IJCSN International Journal of Computer Science and Network, Volume 5, Issue 2, April 2016
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 1.02

An Approach for IR using Extraction and Expansion of Micropost

[1] Priya Mundada, [2] Dr. Manoj Chandak

[1] M.Tech, Computer Science and Engineering, RCOEM Nagpur,440013, Maharashtra, India.

[2] H.O.D., Computer Science and Engineering, RCOEM Nagpur,440013, Maharashtra, India.

Abstract - Query expansion is the process of supplementing additional terms or phrases to the original query to improve the retrieval performance. Nowadays retrieval performance of a search engine plays a vital role. Along with the retrieval performance precise information is also required. It is very difficult to get the precise information which is actually required by the user through Micro post or short comment. Micro post is a form of short comment which people generally give on social networking site to interact with their friends and share the information. The information is written using least number of words. Since there is less number of keywords to retrieve the information we need to expand the micro post this is known as query expansion. It has been suggested as an effective way to resolve the short query and word disambiguation problems. This query expansion helps to retrieve the precise information from the large data. After expanding the micro post we can understand the actual sense of users' query. The proposed system will expand the micro post which will help in specific Information Retrieval.

Keywords - *Query Expansion, Micro Post, Lingo Words, Information Retrieved.*

1. Introduction

Query expansion (QE) is the process of reformulating the query to enhance retrieval performance in information retrieval operations. Query expansion involves evaluating a user's input i.e. what words were typed into the search query area and sometimes other types of data and expanding the search query to match additional documents. Today's world is the era of social networking. People regularly interact with each other through messaging on social networking websites. They do not intend to put lots of efforts in typing the whole matter. To save time they get habituated to write the message in a very short and precise manner. This same scenario happens when they try to put there query to search engine. Hence, it is generally observed that web users put very short query to search engines. The above scenario can create lots of complications when it comes to search engine. It becomes difficult for the search engine to optimize it in an efficient way. Micro post is a very short piece of information which is generally used to share the message amongst friends. This micro post contains the lingo words i.e. the short form of words which is generally used in text messaging to save the time. But these are not the dictionary words which can be used for the Information Retrieval. Search engine is not aware with the lingos which user inputs. Moreover, these lingos differ from person to person. Sometimes the grammar of the sentence is also not in the correct format which may decrease the efficiency of the retrieved information. While writing text messages user is least bothered about the grammar in which it is typing.

There are some situations in which user has not provided sufficient amount of words which may lead to misconception to retrieve the information. Micro post contains very few words as explained above. Hence we need to expand the micro post which is given by the user i.e. Query Expansion. In Query Expansion the lingo words are replaced with the original dictionary words which will help in retrieving the precise information. Also the grammar mistakes will be corrected if any. The micro post will be embedded with the required words which will transform the micro post into the expanded query.

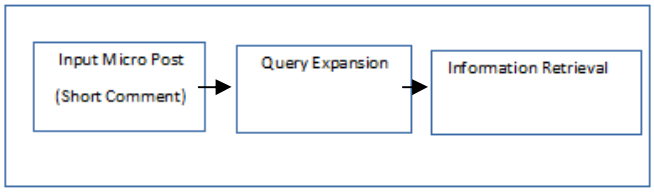

Fig 01: Block Diagram

In this paper a new method of query expansion is suggested which will co-relate the user's micro post and the related retrieved document. This will increase the efficiency of the Information Retrieved. Section II discusses about the literature survey i.e. the existence technologies for query expansion. This helps to expand the idea of the proposed system in terms of using new techniques for the proposed methodology. It also put light on the flaws in the existence technologies and provides some idea to overcome the limitations. Section III is the proposed system. It discusses all the phases used in the proposed along with the approach used in it. Section IV provides information regarding the dataset used. Section V tells the advantages of the proposed system in each phase. Section VI is the conclusion.

2. Literature Survey

The very first technique for query expansion is Global analysis. It produced consistent and effective improvements through query expansion. The earliest global analysis technique is Term Clustering [15] which creates cluster of document term based on their co-occurrences. Queries are expanded by the terms in the same cluster. Other well-known global techniques include Latent Semantic Indexing [4], similarity thesauri [11], and Phrase Finder [8]. Moreover, since it only focuses on the document side and does not take into account the query side, global analysis cannot address the term mismatch problem well.

On the other hand, local analysis uses only some initially retrieved documents for further query expansion. One of the popular local analysis techniques is relevance feedback [12, 14], which modifies a query based on user's relevance judgments of the retrieved documents. Expansion terms are extracted from the relevant documents. If user provides proper and accurate relevant judgment then this technique works effectively. But in general scenario user does not provide the proper relevant judgment. Local feedback mimics relevance feedback by assuming the top-ranked documents to be relevant [3, 13].

Some of the other techniques are re-ranking the retrieved documents using automatically constructed fuzzy Boolean filters [10], clustering the top-ranked documents and removing the singleton clusters [9], clustering the retrieved documents and using the terms that best match the original query for expansion [2]. In addition, recent TREC results show that local feedback approaches are effective and, in some cases, outperform global analysis techniques [17].
Expansion terms are extracted from the top-ranked documents to formulate a new query. In recent years, clustering the top-ranked documents and removing the singleton clusters [9], clustering the retrieved documents and using the terms that best match the original query for expansion [2] techniques are used. Retrieving top-ranked document has its own drawback. If the retrieved bulk of o-ranked document is irrelevant then the words added to the query are likely to be unrelated to the search topic. This will result to the worse query expansion.

There are some of the techniques which talks about expanding the query using the Thesaurus [19]. The approach is to use the synonyms and the linguistics from the thesaurus. It also helps in general Natural Language Processing having the similarity. Well constructed thesaurus has been maintained to get the effective results. Similarity measures such as dice coefficient, Jacquard coefficient, use of semantic network and ontology, use of soft computing methods like Genetic Algorithm and Neural Network has been used for thesaurus construction.
AQE(Automatic Query Expansion) is a technique which is very strong in retrieving and ranking the documents[20]. It deals with various aspects regarding the needs and effectiveness of AQE. It also put light on various application of AQE such as Question Answering System, Multimedia Information Retrieval, Information Filtering and Cross Language Information Retrieval.

3. Proposed System

A new approach has been used for Query Expansion. Very first phase of our proposed system is the Dataset cleaning. We are in need to clean our dataset for retrieving the short comment. After cleaning the data we will be keeping this data for further use. In the proposed system the input will be a micro post i.e. a short comment (provided by the user). This short comment will contain the lingo words. These lingo words will be replaced with the original dictionary words. The lingos differ from individual to individual.

For this purpose own dictionary has been created. This dictionary is maintained and can be expanded as on requirement. The words are added after testing the lingos of different personalities. After replacing the lingo words the grammar of the micro post will be corrected. For the grammar correction different sentence pattern is used. These grammar patterns are made on the Assertive or Declarative type of sentences. The patterns are based on the eight parts of speech. The place of a word is decided on the basis of the sense of that word i.e. to which part of speech it belongs. After the grammar correction the required words are added in the query by using n-gram algorithm. We will be using the data which we have already cleaned in the first step. This cleaned data consists of shorts comment. By referring this clean data we will place the words which will fit our query. There may be the

case in which the grams are not added because the query may be sufficient for the optimization. Working on single word can change the meaning and sense of retrieved information.

Hence, once the query is expanded phrases (Noun Phrase and Verb Phrase) will be extracted from this query. To extract the phrases Parsing technique will be used.

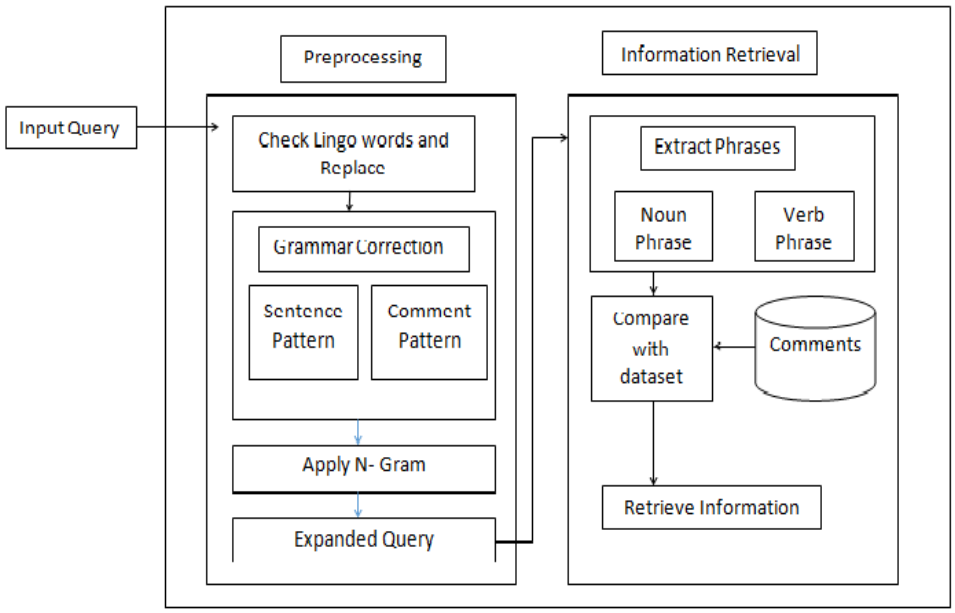

Fig 2: Proposed System Framework

After parsing the expanded query we will get the parsed tree. Through this parsed tree we have to extract the required phrase. Then this phrase will be compared with the dataset. The best matched will be retrieved and it will be categorizes. This categorization will tell whether the query is positive or negative.

4. Module Division

4.1 Lingo word checker

For this module we have created our own lingo word dictionary. From this dictionary the lingos in input query will be replaced with its equivalent meaningful word.
Eg. “Gr8 htl” will be replaced by “Great hotel”.

The pseudo code for this module is as follows:

Start
For each word in a query {
For each entry in dictionary {
Compare input word with dictionary word;
If(input word==word in dictionary){
Replace input word with its meaningful equivalent word from dictionary;
}
}
}

4.2 Grammar Correction

Here we correct the input query grammatically. For this purpose we have used Sentence Pattern for Assertive or Declarative sentence and Comment Pattern for input comments. Sentence Pattern are the rules defined for the Assertive or Declarative sentence construction. Comment Patterns are the rules defined for the comments which excludes subject.

The pseudo code for this module is as follows:
Start
Get input query;
Check which set of rules the input query follows.
If (input query==sentence pattern) {
Formulate according to Sentence Pattern;
}
Else {
Formulate according to Comment Pattern;
}

4.3 N-Gram

After grammar correction we apply N-gram algorithm to our cleaned data of short comment for extracting the key phrases. As on requirement we will populate the query with the retrieved phrases. This is the actual Query Expansion phase.

4.4 Information Retrieval

This is the final module of the proposed system. Here the phrase will be extracted from the input query as well after Query Expansion. This phrase will be compared with the each document of the dataset. Whichever document will give the maximum similarity will be retrieved.

The pseudo code for this module is as follows:

Start
Get input query phrase;
For each document in the dataset{
Compare the phrase;
If(input phrase==document phrase){
Increase the count;
Get the document with the maximum count;
}
}

This is how the information will be retrieved.

5. Dataset Description

The dataset of hotel review is used for extracting the actual micro post. This dataset contains the reviews of hotels according to the city. There are 10 cities in the dataset. Each city contains the review of hotel in that particular city in a text file. Each hotel review is maintained in one text file. Hence we have several files of hotel review for each city. Each text file is associates with number of reviews. Each review is placed in new line. A single review is separated in 3 parts by a tab i.e. Date, Short Comment and Long Comment. More than 2,00,000 micro posts are there in this dataset. Total size of the Dataset is 480.9 MB. This is a huge dataset containing large number of reviews. It can be used in multiple applications. Even the dataset cleaning process used in our proposed system which separates the large comments and short comments can be proven useful for many other Information Retrieval applications.

6. Advantages of Proposed System

The proposed system is designed in such a way that it will retrieve the information according to the users' requirement. It has the mechanism of lingo word correction which will increase the accuracy aspect of the retrieved information. The grammar correction phase corrects the grammar of the inputted query which enhances the query in word sequence format. For this grammar correction we have used sentence patterns for assertive or declarative sentence which helps to correct the statement word sequence according to parts of speech. Then next is the comment pattern which helps to place the adjectives and adverbs to the correct place. N-gram will help to suggest the missed words from the query. This will again help in expanding the query to enhance the performance. Then the phrase extraction technique is used to retrieved the information which will be proven advantageous in the context of retrieving the appropriate information.

7. Conclusion

In today's world people are connected with each other through social networking media. Hence they are more addicted to less typing and least bothered about the grammar. But complications occur when it comes to search engine. It is very difficult to optimize this kind of user query which is very short. Hence the proposed system describes and tries to overcome the above discussed problem regarding the query optimization. Query Expansion is a proposed technique to get the actual sense of the asked query and then retrieve the information. Query expansion will help to retrieve the precise information from the micro post. This expanded query will help the user to get accurate information.

References

[1] Attar, R. and Fraenkel, A.S. 1977. Local feedback in full-text retrieval systems. J. ACM 24, 3 (July), 397-417.

[2] Buckley, C., Mitra, M., Walz, J. and Cardie, C. 1998. Using clustering and superconcepts within SMART. Proceedings of the 6th text retrieval conference (TREC-6), E. Voorhees, Ed.107-124.NIST Special Publication 500-240.

[3] Buckley, C., Salton, G., Allan, J., and Singhal, A., 1995, Automatic query expansion using SMART, TREC 3.Overview of the Third Text Retrieval Conference (TREC-3),pages 69--80. NIST, November 1994. http://trec.nist.gov/.

[4] Deerwester, S., Dumai, S.T., Furnas, G.W., Landauer, T.K.and Harshman, R. 1990. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. 41, 6, Pages 391-407.

[5] Direct Hit website. http://www.directhit.com/.

[6] Furnas, G.W., Landauer, T.K., Gomez, L.M. and Dumais,S.T. 1987. The vocabulary problem in human-system communication. Commun. ACM 30, 11 (Nov. 1987), Pages964-971.

[7] Hull, D., 1993, Using statistical testing in the evaluation of retrieval experiments. In Proceedings of the ACM SIGIR,pages 329--338, Pittsburgh, PA, June 1993.

[8] Jing, Y., Croft, W.B., 1994, An association thesaurus for information retrieval, in Proceedings of RIAO 94, 1994, pp.146-160.

[9] Lu, A., Ayoub, M. and Dong, J. 1997. Ad hoc experiments using EUREKA. TREC-5, Pages 229-240.

[10] Mitra, M., Singhal, A. and Buckley, C., 1998, Improving Automatic Query Expansion. In Proc. of the 21st Annual Int.ACM SIGIR Conf. on Research and Development in Information Retrieval, pp 206--214, Melbourne, August 24 -28 1998.

[11] Qiu, Y. and Frei, H., 1993, Concept based query expansion. In Proc. of the 16th International ACM SIGIR Conference on R & D in Information Retrieval, pages 160--169. ACMPress, New York.

[12] Rocchio. J. 1971. Relevance feedback in information retrieval. The Smart Retrieval system---Experiments in Automatic Document Processing. G. Salton. Ed. Prentice-Hall Englewood Cliffs. NJ. pp.313-323.

[13] Ricardo Baeza-Yates and BerthierRibeiro-Neto. 1999. Modern Information Retrieval. Pearson Education Limited, England, 1999.

[14] Salton, G. and Buckley, C. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science. 41(4): pp. 288-297, 1990.

[15] Sparck Jones, K. 1971. Automatic keyword classification for information retrieval. Butterworths, London, UK.

[16] Wen, J.-R., Nie, J.-Y. and Zhang, H.-J. 2000. Clustering User Queries of a Search Engine. WWW10, May 1-5, 2001, Hong Kong.

[17] Xu, J. and Croft, W.B. 1996. Query expansion using local and global document analysis. In Proceedings of the 19th International Conference on Research and Development in Information Retrieval, pages 4--11, 1996.

[18] Xu, J. and Croft, W.B. 2000. Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems Vol.18, No.1, January2000, Pages 79-11.

[19] "THESAURUS AND QUERY EXPANSION" International Journal of Computer science & Information Technology (IJCSIT), Vol 1, No 2, November 2009.

Biographies:

Priya Mundada completed her Bachelor of Engineering in Computer Science & Engineering in 2014.She is pursuing her Masters in Technology in Computer Science and Engineering from Shri Ramdeobaba College of Engineering and Management, Nagpur-440013. Her areas of interest include Information Retrieval.

Dr. Manoj Chandak received the Masters in Computer Science and Engineering as a first merit holder. He is Ph.D. in Computer Science and Engineering with over 20 years of teaching experience. He has 29 research paper publications and many presentations in conference, to his credit. His research interests include Design & Analysis of Algorithm, Advance Algorithms.