ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

Enhancing the Performance of Information Retrival System using ESBIR Algorithm

Akash D. Waghmare [1], D. D. Puri [2], N. Y. Suryawanshi [3] , P. D. Sharma [4]

Assistant Professor, Department of Computer Engineering, SSBT COET, Bambhori Jalgaon, Maharashtra, India[1]

Assistant Professor, Department of Computer Engineering, SSBT COET, Bambhori Jalgaon, Maharashtra, India[2]

Assistant Professor, Department of Computer Engineering, SSBT COET, Bambhori Jalgaon, Maharashtra, India[3]

Assistant Professor, Department of Computer Engineering, SSBT COET, Bambhori Jalgaon, Maharashtra, India[4]

ABSTRACT: Information Retrieval is the process of obtaining information relevant to an information need from a collection of information resources. Documents that satisfy the given query in the judgment of the user are said to be relevant otherwise non-relevant. An IR engine uses the query to classify the documents in a given collection and returns a subset of documents. Sometimes the relevant documents may not contain the keywords specified in the query. The absence of the given term in a document does not necessarily mean that the document is not relevant because more than one term can be semantically similar although they are lexicographically different. In this paper a noval appraoch, Extended Semantic Based Boolean Information Retrieval using Synonyms and Antonyms (ESBIR), is proposed to retrieve the documents with semantically similar terms to enhance the performance of Information Retrieval System by improving the recall and precision.

KEYWORDS: Information Retrieval, Semantic, Synonyms, Antonyms, Precision, Recall.

I. INTRODUCTION

The system of Information Retrieval (IR) deals with storing, maintaining and searching of information. The information that is retrieved should be relevant to the need of the user. Information Retrieval can be defined as the activity of obtaining information or data relevant to an information needs from a collection of information resources [1]. IR provides the required information according to the user query within a collection of data. The Corpora that are available on the internet and information stored in the databases of libraries and companies are huge in recent years. This leads to the difficulty in searching of information needed from the vast data that exists. The need to efficiently organize, search and manage textual corpora for knowledge has brought a new interest in the process of information retrieval (IR) and data mining strategies.

In IR system, the two important factors that directly affect the efficiency of the retrieval results is the approach to represent text and the measure to evaluate the similarity between query and documents. The main goal 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. Sometimes the documents may contain the opposite words of the given words preceded by the words like ‘not’, ‘im’, ‘un’, ‘ir’ etc. For example, the word ‘good’ has the meaning ‘not bad’. So, the document contains the word ‘not bad’ is also a relevant document for the user. Retrieval, by classical information retrieval models (e.g., Vector Space, Probabilistic, Boolean) [2] are 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 Word Net database, antonyms and stemming algorithm Word Net is an on-line lexical reference system.

The paper is organized as follows. Section 2 describes previous work in Semantic Based Boolean information model using Word Net. The task of proposed work, its framework and algorithm along with the steps used defines in section 3. Section 4 contains the discussion. And finally, the conclusion in Section 5.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

II. RELATED WORK

The significance of Boolean Information Retrieval (BIR) has been revealed in many retrieval systems because of its simplicity [3]. Most of the commercial IR systems use this Boolean model to predict that each document is either relevant or not relevant [4]. 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 [5,6,7]. Boolean queries [8] however, are easy for information professionals to manipulate and are essentially self-documenting in that they define precisely the set of documents that are retrieved. The semantic retrieval [9] approach is used to discover semantically similar terms using Word Net. In many works, Word Net 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 [10] is a context sensitive suffix removal algorithm. Removing suffixes by is an operation which is especially useful in the field of IR. Word Net expansion technique was used [10] over a collection with minimal textual information. Philip resniket. al. [10] annotated the Biblical text to create the aligned corpus such as Bible for linguistic research which also includes the automatic creation and evaluation of translation lexicons and similar tagged text. It has the feature of parallel translations over huge number of languages. It also represents the comparison with dictionary and corpus resources for modern English. Thus, it makes the Bible a multilingual corpus which considered to be a unique resource for linguistic research. R. Thamarai Selvi et. al. [11]&[1] proposed an algorithm based on Boolean Information Retrieval (BIR) that has the significance of its simplicity. In this proposal, Boolean model is used to predict whether each document is relevant or not. It refers an online lexical reference called Word Net to find the semantically similar terms. Stemming is the process which is used for reducing inflected words to their stem, root or base form.

III. PROPOSED WORK

Sometimes, the document may contain the opposite words of the given words in the query, preceded by the words like ‘not’, ‘im’, ‘un’, ‘ir’ etc. For example, the word ‘irrelevant’ has the meaning ‘not relevant’. Hence, the document contains the word ‘not relevant’ is also a relevant document for the user. In this proposed algorithm called “Extended Semantic Based Information Retrieval Algorithm” information retrieval is done in an effective way by increasing the number of relevant documents, improving the precision and recall, and reducing the time taken for retrieving information. Fig. 1 shows the complete architecture of the proposed algorithm.

3.1 ESBIR Framework

In information retrieval, the user formulates any number of arbitrary queries and applies them to a fixed collection. The task of information retrieval, finding documents within in a corpus that are relevant to an information need specified using a query. The framework for ESBIR is given below in the Figure 1.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

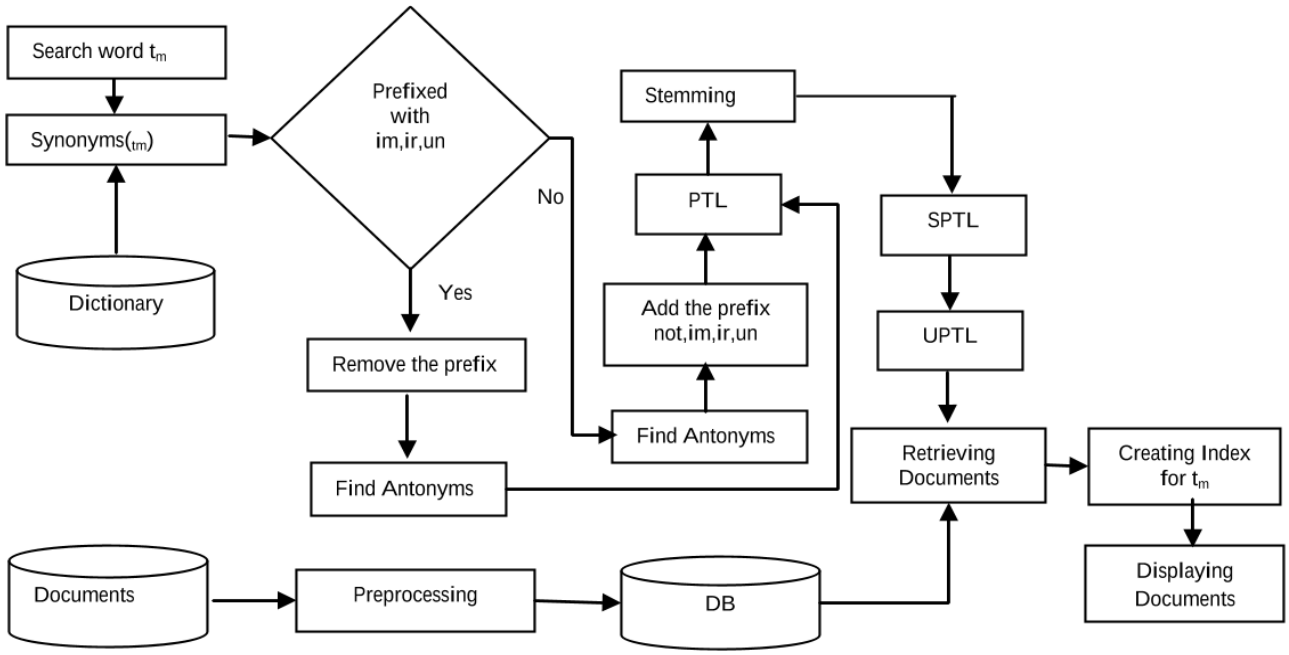
International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

Figure 1



3.2 ESBIR Algorithm

ESBIR retrieves documents from the document set by finding synonyms and antonyms of the given term using Word Net data base to find more similar documents. The proposed ESBIR algorithm is given below.

Input : Documents corpus
Output : Relevant documents
Algorithm:
Step 1: Preprocessing of the documents
Step 2:Enter the query term tm
Find the semantically similar terms si and assign in T
Step 3: For each si in T
If si is prefixed with “im” or “ir” or “un” Remove the above said prefix from si
Find the antonyms of si and assign to PTL
else
Find the antonyms of si , add the suitable prefix “not” or “im” or “ir” or “un” and assign to PTL.
Step 4:User selects the words from PTL and assign to UPTL
Step 5:Find the stemmed word sti
for each si, in UPTL and assign to SPTL
Step 6:for each st i in UPTL
find the document d and put them in Di
create index for the term
Step 7:for each dj in Di
display the documents dj

IV. EVALUATION OF ESBIR AND SBIR

The number of documents retrieved by ESBIR is larger than the documents retrieved by using the SBIR algorithm. Table 1 shows the number of documents that are retrieved by using UFSBIR and ESBIR.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 6, Special Issue 1, January 2017

International Conference on Recent Trends in Engineering and Science (ICRTES 2017)

20th-21st January 2017

Organized by

Research Development Cell, Government College of Engineering, Jalagon (M. S), India

Table1: Number of Documents Retrieved

Query	UFSBIR	ESBIR
bad	604	629
impossible	184	193
uncover	327	332
close	210	223
accept	161	169
reject	108	149
weak	79	89
forget	74	82

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.

V. CONCLUSION

ESBIR is proposed to enhance the performance of Semantic Based Boolean Information Model. Since the proposed algorithm considers the synonyms as well as the antonyms of the search word, it creates the possibility of retrieving more number of documents. The precision and recall values that can be calculated which confirm the improved performance of ESBIR over SBIR.

REFERENCES

[1] R. Thamarai Selvi, E. George Dharma Prakash Raj, "An Approach to Improve Precision and Recall for Ad-hoc Information Retrieval Using SBIR Algorithm" IEEE Conference, 27 March 2014 ISSN: 978-1-4799-2876-7
[2] 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, September 2012, ISSN: 2046-6439.
[3] R.B.-Yates and B.R.-Neto. Modern Information Retrieval. Addison Wesley Longman, 1999.
[4] Salton, G., McGill,M., Introduction to Modern Information Retrieval, McGraw-Hill, New-York, 1983.
[5] H. Turtle. Natural language vs. Boolean query evaluation: a comparison of retrieval performance. In SIGIR, 1994.
[6] X. Xue and W. B. Croft. Transforming patents into prior-art queries. In SIGIR, 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, 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] R. Thamarai Selvi, Dr. E. George Dharma Prakash Raj. "UFSBIR:A Semantic based Boolean Information Retrieval Algorithm with User Feedback", International Journal of Information Systems, Vol.I, 2014, pp.36-40.