Information Retrieval System using BM25, Pivoted Normalization and CombSUM Method

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

Abstract

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

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

1. Introduction

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

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

2. Related Work

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

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

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

3. Background Theory

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

3.1. Indexing

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

3.1.1. Inverted Index

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

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

Table1: Example of Inverted Index

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

3.2. Okapi (BM25)

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

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

Require: query Q and Document Collection

BEGIN

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

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

END

Figure 1:Pseudo-code for Okapi (BM25)

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

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

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

3.3. Pivoted Normalization Method

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

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

END

Figure2: Pseudo code for Pivoted Normalization Method

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

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

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

3.4. CombSUM

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

1. scored-based
2. ranked-based

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

For document (i)

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

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

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

4. Implementation and Design of the Proposed System

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

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

4.1. Pre-Processing

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

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

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

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

Inverted index is used in this proposed system.

4.2. Proposed system for Admin and User

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

1) Admin portion, and
2) User portion

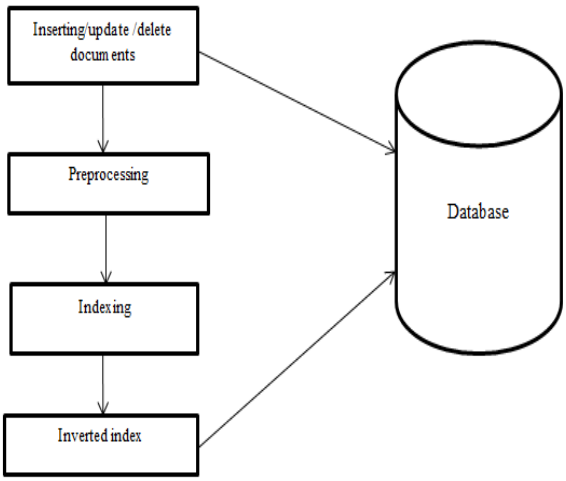

Figure 3: System Flow (For Admin)

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

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

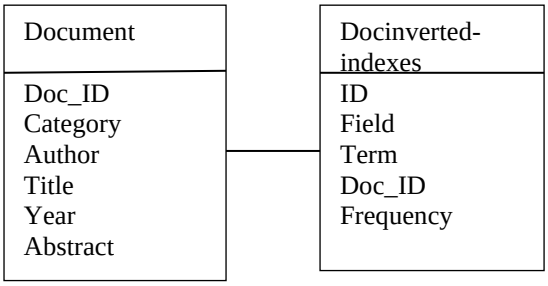

Figure 4: Database Design of the System

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

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

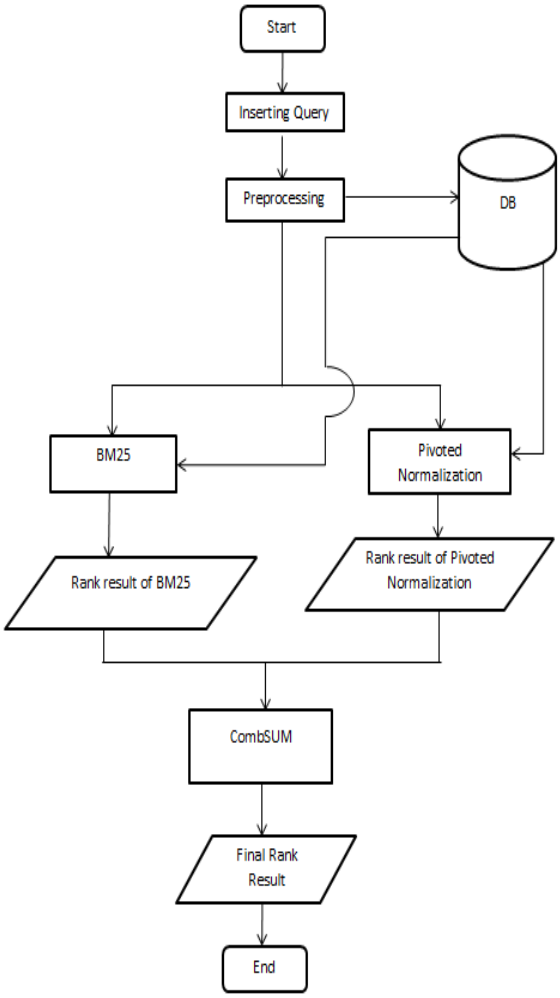

Figure 5: System Flow (For User)

5. System Evaluation

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

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

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

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

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

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

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

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

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

6. Conclusion

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

REFERENCES

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

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

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

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

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

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

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