Asian Journal of Information Technology 15 (23): 4844-4850, 2016
ISSN: 1682-3915

Enhanced Filter Based Personalized Semantic Search

P. Perumal, M.S. Geetha Devasena and R. Ramya
Department of CSE, Sri Ramakrishna Engineering College, Coimbatore, India

Abstract: The web search engines are used to extract query specific information from this massive pool of World Wide Web. A large number of different search engines are available for the user to satisfy their needs. Every search engine uses its own specific algorithm to rank the list of web pages returned by the search engine for the users query, so that the most relevant page appears first in the list. Users think which search engine should be selected for searching corresponding data to any query topic for efficient search. For decision making on the basis of search result, users want to know whether they are significantly different or not. The internet contains vast amount of information that the search engines are able to provide search results that are based on page ranks. But the search results are not related to one particular user's environment. In this project, a new system called as enhanced filter based personalized semantic search which would be able to provide results for search query that relates to a particular user's environment, his area of interests, his likes and dislikes, the data the he/she might have found to be useful for him while searching for providing personalized search results. This process can be able to make applicable for each and every registered user in this application. User can give their basic information in their profile and get benefits from their each and every search. By this way the user can obtain results faster and more accurately. Once the search is complete one can bookmark the links of the search so that it is stored in one's profile so that it can be used for further reference one can also share the link to other profile by either sending the link as a mail or sharing with other users. Hence a more clear and informative search is done according to one's interest and domain.

Key words: Semantic, query, optimization, personalized search, domain

INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing into useful information. Data mining gives rise to many applications as given (Padhy *et al*., 2012). Web mining is the application of data mining techniques to discover patterns from the Web. Web mining is an emerging new area under the data mining. The application of data mining techniques is extract knowledge from Web data, in which at least one of the structure or usage (Web log) data is used in the mining process. The Web has huge amount of information. From the huge amount of information, mining process will help to find the information retrieval and user access pattern of the Web server. A more detailed study on web mining is cited in Pranit and Sheetal (2015). Web mining may be divided into three categories as shown in Fig. 1.

Web content mining: Web content mining essentially is an analog of data mining techniques for relational databases, since it is possible to find similar types of knowledge from the unstructured data residing in Web documents. Those efforts can be grouped into two sub categories namely.

Agent based approach: The agent approach uses so called Web agents to collect relevant information from the World Wide Web. There are three subtypes for the agent based approach: Intelligent Search Agents, Information Filtering/Categorization and Personalized Web Agents.

Database approach: The database approach for Web mining tries to develop techniques for organizing semi-structured data stored in the Web into more structured collections of information resources. Standard databases querying mechanisms and data mining techniques can be used to analyze those collections then. The database approach can be divided into two subtypes: Multilevel databases and web query systems.

Web structure mining: Most of the Web information retrieval tools only use the textual information, while ignore the link information that could be very valuable.

Corresponding Author: P. Perumal, Department of CSE, Sri Ramakrishna Engineering College, Coimbatore, India

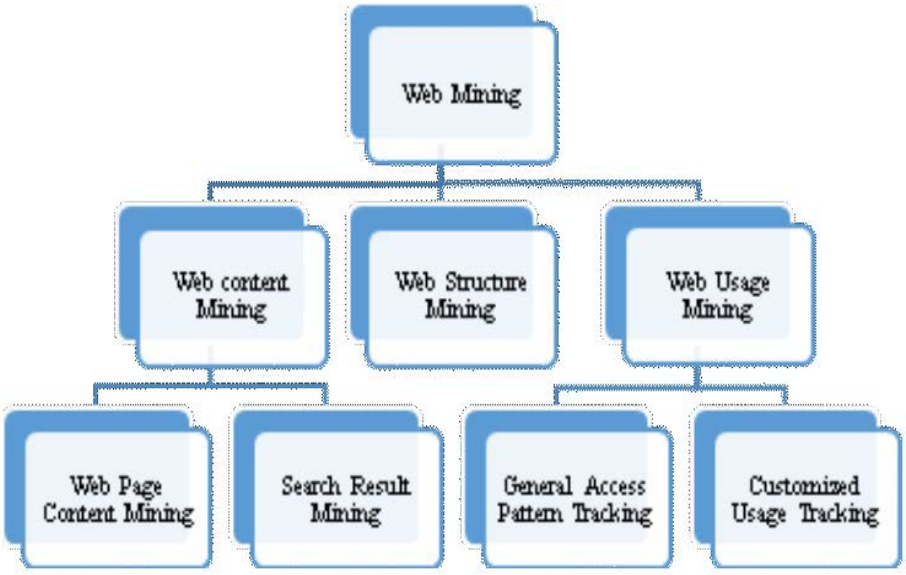

Fig. 1: Web mining classification

Technically, Web content mining mainly focuses on the structure of inner-document while web structure mining tries to discover the link structure of the hyperlinks at the inter-document level.

Web usage mining: Web usage mining tries to discover the useful information from the secondary data derived from the interactions of the users while surfing on the Web. The potential strategic aims in each domain into mining goal as: prediction of the user's behavior within the site, comparison between expected and actual Web site usage, adjustment of the Web site to the interests of its users.

Semantic search is the process of typing something into a search engine and getting more results than just those that feature the exact keyword you typed into the search box. Semantic search will take into account the context and meaning of your search terms. There are several mechanisms that use semantic search as cited in semantic engines (Madhu *et al*., 2011). It about understands the assumptions that the searcher is making when typing in that search query.

Literature review: Some of the most successful and elegant approaches to Web information retrieval are based on the realization of the importance of the link structure of the web. Recently, several methods for the personalization of web search engines have been proposed. The following projects based on web search engine personalization are surveyed to give better idea about the search engine practically.

A new concept in his study on best keyword search (Deng *et al*., 2015) about a baseline algorithm and Keyword-NNE. The baseline algorithm is inspired by the methods of Closest Keywords search which is based on exhaustively combining objects from different query keywords to generate candidate keyword covers. To attack this drawback, this work proposes a much more scalable algorithm called keyword nearest neighbor expansion (Keyword-NNE). Compared to the baseline algorithm, Keyword-NNE algorithm significantly reduces the number of candidate keyword covers generated. When the number of query keywords increases, the performance of the baseline algorithm drops dramatically as a result of massive candidate keyword covers generated. Denial of search attacks will not properly implement so that the hit ratio of the websites are not function correctly. Search engine produces more junk or advertisement websites from the top level of the search. A new technique for users in his paper "Effective Filtering of Query Results on Updated User Behavioral Profiles in Web Mining" (Deng *et al*., 2015; Sadesh and Suganthe, 2015).To improve the query result rate on dynamics of user behavior over time, Hamilton Filtered Regime Switching User Query Probability (HFRS-UQP) framework is proposed. HFRS-UQP framework is split into two processes where filtering and switching are carried out. The data mining based filtering in the research work uses the Hamilton Filtering framework to filter user result based on personalized information on automatic updated profiles

through search engine. Maximized result is fetched that is filtered out with respect to user behavior profiles mentioned in Priya and Sakthivel (2013), Farooqui *et al.* (2012) and Thakur *et al.* (2011). Experiment on factors such as personalized information search retrieval rate, filtering efficiency and precision ratio are to be measured. No user defined Login. Basic search engine should not produce a customized search.

A new method for Web data extraction with three phases in their study "An Implementation of Web Personalization using Web mining techniques" (Estelles *et al.*, 2010; Abdul *et al.*, 2014). In the first phase list of Web documents are selected, second phase documents are preprocessed, in the final phase results are presented to users. The main is to extract patterns based on user interest using a collection of Web documents by creating Web cube.

Further there are several other concepts related to semantic mining as sited in Zhang and Dong (2002) which deals with the efficiency of semantic search and searching techniques related to web mining. Some weaknesses such as browsing information without taking its meaning into account have recently appeared in Web Services. This creates a need for a new Web with more relevance to the user. Semantic Web is actually an extension of the current one in that it represents information more meaningfully for humans and computers alike (Ramulu *et al.*, 2012; Ashlesh and Shripadrao, 2015). It enables the description of contents and services in machine-readable form and enables annotating, discovering, publishing, advertising and composing services to be automated (Thakur and Pandey, 2012; Dwivedi *et al.*, 2013).

Proposed system: According to the proposed system the web crawler will be concentrated more Because of its tricky performance and reliability issues and even more importantly, there are social issues. Crawling is the most fragile application since it in Volves interacting with hundreds of thousands of web servers and various name servers which are all beyond the control of the system. Crawling has a more detailed study cited in which explains the crawling mechanisms and its implementations.

In order to scale to hundreds of millions of web pages, Google has a fast distributed crawling system. A single URL server serves lists of URLs to a number of crawlers. Because of the immense variation in web pages and servers, it is virtually impossible to test a crawler without running it on large part of the internet. Invariably, there are hundreds of obscure problems which may only occur on one page out of the whole web and cause the crawler to crash, or worse, cause unpredictable or incorrect behavior. Systems which access large parts of the internet need to be designed to be very robust and carefully tested.

The project aims to bring results for a search query that relates to a particular user's environment, location and type of data by analyzing the contents of a page, its source and the credibility of the results in response to a query. To make this search based filter in Google engine a license is required from Google. To obtain this license we have to pay a certain amount to gain that license. This license is obtained from the payment made at this link https://www.google.com/work/search/products/gss.html#pricing_content.Once the license to work on Google engine is done any changes can be made to it.This process can be able to make applicable for each and every registered users in this application. Users can give their basic information in their profile to get benefits from their each and every search.

In Fig 2, the user first searches for a query in the search engine. The algorithm checks whether the filter is set or not. If the filter is set it checks for the principal quantum keyword and compares with the database of the search engine. Then it checks for the rating of each link and displays the query results. The user can bookmark the link and it automatically gets registered in the user profile.

Keyword nearest neighbor expansion (Keyword- NNE): The algorithm Keyword-NNE selects one query keyword as principal query keyword. The objects associated with the principal query keyword are principal objects. For each principal object, the local best solution known as local best keyword cover (lbkc) is computed. Among them, the lbkc with the highest evaluation is the solution of BKC query. Given a principal object, its lbkc can be identified by simply retrieving a few nearby and highly rated objects in each non-principal query keyword. Compared to the baseline algorithm, the number of candidate keyword covers generated in keyword-NNE algorithm is significantly reduced. The in-depth analysis in paper and reveals that the number of candidate keyword covers further processed in keyword-NNE algorithm is optimal and each keyword candidate cover processing generates much less new candidate keyword covers than that in the baseline algorithm.

MATERIALS AND METHODS

Hamilton Filtering Regime Switching User Query Probability (HFRS-UQP): Filtering and switching that uses the Hamilton Filtering Regime Switching User Query Probability (HFRS-UQP) to filter user result based on personalized information on automatic updated profiles

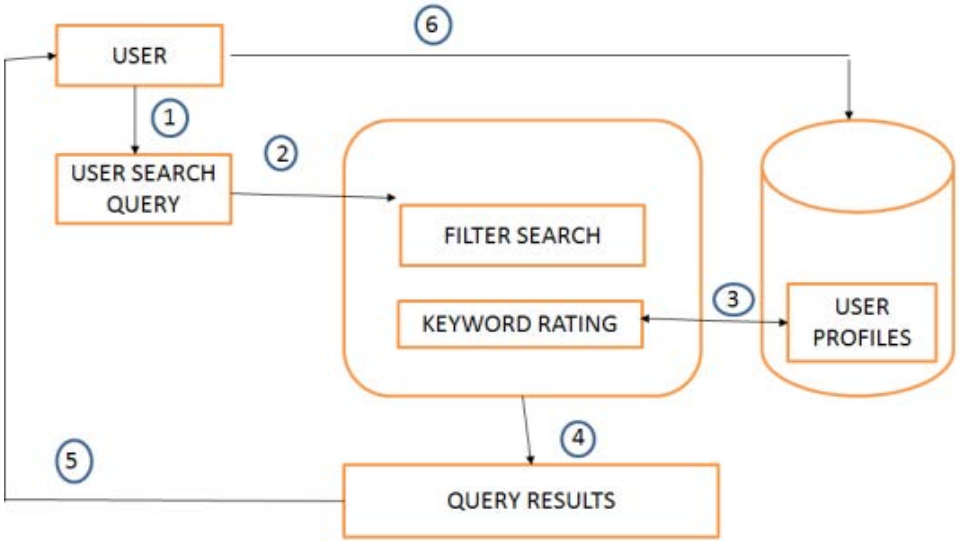

Fig. 2: System architecture

through search engine. The switching performs accurate filtering updated profiles using factors such as personalized information search retrieval rate, filtering efficiency and precision ratio. User results are filtered based on personalized information to help web client users to satisfy the web user specific needs.

Filter based personalized search: The essence of Semantic Search go beyond keyword search taking into account both user context and assumptions about the underlying meaning of data in an effort to return more relevant results and present them in a more appropriate way, always aiming for greater success in information-seeking endeavors. Semantic Search vendors do is focus on specific subject domains in order to provide targeted user experiences. The Search Engine updates is involving with admin process where the data center is available with huge number of data sets according to the user define search. The updating in profile change identifies the second- and higher-order association of query result on the updated profiles. A profile table structure is created in order to update the data which will be retrieved according to the user define search. A site will be updates in the data set with the prior http link, data content and hit ratio.

Filter based search engine: This module involves the administrator process where the data centre is available with huge number of data sets according to the user define search. A search is done based on filters such as type, location, area of interest , occupation and language The algorithm keyword nearest neighbor expansion (Keyword-NNE) and Hamilton Filtering Regime Switching User Query Probability (HFRS-UQP) is used to set the filters where users can interact with the search engine. Initially the search will be done according to the profile information of the user, an interface process is implemented here in order to get the relevance feedback information from the user define output.

The high performance of keyword-NNE algorithm is due to that each principal node (or object) only retrieves a few keyword-NNs in each non-principal query keyword. Basically a search engine will search according to the Meta data as well as according to the hit ratio. Here the database is designed according to the user defined profile so that retrieval will not be more complex and time consuming.

Prolific bookmark optimization: The Prolific Bookmark Optimization is an advanced search engine technique which will optimize according to the user internal search. There are several tools for bookmarks as mentioned in the paper bookmarking tools. But if in case of user searching ay of their related links of filter the searched links will be displayed at the top most searches from their next search. The user will be able to directly bookmark each link or use the save bookmark option to add a category and description. The user can also send the bookmarked link to others by sharing it or sending a mail to the respective person. These bookmarks are saved in the user profile with their respective dates, time and their domain so that it can be used in future reference. The bookmarked link can be sent to other users only those who have a registered profile.

Table 1: Process description of search engine

Process	Output dataflow	Description
Web crawling		
Start page, Queue of URL's to be Crawled	Temporary document files	Downloading the files of URL's as temporary files
Text extracting		
Temporary document files	A queue of URL's to be Crawled, Single Words of Each Document, URL Information(URL Entity)	It discards the HTML format in the document and extracts all the text words for each temporary file
Create inverted index		
Single words of each document	Inverted index	It creates inverted index
Ranking		
Inverted Index	Inverted index with rank	It ranks each word of inverted index
Search query analyzing		
Search query	Word of search query	It determines which method of search is adopted in the searching. They are keyword search, phase search, wildcard search and Boolean search. External user can also perform multi-method search. E.g. Boolean and wildcard search
Searching		
Words of search query	Search results	It searches for each portion of the query from the urlInfo and wordID tables
Search results ordering		
Search results	Final search results	It rearranges the order of search results by the ranking numbers (rank)

Process description: The process of how data flows in the search engine and what the methods and conditions the search engine handles are cited in the Table 1 as follows:

Pseudocode for filter based personalized semantic search

Algorithm:
Input: set of filter, query and database (F, S and D)
Output: Accurate Search Results
Condition: Profile Filter (F)
Step 1: Check filter option and analyze each search word.
Step 2: If filter (F) is set,
    Select the principal keyword
    Match the keyword (S) and filter (F) with the dataset (D)
    Filtering is performed by Log likelihood distribution function
Step 3: Otherwise
    Select the principal keyword (S)
    Results that match only the keyword are retrieved
Step 4: Ranking is done with the results by finding the semantic distance of each link
Step 4: DegreeOfmatch () for each link is obtained
Step 5: Ranking is placed in the tree using depth first search analysis
Step 6: The search result is retrieved and displayed as per the relevance of tree order.
The solutions can be achieved with the basis of matching degree, multi-level matching and the combination in different levels of range of the function degreeOfmatch ().

RESULTS AND DISCUSSION

Performance analysis: A detailed study on Google Search Engine in the papers and is made. This is compared with the proposed filter based engine is taken and analyzed. It shows that using filters one can get more accurate result at a much faster rate. The proposed work is analyzed using the SEO tool called Keyword Battle (http://www.webseoanalytics.com/members/seo-tools/keyword-competition-checker.php) which lists the rank of each URL link related to the search. It is also found that though one used filters or not there is a possibility that sometimes the search engine might miss certain URL even if its Google search engine.

Table 2: Presence of top 10 URL links

	Presence of No. of Top 10 URL	
Keyword (SW)	With filter	Without filter
1	10	9
2	10	8
3	10	9
4	9	9
5	10	10
6	9	10
7	8	8

After the analysis it is found that some links are missed in both the cases but the search with keyword have a minimum error case than the search without keyword proving to be better and more accurate. The analysis is based on both technical and non-technical. The technical search words are related to data structure, data mining and security, while non-technical involves music, books, rent and medicine. Here SW represents the search word that is given in the engine. Table 2 shows the efficiency of the proposed work.

Figure 3 specifies the relationship between the search word and the number of URL's. The performance of number of URL's with filter and number of URL's without filter is analyzed. Table 3 shows that one is able to get better results using the filters rather the search without using filters.

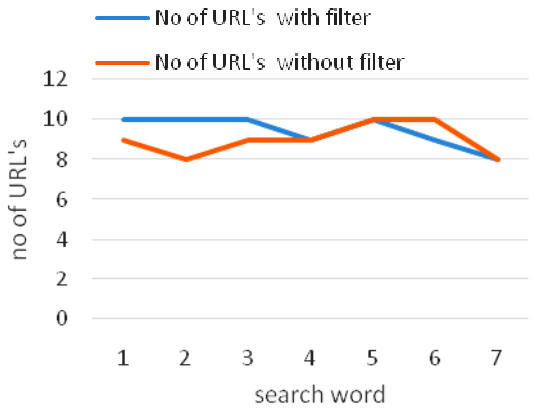

Fig. 3: Presence of top 10 URL links

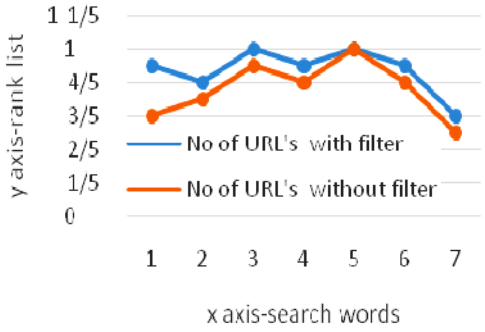

Fig. 4: Top 10 URL ranks

Table 3: Ranking of Best Match URL for top 10

	Best match URL		Accuracy of top 10 URL (%)	
Keyword	Ranking with filter	Ranking without filter	With filter	Without filter
SW1	9/10	6/10	90	60
SW2	8/10	7/10	80	70
SW3	1	9/10	100	90
SW4	9/10	8/10	90	80
SW5	1	1	100	100
SW6	9/10	8/10	90	80
SW7	6/10	5/10	60	50

Figure 4, for Table 3 is given below and it shows that the proposed work shows better results than the existing system.

The analysis of the percentage between the search engine with filter and without filter is shown below. Here, it is found that the search with filter gives an overall average of 87.17% accuracy whereas the search without filter gives an average of 75.71% accuracy.

Another analysis of the first top 10 ranking of both the search engines for the same keyword is taken and the list of URL obtained are tabulated as shown in Table 3. Thus, Fig. 5. the results show that the proposed research gives much more efficient results and saves more time than the existing work.

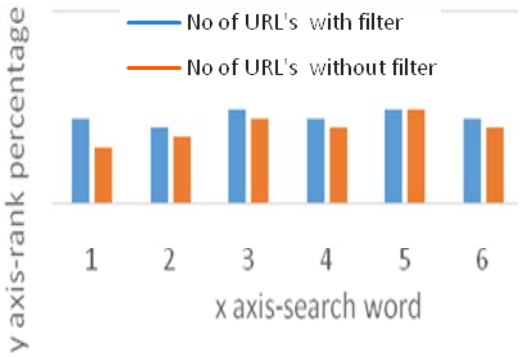

Fig. 5: Percentage of presence of top ranking URL's

CONCLUSION

The design and implementation, as well the analysis, of efficient and effective Web Search Engines (WSEs) are becoming more and more important as the size of the Web has continually kept growing. Furthermore, the development of systems for Web Information Retrieval represents a very challenging task whose complexity imposes the knowledge of several concepts coming from many different areas such as databases, parallel computing, artificial intelligence, statistics, etc. Here, the testing has been done in the local host successfully. An analysis of several efficient algorithms for computing approximations of optimal assignment for a collection of textual documents that effectively enhances the compressibility of the index built over the reordered collection.

REFERENCES

Abdul, L., L. Madiha, H. Muhammad, R. Imran and Y. Usman, 2014. The financial performance analysis of google Inc. V/S industry technology. Res. J. Finance Accounting, 5: 103-110.

Ashlesh, S.P. and B. Shripadrao, 2015. A survey on best keyword cover search. Int. J. Innovative Res. Comput. Commun. Eng., 3: 11834-11837.

Deng, K., X. Li, J. Lu and X. Zhou, 2015. Best keyword cover search. IEEE. Transac. Knowl. Data Eng., 27: 61-73.

Dwivedi, N., L. Joshi and N. Gupta, 2013. Statistical analysis of search engines (Google, Yahoo and Altavista) for their search result. Int. J. Comput. Theor. Eng., 5:298-301.

Estelles, E., E.D. Moral and F. Gonzalez, 2010. Social bookmarking tools as facilitators of learning and research collaborative processes: The Diigo case. Interdiscip. J. E. Learn. Objects, 6: 175-191.

Farooqui, M.F., M.R. Beg and M.Q. Rafiq, 2012. An extended model for effective migrating parallel web crawling with domain specific and incremental crawling. Int. J. Web Serv. Comput., 3: 85-93.

Madhu, G., D.A. Govardhan and D.T. Rajinikanth, 2011. Intelligent semantic web search engines: A brief survey. Int. J. Web Semantic Technol., Vol.2,

Padhy, N., P. Mishra and R. Panigrahi, 2012. The survey of data mining applications and feature scope. Int. J. Comput. Sci., Eng. Inform. Technol., 2: 43-58.

Pranit, B.D. D.M. Sheetal, 2015. Web data mining techniques and implementation for handling big. Int. J. Comput. Sci. Mob. Comput., 4: 330-334.

Priya, V.S. and S. Sakthivel, 2013. An implementation of web personalization using web mining techniques. Int. J. Comput. Sci. Mob. Comput., 2: 145-150.

Ramulu, V.S., C.N.S. Kumar and K.S. Reddy, 2012. A study of semantic web mining: Integrating domain knowledge into web mining. Int. J. Soft Comput. Eng., 2: 522-524.

Rastgoo, V., M.S. Hosseini and E. Kheirkhah, 2014. Semantic web-based software engineering by automated requirements ontology generation in SOA. Int. J. Web Semantic Technol., 5: 1-11.

Sadesh, S. and R.C. Suganthe, 2015. Effective filtering of query results on updated user behavioral profiles in web mining. Sci. World J., 2015: 1-8.

Thakur, M. and G.S. Pandey, 2012. Performance based novel techniques for semantic web mining. Int. J. Adv. Res. Comput. Commun. Eng., 9: 317-327.

Thakur, M., Y.K. Jain and G. Silakari, 2011. Query based personalization in semantic web mining. Int. J. Adv. Comput. Sci. Appl., 2: 117-123.

Zhang, D. and Y. Dong, 2002. A novel web usage mining approach for search engines. Comput. Netw, 39: 303-310.