DISCOVERING GENERALIZED ASSOCIATION RULE IN WEB USAGE MINING BY FREQUENT PATTERN TREE (FP-TREE)

HAN NI NI MYINT THU

M.C.Sc. **FEBRUARY 2019**

DISCOVERING GENERALIZED ASSOCIATION RULE IN WEB USAGE MINING BY FREQUENT PATTERN TREE (FP-TREE)

By

HAN NI NI MYINT THU

B.C.Sc. (Hons:)

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree

of

Master of Computer Science

(M.C.Sc.)

University of Computer Studies, Yangon

FEBRUARY 2019

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincere thanks to those who helped me with various aspects of conducting research and writing this thesis. To complete this thesis, many things are needed like my hard work as well as the supporting of many people.

First and foremost, I would like to express my deepest gratitude and my thanks to **Prof. Dr. Mie Mie Thet Thwin**, Rector, the University of Computer Studies, Yangon, for her kind permission to submit this thesis.

I would like to express my appreciation to **Dr. Thi Thi Soe Nyunt**, Professor and Head of Faculty of Computer Science, the University of Computer Studies, Yangon, for her superior suggestion, administrative supports and encouragement during my academic study.

My thanks and regards go to my supervisor, **Dr. Khine Khine Oo,** Professor, Faculty of Information Science, the University of Computer Studies, Yangon, for her support, guidance, supervision, patience and encouragement during the period of study towards completion of this thesis.

I also wish to express my deepest gratitude to **Daw Htet Htet Aung**, Lecturer, Language Department of the University of Computer Studies, Yangon, for her editing this thesis from the language point of view.

Moreover, I would like to extend my thanks to all my teachers who taught me throughout the master's degree course and my friends for their cooperation.

I especially thank my parents, all of my colleagues, and friends for their encouragement and help during my thesis.

STATEMENT OF ORIGINALITY

I hereby certify that the work embodies in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

---------- ----------

Date Han Ni Ni Myint Thu

ABSTRACT

Web mining techniques can be used to search for web access patterns, web structures, regularity and dynamics of web contents. Web usage mining analyzes Web log files to discover user accessing patterns of Web pages. Log file data can offer valuable insight into web site usage. It reflects actual usage in natural working condition, compared to the artificial setting of a usability lab. This system presents web log mining based on hierarchy of web usage data by generalized association rule. Multi-level association rule is used for implementation of generalized association rule. In this system, Web log database is used to store web access records in log files collected from web server. And web log databases are constructed via a process of data cleaning, data transformation. By using Frequent Pattern Tree (FP- Tree), the system generates rules from web log data, and reduces counting phase of association rule since it stores the pre-computed count values. Frequent patterns are generated instead of page item-sets. The generated frequent patterns can later be applied to improve web site management, decision making process.

This system is implemented using ASP.Net programming language with Microsoft SQL Server 2008 R2 Database Engine.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Web plays a consequential sector and information dissemination medium. There is a log of data to trace any history of communications transaction. Web mining is data mining mechanism to analyse and exploit subsidiary data documents and accommodations from the World Wide Web. Web utilization mining is the application that utilizes data mining to search intriguing of data user's utilization patterns on the web.

Web servers maintain a (Web) log ingress for every single access they get in which they preserve the requested URL, the request originated IP address. Weblog data are extracted from web log files and need to be cleaned and transformed. And then these weblog data are loaded into a data warehouse for mining multilevel predicated patterns. Sodality rule is applied to get a relationship between each dimension of weblog data. In this system, mining sodality rules in a hierarchical database will be applied in web utilization mining.

Sodality rule mining has magnetized concern in both analysis and domain areas. The multi-level mining sodality rule is a component of the consequential branches. Mostly, at every single level of hierarchy concept, multilevel rules will be searched by tracing the database perpetually. It affects the mining algorithms' efficiency, integrality, and excitability on the multilevel rules. This system describes exploiting multilevel rules predicated on various levels of hierarchy by configuration and searching frequent item-sets extracted from primitive items of data. Its searching process is predicated on hierarchies of dynamic concept.

1.1 Web Mining

Web mining is a very sultry research concept that cumulates two of the active research environments: (1) Data Mining, and (2) World Wide Web. They comprise the revelation and analysis of data, documents, and various media from the Web. Web mining is applying data analysing techniques to discover and draw the utilizable data from Web documents and accommodations. Research on extracting regularities in the improvement of the users processing a website can ameliorate the performance of the system and also improve the quality and broadcasting information on Internet

accommodations to the cessation utilizer, and determine the number of dedicated customers for e-commerce. Web Mining is the domain of data mining mechanism on web data. Web mining is categorized into three categories. They are:

1. Web content mining play with the utilizable revelations contents of web and accommodations,
2. Web structure mining is studying of the hyperlinks structure within the web itself, and
3. Web utilization mining additionally kenned as Weblog mining is the application of data processing techniques that search the utilization forms of users from the web data.

1.2 Objectives of the Thesis

The main objectives of the proposed system are:

- To learn web mining algorithms, web utilization mining, and applications of web utilization mining algorithms
- To understand frequent mining algorithms, the process of sodality rule mining and hierarchy of sodality rule
- To apply frequent pattern mining FP-growth algorithm in web utilization mining for pattern revelation and pattern analysis
- To implement the generalized sodality of web utilization mining for the process of future web presage.

1.3 Related Works

Discovering utilization pattern from web log data by sodality rule mining is presented in [1]. SOTrieIT algorithm and Apriori algorithm are utilized for mining sodality rule. In this thesis, sodality rule is engendered for browsing patterns of a web page. Those rules can later be applied recommender system and soothsaying utilizer comportments. Impotency of Apriori algorithm is time-consuming.

Apriori algorithm is applied to data cube so it reduces the counting step of the algorithm since the data cube stored in pre-computed value [8]. Discovering sodality rules from OLAP data cube with daily downloads of Folklore Materials is presented. In their system, sodality rules predicated on dimensions in lieu of browsing page are engendered. It is expeditious in processing since data cube stores pre-computing counts.

The analysis is accentuated for all types of documentsrather than browsing pages. Besides, only four dimensions of documents are analysed in their system.

Web server log files are acclimated to discover the sodality rules and present the relationship of the dimension of these log files. Sodality rule mining method on OLAP Cube is presented. In their system, variants of OLAP database, ROLAP, MOLAP, and HOLAP are presented. Then student information databases are habituated to discover sodality rule from the data cube. It will give the frequent items and rules from the data cube. [7]

Web mining is described and accentuate on web utilization mining. Raw web log files are pre-processing and stored in the warehouse database. The data cube is engendered from the warehouse database and find the sodality rules from it so surmounts the quandary of Apriori's algorithm.

1.4 Organization of the Thesis

The organization of the thesis can be divided into five chapters. They are as follows:

In Chapter 1, the introduction of the system, the objectives of the thesis, related works and thesis organization are described.

Chapter 2 presents the theoretical foundation concerning web usage mining. Web usage mining is web log mining which reveals the knowledge hidden in the log files of a web server.

Chapter 3 describes the design of the proposed system. This section will present web usage mining by generalized association rules. Generalized Association Rule – Rules based on Taxonomy / Hierarchy of product or page in this web usage mining Taxonomy and hierarchies are defined based on attributes of web log records such as browsers, referrer, status, URL and so on.

Chapter 4 expresses the step by step implementation of the system. This system presents generating association rule patterns from weblog data. In order to analyse the web usage patterns, association rule mining algorithm is applied to weblog. Relationships of web server log are displayed as output in this system.

Finally, Chapter 5 presents the conclusions, advantages of system and limitations and further extensions of the system.

CHAPTER 2
BACKGROUND THEORY

The Web serves as a widely distributed information centre for news, advertisements, business management, academic, government, e-commerce, and so on. The Web also contains a large number of information links and numbers of the Web page for processing information to give enormous data for mining. Web mining can improve the efficiency of the search engine because of mining on the web may determine on relative Web pages, classification on page usage, and subtleties raised Web search which is based on keyword [3].

Because of the mining of web, consumer-facet aspect and server-aspect statistics can accumulate consumer desire facts, or won the database relation of an employer (which includes data for business records or web-associated data). Mining on web records might also fluctuate now not handiest the records supply area, but also available facts type, the collected facts populace phase, and its implementation. In web utilization mining, there are various kinds of information can be used.

Content: web pages' actual data, i.e. the data structure is designed to suit the user need. There may be an unlimited amount of data, text, image, etc.

Structure: The statistics that affords the components of the agency. Shape of sub-web page information includes sizable html or xml control in a web page. It additionally describes a tree shape, wherein the basis of the tree is the <html> tag. The fundamental form of sub-page statistics structure is hyper-hyperlinks speaking between web pages.

Usage: The tracking of the consumer's utilization history, for example, IP address, session, and the accessed date-time.

User Profile: User profile is the information of web utilized user's information for the management of user preference prediction.

Internet mining is a way that is used to examine and searching information that consists of at the web page of devoted servers. The principle aim of mining on web records is to guide a technique to method the net information more efficaciously and exactly. The remaining concern method is to go looking the information that is extracted from the customers' activities which can be saved in log files (e.g. Web caching estimate. Therefore, internet utilization analysing may be classified into three

sorts of lessons based on the analysed person necessities. The categorised instructions are (i) mining on the content of the internet, (ii) mining at the shape of the net and (iii) mining on the use of the net.

Internet content material mining is the challenge of discovering useful information available on-line. There are one-of-a-kind styles of internet content cloth that could offer beneficial records to users, for example multimedia records, structured (i.e. Xml files), semi-based totally (i.e. Html files) and unstructured facts (i.e. Simple textual content). The purpose of web content material fabric mining is to provide an efficient mechanism to help the customers to discover the records they are searching for. Internet content mining includes the undertaking of organizing and clustering the files and imparting search engines like Google and Yahoo for gaining access to the one-of-a-type documents by key phrases, classes, contents and lots of others.

Web structure mining is the approach of coming across the shape of links in the web. Nearly, at the same time as net content material mining specializes in the internal-report facts; internet structure mining discovers the link structures on the inter-report diploma. The aim is to identify the authoritative and the hub pages for a given undertaking. Authoritative pages include useful data and are supported by means of numerous hyperlinks pointing to it, which means these pages are distinctly referenced. If a page has numerous referencing hyperlink techniques, the content material of the page is beneficial, best and possibly reliable. Hubs are web pages containing many links to authoritative pages, for that reason, they assist in clustering the authorities. Internet form mining may be accomplished handiest in an unmarried portal or also on the complete internet. Mining the structure of the internet allows the assignment of internet content material fabric mining. Using the data about the structure of the web, the document retrieval may be made extra green, and the reliability and relevance of the observed documents can be extra. The graph form of the internet may be exploited through internet shape mining so as to improve the overall performance of the records retrieval and decorate category of the documents.

Web utilization mining is the challenge of discovering the sports of the customers on the equal time as they're surfing and navigating via the net. The purpose of information navigation choices of the website online visitor is to beautify the brilliant of virtual trade services (e-exchange), to customize the internet portals or to improve the net form and net server performance. In this example, the mined data are the log

documents which can be seen as the secondary information at the internet wherein the documents on hand through the net are understood as number one information.

There are three kinds of log files that may be used for net utilization mining. Log documents are stored at the server facet, at the purchaser side and at the proxy servers. With the aid of getting multiple vicinity for storing the data of navigation varieties of the clients makes the mining method more hard. In reality, reliable effects might be obtained best if one has records from all three styles of log record. The cause for this is that the server element does no longer incorporate records of these net page accesses which may be cached at the proxy servers or on the customer aspect. Besides for the log report on the server, that on the proxy server presents additional statistics. However, the net web page requests saved within the consumer facet are missing. Yet, it is intricate to collect all the statistics from the purchaser thing. For this reason, most of the algorithms work primarily based best on the server component statistics. A few usually used records mining algorithms for internet utilization mining are association rule mining, sequence mining and clustering.

2.1 Web Content Mining

Web content material cloth mining describes the automated seek of information resource available on line and includes mining net records contents. In the internet mining vicinity, net content fabric mining basically is an analogue of statistics mining strategies for relational databases, considering that it's miles possible to discover comparable forms of knowledge from the unstructured information dwelling in net files. The net document usually carries severe varieties of facts, along with text, picture, audio, video, metadata, and links. Some of them are semi-dependent in conjunction with html documents, or greater dependent facts just like the statistics inside the tables or database generated html pages, however most of the information is unstructured text statistics [14]. Web content material mining additionally known as text mining is commonly the second one step in web information mining. Content fabric mining is the scanning and mining of textual content, pixels, and graphs of an internet web page to decide the relevance of the content to the quest query.

2.2 Web Structured Mining

Web-based mining makes a speciality of the hyperlink based on the net. The intention of internet based mining is to generate structural précis approximately the web page and web pages. Technically, net content mining mainly makes a distinctiveness of the established of internal-record, even as net installed mining tries to find out the link structure of the links on the inter-document level. Based totally on the topology of the hyperlinks, internet installed mining will categorize the internet pages and generate the data, at the side of the similarity and courting among remarkable internet websites. In famous, if a web page is hooked up to each different net page right away, or the internet pages are associates, we would like to find out the relationships amongst the only internet pages. The members of the family may be fall in one of the kinds, inclusive of them related by using synonyms or ontology, they'll have similar contents, and both of them can also take a seat within the equal internet server therefore created by using the equal man or woman. Every other task of net form mining is to find out the man or woman of the hierarchy or community of hyperlinks within the websites of a specific domain. This may help to generalize the float of facts in net websites which can constitute a few unique areas; consequently, the query processing is probably less complex and extra green [13]. The project for net dependent mining is to deal with the established of the hyperlinks within the internet itself. Net structure can be handled as part of internet content cloth so that internet mining has classes: web content material mining and internet usage mining.

2.3 Web Usage Mining

Web utilization mining attempts to discover the useful statistics from the secondary facts derived from the interactions of the customers even as browsing at the internet. Net utilization mining mines blog statistics to discover purchaser get admission to styles of net pages. It makes a distinctiveness of the techniques that could anticipate individual behaviour whilst the patron interacts with the net. Analyzing and exploring regularities in weblog records can understand functionality clients for virtual alternate by way of manner of predicting the purchaser’s behaviour within the website, evaluating among expected and real web page utilization, enhance the fine and transport of net records offerings to the surrender user, improve internet server gadget performance and adjustment of the internet net page to the hobby of its consumers. A

web server normally registers an (internet) log entry, or weblog get entry to, for absolutely everyone gets entry to of a web page. It consists of the URL requested, the IP address from which the request originated, and a timestamp. For net-based totally e-exchange servers, a huge amount of net get admission to log data is being amassed. Famous web sites may additionally register weblog facts within the order of masses of megabytes each day. Weblog databases offer rich records about net dynamics. For this reason it's miles essential to increasing today's blog mining strategies [3]. Internet utilization mining is parsed into three one-of-a-kind segments:

- **Pre-processing:** plays a chain of processing of net log report protecting statistics cleansing, person identification, and session identity.
- **Pattern discovery:** application of numerous facts mining techniques to processed records like statistical evaluation, affiliation, clustering, sample matching and so on.
- **Pattern analysis:** once patterns have been located from weblogs, dull regulations are filtered out. The analysis has achieved the usage of information question mechanism together with square or records cubes to perform OLAP operations [8].

2.3.1 Data Pre-processing

Data pre-processing targets to reformat the authentic internet logs understand all internet get admission to durations. The net server typically registers all of the clients get right of entry to sports activities via the internet server logs. Because of the only of a type server placing parameters, they produce many forms of internet logs. Nevertheless, the log documents incorporate comparable number one facts along with patron IP address, request time, requested URL, http recognition code, referrer and so on and so on.

Generally, numerous pre-processing obligations are required to be executed in advance than using net mining algorithms at the internet server logs. The precise server logs are wiped smooth, formatted, and then grouped into meaningful lessons earlier than observe the data mining strategies. Commonly, pre-processing includes data cleansing, person identification and session identification [8].

2.3.1.1 Data Cleaning

The information cleaning manner eliminates the statistics tracked in internet logs which might be useless or inappropriate for mining purposes. The request processed through auto engines like Google, inclusive of crawler, spider, and robot, and requests for graphical net page content material (e.g., jpg and gif) are deleted due to the truth these picture documents are vehicle-downloaded with the requested pages. The principle of statistics cleaning is to reduce extraneous items, and those kinds of strategies are of importance for any sort of net log evaluation not best records mining. In keeping with the cause of various mining packages, extraneous data on internet get right of entry to log may be removed during information cleansing. For the reason that goal of web usage mining is to get the individual's get right of entry to patterns, following sorts of information are pointless and should be eliminated:

- The records of pictures, movies, and the layout records: the information have filename suffixes of a gif, jpeg, css, and so forth, which can be located inside the URL region of every file.
- The facts with the failed http reputation code: through grouping the popularity area of every record in the net access log, the statistics with reputation codes over 299 (or) below two hundred are removed.

2.3.1.2 User Identification

The user identity technique analyses the log document and clusters the users in order that all and sundry inside the identical institution has the identical get right of access to traits. This approach is to categorize who get entry to web website online and which pages are accessed. The aim of session identification is to divide the net web page accesses of anyone at a time into individual training. Particular customers may additionally moreover have the equal IP cope with inside the log. A referrer-based approach is proposed to solve the ones problems in this take a look at. The rules observed to differentiate person periods may be defined as follows:

The one-of-a-type IP addresses distinguish one of the a-type-users. If the IP addresses are equal the unique browsers and operation structures suggest one of the a-type-clients. Inside the individual identity section, the unique customers are tremendous and one of the a-type-customers is identified. This can be executed in various

approaches like the use of IP addresses, cookies and so forth. It’s tough because of safety and privacy use the following heuristics to perceive the patron:

- Each IP deal with represents one patron.
- For more logs, if the IP cope with is the same, but the agent log indicates a trade in browser software or working system, and IP address represents a one of a kind person.
- If a page is asked that isn't always at once available thru a link from any of the pages visited by using the consumer, there may be every other individual with the equal IP address.

2.3.1.3 Session Identification

As quickly as the log documents had been wiped clean, the following step within the records pre-processing is the identification of the consultation. Session identification is the technique of segmenting the individual activity log of each consumer into agencies of web page references within the direction of 1 logical length called session. The purpose of consultation identification is to divide the page accesses of everyone, who's probable to visit the net page greater than once, into character training. The technique to understand someone session encompasses timeout mechanism and maximal ahead reference.

- If there may be a latest client, there is a modern consultation.
- In a single patron consultation, if the referring page is null, there is a modern day session.
- If the time among web page requests exceeds a effective restriction (30 minutes), its miles assumed that the customer is starting a new consultation.

2.3.2 Pattern Discovery

Pattern discovery attracts upon the algorithms and strategies from numerous studies regions, which includes information mining, gadget mastering, information, and sample recognition. The subsections of sample discovery are detail mentioned as follows.

2.3.2.1 Statistical Analysis

Statistical strategies are the maximum commonplace technique to extract know-how about web page traffic to an internet web page. By way of analyzing the consultation report, one could perform distinct kinds of descriptive statistical analyses (frequency, advise, median, and many others.) on variables collectively with page views, viewing time and length of a navigational course. Many net web page site visitors' analysis equipment produces a periodic report containing statistical records consisting of the most frequently accessed pages, commonplace view time of a page or not unusual period of a course via a site. With the useful resource of analysing the statistical statistics contained within the periodic net machine file, the extracted report may be doubtlessly useful for boosting the gadget overall performance, improving the protection of the system, facilitation the website on line trade assignment, and imparting assist for advertising selections.

2.3.2.2 Association Rules

Association rule generation can be used to relate pages that are most customarily referenced together in a single server session. Inside the context of net utilization mining, association guidelines confer with sets of pages which might be accessed together with a aid value exceeding some distinctive threshold. The support is the share of the transactions that incorporate a given sample. As an instance, affiliation rule discovery the usage of the Apriori algorithm can also reveal a correlation among users who visited a web page containing digital products to folks that get admission to a web page about carrying equipment. The net designers can restructure their web sites successfully with the help of the presence or absence of the association policies. The affiliation guidelines may also serve as a heuristic for pre-fetching files so that you can reduce person-perceived latency when loading a page from a faraway site.

2.3.2.3 Clustering

Clustering is a technique to organization collectively a hard and fast of items having similar characteristics. Within the internet usage area, there are styles of exciting clusters to be determined: utilization clusters and page clusters. Clustering of customers has a bent to set up groups of customers showing comparable surfing patterns. Such statistics is in particular beneficial for inferring consumer demographics as a manner to

perform market segmentation in e-trade programs or provide custom designed net content material to the makes use of. But, clustering of pages will find out businesses of pages having associated content material cloth. This record is beneficial for internet search engines like Google like Google and web assist vendors.

2.3.2.4 Classification

Classification is the approach to map a data object into one in every of numerous predefined lessons. In the web domain, one is inquisitive about developing a profile of customers belonging to a particular magnificence or class. This calls for extraction and desire of capabilities that best describe the houses of a given class or category. The class can be completed through the usage of supervised inductive gaining knowledge of algorithms which includes selection tree classifiers, naïve Bayesian classifiers, ok-nearest neighbour classifiers, assist vector machines and so forth. For instance, a class on server logs may additionally cause the invention of exciting recommendations such as 30% of users who placed an online order in /product/track are within the 18-25 age group and stay on the west coast.

2.3.2.5 Sequential Pattern

The technique of sequential pattern discovery attempts to find out inter-consultation patterns such that the presence of a set of devices is followed via any other object in a time-ordered set of classes or episodes. Through using this method, net marketers can assume destiny go to patterns for you to be useful in setting classified ads aimed at effective person agencies. Distinctive forms of temporal evaluation that can be finished on sequential patterns consist of trend evaluation, alternate issue detection, or similarity evaluation.

2.3.2.6 Dependency Modelling

Dependency modelling is every other beneficial pattern discovery challenge in web mining. The purpose is to set up a version this is able to constituting widespread dependencies some of the numerous variables within the internet domain. As an instance, one can be interested to assemble a version representing the wonderful tiers a traveller undergoes whilst shopping in a web preserve based completely on the moves chosen. There are numerous probabilistic studying strategies that can be employed to

version the browsing behaviour of clients. Such techniques include hidden Markov models and Bayesian belief networks. Modelling of net usage patterns will not handiest provide a theoretical framework for analyzing the behaviour of users however is probably useful for predicting destiny web aid consumption. Such information may additionally moreover assist increase strategies to increase the profits of products furnished via the internet net page or beautify the navigational consolation of clients.

2.3.3 Pattern Analysis

Pattern analysis is a very ultimate degree of the whole internet utilization mining. The purpose of this approach is to cast off the irrelative guidelines or patterns and to extract the thrilling rules or patterns from the output of the pattern discovery system. There are two maximum not unusual tactics for the sample evaluation. The maximum not unusual shape of pattern analysis includes a understanding query mechanism which includes rectangular. Every different technique is to load utilization information proper into a statistics dice in order to perform OLAP operations. Visualization strategies, consisting of graphing patterns or assigning colours to distinctive values, can often highlight widespread styles or dispositions inside the records [12].

2.4 Web Server Logs

Web servers hold log entries for all verbal exchange which might be having access to their web sites, and the dimensions of those log files are developing by tens of megabytes each day. Server logs reveal the large quantity of facts about visitors, server behaviour, adjustments in websites, and ability blessings of new technical tendencies. It does not report reader behaviours like common backtracking or not unusual reloading of the equal useful resource. On the arena large internet (www), logs of http site visitors are recorded constantly as a characteristic of most beginning web servers as well as intermediate proxies. The primary function of those logs is to chronicle the operation of these structures. Logs generated with the aid of way of those systems are also used for characterization, evaluation and utilization reporting.

Net server logs are undeniable text (ASCII) document that is unbiased from the server platform.

There are differences among server software program, however historically there are four varieties of server logs:

1. Switch log
2. Agent log
3. Error log
4. Referrer log

The first sorts of log files are trendy. The referrer and agent logs might also or won't be "grew to become on" at the server or can be delivered to the switch log record to create a "prolonged" log file format.

2.5 Web Mining in Societal Benefit Areas

Web mining may benefit the ones companies that want to make use of the web as a knowledge base for supporting preference-making. Pattern discovery, evaluation, and interpretation of mined patterns can also additionally lead to better decisions for the company and for the provided offerings. E-commerce and e-business agency have been two fields empowered through using internet mining having masses of applications for increasing income and doing intelligence-corporation. Plenty of internet mining packages determined in the literature described the effectiveness of the software from the internet management factor of view. The goal in those applications is taking benefit of the mined information from the users to increase the advantages for the organization.

In our technique the attention is on social useful regions from net mining, so our factor of view is on net mining applications that could assist customers or group of customers. An apparent societal gain is that internet mining research efforts bring about person (or organization of users) pride through providing correct and applicable records retrieval; via using imparting customized records; with the resource of mastering approximately person’s demands in order that services can goal particular corporations or even individual customers; and via presenting custom designed offerings. Framing web mining in societal useful regions, we want to outline what areas or domains are considered as socially useful, a difficult query which routes to philosophical and nice of lifestyles factors. The viable wide variety of domains is large. Material nicely being, health, productivity, intimacy, safety, community, and emotional properly-being are domain names suggested in that are related to satisfactory of lifestyles. Within the

domains: individual and households, network participation, environment, fitness, education, bodily safety, monetary protection, subculture and entertainment, social offerings, and authorities are symptoms investigated for the quality of lifestyles.

On this paintings, we frame social beneficial areas primarily based on seek consequences wherein we first diagnosed net mining research in numerous areas that may be of social interest [5]. Then we attempted to the employer the identified research into associated categories. The cease end result for our societal beneficial regions research divides net mining studies into six social useful areas:

1. E-learning to know, in which academic improvement is the gain;
2. E-government, wherein development of government offerings to citizens is the gain;
3. E-politics and e-democracy, where network participation is the gain;
4. Helpdesks and recommendation structures, in which the improvement of social offerings is the advantage;
5. Virtual libraries, where improvement of productivity via diffusion of ideas is the gain;
6. Security and crime research, in which public protection is the advantage.

The primary three regions are taken into consideration to be a part of what is extensively called e-offerings. We can see in the next classes on how current web mining research is associated with the above social useful areas.

2.5.1 E-Learning

Web mining may be used for enhancing the getting to know gadget in e-getting to know environments. Applications of web mining to e-studying are usually internet-utilization primarily based packages. Brin S introduces a framework, in which they use logs to analyse the navigational conduct and the general performance of e-learning just so to customize the analyzing content material of adaptive learning environment in an effort to make the learner attain his analyzing goal. Zaiane studies using machine getting to know strategies and net utilization mining to beautify net-based totally learning environments for the educator to better observe the getting to know the procedure and for the learners to assist them in their mastering venture. University college students' net logs are investigated and analysed in Cohen and Nachmias in a web-supported guidance model.

2.5.2 E-Government

The manner with the aid of groups that engage with residents for pleasing consumer (or corporation of customers) selections results in higher social services. The principle traits of E-government' systems are associated with using technology to deliver offerings electronically that specialize in residents needs by way of supplying properly enough facts and more ideal services to residents to aid political behaviour of government. Empowered through way of web mining strategies e-authorities' structures might also moreover provide custom designed offerings to citizens ensuing to person pleasure, nice of offerings, guide in citizens' choice making, and in the long run leads to social advantages. Such social blessings an entire lot depending at the agency's willingness, understanding, and capability to transport on the quantity of using net mining.

The E-government measurement of a set is generally carried out frequently. E-government maturity fashions describe the web stages a company goes via time, turning into extra mature in using the net for presenting higher offerings to residents. The adulthood degrees start from the organisation's first try to be online aiming at publishing useful to citizens' records and flow to higher maturity ranges of being interactive, making transactions and ultimately reworking the functionality of the business enterprise to operate their business organisation and offerings electronically via the net. Adulthood tiers defined in literature doesn't have an internet mining size, which we remember that have to be the climax in maturity tiers. Riedl states that through using interviews and internet mining, the real access to statistics thru citizens need to be tracked analysed and used for the redesign of e-government' data services.

The E-government literature is famous that simplest currently internet mining has attracted researchers in e-government programs. Fang and Sheng gift an internet mining technique for designing higher net portal for e-government. Hong and lee suggest a wise web records gadget of presidency primarily based on web usage mining to assist deprived users to make specific desire-making for their profits enhancements. Within the fitness quarter upgrades of fitness services thru exquisite labelling of medical net content fabric are investigated.

2.5.3 E-Politics and E-Democracy

E-politics gives political statistics and politics "on demand" to the citizens improving political transparency and democracy, benefiting events, applicants, citizens, and society. Election campaigners, occasions, participants of parliament and individuals of nearby governments on the internet are part of e-politics. Regardless of the importance of e-politics in a democracy there's a restricted net mining methods to fulfil citizen dreams. We diagnosed in the literature research that handiest refers in mining political corporations, being net-shape based definitely efforts. Link assessment has been used to estimate the scale of political web graphs, to map political events' community at the internet and to analyze the U.S. Political blogosphere. Political internet linking is also studied by using foot. Throughout the U.S. congressional election's marketing campaign season at the internet.

2.5.4 Helpdesks and Recommendation Systems

Advice systems are based on consumer modelling especially derived from content material fabric-based on studying or from collaborative filtering. Content-primarily based definitely mastering makes use of a consumer's past usage conduct and acts as an indicator of his future behaviour. Collaborative filtering is based on rankings of consumer favours, like score track or films, so as, that rating data of a patron may be associated with comparable alternatives of other clients. So a client is classified in a person model, wherein pointers can be addressed to customers consistent with favours of different human beings from an appropriate categorized consumer version. Hybrid recommendation structures that take benefits from each collaborative filtering and content material-primarily based totally learning were moreover investigated within the literature. Galina Bogdanova indicates an advice version for predicting consumer alternatives based on common clustering algorithms in a citizen internet portal.

A blog is a magazine-style website normally written with the aid of an unmarried consumer, wherein entries are presented in reverse chronological order. Referral web is a assignment that mines social networks from the internet via the use of collaborative filtering for figuring out professionals that could solution questions requested by means of the use of human beings. Park JS used neural networks provide accurate web tips based on a committee of predictors. Created web page-prompter an

agent-based completely advice model that helps customers navigate an internet site via analysing and learning from internet usage mining and person behaviour.

The exciting a part of net page prompter is that it combines three particular system studying techniques: association rules, clustering and choice trees for reaching its challenge. Park JS also makes use of clustering implemented to internet utilization mining for developing community unique directories to provide customers an extra personalized view of the internet in line with their options and can be assisted by using the use of those directories as beginning elements on their navigation. Sarawagi created an e-mail answering assistant by means of using semi-supervised textual content kind.

2.5.5 Digital Libraries

Digital libraries provide precious facts distributed all over the globe without usually having the want to be bodily present in a conventional library constructing. Internet mining studies aiming to higher offerings on virtual libraries have been identified within the literature. Agrawal R uses clustering for coming across lacking hypertext hyperlinks in Wikipedia, the biggest Encyclopedia on the net this is created and changed through many volunteer authors. Net mining on Wikipedia is likewise investigated thru Galina Bogdanova. Al. Brin S uses clustering for detecting group of entities, like authors, from links and resolving the co-reference trouble of more than one reference to the identical paper in self-reliant quotation indexing engines, like Chen. Chen is a crucial aid for laptop scientists for searching digital versions of papers. Chau. Al work also is based totally on each different famous virtual library in laptop technological information community, the library at http://dblp.Uni-trier.De. Their work is associated with system gaining knowledge of function extraction algorithms in case you want to find out hidden groups in heterogeneous social networks.

Graph analysis for coming across net groups may be modelled with the aid of the usage of Bayesian networks as confirmed through way of Goldenberg and more for figuring out co-authorship networks. Content material-based totally book advice device is proposed through cash, and Roy based totally on net pages of the Amazon on-line digital maintain. Large portals with information up to date regularly consistent with day encompass rich statistics and may be taken into consideration as part of virtual libraries inside the way that newspaper articles are indexed and to be had to readers in traditional libraries. News web sites are huge portal web sites that boom their content material on

a daily foundation. For such sites, the interpretation of internet content material cloth to significant content material that can be classified into semantic lessons in order to make both information retrieval and presentation less difficult for people and institution of users will be very essential. Liu et. Al uses fuzzy clustering to discover meaningful information styles from internet news circulation records.

2.5.6 Security and Crime Investigation

Web mining strategies are used for figuring out cyber-crime actions like internet fraud and fraudulent web sites, unlawful online playing, hacking, virus spreading, child pornography distribution and cyber terrorism. Chen, Chung, et. al., the word that clustering and class strategies can monitor identities of cyber-criminals, whereas neural networks, selection timber, genetic algorithms and assist vector machines can be used to crime patterns and network visualization. Chen. Offer an in-intensity study on techniques toward terrorist groups at the web for predictive modelling, terrorist network assessment, and visualization of terrorists' sports activities, linkages, and relationships. Wang primarily based on the character's online sports use predominant cluster evaluation to perceive a small wide variety of number one subjects from lots and thousands of navigational statistics in an approach that may be useful in protection in the direction of terrorism. Dong Xin executed a gadget that can be beneficial to secure internet surfing in college, domestic and administrative centre. The tool monitors and filters net get proper of entry to with the aid of way of making use of internet mining for performing internet facts class to be able to classify web facts in a "whitelist" of allowed pages or blacklist of blocked internet pages. Social networks extracted from instant messaging with the aid of the usage of using clustering are investigated via Yinbo Wan.

2.6 A Survey on Association Rule Mining

Facts mining is the analysis step of the KDD (know-how discovery and information mining) method. It is defined as the system of extracting interesting (non-trivial, implicit, formerly unknown and beneficial) data or styles from massive facts repositories together with: a relational database, statistics warehouses, and many others. The aim of the records mining manner is to extract data from a record set and remodel it into an understandable shape for further use. Record mining has been given much

interest in database groups due to its wide applicability. The problem of mining affiliation rules from transactional database become delivered in [11].

The idea ambitions locate frequent patterns, thrilling correlations, and associations amongst sets of items in the transaction databases or different facts repositories. Association rules are getting used extensively in various regions along with telecommunication networks, risk and marketplace management, stock control, clinical diagnosis/drug checking out etc. Affiliation rule is the statements that locate the relationship between statistics in any database. Affiliation rule has components "antecedent" and "consequent".

In an example for bread => eggs, bread is the antecedent and egg is ensuing. The antecedent is the item this is determined inside the database, and the consequent is the item that is discovered in combination with the primary. An extra formal definition can be given as: let i = i_1, i_2, …, in be a hard and fast of gadgets. Permit d be a set of project applicable statistics transactions wherein each transaction t is fixed of gadgets such that t ¬ i. A completely unique tid is related to every transaction. Permit a be a hard and fast of objects. A transaction t is said to contain a if and best if a ¬ t. An association rule is the implication of the form a ¬ b, in which a ¬ i, b ¬ i, and a ∩ b = null. Association rule mining is executed to find out affiliation rules that fulfil the predefined minimal assist and confidence from a given database. The hassle of locating affiliation rule is normally decomposed into sub-issues as shown in figure 2.1.

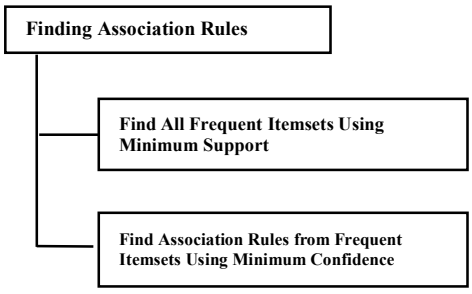


Figure 2.1 Generating Association Rules

As shown in figure 2.1, one sub hassle is to find the ones item-sets whose occurrences exceed a predefined threshold within the database, those item-sets are known as large or frequent item-sets. The second sub-problem is to generate association guidelines from those large item-sets with the constraints of minimal self -belief. Assume one of the large item-sets is t_k, t_k = i_1, i_2, …, i_k, association policies with this item-sets are generated within the following manner: the primary rule is i_1, i_2, …,

$i_{k-1}\neg\, i_k$, by way of checking the confidence this rule is determined as thrilling or no longer. Then the remaining regulations are generated via deleting the final objects within the antecedent and putting them to the consequent, thereafter the confidences of the new rules are checked to decide their interesting-ness. This system is repeated until the antecedent will become empty. On account that the second one sub trouble is quite simple, maximum of the researchers' consciousness on the primary sub-problem.

The primary sub-problem can be similarly divided into two sub-problems: candidate large item-sets generation and common item-sets technology. The item-sets whose assist exceed the assist threshold are referred to as massive or common item-sets and people item-sets that are anticipated or have the desire to be huge or common are referred to as candidate item-sets. The two thresholds on which arm technique is primarily based are referred to as minimal assist and minimal self-belief respectively. Assist is described as the proportion of records that incorporate a ¬ b to the entire quantity of records within the database. Let us count on the support of an item is 0.1%, it means simplest zero.1 percent of the transaction include this item. Self-belief of an affiliation rule is described because of the fraction of the range of transactions that contain a ¬ b to the entire range of information that comprise a. Self-belief is a degree of energy of the affiliation rules, assume the self-assurance of the affiliation rule a¬ b is 80%, it approaches that 80% of the transactions that contain an also incorporate b together.To illustrate this concept, a small example from the supermarket area has been used. The set of objects is i=bread, egg, butter, cheese and a small database (table 2.1) containing the objects (1 represents that object is the gift and 0 represents that item isn't present in a transaction). An instance rule for the grocery store can be bread, egg => butter meaning that if bread and egg are bought, customers also purchase butter.

Table 2.1 Sample Database for Finding Association Rule

T	Bread	Egg	Butter	Cheese
T1	1	1	0	0
T2	1	1	1	0
T3	1	0	1	1
T4	0	1	1	0
T5	1	1	0	0

In the example database, the object set-bread, egg, butter has a guide of 1/5=zero.2 since it happens in 20% of all transactions (one out of five transactions).

The guideline bread, egg=> butter has a self-belief of 0.2/0.Four = zero. Five, which means that that for 50% of the transactions comprise bread and egg (50% of the times a consumer buys bread and egg, butter is offered as well).

2.6.1 Generalized Association Rule Mining Algorithm

Many algorithms for generating association rules are presented over time. Some of the well-known algorithms are Apriori, FP-growth, AIS, Apriori-TID, Apriori Hybrid, Partitioning algorithms, Tertius Apriori Algorithm and many more. Some of the parallel association rule mining algorithms based on Data and Task include CD (Count Distribution), PDM (Parallel Data Mining), HPA (Hash-based Parallel Mining of Association Rules) and PAR (Parallel Association Rules) and many more. In general, a set of items (such as antecedent (LHS) or the consequent (RHS) of a rule) is called an itemset. The length of an itemset is given as the number of items contained in an itemset. Itemsets of some length k are called k-itemsets.

Generally, an association rules mining algorithm contains the following steps:

1. The set of candidate k-itemsets is generated by 1-extensions of the large (k-1)-itemsets generated in the previous iteration.
2. Support for the candidate k-itemsets is generated by a pass over the database.
3. Itemsets that do not have the minimum support are discarded and the remaining itemsets are called large k-itemsets.

This process is repeated until there are no more large itemsets in the database. The most commonly used approach for finding association rules is based on the Apriori algorithm. The efficiency of the level wise generation of frequent itemsets is improved by using the Apriori property which says that all nonempty subsets of a frequent itemset must also be frequent [11].

2.6.2 Issues and Challenges

Lots of studies paintings have been finished within the subject of affiliation rule mining and numerous authors have proposed particular algorithms in this concern. Despite the fact that there exist many troubles and traumatic situations in this vicinity which want to be solved so as to get the entire advantage of this technique. The principle drawbacks of the affiliation rule mining algorithms are:

- Acquiring non-exciting guidelines

- Huge form of decided hints
- Low set of guidelines normal overall performance surrender clients of association rule

Mining equipment stumble upon numerous issues which encompass the algorithms do now not continually go returned the outcomes in less expensive time. It is also found that the set of affiliation policies can unexpectedly turn out to be unwieldy, particularly while we decrease the frequency requirements. Extracting all association regulations from a database requires counting all viable combinations of attributes. Aid and self-belief factors can be used for obtaining exciting regulations which have values for these elements extra than a threshold value.

In most of the strategies, the self-belief is determined as soon as the relevant help for the policies is computed. But, at the same time as the quantity of attributes is large computational time increases exponentially. For a database of m facts and n attributes, assuming binary encoding of attributes in a report, the enumeration of a subset of attributes calls for m*2n computational steps. For the small cost of "n" conventional algorithms are clean and inexperienced but for big values of n, the computational analysis is infeasible. The important thing element that makes association rule mining practical is the min-sup i.e., the minimum guide distinct through the man or woman. It's far used to prune the boring guidelines. But using first-rate a single min-sup approach that is all the objects in the database are of the identical nature. This cannot be the case all the time. For example, in retailing corporation clients often purchase those gadgets which have fewer fees at the same time as the objects which have a better fee won't be presented too frequently. On this form of state of affairs, if the min-sup is about too immoderate, the generated policies will incorporate most effective the ones guidelines containing most effective the ones gadgets that have low rate and make contributions less to the income of the agency.

However, if the min-sup is ready too much less, many meaningless common patterns may be generated for you to overload the choice makers. This sort of state of affairs is known as uncommon item problem [2]. Affiliation rule mining has been very a success in severe fields like commercial, social and human sports activities. However, this technique poses a chance to private-ness. User will effects divulge other's statistics with the aid of the use of this approach. So in advance than releasing the database the touchy information need to be hidden from unauthorized get right of entry to. It has

been discovered that one of the current technical demanding situations on this place is the improvement of strategies that include security and private-ness problems.

The affiliation rule hiding trouble targets at sanitizing the database on this kind of way that via association rule mining one will not have the capacity to show the sensitive records and best the non-touchy information is probably mined [15].

2.6.3 Performance Review

Many algorithms for generating affiliation guidelines had been offered over time. Some of the widely known algorithms are Apriori, FP-boom, AIS, Apriori-tid, Apriori hybrid, partitioning algorithms, FP-growth set of rules, Tertius algorithm and plenty of more. The advantages and downsides of some of the affiliation rule mining algorithms are discussed in table 2.2, the AIS algorithm became the first set of rules to generate all huge itemsets in a transaction database. The algorithm is used to find qualitative guidelines. This technique is limited to the handiest object in the consequent. The ais set of rules makes a couple of passes over the database. The primary trouble of the AIS set of rules is that it generates too many applicants that later emerge as small [1]. Every other disadvantage of this set of rules is that the records systems required for keeping huge candidate itemsets are not distinct.

The Apriori set of rules developed by [1] is the most well-known affiliation rule algorithm. Apriori way “from what comes before” and makes use of breadth-first seek technique. Its implementation is less difficult than different algorithms and consumes less memory. But it has sure risks also. It best explains the presence and absence of an item in transactional databases and requires a massive quantity of database test. Furthermore, the minimal support threshold used is uniform and the variety of candidate item-units produced is big. To overcome some of the bottlenecks of the Apriori algorithm FP-growth set of rules turned into designed which is primarily based on tree shape.

Table 2.2 Performance Review of Some Algorithms

Association Rule Mining Algorithm	Advantages	Disadvantages
AIS	1. An estimation is used in the algorithm to prune those candidate itemsets that have no hope to be large. 2. It is suitable for low cardinality sparse transaction database.	1. It is limited to only one item in the consequent. 2. Requires multiple passes over the database. 3. Data structures required for maintaining large and candidate itemsets is not specified.
Apriori	1. This algorithm has least memory consumption. 2. Easy implementation. 3. It uses Apriori property for pruning therefore, itemsets left for further support checking remain less.	1. It requires many scans of database. 2. It allows only a single minimum support threshold. 3. It is favourable only for small database. 4. It explains only the presence or absence of an item in the database.
FP - growth	1. It is faster than other association rule mining algorithm. 2. It uses compressed representation of original database. 3. Repeated database scan is eliminated.	1. The memory consumption is more. 2. It cannot be used for interactive mining and incremental mining. 3. The resulting FP-Tree is not unique for the same logical database.

2.7 Ethics of Web Mining

By means of searching at web mining from a moral angle, we will discover an area of tension, amongst advantages on the one hand and disadvantages on the opposite. As ethics is the department of philosophy concerned with the character of morals and moral assessment, an ethical angle will enhance questions like what is right or incorrect, what's beneficial or dangerous. Moral research specialise in three styles of problems.

First, there are conditions in which normative ideas are really ignored. Then there are moral issues regarding new problems (varieties of issues that don't wholesome present times) where it's far a query of approaches traditional thoughts may be implemented. The 0.33 kind of ethical conditions deal with the elegance of normative

conflicts. A normative war appears every time there are each appropriate and horrific facts to remember. The problem of net mining is a normative warfare wherein brilliant refers back to the advantages of net mining and horrific refers to its viable risky implication; in one of a kind phrase the ethical values which can be threatened. Values are the middle belief or dream that manual or encourage attitudes and movements and decide how humans behave in sure conditions. As ethics is a reflection on morality, ethical values can be described as that which topics verify as ethical in human conduct. Therefore, moral values have a normative function and are the motive of ethical human behaviour. A fee can be visible as a international goal. This sort of aim needs to be pushed with the aid of a manner, presented by using extra particular norms. As an example, the price of privateness is driven with the aid of norms like respecting someone's personal lifestyles and now not misusing someone's non-public information. Norms might be meaningless without values. Expertise observed after mining the web, must pose a hazard to humans, while as an example non-public statistics is misused.

However, it is this same data component that might suggest lots of different benefits, as it's far of immoderate price to all forms of packages regarding planning and manipulate. Kosala, Blockeel, and Neven have already defined a few particular blessings of net mining, like enhancing the intelligence of engines like Google. Net mining can also make a contribution to marketing intelligence by way of analyzing the net person's online behaviour and turning this statistics into advertising understanding. It ought to be stated that moral problems should arise from mining net facts that do not incorporate non-public facts at all, along with records on one-of-a-kind types of animals, or technical statistics on cars.

However, this phase is confined to net mining that does in some manners involve private data. We can handiest observe the viable moral damage that can be finished to people, because of this that harm completed to businesses, animals, or extraordinary topics of any kind are not part of the scope of this remark. Considering, most internet mining packages are presently discovered in the private sector this will be our vital consciousness. So web mining regarding non-public information might be regarded from a moral angle in an enterprise context. Be conscious that this system isn't intended to be of a moralistic nature. Inside the moral angle of this normative battle, it is miles certainly recognized that web mining is away with a massive number of correct traits and functionality. However, analyzing the viable threats may additionally create

a positive consciousness, major to a well-considered software and further development of this technique.

2.8 Web Mining Applications

Web mining extends evaluation masses similarly by the way of combining the different organisation's information with internet site traffic's statistics. This allows accounting, customer profile, stock, and demographic information to be correlated with internet surfing, which answers complicated questions together with:

- The individuals who hit our web page, how many bought something?
- Which advertising and marketing campaigns resulted in the maximum purchases, now not sincerely hits?
- Do the internet site visitors match a fine profile? Can the visitors be capable of use this for segmenting my marketplace?
- Realistic applications of net mining era are considerable and are in no way the restriction to this era. Web mining gadget may be extended and programmed to reply almost any query.
- Internet mining can offer corporations managerial perception into vacationer profiles, which help pinnacle control take strategic actions for that reason.

Moreover, the agency can acquire some subjective measurements via net mining on the effectiveness in their advertising and advertising and marketing campaign or advertising and marketing research, so you can help the industrial agency to beautify and align their advertising techniques well timed. As an example, the organization may have a list of goals as the subsequent:

1. Boom average internet web page views per session;
2. Growth not unusual earnings consistent with checkout;
3. Lower merchandise again;
4. Boom quantity of referred clients;
5. Increase brand attention;
6. Growth retention rate (which includes the range of visitors which have once more within 30 days);
7. Reduce clicks-to-close (common web page perspectives to carry out a purchase or obtain preferred data);
8. Boom conversion price (checkouts consistent with go to).

The enterprise can discover the strength and weakness of its internet advertising and marketing campaign through internet mining, and then make strategic changes, collect the feedback from net mining over again to look the development. This gadget is an on-going non-stop manner.

2.9 Advantages of Web Mining

Web mining is appealing to organizations due to numerous benefits. Within the maximum stylish sense, it could contribute to the growth in earnings, be it through, in reality, selling extra services or products, or with the resource of minimizing the fees. So that it will try this, marketing intelligence is needed. This intelligence can be recognition on advertising and marketing strategies and aggressive analysis, or at the reference to the customers. The one-of-a-type kinds of internet statistics which might be in some way associated with clients will then be labelled and clustered to build unique client profiles. This now not best helps corporations to maintain modern customers by way of being able to offering extra custom designed offerings but, however, it also contributes in the look for capability customers.

Web mining is an interest that has boosted groups and corporations a further deal. Statistics lets in corporations and corporation human beings to deliver efficient facts referring to the characteristic of the corporation or agency and characteristic capability. A number of the data may be collected from the collective information of lifetime client value merchandise, cross strategies in advertising and promotional campaigns.

The information accumulated allows the groups with the ability to produce effects in a manner this is greater effective and excellent to their corporations that permits you to lead to a boom in earnings. Utilization records can be used growing advertising and marketing competencies with the intention to sell the opposition help in selling the organization's offerings or a product on an excessive degree.

Usage mining is valuable, however no longer most effective to corporation using net or on-line advertising and marketing. But additionally to e-corporations who have business enterprise based totally absolutely on-website online traffic being provided with the useful resource of Search Engine Optimization (SEO). The patron of this form of mining permits the gathering essential data from clients trafficking to the internet site. This could allow great prolonged to finish the evaluation of a go along

with the glide of a organisation's product. E-commercial enterprise is dependents of this kind of information to be in a characteristic to direct the enterprise to powerful internet servers to sell their product and services.

Web mining additionally permits web-based totally establishments to offer better get proper of entry to exclusive offerings or commercials. At the same time as a long time order makes a industrial for services, it does offer, however, are supplied through way of exquisite agencies. Usage of mining statistics will deliver most effective to the only path to the one's portals. There are three uses for mining on this style.

The number one is processing used to finish discovery. That is moreover the hardest use due to the truth most effective bits of information like private information, IP addresses and placement websites are to be had. With this small quantity of information that is available, it receives harder to trash the user thru a domain, this program can be capable of observing users at some point of pages of websites. Content material cloth use is content fabric processing, this encompasses some conversations or net records like textual content, pixels, scripts and special useful forms. With the clustering and categorization of the internet page allows in web page facts primarily based on titles, specific contents, and images to be had.

The one third of use is hooked up processing. That is a mixture of an assessment of the shape of each web page contained on an internet site. This structured system may be difficult if resulting in a brand new shape this is to be accomplished for every web page. Assessment of this usage data will deliver the companies the records this is required to give an effective presence to their cutters.

The records that are accumulated may additionally integrate registration of the individual, get admission to logs, and records that might result in higher net website shape. Consequently, presenting the employer to the maximum valuable in online marketing and advertising and marketing, those can provide a few advantages of outside advertising of an enterprise and its products, services and common control.

2.10 Analysis Summary on Web Mining

There are various types of web mining such as Web Content Mining, Web Structure Mining and Web Usage Mining. Each of such mining is analyzed on the following point of views as shown in table 2.3.

Table 2.3 Comparison of Web Mining Types

	Web Content Mining		**Web Structure Mining**	**Web Usage Mining**
	IR View	**DB View**		
View of Data	• Unstructured • Structured	• Semi-structured • Web site as DB	• Link Structure	• Interactivity
Main Data	• Text Documents • Hypertext Documents	• Hypertext Documents	• Link Structure	• Server Logs • Browser Logs
Representation	• Bag of words, n-gram terms • Phrases, Concepts or Ontology • Relational	• Edge Labelled Graph • Relational	• Graph	• Relational Table • Graph
Method	• Machine Learning • Statistical (including NLP)	• Proprietary Algorithms • Association Rules	• Proprietary Algorithms	• Machine Learning • Statistical • Association Rules
Application Categories	• Categorization • Clustering • Finding Exact Rules • Finding Patterns in Text	• Finding Frequent Sub-Structures • Website Schema Discovery	• Categorization • Clustering	• Site Construction • Adaption and Management

CHAPTER 3
MINING MULTILEVEL ASSOCIATION RULES FROM PRIMITIVE FREQUENT ITEMSETS

Association rule mining observes the concern on association or relationship between a large pair of data items. Because the large quantity of data is stored in the system of real application, relationship association are more concern. Association rules mining can assist decision making in business, and user desire marketing strategies development. The problem of mining in the analysis of market basket can be presented as:

In the transactions of database D,

- one transaction presents an item-set;
- all rules have searched the correlation of an itemset with another itemset.

Agrawal and Srikant [1] present an algorithm for rule mining quick association. Apriori observes every rule of X∪Y which meet given support and output parameters of confidence, where X, Y is sub-set of I, $X \cap Y = \emptyset$, and $I = \{i_1, i_2, \ldots, i_m\}$ be a unique itemset in the D database. The support determined by P (X | Y) is the transactions' percent in D which consists of X ∪ Y. The denoted confidence followed by P (Y|X) is the transactions' percent in D, Y is in X. For instance, from the dataset of dairy product of "ORLEANS BAKERY 'BREADS ON OAK' DITCHES MEAT, DAIRY, AND EGGS AND GOES VEGAN", [2% of support, 60% of confident] show "60 percent of the buyers buy sunshine-bread and daisy milk, and 2 percent of transactions under research prove that daisy-milk & sunshine-bread were bought together." But, many domains, it is not easy to search exact rules of association between items of data at lower layers of abstraction due to the data in multi-scope.

Multi-level association rules mining algorithms proposed for some case are:

- Specific support requirement for dynamic hierarchy;
- Efficiency of algorithm can't satisfy requirements of the real domain;
- Different levels association may be lost.

To meet the described cases, this introduced another way as follows: firstly, atomic level frequent itemsets searching from prime data items by FP-growth called

atomic frequent item sets. Then, multi-level frequent itemsets mining divided into two types:

- **Same-level** frequent item sets for all set of items at the same level concept, and
- **Cross-level** frequent item sets for all set of items at different levels of concept.

The multilevel association rules are searched on analysis of atomic level frequent itemsets, on behalf of normal ways from the database again. So, it is a good manner to reduce the complexity and I/O cost during extracting single-level and lateral-level rules of the association. But some restrictions have to do for some in-accurate which is in the support and confidence computation, and some rules of the frequent association may lose, some facts can get to reduce this type of error. The meet of support and confidence in-accuracy will be discussed in chapter 4. Moreover, this way assists multi-hierarchy that can be constructed based on the user knowledge input.

3.1 Method for Mining Multilevel Association Rules

A technique for rules of multilevel association searching uses FP-growth to search frequent itemsets from the prime data at the atomic layer. The encoding method which is used to encode the items by the given concept of hierarchy uses an encoded string to show hierarchy position. This way is suitable for binding items based on unique pre-fixes encoding, and minimum bits are required. Encoding generates frequent single items, and no additional is required.

To discuss encoding way, Figure 3.1 presents a concept hierarchy from a shop database. According to above mention, “Wonder whole wheat bread” item can be defined as “243”, where first digit “2” for “bread” at level-2, the digit “4” represents “Whole wheat (bread)” at level-1, and the digit “3” for “Wonder” at level-0. The items “milk” and “skim milk” can transform as “1**” and “11*”. So this can bind “121” and “122” into “12*”. This way of binding helps for frequent itemsets analysis.

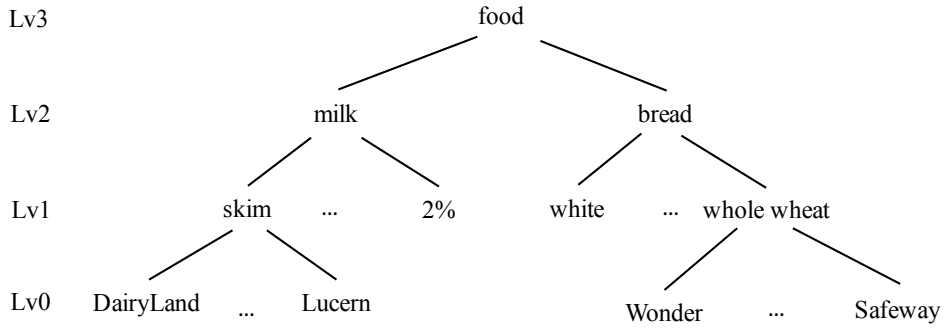


Figure 3.1 Concept Hierarchy Examples

The remaining major steps are as follow:

1. Group and merge frequent high-level item-sets mining at the atomic level;
2. Manipulate each candidate support, and select supported output standard;
3. Generate multilevel frequent item-sets association rule.

Based on the steps (1) and (2), the described steps are iteratively processed until further frequent item sets can't be found. Before multilevel rules generating, FP'-tree idea and its relation mining and frequent itemsets calculation are pre-processed.

FP'-tree is a conceptual transformation of the lower level to higher level FP-tree. The FP'-tree idea [3] applies a different way for construction.

Supported atomic level FP-tree and threshold of level l for constructing FP(l)-tree is summarized as follows.

1. Transform all nodes in the atomic level of FP-tree into level l.
2. For repeated nodes, the top one will not be changed and the others will be discarded.
3. Bind paths by executed identical nodes of support counts.
4. Discard the same item in the header; calculate the count of support for FP (l)-tree relative nodes.
5. Header table items are sort by frequency descending order.
6. Configure the same order items in the header table, and bind the paths as in step 3.
7. Frequency unsatisfied item from level l, not only delete from the header table, but also delete the relative nodes.
8. Justify links of the node.

Step 1, FP-tree transforms level l items by encoding with "*". Step 2 and 3 executes on redundant nodes. For these nodes, if in a single path, delete them without including the top. If in the different link, merge the paths related to it. Phase 4 and 5 are to modify the header table by merge nodes in the tree. The goal is to filter the FP(l)-tree in the form of a normal FP tree.

3.2 Constructing FP-tree at the Atomic Level (Level 0)

Figure 3.2 shows an atomic level (level 0), and assume the given value of level 1 is 3. FP (1)-tree generation is shown below.

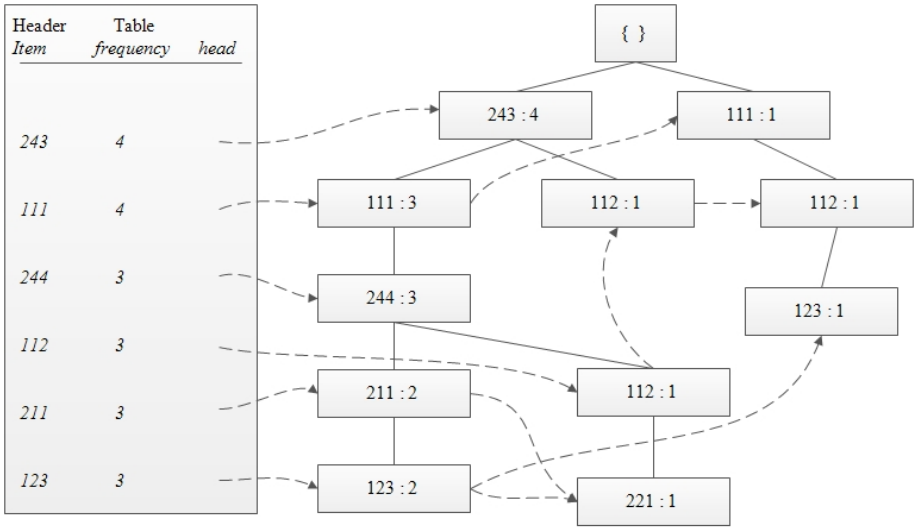


Figure 3.2 FP (0)-tree (Atomic Level)

First, items are transformed into level 1 concept, for instance, both “243” and “244” in Figure 3.2 be transformed to “24*”; and remove the duplicate nodes, that is “24*:4” and “24*:3” are redundant in the single path of tree and the top is “24*:4”, the system maintains “24*:4” and remove “24*:3”. Then, the changes are shown in figure 3.3.

3.3 Uncompleted FP (1)-tree

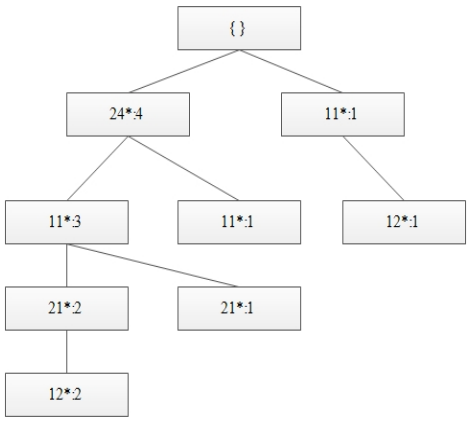


Figure 3.3 Uncompleted FP (1)-tree (Step 1-2)

Clearly, some links need to be combined as shown in Figure 3.3. Such as the link flow from “24*:4”, and the two links flow from “11*:3”. They are combined with

the relative nodes. The unique nodes in other links are combined by summing the related items support count. For instance, the items “11*:3” and “11*:1” can be combined as “11*:4”. So after processing of phase 3, the modified FP-tree is shown in Figure 3.4.

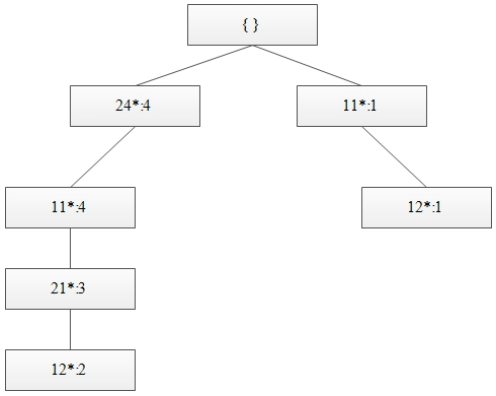


Figure 3.4 Uncompleted FP (1)-tree (step 3)

The next phase is to continue the phases 4 and 5. FP-tree nodes are arranged as the order of the item in new header table in Figure 3.5. After item arrangement, regenerate the path as “11*:4”, “11*:1” into “11*:5”. All the satisfied items of supported threshold do not need to move any items in FP-tree. The final FP (1)-tree is shown in Figure 3.5.

3.4 Completed FP (1)-tree

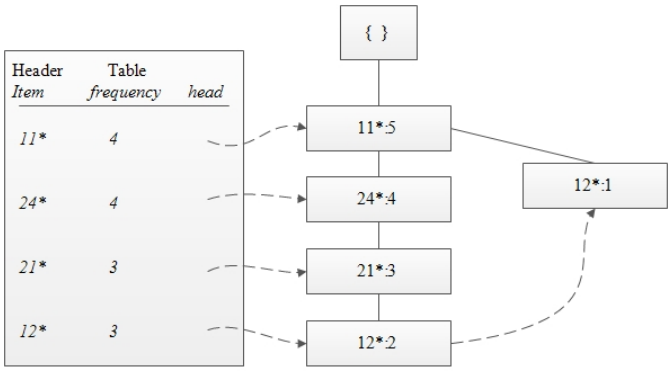


Figure 3.5 Completed FP (1)-tree

In order to skip retracing the database, FP (l)-tree frequent itemset is calculated as the support of a candidate. Assume the frequent itemset “2**, 11*” is transformed

based on Figure 3.1. It determines the threshold count of level-2 is 3. The item set "2**, 11*" can be manipulated by the following according to FP (2)-tree.

1. "11*" of header and child links of "11*" to search "24*" and "21*" in the links "11*" flow in FP (1)-tree 2** is not existing, which "24*" and "21*" are related.
2. Search "24*: 4", "21*: 3", and count 4 for maximal support as 2**. So, the support count for "2**, 11*" is 4. This meets the threshold support; therefore, it is determined as a frequent item set.

To generate frequent primitive item sets to frequent multilevel itemsets, frequent primitive item-sets firstly generate frequent high-level itemsets. Based on the encoding result, grouping and merging are made on the frequent low-level itemsets. The merging is made by matching degree.

Definition 1: For any two given items of length n (n ≥ 2): A = $a_1 a_2 \dots a_n$, and B =$b_1 b_2 \dots b_n$, item degree matching is made as follows:

$$\mathrm{MI(A,B)} = \begin{cases} \sum_1^n \mathrm{m}(\mathrm{a_i},\ \mathrm{b_i}),\ \mathrm{a_1 = b_1} \\ 0,\quad \text{otherwise} \end{cases} \quad \text{..... (1)}$$

Where m (a, b) (a belong to A, b belong to B) is two positions matching in items, and it is manipulated by

$$\mathrm{m(a,b)} = \begin{cases} 0, & \mathrm{a \neq b} \\ 1, & \mathrm{a = b} \end{cases} \quad \text{..... (2)}$$

Definition 2: For sets of total n items (n ≥ 2): I1: $A_1, A_2, \dots, A_i, A_{i+1}, \dots, A_n$, I2: $B_1, B_2, \dots, B_i, B_{i+1}, \dots, B_n$, and the items are arranged by their position in given idea of hierarchy, from left to right, and from down to top. The itemset degree matching is made as

$$\mathrm{MR(I_1, I_2)} = \sum_1^n \mathrm{MI(A_i, B_i)} \quad \text{..... (3)}$$

Same length item sets can combine to extract frequent itemsets by matching item set degree which satisfies the threshold. The main concept of merging is to make the highest similarity between items. After the selection extraction is finished, manipulate their respective support values. Recursively executed generation and verification process can gain certified frequent item which has the higher level frequency. FP(l)-tree atomic level frequent itemsets can give threshold value at every layer, and item set threshold matching of the algorithm for extracting frequent multilevel item sets —MFI steps:

1. Every frequent itemsets are organized based on the amount of the items. The total numbers of items for each item set in a group should be uniform.
2. Calculate the degree of matching for each pair of item sets, combine the related item sets to extract frequent item set if it satisfies the provided threshold by the following:
   (a) Search matched item
   (b) For matched item pair, extract a new item by holding match digits and replace the remaining with "*", and if it does not exist, insert it to the new item set.
3. Remove the frequent candidates.
4. Get the verified frequent itemsets from candidates by algorithm CSI.
5. To get new frequent itemsets, do1~4 again and again until new candidate can't be generated.

3.5 Association Rule

Association rule mining is to find interesting associations or correlation relationships among a large set of data It is an implication from $X \rightarrow Y$, where $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_n\}$ are sets of items with $X \cap Y = \varnothing$. The rule $X \rightarrow Y$ has support x if s% of all item-sets contain $X \cup Y$.

$$\text{Support}(X \rightarrow Y) = \frac{\text{Number of tuples containing both X and Y}}{\text{Total number of tuples}} \quad \dots\dots (4)$$

The rule $X \rightarrow Y$ has confidence c if c% of item-sets that contain X also contain Y.

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Number of tuples containing both X and Y}}{\text{Number of tuples containing X}} \quad \dots\dots (5)$$

Mining association rule is to generate all association rules that have support and confidence greater than the user-specified minimum support min_sup and minimum confidence min_conf respectively.

It has two processes:

- Find the set F of all item-sets with support above minimum support min_sup. These item-sets are called frequent item-sets.
- Use the frequent item-sets to generate the desired rules by eliminating those that do not achieve min_confidence.

CHAPTER 4
SYSTEM DESIGN AND IMPLEMENTATION

4.1 The System Overview

This thesis will present web utilization mining by generalized sodality rules. Generalized Sodality Rule – Rules predicated on Taxonomy / Hierarchy of product or page in this web utilization mining Taxonomy and hierarchies are defined predicated on attributes of weblog records such as browsers, referrer, status, URL and so on. Web server logs are pre-processed including following steps: Cleaning, Utilizer Identification, and Session Identification. Then those pre-processed log records are stored in the transaction database. Then multi-level sodality rule is applied predicated on the taxonomy of weblogs. Frequent patterns are engendered as an output of the system as shown in figure 4.1.

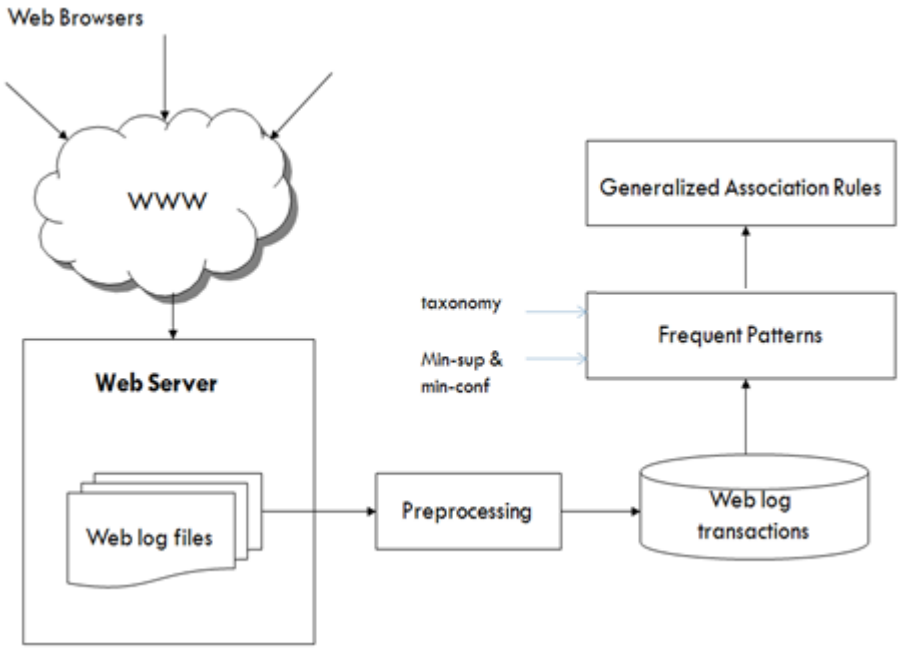


Figure 4.1 System Overview

4.2 Implementation of the System

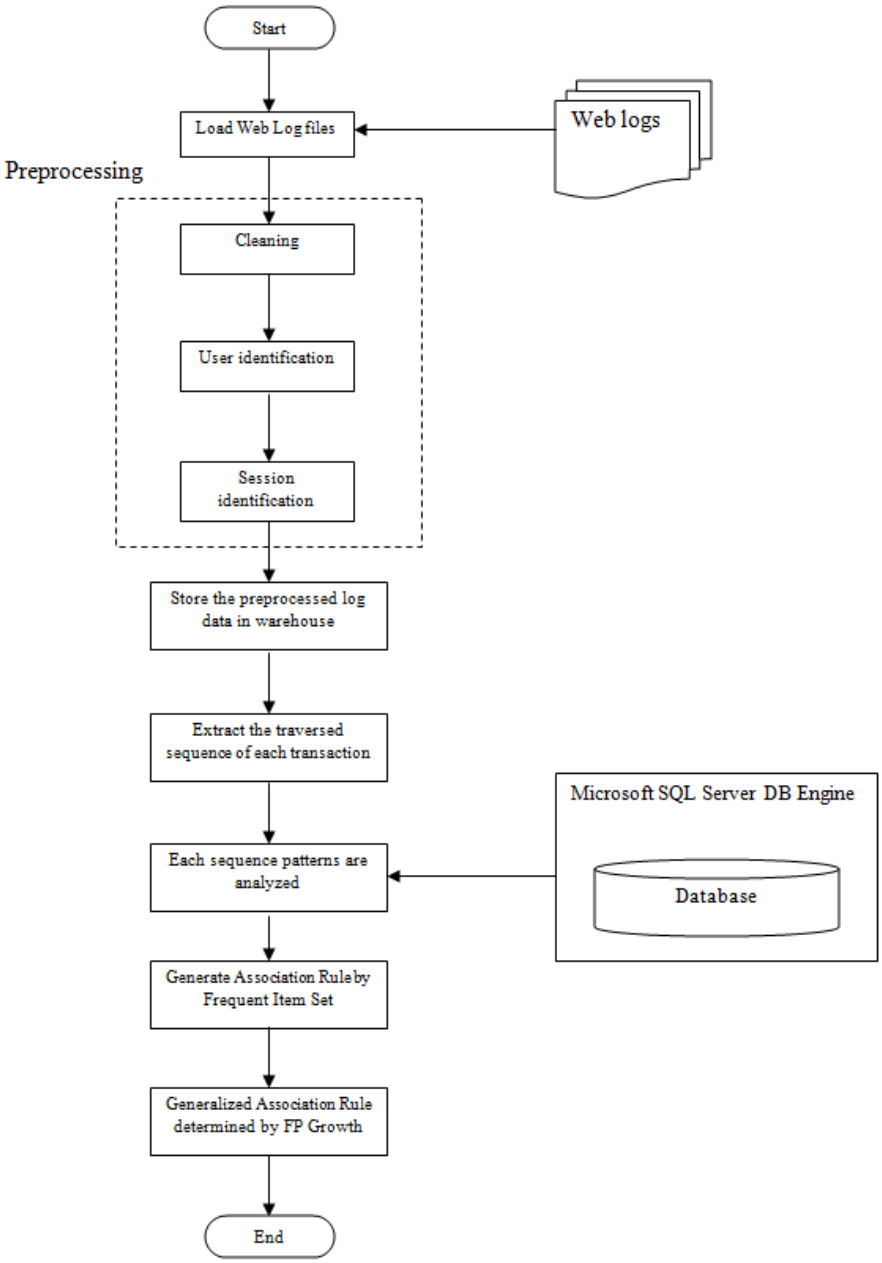


Figure 4.2 The Flow Diagram of the System

At the commencement of the system, weblog must be loaded to be analysed as shown in figure 4.2. In the cleaning phase, image URL is abstracted. Image URL is detected by utilizing its file extension whether it is .gif, .jpg, .bmp, .png, etc., for example, If (URL. ends with (".gif"), abstract the sentence. Weblog records (with date, access time, client IP, URL, method, status) are stored in the recollection. Then, the system extract the traversed sequence of each transaction and each sequence patterns

are analyzed which is stored in the database. After that, the system generates the association rule based on the frequent item set. Finally, the system extracts the generalized association rule by FP growth.

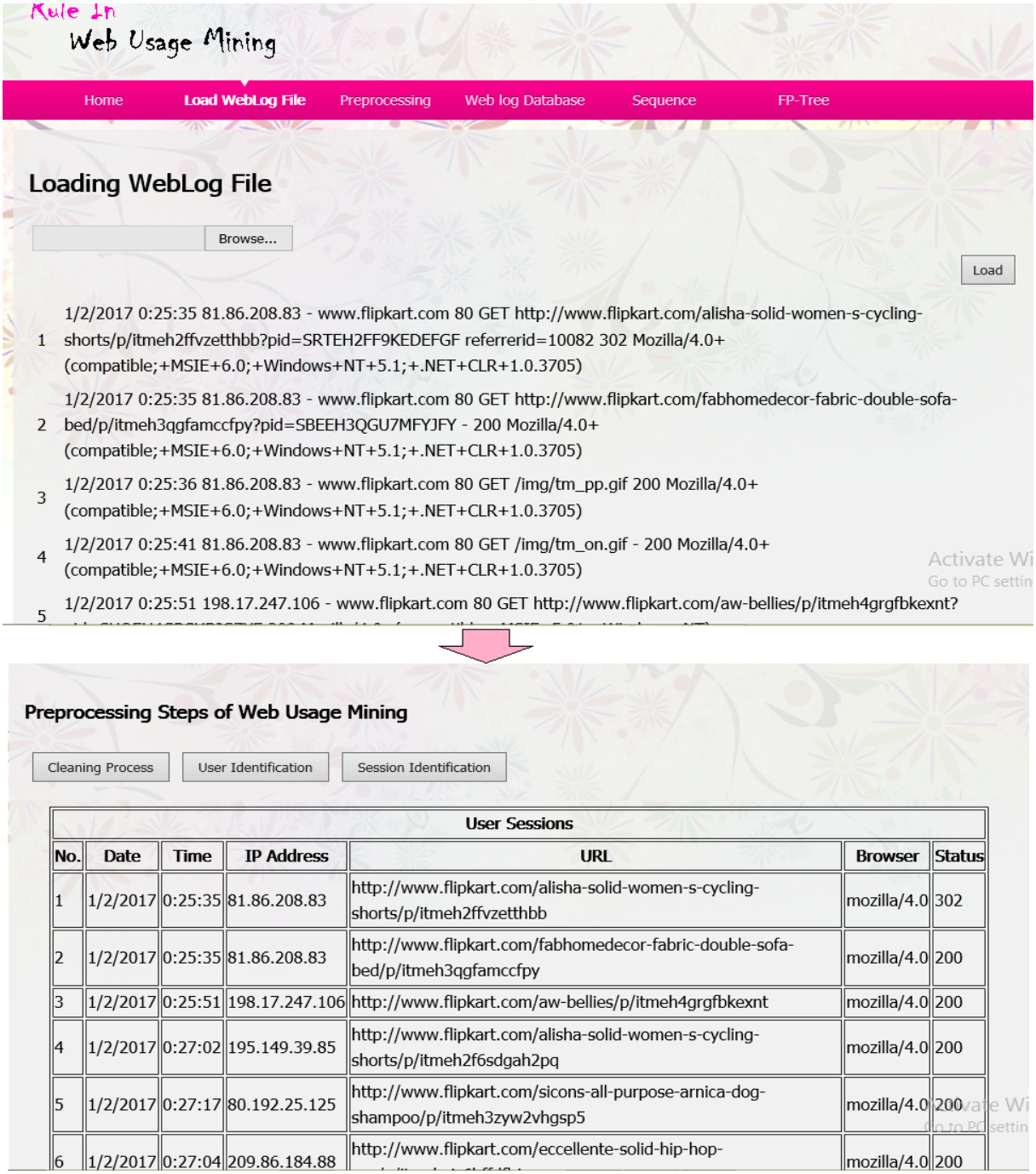
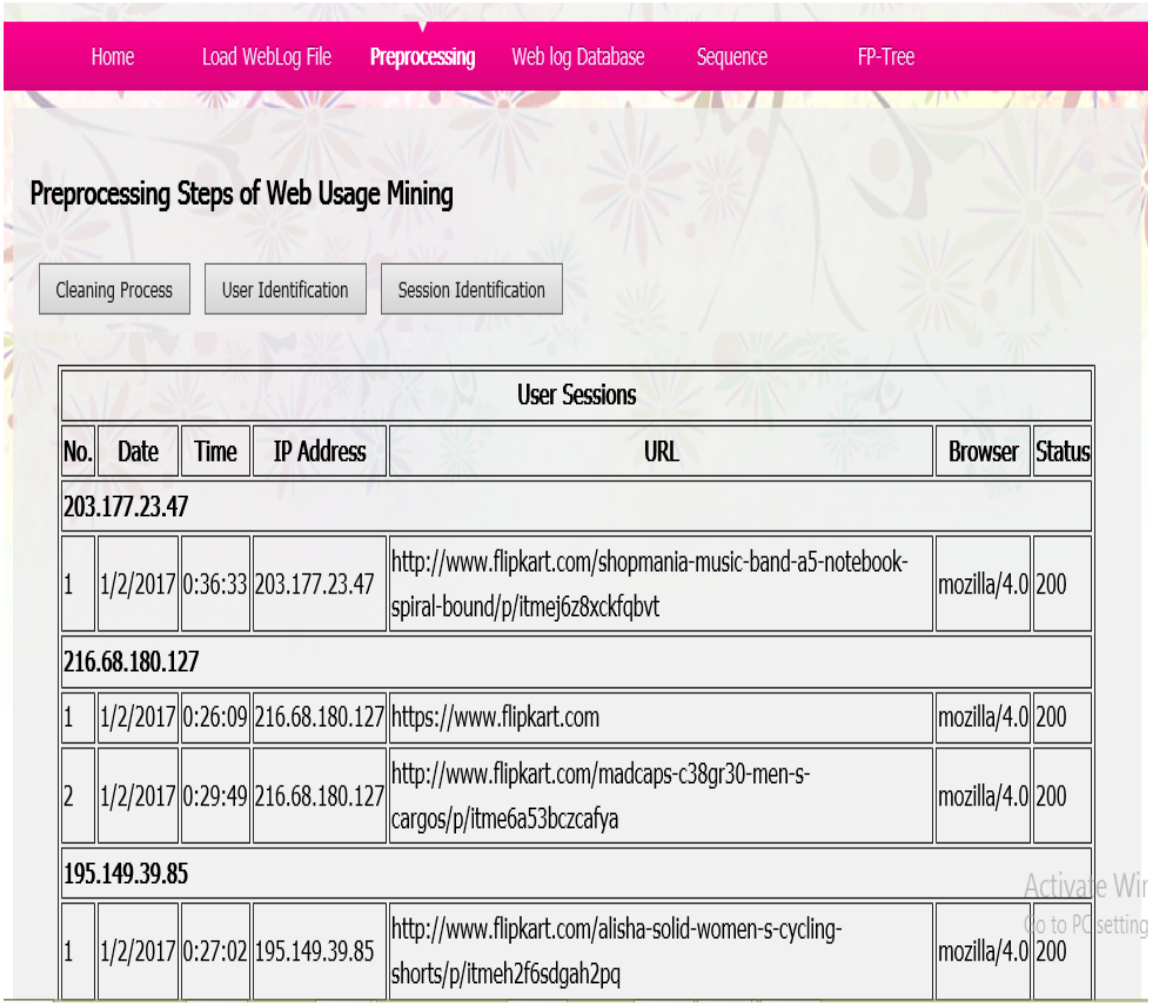


Figure 4.3 Cleaning Phase of Pre-processing

In the user identification phase, records with same client IP are considered to be one utilizer and those records are grouped together as shown in figure 4.3. For

example, log records with client IP address 192.168.0.1 are surmised to be accessed by a single utilizer and those records are grouped together.



User Sessions						
No.	Date	Time	IP Address	URL	Browser	Status
203.177.23.47						
1	1/2/2017	0:36:33	203.177.23.47	http://www.flipkart.com/shopmania-music-band-a5-notebook-spiral-bound/p/itmej6z8xckfqbvt	mozilla/4.0	200
216.68.180.127						
1	1/2/2017	0:26:09	216.68.180.127	https://www.flipkart.com	mozilla/4.0	200
2	1/2/2017	0:29:49	216.68.180.127	http://www.flipkart.com/madcaps-c38gr30-men-s-cargos/p/itme6a53bczcafya	mozilla/4.0	200
195.149.39.85						
1	1/2/2017	0:27:02	195.149.39.85	http://www.flipkart.com/alisha-solid-women-s-cycling-shorts/p/itmeh2f6sdgah2pq	mozilla/4.0	200

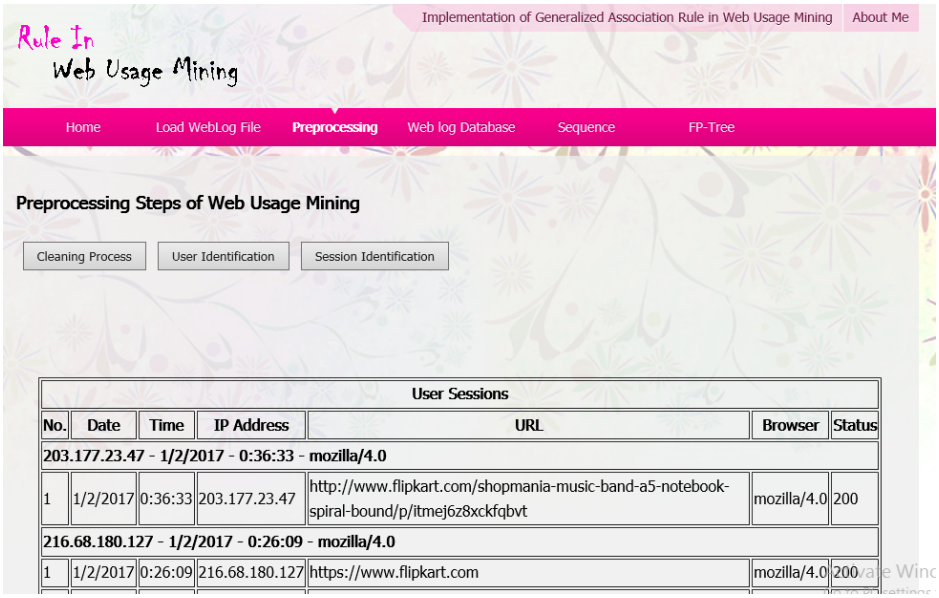
Activate Wir
Go to PC setting

Figure 4.4 User Identification Phase of Pre-processing

The main conception of session identification is that access records with different times may be different sessions. Hence, in the session identification phase, records with the same time group from the utilizer group are grouped together again to get the sessions. For example, client's IP adress 192.168.0.1 with different access time records are summarized to be different sessions as shown in figure 4.4.

4.3 Pre-processed Web Log Sessions

After the pre-processing steps, the weblog records are extracted from web log file as follows. Example web log records are shown in Figure 4.5.



User Sessions						
No.	Date	Time	IP Address	URL	Browser	Status
203.177.23.47 - 1/2/2017 - 0:36:33 - mozilla/4.0						
1	1/2/2017	0:36:33	203.177.23.47	http://www.flipkart.com/shopmania-music-band-a5-notebook-spiral-bound/p/itmej6z8xckfqbvt	mozilla/4.0	200
216.68.180.127 - 1/2/2017 - 0:26:09 - mozilla/4.0						
1	1/2/2017	0:26:09	216.68.180.127	https://www.flipkart.com	mozilla/4.0	200

Figure 4.5 Pre-processed Web Log (Phase of Session Identification)

4.4 Generalized Association Rule (GAR)

Sodality rules engendered from mining data at multiple levels of abstraction are called multiple-level or multilevel sodality rules. For each level, any algorithm for discovering frequent itemsets may be utilized. In this system, FP-Magnification algorithm is utilized. Generalized sodality rule is one of the commonly used web utilization mining technique. Concept hierarchy is utilized to illustrate the relationship between options provided by technician utilizer (site admin or the one who kens their domain very well). Concept hierarchy shows the set of relationship between different items, generalized sodality rules sanction rules at different calibres. Generalized sodality rules were applied to mine the utilizable patterns counting. From the server logs, the hierarchy of the websites is resolute. Comparing with the standard sodality rules, generalized sodality rules sanction rules at different calibres.

4.5 Generalized Association Rule in Web Usage Mining

As more organizations view the Web as an integral part of their operations and external communications, interest in the quantification and evaluation of Web site

utilization is incrementing. Server logs can be habituated to glean a certain amount of quantitative utilization information. Compiled and interpreted opportunely, log information provides a baseline of statistics that designate utilization levels and support and/or magnification comparisons among components of a site or over time. Such analysis additionally provides some technical information regarding server load, unwanted activity, or unsuccessful requests, as well as availing in marketing and site development and management activities.

4.6 FP-Growth in Generalized Association Rule

Generalized Sodality Rule by FP-Magnification has the following processes:

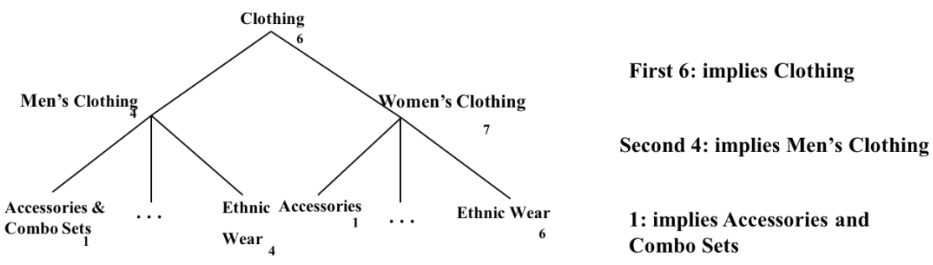
1. Encoding Transaction Table – The first step of GAS is the encoding process which assigns the number to each item in the taxonomy. Pages are encoded with 1, 2, 3, ..., and so on according to the level they stand in the pre-processing steps of web utilization mining.
2. Building Primitive Level FP-Tree.
3. Generate candidate high-level frequent item sets through merging and grouping the frequent itemsets at the primitive level.
4. Calculate the fortification of each candidate, and cull only ones satiating the minimum support.

4.7 FP-Growth in GAR – Step 1 (Encoding)

Table 4.1 Tables for Encoding

Encode the database with layer information

GID	Category	Content	Brand
641	Clothing	Men's Clothing	Accessories & Combo Sets
671	Clothing	Women's Clothing	Accessories



Generalized database stores hierarchy information encoded transaction table in lieu of the pristine transaction table. Ergo, each item is encoded as a sequence of digits in the transaction table as shown in figure 4.6. In GAR, pages are at the Primitive level in the hierarchy and level of taxonomy is tenacious by utilizer input. In generalized sodality rules, FP-Magnification is utilized to mine from primitive items to get frequent item-sets at the primitive level. The encoding scheme is utilized to encode the items according to the concept hierarchy, which uses encoded string to represent a position in a hierarchy. In this system, encoded string is utilized in building FP-tree and the description is only needed for the final exhibit. The encoding process is done according to Encode Tables which is shown in Table 4.1.

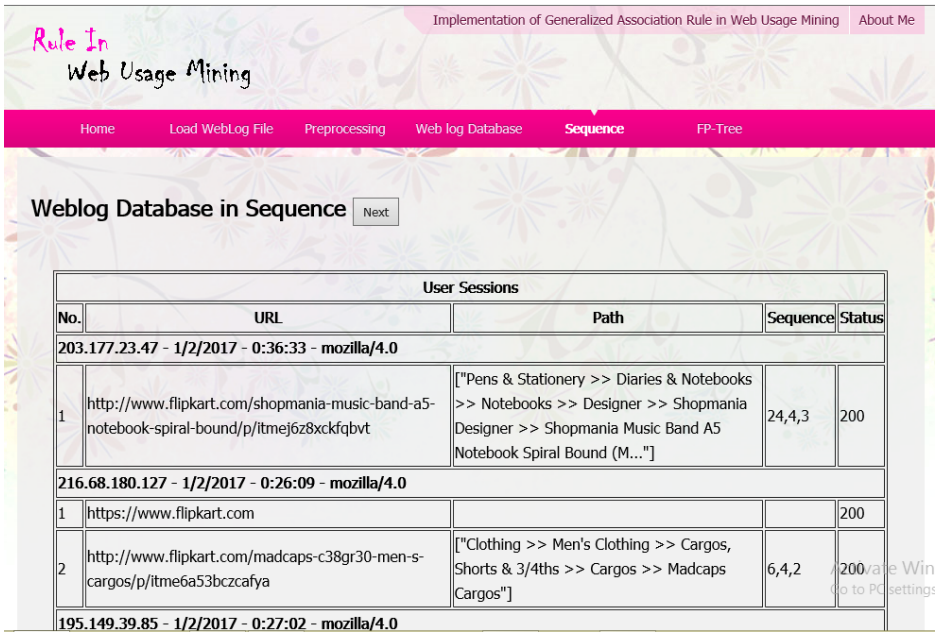


Figure 4.6 Encoded Sequences in Weblog Database

4.8 FP-Growth in GAR – Step 2 (Building FP-Tree for Primitive Level)

Mundane FP-Magnification: In building header table and tree, concepts are not considered; only how many times that page is navigated (count of visit); regardless of browser, method, status and so on.

FP-Magnification with Concept (GAS): In case of taxonomy, web pages including taxonomies are counted and preserved in header table and FP tree.

First, transform the form of items to concept level 1and then abstract the reiterated nodes in the same path of FP-tree. Conspicuously, some inclusive paths need

to be merged such as the two paths derived from node “24*:4”, and the two paths derived from node “11*:3”. Users merge them by merging the corresponding nodes. The identical nodes in different paths are merged through summing the fortification count of the corresponding items. For example, the nodes “11*:3” and “11*:1” can be merged as “11*:4”. After step 3, the transmuted FP-tree is achieved.

The next step is to update the header table as described in steps 4 and 5. And then the nodes in the FP-tree are sorted as the item order in incipient header table. After sorting, the inclusive paths are engendered again, such as nodes “11*:4” and “11*:1” can be merged to “11*:5”. All items slake the fortification threshold, so there is no desideratum to abstract any items in header table or nodes in FP-tree. The final FP (1)-tree is shown in Figure 4.7.

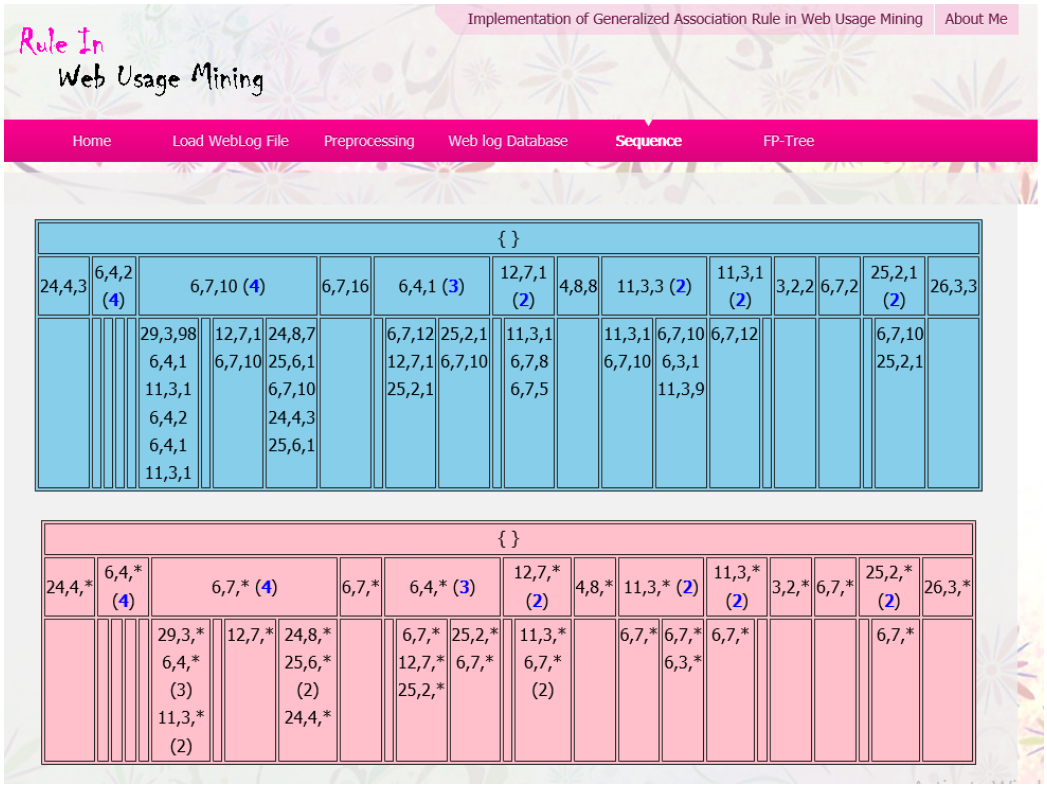


Figure 4.7 Complete FP Tree

4.9 FP-Growth in GAS – Step 3 Detail Explanation (Generating High-Level Frequent Item-sets)

Step 3 is processed by the following steps:

1. Transform all nodes in the atomic level of FP-tree into level l.

2. For repeated nodes, the top one will not be changed and the others will be discarded.
3. Bind paths by executed identical nodes of support counts.
4. Discard the same item in the header, calculate the count of support for FP(l)-tree relative nodes.
5. Header table items are sorted by frequency descending order.
6. Configure the same order items in the header table, and bind the paths as in step 3.
7. Frequency unsatisfied item from level l, not only delete from the header table, but also delete the relative nodes.
8. Justify links of the node.

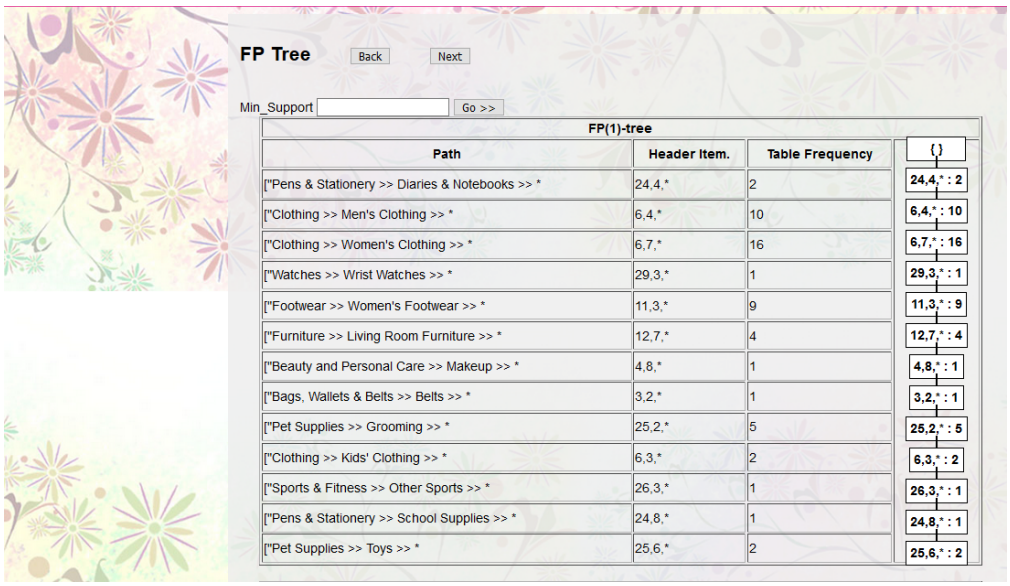


Figure 4.8 System Generated Association Rules

Min Sub section is used to mine the user desire supported frequency of the system as shown in figure 4.8. In figure 4.9, user submitted Min Sub value is 10. So, the system generates the mining results which are greater than or equal value of user submitted value of frequency.

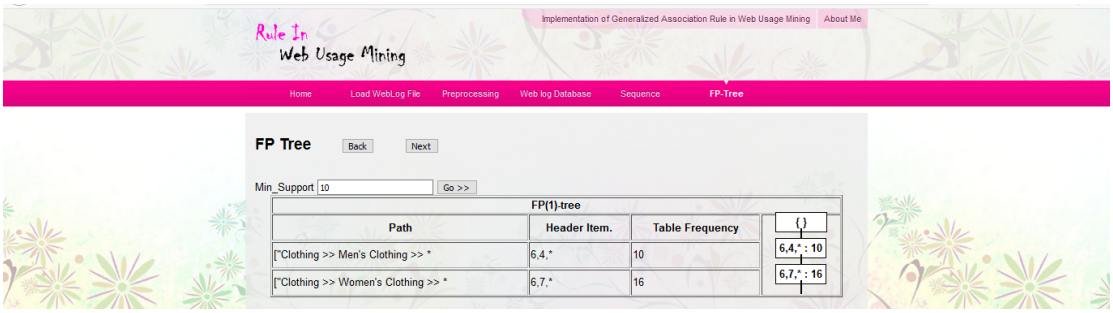


Figure 4.9 System Generated Association Rules with Min Support

4.10 Detailed Explanation of Generalized Association Rule Mining (Emphasized on the Path)

Based on the experiment of figure 4.10, the most frequent item-sets of the imported web log is Clothing which has frequency count 28. The count 28 is composed of three generalized association rules:

1. "Clothing>> Men's Clothing",
2. "Clothing>> Women's Clothing", and
3. "Clothing>> Kids' Clothing".

For more detail comprehension, generalized association rule 1(frequency count 10) is generalized from the association rules:

(a) ["Clothing >> Men's Clothing >> Cargos, Shorts & 3/4ths >> Cargos >> Blimey Cargos"] which has frequency count 5, and

(b) ["Clothing >> Men's Clothing >> Accessories & Combo Sets >> Bandanas >> At Ur Door Bandanas >> At Ur Door Men's Printed Bandana (Pack of 2)"] which has frequency count 5.

By generalizing the association rule with the depth of 2, the two association rules (rule "a" and "b") are generalized as a single rule ➔"Clothing>> Men's Clothing". So, the generalized association rule can support for the quick examination of the system's popular menu items. For the development of the existing business, the generalized association can assist which category must be keep as higher priority and which category must be discarded for future business plan.

4.11 Generalized Association Rule with Simplest Form (Emphasized on the Header Item)

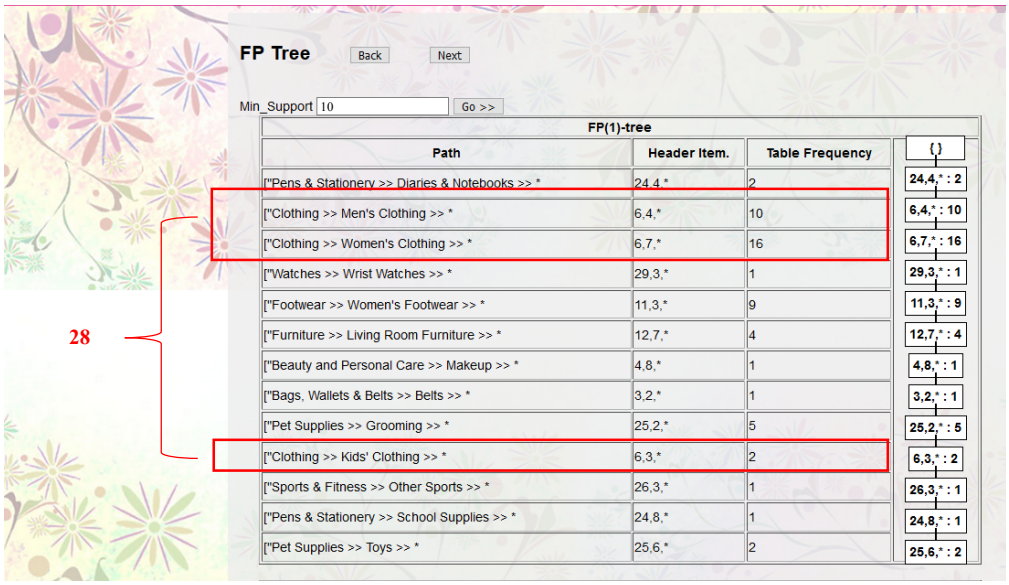


Figure 4.10 System Generated Association Rules (Single Level)

In figure 4.10, generalized association rules for Clothing tab "tab number 6" has been generated for three association rules: such as 6, 4, *: 6, 7, * and 6, 3, *. Those three rules are summarized as two level reductions of generalized association rules. The summarized two level association rule of Clothing "tab number 6" is merged as single rule but all of the frequency measure for Clothing tab are 28 as shown in figure 4.11.

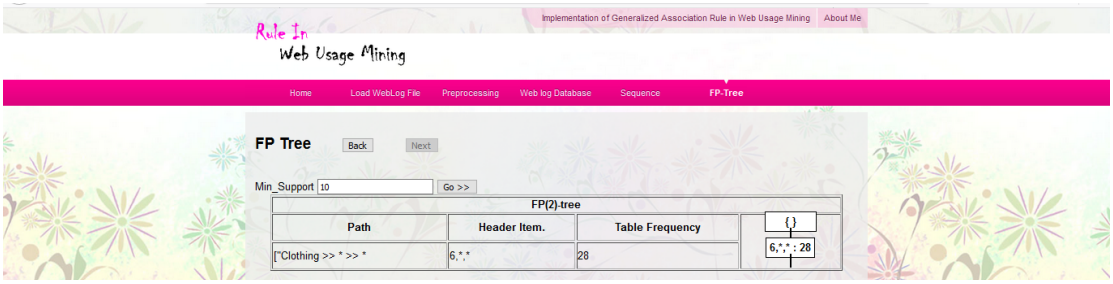


Figure 4.11System Generated Association Rules (Two Level)

4.12 Experimental Result

The proposed multilevel sodality rule algorithm (MAR) utilizing have been tested some datasets from the authentic world. The datasets are adopted from a www.flipkart.com, which consists of daily transaction records. Flipkart is a website of online store which deals for clothing, sport wear, foot wear, kid's clothing, and so on

for worldwide and daily transactions are enough for the proposed analysis. Dataset 1 recorded 1207 transactions in a week, and dataset 2 recorded 3509 transactions in a month. Both of them contain 611 items (the category names). In this experiment, these items can be divided into 3 concept levels. Level 0 is the primitive level, and level 1 is a higher calibre, which abstracts the primitive items to about 100 categories according to their product types. The system can be additionally implemented the FP-magnification algorithm at each independent concept level of these two datasets respectively. The same-level rules only associate the items in the same concept level. It could be found out the page with the highest probability to increment the popularity. Since the FP-magnification predicated algorithms customarily mine an independent concept level every time, the cross-level sodality cannot be ascertained. The cross-level rules are very arduous to mine by some level independent algorithms. The potential of our algorithm is that MAR would be more efficient when the total number of concept levels is astronomically immense or the concept hierarchy is dynamically transmuting.

CHAPTER 5
CONCLUSION, LIMITATION AND FURTHER EXTENSION

5.1 Conclusion

Discovering the sodality rule is a paramount data mining function. This system presents engendering sodality rule patterns from weblog data. In order to analyse the web utilization patterns, sodality rule mining algorithm is applied to the weblog. Relationships of web server log are exhibited as output in this system. Information is provided to better accommodate the website predicated on user's needs. By the study of the result, the web system can be reorganized the web site structure to access the associated link without delay. The current trend of the web utilizer can be analysed. Users can be proposed the approach of mining multilevel sodality rules by organizing and extracting frequent itemsets at the atomic level, which amends the mining efficiency without making the supplemental deviations to the FP(l)-tree while realizing the mining of cross-level sodality rules. Furthermore, this approach can fortify dynamic hierarchies predicated on divergent views of organizing items, which sanction different users to get their desired sodality rules from a customized perspective. Some preliminary tests have been carried out for performance evaluation of the incipient method in terms of running time and obtained promising results. Users discussed the issues of the calculation of rule support, the integrity of the exploited result and the trade-off between mundane sense and specific patterns and suggested possible solutions.

5.2 Benefits of the Proposed System

The proposed approach can be used to develop an association rule mining system. By introducing the dynamic feature of construction of concept hierarchy, it can also be encapsulated as a learning component and integrated with the know ware system to provide the knowledge source for other intelligent components. Strong associations discovered at higher concept levels may represent common sense knowledge, which is too generalized to be used in real application. With this understanding, this proposed

system can make a trade-off between the any kinds of rules. The cross-level rules can be viewed as a kind of trade-off in some sense.

5.3 Limitation and Further Extension

The system inhibition is the integrality of the result. For a given support threshold, the same-level sodality rules exploited by the MAR algorithm may be a subset of the rule set mined by the FP-magnification algorithm without any pruning strategy, because the MAR algorithm acclimates the bottom-up strategy. This betokens if the item sets at level l+1 (l≥0) are frequent, then they must be engendered as parents from some frequent itemsets at level l. In fact, there may additionally be some frequent itemsets at the calibre l+1, which are not the parent of any frequent item set at the calibre l. The MAR algorithm may miss these no-parent frequent itemsets at the calibre l+1. The result of the analysis is analyzed on the periodical data set of India's biggest online store: "Filpkart". This system can be extended for many type of data for any type of company data set to become a general purpose generalized association rule generating system for any type of data set.

REFERENCES

[1] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules", in Proceedings of the 1994 international conference on very large databases (VLDB'94), Santiago, Chile, pp 487–499, 1994.

[2] S. Brin, R. Motwani, J.D. Ullman and S. Tsur S, "Dynamic itemset counting and implication rules for market basket analysis", in Proceeding of the 1997 ACM-SIGMOD international conference on management of data (SIGMOD'97), Tucson, AZ, pp 255–264, 1997.

[3] H. Chen and M. Chau, "Web Mining: Machine Learning for Web Applications", Annual Review of Information Science and Technology (ARIST), 38:289-329, 2004.

[4] H. Chen, W. Chung, J.J. Xu, G. Wang, Y. Qin and M. Chau, "Crime Data Mining: A General Framework and Some Examples.", Computer, 37(4):50-56, 2004.

[5] H. Chen, J. Qin, E. Reid, W. Chung, Y. Zhou, W. Xi, G. Lai, A. Bonillas, and M. Sageman, "An Overview of Web Mining in Social Benefit Areas", Proceedings of the 7th International Conference on Intelligent Transportation Systems (ITSC), Washington D.C., October 3-6, 2004.

[6] D.W. Cheung, J. Han, V. Ng, A. Fu and Y. Fu, "A fast distributed algorithm for mining association rules", In Proceeding of the 1996 international conference on parallel and distributed information systems, Miami Beach, Florida, pp 31–44, 1996.

[7] A. Cohen and R. Nachmias, "A Quantitative Cost-Effectiveness Model for Web-Supported Academic Instruction", The Internet and Higher Education, Vol. 9, No 2, pp.81–90, 2006.

[8] B. Galina and T. Georgieva, "Discovering the Association Rules in OLAP Data Cube with Daily Downloads of Folklore Materials", International Conference on Computer Systems and Technologies, 2005.

[9] Hsu Mon Thet Wai, "Discovery of user access pattern from weblog based on Association Rule Mining", M.C.Sc. Thesis, University of Computer Studies, Yangon, 2010.

[10] J. Han, H. Cheng, D. Xin, and X. Yan, “Frequent pattern mining: current status and future directions”, Data Min Knowl Disc, 15:55–86, DOI 10.1007/s10618- 006-0059-1, 2007.

[11] J. Han, M. Kamber and J. Pei, “Data Mining: Concepts and Techniques”. 3rd ed. Waltham: Morgan Kaufmann, 2012.

[12] J. Jadav Jigna and M. Panchal, “Association Rule Mining Method On OLAP Cube”, In proceeding of International Journal of Engineering Research and Applications, Vol. 2, Issue 2, pp.1147-1151, Mar-Apr 2012.

[13] J.S. Park, M.S. Chen and P.S. Yu, “An effective hash-based algorithm for mining association rules”, in Proceeding of the 1995 ACM-SIGMOD international conference on management of data (SIGMOD’95), San Jose, CA, pp 175–186, 1995.

[14] S. Sarawagi, S. Thomas, R. Agrawal, “Integrating association rule mining with relational database systems: alternatives and implications”, in Proceeding of the 1998 ACM-SIGMOD international conference on management of data (SIGMOD’98), Seattle, WA, pp 343–354, 1998.

[15] A. Savasere, E. Omiecinski and S. Navathe, “An efficient algorithm for mining association rules in large databases”, in Proceeding of the 1995 international conference on very large databases (VLDB’95), Zurich, Switzerland, pp 432–443, 1995.

[16] H. Toivonen, “Sampling large databases for association rules”, In Proceeding of the 1996 international conference on very large databases (VLDB’96), Bombay, India, pp 134–145, 1996.

[17] Y. Wan, Y. Liang, and L. Ding, “Mining Multilevel Association Rules from Primitive Frequent Itemsets”, Journal of Macau University of Science and Technology, Vol.3, No.1, 2009.