AN APPROACH FOR IR OF HOTEL REVIEWS USING EXTRACTION AND EXPANSION OF MICROPOST

[1]PRIYA MUNDADA, [2]MANOJ CHANDAK

[1]Student, Computer Science and Engineering, [2]H.O.D., Computer Science and Engineering,
Shri Ramdeobaba College of Engineering & Management, Nagpur, India

Abstract— The process of supplementing additional terms or phrases in the original query to improve the retrieval performance is known as Query expansion. In today's world, the data retrieval performance of a search has an important role in every field. Along with the retrieval performance, accurate information is also required. It is very difficult to get the precise information which is actually required by the user through Micro post or a 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 the least number of words. Since there are 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, Micropost, Lingo Words, Information Retrieved.

I. INTRODUCTION

Nowadays it has become a trend to post all the current updates on social media in order to keep them updated in the current world. Today's world is a virtual world. It deals with a large amount of textual data in the form of posts and comments. But these posts are not in the proper format to retrieve knowledgeable data or complete information. Hence, query expansion is required. The process of reformulating the query to enhance retrieval performance of information retrieval operations is called as Query expansion (QE). It involves evaluating a user's input i.e. evaluating the user query on the basis of words typed in the query area. 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 may happen when they try to put their query to the search engine. Hence, it is generally observed that web users put the very short query to search engines. This 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. The search engine is unaware 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 the user is least bothered about the grammar. It is also possible when the user does not provide proper information which will surely lead to wrong output or misconception. 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.

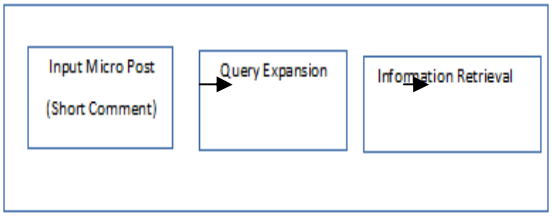

Figure 1: Block Diagram

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 the background work for query expansion. This helps to expand the idea of the proposed system in terms of using new techniques for the proposed methodology. Section III is the overview of the algorithm used. Section IV explains the module in which the project is divided. It discusses all the phases used in the proposed system along with the approach used in it. Section V is the proposed system. Section VI provides information regarding the dataset used. Section VII provides the advantages of the proposed system in each phase. Section VIII is the experimental analysis where the result is analyzed. Section IX is the conclusion.

II. BACKGROUND

The 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 a cluster of document term based on their co-occurrences. Queries are expanded by the terms in the same cluster. On the other hand, the local analysis uses only some initially retrieved documents for further query expansion. Expansion terms are extracted from the relevant documents. If the user provides proper and accurate relevant judgment then this technique works effectively. But in general scenario user does not provide the proper relevant judgment. The other techniques are re-ranking, 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].

Expansion terms are extracted from the top-ranked documents to formulate a new query.

There are some of the techniques which talk 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.

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 puts light on various applications of AQE such as a Question Answering System, Multimedia Information Retrieval, Information Filtering and Cross-Language Information Retrieval.

III. OVERVIEW

In this section, we are going to have an overview of the algorithm used in the project. Keyphrases provide semantic metadata that summarize and characterize documents. The Key Extraction Algorithm [KEA], an algorithm for automatically extracting keyphrases from the text. KEA identifies the candidate keyphrases, using lexical methods, calculates feature values for each candidate, and uses a machine-learning algorithm to predict which candidates are good keyphrases. The machine learning scheme first builds a prediction model using training documents with known keyphrases, and then uses the model to find keyphrases in new documents. We use a large test corpus to evaluate KEA's effectiveness in terms of how many author-assigned keyphrases are correctly identified.

In this project, the KEA is divided into following steps. These steps are performed on the clean dataset of comments of the different hotel. These are the short comments i.e. the Micro-post:

1. Text Processing

After the tokenization is done the symbols are removed. The case of data is kept as it is because the named entities are recognized by the first upper-case letter.

2. Then we apply POS tags. This will tag each word to its respective Parts Of Speech.

3. Then we apply bi-grams on the tagged data.

4. After that, we keep only phrase patterns i.e. Noun-Noun, Adjective-Noun, Named Entity- Noun or Noun-Named Entity. All low-frequency words are removed except for named entities.

5. Now the patterns which satisfy the user query words are provided as the suggestions.

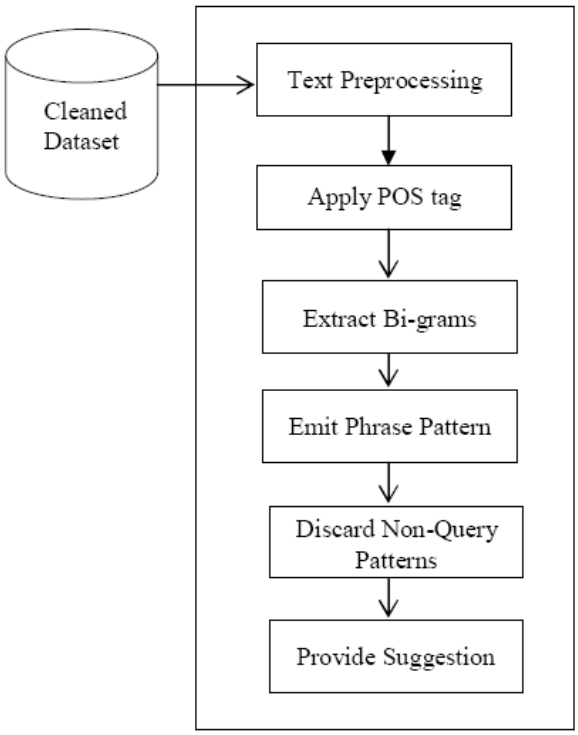

Figure 2: Block Diagram of KEA

IV. MODULE DIVISION

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.

For Example: "nyc locatn" will be replaced by "nice location".

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;
}
}
}

2. Grammar Correction

In Grammar Correction, the input query is corrected grammatically. For this, we have used Sentence Pattern for Assertive or Declarative sentence and Comment Pattern for input comments. Sentence Patterns are the rules defined for the Assertive or Declarative sentence construction. Comment Patterns are the rules defined for the comments which exclude the 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;
}

3. N-Gram

After grammar correction, we apply N-gram algorithm to our cleaned data of short comment of the dataset for extracting the key phrases. These are the bi-grams of defined pattern (Noun-Noun, Adjective-Noun, and Named Entity-Noun). As on requirement, we will populate the query with the retrieved phrases.

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 to 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.

V.PROPOSED SYSTEM

A new approach has been used for Query Expansion. The first phase of our proposed system is the Dataset cleaning. The dataset needs to be cleaned 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 per requirement. The words are added after testing the lingos of different personalities.

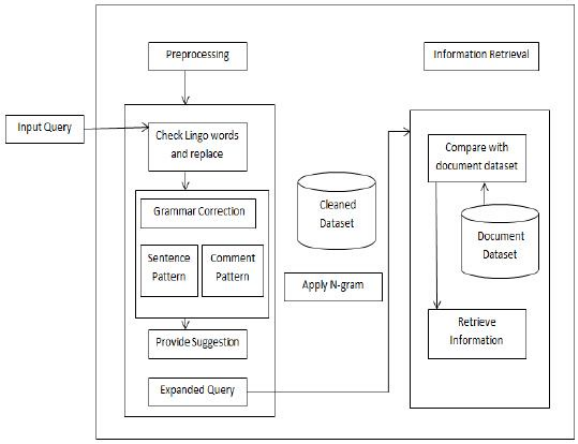

Figure 3: Proposed System Framework

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 consist of a short 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 a single word can change the meaning and the sense of retrieved information. Hence, once the query is expanded phrases (Noun Phrase and Verb Phrase) will be extracted from this query. 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 categorized. This categorization will tell whether the query is positive or negative.

VI. 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 the 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 associated with a number of reviews. Each review is placed on a new line. A single review is separated into 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. The total size of the Dataset is 480.9 MB. This is a huge dataset containing a 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.

VII. ADVANTAGES OF PROPOSED SYSTEM

The main advantage of the proposed system is that it will provide the information only of user interest. Above explained various approaches have been used to improve the retrieval performance. It also increases the query optimization process and gets the relevant information related to the user query. Another advantage is that corrected grammar query leads to more precise information retrieval process.

VIII. EXPERIMENTAL RESULT ANALYSIS

We have performed some experiments on the project and checked the analysis of the proposed system. Here we have divided the results according to the query provided by the user. To calculate the accuracy of the system we have used recall and precision methods.
We have divided the result into three categories as per the query provided.
A – Number of relevant files retrieved according to query.
B – Number of relevant files that are present in the dataset but are not retrieved.
C – Number of irrelevant files retrieved that is the files that are not related to the query provided. (Eg. Input query is good hotel and output has good food).

Result retrieved for query are of two types:

i) Grammar Correction
ii) Without Grammar Correction

i) Grammar Correction

In the case of Grammar Correction, the query provided by the user is grammatically corrected. After that further processing is done.
We have performed few experiments to check the accuracy of the system. They are as follows:

1. Good location (A-1733,B-0,C-0)
2. Good food (A-889,B-0,C-0)
3. New hotel (A-381,B-0,C-237)
4. Wonderful Stay(A-722,B-0,C-0)
5. Disappointing hotel(A-6,B-0,C-1)

Total A = sum of A for all comments / total comments
= 3731/5 = 746.2
Total B = sum of B for all comments / total comments
= 0/5 = 0
Total C = sum of C for all comments / total comments
= 238/5 = 47.6
Recall = A/A+B
= (746.2/ (746.2+0))*100 = 100
Precision = A/A+C
= (746.2/ (746.2+47.6))*100 = 94
Accuracy = 2 * P*R/P+R
= 2*(94*100)/94+100
Accuracy = 96.90

Here we have taken few words such as Good location as our input query and according to that, the files are retrieved, these files are then categorized as A or B or C as per the output. The files that are retrieved according to the query are 1733 i.e. A=1733, there are no files that are present in the dataset but did not retrieve thus, B=0 and there are no files that are irrelevant to the query C=0.
After this, an average of results of all the queries are done and then recall and precision is calculated. And by using this recall and precision accuracy of the results is calculated.

ii) Without Grammar Correction:
In the case of Grammar Correction, the query provided by the user is processed as per the flow mentioned above.
We have performed few experiments to check the accuracy of the system. They are as follows:

1. Location good (A-117,B-12,C-468)
2. Food good (A-116,B-13,C-1)
3. Hotel new (A-15,B-18,C-40)
4. Wonderful Stay (A-38,B-20,C-3)
5. Hotel disappointing (A-0,B-21,C-1)

Total A = sum of A for all comments / total comments
= 286/5
= 57.2
Total B = sum of B for all comments / total comments
= 0/5

= 0

Total C = sum of C for all comments / total comments

= 513/5

= 102.6

Recall = A/A+B

= (57.2/ (57.2+0))*100

= 100

Precision=A/A+C

= (57.2/ (57.2+102.6))*100

= 35.79

Accuracy = 52.71

Table 1: Comparison Table

	Recall	Precision	Accuracy
Without Grammar correction	100	35.79	52.71
With Grammar Correction	100	94	96.9

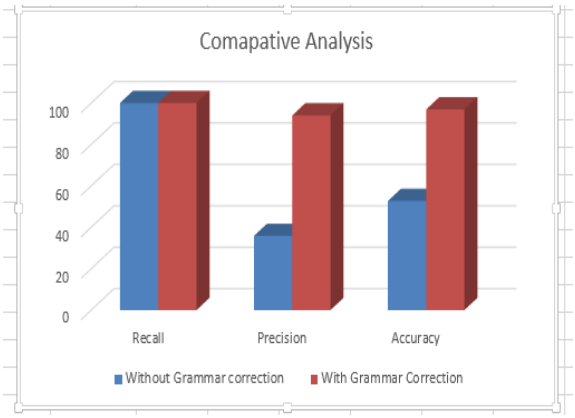

Figure 4: Analysis Graph

Here we have taken few words such as location good as our input query and according to that, the files are retrieved, these files are then categorized as A or B or C as per the output. The files that are retrieved according to the query are 117 i.e. A=117, files that are present in the dataset but did not retrieve are 12, B=12 and there are 468 files that are irrelevant to the query thus C=468.

After this, an average of results of all the queries is done and then recall and precision is calculated. And by using this recall and precision accuracy of the results is calculated.

CONCLUSION

It is very difficult to retrieve related information from the inappropriate information. We need to have the proper query in an appropriate format for information retrieval. Query Expansion is a proposed technique to get the actual sense of the asked query and then retrieve the information. Various approaches have been used to optimize the retrieved query. Query expansion will help to retrieve the precise information from the micro post. This expanded query will help the user to get accurate information.

FUTURE SCOPE

Query Expansion is a process in which optimization of retrieved information can be achieved. Hence, it can be used in Search Engine Optimization. Here we have used only Assertive/Declarative sentence patterns. We can increase the accuracy by implementing good NLP techniques by applying other forms of grammar correction for sentences like Interrogative and Exclamatory Sentence.

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 super concepts 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, ImprovingAutomatic Query Expansion. In Proc. of the 21st Annual Int.ACM SIGIR Conf. on Research and Development inInformation 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 inAutomatic 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. ClusteringUser 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 19thInternational Conference on Research and Development inInformation Retrieval, pages 4--11, 1996.

[18] Xu, J. and Croft, W.B. 2000. Improving the effectiveness of information retrieval with local context analysis. ACMTransactions 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.

★★★