Keyword Query Based Clustering Technique for Efficient Web Page Retrieval

SaravanaKumar K[#1], Kauser Ahmed P[*2], Deepa K[#3]

School of Information Technology & Engineering

[*]*School of Computer Science & Engineering*

VIT University, Vellore 632014, TamilNadu, India

[1]ksaravanakumar@vit.ac.in, [2]kauserahmed@vit.ac.in, [3]deepa.k@vit.ac.in

Abstract

In this paper, a keyword query based clustering technique is proposed for efficient web page retrieval. The advantage of keyword query is that the user will get the refine and domain based information based on the clustering formation. The existing system lacks in performance due to their inefficient clustering result. This paper focuses on two major challenges of keyword search for efficient web page retrieval. First, Generate alternate queries for main query; second, how to make clustering of relevant and irrelevant data. The purpose of this paper is to design a novel algorithm for keyword query and develop clustering technique based on the corpus created from searched documents. Proposed algorithm cluster the search results and show that our method out performs when compared to traditional search engines.

Keywords: Information retrieval; Clustering; Alternate Queries.

1. INTRODUCTION

Information is growing extremely in the international network and World Wide Web is considered as knowledge repository. With the rapid growth of information on WWW, it is difficult to manage large corpus and convince the user with suitable information [3]. Because of this, users are looking for meaningful search results using various search engines within stipulated times [2]. Figure 1 shows architecture of typical search engine [3].

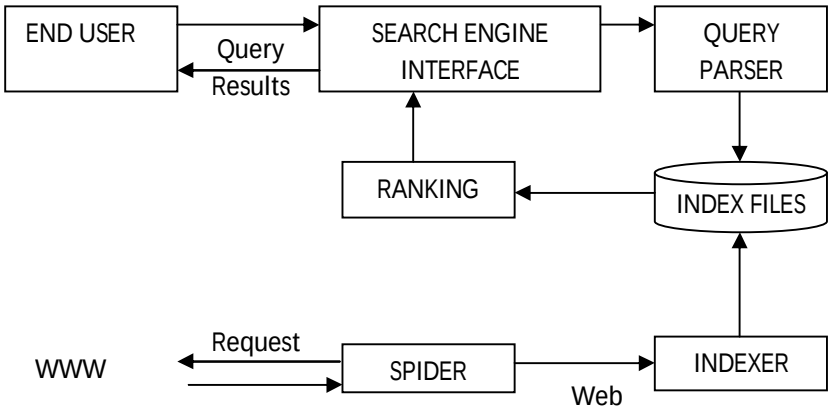

Figure 1: SEARCH ENGINE ARCHITECTURE

Search engine components consists of crawler that traverses the web to download the web pages, indexer to parse the web pages and ranking to display the relevant web pages at the top [3].When a user enter a query using query interface, the query processor process the query through index and returns the result with most relevant pages at the top. The success of a search engine depends on indexing and ranking mechanism. Keyword based search is the well known information retrieval mechanism for data on the World Wide Web. The rest of the paper is as follows; Literature survey is done in section II, Proposed model and methodology is explained in section III, Implementation details are discussed in the section IV, Experimental results are discussed in section V. Finally the paper is concluded in the conclusion and future work section.

ISBN 978-93-82338-91-8 © 2013 Bonfring

2. LITERATURE SURVEY

TABLE 1: Operational Constructs and Its Referential Sources

Construct	Referential source	View point	Research gap
Keyword retrieval	S. Kim et al. / Expert systems with Applications 39 (2012)	Graph ranking is done in this paper.	Semantic keyword search.
Keyword search	C. Kim et al. / Expert Systems with Applications 39 (2012)	Three factors are discussed in this paper. 1. Keyword search advertising factors. 2. Effects of individual keyword advertising. 3. Keyword efficiency.	Keyword ranking.

Ranking of web search engine results depends on the keywords which we are using in query .If the query consist of meaningful keywords, then the search result will be fruitful and produce high relevance result. Keyword based search engine is not efficient because the semantic of the keyword is not considered while search the data in the web. Saravanakumar [9], proposed the alternate query construction mechanism but the context related semantic is not discussed. Alternate query generation and expansion will helps the user in finding the semantic of the query and to get efficient search result in stipulated time. Hamada [14] proposes for suggesting a list of queries that are related to the user query. The related queries are like the query logs based on previously issued queries by the users. The proposed method furnishes the clustering techniques in which semantic queries are identifies but the expansion of queries is not done. This facility provides some queries which are related to the queries submitted by users in order direct them toward their required information.

3. PROPOSED FRAMEWORK & METHODOLOGY

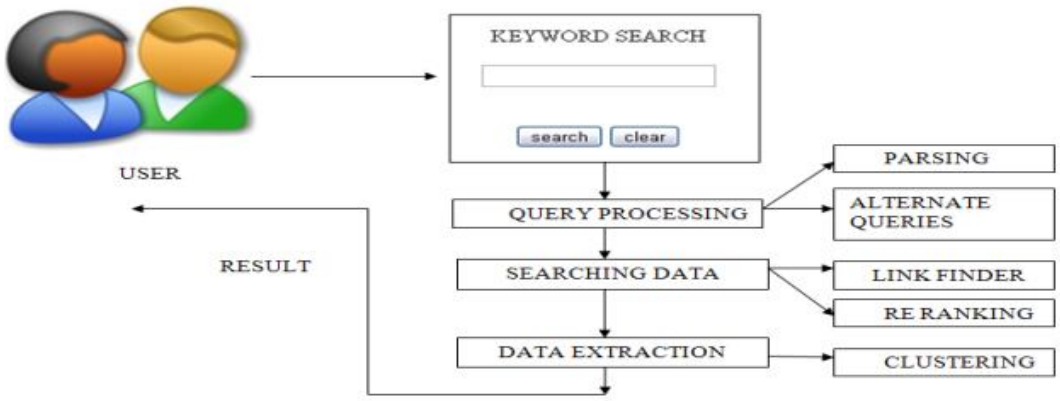

Figure 2: Proposed Frameworks

Figure 2 represents the framework of proposed model. The methodology of the proposed model is as follows. The proposed system consists of three main modules. Each module is treated as an independent module and can be implemented individually. The first module, called the query processing. The second module is about searching data. In the third module, data extraction is done.

3.1. Query Processing

The first module is query processing. Query processing consists of two parts, pre-processing and alternate query generation. In pre processing, the user query is converted into tokens so that the semantics of each word can be identified easily. Second part

ISBN 978-93-82338-91-8 © 2013 Bonfring

consists of alternate query generation in which the similar queries are generated for user query. Then English lexical database WordNet is used to generate possible alternate sub queries for main query.

The aim of this module is to investigate novel query representations for processing and generating alternate queries and finding the semantics of each and every word in the query so that the relevancy can be high. Understanding, pre processing and re-writing user queries are considered as the main core in web page retrieval.

3.2. Searching Data

The second module is data searching. Data searching consist of two parts. Ranking and Finding the relevant links for the given query. The rank of a web page depends on the keyword available in the web page search result. Not so long ago, search engines based their indexing chores on the amount of keyword incorporated on a web page. That means the larger the keyword density of a page, the higher it would rank on search engines.

The second part of this module is to find the relevant links to the user query. Based on the main query and the alternate query generated by WordNet , we will search the result for all the queries using Google. Link finder will take the result of main query and find the presence of all links in the alternate query results. If the link is available in main query and alternate query, we will increase the link weight based on the number of times the link is available in the alternate queries. The re ranking is done based on the position of the links available in the alternate queries.

3.3. Data Extraction

This module consists of two parts, re ranked documents and cluster result list. Re ranking of documents is done based on the number out links to a search web page and the number similar queries available in alternate queries. Re ranking of documents is done to display the web pages based on ranking in the top list.

Data extraction is done based on the clustering of the search results which will find the presence of out links from the main query to all the possible sub queries. K Means clustering is done here and we will find that the relevant data and irrelevant came in different clusters. The similarity is high when the cluster consists of relevant data and the similarity is low the when the cluster data consist of irrelevant data.

4. IMPLEMENTATION OF THE SYSTEM

4.1 Algorithm development

The proposed framework in chapter three can be implemented with the help of three algorithms shown below. In the given algorithms, K_Q is the Query Keyword which is the input, K_i means set of alternate Queries, P is the Links of web pages, O represents the Output generated by algorithms 1 i.e., set of web links, RR represents the Re Ranking of listed web links, AQ represents Alternate Query, OL is OutLinks, CD means Cluster Documents and CS represents Cosine Similarity value found between several documents. Proposed Algorithm 1 explains the query processing part which consists of query processing and alternate query generation [9]. The output of this algorithm is the set of related links to the given query. Algorithm 2 Finds the link and re ranking is done based on the presence of the link in the alternate queries. Algorithm 3 makes the clustering of search results based on the corpus and finds the cosine similarity to furnish the expected result.

Algorithm 1 - Query Processing

Algorithm GQ (K_Q, O)

INPUT: K_Q

OUTPUT: O

1: $O \leftarrow \emptyset$

2: $\forall$ K_Q Genarate K_i

3: $\forall$ $K_i \in K_Q$ do

4: Retrieve pages from K_i & K_Q and find P

5: $P \leftarrow$ Set of all possible Links

ISBN 978-93-82338-91-8 © 2013 Bonfring

6: ***if*** *p contains all* K_i & K_Q *then*

7: $O \leftarrow O \cup P$

8: ***end if***

9: Display O

Algorithm 2 - Link Finder

Algorithm LF(O, RR)

Input : O

Output : RR

1: $\forall$ $L \in O$

2: Increment RR if L present in AQ

3: Find the position in AQ

4: Display RR based on new position

Algorithm 3 - Clustering of Search result

Algorithm C (RR, CS)

Input : RR

Output: CS

1: $\forall$ $RR \in OL$

2: Find D based on OL

3: $\forall$ $OL \in D$

4: Increment D weight $\forall$ *OL*

5: make Corpus of D

6: Cluster the result of D

7: Find CS of CD

8: Display CS.

5. EXPERIMENTAL RESULTS

In this section, we present the detail of the experiment conducted on main query search results .All the programs used for the experiments were written in Visual Basic and Python. As an initial stage, the user written query is converted into set of alternate queries and an example is given in Table 2 below [9]. After identifying a set of relevant alternate queries, we search for the results of every alternate query. Later the results are analyzed for the availability of domains of the main query's links as the out-links.

ISBN 978-93-82338-91-8 © 2013 Bonfring

TABLE 2: The user query and the list of constructed alternate queries

Main Query
American basketball team
Alternate Queries generated:
American basketball team
American basketball squad
American basketball game team
American basketball game squad
American English basketball team
American English basketball squad
American English basketball game team
American English basketball game squad
American language basketball team
American language basketball squad
American language basketball game team
American language basketball game squad

All the documents in the result set which showed the core page of the main query results' links as out-links are grouped together. This set of documents are treated relevant and considered as a corpus. This corpus is later used to make the clusters among relevant and irrelevant data. The following example shows how the document clustering is done using Pattern web mining module available in Python [17].

```
>>> from pattern.vector import Document, Corpus, KMEANS

>>> d1 = Document('abc ')  >>> d2 = Document('xyz ')  >>> d3 = Document('abc ')  >>> d4 = Document('abc ')

>>> d5 = Document('xyz ') >>> d6 = Document('abc ') >>> d7 = Document('abc ') >>> d8 = Document('xyz ')

>>> Corpus = Corpus ((d1, d2, d3, d4, d5, d6, d7, d8))

>>> print Corpus.cluster(method=KMEANS, k=2)
```

Following are the clusters we got based on the out links found in the complete set of alternate queries which refer the main query's results.

C1 : [6,2,5,3,4,1,7] C2: [8]

Here the numbers 6,2,5,3,4,1,7 represents cluster 1 and 8 represents cluster 2 documents which consist of links which point to the main query's suggested links. Further, cosine similarity is calculated using Pattern as follows to find the relevancy.

```
>>> from pattern.vector import Document, Corpus

>>> d1 = Document(' ') >>> d2 = Document(' ') >>> corpus = Corpus(documents=[d1,d2])  >>> print corpus.similarity(d1,d2)
```

Based on the Cosine similarity we found out that the cluster C1 has relevant data when compared to C2. Table 3 gives the expected result. Due to the size of the clustered documents, we specified as just clusters of documents 6, 2, etc. The proposed algorithm out performs when the data is clustered.

ISBN 978-93-82338-91-8 © 2013 Bonfring

TABLE-3: Cluster formation and cosine similarity results

Cluster .No.	Document No. in clusters	Cosine Similarity
1	6,2,5,3,4,1,7	0.62630490768)
2	8	0.613493654634

6. CONCLUSIONS AND FUTURE WORK

We have presented a new approach for keyword query based graph clustering technique for efficient web page retrieval. The proposed system will retrieve only those pages or documents which are relevant to all the keywords of the query. The proposed system suggested less amount of documents compared to any other systems as result. Though the system clustered the documents well, it involves higher processing in scanning all the documents suggested for all the alternate queries, which in some cases may involve irrelevant alternate queries and the results. This problem could be eliminated if we count only the most relevant alternate queries for any given query using rough set. We would do something on this as our future work to reduce the target search space. Soft computing based clustering technique is the future scope of this work.

References

[1] Seung Kim, Wookey Lee , Nidhi R. Arora , Tae-Chang Jo , Suk-Ho Kang, "Retrieving keyworded subgraphs with graph ranking score," Expert Systems with Applications 39 (5), pp. 4647–4656, 2012.

[2] Rashid Ali, M.M. Sufyan Beg, "An overview of Web search evaluation methods," Journal of Computers and Electrical Engineering 37(6), 835–848, 2011.

[3] Dilip Kumar Sharma, A. K. Sharma, "A Comparative Analysis of Web Page Ranking Algorithms," International Journal on Computer Science and Engineering Vol. 02, No. 08, pp. 2670-2676, 2010.

[4] Cookhwan Kim , Sungsik Park , Kwiseok Kwon , Woojin Chang , " How to select search keywords for online advertising depending on consumer involvement: An empirical investigation," Expert Systems with Applications 39 (2012) 594–610.

[5] Sherif Sakr, Ghazi Al-Naymat,"Graph indexing and querying: a review," International Journal of Web Information Systems, Vol. 6 Issue: 2 pp. 101 – 120, 2010.

[6] Gaurav Bhalotia_,Arvind Hulgeriy,Charuta Nakhe, Soumen Chakrabarti, S. Sudarshan, "Keyword Searching and Browsing in Databases using BANKS," 18th International Conference on Data Engineering (ICDE.02), pp. 431 – 440, 2002.

[7] Hao He, Haixun Wang, Jun Yang and Philip S. Yu, "BLINKS: Ranked Keyword Searches on Graphs," ACM SIGMOD'07, pp. 305 – 316, 2007.

[8] M.F. Porter, 1980, "An algorithm for suffix stripping," Program, 14(3) pp. 130_137, 1980.

[9] K. Saravanakumar, K. Deepa, "Alternate Query Construction Agent for Improving Web Search Result using WordNet," International Conference on Computational Intelligence and Communication Systems, pp.117 - 120, 2011.

[10] Sakr S, Al-Naymat G., "Efficient relational techniques for processing graph queries," Journal Of Computer Science And Technology 25(6): 1237–1255 Nov. 2010.

[11] Nursel Yalçin, Utku Köse, "What is search engine optimization: SEO?" Procedia Social and Behavioral Sciences 9, pp. 487–493, 2010.

[12] Anusree Ramachandran, R.Sujatha, "Semantic search engine: A survey," International Journal of Computer Technology and Applications, Vol 2 (6), pp.1806-1811, 2011.

[13] A. M. Riad,Hamdy K. Elminir, Mohamed Abu ElSoud, Sahar. F. Sabbeh," PSSE: An Architecture For A Personalized Semantic Search Engine", International Journal on Advances in Information Sciences and Service Sciences Volume 2, Number 1, pp.102-112, March 2010.

[14] Hamada M.Zahera, Gamal F. El Hady, Waiel.F Abd El-Wahed, "Query Recommendation for Improving Search Engine Results," International Journal of Information Retrieval Research, volume 1 issue 1, pp. 45 – 52, 2011.

[15] Thanh Tran, Philipp Cimiano, Sebastian Rudolph, Rudi Studer, "Ontology-based Interpretation of Keywords for Semantic Search," 6th International Semantic Web and 2nd Asian conference on Asian Semantic Web Conference, pp. 523 – 536, 2007.

[16] Giansalvatore Mecca , Salvatore Raunich, Alessandro Pappalardo, "A new algorithm for clustering search results," Data & Knowledge Engineering 62 , pp. 504-522, 2007.

[17] De Smedt, T., Daelemans, W., "Pattern for Python," Journal of Machine Learning Research, 13, pp. 2031 – 2035, 2012.

ISBN 978-93-82338-91-8 © 2013 Bonfring