Semantic Similarity Measures for Best Keyword Search

Shukrali Sawant[1] Amol Rajmane[2]

[1,2,3]Department of Computer Science and Engineering

[1,2]Ashokrao Mane Group of Institution Shivaji University, Kolhapur, Maharashtra, India

Abstract— As in various search mechanism keyword search is used which provides a simple but user friendly interface to retrieve information. Web search engines are widely used for searching textual documents. An interesting problem known as Closest Keywords search is to query records, called keyword cover, which together cover a set of query keywords and have the inter-record score. In recent years, the availability and importance of keyword re-rank in record evaluation for the better decision making is increasing. This motivates us to investigate a generic version of Closest Keywords search called Best Keyword Cover (BKC) which considers inter-record score as well as re-rank. Also the exact measurement of semantic similarity between words is essential for various tasks such as, information retrieval and synonym extraction. It should be able to understand the semantics or meaning of the words. But in some cases user enters input as a phrases or meaning of the words then it is critical to find the accurate keyword so WNRD (Word-Net Reverse Dictionary) is use. When the set of query keywords is generated by query generation different results are found, to get the most probable result from the given set k-Nearest Neighbour is use on the basis of inter-record score.

Key words: Keyword Search, k-Nearest Neighbour, Semantic Similarity, Word-Net Dictionary, WNRD

I. Introduction

Searching for information is an indispensable component of our lives. Web search engines are widely used for searching textual documents. Semantic similarity measures play an important role in the extraction of semantic relations. Numerous information retrieval and natural language processing applications require knowledge of semantic similarity between words or terms. Computer being a syntactic machine, it cannot understand the semantics. So always an attempt is made to represent the semantics as syntax. There are various methods proposed to find the semantic similarity between words. Some of these methods have used the precompiled databases like WordNet.

The Semantic Decomposition, WNRD and pattern mining techniques are also use to improve performance of information retrieval. A WNRD performs a reverse mapping i.e., given a phrase describing a desired concept, it provides words whose definitions match the entered definition phrase, as opposed to a regular (forward) dictionary that maps words to their definitions.

All the resulted words found from Semantic Decomposition, WNRD and pattern mining techniques, are processed with k-Nearest Neighbour, where score is calculated depending on this inter-record score result is re-rank and send to the user as a final result. The Best Keyword Cover (BKC) query which considers inter-record score as well as keyword re-ranks. It is motivated by the observation of increasing availability and importance of keyword score and re-ranking in decision making.

k-Nearest Neighbour is one of the most popular algorithms for text categorization. Proposed work mainly focuses on finding top k-Nearest Neighbours, where each node has to match the whole querying keywords. We propose an improved k-Nearest Neighbour algorithm, which uses different numbers of nearest neighbours for different categories, rather than a fixed number across all categories.

II. Literature review

Tao Guo, Xin Cao, Gao Cong, [1] prove that the problem of answering mCK queries is NP-hard. They have devised a greedy algorithm that has an approximation ratio of two. That a mCK query can be approximately answered by finding the circle with the smallest diameter that encloses a group of objects together covering all query keywords. It is proved that the group enclosed in the circle can answer the mCK query with an approximation ratio of p2/3. Based on this, they have developed an algorithm for finding such a circle exactly, which has a high time complexity.

Ke Deng, Xin Li, [2] BKC query have provided an additional dimension to support more sensible decision making. The baseline algorithm generated a large number of candidate keyword covers which leads to dramatic performance drop when more query keywords are given. They proposed a much more scalable algorithm called keyword nearest neighbour expansion (keyword-NNE).

S. Bergamaschi, E. Domnori [3] introduced the novel framework for keyword search in relational databases. In contrast to traditional keyword a search technique that requires access to the actual data stored in the database in order to build an index, their technique uses intentional knowledge.

Sheetal A. Takale,Sushma S. Nandgaonkar [4] introduced the new approach for measuring semantic similarity between words using the snippets returned by Wikipedia and five different similarity measures of association. Snippets in Wikipedia are used to measure semantic similarity between words. The result demonstrates that the snippets in Wikipedia have a significant influence on the accuracy of semantic similarity measure between words.

We have reviewed different techniques like semantic similarity measure, BKC query processing, and keyword-NNE. In semantic similarity measures five different similarity measures of association are used for measuring similarity between words. From this concept we are going to use semantic similarity measure for finding the semantic word of related query.

In BKC query processing they have proposed a system which finds the inter object distance and apply Keyword-NNE algorithm for more efficient result. By observing this technique we are going to use inter record score method. And k-Nearest Neighbours will be applied on that score.

III. PROPOSED WORK

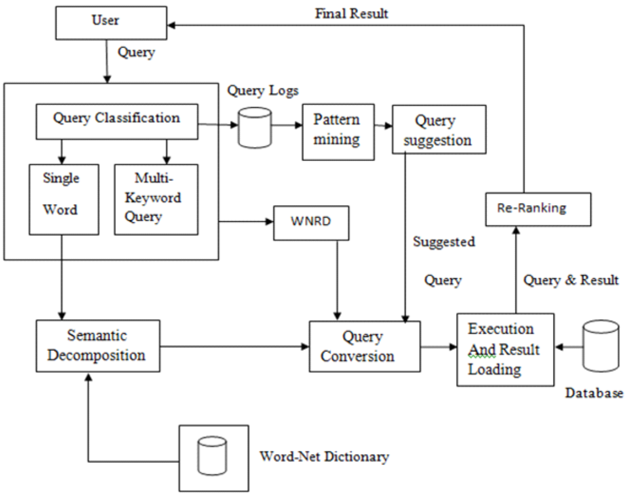

Fig. 1: Architecture of Semantic Similarity Measures for Best Keyword Search

Proposed work is based on Semantic Similarity Measures for Best Keyword Search which is having improved performance of information retrieval.

When user enters the query, it is classify at the Query Classification module, then single word and multi-keyword query is found. The single word query is send to the Semantic Decomposition module in that Word-Net Dictionary is use to find the semantic words of that query. The multi-keyword query is send to the WNRD module. WNRD is use to find the single word of that multi-keyword.

The whole keyword cover is stored into Query Conversion module. Query Classification module give queries to store them into query log. These log records are used for creating patterns. The resulted pattern is compared with current query given by the user and query suggestion related to that is given to query conversion text file.

In Query Conversion module the result is retrieved from semantic decomposition, WNRD. These resulted words are processed to calculate their power set. After calculating the power set values, these values are stored into text file with query suggestion.

Execution and Result Loading module gives the most appropriate keyword identification using k-Nearest Neighbour. The set of keywords is given to Execution and Result Loading module from query conversion text file, current set is compare with stored database and found result is given to re-ranking module. Then according to high-low score the results are obtained on which raking is performing to give the user best optimize response to his request.

IV. METHODOLOGY

A. *Semantic Decomposition and WNRD:*

The Semantic Decomposition module measures the semantic similarity. The single word query is send to the semantic decomposition module where the Word-Net Dictionary is use to find the semantic word of that query.

Reverse Dictionary is use i.e, WNRD (Word-Net Based Reverse Dictionary). In WNRD module performs a reverse mapping.

B. *Pattern Mining:*

Data mining will be perform on resulted pattern with the help of apriori algorithm and association rule mining. The resulted pattern should be compare with current query given by the user and query suggestion related to that is given to query conversion text file.

C. *k-Nearest Neighbour Algorithm:*

Execution and Result Loading Module uses the k-Nearest Neighbour Algorithm. When the set of query keywords would be generated by query generation different results are found, to get the most probable result from the given set k-Nearest Neighbour is use on the basis of inter-record score.

V. CONCLUSION

k-Nearest Neighbour algorithm is proposed to solve the text categorization problem statically. The algorithm uses a predefined set of documents with known categories. In this paper we have reviewed Semantic Decomposition, WNRD and Pattern Mining techniques. These techniques can be apply to improve performance of information retrieval.

REFERENCES

[1] Tao Guo, Xin Cao, Gao Cong, “Efficient Algorithms for Answering the m-Closest Keywords Query”, *SIGMOD*’15, May 31–June 4, 2015.

[2] Ke Deng, Xin Li, Jiaheng Lu, and Xiaofangzhou “Best Keyword Cover Search”, *IEEE Trans. Knowl. Data Eng*., vol. 27, no. 1,. 2015, pp.

[3] Sonia Bergamaschi, Elton Domnori, Francesco Guerra, Raquel TrilloLado, YannisVelegrakis, “Keyword Search over Relational Databases: A Metadata Approach”, *SIGMOD*’11, June 12–16, 2011.

[4] Sheetal A. Takale, Sushma S. Nandgaonkar “Measuring Semantic Similarity between Words Using Web Documents”, *International Journal of Advanced Computer Science and Applications*, Vol. 1, No.4 October, 2010.