International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307

Efficient Methods of Personalized Web Search, Techniques and Privacy

S.Geetharani *[1], Keerthika[2]
[1]Assistant Professor, [2]Research Scholar
Department of Computer Science, PSG College of Arts and Science, Coimbatore, Tamilnadu, India

ABSTRACT

Web search engines play an important role in web life. Generic web search engines are not suitable for identifying different needs of different customers. Personalized web search (PWS) is designed to provide different search results for different users. Personalization aims to provide users with what they need either by asking explicitly or implicitly. Several personalized web search models were developed based on web link structure, web contents, user queries, user profiles, browsing history etc. Personalized search has been a most important research area and many techniques have been developed and tested, still many issues and challenges are yet to be explored. User's information safe and ensuring privacy, search engines should provide security mechanism. This paper concentrates on the many personalized web search approaches understand the web personalization processes, benefits, limitations and future trends.

Keywords: Data Pre-processing, Personalized web search, Page Ranking Strategies, Personalization Techniques, Privacy.

I. INTRODUCTION

World Wide Web (WWW) is largest, commonly used and most accessible source of information. Search engine contains a large amount of miscellaneous data. Hence it is always difficult to extract the related information from this huge dataset. Mostly the single short query contains multiform meanings.

Personalization of web search is the process of customizing web search results based on users past behaviour. Most of the queries submitted to search engines are short and have ambiguity. Every user 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.
Personalization of web search can be done at either server side or client side. Many problems arise 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.
The most common difficulties encountered when searching the Web are:

i) Problems with the data itself
ii) Problems faced by the users trying to retrieve the data they want
iii) Problems in understanding the context of search requests and
iv) Problems with identifying the changes in user's information need.

PWS can be categorized into two types:
One is click-log-based methods and other profile-based ones. The click log based methods are based on just selecting the clicked pages in the user's query history. The main drawback of this method is that it works on repeated set of queries by the users only.

Profile based method has more effectiveness in improving the quality of web search with increasing usage of personal and behaviour information to profile its users, which is usually gathered implicitly from query history, browsing history, click-through data, bookmarks, user documents and so forth. The main drawback of this method is that it requires the user personal data to be send to the server; hence this privacy issue makes the user uncomfortable.

II. WEB DATA PREPROCESSING

Data pre-processing is the process to convert the raw data into the data concepts necessary for the further applying it in building user profiles. It identifies unique users and their session data. A Session data are the different information source utilized in the personalized web search process. It could be in any one of the following forms [1]. (i) Web Page: A document on the World Wide Web and each page is identified by a unique URL .The content of the page can be a simple text, images or structured data such as information retrieved from the databases.(ii) Web Structure: Hyper link structure of the web pages thereby becomes a directed graph. The nodes are the web pages and the directed edges connect different pages. (iii) Web Usage Data: It is a web site usage representation in terms of visitors IP address, date and time of Access, complete path (files or directories) accessed, referrers' address, and other attributes that can be included in a Web access log. (iv) User profile data provide information about the users of a Web site. The user profile contains Demographic information (such as name, age, country, marital status, education, interests etc. For each user of a Web site, as well as information about users' interests and preferences. Such information is acquired through registration forms or questionnaires, or can be inferred by analysing Web usage logs.

III. OVERVIEW ON PERSONALIZED WEB SEARCH

A. Personalized Web Search

A new technique on Personalized Web search can serve the different search results for different users, based upon their interests, preferences, and information needs. User information can be specified by the user or can be automatically learn from a user's historical activities. Personalized web search can be achieved by checking content similarity between web pages and user profiles. Personalized web search can improve the performance of web search. Personalized web search can be implemented on either server side or client side. For server-side personalization, user profile are created, updated, and stored on the search engine side. User information is directly incorporated into the ranking process, or is used to help process initial search results. For client-side personalization, user information is collected and stored on the client side, usually by installing a client software or plug-in on a user's.

B. Personalized Search Based on User Search Histories

User profiles, descriptions of user interests, can be used by search engines to provide personalized search results. Many approaches to creating user profiles collect user information through proxy server or desktop bots. Personalization is the process of presenting the right information to the right user at the right moment. Systems can learn about user's interests collecting personal information, study the information, and storing the results in a user profile. Information can be captured from users in two ways. Explicitly, for example asking for feedback such as preferences or ratings; and implicitly, for example observing user behaviour's such as the time Spent reading an online document.

C. Personalized Concept-Based Clustering of Search Engine Queries

Concept based profiling method that captures the user's conceptual preferences in order to provide personalized query suggestions. Two new strategies are used to achieve this goal. First develop online techniques that extract concepts from the web-snippets of the search result returned from a query. Second a new two phase personalized agglomerative clustering algorithm that is able to generate personalized query clusters.

D. Click-Based Methods (PClick)

PClick is good in capturing user's positive preferences. When the user searches for the query "apple," the concept space derived from our concept extraction method contains the concepts "Macintosh, "iPod," and "fruit." If the user is indeed interested in "apple" as a fruit and click on pages containing the concept "fruit," the user profile represented as a weighted concept vector should record the user interest on the concept

"apple" and its neighbourhood (i.e., concepts which having similar meaning as "fruit"), while downgrading unrelated concepts such as "Macintosh," "iPod," and the neighbourhood.

E. Personalization based on User Positive and Negative Preferences

Most commercial search engines give the same results for the same query, not considering the user's interest. User profiling is a fundamental component of any personalization application. Most existing user profiling strategies are based on object that users are interested in (positive preferences), but not the objects that users dislike (negative preferences).

User Profile	Description
Click-Based	Which capture only Positive preference
Joachims-c	Which capture only negative preference and consider only un clicked page above clicked page
mJoachims-c	Which capture only negative preference and consider only un clicked page both above and below clicked page

Table 1. User positive and negative preference

F. Location *based ranking method (LBRM)*

A Location-based Ranking Method (LBRM) is proposed for ranking search results based on the location effects in the search engine. Users have to give the queries from different locations and retrieve the results. The proposed method incorporates three modules. The first module is similarity identification module. If the user submits the query from a particular location, the search engine provides the results. Firstly, the user locations are identified by the geographic information and get the locations. The similarity value is identified among the locations and retrieved pages. The two databases are derived called Location-page Database (LPD) and Page location Database (PLD) for the similarity identification. Then the frequent retrieval patterns are retrieved by computing the support value. The support value denotes the frequent.

IV. PERSONALIZED PAGE RANKING STRATEGIES

There are several ways to retrieve the documents relevant to the query. The research efforts on re-ranking web search results are categorized into the following classes of strategies.

(i) Explicit relevance judgments
(ii) Implicit relevance judgments
  (a) Content-based implicit measures
  (b) Behaviour-based implicit measures

A. Explicit Relevance Judgments

The trouble-free way to verify whether a result retrieved for a query is relevant to the user is to explicitly ask that user. Explicit judgment allows us to scrutinize the uniformity in relevance assessments across judges in a controlled setting. Advantage of this method allows us to examine the consistency in relevance assessments across judges in a controlled setting. Following are the limitations (i) It is cumbersome for people to give explicit judgment because it consumes additional time and effort from the users. (ii)It is difficult to gather sufficient data to generalize across a broad variety of people, tasks and queries. (iii)It is captured outside an end-to-end search session.

B. Implicit Relevance Judgments

Implicit data can be generated by users' interaction with their service. Implicit measures are easier to collect and allow us to explore many queries from vast variety of searchers. The two most common implicit measures used for personalization are (a) Content-based Implicit Relevance Judgments, (b) behaviour-based Implicit Relevance Judgments.

- Content-based implicit relevance judgments

This type of measure uses a textual representation of users' interest to deduce the results which are relevant to their current need. Content-based profile captures all of the information created, copied or viewed by an individual. It also includes web pages viewed, email messages sent or received, calendar items and documents stored on the client machine. Benefit of using this method is information about millions of

users and millions of queries can be obtained and shows better performance than pure text-based algorithm and content-based algorithm. But, it is having following disadvantages

(i) Activities of users are influenced by presentation of results
(ii) Performance is lower than regular Web Ranking methods
(iii) Currently updated information in the web repositories will not be reflected in the web search results dynamically.

Topical Interest based Ranking covers the spatial factors such as Queries used, Query usage count, Relevancy between the query and the document, query and the user profile, context of the query with reference to the ontologies or web dictionaries. The above factors normally support to develop the knowledge based user models. Sieg et al. utilized the user context to personalize search results by re-ranking the results returned from a search engine for a given query. Re-ranking the search results based on the interest scores and the semantic evidence in the user profile is done. A term-vector r is computed for each document r Є R, where R is the set of search results for a given query. The term-weights are obtained using the tfidf formula.

To calculate the rank score for each document, first the similarity of the document and the query is computed using a cosine similarity measure. Then, the similarity of the document with each concept in the user profile to identify the best matching concept is computed. Once the best matching concept is identified, a rank score is assigned to the document by multiplying the interest score for the concept, the similarity of the document to the query, and the similarity of the specific concept to the query. If the interest score for the best matching concept is greater than one, it is further boosted by a tuning parameter .Once all documents have been processed, the search results are sorted in descending order with respect to this new rank score.

- Behaviour-based implicit relevance judgments

This type of measure uses people's behaviour such as their past interactions with search result lists, click-through data from the logs etc. The Performance is better than pure text-based algorithms. Some of the disadvantages are (i) Intent for a query may vary widely among each individual. (ii)Performance is lower than behaviour based and other web ranking methods. The goal of Collins et al. was to show how modelling reading proficiency of users and the reading difficulty of documents can be used to improve the relevance of Web search results. Web users differ widely in their reading proficiency and ability to understand vocabulary, depending on factors such as age, educational background, and topic interest or expertise. Hence it is clear that there is a need for improvement in ranking search results at an appropriate level of reading difficulty. To address this problem, they described a tripartite approach based on user profiles, document difficulty, and re ranking. First, the snippets and Web pages can be labelled with reading level and combined with Open Directory Project (ODP, www.dmoz.org) category predictions. Second, they described how a user's reading proficiency profile may be estimated automatically from their current and past search behaviour. Third, they use this profile to train are ranking algorithm that combines both relevance and difficulty in a principled way and which generalizes easily to broader tasks such as expertise-based re ranking. In this view, the overall relevance of a document is a combination of two factors: a general relevance factor, provided by an existing ranking algorithm, and a user-specific reading difficulty model, based on the gap between a user's proficiency level and a document's difficulty level. While users may self-identify their desired level of result difficulty, such information may not always be provided. They investigate methods for estimating a reading proficiency profile for users based on their online search interaction patterns. The reading level of user can be defined by,

$$p(\text{u likes level of d} \mid r_u, r_d) = \exp(-(r_d - r_u)^2)$$

Where, u - user; d - document; r_u - reading level of user u; r_d - reading level of document d.

C. PageRank algorithm

PageRank is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages.

Page Rank algorithm was proposed by Brin and Page at Stanford University. For the most of web pages these ranking algorithms can be used repeatedly. During the processing of a query, search algorithm merges pre calculated. As a result, the process of ranking can be completed by Page Rank. This score along with the text matching scores is used to gain an overall ranking score for each web page. Page Rank algorithm function is related to the link structure of the web pages. The concept of Page Rank algorithm is if a page surrounds an essential links on the way to it, then the links of this page near the other page are also to be assumed as imperative pages. The Rank score conclusion can be restricted on the back link of the Page Rank. When the addition of the ranks in the back links are high, then the page holds a high rank as well.

Preference mining and machine learning to model users clicking and browsing behaviour are employed by a method, which was proposed by Joachims. Users clicking and browsing behaviour are modelled by Machine learning and Preference mining. These models are employed by using a method, which was proposed by Joachims. During query processing, the relations are lost and given keywords are treated as individual keywords, thus creating the major problem of isolated keyword matching. Though the ranking of the retrieved web pages has not accounted for relations, such that it is purely based on link analysis like PageRank and some on page relevance factors.

A combination of spying technique and novel voting procedure is employed for determining user's document preferences from the click through data by an algorithm. In order to learn the user behaviour model as a set of weight features, RSVM algorithm is also employed by them. More recently, explicit feedback (i.e., click through data, individual user behaviour etc.) from search engine users is noisy was suggested by Agichteinet al. In the following sections we proposed user profile strategies and ranking algorithm for inbound and outbound links and the relevancy of pages can be returned.

V. PERSONALIZED TECHNIQES

User profiling

- server side implementation
- client side implementation
- Content analysis

A separate user profile should be maintained for each user. User profile consists with technical, demographical and geographical information of users. Previously visited pages, total visit time, number of visits, used links, age, gender, education, IP addresses and bookmarks etc.

- Server side implementation

Search engine has to maintain user profiles by using its resources. Engine can use its all resources to optimize the search results. Allocate a huge amount of memory and computing processes to maintain millions of user profiles.

- Client side implementation

Users are the responsible parties for maintaining their user profiles.an installed software should be used to facilitate. Violation of privacy and security can be preserved as much as possible. Cost of storage and computing processes are distributing among users. Limitation of network bandwidth.

- Content analysis

Content analysis is under user profiling technique. Check the similarity between web pages and user profile details. User interested topics and title or content of the web pages are much concerned.

Hyperlink analysis

Most of leading search engine uses this method. Crawling and ranking concepts. PageRank and biased PageRank approaches.

Community based PWS

Avoid the handling of separate user profile for each user. Search engine has to find the users who have similar kinds of interests. Effective identification increases the productivity of the collaborative web search.

Despite of having various advantages of personalized search, there is no large-scale use of personalized search services currently. Personalized web search faces several challenges that hinder its real-world large-scale applications:

- Privacy is an issue.
- Users are not static.
- Queries should not be handled in the same manner with regard to personalization.

VI. PRIVACY

There are two classes of privacy protection problems for PWS in general. One class includes those works, treat privacy as the identification of an individual. The other includes those consider the sensitivity of the data, particularly the user profiles, exposed to the PWS server.

A. Identification of An Individual Typical works in the literature of protecting user identifications (class one) try to solve the privacy problem on different levels, including the pseudo-identity, the group identity, no identity, and no personal information [4]. Solution to the first level is proved fragile. The third and fourth levels are impractical due to high cost in communication and cryptography. So the existing efforts focus on the second level.

Online anonymity: It works based on user profiles by generating a group profile of k users. Using this approach, the linkage between the query and a single user is broken.

Useless user profile (UUP): This protocol is proposed to shuffle queries among a group of users who issue them. As a result any entity cannot profile a certain individual. These works assume the existence of a trustworthy third-party anonymizer, which is not readily available over the Internet all the time in large number.

Legacy social networks: Instead of the third party to provide a distorted user profile to the web search engine, here every user acts as a search agency of his/her neighbours'. They can decide to submit the query on behalf of who issued it, or forward it to other neighbour's.

B. Sensitivity of Data The solutions in class two does not require third-party assistance or collaborations between social network entries. In these solutions, users only trust themselves and cannot tolerate the exposure of their complete profiles to an anonymity server. (i) Statistical Techniques: To learn a probabilistic model, and then use this model to generate the near-optimal partial profile. One main limitation in this work is that it builds the user profile as a finite set of attributes, and the probabilistic model is trained through predefined frequent queries. These assumptions are impractical in the context of PWS. (ii) Generalized Profiles: Proposed a privacy protection solution for PWS based on hierarchical profiles. Using a user specified threshold, a generalized profile is obtained in effect as a rooted sub tree of the complete profile. C. Issues The shortcomings of current solutions in class one is the high cost introduced due to the collaboration and communication. The statistical methods builds the user profile as a finite set of attributes, and the probabilistic model is trained through predefined frequent queries in class two. These assumptions are impractical in the context of PWS and the generalized profile does not address the query utility, which is crucial for the service quality of PWS.

VII. CONCLUSION

This paper presents a survey report of different methods to help in inferring user's information needs of Personalized Web Search. It also covers issues like need of personalized web search, how personalized web search can be implemented, what are challenges in it, privacy and security issue of it and existing system of personalized web search. Thus the motivation behind the personalization is to enhance quality of rankings.

VIII. REFERENCES

[1]. J. Jayanthi, DR.S.Rathi, "Personalized Web Search Methods-A Complete Review", Journal of Theoretical and Applied Information Technology 30 Th April 2014. Vol. 62 No.3

[2]. M. Spertta and S. Gach, "Personalizing Search Based on User Search Histories", Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI), 2005.

[3]. Himani Arya, Jaytrilok Choudhary, Deepak Singh Tomar, "A Survey on Techniques for Personalization of Web Search", International Journal of Computer Applications (0975 – 8887) Volume 94 – No. 18, May 2014.

[4]. Esmita Gupta, Prof.Deepali vora"Survey on Privacy Preservation in Personalized Web

Environment", International Journal on Recent and Innovation Trends in Computing and Communication Volume: 3 Issue: 4.

[5]. Prashanthi.G, "Personalized location based search engine using click throughs", IJCSIET International Journal of Computer Science information and Engg, Technologies.

[6]. Sofia Sayed, Reeba R "A Survey of Web Page Personalization in Web Search Engine", International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-1, January 2016.

[7]. Charanjeet Dadiyala, Prof. Pragati Patil, Prof. Girish Agrawal" Personalized Web Search", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.

[8]. Vijalakshmi Kakulapati, Dr.D. Vasumathi, Sudarson Jena, "Survey on Web Search Results Personalization Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013.

[9]. Y. Xus, K. Wang, B. Zhang, and Z. Chen, "Privacy Enhancing Personalized Web Search," Proc. World Wide Web (WWW) Conf., 2007.

[10]. Liu F, Yu C, Meng W. Personalized Web Search by Mapping User Queries to Categories. Proc Int'l Conf Information and Knowledge Management (CIKM); 2002.

[11]. S.Geetha Rani and M.Sorana Mageswari "A Link-click-concept based Ranking Algorithm for Ranking Search Results "Indian Journal of Science and Technology, October 2014.

[12]. S.GeethaRani"A New Ranking Algorithm for Ranking Search Results of Search Engine based on Personalized User Profile"International Journal of Computer Applications, July 2013