International Journal of Research in Advent Technology (E-ISSN: 2321-9637) Special Issue
1st International Conference on Advent Trends in Engineering, Science and Technology
"ICATEST 2015", 08 March 2015

A Novel Framework Model for User Specified Personalized Web Search Supporting Privacy Concern

Ms. A. S. Patil1, Prof. M.M.Ghonge2
[1]*M.E Scholar (CSE), JCOET, Yavatmal,ankita.patil5@gmail.com,9665751977*
[2]*Assistant Professor,JCOET,Yavatmal,mangesh.cse@gmail.com,9096449280*

Abstract: Searching is one of the common task performed on the Internet. Search engines are the basic tool of the internet, from where one can collect related information and searched according to the specified query or keyword given by the user, and are extremely popular for recursively used sites. The information on the web is growing dramatically. The users have to spend lots of time on the web finding the information they are interested in. Today, the traditional search engines do not give users enough personalized help but provide the user with lots of irrelevant information. In such case, Personalized Web Search (PWS) has demonstrated its effectiveness in improving the quality of various search services on the Internet. However, evidences show that users' are not willing to disclose their private information during search has become a major barrier for the wide use of PWS. The proposed PWS framework called UPS not only gives information about privacy protection in PWS applications that model user preferences as hierarchical user profiles but also can adaptively generalize profiles by queries while respecting user specified privacy requirements.
Keywords –Privacy protection, Personalized Web Search, UPS framework, profile

1. INTRODUCTION

Over recent years, the World Wide Web has become a new communication medium with Web information access. The web search engine is the most important portal for ordinary people looking for useful information on the web. This incorporates with informational, cultural, social and evidential values to be specific. With the existence of various Search Engines e.g. Google, Yahoo and many more, the users are tend to use them for retrieving their desired Web pages and their information. Although today's search engines can meet a general request, they cannot distinguish different users' specific needs well. So, users generally experience failure and improper results when search engines return irrelevant results that do not meet their real intentions. A typical search engine provides similar set of results without considering of who submitted the query. Therefore, the requirement arises to have personalized web search system which gives outputs appropriate to the user as highly ranked pages and provide customized results depending on each user's interests.

Personalized Web Search (PWS) is a general category of search techniques which aims to provide better search results, according to individual user needs. So, for this user information has to be collected and analysed so that the perfect search results required for the user behind the issued query is to be given to the user. 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 [1]

1.1 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. [1] The need for personalization arises due to the two reasons: 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.

1.2 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 [2].

Web search personalization systems use gathered information about user from profiles, cookies and to conduct and revise the search to maximize the user satisfaction. 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 behavior, 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. This is being done in the background automatically by the search system.[1][2]

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. [2]

1.3 Solutions to Personalized Web Search (PWS)

The solutions to Personalized Web Search (PWS) can generally be categorized into two types, first is click-log-based methods and second is profile-based ones. The click-log based methods are simple and straightforward: This method performs the search based upon clicked pages in the user's query history. Although this method has been demonstrated to perform consistently and considerably well [2], it can only work on repeated queries from the same user, which is a strong limitation and restricted for certain applications. In contrast, profile-based methods improve the search experience with complicated user-interest models generated from user profiling techniques. Profile-based methods can be proved more effective for almost all sorts of queries, but are reported to be improper under some situations.[1].Although there are reasons and considerations for both types of PWS techniques, the profile-based PWS has proved its more effectiveness in improving the quality of web search recently, with increasing usage of one's personal and behavioral information to profile its users, which is usually gathered implicitly with the help of query history , browsing history, click-through data , bookmarks, user documents , and so on. Unfortunately, such type of collected personal data can easily reveal a entire scope of user's private life. Protecting privacy issues rising from the lack of protection for such data, not only raise panic among individual users, but also downs the data-publisher's enthusiasm in offering personalized service. In fact, privacy concerns have become the major barrier for wide use of PWS services. [3]

2. LITERATURE REVIEW & RELATED WORK

This paper focuses on the literature of profile-based personalization and privacy protection in PWS system.

2.1 Profile-Based Personalization

Previous works on profile-based PWS mainly focus on improving the search utility. The basic idea of these works is to tailor the search results by referring to, often implicitly, a user profile that reveals an individual information goal. In the remainder of this section, we review the previous solutions to PWS on two aspects, namely the representation of profiles, and the measure of the effectiveness of personalization.[3]

In the proposed UPS framework, we do not focus on the implementation of the user profiles. Actually, our framework can potentially adopt any hierarchical representation based on taxonomy of knowledge. As for the performance measures of PWS in the literature, Normalized Discounted Cumulative Gain (nDCG) [4] is a common measure of the effectiveness of an information retrieval system. It is based on a human graded relevance scale of item-positions in the result list, and is, therefore, known for its high cost in explicit feedback collection. To reduce the human involvement in performance measuring, researchers also propose other metrics of personalized web search that rely on clicking decisions, including Average Precision (AP), Rank Scoring and Average Rank[4] . We use the Average Precision metric to measure the effectiveness of the personalization in UPS. Meanwhile, our work is distinguished from previous studies as it also proposes two predictive metrics, namely personalization utility and privacy risk, on a profile instance without requesting for user feedback.

2.1.1 Privacy In Profile-Based PWS

To protect user privacy in profile-based PWS, two important and contradicting issues during the search process should be considered. The first issue is that, they attempt to improve the search quality with the personalization utility of the user profile. On the other hand, they need to hide the privacy contents existing in the user profile to place the privacy risk under control. Sometimes people are willing to compromise privacy if the personalization by supplying user profile to the search engine yields better search quality. In an identical situation,

significant gain can be obtained by personalization at the expense of only a small (and less-sensitive) portion of the user profile, namely a generalized profile. Thus, user privacy can be protected without compromising the personalized search quality. In general, there is a compromise between the search quality and the level of privacy protection achieved from generalization.[5]

2.2 Privacy Protection in PWS System

Generally there are two classes of privacy protection problems for PWS. One class includes those 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.[6] 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. Solution to the first level is proved to fragile. The third and fourth levels are impractical due to high cost in communication and cryptography. Therefore, the existing efforts focus on the second level. Using this approach, the linkage between the query and a single user is broken.

The solutions in class two do 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 an anonymity server. Statistical techniques use a probabilistic model, and then this model is used to generate near-optimal partial profile. One main limitation in this work [7] 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. A privacy protection solution for PWS is based on hierarchical profiles. Using a user-specified threshold, a generalized profile is obtained in effect as a rooted subtree of the complete profile. Unfortunately, this work does not address the query utility, which is crucial for the service quality of PWS. For comparison, our approach takes both the privacy requirement and the query utility into account. [8]

A more important property that distinguishes the proposed framework from [9] is that we provide personalized privacy protection in PWS. A person can specify the degree of privacy protection for her/his sensitive values by specifying "guarding nodes" in the taxonomy of the sensitive attribute. Motivate by this, we allow users to customize privacy needs in their hierarchical user profiles. [10]

2.2.1 Levels of Privacy Protection in Personalized Search

Privacy protection varies according to users requirements. Sometimes users may not want anyone else to know or hold any of their personal information, while some users may be willing to share some personal information for better search results or services. Thus the level of privacy protection is required to be given for different users to accommodate different preferences for the tradeoffs of personalization and privacy protection.

2.2.1.1 Level I: Pseudo Identity

A personalized web search system has Level I privacy protection (Pseudo Identity) if:
i) The user identity $ID(U)$ is replaced by a pseudo identity $IDp(U)$ which contains less personally identifiable information than $ID(U)$ does and hence supporting privacy protection.
ii)The description of user information needs $TEXT(N; i)$ can be aggregated according to $IDp(U)$ at the search engine side.

$ID(U)$ can generally be mapped to a single or a small group of users (e.g., family members) with the help of public databases. For example, given an IP address, geographic information such as city and state can be known. With a pseudo identity $IDp(U)$, such mapping is not available and some personal information such as the location of the user is protected. From the viewpoint of personalized search, a pseudo identity $IDp(U)$ can still be used to group all the descriptions of user information needs to build a user profile without needing $ID(U)$[5]. The content of user profile such as queries and click through is intact at the search engine side.

Level I is the lowest level of privacy protection. Because of the removal of $ID(U)$, which may otherwise be used to directly identify a user, some people who do not care much about privacy may accept this level of privacy protection. Unfortunately, this level is not enough to protect a user's privacy because it allows aggregation of all the information need descriptions of a user, which can in turn facilitate identification of the user. Since queries directly indicate a user's interests, being able to group many queries from the same user makes it quite possible to identify a user.

2.2.1.2 Level II: Group Identity

A personalized web search system has Level II privacy protection (Group Identity) if:
i) A group of users share a single user identity $ID(U)$.
ii) The description of user information needs $TEXT(N; i)$ is aggregated at the group level according to $ID(U)$.

This level of protection is achieved when a group of users send their profiles to the search

engine in such a way that the search engine can only build a group user profile for the group instead of a user profile for each single user.In this case, personalized web search can not be done at the individual user level, but is possible at the group level. This may reduce the effectiveness of personalization because a group's information need description is used to model an individual user's information need. However, if the group is appropriately constructed so that people with similar interests are grouped together, we may have much richer user information to offset the sparse description of individual user information needs. Thus the search performance may actually be improved because of the availability of more information from the group profile.

Level II has higher privacy protection than Level I. At this level, one cannot construct an individual user profile. Instead, only an aggregated profile for a group of users can be constructed.

Since the identity information of an individual user *ID*(*U*) is lost in a group of identity, and the description of user information needs *TEXT*(*N; i*) is also mixed with those of other users, it is difficult to infer true information needs of any individual user if the group is appropriately constructed.

A common way to implement the Level II privacy protection is to set up a proxy for a group of users and all the users would communicate with the search engine through the proxy. Currently, there are many public proxy servers available on the Internet.

2.2.1.3 Level III: No Identity

A personalized web search system has Level III privacy protection (No Identity) if:
i) The user identity *ID*(*U*) is not available to the search engine.
ii) The description of user information needs *TEXT*(*N; i*) can not be aggregated on the search engine side, even at the group level.

At Level III, a search engine can not know *ID*(*U*) of individual users at all, thus it has no way to aggregate the description of user information needs. At this level, however, it would be impossible to build a user profile on the search engine side, even at the group level. Since the search engine does not have the user profile, personalized search must be supported on a user's own computer. Specifically, the user profile *P*(*U*) can be kept on the personal computer of the user *U*.

Personalized search can be achieved by combining general Web search with a local, personalized reranking of results. A possible way to implement Level III privacy protection is through the anonymous network.

Level III has a higher privacy protection than Level II. At Level III, it is impossible for the search engine to aggregate any information about the individual user, even at the group level. However, some user information is still kept at the search engine side. For example, the original user queries may be kept at the search engine side. Although a user's query generally does not explicitly contain personal identity *ID*(*U*), it sometimes contains quite sensitive information (It is known that some queries contain social security numbers.) It is thus still possible to infer a user's identity just from a query.

2.2.1.4 Level IV: No Personal Information

A personalized web search system has Level IV privacy protection (No Personal Information) if:
i)Neither the user identity *ID*(*U*) nor the description of user information need *TEXT*(*N*) is available to the search engine.

At Level IV, a search engine does not know *ID*(*U*) of an individual user or the description of user information need *TEXT*(*N*) at all. However, the search engine can still return the normal search results to the correct user. Thus the user privacy is fully protected. On the surface, it appears to be impossible to achieve this level of privacy protection. However, cryptography methodology may be applied to realize this ultimate level of privacy protection.

Another possibility for achieving the Level IV privacy protection is that a search engine would be required by law to guarantee that it does not store any user information (*ID*(*U*) or *TEXT*(*N*)). That is, the search engine will have no memory of any activity of a user, even though it would still respond to a user search request directly. This scenario can be considered to be equal to the scenario that the search engine does not know any information about the user. As in the case of Level III privacy protection, since a search engine cannot construct any kind of user profile, personalized search must be supported on the user's computer.

Level IV has the highest level of privacy protection for personalized search. However, it may also have the highest cost due to higher communication cost and encryption/decryption cost, which will delay real-time response.

3. ANALYSIS OF PROBLEM

The problems with the existing methods are explained in the following observations:[3]

1. The existing profile-based PWS do not support runtime profiling. A user profile is typically generalized for only once offline, and used to personalize all queries from a same user indiscriminatingly. Such "one profile fits all" strategy certainly has drawbacks given the variety of queries.

It is proved that Profile-based personalization may not even help to improve the search quality for some adhoc queries, though exposing user profile to a server has put the user's privacy at risk. A better approach is to make an online decision on:

a. whether to personalize the query (by exposing the profile) and

b. what to expose in the user profile at runtime. Until now no previous work has supported such feature.

2. The existing methods do not take into account the customization of privacy requirements. This probably makes some user privacy to be overprotected while others insufficiently protected. For example, in all the sensitive topics are detected using an absolute metric called surprised based on the information theory, assuming that the interests with less user document support are more sensitive.

3. Many personalization techniques require iterative user interactions when creating personalized search results. They usually refine the search results with some metrics which require multiple user interactions, such as rank scoring, average rank, and so on[4].This paradigm is, however, infeasible for runtime profiling, as it will not only pose too much risk of privacy breach, but also demand prohibitive processing time for profiling. Thus, we need predictive metrics to measure the search quality and breach risk after personalization, without incurring iterative user interaction.

4. OBJECTIVES

The main objectives of proposed framework are summarized below:

i) The proposed a privacy-preserving personalized web search framework UPS, which can generalize profiles for each query according to user-specified privacy requirements.

ii) Relying on the definition of two conflicting metrics, namely personalization utility and privacy risk, for hierarchical user profile, the proposed framework provides privacy-preserving personalized search as –Risk Profile Generalization, with its NP-hardness proved.

iii) The framework provides an inexpensive mechanism for the client to decide whether to personalize a query in UPS. This decision can be made before each runtime profiling to enhance the stability of the search results while avoid the unnecessary exposure of the profile.

5. PROPOSED WORK

The problems in existing methods are addressed in our UPS framework. The framework assumes that the queries do not contain any sensitive information, and aims at protecting the privacy in individual user profiles while retaining their usefulness for PWS.

5.1 System Architecture of UPS

User Customizable Privacy Preserving Search UPS is distinguished from conventional Personalised Web Search in that it:

1) Provides runtime profiling, which in effect optimizes the personalization utility while respecting user's privacy requirements;

2) Allows for customization of privacy needs;

3) Does not require iterative user interaction.

UPS consists of a non trusty search engine server and a number of clients. Each client (user) accessing the search service trusts no one but himself/ herself. The key component for privacy protection is an online profiler implemented as a search proxy running on the client machine itself. The proxy maintains both the complete user profile, in a hierarchy of nodes with semantics, and the user-specified (customized) privacy requirements represented as a set of sensitive nodes. [3]

The framework works in two phases, namely the offline and online phase, for each user. During the offline phase, a hierarchical user profile is constructed and customized with the user-specified privacy requirements. The online phase handles queries as follows:

1. When a user issues a query qi on the client, the proxy generates a user profile in runtime in the light of query terms. The output of this step is a generalized user profile Gi satisfying the privacy requirements. The generalization process is guided by considering two conflicting metrics, namely the personalization utility and the privacy risk, both defined for user profiles.

2. Subsequently, the query and the generalized user profile are sent together to the PWS server for personalized search.

3. The search results are personalized with the profile and delivered back to the query proxy.

4. Finally, the proxy either presents the raw results to the user, or reranks them with the complete user profile.

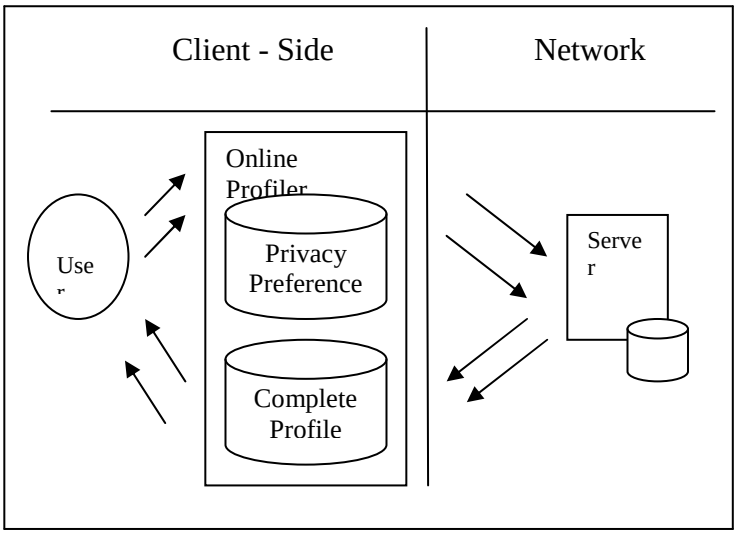

Fig.1: Proposed system of personalised web search

The framework for UPS focuses on structure of user profile and customized privacy requirement

5.2 User Profile

Consistent with many previous works in personalized web services, each user profile in UPS adopts a hierarchical structure. Moreover, our profile is constructed based on the availability of a public accessible taxonomy, denoted as R, which satisfies the following assumption.

Assumption 1.

The repository R is a huge topic hierarchy covering the entire topic domain of human knowledge. That is, given any human recognizable topic t, a corresponding node (also referred to as t) can be found in R, with the subtree subtr(t,R) as the taxonomy accompanying t.

The repository is regarded as publicly available and can be used by anyone as the background knowledge.

Assumption 2.

Given a taxonomy repository R, the repository support is provided by R itself for each leaf topic.

Definition 1 (USER PROFILE/H).

A user profile H, as a hierarchical representation of user interests, is a rooted subtree of R. The notion rooted subtree is given in Definition 2.

Definition 2 (ROOTED SUBTREE).

Given two trees S and T , S is a rooted subtree of T if S can be generated from T by removing a node from T

A diagram of a sample user profile is illustrated in Fig. 2a, which is constructed based on the sample taxonomy repository in Fig. 2b. We can observe that the owner of this profile is mainly interested in Computer Science and Music, because the major portion of this profile is made up of fragments from taxonomies of these two topics in the sample repository. Some other taxonomies also serve in comprising the profile, for example, Sports and Adults.

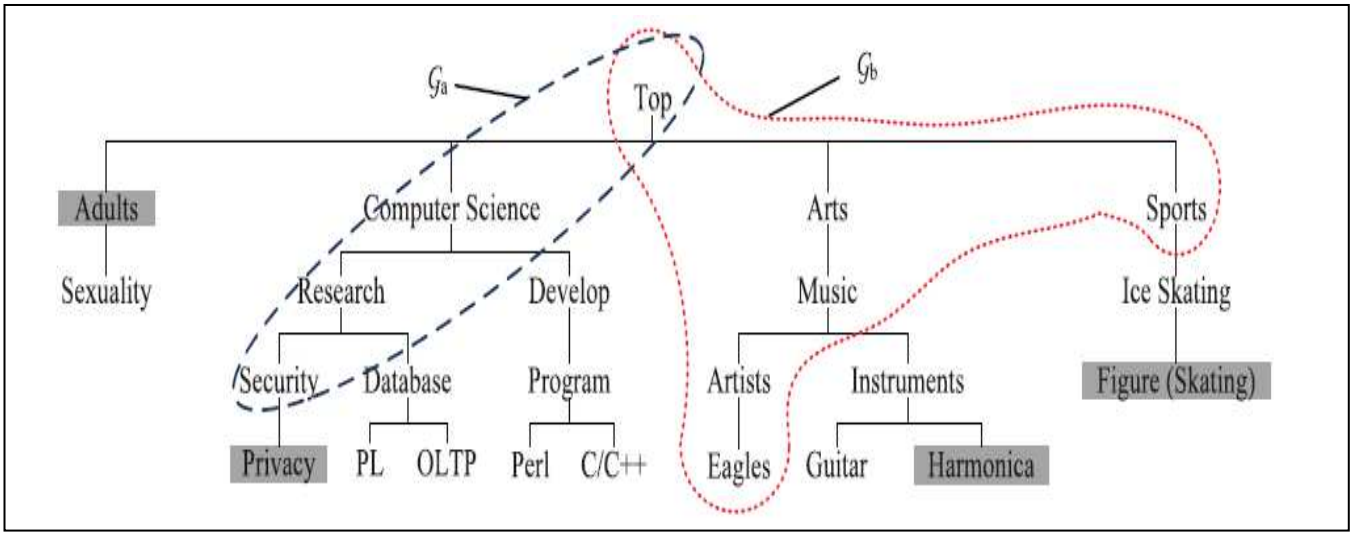

Fig: 2 :a)Sample Use Profile

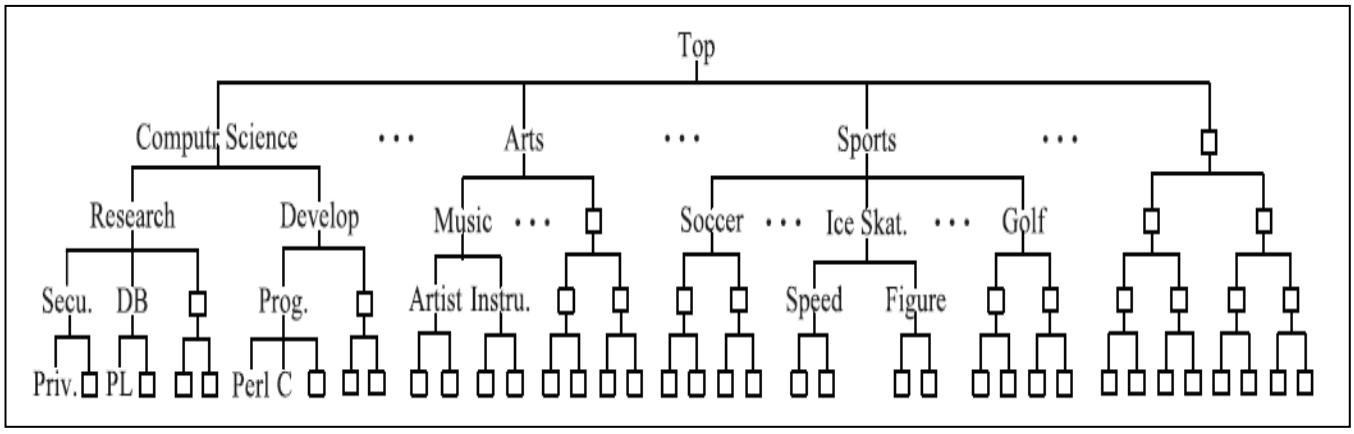

Fig. 2 : b) Sample Taxonomy Repository

5.3 Customized Privacy Requirements

Customized privacy requirements can be specified with a number of sensitive-nodes (topics) in the user profile, whose disclosure (to the server) introduces privacy risk to the user.

Definition 3 (SENSITIVE NODES/S).

Given a user profile H, the sensitive nodes are a set of user specified sensitive topics. In the sample profile shown in Fig. 2a, the sensitive nodes S={Adults; Privacy;Harmonica; Figure (Skating) are shaded in gray color in H.

It must be noted that user's privacy concern differs from one sensitive topic to another. In the above example, the user may hesitate to share her personal interests (e.g., Harmonica, Figure Skating) only to avoid various advertisements. Thus, the user might still tolerate the exposure of such interests to trade for better personalization utility. However, the user may never allow another interest in topic Adults to be disclosed. To address the difference in privacy concerns, we allow the user to specify a sensitivity for each node.[11]

Definition 4 (SENSITIVITY/sen(s).

Given a sensitive-node s, its sensitivity, i.e., sen(s), is a positive value that quantifies
the severity of the privacy leakage caused by disclosing s. As the sensitivity values explicitly indicate the user's privacy concerns, the most straightforward privacy preserving method is to remove subtrees rooted at all sensitive-nodes whose sensitivity values are greater than a threshold. Such method is referred to as forbidding.

Conclusion

The remarkable development of information on the Web has forced new challenges for the construction of effective search engines. Personalized search is a promising way to improve the accuracy of web search, and has been attracting much attention recently. However, effective personalized search requires collecting and aggregating user information, which often raises serious concerns of privacy infringement for many users. This seminar provides information on User customizable Privacy preserving Search framework-UPS for Personalized Web Search. UPS could potentially be adopted by any PWS that captures user profiles in a hierarchical taxonomy. The framework allowed users to specify customized privacy requirements via the hierarchical profiles while respecting user specified privacy requirements.

REFERENCES

[1] 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

[2] 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

[3] Lidan Shou, He Bai, Ke Chen, and Gang Chen,"Supporting Privacy Protection in Personalized Web Search", IEEE Transactions On Knowledge And Data Engineering Vol:26 No:2 Year 2014

[4] J.Jayanthi, M.Ezhilmathi, S. Rathi, " Evaluating the Effectiveness of Web Search Metrics" International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 4– No.10, December 2012

[5] Xuehua Shen, Bin Tan, ChengXiang Zhai," Privacy Protection in Personalized Search", ACM SIGIR Forum , Vol.41 No.1 June 2007

[6] Yabo Xu, Benyu Zhang, Zheng Chen, Ke Wang, "Privacy-Enhancing Personalized Web Search" 2007

[7] Avi Arampatzis, Pavlos Efraimidis, and George Drosatos,"Enhancing Deniability against Query-Logs", ECIR 2011, LNCS 6611, pp. 117–128, Springer-Verlag Berlin Heidelberg 2011

[8] Jordi Castell`a-Roca, Alexandre Viejo, Jordi Herrera-Joancomart, "Preserving User's Privacy in Web Search Engines", Elsevier Preprint, 16 May 2009

[9] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2006.

[10] 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.

[11] T.Sathiyabama, Dr. K. Vivekanandan, "Personalized Web Search Techniques –A Review",Global Journal of Computer Science and Technology Volume 11 Issue 12 July