.

[5] Myanmar Language Commission, "Myanmar-English Dictionary", 11th ed., University Press, 2011.

[6] Grammar, Burmese Language. http://en.wikipedia.org/wiki/Burmese_Language

[7] P. Ekman, "An argument for basic emotions", Cognition and Emotion, 6(3), pp. 169-200, 1992.

[8] P. Ekman, "Emotion in the Human Face", Oxford University Press, 2005.

[9] P. Ekman, and W.V. Friesen, "Unmasking the Face: A Guide to Recognizing Emotions From Facial Expressions", Marlor Books, 2003.

[10] T. Latter, "A Grammar of the language of Burmah", Baptist Mission Press, 1845.

[11] T. S. Ko, "Elementary Handbook of the Burmese language", Rangoon: American Baptist Mission Press, 1924.

Networking and Security

Data Security Based on the IPSec VPN with Filtering Security Algorithm

Zar Ni, Ei Phyo Min

University of Computer Studies, Yangon
zarniaungdala@gmail.com, eiphyo.ucsy@gmail.com

Abstract

The main objective of the system is improving for the secure link and data security between the Server Farm and Branch offices. IPSec VPN is used for the site to site communication network on each gateway. Filtering Security Algorithm is used for the Server Farm. Site to Site IPSec VPN operation secured for data transfer. So attacker can't easily attack from external network. Filtering Security algorithm is used for the more efficient data transfer on end to end secure services. Filtering security algorithm can show the blocking information and protect from the unauthorized users. By using the IPSec VPN and Filtering Security Algorithm, can get the cost effective and secure network. Users can get the secure data transferring and easily access to Server Farm through the Internet. Server can easily maintain and control the accessing clients. This paper analyzed the impact of integrating IPSec VPN and Filtering Security Algorithm. This integration secured the flow traffic of networks, and was increase the data security of server farm. The System is running on the virtual interface and show that the results in simulation tools.

Keywords: IPSec VPN, Filtering Security Algorithm, Virtual box, GNS3, wireshark.

1.Introduction

Todays, Data security is very important for the Telecommunication Companies, Private Banks, Internet Service Providers and so on. Network Communication is enhancement and people can transfer the important data on the network. On the other hand, other organization and attacker can attack and hack with the high disservice technologies. With the help of technologies, we can easily communicate and secure data transfer to all over the world. The main advantages of the data security are more reliable the data transferring and reliable network transaction. In this paper, IELTs Main Office Server Farm and Branch Offices are transferring the important data through the Internet. The Site to Site IPSec VPN is used for the communication network and Filtering security algorithm is used for the data security. Virtual Private Network (VPN) is a communication network by which a user can tunnel through another network by using the global Internet or by Intranet with strong security features. It provides for security, encryption, decryption and tunnel. The IPSec has been a standard method for the VPN technologies. IPSec is a collection of some special Internet protocol to supply a secure data transfer over the network layer. It protects all application traffic over an IP network. Site to Site IPSec VPN can protect form the public interface attacks [1,2]. Filtering Security Algorithm is used for data security of the private interface. SSL (Secure Socket Layer) and MAC (Media Access Control) Address Filtering Algorithm is combined in Filtering Security Algorithm. SSL is a standard security technology, it using more secure for web server and web browser, or mail server and mail client .MAC Address Filtering Algorithm is important for prevention unauthorized users. So, Site to Site IPSec VPN and Filtering security algorithm is used for data security between the IELTs Main Office Server Farm and Branch Offices.

2.Background Theory

The background theory of this system is including the IPSec VPN and Secure Socket Layer.

2.1 IPSec VPN

IPSecurity VPN (IPSec VPN) is provided the security between two gateway routers, or between a client and gateway. IPSec provides two different modes: Transport Mode and Tunnel Mode. Transport Mode uses only for host-to-host security,

provides protection for the payload of IP packet. Tunnel Mode provides security between two networks by protecting the entire IP packet. IPSec is based on two encapsulation protocols: Authentication Header (AH) and Encapsulation Security Payload (ESP). AH provides origin authentication, data integrity and anti-packet repetition. ESP protocol provides data confidentiality, authentication and an optional replay protection service. ESP modifies the original IP packet inserting a new ESP header and a packet trailer. The ESP header is not encrypted but a section of the trailer and the complete data payload are encrypted. ESP protocol authenticated the ESP header, Original IP header, IP payload and ESP trailer. Encapsulated IP packet with ESP When the destination node receives the IPSec packet [3,4]. The Security Association (SA)is an agreement or a contract between two IPSec peers or endpoints. The SA contains all the information required for the two peers to exchange data securely. Internet Key Exchange (IKE) Allows for automatic negotiation and creation of IPSec SAs between IPSec peers. IKE is a hybrid protocol that gives different services to IPSec such as: IPSec peer authentication, security association agreement and key generation /regeneration for cipher algorithms used by IPSec. IKE negotiates the IPSec SAs. This process requires that IPSec peers get authenticated first with the help of digital certificates or pre-shared keys. After doing this, IKE can take further actions for the negotiation of IPSec SA [5].

2.2 SSL

Secure Sockets Layer (SSL) is a standard security technology for establishing an encrypted link between a server and a client, use in a web server (website) and a browser, or a mail server and a mail client. Eg (HTTPS: HTTP over SSL). SSL Security protocol provides data encryption, server authentication, message integrity, and optional client authentication for a TCP/IP connection. SSL is a tunneling protocol that allows a proxy server to act as a tunnel between the client and the server. SSL runs at the application layer and provides secure transaction of data such as credit card details, between a client and an E-commerce server. SSL uses certificates, private/public key exchange pairs and Diffie-Hellman key agreements to provide privacy (key exchange), authentication and integrity with Message Authentication Code (MAC). This information is known as a Cypher Suite and exists within a Public Key Infrastructure (PKI). SSL secures millions of peoples' data on the Internet every day, especially during online transactions or when transmitting confidential information. SSL consists of two phases: handshake and data transfer. Both client and server use a public key encryption algorithm to determine secret-key parameters during the handshake phase. In the transfer phase, however, clients and server use secret key to encrypt and decrypt successive data transmissions [6,7,8].

3.Network Design

GNS3 and Virtual Box simulators are used in this network design. Windows operating systems were run in Virtual Box. Routers and Switches were run in GNS3. GNS3 and Virtual Box simulators were connected for using instead of real devices. Two types of department namely IELTs Main Office Server farm (IELTs Server farm) and Branch offices are used in this system. Branch offices are Yangon Branch Office, Mandalay Branch Office and Bangkok Brach office.

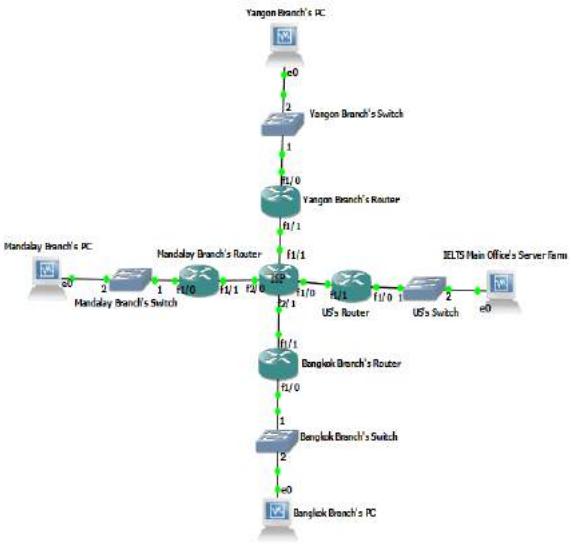


Figure 1. Network design

In figure 1, IELTs Main Office Server farm connects US's Router via US's Switch. Yangon Branch office connects Yangon Branch's Router via Yangon Branch's Switch. Other branches are also connecting respectively. ISP router is used for instant of Internet. US's router and Branch routers connect with ISP router. US's router and Branch offices routers are interconnected with the Site to Site IPSec VPN tunnel.

3.1 Parameters used in the network

In this system, Site to Site IPSec VPN and NAT are configured for each router. Two phase of IPSec VPN are ISAKMP (Internet Security Association and Key Management Protocol) phase 1 policy parameters and IPSec phase 2 policy parameters. I choose these policy parameters for this system. Show in this figures 2 and 3.

Parameters		Use in all Router
Key distribution method	Manual or ISAKMP	ISAKMP
Encryption Algorithm	DES,3DES or AES	AES
Hash algorithm	MD5 or SHA-1	SHA-1
Authentication	Pre-shared keys or RSA	Pre-share
Key exchange	DH Group 1,2 or 5	DH 2
IKE SA lifetime	86400sec or less	86400sec

Figure 2. ISAKMP phase 1 policy parameters

Parameters	US's	Yangon's	Mandalay's	Bangkok's
Transform set	Server	yangon	mandalay	bangkok
Peer IP address	199.199.1.5 199.199.1.9	199.199.1.6	199.199.1.10	199.199.1.14
Network to be encrypted	192.168.2.0 192.168.3.0	192.168.1.0	19.168.1.0	192.168.1.0
Crypto Map name	server-MAP	yangon-MAP	mandalay-MAP	bangkok-MAP
SA Establishment	86400sec	86400sec	86400sec	86400sec

Figure 3. IPSec phase 2 policy parameters

4.Proposed System

The System is designed for IELTs Office Server Farm and Branch Offices (Figure 4). This system gets secure link and security improve for IELTs sever farm. In the network security uses the IPSec, it has been a standard method for the VPN technologies. That supplies a secure transport medium for private networks through the public network. In this network, Site to Site IPSec VPN is used for interconnection. Site to Site IPSec VPN protects from the public interface attacking in this system. Site to Site IPSec VPN uses OSPF routing protocol. Organizational users must use the internet. So, NAT (Network Address Translation) is used for Internet. NAT translates the original IP address. When NAT translates the IP address, Site to Site IPSec VPN tunnels cannot be used in this network because this system uses the Site to Site IPSec VPN tunnels in every organization's gateways. So, the IP address is translated with NAT but except the IPSec VPN tunneling address. And then, Organization can use the Internet and IPSec tunneling. This network design is for the reliable data transfer through the Internet.

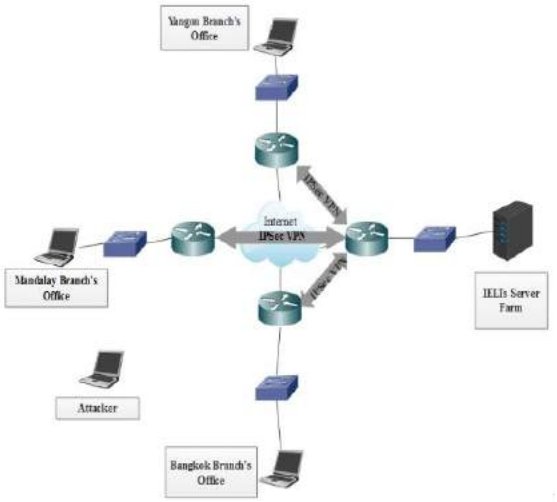


Figure 4. System Overview

IELTs Database Server and IELTs Socket Server in IELTs Server Farm are created. Branch offices (client) and IELTs Server Farm (server) used the Socket for interconnection. The Socket is used instead of Web for accessing to the IELTS server. Web Server can't get the MAC address of requesting computer. By using the Socket, the physical address and information of unauthorized users can be received. MAC address is important for the network user. MAC address is controlled for the data security. Because, that is unique, the Filtering Security algorithm is used when client login. Client must fill the username and password and then he clicks the login button. In this time, the Filtering Security Algorithm achieves the client's username/password and his MAC address. These algorithm double checking with incoming data and predetermine data of IELTs database. If the data are equal, he can connect with IELTs Server Farm else he will be blocked from IELTs Server Farm. Filtering Security Algorithm can show the unauthorized user's information (MAC address, incoming time, login branch, username and password) to IELTs Server Farm. The proposed system supports the six kinds of advantages for the network. They are secure data transfer, reliable over the network transaction, more efficient data security in server farm, more flexible use in secure socket, easily maintain in server farm and an inexpensive for security features. Disadvantages of this system is usage of more transferring time.

4.1 Design of The System

IELTS Server Farm can manage the all database tables. IELTs database server has six tables, they are:

1. blockmac (Blocking database table)
   Include (MAC address, Incoming time)
2. blocktimelimit (Block time limit database table)
   Include (MAC address, Incoming time)
3. branchlogin (Branch login database table)
   Include (Branch name, Username, Password, MAC, Time)
4. mac (Organization's MAC address and username password table)
   Include (Branchname, Username, Password, MAC)
5. registration (Branch register database table)
   Include (Fname, Lname, NRC, Email, Module, Testdate, Cost, Branchname)
6. result (Result database table)
   Include (Fname, Lname, NRC, Testdate, L, R, W, S, Avgscore, Branchname)

IELTs Socket Server is created in IELTs Server Farm. It maintains and checks the Branch Management operation and Branch login status operation. The branch Management action can check easily the login predetermine information of our organization. Branch office users can also add, delete and update of the branch's name, password and MAC address to the database table. In the Branch login status operation has two parts. They are view success login and view unauthorized login. View success login shows the about of branch information (Branch names, user name, password, Mac address and login time). View unauthorized login has the MAC address of unknown PC and login time. If the unsuccessful login time is over three times, this unauthorized user's information was added in blocktimelimit table of IELTs database server. Server site can control the block timing limit for unauthorized user. If the real organizational users are being blocked, server site can manage the user. IELTs Server Farm creates the client windows application and form. This is Data Security socket, is used for the Brach Offices. This socket can show the student's result and it manages the registration list. Branch Office's client can create the new student registration list in this socket. In creating the new student registration list, students pay their hardcopy information to the branch client. IELTs Server Farm's data transmission action has included the SSL operation and IPSec VPN operation. If Branch Offices want to access the IELTs Server farm, this must be done according to these flow chart (figure5).

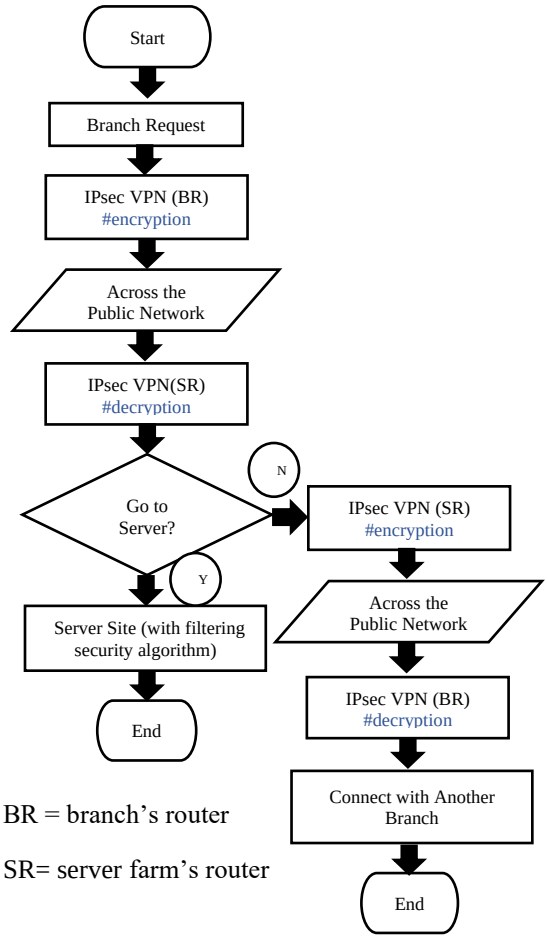


Figure 5. Over view of Branch Client requesting to IELTs Server farm

In this flow chart has two ways. First way, Branch client can connect to the IELTs Server Farm. When Branch client requests to the socket, destination address has been included in Site to Site IPSec VPN tunnel processing. And then this request packet across to the internet. The requesting packet reaches the US's router and it checks the destination address. If this address is IELTs server Farm, this packet must through the filtering security algorithm. Second way, Branch's client connects to the another branch offices client. This network traffic must be through the site to site IPSec VPN tunnel. Figure 6 represents the flow chart of Filtering Security algorithm. When request computer (client) requests to IELTs Server Farm, must through the Filtering Security algorithm.

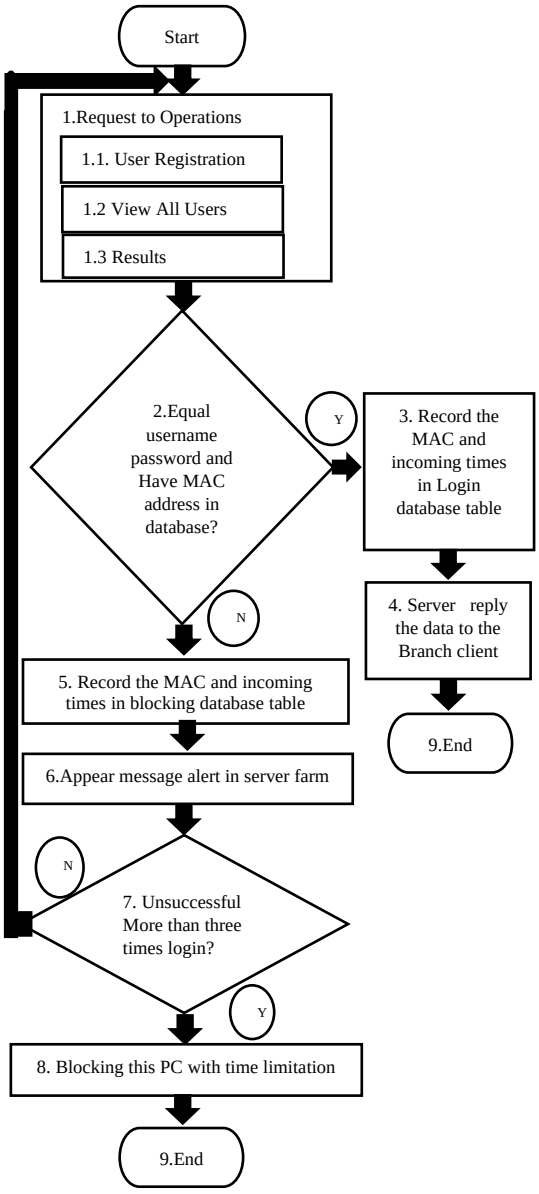


Figure 6: Filtering security algorithm Flow chart of IELTs Server Farm

Step1: Request computer can choose the three operations of Data Security Socket. In choosing any action, Branch client must fill username and password.

Step2: Filtering Security Algorithm gets the request computer's hardware address (MAC address) and username/password. This Algorithm compares with the predetermine database's MAC address and the request computer's MAC address and then check the username and password. When they are equal, go to step3. Else go to step 5.

Step3: Request computer's MAC address and incoming time are recorded in branchlogin database table.

Step4: This PC can access with server farm. Go to step 9.

Step 5: If the request computer MAC address has not in predetermine database, request computer's MAC address and incoming times are recorded in blocking database table.

Step 6: By using this algorithm, the Message alert appears in server farm's command prompt for checking the unauthorized user.

Step 7: When the request computer logins wrong username and password, it can try login again. If login time is more than three times, go to step 8. Else, go to step 1.

Step 8: This request computer will be blocked from the server farm. During the block limitation time, it can't try to login again.

Step 9: Filtering Security Algorithm processing is ending.

5. Experimental Results of Data Security

This section represents the security level of the purpose system. IELTs Server Farm and Branch Offices are transferring the data on the network. If the attackers attack organization on the public network, they can receive only the ESP encapsulation payload. Attackers can't know what is source and destination. They can't receive the real data because transferring packet is encrypted and digested by IPSec VPN operation. When Mandalay Branch client connects to Data Security Socket, other organization's users sniff the transferring packet in the public network. If they use wireshark packet sniffing tool, Site to Site IPSec VPN is how to prevent the transferring packet shown in figure 7. In this transferring, tool can't sniff the stream of TCP/UDP/SSL.

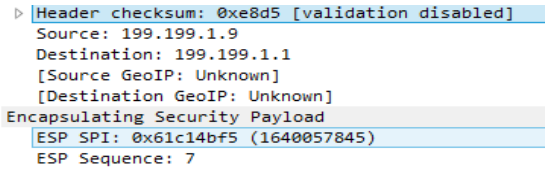


Figure 7. IPSec VPN prevent the transferring packet in public network

If the attacker tries to attack in the private interface, Filtering Security Algorithm can prevent him. MAC address filtering Algorithm is used for checking the login condition. In Branch's PC login case, user can easily access to IELTS Server Farm. If Branch's client login is fault username or password, this PC can be blocked and recorded in blocking database table. But, that PC can try three times consecutive login. Message alert shows in IELTs Server Farm's command prompt. In unauthorized user or attacker login case, He can't access. Even the username and password is true. Because filtering security algorithm checks the MAC address. If this MAC address is not matched the predetermined MAC address, this PC will be blocked and recorded from IELTS server farm. And then message alert shows in IELTs Server Farm's command prompt. This figure 8 is the alert of IELTs server farm. Message alert has included the attacker's MAC address, username, password and known the attacker login form where branch.

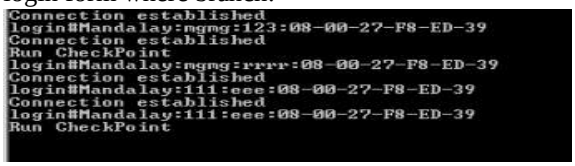


Figure 8. Message alert in IELTs server farm

If Branch's PC or unauthorized user are more than three times loin unsuccessful, this PC will be blocked and time limitation for next time login form the IELTS server farm. Figure 9 represents the client is unsuccessful login over three times.

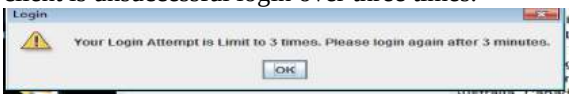


Figure 9. Time limitation Message alert

When Attackers reaches Branch's private network, they sniff the transferring data on the network. But attackers can't sniff the real data because SSL is used on data. Attackers will get the TCP stream of only encryption data. Figure 10 represents the attackers have sniffed the result of wireshark packet sniffing tool when Mandalay Branch offices look up the IELTs students result.

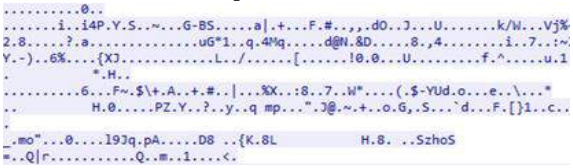


Figure 10. System prevention transfer data in Private interface

6.Conclusion

This paper introduced the improving data security. The proposed system proved the reliable data transfer between the private network and public network. By using the IPSec VPN, the users can get strong security for router to router making it easier to maintain the network. IPSec VPN processing has encryption time, digesting time and encapsulation time. So data transfer rate is slower than using only OSPF routing protocol. For end to end data security, the Filtering security algorithm needs to be added in server farm. According to this proposed system, the users can get secure data transfer link for their data, reliable network transition. The users can also get the information of Attackers and can know who attack their network. So this system exactly prevents from attacking of the unauthorized user and attacker. This System can easily add and maintain the new network when the organization's need. Hardware and software firewall are very expensive for network security. So, Site to Site VPN and Filtering Security Algorithm are used in this purpose system. This way can reduce the price of data security and network security. This System gets the cost effective. The system is very suitable for education site, market site, banking site and travel servicing site. In future extension, attacker or hacker will encounter return-attack by using the better security feature.

References

[1]Douglase E.Comer "Internetworking with TCP/IP Principles ,protocol ,And Architecture Sixth Edition" volume 1,2014.

[2] Yi Yang, "Virtual Private Network Management", May 2011

[3] Powell, J. Myles, "The Impact of Virtual Private Network (VPN) on a Company's Network" (2010).

[4] Aruna Malik ,Harsh K Verma ,Raju Pal "Impact of Firewall and VPN for securing WLAN" May 2012

[5] Masqueen Babu, "Performance Analysis of IPSec VPN over VoIP Networks Using OPNET" September 2012.

[6] Brhrouz A.Forouzan , "Cryptography and Network Security", International edition 2008.

[7]Mohammed A. Alantheer, "Secure Socket Layer (SSL) Impact on Web Server Performance" (September 2014).

[8] Li Zhao, Ravi Iyer, Srihari Makineni, Laxmi Bhuyan "Anatomy and Performance of SSL Processing.

Meeting Room Management System on peer-to-peer Overlay Network using Event-Based Routing

Su Yadanar Than Htike, San Thida
University of Computer Studies, Yangon
suyadanar442@gmail.com

Abstract

With the advent of the Internet, it became possible to build large-scale distributed applications because of the existence of a global, packet-based communication infrastructure. This increase in application scale by several orders of magnitude results in systems with millions of nodes that need to communicate and cooperate in order to achieve a common goal. This system presents an efficient method for building large-scale distributed systems (Case Study: meeting room management system). The propose method employs two underlying techniques: event-based communication and peer-to-peer overlays. Event-based communication offers many-to-many communication style while peer- to-peer overlays provide scalable, self-organizing and decentralized substrate for distributed applications.

Key words: ***Large-scale distributed system, Event-based communication, peer-to-peer***

1. Introduction

The large boom in computing electricity and the emergence of high-velocity PC networks have made allocated large-scale systems viable and computing allocated today is widely deployed while building huge software program systems. Such systems, usually a few variations of the traditional request / reply system, use distinctive interaction models. The main project with this model of interaction is its synchronous existence, resulting in very static archite ctures.

Any other inherent request / reply problem is its point-to-point communication version, which creates close coupling between speaking parties. A new methodology is crucial as it creates large-scale decentralized networks that have to scale up to the huge internet world of today. A model known as a fully communicate or put up / subscribe paradigm based on an occurrence offers a great choice to try this and is becoming more and more crucial. The simple principle is that the system is made up of publishers who send events, and subscribers who participate in successful occasions [4].

This model helps with many-to-many verbal exchanges and high decoupling between the different events. Fully daily interaction is a powerful version. It also contains the apex of an activity or message agent overlay network. Message agents form distribution trees that can be used to route the event to the subscribers from the publisher. Such agents rely on the routing of the utility degree because the multicast network degree has not been commonly used now.

This system aimed to develop the meeting room system by using event-based communication. Meeting room management systems consist of software for conference room scheduling and interfaces. The software allows company employees or guests to book meetings online or through an application.

Software-based conference room booking systems may be used in hospitality, in studios or spas, or in enterprises for employee booking. Another popular usage is in shared workspaces, for reserving rooms or just desk space.

Meeting Room Booking Systems feature the following capabilities:

- Book conference rooms electronically
- Set reservations or book impromptu meetings
- Seating assignments, or hot desks
- Visitor management (e.g. visitor details, sign-in, reminders)
- Repository for room information (e.g. amenities, photos, etc.)
- Floor plans and maps for guests or facilities
- Check-in using app, or clean up ghost (i.e. abandoned) meetings from schedule
- Real-time availability detection, up-to-date floorplans
- Search for available meeting rooms with filters (e.g. location, time available)
- Usage analytics (e.g. utlization, busiest times, etc.)

- Booking available on the web, mobile app, or a tablet
- Room issue reporting (e.g. equipment failure) through the app
- Integrate with digital signage or room displays / interfaces
- Integrate with commonly used calendar or productivity apps

2. Related Work

The related works of concurrency controls are discussed in this session.

In this system [1], processes at each exit and entrance maintain the number of vacancies in the car parking. Firstly, the process exams the kind of event (getting into or leaving) requested for updating the wide variety of vacancies. If the kind of process is a "getting into event", the system will test any vacancies within the automobile parking. If there's no emptiness, the system will display the "unavailable message" to the consumer. Otherwise, the process will multicast requests to all different tactics and watch for their replies. After you have all reply, the process decreases the range of vacancies by1 and lets the car input. Then, it is multicast the up to date facts to all different hosts. And it replies all requests saved inside the local queue. Sooner or later, the process leaves the essential phase. If the type of technique is "leaving manner", the process makes requests to all different processes and waits for his or her replies. Upon getting all replies, the technique increases the quantity of vacancies by1 and shall we the car depart. Then, it is multicasts the up-to-date information to all different approaches. And it replies all requests saved in the neighborhood queue. In the end, the method leaves the crucial phase.

Event oriented model was introduced in distributed system since 1970s [2]. Many software and operating system use graphical consumer interface based totally on occasion. Dealing room system turned into designed based totally on the issue of event and notification. This system affords the present price of stock on market. Whilst the adjustments on price occur, it could tell to consumer straight away by sending notification rubdown. As it turned into based on the event and notification paradigm, consumer does not need to couple with the gadget and the conversation between the device and users is asynchronous.

Internet Indirection Infrastructure, i3, is another rendezvous-based communication abstraction [9]. It offers an application level multicast architecture that can be seen as a sort of topic-based publish-subscribe system. Each packet is associated with an identifier which is used by sources (publishers) to send the packet, and by receivers (subscribers) to express their interest in certain packets. This effectively decouples the sending from the receiving, that is, provides indirection. Subscribing in i2 is done by inserting a trigger to the system that tells who's interested and in which topic (id). Each id has its own rendezvous node in the network where the triggers are stored.

3. Background Theory

The huge increase in computing power and the emergence of high-speed computer networks have made large-scale distributed systems feasible and today distributed computing is widely deployed when building large software systems. These systems use different communication models, typically some variation of the traditional request/reply paradigm.

The major challenge with this communication model is its synchronous nature which leads to very static architectures. Another inherent problem with request/reply is its point-to-point communication model which creates tight coupling between communicating parties [10, 11]. A different approach is necessary when building largescale distributed services that must scale to today's Internet wide environment. A model called "event-based communication" or "publish/subscribe paradigm" offers a good option to do this and is becoming more and more important. The basic principle is that the system consists of publishers that publish events, and subscribers that subscribe to certain events. Subscription can simply be topic-based, but more advanced schemes, such as content-based and type-based schemes, have been proposed. Events are delivered asynchronously to the receivers. This model supports many-to-many communication and high decoupling between the different parties. [9]

Event-based communication is a powerful model. It is usually implemented on top of an overlay network of event or message brokers. Message brokers form event dissemination trees that are used to route the event from the publisher to the subscribers. These brokers rely on application level routing since the network-level multicast has not been widely deployed. Traditional publish/subscribe systems have been built on quite static overlays and

do not take advantage of the underlying network topology very well [7].

The questions about continuously changing, dynamic and fault-tolerant network architectures have been addressed in the peer-to-peer networking research. Several propositions of scalable, decentralized and self-organizing overlay networks of communicating peers have emerged. These architectures are often based on a mechanism to route packets to a selected peer in the network in an efficient manner. The efficiency is usually measured in terms of physical network utilization, event dissemination latency and routing state kept in the brokers [5]. This system will present the event-based communication paradigm as basis, improving it by using a sophisticated mechanism to build the underlying overlay network of message brokers.

3.1. Event-Based Communication

Numerous verbal exchange models frequently lack good transparency in time, space and synchronization, which un-appropriate for big-scale disbursed computing. Occasion- based totally communique offers an efficient approach to clear up those problems.

A perfect introduction to periodic, mainly focused verbal exchange, often referred to as the model of submission / subscription. The basic function of subscribers is to remind the computer of their participation in positive activities. This is usually done by subscribe ( ) activity on a few forms of event notification service. This form of service can be seen by publishers and subscribers as an impartial mediator. This carrier is also commonly referred to as a message or event broker, and an ordinary pub / sub network consists of a variety of them forming an overlay group over a community layer of the body.

When a publisher wants to publish an event, it calls the event service to notify ( ) operation, which is then responsible for routing the event through the event broker network to the interested subscribers. The concept of event service provides three-dimensional decoupling between the event publisher and subscribers: space, time and synchronization [9].

Space decoupling; A decoupled architecture allows each component to perform its tasks independently of the others, while also enabling structural variations between source and target.

Time decoupling results in independence of actions in time so that the communicating parties do not need to be active at the same time.

When sending or receiving events, publishers and subscribers are not blocked. These properties of decoupling result in a very transparent communicatio n infrastructure that scales well and supports a large number of subscribers and publishers.

3.2. Peer-to-Peer Overlay Networks

Peer-to-peer overlay networks have emerged as an efficient communique approach. The peer-to-peer networks offer numerous exciting aspects like fault- tolerance and absolutely decentralized manage. This allows them to adapt nicely to exceedingly dynamic networks, ie., networks, wherein nodes may additionally be a part of and leave at will each time, and they scale well to nowadays internet-extensive surroundings.

3.3. Administrability

An event-based middleware is a complex dispensed tool with a lot of additives in itself. Any such machine's smooth administration is a crucial necessity, particularly since any large-scale system could also evolve considerably over its lifetime. For example, a fully middleware based on occasion can also build a complex software-stage overlay culture, the topology of which is not considered in advance. The implementation now not only defines the system's exact operation, but also influences the efficiency of the collection of information [5].

Whilst new additives are delivered to growth overall performance or availability, the occasion-primarily based middleware has to evolve without significant quantities of human intervention.

Through having components as autonomous as possible, the administration effort can be reduced. Self-adapting systems can relieve the administrator from many decisions, thereby enabling system management assignment. Because of the fact that peer-to-peer systems are developing themselves with the help of classification, peer-to-peer approaches are again a sensible choice. To add a brand new component to the event-based middleware, a count number of "plug it in" must be handy [2].

4. Event Publication and Subscription

Each scribe node may be a publisher that creates a topic, or a subscriber of a certain subject

matter. Any node having the specified credentials also can put up activities to a topic. Nodes can publish, create and enroll in many subjects at the identical time. The simple API (Application Program Interface) provided by scribe infrastructure includes the following strategies:

• CREATE (credentials, topicId)

• SUBSCRIBE (credentials, topicId, eventHandler)

• PUBLISH (credentials, topicId, event)

Submit-subscribe combined to an efficient overlay which gives an effective occasion dissemination mechanism. Scalability and expressiveness are two vitals, and incredibly contradictory dreams for this type of gadget. Expressiveness right here means the device's functionality to provide an effective information version and filtering primarily based on it. Scalability refers back to the potential to support a huge range of subscribers, publishers and activities. Then, it is also way that the device should help heterogeneous extensive-location networks whose requirements are completely distinct from those of a neighborhood-area community [3, 6].

4.1. Creating Topic

Every subject has a correct topicId. The scribe node with nodeId that is numerically closest to the topicId acts as a meeting point for the related topic. When creating an issue, a scribe node asks pastry to path the CREATE message with the subject I d to the nearest number node, the rendezvous node. The appointment reviews the credentials and adds the subject to its list of topics. The topicId is actually a hash of the subject's request to use a collision-free hash function in conjunction with the name of its author. It means that on the underlying network, the ids are dispensed equally [6].

4.2. Subscribing

If a Scribe node decides to subscribe to a subject, the SUBSCRIBE message will be sent using the route method where the topicId is used as the password. This brings the message to the rendezvous of the subject. Every node along the route investigates whether the topic is already widely subscribed. If the subscription is for an unknown issue, the subscription will be sent forward to the rendezvous as an entry for it is far produced to a children's desk to consider the subscription [11].

This mechanism of subscription is very scalable to a large number of topics and subscribers per topic. In addition, many of the nodes within the network are adequately dispensed with the forwarding load, and the rendezvous factors in particular are not overloaded. Local houses are lowering the traffic created by the network [8].

4.3. Publishing an event

If a publisher wants to publish an event, it calls on Pastry to use the topicId as the key to route the event to the rendezvous point. In the peer-to-peer overlay network, the rendezvous performs access control and forward the event to the multicast.

4.4. The Algorithm of the Proposed System

```
Let Subscriber = S; Publisher  = P;
   Event Broker = EB; Space decoupling = SD;
   Time decoupling = TD;
BEGIN
  Step1: Accept the event to operate;
  Step2: The system check the event owner is P or S;
  If (Event Owner == P)
   {
     If (Communicating parties [Ps] which are active
        at the same time on same event)
     {
        Solve the same activation event by TD;
        // Publishers and subscribers are not blocked
          while sending or receiving events.
      }
    End If
        Operate the accepted event;
        Determine the Ss to send the notification by
        the SD;
        Calls Notify ( ) operation;
        EB network → Ss;
      }
      Else If (Event Owner == S)
      {
```

S (interest in certain events) → System.

Call Subscribe ( ) operation;

Executed on some kind of event notification service;

Respective P publishes the S requested message;

}

End If

END

5. Implementation of the System

Personnel participants want which will log in and e-book meetings, even when they're far from the workplace. Advancements in facts propagation technology have also brought about an increase in assembly room booking system utilization. Blessings of assembly room booking management device:

Save time

Prevent errors

Monitor meeting room usage

Streamline visitor management

The detail processing steps of the proposed system are shown in the following figure [figure1].

In this meeting room management system, each peer station can accept the user request and checks the concurrency with the other peers' holding events. If there is no concurrent processing for the user request, the publisher of the user request holding peer publish the transaction processing event locally. After updating locally, the event broker of the peer inform the other peers' subscribers and then propagate the data update to all peers consecutively.

If there is concurrent processing for user request, the system grant a chance to the user to choose the other meeting room or change the appointment time for that room or can exit from the system.

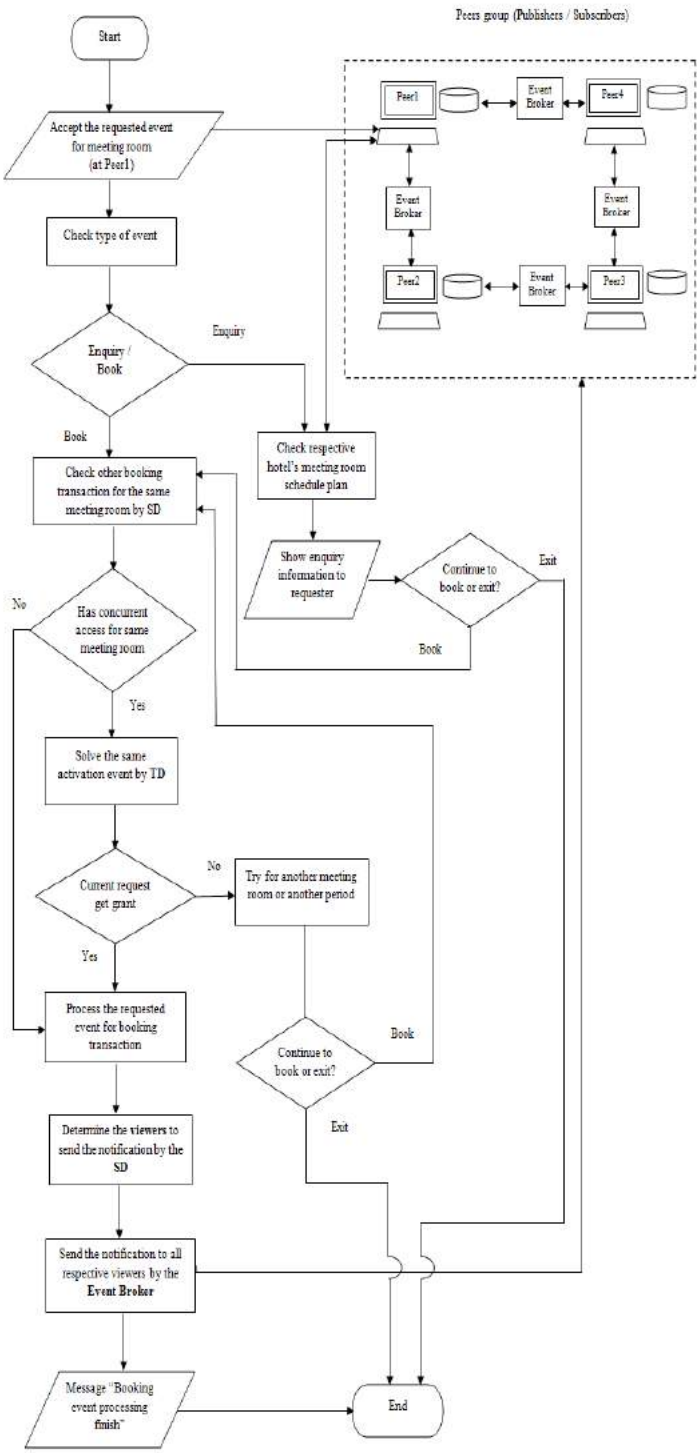


Figure 1: The System Flow

6. Conclusion

This system presents the publish-subscribe model and how to enhance it by linking message brokers with peer-to-peer overlay networks. Publish-subscribe offers an efficient model for communication by loose decoupling between subscribers and publishers. The overlay culture that

links the brokers is the crucial part of this system. The meeting room booking management system will offer all clients accessibility, accuracy and continuity to the records by controlling the scribe to event-based routing.

REFERENCES

[1] Thida Kyaw, "Monitoring the Number of Vacancies in a Car Parking by Using Distributed Mutual Exclusion", , M.C.Sc 2011, University of Computer Studies, Yangon.

[2] Su Su Aung, "Event Notifications among Distributed Objects in Dealing Room System", M.C.Sc 2012, University of Computer Studies, Yangon.

[3] L. F. Cabrera, M. B. Jones, and M. Theimer. Herald: Achieving a global event notification service. pages 87–94, 2001.

[4] A. Carzaniga, D. S. Rosenblum, and A. L. Wolf. "Design and evaluation of a wide-area event notification service". ACM Transactions on Computer Systems, 19(3):332–383, 2001.

[5] M. Castro. An evaluation of scalable application-level multicast built using peer-to-peer overlay networks, 2003.

[6] M. Castro, P. Druschel, Y. Hu, and A. Rowstron. Topologyaware routing in structured peer-to-peer overlay networks, 2002.

[7] M. Castro, P. Druschel, A. Kermarrec, and A. Rowstron. SCRIBE: A large-scale and decentralized application-level multicast infrastructure. IEEE Journal on Selected Areas in Communications (JSAC), 20(8):100–110, 2002.

[8] G. Cugola, E. D. Nitto, and A. Fuggetta. The jedi event-based infrastructure and its application to the development of the opss wfms. IEEE Trans. Softw. Eng., 27(9):827–850, 2001.

[9] P. T. Eugster, P. A. Felber, R. Guerraoui, and A.-M. Kermarrec. The many faces of publish/subscribe. ACM Comput. Surv., 35(2):114–131, 2003.

[10] G. Mühl. Large-Scale Content-Based Publish/Subscribe Systems. PhD thesis, 2003.

[11] G. Perng, C. Wang, and M. Reiter. Providing contentbased services in a peer-to-peer environment. In DEBS '04: Proceedings of the 2nd international workshop on Distributed event-based systems, pages 74–79. ACM Press, 2004.

Dual Axis Solar Tracking System Using PIC16F887 Microcontroller

Theint Zin Zin Moe, Khin Than Mya
University of Computer Studies, Yangon
theintzinzinmoe@gmail.com, khinthanmya@ucsy.edu.mm

Abstract

The use of alternative sources of energy is widely used in all over the world. Alternative energy is a clean energy source that is derived from natural and renewable source of energy such as solar, wind, tides, geothermal, hydrogen and so on. Sun is also main source of various energies; the light energy is the best way to treat a remark value. The solar panel converts the light energy that stored into electrical energy, which power can use electrical appliance directly or store into battery. The efficiency of solar panel can be exploited by aligning the solar panel according to the sun rotation. In this system, two-axis tracking system is implemented by using PIC microcontroller. The controlling of motor speed depends on the intensity of light receiving from the sensor. PID control is used to control the speed of the stepper motor. In this thesis, Light Dependent Resistor is used to detect and measure the intensity of light. Stepper motor is used to rotate the solar panel to sense maximum sun light location recorded by the LDRs. PIC16F887 microcontroller is used to control the solar tracking system.

Keywords: PIC microcontroller, PWM (Pulse Width Modulation), Dual Axis tracking

1. Introduction

Alternative energy is a type of energy that is achieved from natural resources such as sunlight, wind, rain, tides and geothermal heat. The usage of energy sources that are renewed rapidly growing because of the fact that fossil fuel price fluctuates, environmental problem such as pollution and climate change, controlling of deforestation, ozone layer protection and mining of CO2 production [1].

Renewable energy source can get anywhere and can be developed using various technology. This type of renewable energy sources is clean and decrease environmental problem. Renewable energies have been widely used in residential, industrial, remote and transportation areas. Solar energy also has the potential to be the major energy supply in the future.

Solar Panels are used to take up the rays of sun and convert that into electricity or heat. A solar panel is actually a collection of solar cells.

Solar tracking is used to rotate the solar panel to face the sun. When it is rotated according to the time, solar tracker is recorded the amount of light intensity. There are various types of popular solar trackers but the two basic categories of tracker widely used are single axis and dual axis tracker [2].

2. Related Work

A solar tracker is a device used for orienting solar panel towards the sun by using the light sensor connected with the motor, hence to get maximum radiation at the solar panel. Solar tracking system can be used for industrial and household. Arduino Uno is used for controlling the system. Five light dependent Register sensors is used for sensing the position of the sun. Two servo-motor is used to direct the position of the panel. Dual-axis solar tracking system generate 40% more power than single axis solar tracking system [2].

A solar tracking system is a monitoring system which aims for solar panels to operate by tracking the sun at full efficiency during the day. There is Atmega 328 microcontroller, DC motor and LDR sensor use in this design. Atmega 328 microcontroller controlled the solar panel direct to the sun position and command the motor to get the maximum efficiency. PID controllers are widely used in industrial application. For PID controller has been implemented with physical design. Fuzzy controller is suitable for human decision making process. For fuzzy logic controller has been implemented using MATLAB [3].

3. Background Theory

In this section represent control system, PID control system and pulse-width modulation method.

3.1. Control System

A control system is a set of mechanical or electronic devices that regulates other devices or systems. Typically, control systems are computerized. A control system is an interconnection of components related in such command, direct or indirect. Two type of control system are open loop

control and closed loop control system. In this system, closed loop control system is used.

3.2. PID Control System

A proportional–integral–derivative controller is the algorithm that calculates the measured speed by adjusting to reach the desired speed with minimal delay and overshoot, by regulating the power output of the engine. The PID control equation involves; the Proportional P, the Integral I, and the Derivative D as shown in Figure 1. PID controller calculate error (e (t)) value as the difference between a measured process variable and the desired set point. The controller attempts to minimize the error by processing the control inputs [3].

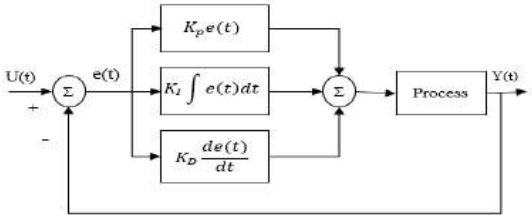


Figure 1: PID Control System

3.3. Pulse Width Modulation Method

Pulse-width modulation (PWM) is used for controlling the amplitude of digital signals in order to control devices and applications requiring power or electricity.

PWM is particularly suited for running inertial loads such as motors, which are not as easily affected by this discrete switching, because they have inertia to react slowly. The PWM switching frequency has to be high enough not to affect the load, which is the resultant waveform perceived by the load must be as smooth as possible. For stepper motor control, the duty cycle is fixed and the frequency varied. By varying the frequency, the speed of the stepper motor may be changed as shown in Figure 2.

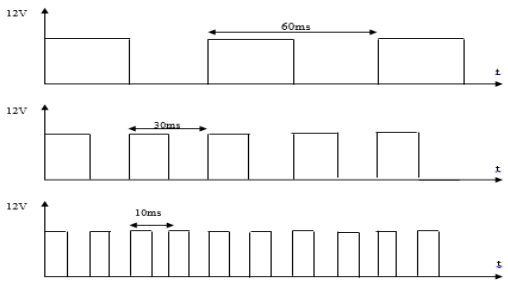


Figure 2: Controlling For Stepper Motor Speed

4. System Design and Implementation

In this section intended to present the implementation of the solar tracking system. In system overview, describe the block diagram of the solar tracking system and describe briefly about the components which used in system.

4.1. System Overview

PIC microcontroller, LDR (light dependent resistor), stepper motor, stepper motor driver, a solar panel, limit switches are containing in this system. The stepper motor driver control circuit is designed to make correct direction of motor rotation depending on the difference between sensors.

LDR sensors are used for sensing the strength of the sun light to get the maximum power output. Limit switch is used to detect the rotating panel to reach the limit position. Stepper motor is used to reach the desired position. PWM and PID control is applied to manage the correct speed of the motor.

PIC16F887 microcontroller controlled the speed of the motor and the overall system component as shown in Figure3.

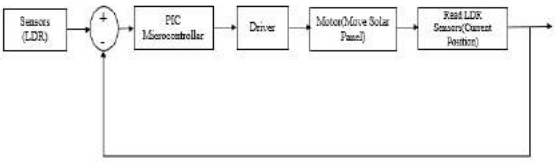


Figure 3: Block diagram of the Solar Tracking System

5. Hardware Implementation

This system is used to point the solar panel to track the location of the sun during the day as to get the highest efficiency of the power output. The hardware elements applied here are solar panel device, LDR (Light Dependent Resistor), stepper motor driver, stepper motor, Microcontroller and Limit Switch. The mechanical hardware design is shown in Figure 4.

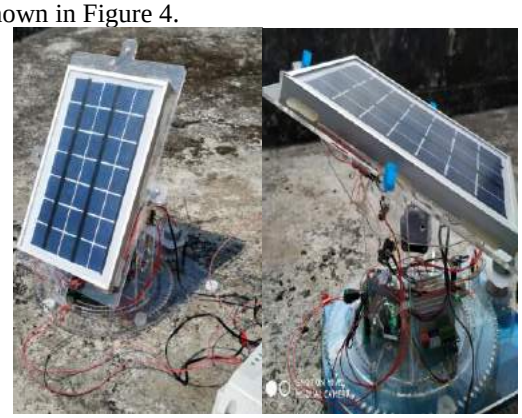


Figure 4: Complete System Design

5.1. Microcontroller

PIC microcontroller is a microcontroller with 35 input/output as shown in Figure 5. Input/output pins can provide up to a maximum 25mA of current. There are 8K ROM, 256 bytes EEPROM memory, 368 bytes RAM memory. Microcontroller is applied as a main component for this system [4]. There are various kinds of PIC microcontroller but here, PIC16F887 is used.

Figure 5: PIC16F887 microcontroller

5.2. Light Dependent Resistor (LDR)

A Light Dependent Resistor (**LDR**) is a device whose resistivity is a value of the incident electromagnetic radiation. The Figure of LDR sensor as shown in Figure 6. The resistance of LDR will fall down when it is increased incident light intensity [5]. In this system, four LDR sensors are placed on the north, south, east and west. Light Dependent resistors connected in series to a 330R resistors.

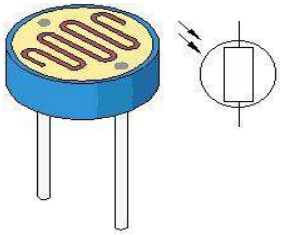
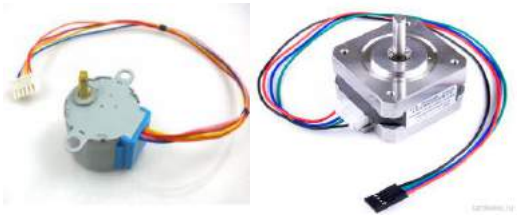


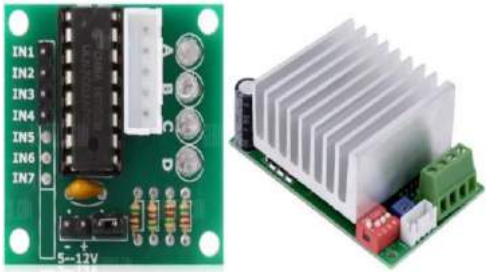
Figure 6: Light dependent resistor

5.3. Stepper Motor

A stepper motor is an electromechanical device which converts electrical pulses into discrete mechanical movements as shown in Figure7. In this system, 28BYJ-48 and JLB stepper motor 17hs1362-p4130 motors are used to drive the solar panel toward the sun. 28BYJ-48 stepper motor was used for horizontal axis. 17hs1362-p4130 stepper motor was used for vertical axis. In this system, two different motor types are used because the weight of the solar panel to lift more powerful motor is required for vertical axis.



(a) 28BYJ-48 and 17hs1362-p4130 stepper motor



(b) ULN2003A and TB6600 stepper motor driver

Figure 7: Stepper motors and stepper motor drivers

5.4. Solar Panel

Solar panels are devices that convert light energy into electrical energy. A solar panel is actually a collection of solar cells, which can be used to generate electricity. There are three kinds of solar panel based on their efficiency, price and temperature coefficient. They are monocrystalline, polycrystalline and thin film. Polycrystalline solar panel was chosen here because it has low cost compare to other types. In this system, 12V 5W solar panel is used as shown in Figure 8. The feature of solar panel as shown in Table1.

Figure 8: Solar Panel

Table 1: Solar Panel Feature

Parameters	Value
Maximum power	5W
Open circuit voltage	20.9V
Short circuit current	0.31A
Maximum power voltage	17.64V
Maximum power current	0.28A

5.5. Limit Switch

A limit switch is an electromechanical device that consists of an actuator mechanically linked to a set of contacts. They can be determined the presence or absence, passing, positioning, and end of travel of an object. In this system, limit switch is used to detect the stepper motor to stop the desire position as shown in Figure9.

Figure 9: Limit switch

6. Software Implementation

For the operations, the system writes a C programming to control motor via PIC16F887. The codes were compiled and uploaded to PIC16F887 microcontroller.

6.1. Operation of the System

The described system is firstly turned on. The data read from Left and Right light dependent resistors and find out the difference between them. If the difference is greater than Gap, the motor will shift to the left and otherwise the motor move to the right. The Left and Right rotation process is run once at the start of the system because the solar panel face to the sun position and then the up and down rotation process for the whole day. Then the data read from Up and Down light dependent resistors and find out the difference between them. If the difference is greater than Gap, the motor will move to the Up and otherwise the motor will move to the Down. A Gap value was chosen 5 by making many tests according to the sensitivity of the sensor. The Up and Down rotation process is run repeatedly during the day.

If there is no match above the state, the solar panel will move to initial position. The overall operating step-by-step are illustrated in the Figure 10.

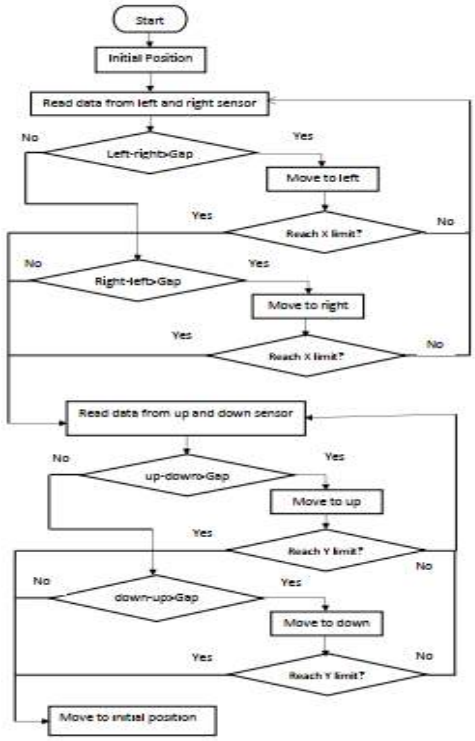


Figure 10: The flowchart of the system

7. Experimental Results

The system results are presented in this section. The result shows the locations of the solar tracker is important fact in obtaining maximum output voltage and current. The output voltage and current of solar tracking system are measured every one hour from 7 A.M to 5 P.M [6]. The solar panel was placed toward the east. The static panel with the same characteristics was tested. For left-right rotation, the solar panel move between 0-180 degree and for up-down rotation, the solar panel move between 0-180 degree.

The measurement of the data was taken in the roof space to get the maximum sunlight. The voltage, current and power output from solar tracking system as shown in Table 2. The voltage, current and power output from without solar tracking system as shown in Table 3. According to the Table2, the output power is obtained maximum power. In Table3, the output power is decreased from the evening.

The results of power output comparison between solar tracking system and the panel without solar tracking system as demonstrate in Figure 11. According to the result, the solar tracking system is more energy produce than without solar tracking system. The solar tracking system is continuously tracked the sun, it can get maximum output power for the whole day.

According to the output result, the output power of the static panel decreases from 12:00 PM.

Table2: Voltage, Current and Power output with solar tracking system

Time	Output Voltage(V)	Output Current(A)	Output Power(W)
7:00 AM	19	0.04	1.14
8:00 AM	19.6	0.04	2.156
9:00 AM	20.0	0.24	4.8
10:00AM	20.1	0.24	4.824
11:00AM	20.1	0.24	4.824
12:00PM	20.2	0.24	4.848
1:00 PM	20.2	0.24	4.848
2:00 PM	20.2	0.23	4.646
3:00 PM	20.1	0.22	4.422
4:00 PM	20.1	0.18	3.618
5:00 PM	20.0	0.17	3.4

Table3: Voltage, Current and Power output without solar tracking system

Time	Output Voltage(V)	Output Current(A)	Output Power(W)
7:00AM	19	0.06	1.14
8:00 AM	19.4	0.11	2.134
9:00 AM	19.8	0.22	4.356
10:00AM	19.58	0.22	4.3076
11:00AM	19.4	0.22	4.268
12:00PM	19.2	0.18	3.456
1:00 PM	17.4	0.04	1.218
2:00 PM	17.1	0.02	0.855
3:00 PM	17.6	0.03	0.528
4:00 PM	17.3	0.02	0.346
5:00 PM	17.0	0.02	0.32

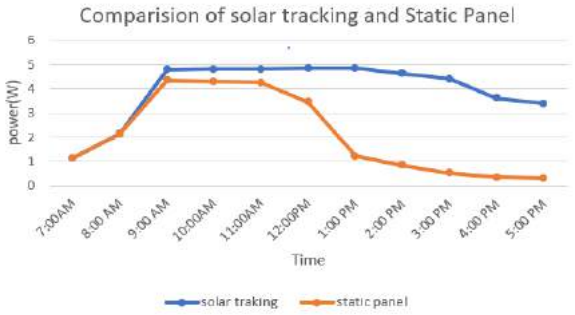


Figure 11: Output power comparison

8. Conclusion

In this system, a solar tracker has been developed to increase the power generated by the solar panel. A PIC microcontroller was used to control the movement of the solar panel. Solar Energy is being widely used, and within some more years it will be very popular. According to the data, the solar tracking system is more efficient than non-solar tracking system in terms of voltage and current. Dual-axis solar tracking system can obtain maximum power output than static panel. It will be used for many purposes, in industries and household as well.

References

[1] N. Othman and et.al, "Performance Analysis of Dual-axis Solar Tracking System", Dec 2013.
[2] M.D Rupani, "Design and Implementation of Dual Axis Solar Tracking System", Journal of Engineering Research and Applications, 2015, Vol5.
[3] G. Gol, "A Comparison of Fuzzy Logic and PID Controller for a Single-axis Solar Tracking System", Feb 2016.
[4] PIC16F887 Data Sheer http://www.microchip.com
[5] S.Ghosh1 and N. Haldar, "Solar Tracking System using AT89C51 Microcontroller and LDR", International Journal of Emerging Technology and Advanced Engineering, 2014, Vol4.
[6] A.Acakpovi and et.al, "Low Cost Two-Axis Automatic Solar Tracking System", Dec 2015.

Smart Water Filling System by Using GSM Network

Nandar Aung Than, Khin Than Mya
University of Computer Studies, Yangon
nandaraungthan1994@gmail.com, khinthanmya@ucsy.edu.mm

Abstract

Automatic system becomes one of the important sectors in today's world. Smart system technology make life easier and more convenient and can control whether you're at work or on vacation. Smart System also provide some energy efficiency saving. Among the energy resource, water is the essential resource for everyone who live in the earth. Because of human neglect, there is annual water loss in many houses without necessary. Automatic water level indicators and controller can solve a solution to this problem. Therefore, this system will present the water level controlling process based on Arduino and SMS (Short Message Service). It also bases on SMS in the GSM network to instantaneously transfer the collected data. In this system, we are going to measure the water level by using floating switch and Time of Flight (ToF) sensor. The ToF sensor measures the water distance from the sensor to the water surface which means that it measures as mm. When the water level is near to the ToF sensor, the system will automatically detect the level and close the motor which is assigned the set point by using the proportional control. The motor pump will control by the user. This will help in reducing wastage of water as electricity. It will be of great benefit to many future smart homes and industries.

Key words: Arduino, GSM Module, Water Pump, Floating Switch, ToF Sensor.

1. Introduction

Water is the main source in every section like variety of industries and agriculture. Water covers 71 percent of the earth's surface. It is significant that water is very important to every creature on earth. The main cause of water wastage is often poor infrastructure, although many other factors influence the scarce use of water.

Therefore, the usage of water controlling system will need for home or office. Most of the people want to control from any places at any time by remotely and want to get easy access to control remotely with their phones. With the development of today technology, a new technology called smart home system has been popular and people are trying to control their any electrical things remotely by using GSM technology.

The common method of water level control is simply to start the pump at a low level when the user commands from the SMS and allow it to run until a higher water level is reached in the water tank. Water level control system is now widely used for monitoring and controlling of other liquid levels, dams and tanks etc. [5]

The water level control is a design of an intelligent automatic level measurement system using PID controller. The system will allow users to measure the set level and control the process of water flowing into the tank. The system will relate to level sensor and GSM module for the control section.

By designing these systems, there are a lot of benefits such as electricity consumption savings, water overflow, manpower usage, etc. Electricity is also become an essential part of our living environments. Hence, we can't afford to misuse any fraction of it. If we leave any of the home appliances switched on, this will lead to energy wastage too. With an aim to solve this problem, a control system based on Arduino device has been developed that can control any electrical equipment at home remotely both for any distances by using mobile technology. Hence, this system can protect our home from any accident caused by overuse of home appliances like electric iron, microwave, oven, water pump, etc. The water pump will be controlled smartly by the system and the user through SMS. Traditional water pump is controlled manually which results wastage of time, energy and resources. So, the system is designed to control water pump automatically by sending a SMS [1].

2. Related Work

In the automatic system, water pump motor is automatically turned on when the water level reaches to the assigned level in the tank and it is automatically turned off when the water level is full. There is PIC18F452, LCD and Buzzer and 10 DIP switches in that design. PIC microcontroller controlled the motor from the main tank, and it pumps the water into another tank. LCD used to show the

water level in both tanks. Buzzer used to create an alert to close the pump or when the water is in rising sate. DIP switches work as sensing the water level [2].

The industrial application of tank system is widely used in chemical processes. The control of liquid level in tanks and flow between tanks is a problem in the process technologies. The process requires liquid to be pumped and stored in tanks, and then pumped to another tank automatically. Mostafa et.al. presented a simulated level control of liquid in two tanks with a different controller's such as Proportional-Integral-Derivative (PID). Various conventional techniques of PID tuning method were tested in order to obtain the PID controller parameters. Simulation was communicated within MATLAB environment to describe the performances of the system [3].

The water level monitoring system with a combination of GSM module to notify the admin via SMS alert using GSM technology. The water level is monitored, and its data sent through SMS only to the defined user mobile's phone upon reaching the assigned level. The system consists of different circuit boards and GSM AT command. It also included the calculation of level height to monitor and notify to user. The difference types of circuit boards are water level detector circuit board and microcontroller GSM circuit board [4].

3. Background Theory

3.1. Control System

A control system is a system which consists of number of devices or sets of devices that managed and connected to perform a specific function, in which the output is controlled by input. Control systems can be classified into two main classes of control action. They are open loop and closed loop.

3.1.1 Open Loop Control System

Control Systems in which the output has no effect on the control action and need to control manually by user are called open loop control system. In an open loop control system, there is no feedback and the output will not be compared with the input as described in Fig (1).

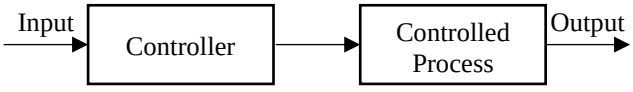


Fig (1). Open Loop Control System

3.1.2 Closed Loop Control System

In a closed loop or feedback control system, the controlled output is measured and compared with the reference input. The difference between the two, called error, is fed into the controller which produces a control signal to reduce the error in Fig (2).

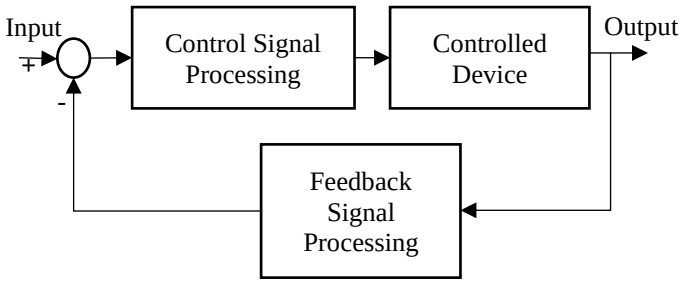


Fig (2). Closed Loop Control System

4. Controller Design

The PID controller is the most common form of feedback. PID controllers are today found in all areas where control is used. PID control is an important ingredient of a distributed control system. The controllers are also embedded in many special-purpose control systems. A PID controller is a controller that includes the proportional element, **"P element"** that accounts for the present value of the error, the integral element, **"I element"** that accounts for the past values of the error and the last derivative element, **"D element"** that accounts for possible future trends of the error, based on its current output of change. Algorithm for PID can be classified as below in Fig (3).

$$u(t)=K_p e(t) + K_i \int_0^t e(t)\,d + K_d \frac{d}{dt} e(t)$$

where u(t) : controller output
K_p : Proportional gain
K_i : Integral gain
K_d : Derivative gain
e : Error
t : Time or Instantaneous time

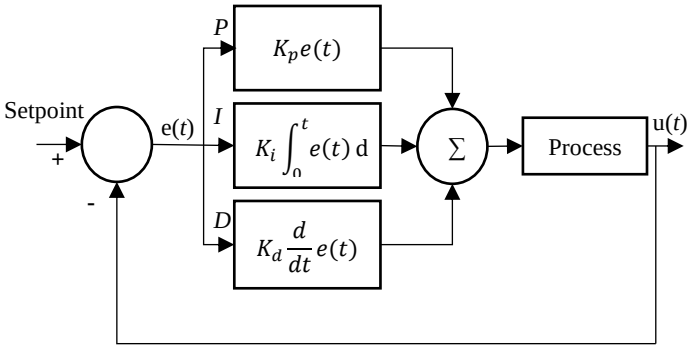


Fig (3). PID Controller Algorithm

5. Overview of the Proposed System

5.1 Comparison of Wireless Technology

Today many wireless technologies are developed and useful. Researchers can easily implement in most controller. They can also choose the technology depend on the system [6]. Among the wireless technology, the most useful and famous are as shown in Fig (4).

Description	Zigbee	Bluetooth	GSM
Range	10-100 meters	10 meters	Large distance
Operating frequency	2.4GHz	2.4GHz	900MHz
Networking Topology	Ad-hoc, peer to peer	Ad-hoc, very small network	By SMS messages
Cost	Expensive	Low Cost	Expensive

Fig (4). Variety of Wireless Technology

5.2 Hardware Implementation

5.2.1 Block Diagram

The following block diagram of the plan is shown in Fig (5). The system is aimed for filling the water only by sending SMS to switch on the motor from the user when the user is at home or not. There are using the updated materials to improve the system. They are Arduino Uno, GSM Module, Water Pump, Floating Switch, ToF Sensor, LCD, SD Card.

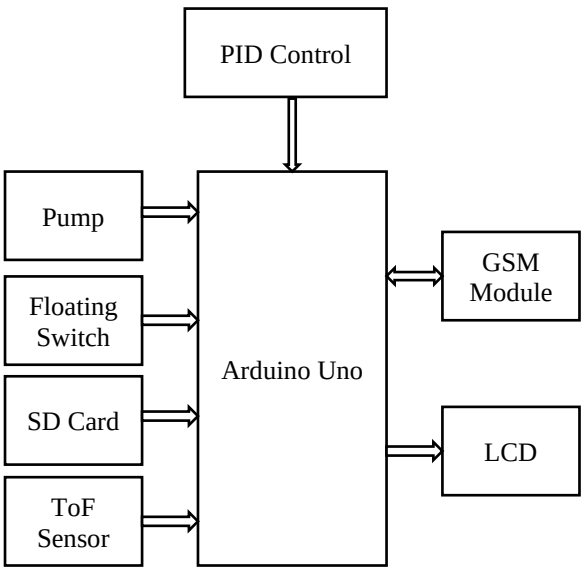


Fig (5). Block Diagram of Water Level Control

5.2.2 Controller

There are many different types of Arduino controller such as Arduino Uno(R3), Lilypad Arduino, Red Board, Arduino Mega(R3), Arduino Leonardo. Among them, Arduino Uno is the most suitable and compatible with other devices.

The Uno is a type of microcontroller board based on the ATmega328. It contains everything needed to support the microcontroller; simply connect it to a computer with an USB cable or power it with an AC-to- DC adapter or battery to get started. It can run on windows, Linux and Mac.

The advantages for using Arduino are easy to purchase, inexpensive, cross-platform, simple to control, clear programming environment, open source for all and extensible software/hardware.

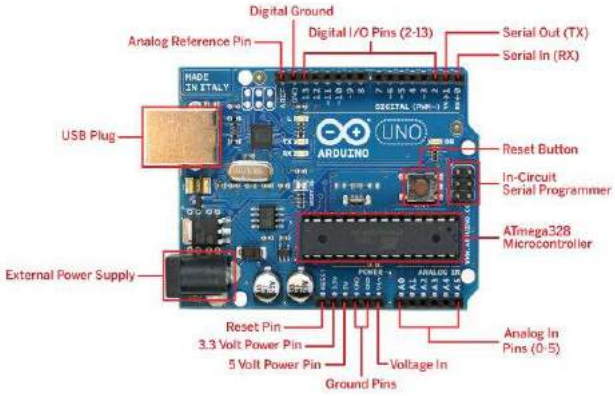


Fig (6). Arduino Uno

5.2.3 GSM Module

The GSM module is a breakout board and minimum system of SIM900A Dual-band GSM/GPRS module. This is reliable wireless module. SIM900A delivers GSM/GPRS 900/1800MHz performance for voice, SMS, Data, and Fax in a small form factor and with low power consumption. It can communicate with controllers via AT commands. The interface is via UART/Serial.

Fig (7). GSM Module

5.2.4 Water Pump

The mini micro submersible water pump is used for water pump as fountain, garden and controlled water hydroponic systems. The motor can easily connect with Arduino pin of 3,5,6,9,10 and 11 which can be used as analog output pins [7].

Fig (8). Mini Submersible Water Pump

5.2.5 Floating Switch

These side mount float switch works well for small tanks. It regulates single and multiple levels, give low-level or high-level readings. The switch is normally open in the vertical or drop position and normally close in the horizontal or up position. The floating switch is compatible with any controller including Arduino, PIC microcontroller and Raspberry Pi, etc. In this system, floating switch will only check the water which is still having or not in the lower tank for filling to upper tank.

Fig (9). Floating Switch for Level Sensing

5.2.6 Time of Flight Sensor (ToF Sensor)

ToF is a high-performance proximity and ranging module housed in the smallest package, providing accurate distance measurement whatever the target reflects unlike conventional technologies. It can easily integrate in Arduino, low power consumption, competitive system cost and proximity sensing.

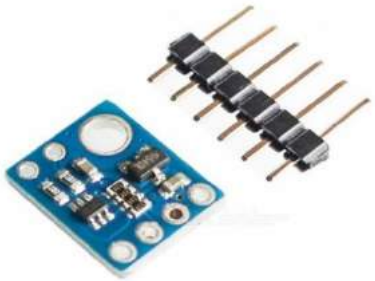


Fig (10). ToF Sensor

5.3 System Design

The following Fig (11) and Fig (12) show the design by user side and system overview.

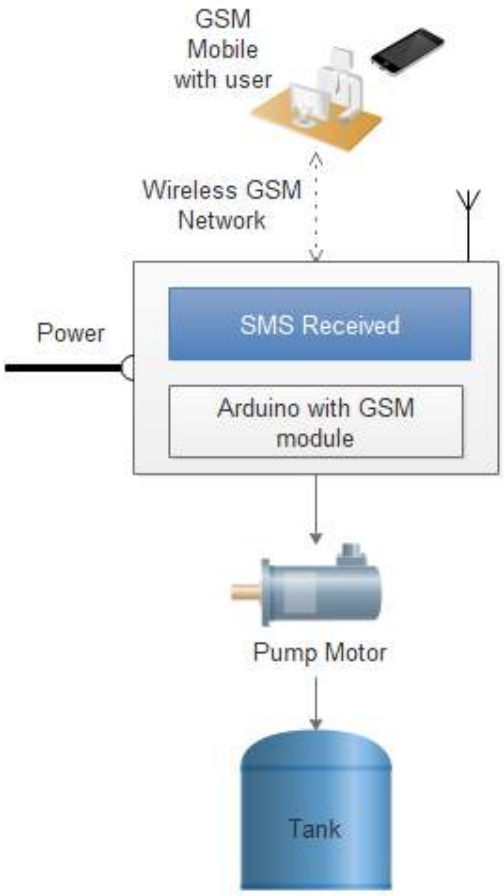


Fig (11). System Design by User View

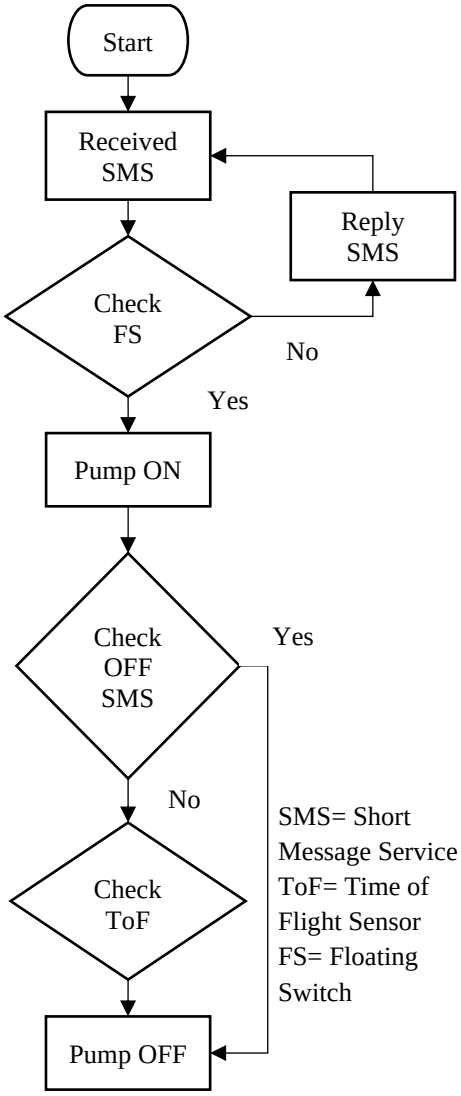


Fig (12): The System Overview

As described in Fig (11) and Fig (12), users can switch on the motor from anywhere. When the user sends "ON" message to the controller, the sensors will check both tanks about the water state. If there's no water in the lower tank or full in the upper tank, the motor will not work.

If above state is not true, the motor will work and filling the water into an upper thank from lower tank. At this time, the floating sensor and ToF sensor will always check in the filling state.

When the user is suddenly switch OFF the pump for no reason even the filling process is not finish yet, the motor will OFF depend on the condition. After the filling process is complete, the motor will turn off automatically or can close by user.

6. Experimental Results

According to the Fig (13), the motor output data is exported from SD Card. The system can check the motor speed because of applying P Value in this design. It means that the power would be on until the target water level is reached, and the power would be removed, so the pump reduces speed.

In proportional control, the power output is always proportional to the error. If the water level is at target distance and the speed of the motor will gradually decrease and stop.

In this system, set a distance to control the stability of motor output between the water level and ToF sensor. When the user drains the water, there'll occur error because of the water wave. Therefore, this system needs to set the constant motor output before the water level reaches the sensor.

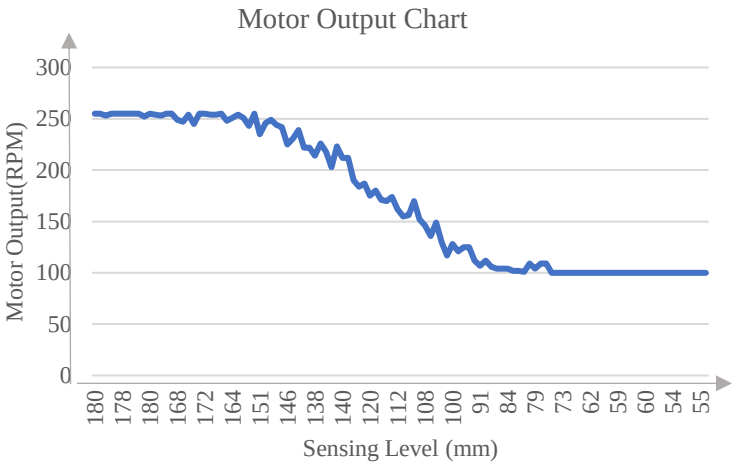


Fig (13): Measured Motor Output Chart

The below Fig (14) and Fig (15) shows respectively the process of filling water process. When the water level is gradually rising, the speed of the motor will slow down because of the proportional theory. Depend upon the received information, user can ON/OFF the motor.

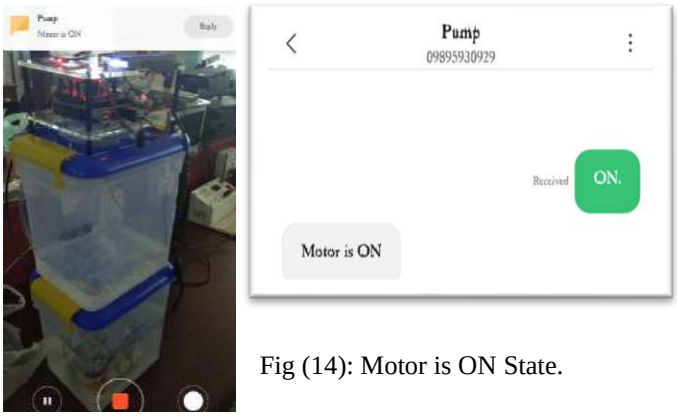


Fig (14): Motor is ON State.

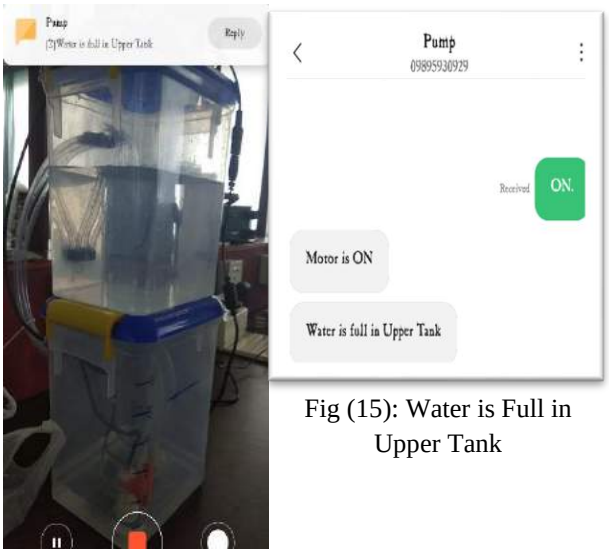


Fig (15): Water is Full in Upper Tank

6.1 Applications

- Used in industries to control the water level automatically.
- Because of using SMS control, devices can be controlled from long distances.
- Can be used by everyone with just the knowledge of test SMS.
- Can be implement with other system in control of water level and alarms for steam boilers, neural network, solar tracking and many systems.

7. Conclusion

The smart water filling system by using GSM Network is the smart system as all processes involve with continuous updates by Arduino controller to the user, via GSM technique which shows with SMS notification in phone. Its advantage is power saver, money saver, automatic, water maximization and easy installation with other system. It also has a disadvantage too like some of the devices will be outdated in future and need to replace with updated devices, no warranty or guarantee. Arduino controller is used as a platform and local materials. In this system, pump motor will ON/OFF automatically by sending SMS to control the level of water. The smart water filling process can be used in agriculture field, apartments, factories, domestic house, etc.

References

[1] Anirban Sarkar, Sadia Sultana and Md.Habibur Rahman, "A Smart Control System of Home Appliances Using SMS", Global Journal of Researches in Engineering, ISSN 2249-4596, vol 17, no 1, April 2017.

[2] Ahmed Abdullah, Md. Galib Anwar, Takilur Rahman, and Sayera Aznabi, “Water Level Indicator with Alarms Using PIC Microcontroller”, American Journal of Engineering Research (AJER), e-ISSN 2320-0846, p-ISSN 2320-0936, vol 4, pp. 988-92, July 2015.

[3] Mostafa A. Fellani, and Aboubaker M.Gabaj, “PID Controller Design for Two Tanks Liquid Level Control System using Matlab”, International Journal of Electrical and Computer Engineering (IJECE), ISSN 2088-8708, vol 5, pp. 436-442, June 2015.

[4] Ayob Johari, Mohd Helmy Abd Wahab, Nur Suryani Abdul Latif et.al. “Tank Water Level Monitoring System using GSM Network”, International Journal of Computer Science and Information Technologies, ISSN 0957-9646, vol 2(3), May 2011.

[5] Ms T.Deepiga, and Ms A. Sivasankari, “Smart Water Monitoring System Using Wireless Sensor Network at Home/Office”, International Research Journal of Engineering and Technology (IRJET), vol 2(4), July 2015.

[6] Comparison of Wireless Technologies (Bluetooth, Wifi, BLE, Zigbee, Z-Wave, 6LoWPAN, NFC, Wifi Direct, GSM, LTE, LoRa, NB-IoT, and LTE-M). https://predictabledesigns.com/wireless_technologies_bluetooth_wifi_zigbee_gsm_lte_lora_nb-iot_lte-m/

[7] Arduino Functions – AnalogWrite. https://thinkcrate.co/#intro

Barrier Avoidance Robot by Fuzzy Logic

Ei Ei Khaing (Computer University Hinthada)

eikhaing.hinthada@gmail.com

Abstract

Barrier avoidance robot primarily fuzzy logic control system can be applied ultrasonic sensor by Arduino microcontroller. Automated Guided Vehicles have many potential applications in manufacturing, medicine, space and defense. The main scope of the system is to automatically changing the direction of robotic vehicle as required whenever any obstacle comes on its ways. The aim of the system is to implement a Barrier Avoidance Robot based on fuzzy logic[2]. It avoids any static barriers in front of the robot. The fuzzy logic controller uses the sensor data as inputs and drives the motor used in the robot for avoiding the barrier. This system uses three ultrasonic sensor and one output[5,10]. Sensor detects the presence of any obstacle and sends the signal to microcontroller which changes the direction of the robot[3]. It uses fuzzy logic controller to impart smooth movement of robot while it tries to avoid barrier. Fuzzy inference engine of fuzzy set is eleven proposed. This system girts with three ultrasonic (HC-SR04) sensors to measure the distance from the barrier.

Keywords: *barrier avoidance, fuzzy logic controller, ultrasonic sensor, Arduino UNO, DC motor, L298N motor driver.*

1. Introduction

Introduces the design of an autonomous barrier -avoidance robot car using ultrasonic sensor. The fuzzy logic system has widely used for one of effective means in unknown and complexes industrial environments. Environment developments of mobile robot have attracted the attention of researchers in the various areas. The barrier avoidance distance can be measured[4]. We here propose another new rule table that is induced from the consideration of the distance to barrier and the angle between the sensor and the target. The sensor gets the data from surrounding area through mounted sensors on the robot. The proposed a new activity to design a fuzzy controller for improve the competent of mobile robot to react to dynamic surrounding. The sensor gets the data from surrounding area through mounted sensor on the robot. Many Studies present Fuzzy logic Controller with 3 input sensors for barrier detection and avoidance robot, normally resulting in dead zones and difficulties in maneuvering. Ultrasonic sensors are used for measure distances around robot from left barrier, right barrier, and center barrier. Specific challenge of design an intelligent controller is in determining what information is needed. Although on option is to utilize probability theory in order to come up with a more realistic model, this still relies on obtaining information about an agent's environment with some amount of precision. In fact, is a unique feature of the design since it enables replacement of existing component with more sophisticate. The design specification was to build a robot which following a line and avoid barrier. The important were the safety considerations for the robot in case of run-away situations. The generality of the accidents occurs on the way because of human intensive driving. The answer to solve this problem is the development of the self-driving equipment. A self-driving vehicle is able to sensor sense its surroundings and go to the direction without human stead. It has capability to minimize damage due to driver demerit.

2.Three portions in this system

- Sensors as input
- Controller
  - Fuzzy Logic Controller
  - Microcontroller
- Actuators as output

Three portions of system is shown in Figure 1.

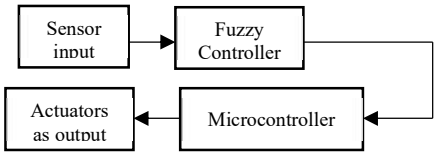


Figure 1. Block diagram of the barrier avoidance robot

Hardware components in this system are

(1) Arduino UNO R_3
(2) L298N motor driver
(3) 4× 6V DC motors for 4 wheel
(4) 3 Ultrasonic sensors

(5) 3× 4V Lithium. Ion Battery model: 18650
(6) 1 battery holder
(7) Car Chassis
(8) Jumper wires

The proposed system architecture is shown in Figure 2.

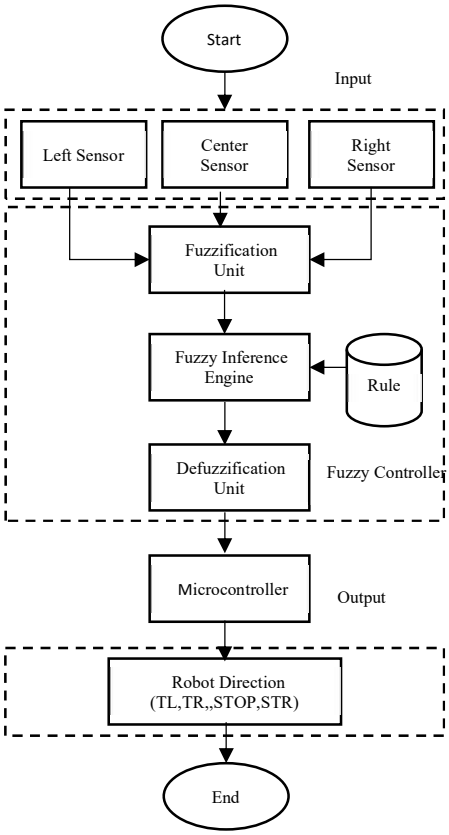


Figure 2. Propose system architecture

2.1. Sensors

A sensor is a machine that produce and responds to some form of input from the physically surrounding. The accurate input could be light, heat, motion, moisture, pressure, or anyone of a great number of other environmental strange event. The output is generally a signal that is changed to human-readable display at sensor position or transmitted electronically over a network for reading or further processing. Two main categories of sensors: analog and digital.

Sensors are quantifying distances mobile robot from facade barrier, left barrier, and right barrier. According to intelligence acquired by robot sensors relevant fuzzy control rules are activated. The outputs of activated fuzzy rules based are combined by fuzzy logic operations of the control and operation of a motor vehicle of the robot.

2.1.1. Ultrasonic sensor (HC-SR04)

Ultrasonic sensor task output sound waves at a frequency too high for person to hear. They then watch the sound to be reflected back calculation distance basis on the time required. They measure distance no damage and are simple to use and regarded as reliable[7,8]. Ultrasonic sensor (HC-SR04) provides 2cm-400cm distance measurement task, the ranging precise can arrive to 3mm. The modules includes ultrasonic sensor transmitters, receiver and control circuit is shown in Figure 3.

2.1.2. HC-SR04 Sensor Features

- Operating voltage: +5V
- Theoretical Measuring Distance: 2cm to 400cm
- Practical Measuring Distance: 2cm to 80cm
- Accuracy: 3mm
- Measuring angle covered: < 15°
- Operating Current: < 15mA
- Operating Frequency:40Hz

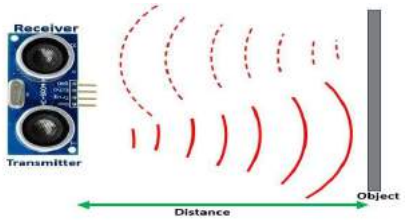


Figure 3. Ultrasonic sensor

$$Distance = \frac{Time \times Speed\ of\ sound\ in\ air(\frac{343m}{s})}{2} \quad (1)$$

2.2. Controller

2.2.1. Fuzzy logic Controller

Fuzzy Logic Controller is a method for decides a not accurate. Intelligent fuzzy logic controller for intelligent robot give authority the robot to repudiate the barrier and improve fined the target ability. While migrating to the barrier avoidance way to go, the robot changes it's heading. Three inputs of membership function in fuzzy logic controller are near, medium and far and using three

distance ultrasonic sensor for the amount of space from robot to left sensor, right sensor, center sensor the barrier. Fuzzy Logic is a solving problems control system methodology. It insert a simple, rule based IF X AND Y THEN Z approach to a solving control problem. The functions which part of it in the rule-based system. It is called *Membership Function*. The values that are issued by Membership functions are known *Linguistic Variables*. Fuzzy Logic allows an agent to profit inexactness in its collected information by allowing for endurance level. It can be particularly important when precise or state of being correct in a measurement is quite expensive. Three main portion events Fuzzy Logic System design are

(1) Fuzzification Unit
(2) Fuzzy Inference Engine
(3) Defuzzification

This process is shown in Figure 4.

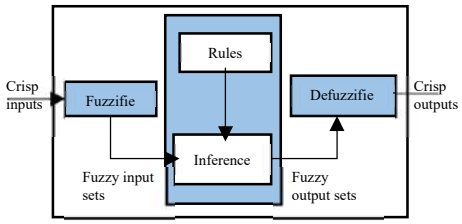


Figure 4. Fuzzy logic system

2.2.2. Fuzzification Unit

Fuzzification is the process of change a crisp value to a fuzzy value that is performed by the use of the data in the knowledge base. Fuzzification is the many types of curves can be seen in literature, Gaussian, triangular, and trapezoidal are the most used in the fuzzification process[1]. The condition of function translates accurate crisp input value into linguistic features variables is shown in Figure 5.

Using min-min method for Fuzzification,

$$£(x)=\text{Min (min, min)} \qquad (2)$$

Fuzzy Inference is the process of a mapping from input set to output set using fuzzy logic. The basic rules used in fuzzy logic are "if ... then ..." the basic rules are determined according to the wanted output. A set of IF-Then rules in Table 1 which contains a fuzzy logic quantity of the master's linguistic mention of how- to success good control.

Table 1. Fuzzy rules

No	Left	Center	Right	Robot direction
1	Near	Near	Near	Stop
2	Near	Medium	Medium	Right
3	Near	Far	Far	Right
4	Near	Far	Far	Straight Forward
5	Medium	Near	Near	Left
6	Medium	Medium	Medium	Straight Forward
7	Medium	Far	Far	Right
8	Medium	Far	Far	Straight Forward
9	Far	Near	Near	Left
10	Far	Medium	Medium	Left
11	Far	Far	Far	Straight Forward

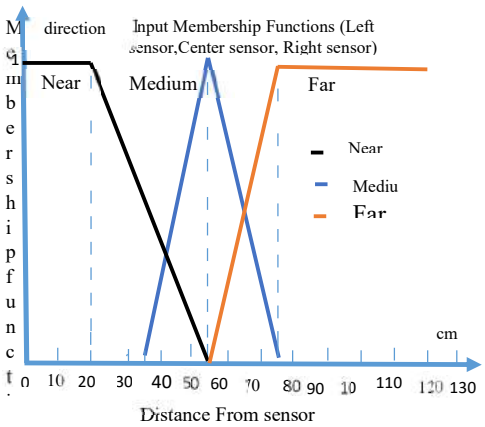


Figure 5. The fuzzy membership functions input

2.2.3. Defuzzification Unit

Defuzzification is the process of inversely a fuzzified output into only crisp value with respect to a fuzzy set is shown in Figure 7. The defuzzified value in FLC (Fuzzy Logic Controller) represents the action to be taken in controls the process. Mathematically, the process of defuzzification is also called rounding it off. Defuzzification is the process of change the degrees of the membership function of linguistic variables output within the linguistic terms and conditions into crisp statistical values[6]. There are many different methods of defuzzification able to be use, including the following:

The following are the known processes of defuzzification.

1. Center of Sums Method (COS)
2. Center of gravity (COG) / Centroid of Area (COA) Method
3. Center of Area / Bisector of Area Method (BOA)
4. Weighted Average Method
5. Maxima Methods

This process is based on Maxima Method(Height Method). It is based on Max-membership principles is shown in Figure 6.

$$\mu_C(x^*) \geq \mu_C(x) \text{ for all } x \in \quad (3)$$

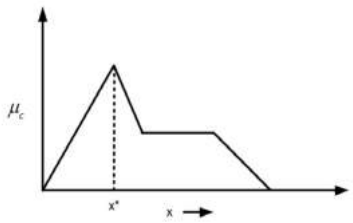


Figure 6. x* is the high place of the output fuzzified set C.

- First of Maxima Method (FOM)
- Last of Maxima Method (LOM)
- Mean of Maxima Method (MOM)

This system is used First of Maxima Method.

First of Maxima Method (FOM):

$$x* = \min\{x \mid C(x) = \max_w C\{w\}\} \quad (4)$$

Combine rule result:
Root-Sum Square Method:
For Defuzzification,

$$\text{Output} = \sqrt{(\sum_i rule^2)} \quad (5)$$

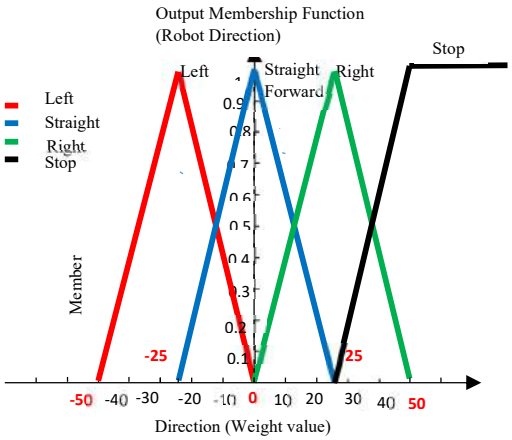


Figure 7. Robot direction output

2.2.4 Microcontroller

Microcontrollers are small computing systems used for low memory and low power consumption intention. A microcontroller consists in a microchip on a circuit board with read-write capabilities and a close-up picture. A microcontroller is a specific type of microprocessor. This development boards are known as Arduino Module, which are open-source prototype platform.

2.2.5 Arduino UNO

The microcontroller board used for this system is Arduino Uno which is based on ATmega328 microcontroller is shown in Figure 8. Arduino Uno ATmega328 microcontroller is describe in a detailed manner. Arduino software is installed in the computer and so that we can edit and upload the program according to the applications. Mostly the Arduino software supporting c++ and c programming languages. It is a programmable logic controller. The Arduino microcontroller is used in art and design as an open source programmable to improvise mutual benefit jobs. It can drive motors, LEDs, sensors and other components. In the process of put a decision, Arduino UNO is used for hardware configuration and programming. Once the code is written in Arduino IDE using C language, it

is uploaded into Arduino IDE and output is serial monitor. The data is collected from each sensor in an Excel sheet, and for integration between Arduino IDE and Excel sheet, another piece of code is written in Arduino IDE to access data from sensors. It has 14- digit input/output Pin (of which 6 can be used as PWM output), 6 analog inputs, a 16 MHZ quartz crystal, a USB connection, a power jack, an ICSP header and a reset button.

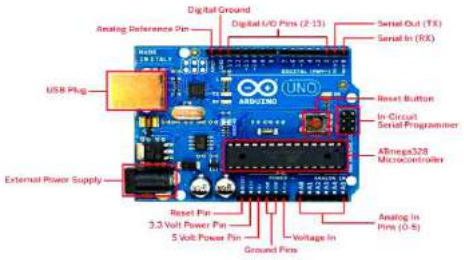
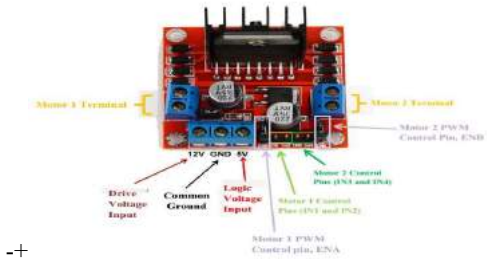


Figure 8. Arduino UNO board

2.3. Actuators

2.3.1. L298N Motor driver

Motor driver module is a simple circuit use for control a DC motor. The L298N motor driver in Figure 9 is a dual H-Bridge motor driver which grant permission direction and speed control of DC Motors at the same time. These L298N motor driver module is a high -power motor driver module for driving Stepper motors and DC motor.



-+

Figure 9. L298N motor driver

2.3.2. DC motor

An electric motor operated by direct current is known as a DC motor. A DC motor converts DC electrical energy into mechanical energy.

3.Working Principle

Depending on the condition of the mobile robot is able to choose the true path. A determination makes process of barrier avoiding the outside limit of an object area detection occurs as a result of a sudden impulse and without premeditation.

Step 1. Start

Step 2. Is any barrier?

(a) If yes, go away step 4.
(b) No, robot straight forward.

Step 3. IF any barrier?

(a) Yes, move robot backward.
(b) No, go away step 2.

Step 4. Range measurement to move right or Left.

IF left distance > right distance

(a) Yes, move left
(b) No, move right

Step 5. Continuous step 2.

Step 6. If power down go to step 7.

Step 7. Stop

The performance of system is measured with

$$\text{Speed} = \frac{Distance}{Time}$$

Table 2. Comparison of distance and speed based on time

No	Distance d (cm)	Time t (s)	Speed (cm/s)
1	692	40	17.3
2	685	38	18
3	670	32	20.9
4	588	25	23.52
5	560	18	31.1

The measurement of this system is shown in Table 2.

4. Circuit connection

There are four DC motors used for making mobile robot car. DC motor is interfaced between Pin 6, Pin 7, Pin 8, Pin 9, Pin 10 and Pin 11 of the L298N motor driver IC. The out 1 and out 2 pin motor drivers for DC motor. Pin A0, Pin A1, Pin A2, Pin A3, Pin A4, Pin A5, Ground, 5V connected between Arduino UNO R3 and three ultrasonic sensors[9]. This circuit diagram is shown in Figure 10.

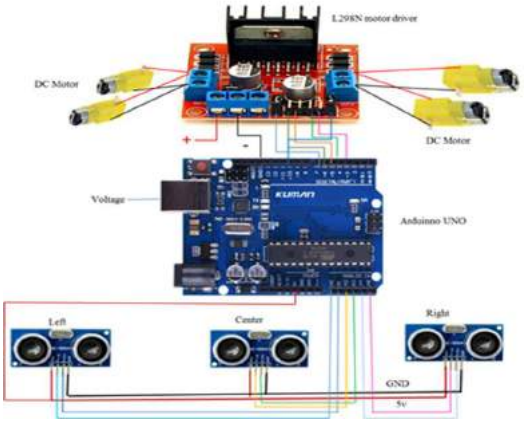


Figure 10. Circuit diagram

The practical implementation of proposed system is shown in Figure 11.

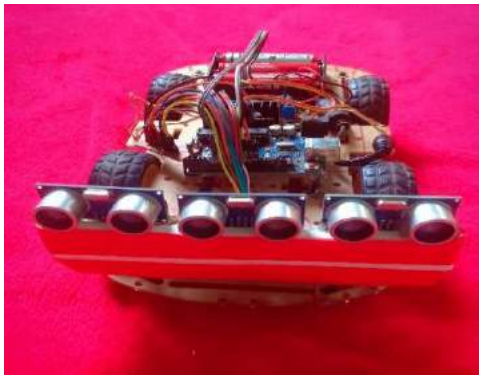


Figure11. Front view barrier avoidance robot

Barrier avoiding robots can be applied in almost all mobile robot navigation systems. They can be used for housework like automatic vacuum clean. This system can also be used in dangerous surrounding, where people pierce could be terrible. The design developed for use in risky area. Robotic service employee transports a heavy burden.

6. Conclusion

This system has accomplishing a purpose integrity completion a fuzzy control system scheme for adjust the heading and mobile robot speed. Algorithm of Fuzzy Logic for barrier avoidance were implemented in the mobile robot. Testing results with many conditions of barrier display the ability of mobile robot to avoid it
and have directed a good performance.

REFERENCES

[1]W.Martin, "Autonomous robot obstacle avoidance using a fuzzy logic control scheme," 2009.
[2]A.Dewan, "The Fuzzy Logic based Obstacle Avoidance Robot".
[3]W.C. Chiang, N.KelKar and E.L.Hall, "Obstacle Avoidance System with Sonar Sensing and fuzzy Logic".
[4]S.H.Lian, "Fuzzy Logic Control of an Obstacle Avoidance Robot".
[5]L.CHIA WOON, "Obstacle Avoidance Robot Applying Fuzzy Control System, july 2014".
[6]Timothy J.Ross, "Fuzzy Logic with Engineering Applications", Fourth Edition.
[7]Md.Arif Istiek Neloy, Vicky Barua, Mithun Das, Parthiba Barua, Shahid Uddin Rahat, Abhijit Pathak, " An Intelligent obstacle and Edge Recognition System using Bug Althorithm."
[8]Hai Prasaath k, "Efficient Wall following Robot with Ultrasonic sensor that Works in both Indoor and Outdoor Environments." August, 2017.
[9]Faiza Tabassum, Susmita lopa, Muhammad Masud Tarek & Dr.Bilkis Jamal Ferdosi, "Obstacle avoiding Robot".
[10]A.SHITSUKANE, W.CHERUIYOT, C.OTIENO MGALAMVURYA, "Fuzzy Logic Sensor Fusion for Obstacle Avoidance Mobile Robot."