International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

World Wide Web Metasearch Using TF-IDF Method

S. P. Phadtare[1], S. B. Magdum[2]

[1]Sharad Institute of Technology Polytechnic, Yadrav, Near Jay- Sangli Naka, Ichalkaranji-416 115, Maharashtra, India
[2]Sharad Institute of Technology Polytechnic, Yadrav Near Jay- Sangli Naka, Ichalkaranji-416 115, Maharashtra, India

Abstract: ***As the storage capacity and the processing speed of search engine is growing to keep up with the constant expansion of the World Wide Web, the user is facing an increasing list of results for given query. A simple query composed of common words sometimes has thousands of results making it impossible for the user to verify all of them, in order to identify a particular site. Even when the list of results is presented to the user ordered by a rank, most of the time it is not sufficient support to help him identify the most relevant sites for his query. The concept of search result clustering was introduced as a solution to this situation. The process of clustering search results consists of building up thematically homogenous groups from the initial list results provided by classic search tools, and using up characteristics present within the initial results without any kind of predefined categories.***

Keywords: clustering, WWW, lexical analyze, stop word, stemming

1. Introduction

Text mining is relatively new research field whose main concern is to develop effective meaningful information with respective given purpose. there are many contexts where large amount of documents have to be managed, browsed, explored, categorized and organized in such a way that information we are looking for accessed in a fast and reliable way. Trying to keep up with continues growth of World Wide Web (WWW) the searching tools are engaged in a permanent race for ever faster development in order to reach better performance.

When the document storage reached considerable sizes the problem of better indexation was addressed. The bigger the storage capacity it becomes the more performing the indexing algorithm had to be in order to keep the web pages properly ordered. but the WWW was still growing with increasingly speed, so the crawler module had to be developed to reach higher speed in finding and downloading new pages and performing the clustering on that pages. Clustering search tools results means grouping them into classes which constructed using the search result characteristics. Vector space model is widely used document representation. It represents each document as a vector with one real valued component. The weighting representation model is based on TF-IDF method. When a query is submitted to a Metasearch engine, decisions are made with respect to underlying search engines to be used, what modifications will be made to the query, and how to score the results. These decisions are typically made by considering only the user's keyword query, neglecting the larger information need. User with specific needs, such as "research paper "or "homepages" are not able to express these needs in a way that affects the decisions made by the Metasearch engine. Users with different needs, but the same keyword query, may search different sub-search engines, have different modifications made to their query, and have results ordered differently. Metasearch is utilizing multiple other search systems to perform simultaneous search. A metasearch engine is a search system that enables metasearch. To perform a metasearch, user query is sent to multiple search engines; once the search results returned, they are received by the Metasearch Engine, then merged into a single ranked list and the ranked list is presented to the user. When a query is submitted to a Metasearch Engine, decisions are made with respect to underlying search engine to be used, what modifications will be made to the query and how to score the results. These decisions are typically made by considering only the user's keyword query, neglecting the large information need.

2. Figure

In figure 1 system provides the interface to the user. User enters the query. Query sends to the search tool like Google, Bing based on user requirement. Search tools provide the results and sends to the web search Clustering Model. Clustering model performs clusters. Clustering results send to the interface and interface sends result to user [1].

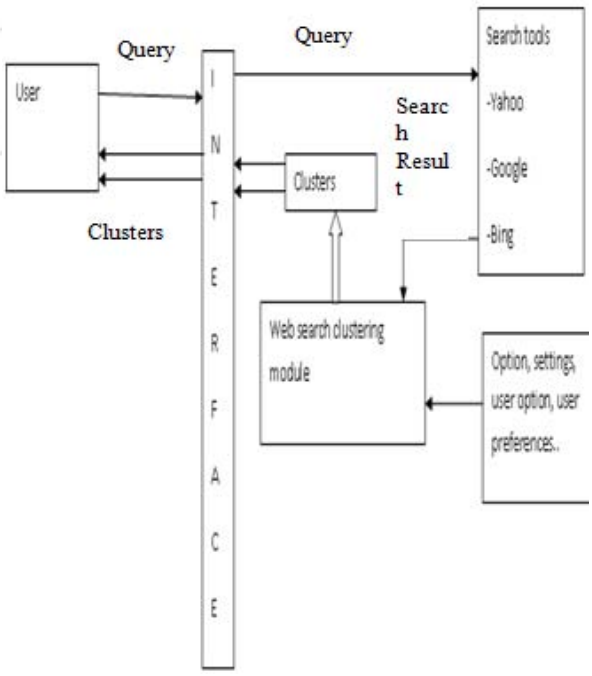

Figure 1: Metasearch Engine

3. Methodology

Clustering is search tool which performs grouping the results into classes which constructed using the search results characteristics [2]. The initial web is provided by a query. The results for that query are collected from various search engines. These results are provided in the form of snippets. The steps such as document preprocessing, vector representation, tag analyze, world root extraction, stop word removal are performed on those snippets. After pre-processing, using K-means method the no of clusters are obtained on the basis of maximum numbers of words occurred repeatedly in the snippets [2]. The cluster number, keyword and description of the query is store in database. This whole module implemented is based on the results collected from Google search engine. Further implementation of our search engine is based on collection of results from various other search engines such as Yahoo, Bing, etc

3.1. Steps in Web Search Result Clustering Process

a) Obtaining the web page list.
b) Document pre-processing.
c) Transforming the documents into vector representations.
d) Constructing cluster like representation for final results

a) Obtaining the Web Page List

The initial web page list is obtained by reuniting all results from all search from all tools used. The module must not over crowd the search tool's resource. In order to obtain the web page list the next steps have to be done.

i. Elimination of multiple links because the results are obtained from more search tools it's more than likely the same site will be returned by more than one tool.
ii. Duplicate web pages will receive the rank from the search tool with the lowest index.

b) Document pre-processing-

The clustering algorithm uses information from the pages in order to determine its subject or characteristics. Most document clustering algorithms use the whole document for this process, but such approach would slow our web page clustering algorithm too much.

c) Transforming the documents into vector representations

We need to transform each document into a vector. The vector will have the same size, turning our result list into an M*N matrix. Each line represents one web page and each column represents one word. The N dimension represents the total number of words that will proceed, from all documents. The dimensions' M represents the remaining number of web pages after the-

i. Tag cleaning eliminating portions of the web document which are strictly related to text formatting.
ii. Lexical analyze the purpose of the analysis is to identify distinct words. The process implies eliminating useless characters such as punctuation marks, comma, sometimes numbers, or characters like:#, $...
iii. Stop word elimination stop word is a word that does not have an informational value. In all languages there are a series of words which are considered stop words, e.g. "on", "and", "the", "in", etc.
iv. Establishing index words an index word, is a word that is representative in the context of the document.

3.2. Steps

i. Collecting results from two search engines, i.e. Google and Bing.[3]
ii. Performing operations such as stemming, stop word removal, Lexical analysis, Elimination of multiple links, duplicate web pages, word root extraction, tag cleaning, etc. on collected results individually.[1][2]
iii. Performing clusters of specific pages based on words appearing frequently in this page. The header of cluster is given name on basis of that word. The final result obtained is stored in database in tabular form which contains the tables namely search results and cluster headers using mysql server.[3]

4. Conclusion

In this paper we have presented an overview and some ranking strategies in Metasearch Engine. We also reported our study on how to merge the search results returned from multiple component search engines into a single ranked list; this is an important issue in Metasearch engine research. We investigated merging algorithms that utilize a wide range of information available for merging, from local ranks by component search engines, search engine scores, title and snippets of search result records to the full documents.

5. Future Enhancement

We have described a new metasearch engine architecture utilizes user preference information in deciding where to search, how to modify the queries, and how to order the results. This approach allows for much greater personalization and higher quality results than a regular metasearch engine, because of the ability to consider more than just the keyword query when making search decisions

References

[1] Architecture of Metasearch Engine that supports user information needs Eric J. Glover 1T2, Steve Lawrence', William P.Birmingham2, C. Lee Giles'
[2] World Wide Web Metasearch Clustering algorithm
[3] W.Meng, C.Yu, and K.Liu,"Building Efficient and Effective Metasearch Engine", In ACM Computing Surveys, 2002 and also cited inside the manuscript (example: page numbers, year of publication, publisher's name etc.).

Author Profile

Ms. Shubhangi Premchand Phadtare received the B.E. degrees in Computer Engineering from Sharad Institute of Technology College of Engineering, Yadrav, in 2014

Ms. Snehal B. Magdum received the B.E. degrees in Computer Engineering from J. J. Magdum College of Engineering, Jaysingpur, in 2013