



A SURVEY ON PRIVACY-ENHANCING PERSONALIZED WEB SEARCH

Seerapu Divya Bhanu Sri¹, B. Hema Nagamani²

M.Tech Student¹, Asst Professor², Dept of CSE

V.S.Lakshmi Engg College for Womens, Matlapalem, Kakinada, India

ABSTRACT

Personalized web search is a promising way to improve search quality by customizing search results for people with individual information goals. However, users are uncomfortable with exposing private preference information to search engines. On the other hand, privacy is not absolute, and often can be compromised if there is a gain in service or profitability to the user. Thus, a balance must be struck between search quality and privacy protection. This paper presents a scalable way for users to automatically build rich user profiles. These profiles summarize a user's interests into a hierarchical organization according to specific interests. Two parameters for specifying privacy requirements are proposed to help the user to choose the content and degree of detail of the profile information that is exposed to the search engine. Experiments showed that the user profile improved search quality when compared to standard MSN rankings. More importantly, results verified our hypothesis that a significant improvement on search quality can be achieved by only sharing some higher-level user profile information, which is potentially less sensitive than detailed personal information.

Keywords: *privacy, personalized search, hierarchical user profile*

I.INTRODUCTION

As the amount of information on the web continuously grows, it has become increasingly difficult for web search engines to find information that satisfies users' individual needs. Personalized search is a promising way to improve search quality by customizing search results for people with different information goals. Many recent research efforts have focused on this area. Most of them could be categorized into two general approaches: Re-ranking query results returned by search engines locally using personal information; or sending personal information and queries together to the search engine [1]. A good personalization algorithm relies on rich user profiles and web corpus. However, as the web corpus is on the server, re-ranking on the client side is bandwidth intensive because it requires a large number of search results transmitted to the client before re-ranking. Alternatively, if the amount of information transmitted is limited through filtering on the server side, it pins high hope on the existence of desired information among filtered results, which is not always the case. Therefore, most of personalized search services online like Google Personalized Search [2] and Yahoo! My Web[3] adopt the second approach to tailor results on the server by analyzing collected personal information, e.g. personal interests, and search histories. Nonetheless, this approach has privacy issues on exposing personal information to a public server. It usually requires users to grant the server full access to their personal and behavior information on the Internet. Without the user's permission, gleaning such information would violate an individual's privacy. In particular, Canada launched the Personal Information Protection and Electronic Document Act 1 in 2001 to protect a wide spectrum of information, i.e., age, race, income, evaluations, and even intentions to acquire goods or services from being released to outside parties. It is also evidenced by a recent survey conducted by Choicestream2 that the privacy fear continues to escalate although personalization remains something most consumers want. The number of consumers interested in personalization remains at a remarkably high 80%; however, only 32% of respondents were willing to share personal information in exchange for personalized experience, down from 41% in 2004. Recent coverage about identity thefts and online security breaches, i.e. AOL search query data scandal, even causes users to be more wary than ever on sharing their private information—even with established, trusted brands.. In practice, however, privacy is not



absolute. There exist already many examples where people give up some privacy to gain economic benefit. One example is frequent shopper card in grocery stores. Consumers trade the benefit of extra saving in the grocery stores versus the creation of a detailed profile of their shopping behavior. As another example, consider a basketball fan. He may not be comfortable broadcasting a weekly work-out schedule, but might not mind revealing an interest on basketball if a search engine can help identify “Rockets” as an NBA team instead of anything related to space exploration. Thus, people may compromise some personal information if this yields them some gain in service quality or profitability. Another important observation is that detailed personal information might not be necessary if it is possible to catch a user’s interests at more general level. In the above example, the times and locations where the user has played basketball would not be relevant in searching for a favorite NBA basketball team. In fact, such unnecessarily detailed information often becomes noise in the search task. Hence, a proper filtering of a user’s private information not only helps protect the user’s privacy but also may help improve the search quality. The key is distinguishing between useful information and noise, as well as striking balance between search quality and privacy protection. Personal data, i.e. browsing history, emails, etc., are mostly unstructured, for which it is hard to measure privacy. In addition, it is also difficult to incorporate unstructured data with search engines without summarization. So, for the purpose of both web personalization and privacy preservation, it is necessary for an algorithm to collect, summarize, and organize a user’s personal information into a structured user profile. Meanwhile, the notion of privacy is highly subjective and depends on the individuals involved. Things considered to be private by one person could be something that others would love to share. In this regard, the user should have control over which parts of the user profile is shared with the server.

This paper targets at bridging the conflict needs of personalization and privacy protection, and provides a solution where users decide their own privacy settings based on a structured user profile. This benefits the user in the following ways:

- Offers a scalable way to automatically build a hierarchical user profile on the client side. It’s not realistic to require that every user to specify their personal interests explicitly and clearly. Thus, an algorithm is implemented to automatically collect personal information that indicates an implicit goal or intent. The user profile is built hierarchically so that the higher-level interests are more general, and the lower-level interests are more specific. In this approach, a rich pool of profile sources is explored including browsing histories, emails and personal documents.
- Offers an easy way to protect and measure privacy. With a hierarchical user profile, the exposure of private information is controlled using two parameters. minDetail determines which part of user profile is protected. Interests in the user profile that does not satisfy minDetail are either too specific or uncommon, are considered private and hidden from the server. expRatio measures how much private information is exposed or protected for a specified minDetail.

The paper is organized as follows: Section 2 reviews related work focusing on personalized search and privacy issues. An overview of the problem is given in Section 3. Our approach is described in Section 4. Experiment results are presented in Section 5. Conclusions are presented in Section.

II.RELATED WORK

In information retrieval, much research is focused on personalized search. Relevance feedback and query refinement [13] [14] harnesses a short-term model of a user’s interests, and information about a user’s intent is collected at query time. Personal information has also been used in the context of Web search to create a personalized version of PageRank [5] [6]. There are still approaches, including many commercially available informationfiltering systems [9] [10], which require users explicitly specify their interests. However, as [13] pointed out, users are typically unwilling to spend the extra effort on specifying their intentions. Even if they are motivated, they are not always successful in doing so. A majority of work focuses on implicitly building user profiles to infer a user’s intention. A wide range of implicit user activities have been proposed as sources of enhanced search information. This includes a user’s search history [12], browsing history [7], click-through data [18] [28], web community [12] [15], and rich client side information [8] in the form of desktop indices. Our approach is open to all kinds of different data sources for building user profiles, provided the sources can be

extracted into text. In our experiments data sources like IE histories, emails and recent personal documents were tested. User profiles can be represented by a weighted term vector [7], weighted concept hierarchical structures [10] [12] like ODP3 , or other implicit user interest hierarchy [11]. For the purposes of selectively exposing users' interests to search engines, the user profile is a term based hierarchical structure that is related to frequent term based clustering algorithms [16][17]. The difference here is that the hierarchical structure is implicitly constructed in a top-down fashion. And the focus is the relationships among terms, not clustering the terms into groups. Privacy concerns are natural and important especially on the Internet. Some prior studies on Private Information Retrieval (PIR) [4], focuses on the problem of allowing the user to retrieve information while keeping the query private. Instead, this study targets preserving privacy of the user profile, while still benefiting from selective access to general information that the user agrees to release. To our knowledge, this problem has not been studied in the context of personalized search. One possible reason for this is that personal information, i.e. browsing history and emails, is mostly unstructured data, for which privacy is difficult to measure and quantify. Some works on privacy issues in the data mining community focus on protecting individual data entries while allowing information summarization. A popular way of measuring privacy in data mining is by examining the difference in prior and posterior knowledge of a specific value [19] [20]. This can be formalized as the conditional probability or Shannon's information theory. Another way to measure privacy is the notion of k-anonymity [21] which advocates that personally identifying attributes be generalized such that each person is indistinguishable from at least k-1 other persons. In this study the notion of privacy does not compare information from different users, but rather the information collected over time for a single user. In addition, this study addresses unstructured data.

III.PROBLEM OVERVIEW

Personal data, i.e. personal documents, browsing history and emails might be helpful to identify a user's implicit intents. However, users have concerns about how their personal information is used. Privacy, as opposed to security or confidentiality, highly depends on the person involved and how that person may benefit from sharing personal information. The question here is whether a solution can be found where users themselves are able to set their own privacy levels for user profiles to improve the search quality.

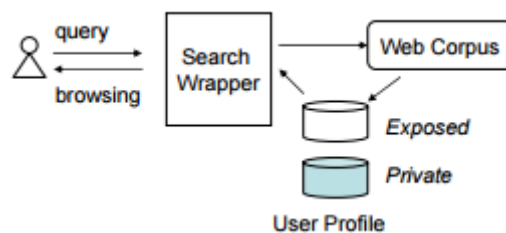


Figure1. System Overview

Figure 1 provides an overview of the whole system. An algorithm is provided for the user to automatically build a hierarchical user profile that represents the user's implicit personal interests. General interests are put on a higher level; specific interests are put on a lower level. Only portions of the user profile will be exposed to the search engine in accordance with a user's own privacy settings. A search engine wrapper is developed on the server side to incorporate a partial user profile with the results returned from a search engine. Rankings from both partial user profiles and search engine results are combined. The customized results are delivered to the user by the wrapper. The solution has three parts: First, a scalable algorithm automatically builds a hierarchical user profile from available source data. Then, privacy parameters are offered to the user to determine the content and amount of personal information that will be revealed. Third, a search engine wrapper personalizes the search results with the help of the partial user profile.

IV.PRIVACY-ENHANCING PERSONALIZED SEARCH

4.1 Constructing a Hierarchical User Profile

Any personal documents such as browsing history and emails on a user’s computer could be the data source for user profiles. Our hypothesis is that terms that frequently appear in such documents represent topics that interest users. This focus on frequent terms limits the dimensionality of the document set, which further provides a clear description of users’ interest. This approach proposes to build a hierarchical user profile based on frequent terms. In the hierarchy, general terms with higher frequency are placed at higher levels, and specific terms with lower frequency are placed at lower levels. D represents the collection of all personal documents and each document is treated as a list of terms. D(t) denotes all documents covered by term t, i.e., all documents in which t appears, and |D(t)| represents the number of documents covered by t. A term t is frequent if |D(t)| ≥ minsup, where minsup is a user-specified threshold, which represents the minimum number of documents in which a frequent term is required to occur. Each frequent term indicates a possible user interest. In order to organize all the frequent terms into a hierarchical structure, relationships between the frequent terms are defined below. Assuming two terms tA and tB., the two heuristic rules used in our approach are summarized as follows:

1. Similar terms: Two terms that cover the document sets with heavy overlaps might indicate the same interest. Here we use the Jaccard function [27] to calculate the similarity between two terms: $Sim(t_A, t_B) = \frac{|D(t_A) \cap D(t_B)|}{|D(t_A) \cup D(t_B)|}$. If $Sim(t_A, t_B) > \delta$, where δ is another user-specified threshold, we take tA and tB as similar terms representing the same interest.

2. Parent-Child terms: Specific terms often appear together with general terms, but the reverse is not true. For example, “badminton” tends to occur together with “sports”, but “sports” might occur with “basketball” or “soccer”, not necessarily “badminton”. Thus, tB is taken as a child term of tA if the condition probability $P(t_A | t_B) > \delta$, where δ is the same threshold in Rule 1.

Rule 1 combines similar terms on the same interest and Rule 2 describes the parent-child relationship between terms. Since $Sim(t_A, t_B) \leq P(t_A | t_B)$, Rule 1 has to be enforced earlier than Rule 2 to prevent similar terms to be misclassified as parent-child relationship. For a term tA, any document covered by tA is viewed as a natural evidence of users’ interests on tA. In addition, documents covered by term tB that either represents the same interest as tA or a child interest of tA can also be regarded as supporting documents of tA. Hence supporting documents on term tA, denoted as S(tA), are defined as the union of D(tA) and all D(tB), where either $Sim(t_A, t_B) > \delta$ or $P(t_A | t_B) > \delta$ is satisfied.

Using the above rules, our algorithm automatically builds a hierarchical profile in a top-down fashion. The profile is represented by a tree structure, where each node is labeled a term t, and associated with a set of supporting documents S(t), except that the root node is created without a label and attached with D, which represent all personal documents. Starting from the root, nodes are recursively split until no frequent terms exist on any leaf nodes. Below is an example of the process. Before running the algorithm on the documents, preprocessing steps like stop words removal and stemming needs to be performed first. For simplification, each document is treated as a list of terms after preprocessing.

D1:sports, badminton
D2:ronaldo,soccer,sports
D3:sex, playboy, picture
D4:sports,soccer,english premier
D5:research, AI, algorithm
D5:research, AI, algorithm
D7:Fox, channel, sports, sex
D8:MSN,search
D9:research,AI,neuro network
D10:personalized,search,google,

Figure 2. An example data source

Three nodes “research”, “sports” and “sex” are left after the merging operations. As we mentioned earlier, every document in S(t) is regarded as a supporting document of term t. And the support of term t, contributed by all documents in S(t), is an indication of the degree of the user’s interest on t. For any document d in S(t), if d appears in n nodes ($n \geq 1$), which was interpreted as d supporting all n terms, the support from d in S(t) is counted only as 1/n. This guarantees the sum of support contributed by each document equals to 1 in spite of the number

of terms it supports. Thus the support of a term t , denoted as $Sup(t)$, is calculated as the sum of the supports from all documents in $S(t)$. In this example, $D7$ appears in both $S(\text{"sports"})$ and $S(\text{"sex"})$, so $Sup(\text{"sports"})=1+1+1+1/2=3.5$, and $Sup(\text{"sex"})=1.5$.

A diagram of the user profile after the first splitting is shown in Figure 3, where the term t and its support $Sup(t)$ are attached to each cluster, with the supporting documents $S(t)$ listed below. Each node on the same level is sorted by $Sup(t)$ in a descending order.

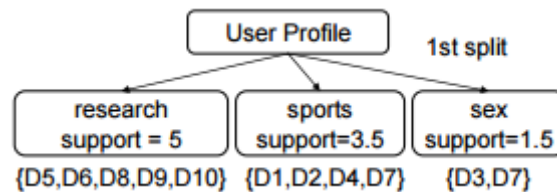


Figure 3. User profile after 1st split.

The node “research” is subsequently examined for further splitting. First $S(\text{"research"})$ is scanned, and the frequency for each term t is counted. Note that any term like “research” that appears in an ancestor node will not be counted again. Frequent terms and their frequency are listed as follows: , , . According to Rule 2, “search” and “personalized” is combined together and the node is labeled “personalized/search” since $Sim(\text{"search"}, \text{"personalized"}) = 2/3 > \delta$. The child nodes after splitting are shown in Figure 4. The splitting can be recursively done until no term is frequent.

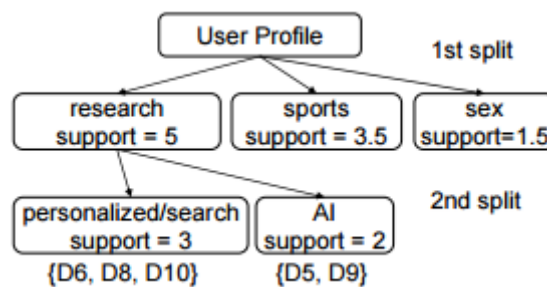


Figure 4. User profile after 2nd split

The formal algorithms are described in Figure 5. $Split(n, S(t), minsup, \delta)$ is called to split a node n . Rule 1 is enforced in line 3-4, and Rule 2 is enforced in line 5-6. In line 9, nodes are sorted in a descending order of the support of term t_i . The reason will be explained in section 4.2. A complete user profile is constructed by calling $BuildUP(\text{root}, D, minsup, \delta)$, where root represents the root node, and D is the set containing all personal documents. $Split(n, S(t), minsup, \delta)$ are recursively applied on each node until no frequent term exists on any leave node.

4.2 Measuring Privacy

According to Alan Westin [23], “privacy is the claim of individuals, groups, or institutions to determine for themselves when, how and to what extent information is communicated to others”. Privacy per se is about protecting users’ personal information. However, it is users’ control that comprises the justification of privacy. With the complete user profile constructed above, an approach without any privacy risk is to grant users full control over the terms in the hierarchy so that they can choose to hide any terms manually as they desire. Unfortunately, studies have shown that the vast majority of users are always reluctant to provide any explicit input on their interests [24]. In order to offer users a more convenient way of controlling private information they would agree to have exposed, two parameters derived from information theory are proposed below. In the following discussion, “interest” and “term” are indistinguishable in the context of the user profile. The support of an interest or a term t is $Sup(t)$, and $S(t)$ represents all the supporting documents for term t . $\sum Sup(t)=|D|$ is for all terms t on the leave node, where $|D|$ represents the total number of supports received from personal data.

Interestingly, this measure matches perfectly with our following observations on users’ privacy concern: the interest with large self-information corresponds to two types of information to which users are usually sensitive to grant access to. The first case is that the interest itself is too specific. Users might not mind telling others

about general interests, i.e. a user likes basketball, but is cautious about letting others know his weekly basketball schedule. The second case is that the interest is general but less popular among all interests. It might represent a private event, i.e. the category “sex” in Example 1. The idea is to protect private information that is either too specific or too sensitive in the user profile. Both kinds could be measured by the support of the interest, under the assumption that the more specific or sensitive the interest is, the larger self-information the interest will carry. This leads to the two parameters for specifying the requirement of privacy protection.

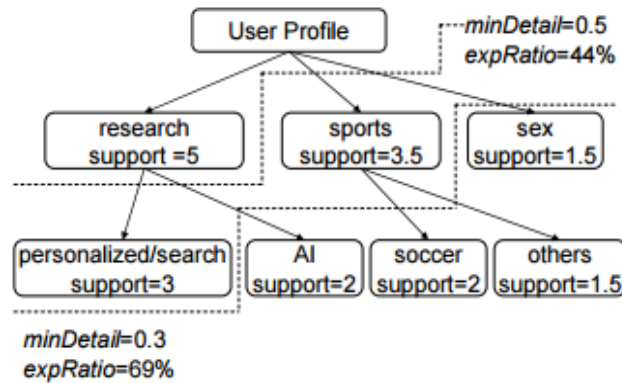


Figure 5. Fully extended user profile

The complete user profile is denoted as U , and $U[exp]$ represents the exposed part of U , or the part above $minDetail$. Since the support for terms decreases monotonically traveling horizontally and vertically, the $U[exp]$ will be a connected subtree of the complete user profile stemming from the user profile root. With the threshold $minDetail$, the user will know exactly which part of the user profile is protected.

4.3 Personalizing Search Results

In order to incorporate the user profile with results returned by a search engine, $U[exp]$ is transformed into a list of weighted terms where a search wrapper calculates a score for each of the returned search results. The final ranking of the search results is decided by the search engine and $U[exp]$. The weight of each term in $U[exp]$ is estimated by applying the concept of IDF(Inverse Document Frequency)Error! Reference source not found.. Given a term t , the weight of t , denoted by w_t , is calculated as:

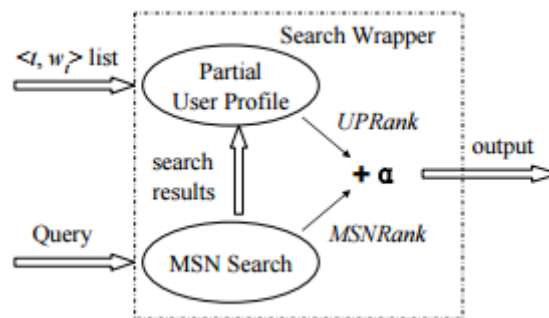


Figure 6. The workflow in the search wrapper

V.CONCLUSIONS AND FUTURE WORK

Personalized search is a promising way to improve search quality. However, this approach requires users to grant the server full access to personal information on Internet, which violates users’ privacy. In this paper, we investigated the feasibility of achieving a balance between users’ privacy and search quality. First, an algorithm is provided to the user for collecting, summarizing, and organizing their personal information into a hierarchical user profile, where general terms are ranked to higher levels than specific terms. Through this profile, users control what portion of their private information is exposed to the server by adjusting the $minDetail$ threshold. An additional privacy measure, $expRatio$, is proposed to estimate the amount of privacy is exposed with the

specified minDetail value. Experiments showed that the user profile is helpful in improving search quality when combined with the original MSN ranking. The experimental results verified our hypothesis that there is an opportunity for users to expose a small portion of their private information while getting a relatively high quality search. Offering general information has a greater impact on improving search quality.



This paper is an exploratory work on the two aspects: First, we deal with unstructured data such as personal documents, for which it is still an open problem on how to define privacy. Secondly, we try to bridge the conflict needs of personalization and privacy protection by breaking the premise on privacy as an absolute standard. There are a few of promising directions for future work. In particular, we are considering ways of quantifying the utility that we gain from personalization, thus users can have clear incentive to compromise their privacy. Also, we suspect that an improved balance between privacy protection and search quality can be achieved if web search are personalized by considering.

REFERENCES

- [1] J. Pitkow, H. Schuetze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel. Personalized search. *Communications of the ACM*, 45(9):50-55, 2002.
- [2] Google personalized search: <http://www.google.com/psearch>
- [3] Yahoo! My Web 2.0: <http://myweb2.search.yahoo.com/>
- [4] W. Gasarch. A survey on private information retrieval. *The bulletin of the European Association for Theoretical Computer Science (EATCS)*, 82:72--107, 2004.
- [5] Glen Jeh, and Jennifer Widom. Scaling personalized web search. In *Proc. of the 12th International World Wide Web Conference (WWW)*, Budapest, Hungary, May 2003.
- [6] T.H. Haveliwala. Topic-sensitive PageRank. In *Proc. of the 11th International World Wide Web Conference (WWW)*, Honolulu, Hawaii, May 2002.
- [7] K. Sugiyama, K. Hatano and M. Yoshikawa. Adaptive Web search based on user profile constructed without any effort from users, In *Proc. of the 13th International World Wide Web Conference (WWW)*, New York, New York, May 2004.
- [8] J.Teevan, S. T. Dumais, and Eric Horvitz. Personalizing search via automated analysis of interests and activities. In the *Proc. of 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, Salvador, Brazil. August, 2005
- [9] Paolo Ferragina, and Antonio Gulli. A personalized search engine based on Web-Snippet hierarchical clustering. In *Proc. of the 14th International World Wide Web Conference (WWW)*, Chiba, Japan, May 2005.
- [10] P. A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschutter. Using ODP metadata to personalize search. In the *Proc. of 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, Salvador, Brazil, August, 2005
- [11] H.R. Kim, and Philip K. Chan. Learning implicit user interest hierarchy for context in personalization. In *Proc. of International Conference on Intelligent User Interface (IUI)*, Miami, Florida, January, 2003.
- [12] M. Speretta, and S. Gauch, Personalizing search based on user search history. In *Proc. of International Conference of Knowledge Management(CIKM)*, Washington D.C., 2004
- [13] P.Anick. Using terminological feedback for Web search refinement: a log-based study. In *Proc. of the 13th International World Wide Web Conference (WWW)*, New York, New York, May 2004.
- [14] K.R. McKeown, N. Elhadad, and V. Hatzivassiloglou. Leveraging a common representation for personalized search and summarization in a medical digital library. In *Proc. of International Conference on Digital Library*, 2003
- [15] A. Kritikopoulos, and M. Sideri. The compass Filter: Search engine result personalization using web communities. In *Proc. of Intelligent Techniques in Web Personalization (ITWP)*, 2003.



AUTHOR'S PROFILE

	<p>SEERAPU DIVYA BHANU SRI is a student of V.S.LAKSHMI ENGINEERING COLLEGE FOR WOMENS. Presently he is pursuing M.Tech [Computer Science and Engineering] from this college and he also completed his B.Tech .</p>
	<p>Mrs.B.Hema Nagamani, working as Asst.Professor in the Dept.of Computer Science and Engineering from V.S.Lakshmi Engineering College, Matlapalem, Kakinada.</p>