American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)

ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

http://asrjetsjournal.org/

Improving the Effectiveness of Information Retrieval System

Su Mon Phyo[a]*, Moe Moe Aye[b]

[a,b]*Department of Computer Engineering and Information Technology, MTU, Mandalay, Myanmar*

[a]*Email: sumonphyo13586@gmail.com*

[b]*Email: moeaye255@gmail.com*

Abstract

With the rapid growth of information and easy access of information, in particular the boom of the World Wide Web, the problem of finding useful information and knowledge becomes one of the most important topics in information and computer science. Information Retrieval (IR) systems, also called text retrieval systems, facilitate users to retrieve information which is relevant or close to their information needs. This research provides an effective IR system for retrieving not only relevant but also related documents. For retrieving relevant documents, Probabilistic Model is applied. For retrieving related documents, the related indexed table is built including extracted keywords and related documents lists. In constructing related index table in the database, Shannon's entropy difference between intrinsic and extrinsic mode is used to extract the highly significant keywords. Entropy threshold value was assigned to 0.5 of normalized entropy difference square (ED^2_{nor}) according to the analytical results. The proposed keyword similarity distance (KSD) function is used to calculate similarity and relations between document pair. The proposed system is implemented by using PHP programming language and MySQL database. The performance of this approach is evaluated by using standard IR metric such as Precision (P), Recall (R), F-measure (F) and Average Precision (AP) on three test datasets (Oshumed, CISI and CRAN). According to the experimental results, the performance of the proposed system using related index table is more effective than the traditional probabilistic model.

Keywords: Information Retrieval; Probabilistic Model, Keyword Extraction; Keyword Similarity Distance; Related Index.

---

* Corresponding author.

1. Introduction

IR is the process of searching and retrieval of information from documents that matches user query. Generally, IR is used to retrieve relevant information such as documents with respect to user query in short response time. Obvious examples include search engines as Google, Yahoo or Microsoft Live Search. Many university, corporate, and public libraries now use IR systems to provide access to books, journals, and other documents. There are too much documents available in dataset and user finds difficult to get related documents he wants. So, in order to ease work of user document retrieval is used. Document retrieval is IR task in which information is extracted by matching text in documents against user query. Documents related to the user query should be retrieved in acceptable time [1].

Several existing IR system can support retrieving only relevant information. Related information cannot be retrieved because these models ignore important relationship between word pair in each document pair. To get the related results, the relation between words should be considered. The relational similarity function between word pair in each document pair is used in the proposed system.

Measuring the similarity between a pair of documents is becoming increasingly important task as the document collections are growing rapidly. Whether clustering a corpus of documents, or searching for relevant documents related to a query text/ document, similarity measure is critical and unavoidable for performing these tasks. Comparing the similarity between documents has many different purposes such as checking plagiarism, classifying text and retrieving relevant information [2]. A similarity measure is a function that assigns a numerical value between 0 and 1 to a pair of objects. The value zero represents that the two objects are totally different, while a value of one suggests that the two objects are identical under the feature sets that are used to represent similarity functions [3].

In this research, a new similarity function called KSD is proposed to calculate the keyword similarity distance and average similarity score.

2. Related Work

Measuring the similarity between words, sentences, paragraphs and documents is an important component in various tasks such as information retrieval, document clustering, word-sense disambiguation, text classification, topic detection, and text summarization.

T.K. Landauer and S.T. Dumais proposed the technique of Corpus-Based similarity so called Latent Semantic Analysis (LSA). A matrix containing word counts per paragraph (rows represent unique words and columns represent each paragraph) is constructed from a large piece of text and a mathematical technique which called singular value decomposition (SVD) is used to reduce the number of columns while preserving the similarity structure among rows. Words are then compared by taking the cosine of the angle between the two vectors formed by any two rows [4].

Wei He and his colleagues [5] proposed a method for measuring semantic relatedness between words by using

lexical context in which first for each word of a word-pair, a lexical context is created using Word Net, which constitutes words that are highly related to the target word. In next step, semantic relatedness between a word and lexical context of another word for an original word-pair is calculated using Web Dice coefficient. Performance of the system was verified through experiment using Miller-Charles benchmark dataset achieving a Pearson correlation coefficient of 0.912. An alternative better method needs to be used to extract lexical context for a word.

K. Lund and his colleagues. proposed the Hyperspace Analogue to Language (HAL) model [6]**.** This model creates a semantic space from word co-occurrences. A word-by-word matrix is formed with each matrix element is the strength of association between the word represented by the row and the word represented by the column. HAL has been successfully applied to query expansion in IR, but has several limitations, including high processing cost and use of distributional statistics that do not exploit syntax.

3. Proposed System

The proposed system aims to provide not only relevant but also related documents by using probabilistic model and related index table including extracted keywords and related documents list. In the system, there are three mainly components namely preprocessing, probabilistic model and constructing related index table. The design of the proposed system is shown in Figure 1.

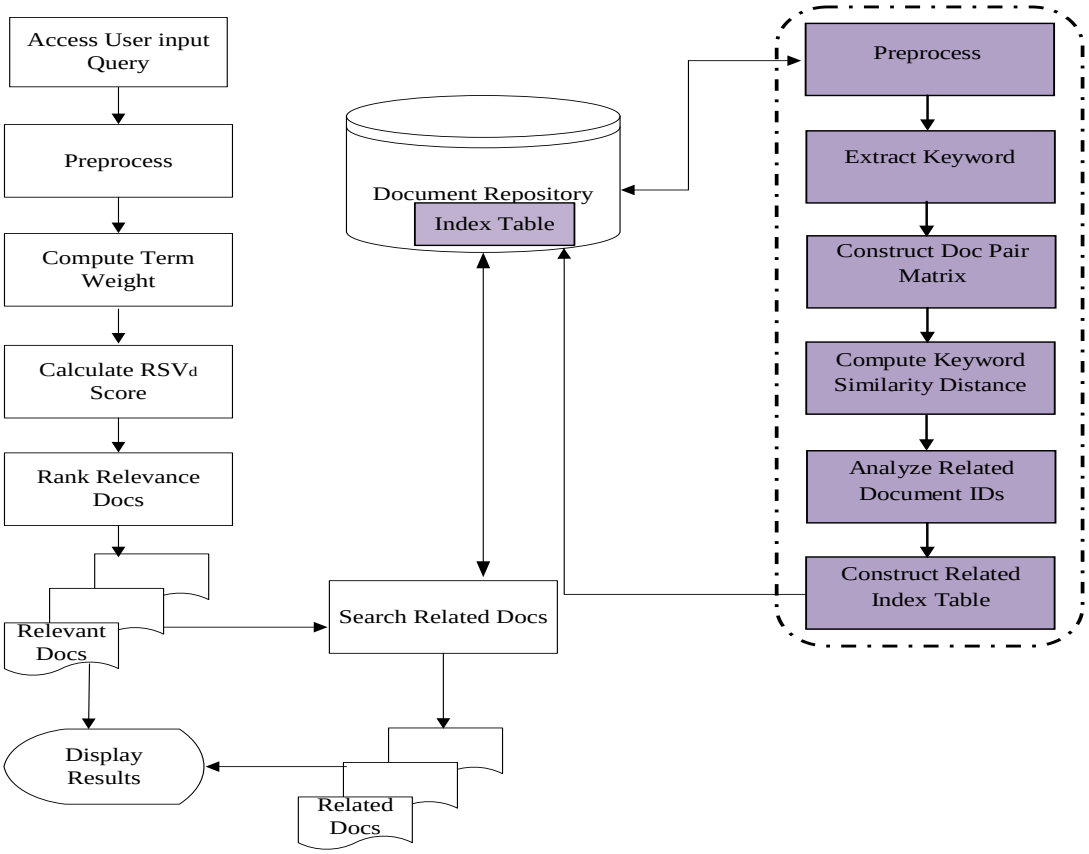

Figure 1: Design of the proposed system

3.1. Preprocessing

User gives the query term to the search engine as request. Firstly, user query and input documents in database are pre-processed. The pre-processing step involves four phases: special character removal, stop words removal, case conversion and stemming. There are 32 special characters such as ~, !, @, #, $, etc. The stop words list contains 543 words in nine syntactic classes (conjunctions, articles, particles, prépositions, prônons, anomalous verbs, adjectives, and adverbs). And all the upper case letters from the entire document are converted to lower case to overcome the case-sensitivity problem in searching. Then, these words are converted to the root form (e.g. “studies” => “study”) by stemming process.

3.2. Probabilistic Model

There are different IR models to retrieve relevant information according to the user’s needs. There are four main models: Boolean model, vector space model, language model and probabilistic model. The IR system has an uncertain understanding of the user query and makes an uncertain guess of whether a document satisfies the query. Probability theory provides a principled foundation for such reasoning under uncertainty. Probabilistic IR models exploit this foundation to estimate how likely it is that a document is relevant to a query. In classical probabilistic approach, the first influential model is the Binary independence model (BIM) and a more modern, better performing model is the Okapi BestMatch25 (BM25) [7]. In this research, BM25 probabilistic model is used for achieving most relevant information.

Okapi BM25 is a probabilistic model that incorporates term frequency (i.e., it’s non-binary) and length normalization. BIM was originally designed for short catalog records of fairly consistent length, and it works reasonably in these contexts. For modern full-text search collections, a model should pay attention to term frequency and document length. BM25 is sensitive to these quantities. In the simplest version of BIM, the RSV (Relevant Status Value) score for document d is just idf weighting of the query terms in the document as in Eqn. (1):

$$RSV_d = \sum_{t \in q \cap d} \log \frac{N}{dft} \qquad (1)$$

The idf term [log N/dft] is improved by factoring in term frequency and document length as describe in Eqn. (2).

$$RSV_d = \sum_{t \in q} \left[\frac{N}{dft} \right] \cdot \frac{(k_1 + 1) tf_{td}}{k_1((1-b) + b \times (\frac{L_d}{L_{ave}})) + tf_{td}} \qquad (2)$$

, where

q : query

N : total number of terms in document d

tftd : term frequency in document *d*

Ld (Lave) : length of document d (average document length in the whole collection)

k1 : tuning parameter controlling scaling of term frequency (*k1* = 2)

b : tuning parameter controlling the scaling by document length (*b* = 0.75)

3.3. Constructing Related Index Table

To construct the related index table, highly significant keywords are extracted by using Shannon's entropy difference between the intrinsic and extrinsic mode. After generating keywords, keyword similarity distance for each pair of documents is computed by utilizing the proposed KSD function and the decision rule. Finally, related index table is constructed and stored in the database.

3.3.1. Keyword Extraction Process

After the preprocessing step, keywords are extracted by using the Shannon's entropy difference between intrinsic and extrinsic mode and ranked according to maximum entropy value. To distinguish the intrinsic and extrinsic mode, the positions of the word occurrences in a text with frequency m are denoted by $t_1, t_2, t_3, \ldots, t_m$.

The distance between two successive occurrences of a word can be calculated as in Eqn. (3):

$$d_i = t_{i+1} - t_i \qquad (3)$$

The average distance (mean value) is constant value defined by Eqn. (4):

$$\mu = \text{total distance / word frequency} \qquad (4)$$

The arrival time differences d_i belongs to the intrinsic mode d^I if $d_i \leqq \mu$. Let $d^I = \{ d_i \mid d_i \leqq \mu \}$ be the union set for all $d_i \leqq \mu$. Thus, the intrinsic modes entropy of a word is defined as in Eqn. (5):

$$H(d^I) = -\sum_{d \in d^I} P_d \log_2 P_d \qquad (5)$$

Let $d^E = \{ d_i \mid d_i > \mu \}$ be the union set for all $d_i > \mu$. The extrinsic mode entropy of a word state as Eqn. (6):

$$H(d^E) = -\sum_{d \in d^E} P_d \log_2 P_d \qquad (6)$$

Thus the entropy differences between the intrinsic and extrinsic mode can be define as in Eqn. (7)

$$ED^q(d) = (H(d^I))^q - H(d^E)^q \qquad (7)$$

The top-N accuracy rate and average precision (AP) for the ED_{nor} metric with different q is vary from q = 2 to q = 5. ED^2_{nor} is used because it achieves a better average top-N accuracy rate and AP than the other metrics (ED^1, ED^3_{nor}, ED^4_{nor} and ED^5_{nor}).

After the keywords have extracted with ED^2_{nor}, top-rank keywords are filtered with the specific threshold value. Threshold values can be assigned between 0 and 1. Among them, the value 0.5 gives the better related result than others according to the experimental results [8]. So, ED^2_{nor} of 0.5 is used to extract the precise value of keywords in the proposed system. Figure 2 illustrates the extracted keywords from ohs1.txt file. The filtered keywords, word count and entropy difference values are demonstrated.

Significant keywords have been extracted from each document and then the document pair matrix (Da,Db) is constructed as depicted in Figure 3 to compare keywords for each document pair. The document pair similarity is computed by using the proposed KSD function.

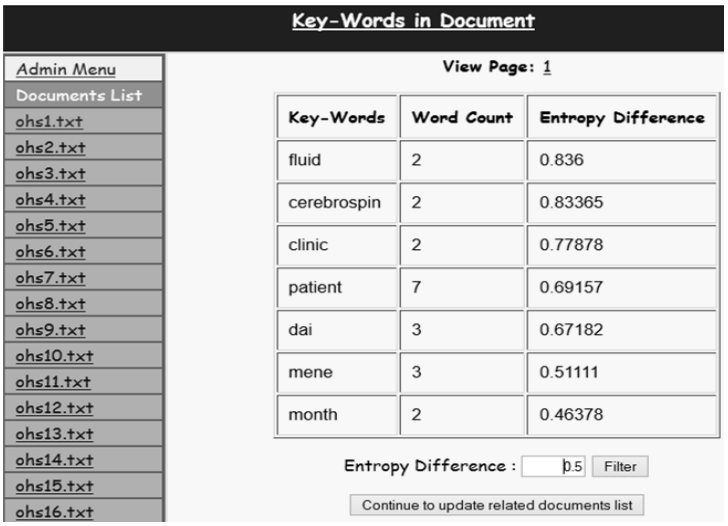

Figure 2: Extracted keywords from ohs1.txt at entropy threshold (0.5)

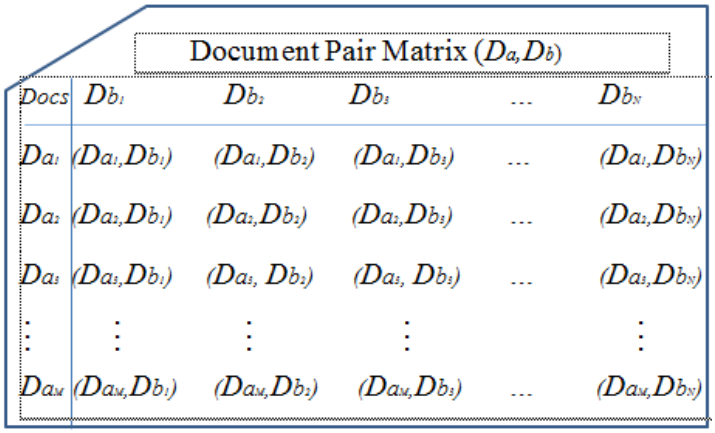

Figure 3: Document pair matrix *(Da,Db)*

3.3.2. Keyword Similarity Distance (KSD)

In the proposed system, a new similarity function is proposed to calculate the keyword similarity distance for each pair of documents. The comparison of each document pair represents Doc*(Da,Db)*. For example, *Da* is compared with D_b for each pair documents in database. According to the Eqn. (8), document pair calculation is the form of symmetric matrix. So, it is needed to calculate the upper half of the distance values and reduce half of the processing time. Because of the left half is synchronized with the upper half of the main diagonal.

The proposed Keyword Similarity Distance (KSD) function is as in Eqn. (8**)**:

$$KSD(D_a, D_b) = \{MN - [\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(w_i - w_j)]\} \% Min(K_{Da}, K_{Db})$$

$$\begin{matrix} \textbf{if } \boldsymbol{w_i = w_j \rightarrow 0} \\ \boldsymbol{w_i \neq w_j \rightarrow 1} \end{matrix} \quad (8)$$

, where

Da = document *a*

Db = document *b*

M = number of keywords in documents *Da*

N = number of keywords in documents *Db*

w_i = one word of i keywords in D_a

w_j =one word of j keywords in D_b

Keyword similarity distance (KSD) value in *(Da,Db)* is the difference between total number of keywords (*MN*) and number of dissimilar keywords between two documents $[\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(w_i - w_j)]$. To get dissimilar keywords, if the words in *Da* and *Db* are the same tends to zero and otherwise, tends to one. Then, the modulo operation (sometimes called modulus) finds the remainder after division of $\{MN - [\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(w_i - w_j)]\}$ by $Min(K_{Da}, K_{Db})$ in order to achieve the same remainder value for same document pair eg., *(Da,Da) and (Db,Db).*

3.3.3. Analyzing Related Documents IDs

After calculating KSD values for all pair of documents, related document list have to analyze from KSD value between document pair by using the decision rule as indicated in Eqn. (9**)**. It can determine whether each

document is related with other or not and made the relation between documents.

$$\textit{If } KSD(D_a, D_b) > Min\ (M, N)/2 \textit{ then return related.} \quad (9)$$

Finally, the related index table including extracted keywords, contents, filtered contents and related IDs as shown in Figure 4 is built. Then, this table is stored in the database for retrieving related information according to the user query.

DocID	DocName	KeyWords	Contents		FilterContents		RelatedIDs
1	ohs1.txt	cerebrospin,clinic,dai,fluid,mening,	Haemoph...	1K	haemophilu influenza me...	647B	29
2	ohs2.txt	aeromona,ascend,bacteri,cholang,cult	Mucosal...	1K	muco intussuscept avoid...	784B	
3	ohs3.txt	abnorm,absorpt,gastrointestin,normal	Gastroi...	1K	gastrointestin function...	930B	13,29
4	ohs4.txt	biologi,growth,improv,metaphys,molec	Epidemi...	962B	epidemiolog bone joint ...	562B	42
5	ohs5.txt	aerob,bone,soft,tissu	The dia...	1K	the diabet foot soft ti...	589B	45
6	ohs6.txt	infect,reduc	Infecti...	602B	infect total joint repl...	346B	14,30,43
7	ohs7.txt	antibodi,ic,lp,respons	Analysi...	1K	analysi immun respon li...	592B	
8	ohs8.txt	grow,high,intestin,larg,pig,protein,	Organ h...	1K	organ hypertrophi respo...	1K	
9	ohs9.txt	avp,children,concentr,fluid,plasma	Normali...	1K	normal plasma arginin v...	877B	
10	ohs10.txt	activ,bacteri,defect,fmlp,function,g	Impairm...	2K	impair neutrophil chemo...	1K	29
11	ohs11.txt	cerebr,patient,varicella	Varicel...	652B	varicella delai hemiple...	428B	26,29
12	ohs12.txt	antibodi,guthri,hiv,posit,pregnanc,r	Prevale...	2K	preval hiv antibodi pre...	1K	
13	ohs13.txt	abnorm,absorpt,gastrointestin,normal	Gastroi...	1K	gastrointestin function...	930B	3,29
14	ohs14.txt	antibodi,cell,cultur,detect,hiv,infe	Resista...	1K	resist infect hiv perip...	926B	6
15	ohs15.txt	characterist,featur,studi	Endemic...	1K	endem kaposi sarcoma hu...	991B	130
16	ohs16.txt	antibodi,igm,measur,patient,platelet	Assessm...	1K	assess platelet antibod...	918B	29
17	ohs17.txt	aid,common,imag,lymphoma,mr,patient,	Use of ...	1K	us ct mr imag distingui...	770B	25,29

Figure 4: Related index table

4. Standard Evaluation Measures for IR System

The standard evaluation metrics such as precision, recall, F-measure and average precision have been used for evaluating performance of the IR system. In the standard IR terminology, these metrics are defined as follows:

Precision (P): The ratio between the number of relevant documents in retrieved documents and the total number of retrieved documents as in Eqn. (10**)**.

$$Precision = |Relevant\ And\ Retrieved| / |Retrieved| \quad (10)$$

Recall (R): The ratio between the number of relevant documents in the retrieved documents and the total number of relevant documents as in Eqn. (11).

$$Recall = |\ Relevant\ and\ Retrieved| / |Relevant| \quad (11)$$

F-measure (F): A measure that combines precision and recall is the harmonic mean of precision (P) and recall(R).The F-measure is computed using the average precision and average recall values as in Eqn. (12).

$$F1 = 2\ |Relevant\ and\ Retrieved| / (|Relevant| + |Retrieved|) \quad (12)$$

Average Precision (AP): Average of the precision values at the points at which each relevant document is

retrieved for each query.

In good IR systems, high recall (but low precision) can get by retrieving all documents for all queries. The fact that recall is a non-decreasing function of the number of retrieved documents and precision is a decreasing function of the number of retrieved documents.

Average precision can be computed at different levels of recall. F-measure (F) is a weighted harmonic mean of precision and recall. Harmonic mean is a conservative average that is when the value of two numbers differs, harmonic mean is closer to their minimum than arithmetic or geometric mean.

5. Results and Discussion

In this research, a retrieval engine is proposed to improve the effectiveness of the IR system. So, the proposed search engine model depicted in Figure 5 can provide both relevant and related information according to the user query.

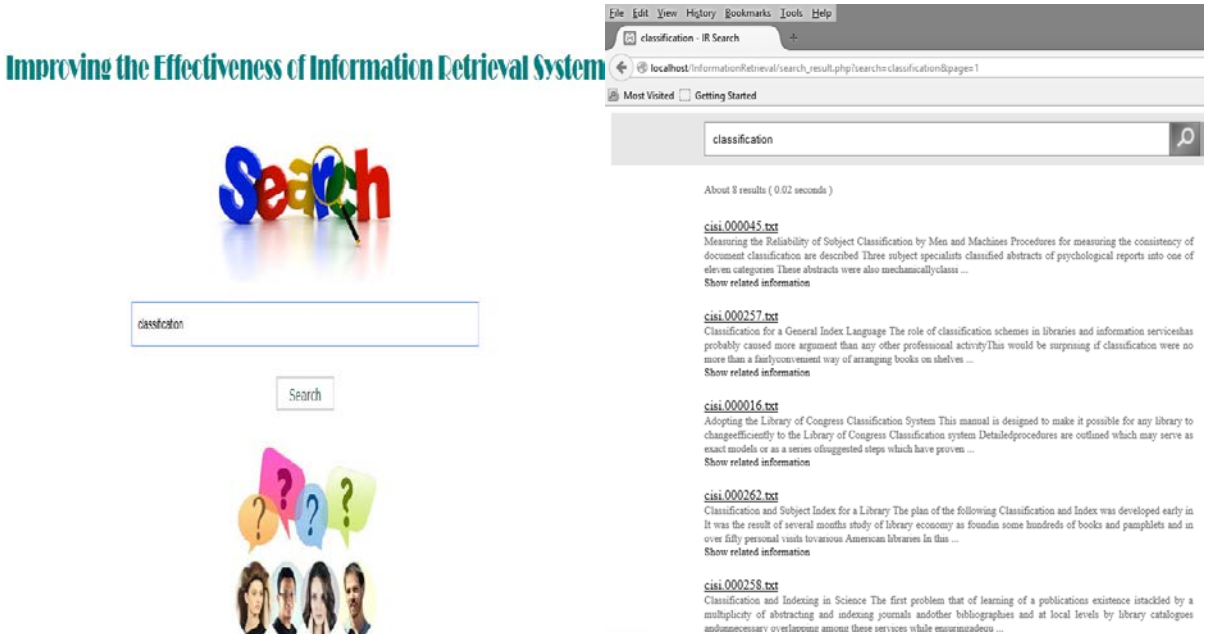

Figure 5: Proposed search engine model

Three standard test datasets, CISI, Oshumed and Cran collections, as shown in Table 1, are used in this research. There are total 150 input documents for different datasets and 15 queries are tested to evaluate the performance measures.

Experiments are run for three test datasets with 50 individuals in each dataset. The results of Precision, Recall, F-measure and Average Precision (AP) values for different queries are described in Table 2.

In the testing results obtained from the experiments, precision values for Quer1 to Query 6 (except Query2) are lower and then slightly higher for Query7 to Query15. But, the highest recall values for every query can be

achieved. This means that the ability of the search to find all of the relevant documents in the dataset. The balance value between precision and recall (F-score) has the significant value (>=0.5) that is the tradeoff between higher recall and lower precision. Finally, the precision value for each query needs to average to get the optimal results of the proposed system illustrated in Figure 6.

Table 1: Details of the document collection

Collection Name	Description	Number of Documents	Number of Queries
CISI	Information Science Abstracts	50	5
Oshumed	Medical Abstracts	50	5
CRAN	Aeronautical Collection	50	5

Table 2: Precision, recall and f-measure values for different datasets

Query	Instances	Relevant Docs	Precision	Recall	F-measure	AP
1.classification	150	6	0.4	1	0.6	0.8
2.retrieval	150	17	0.6	0.7	0.6	0.9
3.indexing	150	8	0.3	1	0.5	0.4
4.information system	150	34	0.4	0.97	0.6	0.8
5.data base	150	29	0.4	1	0.6	1
6.malaria	150	5	0.3	1	0.5	1
7.HIV	150	8	0.7	1	0.8	1
8.antibiotics	150	8	0.7	1	0.8	1
9.diabetic	150	3	0.8	1	0.9	1
10.leukemia	150	3	1	1	1	1
11.aerodynamics	150	9	0.9	1	0.95	1
12.missile	150	3	0.8	1	0.9	1
13.aircrast	150	5	1	1	1	1
14.atmosphere	150	5	0.7	1	0.8	1
15.thermal	150	4	0.6	1	0.75	1

In the proposed system, higher recall value gets for retrieving both relevant and related documents. The harmonic mean (F-measure) that takes into both precision and recall value achieves significant value (more than fifty percent). Moreover, the fact that the average precision for each query reaches maximum value, 1, means that the system can return all documents for completeness of the user query.

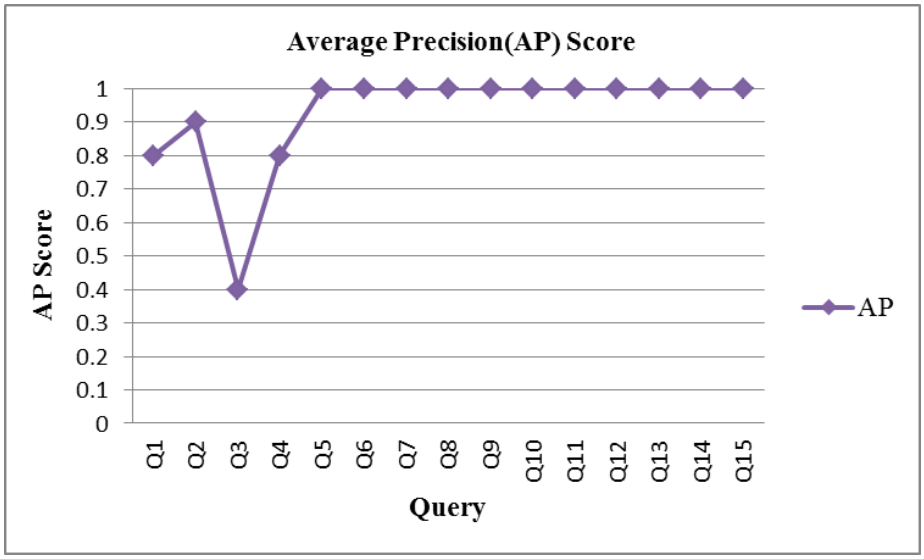

Figure 6: AP score for each query

6. Performance Improvement Using Related Index Table

The performance of the proposed system is evaluated using standard measures as in above section. And, then the evaluation measures are calculated in order to achieve the performance using only probabilistic model and using with related index table for three test datasets is depicted in Figure 7, 8 and 9.

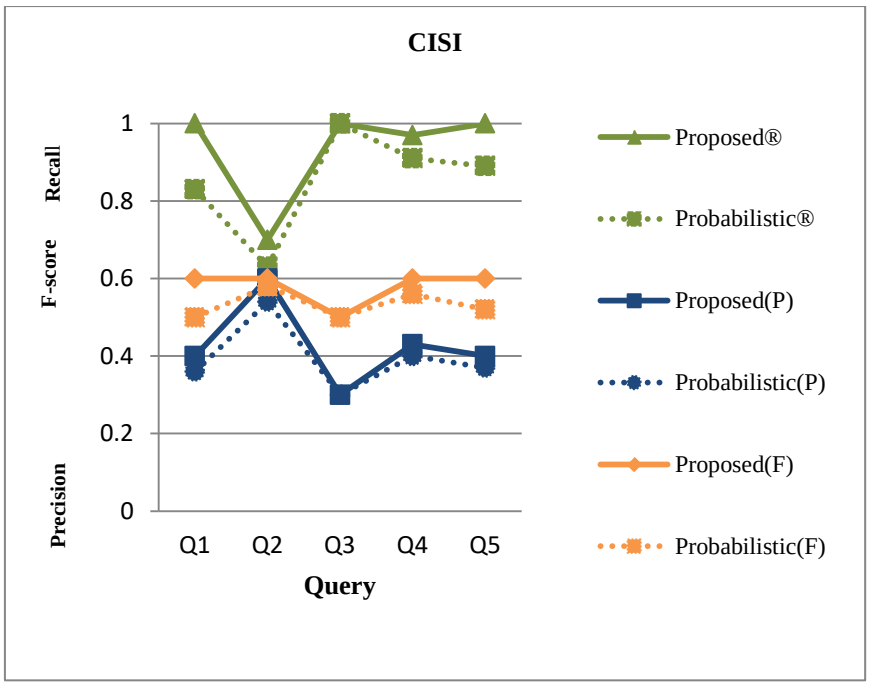

Figure 7: Performance comparison using CISI (Queries1-5)

Precision, Recall and F-score values are calculated for Query 1 to Query 5 using CISI dataset for using only probabilistic model and using with related index table. The improvement of the performance using related index

table is significant (recall (17%) and F-score (10%)) than only probabilistic model although precision value is insignificant (6%). Queries 6 to 10 from Oshumed dataset are also evaluated and the proposed system achieved the performance improvement of 67% than the probabilistic model. The performance of the proposed system and probabilistic model has no different for Queries 11 to 15 (Cran collections). Therefore, the proposed IR system using related index table is more effective than the traditional IR model. The existing IR system using only probabilistic model can't provide both relevant and related information for the searching user needs.

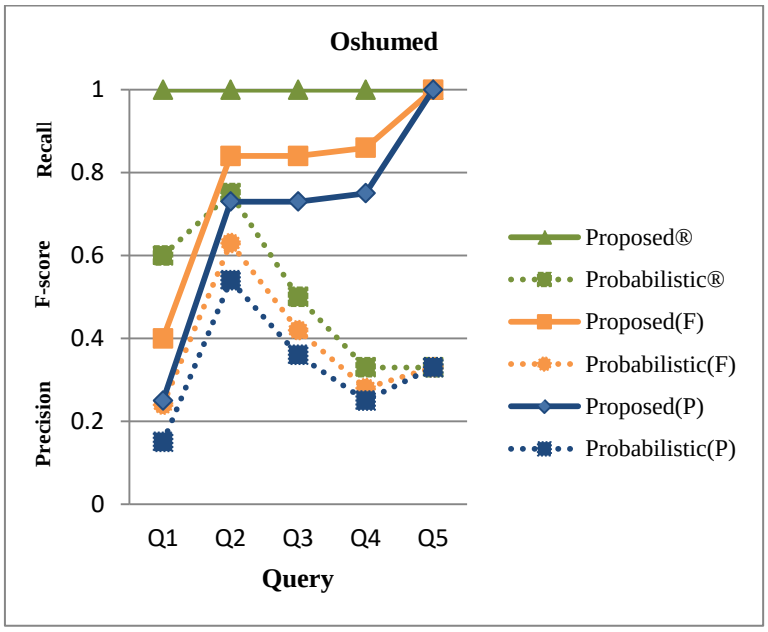

Figure 8: Performance comparison using Oshumed (Queries 6-10)

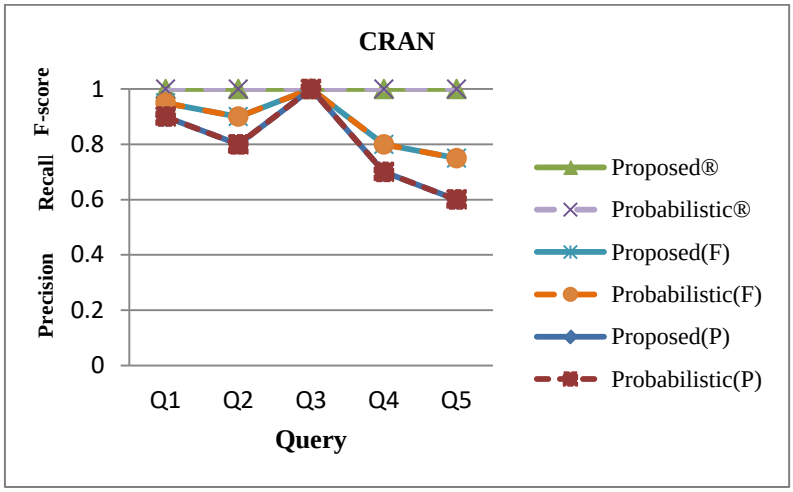

Figure 9: Performance comparison using CRAN (Queries 11-15)

7. Conclusion

The proposed IR system can provide not only relevant but also related documents on evaluation of three test datasets using 15 queries. According to the experimental results, the standard metric recall value for each query

has higher value at each precision value. It means that there are no relevant document missing and the system can satisfy user needs for completed search. Moreover, one measure of performance that takes into account both recall and precision value (F-measure) is effective for achieving more than 50 % performance. The effectiveness of the system has been analyzed using several test collections by standard evaluation metrics and the results show that the proposed system gives the significant values of F-measure and average precision as high as 100%. The performance improvement of the proposed system is also compared with the existing IR model and the results showed that precision, recall and F-score has a significant improvement. The next step in evaluating the performance of the proposed system is to use the created database containing data mining domain. In addition, the pioneering test collection (Cran field) will be used to evaluate precise quantitative measures of IR effectiveness.

References

[1] M. Donge and V. Nandedkar, Information Retrieval using Context Based Document Indexing, International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), vol. 2, issue. 12, Dec 2014.

[2] J. Singh, P. Singh, Y. Chaba, Performance modeling of information retrieval techniques using similarity functions in wide area networks, International Journal of Advanced Research in Computer Science and Software Engineering**,** vol.4, issue.12, Dec 2014.

[3] J. Singh, P. Singh, Y. Chaba, Performance evaluation and design of optimized information retrieval techniques using similarity functions in wide area networks, IJARCSSE, vol.5, issue.1, Jan 2015.

[4] T.K.Landauer and S.T Dumais, “A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge", Psychological Review, 1997.

[5] R. K. Rajpal and Y. Rathore, “A Novel Techinque For Ranking of Documents Using Semantic Similarity”, International Journal of Computer Science and Information Technologies (IJCSIT), vol. 5, 2014.

[6] K. Lund and C.Burgess, “Producing high-dimensional semantic spaces from lexical co-occurrence”, Behavior Research Methods, Instruments & Computers, p. 203-208, 1996.

[7] H. Sch¨utze, Introduction to Information Retrieval, http://informationretrieval.org, Institute for Natural Language Processing, University¨at Stuttgart, Aug 2011.

[8] S. M. Phyo, L. W. Kyi, Developing related index table for effective IR system, The Sixth International Conference on Science and Engineering (ICSE), Myanmar, Dec 2015.