International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
International Conference and Workshop on Communication, Computing and Virtualization (ICWCCV 2015) – www.ijais.org

Personalized Web Search in Web Mining

Sheetal Vishwas Kashid
Pursuing ME
Department of Computer Engineering
Shah and Anchor Kutchhi Engineering College
Mumbai, India

Vishwakarma Pinki
Assistant professor
Department of Computer Engineering
Shah and Anchor Kutchhi Engineering College
Mumbai, India

ABSTRACT
Web is a collection of inter-related files on one or more Web servers. Web mining is the application of data mining techniques to extract knowledge from Web data. Web mining involves the process to apply data mining techniques to extract and uncover knowledge from web documents and services. Web mining has been explored to a vast level and different techniques have been defined for variety of applications like Web Search, Web Classification and Personalization etc. Personalized search is a way to improve the accuracy of web search. However personalized search requires collection and aggregation of user information, which often raise serious problems of breach of confidential data for many users. So, personalized web search in web mining is one of the techniques to improve privacy concerns of the user's data.

Keywords
Personalized web search, web mining, user profile, information retrieval.

1. INTRODUCTION
1.1 Effect Of World Wide Web
Web mining is a very broad research area emerging to solve the issues that arise due to the WWW phenomenon. Various web search engines such as Google, Yahoo etc have made enormous contributions to the web and society. They have solved the problems of the internet users to find information on the web, quickly and easily. With the drastic growth of information available over the Internet, World Wide Web (www) has become main platform for storing and retrieving the information (e.g. cloud computing which provides huge storage space to store private data and user can retrieve his data as and when required) as well as mine the useful knowledge. Every day, WWW serves a platform for exchanging various kinds of information, ranging from research papers, and educational content, to multimedia content and software. Issues that have to be dealt with it are the detection of relevant information, involving the searching and indexing of the Web content, as well as the addressing of the individual users' needs and interests, by personalizing the provided information and services.

Web information
Web information is the data to be collected and used in the context of Web personalization. These data are classified in four categories:

a. ***Content data*** is the structured data that is represented to the end-user. It can be simple text, images, or structured data, such as information retrieved from databases.

b. ***Structure data*** represent the way the content is organized. They can be either data entities used within a Web page, such as HTML or XML tags, or data entities used to put a Web site together, such as hyperlinks connecting one page to another.

c. ***Usage data*** represents a Web site's usage, such as a visitor's IP address, time and date of access, complete path (files or directories) accessed and other attributes that can be included in a Web access log.

d. ***User profile data*** provide information about the users of a Web. A user profile contains demographic information for each user of a Web, as well as information about users' interests and preferences. Such information is acquired through registration forms or feedbacks, or can be analyzed by Web usage logs.

These above categories are the techniques of web mining which we will discuss in next topic.

2. CATEGORIES OF WEB MINING
Web mining is the use of data mining techniques to make the web more useful and more profitable (for some) so as to increase user's interaction with the web efficiently.Three categories of web mining include:

a) **Web Content Mining** Web Content Mining is the process in which user extract useful information like text, images or multimedia from the contents of Web documents. Web content mining changes this retrieved information (content) available on the Web into more structured forms as well as its indexing for easy tracking of information locations. Web content may be unstructured (plain text), semi structured (HTML documents), or structured (extracted from databases into dynamic Web pages).This type of web mining also called as web text mining because text content is more popularly researched.

b) **Web Structure Mining**: The structure of a typical Web graph includes web pages as nodes, and hyperlinks as edges connecting between two or more related web pages. It uses the process of graph theory. Thus, Web Structure Mining can be defined as the process of discovering structure information from the Web. This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level. Web structure mining can be used by search engines such as Google search engines. Web structure mining divided into two kinds, according to type of web structural data:

a) Extracting patterns from hyperlinks in the web.
b) Mining the document structure.

c) **Web Log Mining**: Web Log Mining is the application of data mining techniques to discover usage patterns from web data, in order to understand and serve the needs of web based applications. Log data captures the identity of web users along with their browsing behavior. Web Log mining is the process of identifying browsing patterns by analyzing the user's browsing behavior. This information takes as input the usage data, i.e. the data residing in the web server logs, recording the visits of the users to a web site. Web usage mining applications include: Improving site design, targeting potential customers for E-Commerce, enhancing the quality and delivery of internet information to the end user. Figure 1 shows web mining techniques.

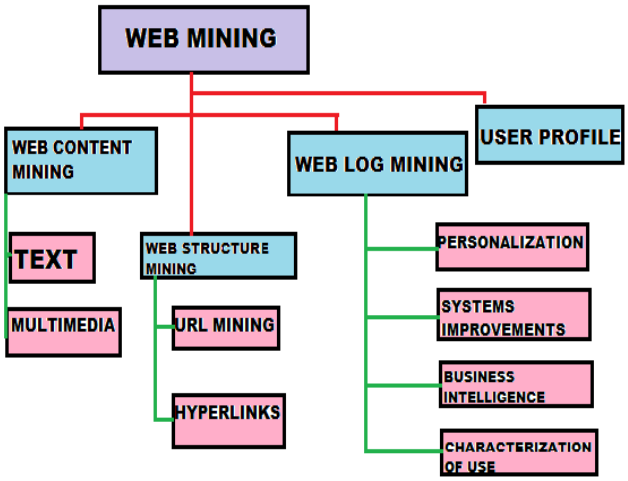

Figure 1: Web Mining Techniques

3. WEB MINING PROCEDURE

Web mining procedure is divided into four stages:

a) **Stage 1 Collection of source data**: Direct source of data in web mining is web log files that are stored on web servers. Web log files record the browsing behavior of the user which includes timestamp of the user when connected to web, queries given by user and click-through data etc.

b) **Stage 2 Data preprocessing: C**ollected data from web may be ambiguous data (data having same name but different meaning), incomplete data (missing contact details) or noisy data (incorrect pin code of a city) or redundant data. Preprocessing provide short and accurate data for data mining. The technique used is to clean web server logs to eliminate unrelated data. It includes data cleaning (removing unnecessary data from web server logs), user identification (by his IP address) and user session identification.

c) **Stage 3 Knowledge Discovery**: There are different pattern discovery techniques like path analysis, association rule discovery, clustering analysis and classification.

d) **Stage 4 Pattern Analysis**: This includes pattern analyzing techniques like visualization tools, OLAP techniques, data and knowledge querying and usability analysis.

4. WEB MINING APPLICATIONS

a) E-Learning: Web mining can be used for improving and enhancing the process of learning in e-learning environments. Applications of web mining to e-learning are usually web based i.e. online and not offline.

b) E-government: Organizations that communicate with citizens of the country for better social services. The main characteristics of e-government systems is related to use of technology to deliver services electronically, focusing on the needs of the citizens by providing information and enhanced services in support of government.

c) E-commerce: A major challenge is to understand visitors or customer's needs. It can improve capacity of service for consumer.

d) Security and crime investigation: Web mining also used for protection of user system or logging information against cyber crimes such as hacking, internet fraud, id theft, online gambling, spreading viruses and child pornography.

5. PERSONALIZED WEB SEARCH

When user issues a query to Google search engine, every query to Google search engine is composed of one or more keywords. In response to a query, Google search returns a page of results. Figure 2 shows example page of Google search engine to the query "cough". We highlight the four results with red boxes. Many results contain >=1 links. There is a primary link which is highlighted with red arrows.

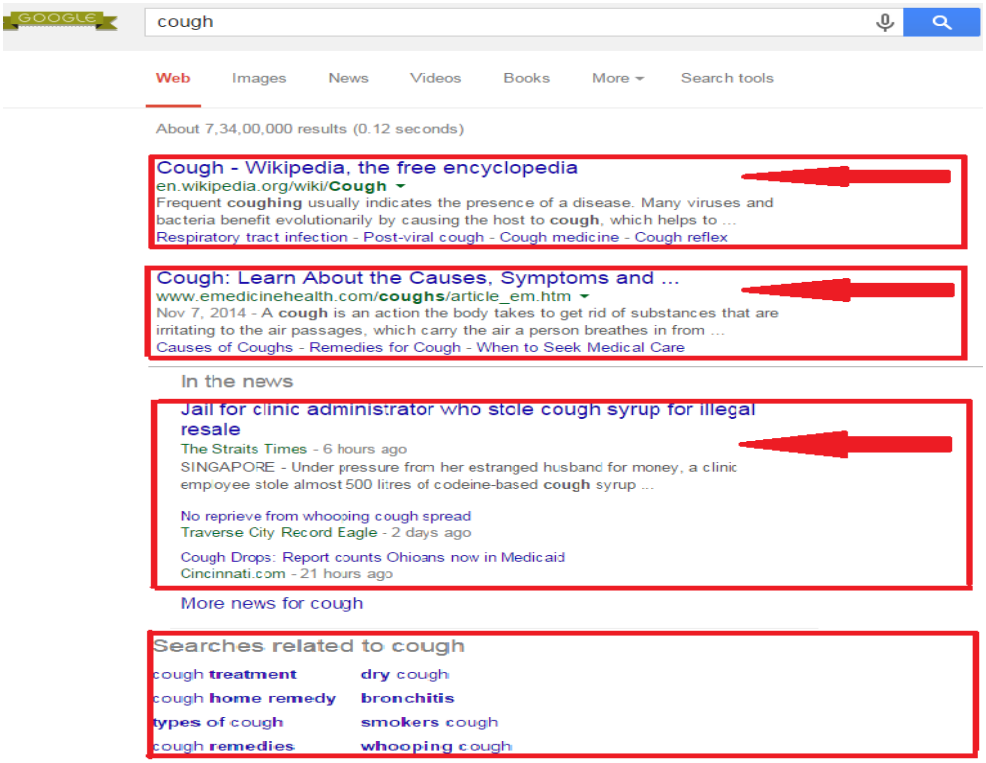

Figure 2: Example page of Google Search results for query "cough"

6. ISSUES OF GENERAL WEB SEARCH ENGINES

The goal of web search engines is to organize the web content and show the users the most important information to the users in response to their queries. This is a highly challenging task giving the huge amount of overload that exists on the web today with millions of users and terabytes of data. The queries provided by the users are often short and poor descriptors of information need and sometimes with ambiguities. Also, usually the users are not patient to see a long list of results and only see the top few results shown by the search engines. This adds to the problem and demands the search engines to show the most relevant results to the user on the top few for the user to really be happy with the search engine.

Till now, the search engine has mostly concentrating on the first factor – Information overload and trying to organize the web content and provide with the results. Some successful systems have been built. However, the focus has been on being able to provide search results. With this progress, the next step is to address the relevance part. This should be pursued in the future. It is challenging because the information need is not expressed properly and there is a lot of information that can be inferred to be relevant (especially from the word to word matches used by the current search engines).It is really important that the search engines understand the information need better in order to provide better results.

A main shortcoming of search engine is "One size fits all" model and are not adjustable to individual users. This is shown in such cases:

a) Different users have different interests. They may have different information needs when issuing the same query. For example a computer science student issuing a query "Java" which is a programming language and tourist or travel agent using same query to find the information about the same query "java" which is an island. When such queries are issued, Search engines return various documents related with it. So it will take time for user to search for his/her relevant document and this makes user unsatisfied. Queries like "java" are ambiguous queries. Thus search engines find difficulty in providing relevant results to ambiguous queries.

b) Sometimes users are not static. His information requirements may change in future. For example a computer science student, while watching television news about "Java Island" may search for the "Java Island".

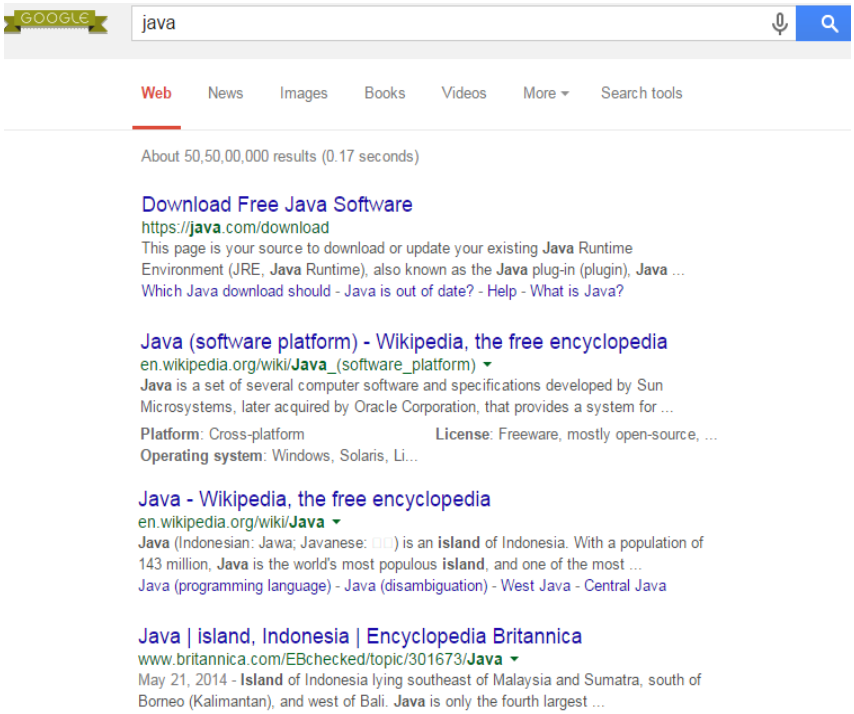

Figure 3: Example page of Google Search results for query "java"

To address the above problems, Personalized Web Search is a best solution, as it provide different search results based on the user's information requirements. So considering the above example Personalized Web Search disambiguates the query by gathering following information:

- User is a computer student not a travel agent.
- User had issued a query "C or C++ Programming language" before issuing a query "Java".

7. CHALLENGES OF PERSONALIZED WEB SEARCH

Personalized web search faces several challenges that retard its real-world large-scale applications:

1. Privacy is an issue. Personalized web search, especially server-side implement, requires collecting and aggregating a lot of user information including query and clickthrough history. A user profile can reveal a large amount of private user information, such as hobbies, vocation, income level, and political inclination, which is clearly a serious concern for users [9]. This could make many people nervous and feel afraid to use personalized search engines because of breach of confidentiality.
2. It is really hard to infer user information needs accurately. Users are not static. They may randomly search for something which they are not interested in. User search histories inevitably contain noise that is irrelevant or even harmful to current search.
3. Queries should not be handled in the same manner with regard to personalization. Personalized search may have little effect on some queries. Some work [1, 2, and 3] investigates whether current web search ranking might be sufficient for clear/unambiguous queries and thus personalization is unnecessary.

8. PERSONALIZED WEB SEARCH AND ITS LITERATURE SURVEY

For a given query, a P**ersonalized Web search** provide various search results for different users based upon their interests, preferences, and information requirements. It differs from general web search, which returns identical search results to all users for identical queries, regardless of user interests and information needs. Personalized web search is an optimistic way to improve search quality by customizing search results for people with individual information goals. The objective of a Web personalization system is to "provide users with the right information they want, without expecting from them to ask for it explicitly".

a) **User Profiling based Personalized Search:** To provide personalized search results to users, personalized web search maintains a user profile for each individual. User Profile stores user interests and information needs. Such information includes: Demographic and Geographical information including age, gender, education, country, address and, interest areas. Search history including previous queries and clicked documents. User browsing behavior when viewing a page such as mouse click, mouse movement, printing and bookmarking. Chirita et al. [1] and Teevan et al. [11] demonstrate that external user data stored in a user client is useful to personalize individual search results. User information is either specified by user (explicit collection of data) or automatically learnt from user's historical activities (implicit collection of data). As large numbers of users are opposed to provide explicit feedback on search results, many works focuses on how to automatically

learn user preferences without involving direct efforts of user [4, 6, 7, 8, and 10].User profile combine user's history and assume it as user's long-term interests. But some work has investigated that such long-term profile is ineffective in some cases. Assume the second case where user will have different needs at different times. So in such cases, personalization based on user's long-term interests may not be effective because same results cannot be returned.

In previous work, personalized search has been identified as one of the major challenges in information retrieval.

i. Search Query logs which consist of logs of searches made by users of search engines. They are usually collected at the search engine server. They typically consist of: user identity (Ip address or anonymous id etc.), search queries, corresponding clickthrough made by the user and click information regarding it like the click time, no of clicks made etc. Sometimes the query logs are also captured on the client side i.e., on the user's computers. Click through data/Query logs have been the most important source for capturing user context for modeling a user. Query logs are mined and queries similar to the current query are used to improve the search results. B.Tan, X. Shen, and C. Zhai, [8] Long-term search history contains rich information about a user's search preferences, which can be used as search context to improve retrieval performance.

ii. Z. Dou, R. Song, and J.-R. Wen, [3] although personalized search has been proposed for many years and many personalization strategies have been investigated, it is still unclear whether personalization is consistently effective on different queries for different users, and under different search contexts. In this paper, we study this problem and provide some preliminary conclusions.

b) **Content Analysis based Personalized Search:** Personalized web search can be achieved by checking content similarity between web pages and user profiles. Some work has represented user interests with topical categories that are either explicitly specified by users themselves, or by classifying implicit user data. In some work [2, 6], a user profile is structured as a concept/topic hierarchy. User-issued queries and user-selected snippets/documents are categorized into concept hierarchies that are accumulated to generate a user profile. Chirita et al. [2] use the ODP (Open Directory Project, http://www.dmoz.org/) hierarchy to implement personalized search.

c) **Hyperlink Analysis based Personalized Search**: Most generic web search approaches rank importance of documents based on the linkage structure of the web. An intuitive approach of personalized web search is to adapt these algorithms to compute personalized importance of documents. A large group of these works focuses on personalized Page Rank. Page Rank, proposed by Page and Brin [5], is a popular link analysis algorithm used in web search.

d) **Community-based Personalized Web Search**: In most of the above personalized search strategies, each user has a distinct profile and the profile is used to personalize search results for the user. These approaches are called community-based personalized web search or collaborative web search. In a community-based personalized web search, when a user issues a query, search histories of users who have similar interests to the user are used to filter or re-rank search results. For example, documents that have been selected for the target query or similar queries by the community are re-ranked higher in the results list.

9. SOFTWAREARCHITECTURE FOR PERSONALIZED SEARCH

Personalized Web Search can be implemented on client side (in the user's computer or a personalization agent) and server side (in the search engine).

a) **No Personalization**: For applications like web searching, client-server architecture shown in figure 4 where client (web browser) sends queries to server (the search engine) directly. The search engine analyses users information need, look into its index structure of documents and returns a ranked list of search results to the client for user's view. A search engine stores search logs or search histories for Anti-Spam and personalization.

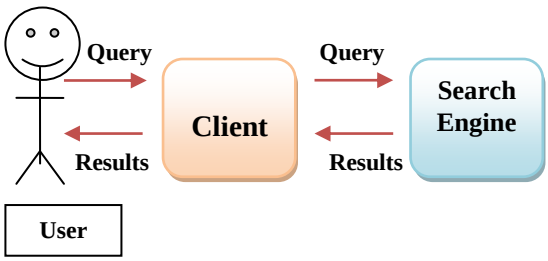

Figure 4: No Personalization

b) **Server-Side Personalization**: In this personalization, personally identifiable information is stored on server side. The search engine develops and updates the user profile either through user's input (specification of user's personal interests) or by collecting user's search histories (queries).The advantage of this architecture is that search engine can use all of its resources to provide proper response to the users. But the main problem in this web searching technique is that user's information needs as well as his personal information log details are exposed to the search engine i.e. on server side. This may disclose the user's privacy and may cause breach of confidentiality.

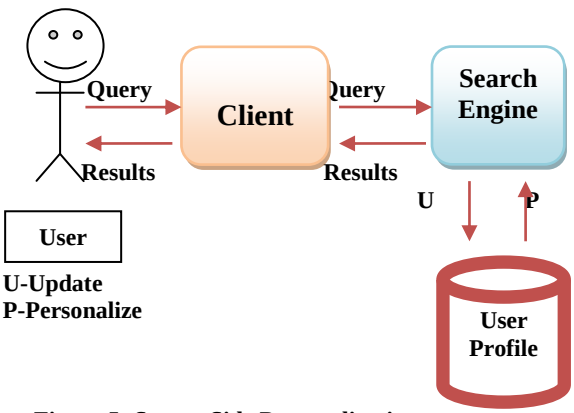

Figure 5: Server-Side Personalization

c) **Client-side Personalization**: In this type of personalization, personally identifiably information is stored on the client side .But after giving a user's queries, user will do query expansion to change original query and then send it further to search engine.

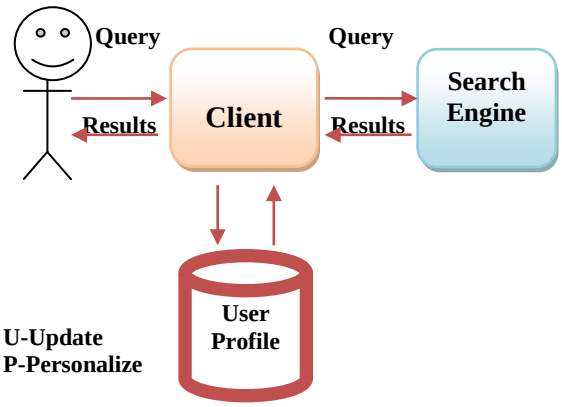

Figure 6: Client-Side Personalization

10. ADVANTAGES AND DISADVANTAGES OF PERSONALIZED WEB SEARCH

Advantages: It enhances the stability of the search quality and it avoids the unnecessary exposure of the user profile.

Disadvantage: All the sensitive topics are detected using an absolute metric called surprisal based on the information theory.

11. CONCLUSION

Web mining methods have strong significance on e-systems. web mining form the basis of marketing and e-commerce systems on the web. It's also be used to provide fast services to the users as well as intelligent websites for improving their businesses. Personalised web search is a promising solution to improve performance of generic web search engines as well as providing relevant results for the given queries.

12. REFERENCES

[1] Chirita P.A., Firan C., and Nejdl W. Summarizing local context to personalize global web search. In Proc. Int. Conf. on Information and Knowledge Management, 2006.

[2] Chirita P.A., Nejdl W., Paiu R., and Kohlschu¨tter C. Using ODP metadata to personalize search. In Proc. 31st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2005.

[3] Dou Z., Song R., and Wen J. A large-scale evaluation and analysis of personalized search strategies. In Proc. 33rd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2007.

[4] Liu F., Yu C., and MengW. Personalized web search by mapping user queries to categories. In Proc. Int. Conf. on Information and Knowledge Management, 2002.

[5] Page L., Brin S., Motwani R., and Winograd T. The pagerank citation ranking: bringing order to the web. Technical report, Computer Science Department, Stanford University, 1998.

[6] Pretschner A. and Gauch S. Ontology based personalized search. In Proc. 11th IEEE Int. Conf. on Tools with Artificial Intelligence, 1999.

[7] Qiu F. and Cho J. Automatic identification of user interest for personalized search. In Proc. 15th Int. World Wide Web Conference, 2006.

[8] Shen X., Tan B., and Zhai C. Implicit user modeling for personalized search. In Proc. Int. Conf. on Information and Knowledge Management, 2005.

[9] Shen X., Tan B., and Zhai C. Privacy protection in personalized search. SIGIR Forum, 41(1):4–17, 2007.

[10] Sugiyama K., Hatano K., and Yoshikawa M. Adaptive web search based on user profile constructed without any effort from users. In Proc. 12th Int. World Wide Web Conference, 2004.

[11] Teevan J., Dumais S.T., and Horvitz E. Personalizing search via automated analysis of interests and activities. In Proc. 31st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval.

[12] Y.Raju et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (6), 2011.

[13] http://www.dmoz.org/.

[14] http://en.wikipedia.org/wiki/Web_mining