A Survey on Techniques for Personalization of Web Search

Himani Arya
Department of Computer Science
MANIT, Bhopal (MP) India

Jaytrilok Choudhary
Department of Computer Science
MANIT, Bhopal (MP) India

Deepak Singh Tomar
Department of Computer Science
MANIT, Bhopal (MP) India

ABSTRACT
The World Wide Web (WWW) is growing exponentially per year thereby search engine provides the low quality of search results. Thus, the users get difficulty in getting the relevant information from the obtained search results. The quality of web search results depends on the information needs of the user and the searching techniques employed in the web search systems. So, the personalization is a general need in web search now-a-days. This paper includes the review of various approaches towards personalization. The approaches include hybrid profiling, personalized click model, ontology based user profiles and fuzzy theory for personalization.

Keywords
Personalization, web search, user profile, ontology.

1. INTRODUCTION
Internet is a global networking infrastructure that connects computers of many smaller networks. It transports the content to the linked computers all over the world. Traveling of data over the internet occurs in some format known as protocols like SMTP, HTTP etc. The content present on Internet in the form of hypertext that references to the other text or hypertext. This collection of documents around the world forms the web also known as World Wide Web (WWW). The Pages in the hypertext document are known as web pages and its location is specified by a uniform resource locator (URL). Thus, web can be treated as the software which allows the user to use the content available on the internet and write their own content.

The web pages on Internet are growing rapidly day-by-day. Thus, an information retrieval tool is needed to find the information form WWW. Web search engine, is an information retrieval tool, needed to look for these web pages. When the user fires a query to any search engine like google or bing, he/she may get even hundreds or more search results related to the query. Generally, theses are divided into a number of search engine results pages (SERPs). From the set of SERPs, user can decide for the link one should try to see if its referenced page contains the desired data. A huge amount of documents is available in web. So, it cannot identify the most important documents for a particular user for a specific query. Due to this reason, the need for personalizing web search arises to find out relevant information for users.

Personalization of web search is the process of customizing web search results based on users' past behavior [1]. Most of the queries submitted to search engines are short [2, 3] and have ambiguity [4, 5]. Every users may have different needs and goals under the same query. Thus the effectiveness of a personalization of web search depends on the query, user and search context [6].

2. NEED OF PERSONALIZATION
Generic Search Engines present the results which are general and not adaptable to individual users. For a particular query fired to the search engine, different results are provided for different users. Search results are organized for every user considering one's interest, preferences and information needs. The need for personalization arises due to the following facts: firstly, different users have different backgrounds and interests. For the same query, they have different information needs and goals. Secondly, User information needs may change over time. Users may have variety of requirements based on the time and circumstances. For example, a zoologist user may use query ''mouse'' to find information about computer peripheral when he/she wants to buy a computer mouse and a computer user may submit same query to find the information about the mouse as rodents while watching any animal tv channel. search engines can not to differentiate between such cases.

3. PERSONALIZATION APPROACH
When applied to search, personalization would involve the following steps: 1. To collect and represent information about the user in order to understand the user's interests. 2. Use this information to either filter the results returned from the initial retrieval process, or directly include this information into the search process itself to select personalized results [7].

Web search personalization systems use gathered information about user from profiles, cookies and to conduct and revise the search to maximize the user satisfaction [8]. The user profiles are created which specifies the user's interests, preferences and information needs to better personalize the search results. There are two ways to generate user profiles- explicit and implicit user profiling. In the explicit approach users create their profiles manually by providing some kind of feedback to a search system. In implicit user profiling, the user profile is created from user's past behaviour, such as by determining the documents they do select for viewing, the duration of time spent viewing a document or page browsing or scrolling actions [9]. This is being done in the background automatically by the search system.

Personalization of web search can be done at either server side or client side. Many problems arises on personalizing the web at server side like server should maintain all the search history for each and every user. It also has to search the history of a particular user when a user submits any ambiguous query. The performance of the server gets down when many users submits the query at the same time. Therefore, most of the techniques employ client side approach as all the search histories and queries are maintained at the client system making the faster way to access the user profile.

4. PERSONALIZED WEB SEARCH METHODS

Many attempts have been made to personalize the web search. personalized search strategies followed includes personalized search based on content analysis, hyperlink structure of the web and user groups.

4.1 Personalized Search Based on Content Analysis

In this approach, the content similarity between returned web pages and user profiles is calculated. The user profiles can be made by users themselves [10, 11] or can be learnt implicitly using user's historical activities. As the user is not always ready to provide their choices explicitly, so most of the work focuses on automatically collecting the preferences from past history. Under content analysis, user profiles can be built using two ways: topical categories [10, 12, 13] and keywords lists [14-18]. In topical categories, a user profile is framed as a hierarchy of concepts or topics. Previously issued queries and user selected documents are used to make concept hierarchy which further generates a user profile. In keywords lists, a list of keywords is used to show the user preferences. User profile is built as a vector of distinct terms and is made by collecting past user preferences both short term and long term preferences [17].

4.2 Personalized Search Based on Hyperlink Analysis

Generic search approaches rank documents depending on the link structure of the web. Thus, page rank algorithms are being used in web search. Page Rank laid emphasis on the fact that important pages are linked to/by many important pages. The PageRank of a page p is defined as the probability that the surfer visited page p. Personalized pagerank algorithm was proposed to personalize web search by page [19] which is the modified version of page rank used to re-rank the search results during personalization.

4.3 Personalized Search Based on User Group

In this approach, the community of like-minded users is formed. So, only the users are responsible to provide the information needed to form the user profiles. Search histories of users who have similar interests with the other user are used to refine the search results. Collaborative Filtering [17] [20] and CubeSVD [20] are some of the group based personalization methods.

5. RELATED WORK

Much research has been done on web search personalization. Various methods have been explored to understand the user's behavior.

5.1 Personalized Click Model through Collaborative Filtering

Shen et al. [21] proposed a personalized click model to explain the click preferences of users. These preferences applies and extends matrix/tensor factorization to connect users, queries and documents together. Click-through logs are used in search engines to learn user preferences for search results. Click prediction specifies the probability that a given document in a search-result list is clicked after a user enters some query. Click data optimizes a search engine performance with low cost but the problem with such data is the position bias. It implies that the document which is positioned higher on the web page will be getting more user clicks than the one placed lower and having more relevance. Personalized click model employs the collaborative Filtering method in it. Then three matrix or tensor factorization set ups are used.

5.1.1 Matrix Factorization Click Model

In previous click models, the position bias issues were considered but the complexity of document relevance has been largely neglected. Individual characteristics of queries and documents were not thoroughly considered. A query and a document has been simply treated as an integrated pair. A matrix factorization click model (MFCM) is proposed to focus on queries and documents interactions through their latent feature vectors. Suppose that a query q is submitted in a session and N documents are fetched, the i-th document being d_i. If there are in total M_q queries and M_d documents, then let $Q \subset R^{FxM_q}$, where the set consists of all possible real-valued matrices with size $F \times M_q$, and $D \subset R^{F*M_d}$ represents the latent factors of queries and documents, respectively. F is the number of factors.

5.1.2 Tensor Factorization Click Model

This introduces a personalised click model which extends MFCM to the user domain to includev Personalization. Suppose that there are M_u users. Let $U \subset R^{F*M_u}$denote the latent factors of user domain. The event of user being personally interested in the i-th document is indicated by N_i. Let α_{uqd_i} denote the probability of event N_i.

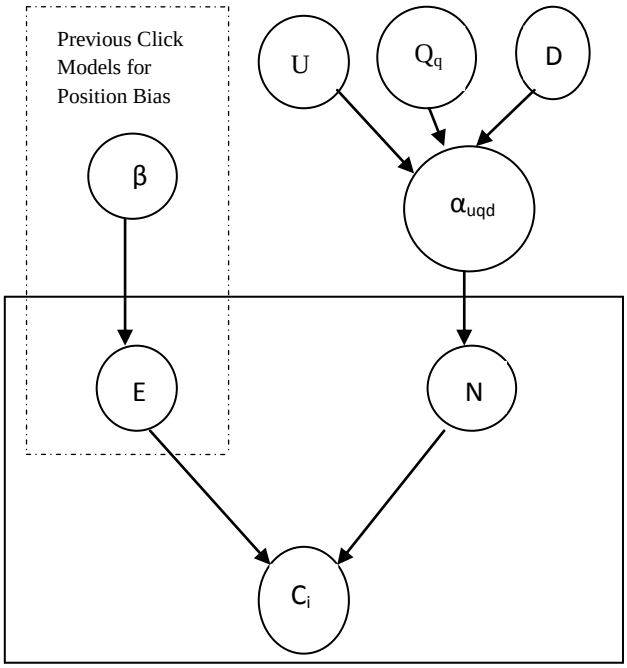

Fig 1: Graphical representation of Personalized Click Model

It extends matrix factorisation to tensor decomposition. The event of click at position i is dependent on the event of user examination and an individual document interest from user u. PCM considers the implicit interactions among users, queries and documents.

5.1.3 Hybrid Personalized Model

HPCM is a combination of PCM and MFCM. It laid emphasis on the interactions between queries and documents, which have been believed to be the dominant part for relevance determination. Then to solve the problem of personalization, only the residuals are factorized using user latent factors to describe personal deviations from the global query-document factor model. The interactions between queries and documents can be viewed as a relevance bias, while the user-query-document relationship may be viewed as user preference variations.

The models are capable enough to handle all kinds of queries including informational queries. MFCM captures latent feature vectors of queries and documents well. PCM enhances the capability of personalization. HPCM achieves the more improvement by combining the strengths of previous two models

5.2 Web Search Personalization: A Fuzzy Adaptive Approach

Mohammad. S Norouzzadeh et al. [22] proposed a personalized adaptive search, client-side approach for the personalization of web search which adapts the results based on user interests. Adaptive means adjusting search results by user's feedback. User's feedbacks can be considered in two types including explicit or implicit. The proposed approach uses implicit feedbacks model to capture user's interests. In this, Rocchio formulation [23] is adjusted to consider membership grade of relevance for each document. In the approach a query vector is associated with each query. The Rocchio formulation [24] retrieves relevant documents, but it cannot prevent some irrelevant results for the search. This query vector is enhanced with user's feedback. Subsequently, this query is adjusted to distinguish relevant documents from irrelevant ones.

A fuzzy variable is also defined to clarify the relevance of each document. Each document can be labelled as not relevant, not many relevant, somewhat relevant, very relevant or etc. To measure the value of fuzzy variable, the abstracts of each document which the search engines has returned and the behaviour of users are considered. Fuzzy set determines the relevance of the document. To calculate the relevance of each document, abstract terms are used and considered all abstracts relatively. Each abstract has some common terms with the query vector, the number of these common terms are counted. The ratio of document relevance –value of fuzzy variable – is calculated by the number of common terms with the query vector.

This is a client side computation provides privacy for users. Using abstracts of the documents reduces the computation time than utilizing full documents. Use of explicit user's feedback would be more helpful and progressive to the approach

5.3 Generation of Ontology Based User Profiles for Personalized Web Search

Jayanthi et al. [25] proposed an approach to create an accurate, ontology-based user profile without the user interaction. Ontology is an explicit specification of concepts and relationships that can exist between them. User profiles are often represented by keyword/concept vectors or a weighted concept hierarchy built. Implicit methods for creating an ontological user profiles are used as user's interests changes over time. For the evaluation of personal ontology annotations like an interest score, are utilised. After annotating each concept with a weight based on an accumulated similarity score, a user profile is created consisting of all concepts with nonzero weights. Using ontology as the basis of the profile allows the initial user behaviour to be matched with existing concepts in the domain ontology and relationships between these concepts.

In the approach, all sites are browsed using the user's own ontology rather than system supplied reference ontology. For the personal ontology, the sample documents are provided by the user i.e. from the browsing history. To create a personal ontology many factors are considered like the term, term frequency, URL's visited, downloads done on a page and the time spent on a page. A profile construction algorithm is also used. The personal browsing system needs to map from reference ontology concepts to the best matching concept in the personal ontology based on query weights and file sizes. The match value between each concept in the reference ontology and the concepts in the personal ontology is calculated. The goal of the mapping phase is to map every concept in the reference ontology to a concept in the personal ontology and also to map related websites to give more personalized search results.

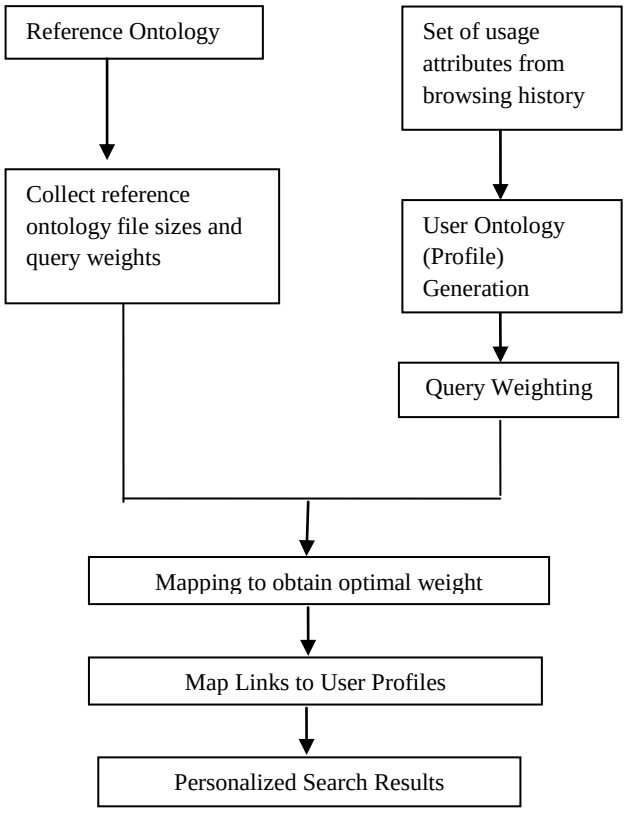

Fig 2: System Architecture

The documents for each concept were merged to create a collection D containing one super document per concept. The super documents were pre-processed to remove stop words and HTML tag. Finally each word is reduced to its root decreasing the dimensionality of the vectors used to represent each concept. The algorithm for creating user profiles consist of three phases, the initialization phase, where the terms extracted and an empty ontology is initialized. The first stage is to create a full ontology from narrower term relations. The full ontology is then pruned by eliminating unnecessary relations in the second stage.

Using this methodology, only the profiles of users who provide a significant enough number of documents are built.

5.4 Hybrid Profiling in Information Retrieval

Mandeep Pannu et al.[26] presented a paper in which a hybrid user profiling system is proposed where both explicit and implicit user profiles are considered to improve the web search effectiveness in terms of precision and recall. The system is content-based and implements the Vector Space Model (VSM). The similarity between user profiles and documents, storing and combining explicit and implicit user profiles and filtering process is handled by VSM. It represents the document and a profile by term vectors. Each dimension of a vector represents a keyword and the vector's value in that dimension determines the importance/weight of that word. Weight values can be applied to each term in document representations, the user query and the profile. The longer documents are not favoured over the short ones and rare terms are given more preference than popular terms. Documents are filtered as the similarity between user profiles and document representation is determined.

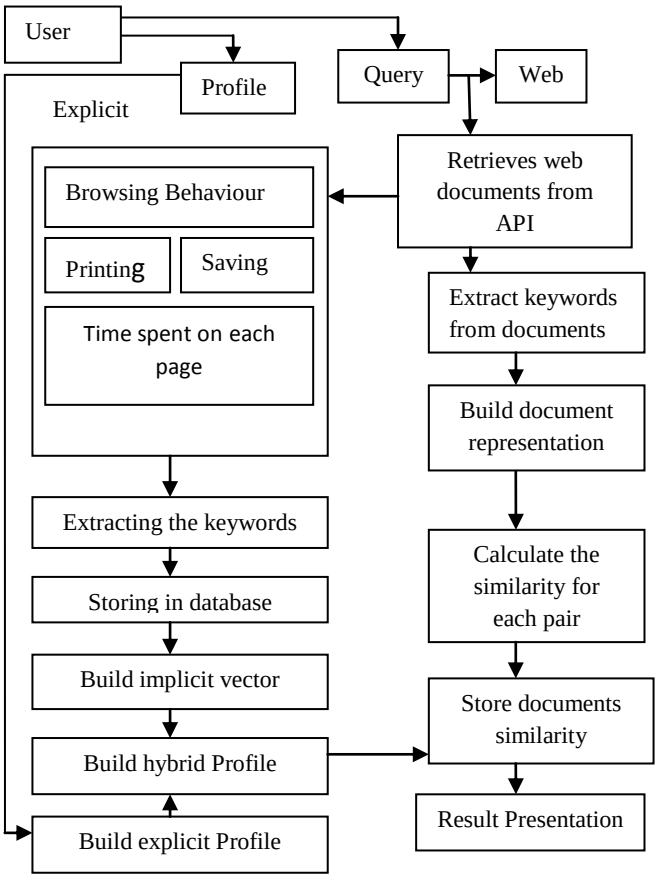

Fig 3: Hybrid System Architecture

The user profile generation occurs explicitly or implicitly. The explicit user profile is created by asking the users themselves. It is stored in a term vector for future use. For implicit user profile, user's activity is monitored constantly. The keywords are extracted from every visited document and given different weights depending on their position within the document. The system also stores the activity type and the time of the event. When the implicit vector is generated, weight should be applied to the representation of that document. Hence, implicit user profiles in form of keywords and weights are formed. In hybrid system, the implicit and explicit user profile vectors generated separately are combined into a single term vector. So, both the vectors are scaled so that weight of explicitly entered keyword is equal to the highest weight of any keyword from implicit profile and then both vectors are added. The combined vector is then normalised by the system and used for searching.

Hybrid profiling enhances the search system performance in terms of precision and recall. But if the interests of the user changes, then it does not prove to be much beneficial as the new profiles have to be created.

5.5 Web Search Personalization using data mining

Mangesh Bedekar et al. [27] presented a system to improve the relevant searches for the experienced user using their profile. Principal elements of web personalization include modeling of web objects and subjects, how the objects and subjects are categorised and performing the match among the objects and/or, determine the set of actions that are recommended to be performed for personalization.

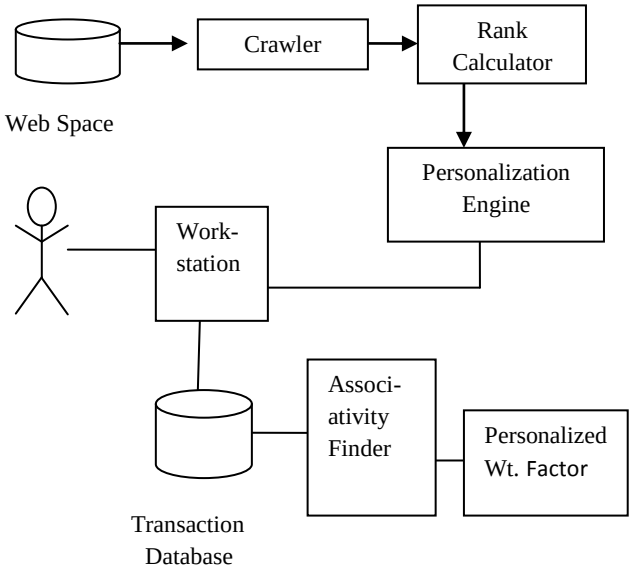

Fig 4: System Architecture

The approach is shown by the architecture using the data mining techniques to make the personalization automatic and dynamic. It was proposed for a single user. For faster ordering of data, all elements are implemented on the client side except Crawler and the Rank Calculator. It includes the following elements :

5.5.1 Crawler

It crawls the web to capture the database. The crawling approach is work based on content and link structure analysis [28]. The information stored reflects the link structure of the web. The crawler is implemented using PHP and MySQL database as backend.

5.5.2 Page Rank calculator

It calculates the page rank factor. It also includes the Personalization factor which is given by personalization engine that makes the rank calculated vary for every user or for every cluster of users.

5.5.3 Personalization Engine

It calculates the personalization factor 'Q' for the user on the client side and then pass this to page rank calculator. A sequence of Data Mining Algorithms is used to obtain the user specific, query specific weight factor.

It is a module that implements the personalization using log-like table file and combines the page rank and the weight factor by using a weighted formula as below

Personalized PageRank = PageRank + WeightFactor

5.5.4 Association Finder

Implements the Apriori Algorithm and calculates the association rules among the query and the URL. The obtained confidence is considered to be the weight factor for that URL. Currently, the user behavior is represented by the past<QUERY, URL> pair.

5.5.5 Data Storage

It was planned to use MySQL database. The user profile information is used to extract rules and calculate the personalization factor. The associativity can be easily extended by adding certain other columns to enhance the effectiveness of the personalization in the future.

5.5.6 User Interface

A simple user interface is there in which the query is keyed, the results are displayed based on Page Rank, and this is how we take the input and calls all other functions to re-order the results.

6. CONCLUSIONS

The WWW is growing day by day. Thus, it is very difficult for user to find the relevant search results from the list of search results returned by search engine. Personalization of web search is a necessity now-a-days to reveal user preferences in search results. In this paper, a survey of personalization has been given. The maximum number of web personalization methods are based on web page content analysis and textual similarity with the user preferences. Some approaches uses ontological user profile for personalization. In all the approaches, it is unclear whether personalization is working efficiently and effectively for all users in different search context.

7. REFERENCES

[1] Eirinaki, M., Mavroeidis, D., Tsatsaronis, G. and Vazirgiannis, M. 2006. Introduction to Semantics in Web Personalization: the Role of Ontologies. EWMF/KDO5.

[2] Silverstein, C., Marais, H., Henzinger, M. and Moricz, M. 2009. Analysis of a very large web search engine query log .SIGIR Forum, 33(1):6–12.

[3] Jansen, B. J., Spink, A. and Saracevic, T. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. Information Processing and Management, 36(2):207–227.

[4] Krovetz, R. and Croft, W. B. 1992. Lexical ambiguity and information retrieval. Information Systems, 10(2):115–141.

[5] Cronen-Townsend, S. and Croft, W.B. 2002. Quantifying query ambiguity. In Proceedings of HLT '02, pages 94–98.

[6] Elbassuoni, S., Luxenburger, J. and Weikum, G. 2007. Adaptive Personalization of Web Search, WISI Workshop.

[7] Susan Gauch, Jason Chaffee, Alexander Pretschner. 2003. Ontology-Based User Profiles for Search and Browsing. The OBIWAN Project.

[8] Feng Qiu, Junghoo Cho. 2006. Automatic identification of user interest for personalized search. WWW: 727-736.

[9] Maimon, Oded; Rokach, Lior (Eds.). 2008. Soft Computing for Knowledge Discovery and Data Mining. Hardcover Springer Science and Business Media, Inc. XIV, 434 p. 74 illus.

[10] Pretschner, A. and Gauch, S. 1999. Ontology Based Personalized Search. Proc. 11th IEEE Int'l Conf. Tools with Artificial Intelligence (ICTAI '99), pp. 391-398.

[11] Chirita, P.-A., Nejdl, W., Paiu, R. and Kohlschu C.¨ tter. 2005. Using ODP Metadata to Personalize Search. Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '05), pp. 178-185.

[12] Liu, F., Yu, C. and Meng, W. .2002. Personalized Web Search by Mapping User Queries to Categories. Proc. ACM Int'l Conf. Information and Knowledge Management (CIKM '02), pp. 558-565.

[13] Liu, F., Yu, C. and Meng, W. Jan. 2004. Personalized Web Search for Improving Retrieval Effectiveness. IEEE Trans. Knowledge and Data Eng., vol. 16, no. 1, pp. 28-40.

[14] Shen, X., Tan, B. and Zhai, C. 2005. Implicit User Modeling for Personalized Search Proc. ACM Int'l Conf. Information and Knowledge Management (CIKM '05), pp. 824-831.

[15] Tan, B., Shen, X. and Zhai, C. 2006. Mining Long-Term Search History to Improve Search Accuracy. Proc. 12fth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '06), pp. 718-723.

[16] Teevan, J., Dumais, S.T., and Horvitz, E. 2005. Personalizing Search via Automated Analysis of Interests and Activities. Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '05), pp. 449-456.

[17] Sugiyama, K., Hatano, K. and Yoshikawa, M. 2004. Adaptive Web Search Based on User Profile Constructed without Any Effort from Users. Proc. 13th Int'l World Wide Web Conf. (WWW '04), pp. 675-684.

[18] Chirita, P.A., Firan, C. and Nejdl, W. 2006. Summarizing Local Context to Personalize Global Web Search. Proc. ACM Int'l Conf. Information and Knowledge Management (CIKM).

[19] Page, L., Brin, S., Motwani, R. and Winograd, T.1998 .The PageRank Citation Ranking: Bringing Order to the Web.Technical report, Computer Science Dept., Stanford Univ.

[20] Sun, J.-T., Zeng, H.-J., Liu, H., Lu, Y. and Chen, Z. 2005. CubeSVD: A Novel Approach to Personalized

Web Search. Proc. 14th Int'l World Wide Web Conf. (WWW '05), pp. 382-390.

[21] Shen, S., Hu, B., Chen, W. and Yang, Q. 2012. Personalized Click Model through Collaborative Filtering. *WSDM'12,* February 8–12, Seattle, Washington, USA.

[22] Mohammad. S Norouzzadeh, Ayoub Bagheri, Mohammad. H Saraee. 2009. Web Search Personalization: A Fuzzy Adaptive Approach. IEEE.

[23] Jackson, P., Moulinier, I. 2002. Natural Language Processing for Online Applications, John Benjamins Publishing Company Amsterdam /Philadelphia.

[24] Menczer, F. 2003. Complementing search engines with online web mining agents Decision Support Systems Volume 35 , Issue2 (May)

[25] Jayanthi .J, Jayakumar, K.S., Surendran, S. 2011. Generation of Ontology Based User Profiles for Personalized Web Search. IEEE

[26] Pannu, M., Anane, R. and James, A. 2013. Hybrid Profiling in Information Retrieval. Proceedings of the IEEE 17th International Conference on Computer Supported Cooperative Work in Design.

[27] Bedekar, M., Deshpande, B. and Joshi, R. 2008. Web Search Personalization by User Profiling. First International Conference on Emerging Trends in Engineering and Technology, IEEE.

[28] Pal, A., Tomar, D.S., and Shrivastava S.C. 2009, Effective focused crawling based on content and link structure analysis. International Journal of Computer Science and Information Security, Vol. 2, No. 1, June 2009