Restructured Framework for Hunting User Goals

Mr. N. Arunachalam[1], S. Karthikeyan[2], C. Kaviarasan[2], A. Bharath[2]

Assistant Professor, Dept of Information Technology, SMVEC[1]
B. tech, Dept of Information Technology, SMVEC[2]
narunachalam85@gmail. com, karthisivait@gmail. com,
caveearasan@gmail. com, bharathalonearjun@gmail. com

Abstract:

In today's world each user expects their search engine to seek results with more accuracy. Data mining is the key process that is involved in extracting information from a data set and transforming it into an understandable structure for further use. Since for a particular query there are meanings in multiple dimensions and even for the same query each user has multiple thinking. The main goal of search is to find the exact meaning of the query and should satisfy the needs of the user. The existing search engines can only show the links that is often used by the people. But it cannot satisfy all the user needs. We here define a framework that would try to portray the multiple view of a query and prioritizes them in a right manner. A dictionary is used to find multiple meaning of query and hence related data is retrieved in an efficient way.

Keywords:-Data mining, search engines, dictionary, framework, user goal

1. Introduction:

In present world, the researchers are turning their focus towards data henceforth the retrieving and storing of data are major areas of focus in data. Mostly the data stored in the database are semi structured i. e. either completely structured or not completely structured. So here comes data mining into play which means the key process that is involved in extracting information from a data set and transforming it into anreasonable structure for further use. But, traditional Information Retrieval techniques become inadequate as the size and complexity of the data increases. Typically, if we truly see only tiny fraction of the existing documents is needed for a single user or individual and without being aware of what could be in the documents; it is pretty risky task of formulating effective queries for analysing and extracting useful information from the data. An effective tool is essential for Users to analyse

and compare different documents, rank the documents according to the relevancy, or to discover patterns across multiple leaflets. Thus, Text Mining has become an increasingly widespread and essential subject in Data Mining.

In web applications, a search engine plays a vital role in seeking information from the World Wide Web. And the results of it, is the corpus that is relevant to the user query. The efficiency of the search engine depends on how well the corpus relates to the query. The search engine has to meet the challenge of satisfying the exact need of the user because each query may have different meaning and each user search form different aspect. Each and every user may have different needs for a same query so, there comes the actual challenge for a search engine. For example if we search for query "apple", it may be a fruit or it may be a commercial product. But the actual aspect of the query may differ from each user. Therefore it is necessary for search engine to display the relevant documents that is related to a query. Generally for a single query there are many meaning, so here we try to focus various meaning of a single word. This improves the accuracy of a query engine and the increases the relevance of documents related to a query.
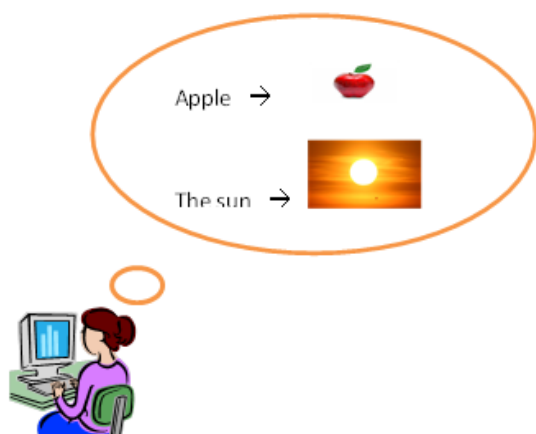

Figure 1. Here the user needs apple to be a fruit and the sun to be the star of the solar system that is located at the centre of the solar system.
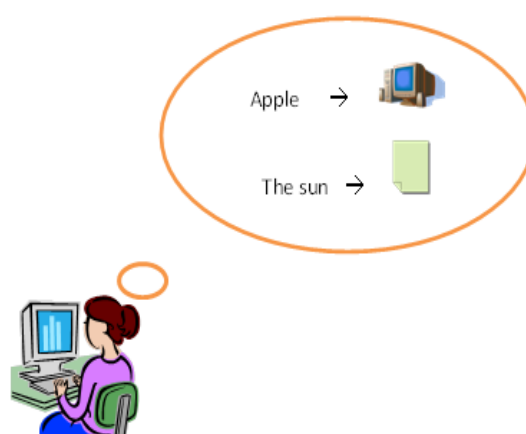
Figure 2. Here the user views apple as the computer product and the sun as English newspaper.

2. Data mining:

Data mining is the process of scrutinising data from different perceptions and briefs it into useful information. Data mining software is one of a number of diagnostic tools for investigating data. The main goal of search is to find the exact meaning of the query and should satisfy the needs of the user. The existing search engines can only show the links that is often used by the people. But it cannot satisfy all the user needs. We here define a framework that would try to portray the multiple view of a query and prioritizes them in a right manner.
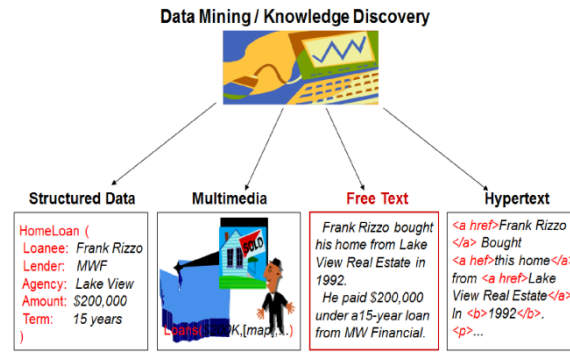
Figure 3. General architecture diagram of data mining.

3. **Framework of our approach:**

We here define a framework that enable the user to see the various view of his/her query. So that he/she can able to choose his/her exact need. Initially the user types his/her query in the search box. Then, the query is made into tokens then the tokens are referred with dictionary to find the various view of that query. And finally we categorise and cluster them into a web page.
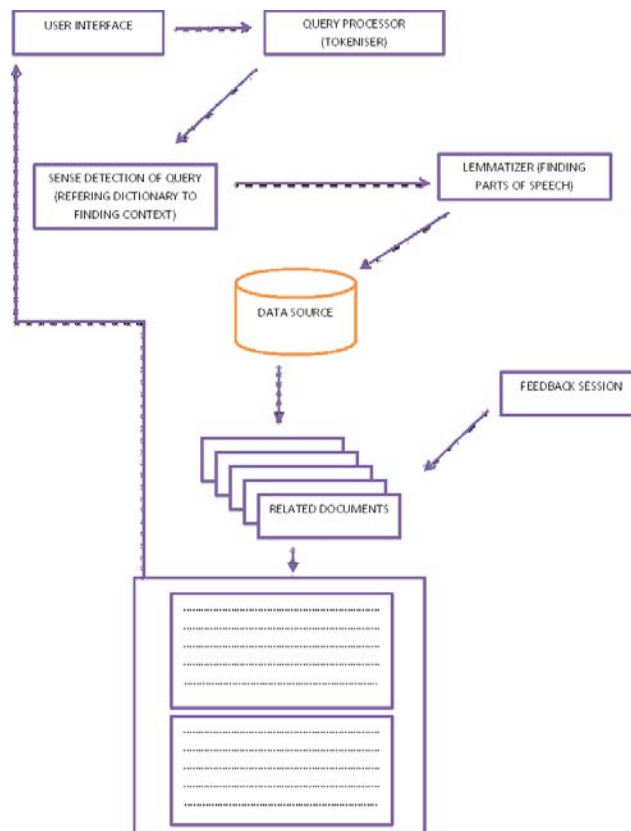

Figure 4. Architecture diagram of our frame work

4. Module of our approach:

A. Module 1:

This module is meant to fetch the input from the user and convert them into a meaningful words. Here we use tokenization concept to seek a set of meaningful words from the user's query. So, tokenization is an important process that is involved in fragmenting the stream text into words, symbols, phrases which are called as tokens. In order to find the various meaning of those tokens. Here we use wordnet2. 0 to refer various context of a particular token
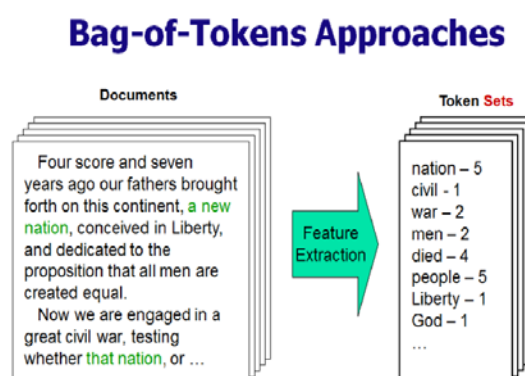
Figure 5. Tokenising process

B. Module 2:

This module is meant to sense the tokens that were tokenised in the previous module such that, we can focus the diversified thinking of the user and can portray it through related links.

C. Module 3:

This module is meant to lemmatize the tokens where lemmatization can be defined as the process of finding the root word of the tokens and organising them, further finding relation between them. Using this process the query is being optimised in a well efficient manner.

D. Module 4:

In this module the lemma of those tokens is passed to the crawler to seek the related URLs from the internet server and the fetched URL are returned back to the user interface. The optimised results are returned back to the UI.

5. Methods used to cluster corpus related to hunt user goals:

Now the optimised results are grouped under various domains by the use of clustering algorithm by finding the relative matching of keywords existing between the documents. We here use k means clustering algorithm to cluster the documents based on the relativeness existing between them..

6. Methodologies used to evaluate relevancy:

We here use some methodologies to find the relevancy between the documents and the optimised query. Here we describe about the method to infer the user search goal and portray them with imperative keywords.

Table 1. set of Abstract keywords

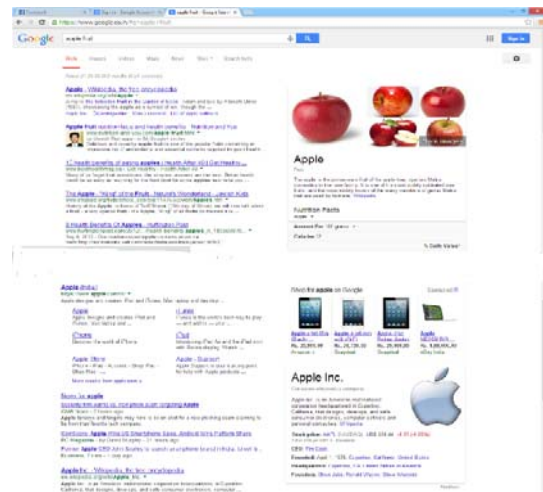| QUERY	Keywords to depict user search goals
Apple	Fruit, computer product
The sun	Star at the centre of solar system, newspaper
Storm	Hurricane, tornado, big data tool
Lamborghini	Car, history, company, overview, automobile

Further In order to restructure the documents we here use the classified average precision method.

Our proposed work, groups the session history to infer the user search goals and the method we apply here will confine the top results based on the relevancy existing between the query and the data existing in the internet. This enables us to portray different context of query

7. Sample output:

Figure 6. The actual output **Figure 7.** Our Modified output

8. Conclusion:

In this paper we define a frame work that portrays the various context of the query of the user. Such that to fulfil all types of user need. Firstly we derive various context of the query that we get from the user then we give diverged context to search engine

that sort out set of results related to the query. We further cluster those results based on the keywords, for this we use k means clustering algorithm. Thus user can get the results in different groups such that user can find what they actually need.

9. References

[1] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines, " Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588-596, 2004.

[2] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log, " Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '00), pp. 407-416, 2000.

[3] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "Varying Approaches to Topical Web Query Classification, " Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development (SIGIR '07), pp. 783-784, 2007.

[4] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay, "Accurately Interpreting Clickthrough Data as Implicit Feedback, "Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '05), pp. 154-161, 2005.

[5] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. ACM Press, 1999.

[6] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "Context-Aware Query Suggestion by Mining Click-Through, " Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '08), pp. 875-883, 2008.

[7] X. Wang and C. -X Zhai, "Learn from Web Search Logs to Organize Search Results, " Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.

[8] J. -R Wen, J. -Y Nie, and H. -J Zhang, "Clustering User Queries of a Search Engine, " Proc. Tenth Int'l Conf. World Wide Web (WWW '01), pp. 162-168, 2001.

[9] H. -J Zeng, Q. -C He, Z. Chen, W. -Y Ma, and J. Ma, "Learning to Cluster Web Search Results, " Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '04), pp. 210-217, 2004.