IJRSET Volume 2, Issue 4 Pages: 1-8

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

SECURED AN IMPLEMENTATION OF PERSONALIZED WEB SEARCH

[1] Mr. A. MANIKANDAN, [2] Dr. A.VIJAYA KATHIRAVAN

[1] Research Scholar, [2] Assistant Professor in Computer Science,

[1] Dept of Computer Science, [2] Dept of Computer Science

[1] Government Arts College, [2] Government Arts College

[1] Salem-636007, [2] Salem- 636007.

Abstract:-

Personalized web search is one of the growing concepts in the web technologies. Personalization of web search is to carry out retrieval for each user incorporating his/her interests. For a given query, a personalized Web search can provide different search results for different users or organize search results differently for each user, based upon their interests, preferences, and information needs. There are many personalized web search algorithms for analyzing the user interests and producing the outcome quickly; Personalized web search (PWS) has demonstrated its effectiveness in improving the quality of various search services on the Internet. However, display show that users' reluctance to disclose their private information during search has become a major barrier for the wide proliferation of SPWS. We study privacy protection in SPWS applications that model user preferences as hierarchical user profiles. We propose a SPWS framework called UPS that can adaptively generalize profiles by queries while respecting user specified privacy exaction. Our runtime generalization aims at striking a balance between two predictive metrics that evaluate the utility of personalization and the privacy risk of exposing the abstraction profile. We present Greedy Algorithm and Rating algorithm, for runtime generalization. We also provide an online prediction mechanism for deciding whether personalizing a query is beneficial. Extensive experiments exhibition the effectiveness of our framework. The empirical results also reveal that Symmetric key and new Advanced Encryption Standard (AES) in terms of efficiency.

Keywords: - Web, Search, Data Mining, UPS, Framework, AES.

1. INTRODUCTION

Data mining is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. It has been defined as, The automated analysis of large or complex data sets in order to discover significant patterns or trends that would otherwise go unrecognized. The goal of data mining is to unearth relationships in data that may provide useful insights. Data mining tools can sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions, performance bottlenecks in a network system and identifying anomalous data that could represent data entry keying errors. The ultimate significance of these patterns will be assessed by a domain expert - a marketing manager or network supervisor - so the results must be presented in a way that human experts can understand. Data mining tools can also automate the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to

identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events. Data mining techniques can yield the benefits of automation on existing software and hardware platforms to enhance the value of existing information resources, and can be implemented on new products and systems as they are brought on-line.The following diagram summarises the some of the stages/processes identified in data mining and knowledge discovery.

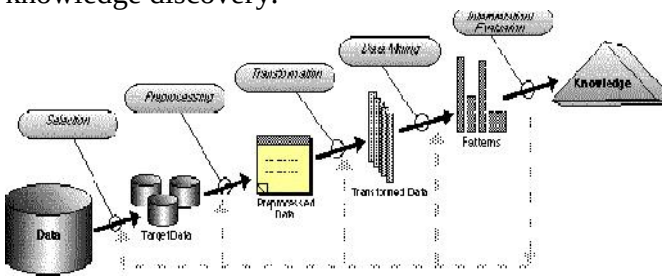


Figure 1.1: Knowledge Discovery Process

The phases depicted start with the raw data and finish with the extracted knowledge which was acquired as a result of the following stages:

Selection - selecting or segmenting the data according to some criteria e.g. all those people who own a car, in this way subsets of the data can be determined.

Preprocessing - this is the data cleansing stage where certain information is removed which is deemed unnecessary and may slow down queries for example unnecessary to note the sex of a patient when studying pregnancy. Also the data is reconfigured to ensure a consistent format as there is a possibility of inconsistent formats because the data is drawn from several sources e.g. sex may recorded as f or m and also as 1 or 0.

Transformation - the data is not merely transferred across but transformed in that overlays may added such as the demographic overlays commonly used in market research. The data is made useable and navigable.

Data mining - this stage is concerned with the extraction of patterns from the data. A pattern can be defined as given a set of facts(data) F, a language L, and some measure of certainty C a pattern is a statement S in L that describes relationships among a subset Fs of F with a certainty c such that S is simpler in some sense than the enumeration of all the facts in Fs.

Interpretation and evaluation - the patterns identified by the system are interpreted into knowledge which can then be used to support human decision-making e.g. prediction and classification tasks, summarizing the contents of a database or explaining observed phenomena. The key to understanding the different facets of data mining is to distinguish between data mining applications, operations, techniques and algorithms.

1.1. Personalization on the Web

Web personalization is a strategy, a marketing tool, and an art. Personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in your content delivery framework to manipulate what information you present to your users and how you present it. Correctly executed, personalization of the visitor's experience makes his time on your site, or in your application, more productive and engaging. Personalization can also be valuable to you and your organization, because it drives desired business results such as increasing visitor response or promoting customer retention. Unfortunately, personalization for its own sake has the potential to increase the complexity of your site interface and drive inefficiency into your architecture. It might even compromise the effectiveness of your marketing message or, worse, impair the user's experience. Few businesses are willing to sacrifice their core message for the sake of a few trick web pages. Contrary to popular belief, personalization doesn't have to take the form of customized content portals, popularized in the mid-to-late 90s by snap.com and My Yahoo!. Nor does personalization require expensive applications or live-in consultants. Personalization can be as blatant or as understated as you want it to be. It's a tired old yarn, but if you hope to implement a web personalization strategy, the first and most important step is to develop and mature your business goals and requirements. It is important to detail what it is you hope to do and, from that knowledge, develop an understanding of how you get from an idea to implementation. You might be surprised to discover that it won't require most of next year's budget to achieve worthwhile results.

Web personalization can be seen as an interdisciplinary field that includes several research domains from user modeling [14], social networks

[19], web data mining [8,13,19], human-machine interactions to Web usage mining[13]; Web usage mining is an example of approach to extract log files containing information on user navigation in order to classify users. Other techniques of information retrieval are based on documents categories' selection [13]. Contextual information extraction on the user and/or materials (for adaptation systems) is a technique fairly used also includes, in addition to user contextual information, contextual information of real-time interactions with the Web. [8] Proposed a multi-agent system based on three layers: a user layer containing users profiles and a personalization module, an information layer and an intermediate layer. They perform an information filtering process that reorganizes Web documents. [3] Propose reformulation query by adding implicit user information. This helps to remove any ambiguity that may exist in query: when a user asks for the term "conception", the query should be different if he is an architect or a computer science designer.

Requests can also be enriched with predefined terms derived from user's profile [8] develop a similar approach based on user categories and profiles inference. User profiles can be also used to enrich queries and to sort results at the user interface level [11]. Other approaches also consider social-based filtering [12] and collaborative filtering. These techniques are based on relationships inferred from users' profile. Implicit filtering is a method that observes user's behavior and activities in order to categorize classes of profile.

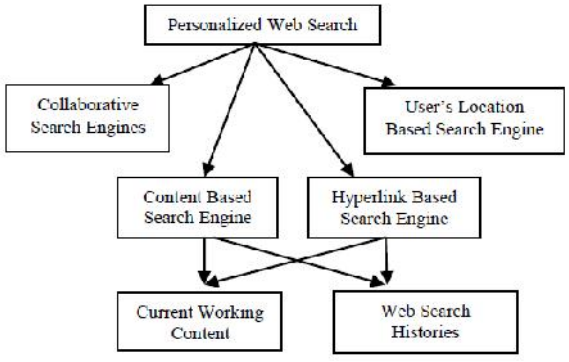


Figure 1.2: Personalized Web Search Approach

Collaborative Search Engines (CSEs) are an emerging trend for Web search and Enterprise search within company intranets. CSEs let users concert their efforts in information retrieval (IR) activities, share information resources collaboratively using knowledge tags, and allow experts to guide less experienced people through their searches. Collaboration partners do so by providing query terms, collective tagging, adding comments or opinions, rating search results, and links clicked of former (successful) IR activities to users having the same or a related information need. Personalized web search can be achieved by checking content similarity between web pages and user profiles. Some work has represented user interests with topical categories. User's topical interests are either explicitly specified by users themselves, or can be automatically learned by classifying implicit user data. Search results are filtered or re-ranked by checking the similarity of topics between search results and user profiles.

2. PROBLEM STATEMENT

The web search engine has long become the most important portal for ordinary people looking for useful information on the web. However, users might experience failure when search engines return irrelevant results that do not meet their real intentions. Such irrelevance is largely due to the enormous variety of users' contexts and backgrounds, as well as the ambiguity of texts. Personalized web search (PWS) is a general category of search techniques aiming at providing better search results, which are tailored for individual user needs. As the expense, user information has to be collected and analyzed to figure out the user intention behind the issued query. The solutions to PWS can generally be categorized into two types, namely click-log-based methods and profile-based ones. The click-log based methods are straightforward they simply impose bias to clicked pages in the user's query history. Although this strategy has been demonstrated to perform consistently and considerably well, it can only work on repeated queries from the same user, which is a strong limitation confining its applicability. In contrast, profile-based methods improve the search experience with complicated user-interest models generated from user profiling techniques. Profile-based methods can be potentially effective for almost all sorts of queries, but are reported to be unstable under some circumstances. Although there are pros and cons for both types of PWS techniques, the profile-based PWS has demonstrated more

effectiveness in improving the quality of web search recently, with increasing usage of personal and behavior information to profile its users, which is usually gathered implicitly from query history, browsing history, click-through data bookmarks, user documents , and so forth. Unfortunately, such implicitly collected personal data can easily reveal a gamut of user's private life. Privacy issues rising from the lack of protection for such data, for instance the AOL query logs scandal, not only raise panic among individual users, but also dampen the data-publisher's enthusiasm in offering personalized service. In fact, privacy concerns have become the major barrier for wide proliferation of PWS services.

3. LITERATURE REVIEW

Zhou et al. [16] proposed a hybrid index structure to handle both content and location-aware queries. The system first detects geographical scopes from web documents and represents the geographical scopes as multiple minimum bounding rectangles (MBRs) based on geographical coordinates. A hybrid index structure is used to index the content and location information of the web documents. A user is required to present their content and location interest in their search queries. A ranker is then employed to rank the search results according to the content and location relevance's using the hybrid index.

Allan et al. [3] define the problem of contextual retrieval as follows: "Combine search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information needs." Effective personalization of information access involves two important challenges: accurately identifying the user context and organizing the information in such a way that matches the particular context. Since the acquisition of user interests and preferences is an essential element in identifying the user context, most personalized search systems employ a user modeling component.

Liu et al. [17] utilize the first three levels of the ODP for learning profiles as bags of words associated with each category. The user's query is mapped into a small set of categories as a means to disambiguate the words in the query. The Web search is then conducted based on the user's original query and the set of categories. As opposed to using a set of categories, Chirita et al. [6] utilize the documents stored locally on a desktop PC for personalized query expansion.

Jeh and Widom [4] proposed a personalized web search by modifying the global PageRank algorithm. Instead of starting from random pages on the web, the "random surfer" starts from a set of preferred pages (such as bookmarks). Hence, the pages related to the preferred pages get higher PageRank score. Gauch and Pretschner [5] presented a system that allows for the automatic creation of structured user profile, and used the user profile to re-rank the search results, their user profiles were built based on an existing category hierarchy.

Gan et. al [8] suggested that search queries can be classified into two types, **content (i.e., non-geo)** and **location (i.e.,geo)**. Typical examples of geographic queries are .hotels hong kong., .building codes in seattle. and .virgina historical sites.. A classifier was built to classify geo and non-geoqueries, and the properties of geo queries were studied in detail. It was found that a significant number of queries were location queries focusing on location information. Hence, a number of location-based search systems designed for equerries have been proposed.

4. EXISTING SYSTEM STRUXCTURE

The existing profile-based Personalized Web Search does 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. One evidence reported in is that profile-based personalization may not even help to improve the search quality for some ad hoc queries, though exposing user profile to a server has put the user's privacy at risk. 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 surprisal based on the information theory, assuming that the interests with less user document support are more sensitive. However, this assumption can be doubted with a simple counterexample: If a user has a large number of

documents about "sex," the surprisal of this topic may lead to a conclusion that "sex" is very general and not sensitive, despite the truth which is opposite. Unfortunately, little prior work can effectively address individual privacy needs during the generalization.

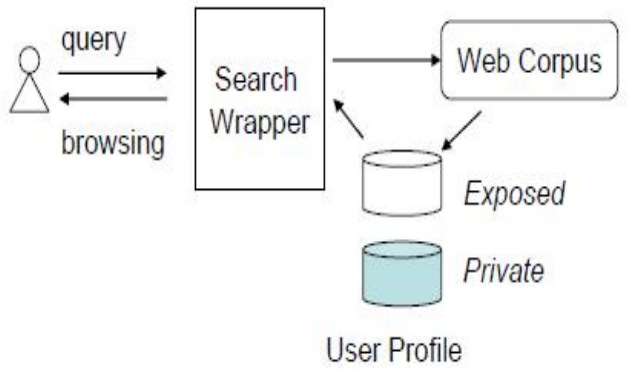


Figure 4.1: Existing System Structure

Figure 4.1 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. Unfortunately, the previous works of privacy preserving PWS are far from optimal. The problems with the existing methods are explained in the following observations:[5] 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 ad hoc queries, though exposing user profile to a server has put the user's privacy at risk. 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. 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.

5. PROPOSED SYSTEM & ITS CONTRIBUTIONS

We propose a privacy-preserving personalized web search framework UPS, which can generalize profiles for each query according to user-specified privacy requirements. Relying on the definition of two conflicting metrics, namely personalization utility and privacy risk, for hierarchical user profile, we formulate the problem of privacy-preserving personalized search as Risk Profile Generalization, with its NP-hardness proved. We develop two simple but effective generalization algorithms, GreedyDP and GreedyIL, to support runtime profiling. While the former tries to maximize the discriminating power (DP), the latter attempts to minimize the information loss (IL). By exploiting a number of heuristics, GreedyIL outperforms GreedyDP significantly. We provide 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.

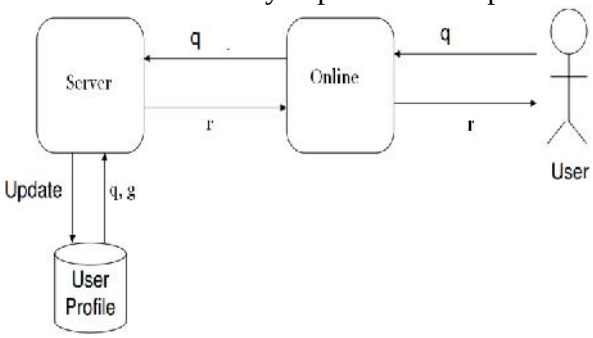


Figure 5.1: General Architecture of Proposed Scheme

, The user profiling process generally consists of three main phases. First, an information collection process is used to gather raw information about the user. Depending on the information collection process selected, different types of user data can be

extracted. The second phase focuses on user profile construction from the user data. The final phase, in which a technology or application exploits information in the user profile in order to provide personalized services.

i) Collecting information about Users

The first phase of a profiling technique collects information about individual users. A basic requirement of such a system is that it must be able to uniquely identify users. The information collected may be explicitly input by the user or implicitly gathered by a software agent. It may be collected on the user's client machine or gathered by the application server itself.

ii) User Profile Construction

User profiles are constructed from information sources using a variety of construction techniques based on machine learning or information retrieval. Depending on the user profile representation desired, different techniques may be appropriate. Profiles may be constructed manually by the users or experts, however, this is difficult and time consuming for most users and would be a barrier to widespread adoption of a personalized service.

iii) Building Concept Profiles

This section describes three representative systems that build user profiles represented as weighted concept hierarchies. Although each uses a different construction methodology, they each use reference taxonomy as the basis of the profile. These profiles differ from semantic network profiles because they describe the profiles in terms of pre-existing concepts, rather than modeling the concepts as part of the user profile itself. Thus, they all require some way of determining which concepts a user is interested in based on their feedback.

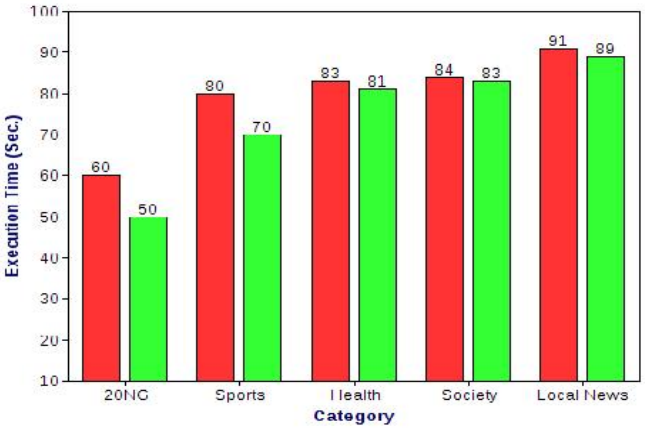
iv) The Greedy Algorithm

A greedy algorithm is a mathematical process that recursively constructs a set Recursion of objects from the smallest possible constituent parts. is an approach to problem solving in which the solution to a particular problem depends on solutions to smaller instances of the same problem. Greedy algorithms look for simple, easy-to-implement solutions to complex, multi-step problems by deciding which next step will provide the most obvious benefit.

6. IMPLEMENTATION & RESULTS

The following performance parameters are commonly used in privacy protection technique evaluation. The existing approach is compared with proposed scheme using these evaluation parameters. The performance of the TC process can be measured by one or more of the following methods.

i) Recall

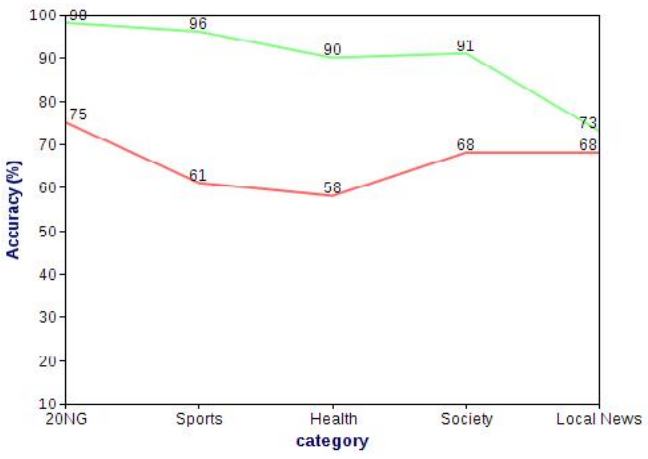


In the above plotting, the red line represents the existing approach and the green line represents the GreedyIL for executing the user profile of the various users with various categories. The existing hierarchical link approach takes more time for extract the result from the dataset.

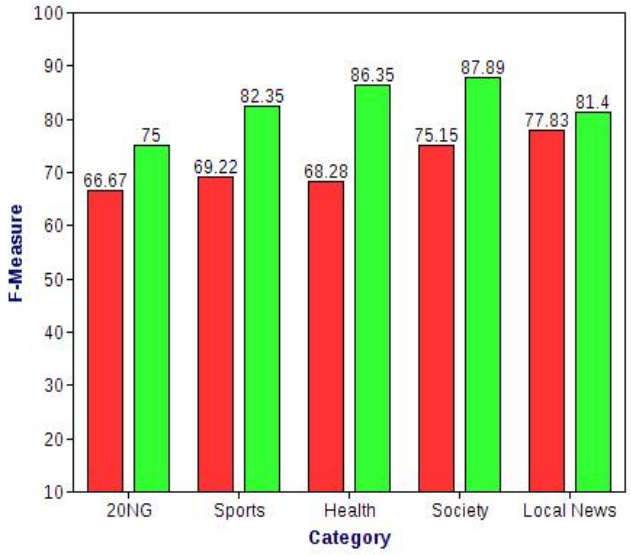
ii) Precision

The precision represents the accuracy of retrieval or categorizing the data. In the above result, the red line represents the existing approach and the green line represents the GreedyIL for executing the user profile. Existing approach accuracy level is poor compare with the GreedyIL Approach of the proposed one.

iii) F-Measure



The F-Measure represents the measure of recall and precision of retrieval or categorizing the data. In the above result, the red line represents the existing approach and the green line represents the GreedyIL for executing the various user profile privacy. These measures are very helpful in evaluating the performance of both frequent and rare categories.



CONCLUSION

This research work presented a client-side privacy protection framework called 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. In addition, UPS also performed online generalization on user profiles to protect the personal privacy without compromising the search quality. We proposed two greedy algorithms, namely GreedyDP and GreedyIL, for the online generalization. Our experimental results revealed that UPS could achieve quality search results while preserving user's customized privacy requirements. The results also confirmed the effectiveness and efficiency of our solution. For future work, we will try to resist adversaries with broader background knowledge, such as richer relationship among topics (e.g., exclusiveness, sequentiality, and so on), or capability to capture a series of queries from the victim. We will also seek more sophisticated method to build the user profile, and better metrics to In particular, we are considering ways of quantifying the utility that we gain from personalization, thus users can have clear incentive to comprise 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 only exposing those information related to a specific query.

REFERENCES

[1] J.S. Breese, D. Heckerman, and C.M. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence (UAI), pp. 43-52, 1998.

[2] G. Chen, H. Bai, L. Shou, K. Chen, and Y. Gao, "Ups: Efficient Privacy Protection in Personalized Web Search," Proc. 34th Int'l ACM SIGIR Conf. Research and Development in Information, pp. 615- 624, 2011.

[3] J. Conrath, "Semantic Similarity based on Corpus Statistics and Lexical Taxonomy," Proc. Int'l Conf. Research Computational Linguistics (ROCLING X), 1997.

[4] J. Castellı´-Roca, A. Viejo, and J. Herrera-Joancomartı´, "Preserving User's Privacy in Web Search Engines," Computer Comm., vol. 32, no. 13/14, pp. 1541-1551, 2009.

[5] P.A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschu¨ tter, "Using ODP Metadata to Personalize Search," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), 2005.

[6] Dou, Zhicheng, Ruihua Song, and Ji-Rong Wen. "A large-scale evaluation and analysis of personalized search strategies." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

[7] Z. Dou, R. Song, and J.-R. Wen, "A Large-Scale Evaluation and Analysis of Personalized Search Strategies," Proc. Int'l Conf. World Wide Web (WWW), pp. 581-590, 2007.

[8] O. Etzioni. The world wide web: Quagmire or gold mine. Communications of the ACM, 39(11):65–68,1996.

[9] S. Gauch, J. Chaffee, A. Pretschner, Ontology-Based User Profiles for Search and Browsing, User Modeling and User-Adapted

Interaction: The Journal of Personalization Research, Special Issue on User Modeling for Web and Hypermedia Information Retrieval, vol. , (2003).
[10] K. Hafner, Researchers Yearn to Use AOL Logs, but They Hesitate, New York Times, Aug. 2006.
[11] A.Krause and E. Horvitz, “A Utility-Theoretic Approach to Privacy in Online Services,” J. Artificial Intelligence Research, vol. 39, pp. 633-662, 2010.
[12] R. Kosala, H. Blockeel “Web mining research: A survey,” ACM SIGKDD Explorations, Vol. 2 No. 1, pp. 1-15, June 2000.
[13] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Workload-aware anonymization,” in KDD ’06, Philadelphia, PA, USA, 2006, pp. 277–286.
[14] A. Pretschner and S. Gauch, “Ontology-Based Personalized Search and Browsing,” Proc. IEEE 11th Int’l Conf. Tools with Artificial Intelligence (ICTAI ’99), 1999.
[15] J. Pitkow, H. Schu¨ tze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel, “Personalized Search,” Comm. ACM, vol. 45, no. 9, pp. 50-55, 2002.
[16] F. Qiu and J. Cho, “Automatic Identification of User Interest for Personalized Search,” Proc. 15th Int’l Conf. World Wide Web (WWW), pp. 727-736, 2006.
[17] L Razmerita,, A.Angehrn, A. Maedche, Ontology based user modeling for Knowledge Management Systems, Proceedings of the User Modeling Conference, Pittsburgh, USA, Springer Verlag, pp. 213-217, 2003.
[18] K. Ramanathan, J. Giraudi, and A. Gupta, “Creating Hierarchical User Profiles Using Wikipedia,” HP Labs, 2008.
[19] Shou, Lidan, et al. "Supporting Privacy Protection in Personalized Web Search."(2012): 1-1.
[20] Srivastava, Jaideep, et al. "Web usage mining: Discovery and applications of usage patterns from web data." ACM SIGKDD Explorations Newsletter 1.2 (2000): 12-23.
[21] K. Sugiyama, K. Hatano, and M. Yoshikawa, “Adaptive Web Search Based on User Profile Constructed without any Effort from Users,” Proc. 13th Int’l Conf. World Wide Web (WWW), 2004.
[22] Shen, Xuehua, Bin Tan, and ChengXiang Zhai. "Privacy protection in personalized search." ACM SIGIR Forum. Vol. 41. No. 1. ACM, 2007.
[23] X. Shen, B. Tan, and C. Zhai, “Implicit User Modeling for Personalized Search,” Proc. 14th ACM Int’l Conf. Information and Knowledge Management (CIKM), 2005.