Volume 3, Issue 6, June 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Design and Implementation of Search Engine Using Vector Space Model for Personalized Search

Mr.Ishwar.N.Bharambe, Prof. Richa.K. Makhijani
Computer Science &Engg.,SSGB,
Bhusawal & NMU, India

Abstract— In Vector Space Model we use the text-retrieval technique that is based on indexing keywords. In this paper, we design & implement Meta search engine using vector space model for personalized search. The search engine helps the machine to learn users' interest, so the personalized Meta search engine can help users to pick up important information for them faster. Vector Space Model helps us to implement personalized search engine, by remembering user's interest and displaying results or their interest on the beginning of the search links

Keywords— Meta-Search Engine (MSE), Vector Space Model (VSM), Uniform Resource Locator (URL), Search Engine (SE).

I. Introduction

The Internet can enable people to get the information more efficiently. This leads to information in enormous forms, and thus network information grows exponentially. A search engine is a web-based tool that enables users to locate information on the World Wide Web. As the Search Engines (SE) gives us lots of information so that SE is more essential in our life. Some Meta Search Engines (MSEs) use proxy log records for accessing user's pattern and store these patterns in the database. A relevance score is measured using some heuristic for each user and the URL that she/he visited. As the MSE gives more accurate information to the users, it is more popular in our day to day life. As all of us know that the meta-search engine can get more information from large sources, there is lots of information that users don't think about. This disadvantage turns to advantage. It makes user to use more time to deal with the information they are not interested in. Against the background, personalized meta-search engine is one way to solve the problem. The meaning of personalization is, search engine that can help users to sort important information for them by using user's interest. Search engine will get the users' interest at the beginning of the results, so it is very convenient for users to access useful information. In this paper we introduce the design and implementation of meta-search engine. We model the results and users' interest according to the Vector Space Model. They can put the users' interest information at the beginning of results, so the users can get the information rapidly.

The prime reason for the SEs indexes the pages is key-words inputted by the user. On the other hand, when user is searching the internet, we quite often may not know the correct and complete set of key words that might have led us to the desired URL. One needs to look into the semantics of the key words. The web search engines generally provide search results without consideration of user interests or context. We propose a personalized search approach that can easily extend a conventional search engine on the client side. This paper suggests a new approach that is based on some model which considers semantic aspects and uses them to implement a Meta-Search Engine.

II. LITERATURE SURVEY

A. *Meta-Search Engine*

It will examine the advantage and disadvantage of various approaches. There are three main directions for implementing Meta Search Engine:

1. Growth in user-interface
2. To sort the results of query
3. Consider the algorithms for indexing of web-page.

The more concentration on user requirements is recommended in the architecture of Meta-Search Engine. The Personalized Meta-Search Engine has been already proposed that provides quick response with re-ranked results after extracting user preference. It uses Naïve Bayesian Classifier for re-ranking. Some MSEs use proxy log records for accessing user's pattern and store these patterns in the database. A relevance score is measured using some heuristic for each user and the URL that she/he visited. A profile is maintained the user which contains currently visited most relevant URLs. Relevance of these URLs with their respective relative position is updated in profile when users visit those links further. Current research also suggests the framework of Meta-search engine based on Agent Technology. An enhanced version of open source Helios Meta-search engine takes input keywords along with specified context or language and gives refined results as per user's need. All the proposed solutions refine search-results up to some extent but they have a serious drawback which is that the user profile is not stationary from this it is observed that we need to consider alternative methods of re-ranking. This is provided by really statistical methods like Latent Semantic Analysis (it is also called as Latent Semantic Indexing) and the newly introduced Probabilistic Latent Semantic Analysis (it is also called Probabilistic Latent Semantic Indexing) which promises to give results that are more correct than those of Latent Semantic Analysis. Thus, the emergence of these algorithms and the need for robust meta- search engines.

Probabilistic Latent Semantic Analysis (PLSA) gives robust results for Information Retrieval when the task is to search the most relevant documents from a given corpus, for a given query. As both of these methods depend on the Vector Space Model, the Vector Space Model is explained prior to both.

B. Vector Space Model

The most of the text-retrieval techniques are based on indexing keywords. Since only keywords are unable to capturing the whole documents' content, they results poor retrieval performance. But indexing of keywords is still the most applicable way to process large corpora of text. After identification of the significant index term a document can be matched to a given query by Boolean Model or Statistical Model. Boolean Model applies a match that relies on the extent. Fig.1 represents of the Doc1 and Doc2 in space of three terms namely "Information", "Retrieval" and "System". Three are perpendicular dimensions for each term represents "**Term-Independence**". This independence can be of two types namely linguistic and statistical.

When the occurrence of a single term does not depend upon appearance of other term, it is called Statistical independence. In Linguistic independence; interpretation of a term does not rely on other any term an index term satisfies a Boolean expression while statistical properties are used to discover similarity between query and document in Statistical Model.

The statistically based "Vector Space Model" which is based on the theme of placing the documents in the n-dimensional space, where n is number of distinct terms or words (as- t_1, t_2…t_n) which constitutes the whole vocabulary of the corpus or text collection. Each dimension belongs to a particular term. Each document is considered as a vector as- D_1, D_2…D_r; where r is the total number of documents in corpora. Document Vector can be shown as following: $\mathbf{D}_r$={d_{1r},d_{2r},d_{3r},……..d_{nr}}.

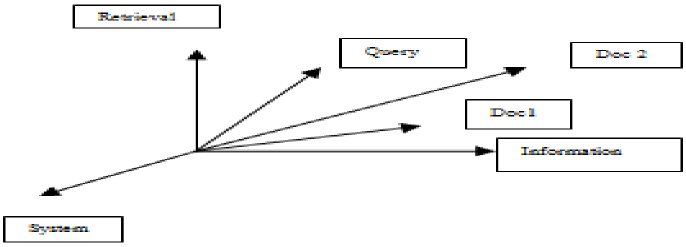


Fig. 1 Document Representation in term space

where d_{ir} is considered to be the i^{th} component of the vector representing the r^{th} document. There are various similarity measures that are proposed and one of them, that is very frequently used, is **Cosine Similarity**.

$$\text{Cos }\theta = \mathbf{Q} * \mathbf{D} / |\mathbf{Q}| * |\mathbf{D}|$$

The above expression represents the cosine of the angle between two vectors in the term space. The relevant document will be that one which is the nearest to given query. In the same way two documents would be considered relevant if they are in neighbour-hood region of each other.

The other measure are 1) Inner Product = $\Sigma\, Q_j * D_j$

2) Dice Coefficient = $2\, \Sigma\, Q_j * D_j / \{ \Sigma\, Q_j^2 + \Sigma\, D_j^2 \}$

3) Jaccard Coefficient =$\Sigma\, Q_j * D_j /\{\Sigma\, Q_j^2 +\Sigma\, D_j^2 -\Sigma\, Q_j * D_j\}$

Each component of document vector is always associated with some numeric-factor which is called weight of that respective term in document. This weight, w_i, can be replaced by term-count or term-frequency (tf_i). This assignment leads to another variation of the model that is called "**Term Count Model**".

III. PROPOSED SYSTEM

In the proposed system, we propose a content ontology to accommodate the extracted content and location concepts as well as the relationships among the concepts. We introduce different entropies to indicate the amount of concepts associated with a query and how much a user is interested in these concepts. With the entropies, we are able to estimate the effectiveness of personalization for different users and different queries.

A. *DESIGN*

This system consists of a JSP front with the composition of the background java program. The user interface using JSP production is used with the user interaction (Figure in step 1), the system obtains the keywords entered by the user. It then turns the query to the URL that can get results from Google (Figure in step 2). Then the page crawling module will search request processing module based on the module generated by the URL of the web pages to crawl (figure in step 4). Due to the page coming from different sources (respectively from Google), each page is independently analysed by engine. This is page by page analysis engine module to extract the key content, such as extracting the results of each of the page URL, title, and text descriptions. (Figure in step 6).

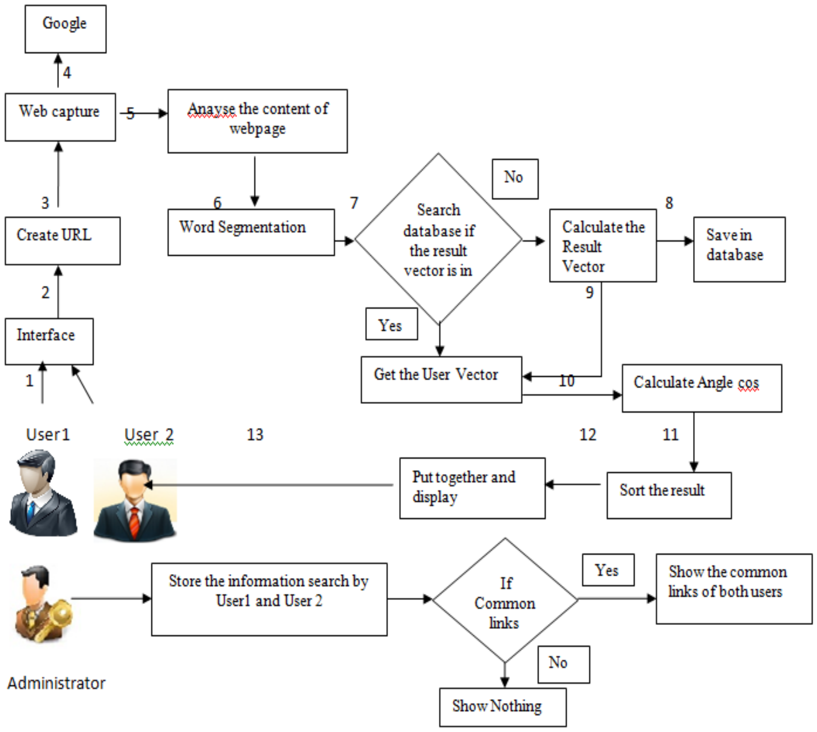


Fig. 2 System Architecture design

Then word segmentation results are achieved by the page analysis module (figure in step 7). Result modelling module will use the result of English word segmentation. Result modelling module will search for the database if it contains the result of URL and its vector. If it is not contains the result, the result modelling module will calculate the result's vector (the detail will show in the model module) and put the result into the database (Figure in step 8). Otherwise, the result modelling will use the vector directly (Figure in step 9). Then the system will get the user's interest vector, this vector will use to calculate cosines of angel between result vector and user vector (Figure in step 10). The system will use these values to sort the results and feedback to users (Figure in step 11, 12, 13). The architectural design of the personalized Meta search engine is shown in the fig. 2

B. *IMPLEMENTATION*

The users will login to the MSE. It will search the information which is necessary for him/her. When the administrator will log in to the to the personalized meta search engine then it will store the information of both the users such as date of searching the information and what they have searched when they were logged in to the meta search engine. And also store the common links and date of the searched common links that is visited by the users.

The personalized meta search engines don't require traversing the network, downloading web documents or building up an index. They mainly consist of member search engine selection, query forwarding, result integration and other algorithms. So, compared to robot based search engines or directory based search engines, the personalized Meta Search Engines have much lower technical doorsill and threshold in development and maintenance.

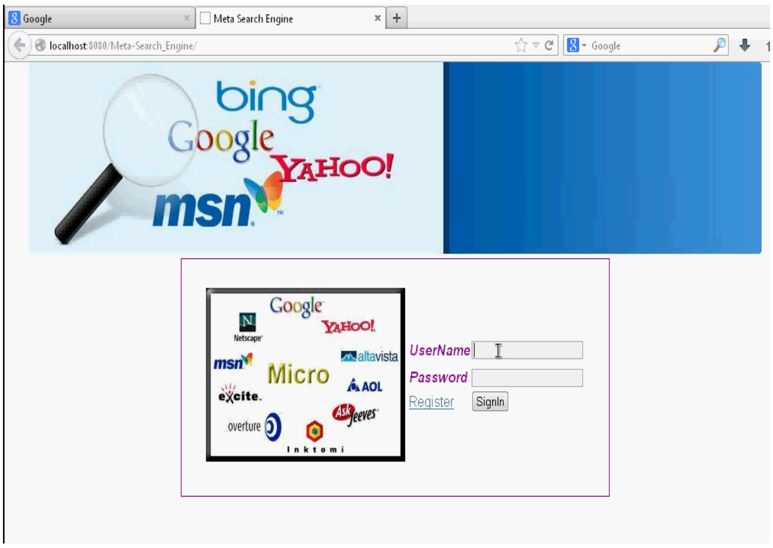
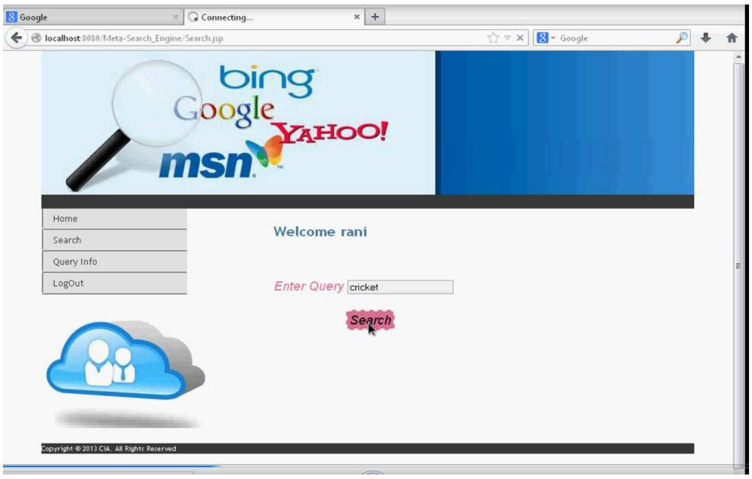


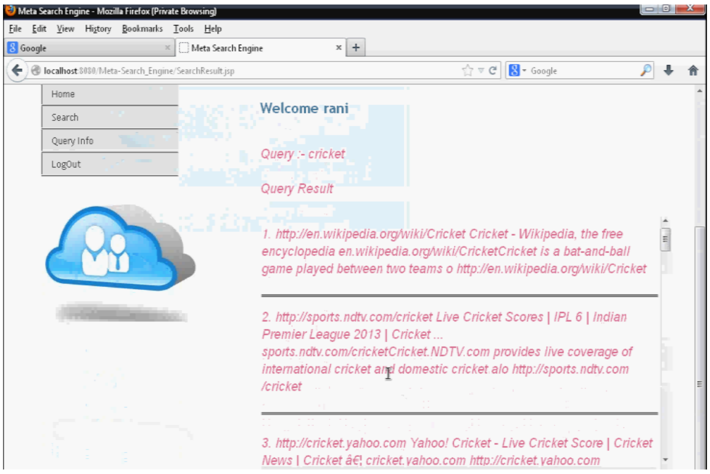
Fig .3 Users login page into Meta Search Engine


Google
Connecting...
localhost:8080/Meta-Search_Engine/Search.jsp
Google
bing
Google
YAHOO!
msn
Home
Search
Query Info
LogOut
Search
Copyright © 2013 CIA. All Rights Reserved

Welcome rani

Enter Query cricket

Fig. 4 User is searching the information from Meta Search Engine


Meta Search Engine - Mozilla Firefox (Private Browsing)
File Edit View History Bookmarks Tools Help
Google
Meta Search Engine
localhost:8080/Meta-Search_Engine/SearchResult.jsp
Google
Home
Search
Query Info
LogOut

Welcome rani

Query :- cricket

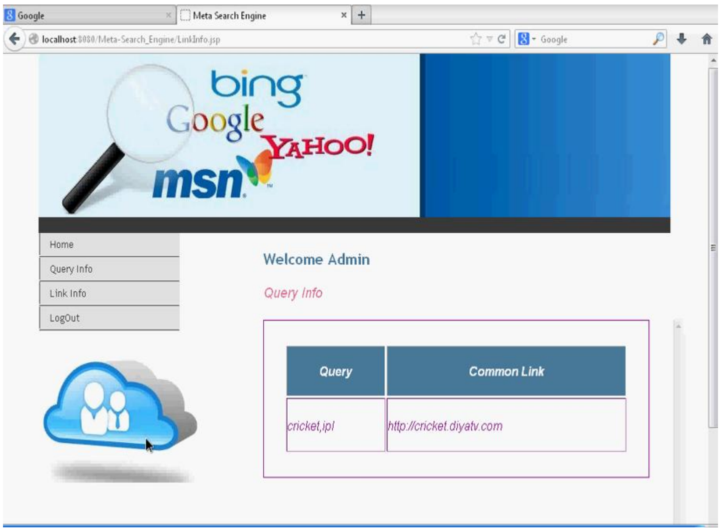
Query Result

1. http://en.wikipedia.org/wiki/Cricket Cricket - Wikipedia, the free encyclopedia en.wikipedia.org/wiki/CricketCricket is a bat-and-ball game played between two teams o http://en.wikipedia.org/wiki/Cricket

2. http://sports.ndtv.com/cricket Live Cricket Scores | IPL 6 | Indian Premier League 2013 | Cricket ... sports.ndtv.com/cricketCricket.NDTV.com provides live coverage of international cricket and domestic cricket alo http://sports.ndtv.com /cricket

3. http://cricket.yahoo.com Yahoo! Cricket - Live Cricket Score | Cricket News | Cricket â€¦ cricket.yahoo.com http://cricket.yahoo.com

Fig.5 Serached links of User



Welcome Admin

Query Info

Query	Common Link
cricket,ipl	http://cricket.diyatv.com

Fig. 6 Administration login with common links of Uesr Searched in MSE

Figure 3, 4, 5 and 6 flaunt the proposed framework in execution.

IV. CONCLUSION

The personalized search provide a common interface and conducts searches in many search engines simultaneously and return results in a uniform format. In present scenario search-engines are really useful devices to extract needed information from Internet. The personalized Meta-Search engines solve the same purpose with big span of coverage and advanced features like maintaining user's profile, filtering results etc. We Proposed MSE is based on refining the results using query expansion while next keywords are suggested by MSE itself without using any dictionary.

References

[1] I.N.Bharambe and R.K.Makhijani,*"Design of Search Engine using Vector Space Model for Personalized Search"*, International Journal of Computer Science and Mobile Computing(IJCSMC), ISSN 2320-088X, Vol.2,Issue. 3,March 2013,pg.63-66.

[2] Jiandong Cao, Yang Tang and Binbin Lou*," Personalized Meta-search Engine Design and Implementation"* Software College Northeast University (NEU) Shenyang, China, 978-1-4244-5540-9/10/$26.00 IEEE 2010.

[3] Abawajy, J.H.; Hu, M.J., *"A new Internet meta-search engine and implementation"*, The 3rd ACS/IEEE International Conference on Computer Systems and Applications, Page(s):103, 2005

[4] Shanmukha Rao, B.; Rao, S.V.; Sajith, G.; *"A user-profile assisted meta search engine",* TENCON 2003 Conference on Convergent Technologies for Asia-Pacific Region Volume 2, Page(s):713 - 717 , 15-17 Oct. 2003

[5] Spink, A.; Jansen, B.J.; Blakely, C.; Koshman, S.; *"Overlap Among Major Web Search engines"*, ITNG 2006 Third International Conference on Information Technology: New Generations, 2006. Page(s):370 – 374, 10-12 April 2006.

[6] A. Gulli, A. Signorini ,*"Building an opensource Meta-Search Engine"*, Special interest tracks and posters of the 14th international conference on World Wide Web WWW '05 ,ACM Press, May 2005.

[7] Zheng Li, Yuanqiong Wang ,Vincent Oria, "*A New Architecture to Web Meta-Search Engine"*, CIS Department ,New Jersey Institute of Tech., Seventh Americas Conference on Information Systems, 2001.