Extended Semantic based Boolean Information Retrieval Algorithm for User-driven Query

Vachhani Upama
ME Student,
Department of Computer Engineering,
GEC, Sector-28,
Gandhinagar, India.

S. M. Shah
Associate Professor,
Department of Computer Engineering,
GEC, Sector-28,
Gandhinagar, India.

***Abstract* - Information Retrieval (IR) is essentially a matter of deciding which documents within a large collection satisfies a user's information need. Those documents are called relevant documents and the documents that are not of the topic specified by the user are said to be non-relevant. An existing SBIR algorithm uses lexical database, WordNet to find synonyms of single-word query term considering that the absence of the given term in a document does not necessarily mean that the document is not a relevant.In this paper, a new algorithm is proposed which works with compound terms and uses modified Porter Stemming Algorithm to solve some stemming errors found in Porter Stemmer Algorithm proposed by M. F. Porter. This will improve the recall as more relevant documents will be retrieved. We propose to involve a user in the search process through interactive feedback for word senses. This will further improve recall by retrieving more user relevant results.**

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

I. INTRODUCTION

The plentiful information stored in online databases can be highly advantageous for both people and automated computer systems that seek information, if it can be retrieved efficiently. Information Retrieval is a procedure of finding the documents in a corpus based on a specific query.The main idea is to locate documents that contain the terms that the users specify in their queries.

A. Boolean Retrieval Model

Most of the classicalinformation retrieval models retrieve the document based on lexicographic term matching only.Boolean retrieval model, based on set-theory, is the very first retrieval model proposed three decades ago. It is kind of exact match model. If the exact term exists in document, then only the document is retrieved otherwise not. In professional search environments such as legalsearch or patent search, users are expecting many retrievalresults, i.e., there is ordinarily an emphasis on recall.Recent surveys have also checked thatprofessional searchers keep on having a solid inclination forBoolean queries because of the precise nature of Boolean model.Popular medical databases like MedLine and PubMed which allows to search for articles on biology and medicines and legal database like Westlaw which allows to search for legal documents are based on Boolean retrieval model. Even many search engines also use this information retrieval model.

B. Query Expansion

Query expansion is an effective way of enhancing performance of information retrieval systems. The basic process is that select new terms which are based on the initial query, and then combine both of them to form a new query [3]. It is more efficient to users for simpler search tasks whereas interactive query expansion is more productive for more complex search tasks. Irrespective of the method used, the key point is to get the best words that are used to expand the query.The aim for query expansion is to reduce the mismatch between query and documents by expanding the query terms using words or phrases which are synonymous to query terms. This has an impact on the recall of most information retrieval systems. Despite the significance of Boolean queries in professional search, there has not been much research on assisting information professionals in expanding search query.

C. Stemming

Generally in IR applications, stemming is done before index is created. Stemming algorithm is a procedure of linguistic normalization, in which the variant forms of a word are reduced to a common form, for example, (operates, operation, operatives, operational) -> oper. The terms extracted from documents are stemmed using some stemming algorithm. The purpose of this step is to remove various suffixes, to reduce number of distinct words, to have exactly matching stems, to save memory space andtime. It is vital to admire that we utilize stemmingwith the expectation of enhancing the performance of IR systems.

D. Inverted Index

Finding information is not the only action that exists in an Information Retrieval (IR) system. Indexing, for instance, refers to how information in the system is represented. The documents are represented through a set of index terms or keywords. These terms are extracted from the text of the documents.

Inverted index is the standard method for supportingqueries on large text databases; there are no practical alternativesto inverted indexes that facilitate with sufficiently fast query evaluation. Apositional inverted index is a two-levelstructure.

The upper level contains all the index terms.For text databases, the index terms are all thedistinct words occurring in the text. Thelower level is a set of postings lists, one per index term.Each posting is a triple of the form:

$$<d,f_{[d.t]}\,,[o_1,......,o_{f[d,t]}]>$$

where d is the identifier of a document containing term t, thefrequency of t in d is $f[d,t]$, and the o values are the positions in d at which t is observed.

It is straightforward to use an inverted index to evaluate a phrase query.Consider the query "run away". Of these terms, "run" is the rarest, andits inverted list is fetched first. The postings are decodedand a temporary structure is created, recording which documentscontain this word and the ordinal word positions ineach document at which it occurs. The remaining term "away"isnext rarest term in input, and is the one to be processed next.For each documentidentifier and word offset in the temporary structure createdearlier, a posting is sought to see whether "away' is in the document one word later.If the search fails for a position in a posting, thatword position is discarded from the temporary structure.The remaining entriesgive documents and word positions at which the phraseoccurs.

II. Related work

The Boolean Information Retrieval (BIR) is still popular in many retrieval systems because of its simplicity and precise nature [9]. The majority of commercial IR systems use Boolean model to predict if the document isrelevant or not. Also, professional searchers keep on using Boolean queries.

Stemming is performed by IR systems on the terms extracted from database in order to remove common suffixes and thus to reduce the words to their base forms. The most common algorithm for stemming English, and one that has been repeatedly shown to be empirically very effective, is Porter's algorithm (Porter, 1980) [4].However, it is still having several limitations, although many attempts have been made to improve its structure.[5] uncovers the inaccuracies that are encountered during the stemming process and the corresponding solutions are also proposed.

WordNet [7] is a large manually constructedcomprehensive thesaurus developed at Princeton University.In WordNet,the basic unit is the Synset, which represents alexicalized concept. Synset consists of nouns, verbs,adjectives, and adverbs. Many IR systems [10] have been developed which uses WordNet for Query Expansion considering the fact that lack of query term in document does not mean that document is irrelevant. If any synonym of the query exists in document then also document is said to be related to the search query. Following figure shows the framework of IR system using WordNet for Query Expansion.

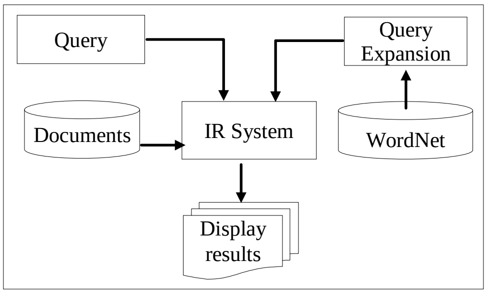

Fig. 1: Query Expansion using WordNet

III. Proposed work

This work developsa system for retrieving information from a static data set and aims to provide documents from within the corpus that are relevant to user's information need. An information need is the topic about which the user desires to know more. [1] retrieves documents from the document set by first finding synonyms of the input query term using WordNet data base and then retrieves the documents for each extracted synonym. The algorithm specified doesn't work for compound word query.Due to which, for some single word query, if one or more synonyms consists of list of words (or phrase), they will be ignored as compound terms cannot be processed by the specified algorithm. For example, if search query given is "escape", one of its synonyms returned by WordNet is "run away". But since SBIR algorithm cannot work for phrases, documents matching "run away" will not be retrieved though they are relevant to the search query. This results in decrease in recall for such words having synonyms containing compound terms or phrases.

User's query, often just a few words, is usually not accurate and clear enough to describe the information needs [8]. The user often knows only vaguely what he/she wants but does not know exactly what he/she is looking for until he/she has seen it.The joint interaction between the system (which shows the user some terms to modify the original query) and the user (whom, after entering the query, assesses the possibility of expanding it, using the system's suggestions) is very effective in retrieving user relevant results. Thus adding interactive synonym feedback to the existing algorithm can improve the recall values compared to SBIR algorithm giving better retrieval results from user's point of view.

A. Modified Porter Stemmer Algorithm

Porter Stemmer Algorithm by M. F. Porter[4] is used [1], but this algorithm has some inaccuracies in certain cases. Example, word "animal" and "animate" are both stemmed to "anim". And if two or more words have same stem it means they have similar meaning and they can be grouped while indexing. But these words "animal" and "animate" are no-where related then also they have same stems. So to remove over stemming errors, we refine the existing stemming algorithm with certain constraints implemented from Improved Porters Algorithm [5]. Also, one more under-stemming error has been found.

Error:
Porter Stemmer algorithm stems

“improve” -> “improv”
“improvise” -> “improvise”

These two words are semantically related then also they have different stems. Hence, though having similar meaning, when searched for ‘improvise’ will not retrieve documents having ‘improve’ and vice versa. So, this stemming error reduces recall. Hence, a new rule has been added for words ending with “ise” to solve the above under-stemming error.

Solution: The following rule has to be added to step-5 in Porter Stemmer Algorithm.

```
case “s”:
if(ends(“ise”)) break;
return;
```

B. *Extended Semantic based Boolean Information Retrieval Algorithm Framework*

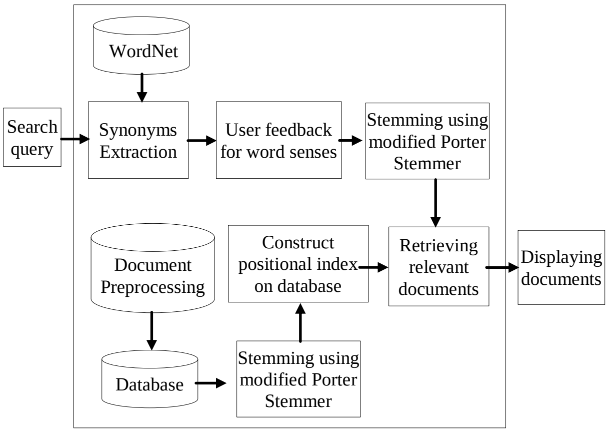
WordNet
Search query
Synonyms Extraction
User feedback for word senses
Stemming using modified Porter Stemmer
Document Preprocessing
Construct positional index on database
Retrieving relevant documents
Displaying documents
Database
Stemming using modified Porter Stemmer

Fig. 2: Framework of Proposed Algorithm

Step 1: Preprocessing

In the very first step, the chapter name, chapter number, verse number and verses are extracted from the documents. This information is stored in mysql database. This step performs the extraction of individual documents from thedocument set and are then stored in the database. And also thekeywords are extracted from the document by eliminatingthe stop words (frequently occurring words) and are stored in the database. This process of separating each term in the documents is called tokenization.

Step 2: Stemming and Creating Positional Inverted Index

After extracting the distinct terms from the database in first step, in this step all those terms are stemmed using modified Porter Stemmer algorithm which solves some of the errors shown by [4]. The idea is to reduce the total number of distinct terms which in turn results in decrease in size of index created on the database and time consumed for creating index will be less. Thus, now positional inverted index is created on the database. It stores the document ids for all documents which contain the term and positions where the term occurs in all those documents.

Step 3: Extracting synsets from WordNet

In this step, synsets are retrieved from theWordNet database for the given input query. The basic unit in WordNet is the Synset, which represents alexicalized concept. Closely related synonyms are gathered into one synset. Synset consists of nouns, adjectives,verbs and adverbs.These synsets are stored in a data structure array.

Step 4: User feedback

Lack of input query word in a documents does not mean that document is irrelevant because it may contain synonym of the input query. But instead of extracting documents matching all the synonyms in all the synsets of the given query, user will be involved in the search process. User will select the synsets which are closely related to his information need.

Step 5: Stemming

The stemming process is done on thesynsets got from previous step. The modified Porter StemmerAlgorithm implemented in java performs this process foreach word from the synsets. These stemmed words will be used toextract the documents from the data base.

Step 6: Retrieving Documents from the database

In this last step, the relevant documents are retrievedfrom the database for each stemmed word/phrase in the array and whole result will be displayed to the user.

IV. EXPERIMENT

The proposed algorithm which improves the performance of Boolean Information Retrieval system is implemented using Javaprogramming language. Extended SBIR is tested on Bible text This collection consists 66 Books, 1189 Chapters, **31,102** Verses and 7,882,80 words. These verses are converted into individual documents and stored inthe MySql database.Following table shows the results by SBIR [1] and after extending SBIR to work with compound terms by implementing positional inverted index:

Query	No of docs retrieved	
	SBIR	Extended SBIR
withdraw	1115	1135
light	892	893
compensate	707	729
escape	212	329
really	288	331
offering	892	898

V.CONCLUSION

“Semantic Based Boolean Information Retrieval(SBIR)” a recently proposed algorithm does query expansion using WordNet before retrieving documents using conventional Boolean model.This algorithm gives higher precision and recall values than original Boolean retrieval model. But still there are chances to improve the performance of algorithm by improving Porter Stemmer algorithm it uses to avoid stemming errors that are shown in certain cases.Further, instead of retrieving documents containing all those terms in all synsets, sometimes the user may want to select the words from the synsets to perform the search process. Hence, presenting word senses returned by WordNet to user can help user to get specific results needed. Thus, Extended SBIR algorithm is proposed which will achieve higher recall and precision as it solves certain stemming errors found in original Porter Stemmer algorithm and implements positional inverted index to work with compound terms.

REFERENCES

(1) Selvi R. Thamarai and E. Raj. "An Approach to Improve Precision and Recall for Ad-hoc Information Retrieval Using SBIR Algorithm." In Computing and Communication Technologies (WCCCT), 2014 World Congress on, pp. 137-141. IEEE, 2014.

(2) Liu, Shuang, et al. "An effective approach to document retrieval via utilizing WordNet and recognizing phrases." Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004.

(3) Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. Vol. 1. Cambridge: Cambridge university press, 2008.

(4) M. Porter. An algorithm for suffix stripping. Program, 14(3):130–137, 1980.

(5) Yamout, Fadi, et al. "Further Enhancement to the Porter’s Stemming Algorithm." Ulm, September 21, 2004 (2004): 7.

(6) Ramasubramanian, C., and R. Ramya. "Effective Pre-Processing Activities in Text Mining using Improved Porter’s Stemming Algorithm." International Journal of Advanced Research in Computer and Communication Engineering 2.12 (2013).

(7) George A. Miller, Richard Beckwith, Christiane Fellbaum, DerekGross, and Katherine Miller. (1993): ‘Introduction to WordNet: AnOn-line Lexical Database’ Revised Version 1993

(8) Kotov, Alexander, and ChengXiang Zhai. "Interactive sense feedback for difficult queries." Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011.

(9) Kim, Youngho, Jangwon Seo, and W. Bruce Croft. "Automatic boolean query suggestion for professional search." Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011.

(10) Barman, Anup Kumar, Jumi Sarmah, and Shikhar Kr Sarma. "WordNet Based Information Retrieval System for Assamese." Computer Modelling and Simulation (UKSim), 2013 UKSim 15th International Conference on. IEEE, 2013.