Advancement in Personalized Web Search Engine with Customized Privacy Protection

Jeena Mariam Saji, Kalyani Bhongle, Sharayu Mahajan, Soumya Shrivastava and Ashwini Jarali

Abstract Technologies are blooming, needs are growing, larger user data is getting aggregated, and thus privacy becomes a matter of concern in this fast paced, technology driven environment. People are relying mostly on Internet for almost everything they work on or experience. The web search engines confuse us sometimes by giving mixed results. Different people may have variant requirements, and search engines provide same results for same queries, but to different people. In this paper, we intend to solve this problem by a technique of generating online user profiles before firing any query. This user profile would store the user details and the search engine would display results according to this generated profile. We use collaborative filtering and ranking function to filter out the pages according to the preferences of user. We intend to add a feature in our system where in, the users will get a chance to handle their degree of privacy. We offer them two friendly buttons—"Private" and "Public". These buttons will decide whether the user wants to share his details with other users or not. A combination of personalization and privacy would surely be worth a good use for the Internet seekers.

J.M. Saji (✉) · K. Bhongle · S. Mahajan · S. Shrivastava · A. Jarali
Department of Computer Engineering, International Institute of Information Technology, Pune, India
e-mail: jeenamariamsaji@gmail.com

K. Bhongle
e-mail: bhonglekalyani@gmail.com

S. Mahajan
e-mail: sharau.mahajan@gmail.com

S. Shrivastava
e-mail: soumya.300195@gmail.com

A. Jarali
e-mail: ashusleek1@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_44

Keywords Customize privacy · Filtering · Generalized profile · Personalized search results · Re-ranking

1 Introduction

Search engines have become a key element for discovering information over the internet. For every problem, we think of internet as a savior. It is often observed that same set of results are displayed to different users for the same query. For example, a doctor wants to search information related to a human face. When he types "face" and hits the search button, he may get Facebook as a search result instead of a human face. Thus in current system, we get mixed results, not the results according to our preferences. Hence it becomes difficult to find for the desired information at one single glance. We often need to go through several other pages in order to find the specific link of information. Sometimes, the results displayed by the search engine may be relevant to the keyword entered by the user, but may not be able to fulfill user's expectations of information need. A user enters the query for which he wants to search information in the search textbox and receives a long list of results or links in lieu of the query entered. The challenge of the search engine is to translate user's simple queries into list of documents that satisfy the different information needs. To overcome this challenge, people came into a conclusion of working with the personalization of search engines.

The profile-based personalized search engine takes the user input, displays the list of results, and also uses the additional information about the user to aid in checking the relevance of the pages. There are various approaches to provide personalization to web search engines. Some of the features determining such approaches are user details, user level interaction, and information which is stored and algorithm which is used to retrieve user details into the search.

The main feature of this paper is that we allow users to control their degree of privacy protection by providing them a Private and Public option in our search engine. These options help the user in deciding their level of privacy according to their requirements. If a user wants to share the browsing queries, he just has to hit the "Public" button and thus he can maintain the transparency accordingly.

In general, our main aim is to develop such a search engine which is privacy protected as well as customized privacy web search engine. The paper is further elaborated into II. Literature Survey and III. Proposed System briefly explaining the purpose of the paper.

2 Literature Survey

Personalization is being accepted by a large set of users to ease the use of web search engines. But despite being proposing it for many years it is still difficult to analyze whether personalization has an adverse effect on all kinds of queries and for different users or not. Dou [1] gives an overview of different problems faced in personalization along with their solutions to it. It is followed that the queries entered by different users often produce the same results altogether, in spite of being variant information goal. A framework based on query logs is developed to ensure massive scale enhancement of personalized search. It is revealed that queries with large click entropy have severe improvement over the common web search. It is seen that different queries has different effectiveness so it is advised that not all the queries should be personalized. The profile-based search strategy mentioned by Z. Dou is not as reliable or effective as the click-based search strategy. It is realized that short-term, long-term contexts and logs are necessary to be analysed for a profile-based search strategy. Thus it is concluded that a combination of both would be reliable.

Deng and Lee [2] presented the personalization of web search engine where the results are displayed according to the preferences set by the users. They introduce us to another mining technique, called Spy Naïve Bayes (NB) which states that the clicked items imply user's choices. It is often seen that for same queries issued by different users, produce same result. However, different users may tend to have different choices for searching a particular query. For resolving this problem of search engine transformation, some research issues are considered. The primary research issue is preference mining which deals with the preference of users of search results from click-through data. Another issue is ranking function optimization which helps in optimizing the retrieval of results according to the user's preferences in search engines.

In the new SpyNB approach, a list of preferences is generated and is fetched by the Ranking Support Vector Machine (RSVM) for optimizing the ranking function for the user. The SpyNB algorithm helps in generating preference fragment pairs used for ranking function. The fragment pairs offer an effective element in making this approach more reliable. Thus it is concluded that SpyNB approach is more productive and flexible than the algorithm existing currently.

Alexander Pretschner and Susan Gauch [3] in paper represented a system where they have explored the ways of incorporating user's interest to the search process in order to enhance the search results. They suggested the method to generate a user profile depending on the way the user would surf the online pages. Combination of three major metrics, time, subject discriminator, and length were used to analyse the user behaviour and create his/her profile accordingly. Here time denotes the amount of time a user spends on a given page, while length refers to the number of characters in the page. On the basis of the analysis done by them, the profiles reflected the user's interest quite well and could be used to deploy more effective information retrieval and filtering. So basically this paper provided a solution to

retrieve more relevant search results by using the profiles created on the basis of surfing history of a user.

In order to overcome the issues related to privacy concern from user's perspective, Krause and Horvitz [4] explored and introduced a study of privacy in personalization, where user has an option to share his/her personal information, in return for expected enhancement in the retrieval of more relevant search result. Krause and Horvitz illustrated the methodology based on the graphical analysis survey of the log that saved user's search history. Through this survey they seek to comprehend the utility of personalization that can be actualized by using user's log-based information to analyse his/her willingness to trade the sharing of their personal data with any online services that they are exposed to. Thus they focused mainly on achieving efficient personalized search service using minimum user information.

Lidan Shou and Chen [5] presented a study paper where they used a user-side privacy protection system which is called UPS for personalized web search. They proposed two algorithms, namely Greedy DP (Discrimination power) and Greedy IL (Information Loss) that was used to generalize the user profile in order to avoid exposure of user's personal information while using the search engine. Privacy risk and utility of personalization were the two major predictive metrics used in their proposed algorithm. Also their experimental results acknowledge that the UPS framework could outperform the existing web framework and provide more effective and efficient solution. For future work they suggested to use the better predictive metrics to improvise the performance of the UPS framework.

Xiao and Tao [6] believed that the existing methodology focus on a universal approach that endeavor the same amount of preservation for all persons, without catering for their actual needs. Motivated by this, they came up with the alternative solution of generalizing the whole web search framework based on the concept of personalized k-anonymity. The method has been explained with the help of careful theoretical study of the user information which is used for the research purpose. They have used QI generalization to generalize the various attributes that are taken as an individual's detail information. In this technique if the user provides n number of attributes in a detail table, then after generalization over these tuples/attributes, only n and k detail would be exposed to the outsider hence preventing any kind of information loss. Here k is the tuples that has been generalized and eliminated from the final set of data.

Also the paper has clearly mentioned the drawback of their proposed solution for providing privacy protection in the personalized system and also focuses on developing more optimal alternative generalization strategies.

3 Proposed System

In this proposed system, we are using various technologies to develop a client side privacy protected personalized web search engine. By calling it a client side protected search engine, we mean that the user will have a control over sharing his profile and browsed logs with other users. In this system, we are using a technique known as collaborative filtering where information and patterns involving collaboration among multiple agents are filtered which lets us know the preference of the users. Using collaborative filtering, ranking and rating will be done on web documents. Our main aim is to generate results according to user's preference and lower the risk of disclosing user's sensitive information. Technologies used in this technique are web crawling, web mining, pattern recognition, and application program interfaces (API's).

In our system, initially the user will create an account on search engine. By creating an account the user will create a profile which will be stored in database server. Privacy is also provided to individual profiles. While creating a profile, the user will provide personal details like address, profession, interests, hobbies, etc. After signing up, the user will login into the respective account and will start browsing by issuing a query. We need a client database server which would take the responsibility of storing the user profiles. Generalization of the profiles [5] would take place alongside and will be sent to the central server. Let us assume these generalized profiles as "G". Generalized Profile (G) will be sent to the web crawler and then the functioning will begin.

When the query is issued, it is first preprocessed and then sent to the World Wide Web where the web crawlers analyse the entered query and crawl to different web

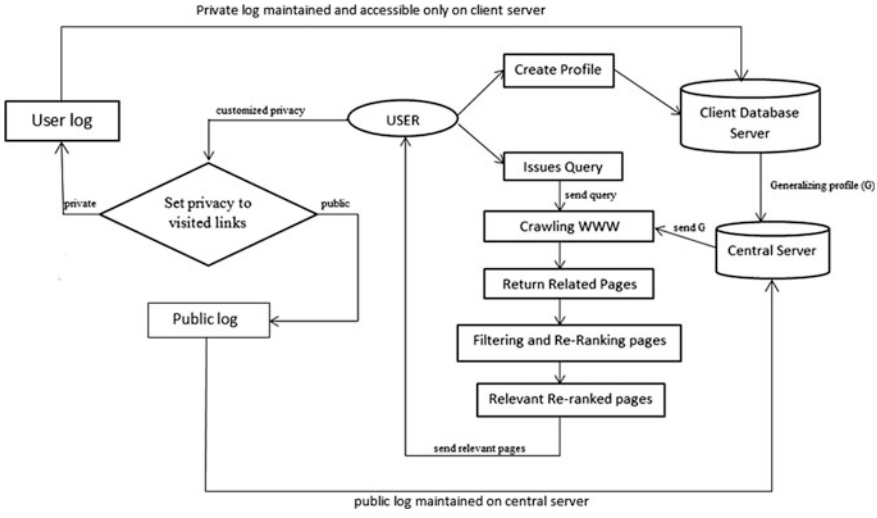

Fig. 1 System flow

pages to collect information from different sources. The web crawler then returns the pages related to the issued query. Since we are using collaborative filtering in our system, the related pages will get filtered. After filtering the crawled pages, ranking function is performed on them. And that is how relevant results are sent back to user. The user-system interaction will be possible because of an API in the middle layer. Figure 1 shows the brief functioning of our system.

In our system, we are using Spy Naïve Bayes algorithm [2] and Deep Search algorithm for reverse searching of relevant document. Privacy is provided to users by providing them with two buttons—"Private" and "Public" respectively. When the "Private" button is clicked, the link being customized to private status will be maintained in the private log, which will be stored on the client server and accessible only to the user. Whereas when "Public" button is clicked, the links with public status will be maintained in public log that is stored in server and will be accessible to all. Thus in this way we combine privacy with personalization in our system making it users decision whether to share his details and visited sites with server which in turn uses the public log for business and similar purposes making this public log consisting user history accessible to other stakeholders like advertisers, researchers, analysts and similar third party member.

3.1 Basic Steps Involved in Personalized Web Search (PWS)

Let us consider the following set of tuples:

P = {Set of user profiles}
Q = {Set of queries given by a user}
R = {Set of Response/search results given back to user}
G = Generalized profile for every P
N = Number of overall results related to Q

- Creating generalized user profile

1. Initially a user will register with the PWS engine by creating his/her own profile P which will consist of attributes like his/her name, gender, age, profession, interest, and other such related personal detail.
2. This profile P will be processed in such a way, that only the attributes which are required for the further processing will be collaborated together to form a new generalized profile G. This is done in order to avoid any kind of user's personal information loss.

The information related to the user profile will be saved on the client-side itself for reducing the privacy concern. While the updated generalized profile G can be saved on the serve-side, since it would be required while filtering the relevant search results as per the user's profession and interest. Storing G on the server side assures the better response time for processing at the same time reduces the

complexity while filtering and ranking the pages as the need of communicating with the client server would be avoided.

- Ranking and providing customized privacy

1. User can further browse a query Q in the search box which in return will send this Q to the main server where the crawling over World Wide Web would be initialized.
2. The process of crawling will give N results related to the query Q send by the user. These N results would be filtered and re-ranked based on the generalized user profile G.

Suppose R[n] is the set of result related to Q returned after crawling, 'i' is the index for every individual page/link in R[n] and 'rank' is the ranking of the page,

Then if R[i] == G && rank == high
Then set R[i] first

Repeat till whole result set R[n] is sorted and re-ranked based on G.

3. Once the filtering is done the user response would be created with completely new set of re-ranked page results R as per the user's interest.
4. For every Q the user would be given the choice to set his status either private or public. The status here specifies whether the user want to share the visited links with the other users.

Here the other users are the stakeholders consisting both the registered PWS users and advertisers.

This way the user would be ensured that his pattern of going through the result set is not being intruded by others and neither any of his personal information is exposed to the outsider.

3.2 Tools and Technologies Used

Tools required in our system includes Windows 2007 or above, JDK 1.7 and Tomcat Apache 7.0 and MySQL database. For our system to be executed, we need minimum hardware which would include Processor Pentium 4 or above and minimum hard disk space of 2 GB.To communicate and get connected we need some hardware interfaces like Ethernet, modem and Wi-Fi router, as well as some software interfaces like web browsers, DB2, Eclipse, servlets, AJAX, JSP, and Operating System.

We make use of communication interfaces like Internet, Web Server, and HTTP protocol basically on the central repository.

3.3 *Objectives of Proposing the System*

- To provide relevant search results based on the users choice.
- To ensure privacy protection to the user's personal information.
- To simplify the filtering process using simple sorting based solution.
- To provide customized privacy service to the user for sharing his/her visited results and search queries with other users.
- To eliminate the unwanted advertisement pop-ups.
- To improve the efficiency of the existing PWS by suggesting optimal solution.

4 Conclusion

The advancement in the technology like web search engine is boundless. Moreover the development of profile-based personalized search engine over a regular web search engine has catered many requirements of the user and inclusion of concept like privacy protection has served for the betterment of the system which has helped reduce various privacy concerns making the PWS more user-friendly. Though the existing PWS helps to retrieve relevant results to the user, there is still the need of improvising and providing more stable solution in order to make the whole system efficient and effective at a time. The other drawback of existing system is that it provides the privacy without knowing whether the user really want to personalize the information and other attributes like browsed query and visited links or share it with other registered users. Our proposed "Customized privacy" can help to overcome this drawback and let the user decide in case he/she wants to share the log with others and allows user to selectively share the information with related online services that is usually used for advertisement and research purposes. For future work, better generalizing strategies can be looked for that can replace the existing strategy and make the system more optimal and efficient.

References

1. Zhicheng Dou, Ruihua Song, Ji-Rong Wen: A Large-scale Evaluation and Analysis of Personalized Search Strategies. ACM Transactions 978-1-59593-654-7/07/0005 (2007)
2. Wilfred Ng, Lin Deng, Dik Lun Lee: Mining User Preference Using Spy Voting for Search Engine Personalization. ACM Transactions on Internet Technologies, Vol. 7, (2007)
3. Alexander Pretschner, Susan Gauch: Ontology Based Personalized Search. Proc. 11th IEEE Intl. Conf. on Tools with Artificial Intelligence, November (1999) 391–398
4. Andreas Krause, Eric Horvitz: A Utility-Theoretic Approach to Privacy in Online Services. Journal of Artificial Intelligence Research (2010) 633–662

5. Lidan Shou, He Bai, Ke Chen, Gang Chen: Supporting Privacy Protection in Personalized Web Search. IEEE Transactions on Knowledge and Data engineering, Vol. 26, NO. 2 (2014)
6. Xiaokui Xiao, Yufei Tao: Personalized Privacy Preservation. Proc. ACM SIGMOD, June (2006)