An Approach to Improve Precision and Recall for Ad-hoc Information Retrieval using SBIR Algorithm

R. Thamarai Selvi
Assistant Professor and Head
Department of Computer Applications,
Bishop Heber College,
Trichirappalli, India
thams_shakthi@yahoo.co.in

E. George Dharma Prakash Raj
Assistant Professor
School of Computer Science, Engineering and Technology,
Bharathidasan University, Trichirappalli, India
georgeprakashraj@yahoo.com

***Abstract* - Information Retrieval is a process of finding the documents in a collection based on a specific topic. The information need is expressed by the user as a query. Documents that satisfy the given query in the judgment of the user are said to be relevant. The documents that are not of the given topic are said to be non-relevant. An IR engine may use the query to classify the documents in a collection, returning to the user a subset of documents that satisfy some classification criterion. There are several search engines to find information in the given repositories containing large amounts of unstructured form of text data. However, the task of ad hoc information retrieval is, finding documents within a corpus like Bible, that are relevant to the user remains a hard challenge. Sometimes the relevant documents may not contain the specified keyword. The lack of the given term in a document does not necessarily mean that the document is not a relevant. Because more than one terms can be semantically similar although they are lexicographically different. In this paper a new algorithm called "Semantic based Boolean Information Retrieval" (SBIR) is proposed to retrieve the documents with semantically similar terms to enhance the performance of Boolean Information Model by improving the recall and precision.**

Keywords - Information Retrieval, Semantic, WordNet, Stemming Algorithm, Boolean Information Retrieval.

I. INTRODUCTION

The abundance of information available in on-line repositories can be highly beneficial for both humans and automated computer systems that seek information, yet posses extremely difficult challenges due to the variety and amount of data available. The necessity of developing effective methods of automated IR has grown in importance because of the large amount of unstructured data. Search engines have become a crucial tool upon which millions of users are dependent for finding desired information. One of the core problems that search engines face in order to satisfy users' information needs is judging whether a piece of information is relevant to a given information need as specified by a text query. Indeed, there are various challenges involved in estimating the relevance of a text span to an information need underlying a query. For example, users often use queries that contain very few terms to describe their information needs and such queries are ambiguous in many cases.

Information retrieval is currently being applied in different application domains such as database systems, Web information search engines etc. The task of ad hoc information retrieval is, finding documents within in a corpus that are relevant to information need specified using a query. The main idea is to locate documents that contain terms that the users specify in queries. The lack of the given term in two documents does not necessarily mean that the documents are not related. Retrieval, by classical information retrieval models (e.g., Vector Space, Probabilistic, Boolean) [1] is based on lexicographic term matching. Therefore, these methods do not retrieve documents with semantically similar terms. In this paper a new algorithm is proposed to retrieve the semantically relevant documents with the use of WordNet database, stemming algorithm and a simple Boolean Information Retrieval model. WordNet is an on-line lexical reference system developed at Princeton University. It attempts to model the lexical knowledge of a native speaker of English. And it can also be seen as an ontology for natural language terms.

The rest of this paper is organized as follows. In Section 2, describes previous work in Boolean information model and in information retrieval using WordNet. Section 3 defines the task of proposed work, its framework and algorithm along with the steps used. Section 4 contains the experimentation and analysis. Finally, the conclusion and future work is given in Section 5.

II. RELATED WORK

The significance of Boolean Information Retrieval (BIR) has been revealed in many retrieval systems because of its simplicity [2]. Most of the commercial IR systems use this Boolean model to predict that each document is either relevant or non relevant [3]. For a number of reasons, both historic and technical, Boolean queries are particularly common in professional search.

The number of studies over the years have shown that keyword queries are often significantly more effective [4,5,6]. Boolean queries [7] however, are easy for

978-1-4799-2877-4/14 $31.00 © 2014 IEEE
DOI 10.1109/WCCCT.2014.68

information professionals to manipulate and are essentially self-documenting in that they define precisely the set of documents that are retrieved. The semantic retrieval [8] approach is used to discover semantically similar terms using WordNet. In many works, WordNet is used to identify similar concepts that correspond to document words. In most cases morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of IR application. In linguistic morphology, stemming is the process for reducing inflected words to their stem, base or root form.

The Porter stemmer [9] is a context sensitive suffix removal algorithm. Removing suffixes by is an operation which is especially useful in the field of IR. WordNet expansion technique was used [10] over a collection with minimal textual information. It was also proposed a new method using WordNet for document expansion [11] for a random walk algorithm for a given full document, over the WordNet graph ranks concepts closely related to the words in the document.

III. PROPOSED WORK

Several search engines have been developed to find information in repositories containing large amounts of unstructured form of text data. Ad hoc information retrieval is finding documents within a corpus that are relevant to information need specified using a query. All classical information retrieval models retrieve the documents based on lexicographic term matching only. But, two terms can be semantically similar although they are lexicographically different. Therefore, retrieval by classical retrieval methods will not retrieve documents with semantically similar terms.

A fat book which many people refer is The Bible which contains many verses as documents. People may want to refer the verses which have the same meaning for the given word. Grepping through text can be a very effective process to find a word since the speed of modern computers is very high and the usage of useful patterns of wildcard pattern matching through the use of regular expressions.

In this paper a new algorithm called "Semantic based Boolean Information Retrieval" (SBIR) is proposed to retrieve the documents with semantically similar terms to enhance the performance of Boolean Information Model by improving the recall and precision.

A. SBIR Framework

Several methods have been implemented and evaluated to improve the performance of search process. The information need is expressed by the user by giving a search word. Documents that satisfy the given query in the judgment of the user are said to be relevant. Documents that are not about the given topic are said to be non-relevant. An IR engine uses the search word to classify the documents in a collection, returning to the user a subset of documents that satisfy some classification criterion. Here the classification criteria is based on the synsets of the given search word.

The Semantic based Boolean Information Retrieval approach to information retrieval provides a novel perspective for approaching the task of ad hoc retrieval. In ad-hoc information retrieval, the user formulates any number of arbitrary queries and applies them to a fixed collection. The task of ad hoc information retrieval, finding documents within in a corpus that are relevant to an information need specified using a query. The framework for SBIR is given below in the Figure 1.

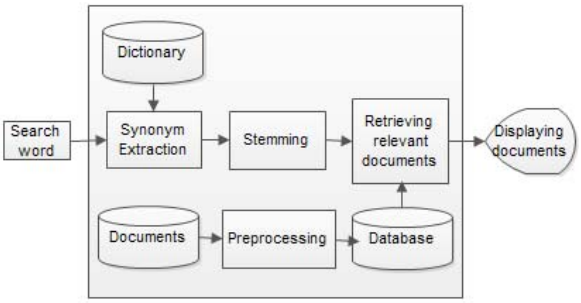

Figure 1 SBIR FRAMEWORK

B. SBIR Algorithm

This work targets to develop a system to address the Information Retrieval for a static data set and aims to provide documents from within the collection that are relevant to an arbitrary user information need. An information need is the topic about which the user desires to know more. SBIR retrieves documents from the document set by finding their synonyms using WordNet data base to find more similar documents are retrieved. The synonyms are then stemmed to find the root words using Porter Stemmer algorithm. And then the documents are retrieved for each stemmed word by Boolean Information Retrieval model. The proposed SBIR algorithm is given below.

```
preprocess the documents
enter a term t_m
  find semantically similar terms s_i and assign in T
  find root word st_i for each s_i, in U and assign in S
  for each st_i in S
  {
   find the documents d and put them in D_i
    }
  for each d_j in D_i
  {
    display the documents d_j
  }
```

The steps given in the algorithm are explained below.

Step 1: Preprocessing

The documents are preprocessed to extract the chapter name, chapter number, verse number and verses. This

information is stored in mysql database. This step is performed to retrieve the individual documents from the corpus and store them in the database. And also the keywords are extracted from the document by eliminating the stop words and stored them in the data base.

Step 2: Retrieving Synsets from WordNet

In the second step, we retrieve the synsets from the WordNet data base for the given keyword. These set of words of stored in a data structure array. WordNet is an on-line lexical reference system developed at Princeton University. WordNet attempts to model the lexical knowledge of a native speaker of English. WordNet can also be seen as an ontology for natural language terms.

Step 3: Stemming Process

In this third step, the stemming process is done on the synsets extracted from the second step. The Porter Stemmer Algorithm implemented in java performs this process for each word from the synsets. Porter's stemmer is more compact and easy to use than Lovins [12]. These words are stored in a vector. These stemmed words will be used to extract the documents from the data base.

The Porter stemming algorithm is a process for removing the commoner morphological and inflexional endings from words in English. Stemming algorithms are used in many types of language processing, text analysis systems, information retrieval and database search systems. Its main use is for term normalization process that is usually done when setting up Information Retrieval systems. Word stemming is an important feature supported by present day indexing and search systems. The idea is to improve recall by automatic handling of word endings by reducing the words to their word roots, at the time of indexing and searching.

Step 4: Retrieving Documents from the data base

In this fourth step, the relevant documents are retrieved from this data base for each word in the array and displaying the results to the user sequentially. In this experiment the Bible (Kings James Version) text is tested. This corpus contains 66 Books, 1189 Chapters, **31,102** Verses and 7,882,80 words. .

IV. Experimentation and Analysis

A. *Experimental Setup*

This SBIR algorithm which improves the precision and recall for Information Retrieval is implemented using Java programming language. SBIR is tested on Bible text. This text contains many small documents called verses. These verses are converted into individual documents and stored in the MySql database. Next the dimension of the documents is reduced using the stop word elimination and all the key terms are stored in the database.

Then the *synsets* are extracted from WordNet using Java APIs and stored in vectors to test the SBIR algorithm. After that, each term in the synset vector is stemmed to root words and stored in another vector. For example, the synsets for the word lord are creator, maker, divine, god, almighty, jehovah and master. These words are stemmed to their root word. After stemming process, the documents are retrieved from the database for all the root words.

B. *Evaluation of BIR and SBIR Algorithms*

The number of documents retrieved after getting the synsets is larger than the documents retrieved by using the keyword alone. And the number of documents retreived after stemming process of each synset is further increased.

Table 1 shows the number of documents that are retrieved by using BIR and SBIR.

TABLE 1 Number of Documents Retrieved

Words	BIR	SBIR
Rejoice	221	986
Faithfully	9	534
Oath	68	136
persecution	13	81
commitment	0	212
Sincerely	3	123
crucification	12	190
Hell	51	60
Jesus	865	1282

Fig. 2 shows the overall performance of BIR and SBIR algorithms. For some queries BIR does not give any document whereas SBIR retrieves the some documents which satisfy the user.

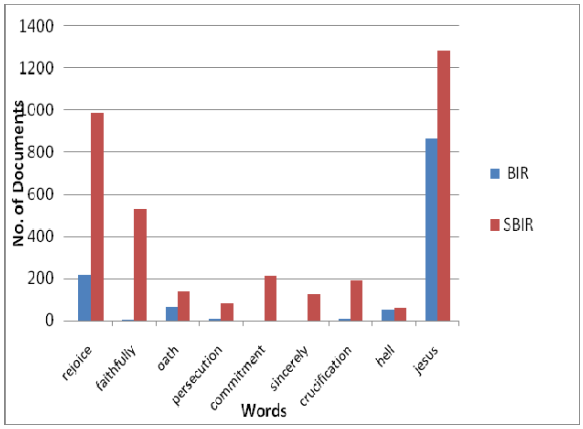

Fig. 2 Analysis of the Performance of BIR and SBIR.

In Information Retrieval, **Precision** and **Recall** are the basic measures used in evaluating search strategies. Recall is defined as the number of relevant documents retrieved

divided by the total number of existing relevant documents and Precision is defined as the number of relevant documents retrieved divided by the total number of documents retrieved by that search. Table 2 shows the confusion matrix.

TABLE 2
SYSTEMATIC AND TRANDITIONAL NOTATIONS OF CONFUSION MATRIX.

	Relevant	**Not Relevant**
Retrieved	*TP*	*FP*
Not Retrieved	*FN*	*TN*

TP=True Positive (Correct Result)
FN=False Negative (Missing Result)
FP=False Positive (Unexpected Result)
TN=True Negative (Correct absence of Result)

The values obtained by the two algorithms BIR and SBIR are entered in the confusion matrix for different keywords and the precision and recall values are calculated by using the formula given below.

$$\text{Recall} = TP / (TP + FN)$$
$$\text{Precision} = TP / (TP + FP)$$

The Table 3 shows the precision and Recall of the two algorithms.

TABLE 3 Precision and Recall of BIR and SBIR

	Precision		**Recall**	
Query	***BIR***	***SBIR***	***BIR***	***SBIR***
Rejoice	0.803922	0.967742	0.041582	0.912779
faithfully	0.777778	0.957854	0.013109	0.93633
Oath	0.909091	0.923077	0.367647	0.882353
persecution	0.625000	0.750000	0.123457	0.740741
sincerely	1.000000	1.000000	0.02439	0.813008
crucification	0.857143	0.874317	0.063158	0.842105
Hell	0.888889	0.909091	0.666667	0.833333
Jesus	0.893617	0.958188	0.655226	0.858034

The performance of BIR and SBIR is represented by precision and recall graph. An algorithm is better than another if it achieves better precision and recall. Fig. 3 and Fig. 4 indicate that SBIR is more effective than BIR.

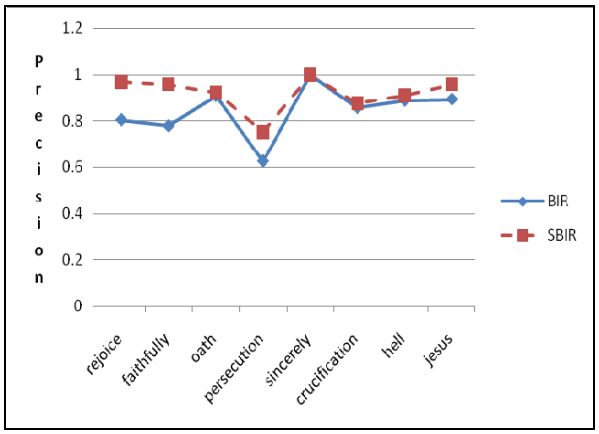

Figure 3 Precision Graph

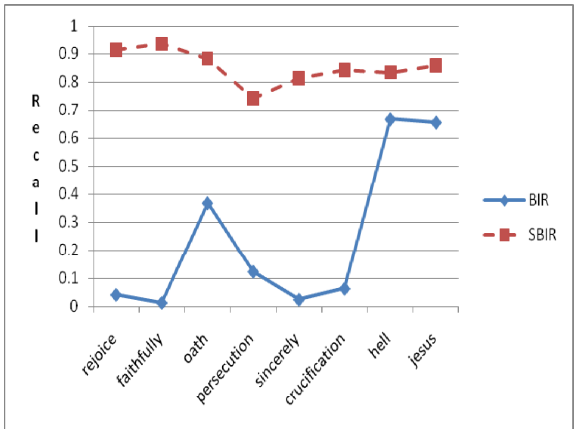

Figure 4 Recall Graph

V. CONCLUSION

In this paper a new algorithm called SBIR is proposed to enhance the performance of Boolean Information Model by improving the recall and precision. The use of stemming algorithm can give the root word to the *synsets* that has been retrieved from WordNet. The number of documents retrieved after getting the synsets is more than the documents retrieved by using the keyword alone. And the number of documents retrieved after stemming process of each synset is further increased. The precision and recall values are calculated and the results reveal that the efficiency of SBIR is due to the contribution of stemmed synonym terms. But still some problems have been identified in the above algorithm. For example, the document may not contain all the words in synsets. And sometimes the user may want to select the words from the synsets to perform the search process. SBIR can also be extended to work with compound terms.

In future, some other algorithms will be proposed to solve these problems and still increase the performance of Information Retrieval.

REFERENCES

[1] R. Thamarai Selvi, E. George Dharma Prakash Raj, "Information Retrieval Models: A Survey" International Journal of Research and Reviews in Information Sciences (IJRRIS) Vol. 2, No. 3, UK, September 2012, ISSN: 2046-6439.

[2] R.B.-Yates and B.R.-Neto. Modern Information Retrieval. Addison Wesley Longman, 1999.

[3] Salton, G., McGill,M., Introduction to Modern Information Retrieval, McGraw-Hill, New-York, 1983.

[4] H. Turtle. Natural language vs. Boolean query evaluation: a comparison of retrieval performance. In SIGIR, 1994.

[5] X. Xue and W. B. Croft. Transforming patents into prior-art queries. In SIGIR, 2009.

[6] X. Xue and W. B. Croft. Automatic query generation for patent search. In CIKM '09, 2009.

[7] Youngho Kim yhkim, Jangwon Seo, W. Bruce Croft Automatic Boolean Query Suggestion for Professional Search SIGIR'11, July 24–28, 2011, Beijing, China.

[8] Fellbaum, C. WordNet. Theory and Applications of Ontology: Computer Applications, 231, PP: 231-243, Springer Science Business Media B.V, 2010.

[9] Giridhar N S, Assistant Professor, Prema K.V, Professor, N .V Subba Reddy, A Prospective Study of Stemming Algorithms for Web Text Mining, GANPAT UNIVERSITY JOURNAL OF ENGINEERING & TECHNOLOGY, VOL.-1, ISSUE-1, JAN-JUN-2011.

[10] Manuel, D., Maria, M., Alfonso, U. L., & Jose, P. 2010. Using WordNet in Multimedia Information Retrieval. CLEF 2009 Workshop, Part II, LNCS 6242, pp. 185–188, Springer-Verlag Berlin Heidelberg.

[11] Eneko, A., Xabier, A. & Arantxa, O. Document Expansion Based on WordNet for Robust IR. COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics, Volume, pages 9–17, Beijing. ACM, 2010.

[12] Deepika Sharma, "Stemming Algorithms: A Comparative Study and their Analysis", International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 4– No.3, 2012.