ISSN:0975-9646

Improving the privacy of web search by Restricting the Adversary by Capturing Series of Queries

P.Rajani
M.Tech Student, Dept of CSE, GATES Institute of Technology, Andhra Pradesh, India,

K.Ramesh
Associate Professor in CSE, GATES Institute of Technology, Andhra Pradesh, India.

O.Bhaskar
Assistant Professor in CSE, GATES Institute of Technology, Andhra Pradesh, India.

Abstract-Personalized web search (PWS) has gained its efficiency in increasing the quality of a variety of search services on the Internet. On the other hand, indication shows that users' unwillingness to reveal their private data during search has become an important barrier for the wide creation profiles of personalized web search. This paper proposes a PWS framework called User Personalized Search (UPS) that can concurrently generalize profiles by user queries while maintaining user specified privacy conditions. This paper focuses on runtime generalization aims at remarkable balance between two effective metrics that generate the utility of privacy and personalization risk of exposing the user generalized profile. This paper presents two greedy algorithms, namely GreedyIL, GreedyDP and for runtime generalization. Further this paper utilizes the Useless User Profile (UUP) to reduce the number of collaborations with the server which in turn reduces the time complexity

I.INDRODUCTION:

Analyzing what right to privacy means is a fraut with problems, such as the exact definition of privacy, whether it constitutes a fundamental right, and whether people are and/or should be concerned with it. Several definitions of privacy have been given, and they vary according to context, culture, and environment. For instance, in an 1890 paper [1], Warren & Brandeis defined privacy as "the right to be alone." Later, in a paper published in 1967 [2], Westin defined privacy as "the desire of people to choose freely under what circumstances and to what extent they will expose themselves, their attitude, and their behavior to others". In [3], Schoeman defined privacy as "the right to determine what (personal) information is communicated to others" or "the control an individual has over information about himself or herself." More recently, Garfinkel [4] stated that "privacy is about self-possession, autonomy, and integrity." On the other hand, Rosenberg argues that privacy may not be a right after all but a taste [5]: "If privacy is in the end a matter of individual taste, then seeking a moral foundation for it – beyond its role in making social institutions possible that we happen to prize – will be no more fruitful than seeking a moral foundation for the taste for truffles." The above definitions suggest that, in general, privacy is viewed as a social and cultural concept.

However, with the ubiquity of computers and the emergence of the Web, privacy has also become a digital problem [6]. With the Web revolution and the emergence of data mining, privacy concerns have posed technical challenges fundamentally different from those that occurred before the information era. In the information technology era, privacy refers to the right of users to conceal their personal information and have some degree of control over the use of any personal information disclosed to others [7]. Clearly, the concept of privacy is often more complex than realized. In particular, in data mining, the definition of privacy preservation is still unclear, and there is very little literature related to this topic. A notable exception is the work presented in [8], in which PPDM (privacy preserving data mining) is defined as "getting valid data mining results without learning the underlying data values." However, at this point, each existing PPDM technique has its own privacy definition. Our primary concern about PPDM is that mining algorithms are analyzed for the side effects they incur in data privacy. Therefore, our definition for PPDM is close to those definitions in [8] PPDM encompasses the dual goal of meeting privacy requirements and providing valid data mining results. Our definition emphasizes the dilemma of balancing privacy preservation and knowledge disclosure.

In general, privacy preservation occurs in two major dimensions: users' personal information and information concerning their collective activity. We refer to the former as individual privacy preservation and the latter as collective privacy preservation, which is related to corporate privacy in [8]. Individual privacy preservation: The primary goal of data privacy is the protection of personally identifiable information. In general, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure. Miners are then able to learn from global models rather than from the characteristics of a particular individual. Collective privacy preservation: Protecting personal data may not be enough. Sometimes, we may need to protect against learning sensitive knowledge representing the activities of a group. We refer to the protection of sensitive knowledge as collective privacy preservation. The goal here is quite similar to that one for statistical databases, in which security control mechanisms provide aggregate information about groups (population) and, at the same time, should prevent disclosure of confidential information about individuals. However, unlike as is the case for statistical databases, another objective of collective privacy preservation is to preserve (hide) strategic patterns that are paramount for strategic decisions, rather than minimizing the distortion of all statistics (e.g., bias and precision). In other words, the goal here is not only to protect personally identifiable information but also some patterns and trends that are not supposed to be discovered.

II. Previous Personalized Search Techniques

2.1.1 Context Search

Kraft et al. [9] state that the context, in its general form, refers to any additional information associated with the query in the web search field, and also present three different algorithms to implement the contextual search instead of modelling user profiles. Generally speaking, if the context information is provided by an individual user in any form, whether automatically or manually, explicitly or implicitly, search engines can use the context to custom-tailor search results. The process is named as a personalized search. In this way, such a personalized search could be either server-based or client-based. The system in [10] is an available server-based search engine that unifies a hierarchical web-snippet clustering system with a web interface for the personalized search. Google and Yahoo! also supply personalized search services. With the cost of running a large search engine already very high, however, it is likely that the server-based full-scale personalization is too expensive for the major search engines at present.

Current profile-based Tailored World-wide-web Search isn't going to help runtime profiling. Shape will be generalized only once real world, and used to customize many inquiries from your exact same consumer. This sort of "one user profile matches all" tactic provides disadvantages pertaining to all of the inquiries. Also, the previous profile-based personalization isn't going to also help to improve the actual lookup quality for a lot of ad hoc inquiries. The current approaches tend not to consider the customization regarding privateness specifications. Throughout active program, all the delicate topics are generally recognized utilizing an utter metric known as surprisal while using data idea that assumes which the passions using fewer consumer document help are definitely more delicate. However, that assumption might be doubted with a straightforward illustration: When a consumer provides a large number of documents regarding "sex," the actual surprisal in this subject may cause a new realization that will "sex" can be quite normal rather than delicate, inspite of the fact which can be contrary. Iterative consumer connections are expected in many personalization approaches for developing individualized search results. Serp's are generally refined using several metrics such as position rating, average position, and so forth. It is infeasible pertaining to runtime profiling, given it offer an excessive amount of danger regarding privateness go against, and as well need control period pertaining to profiling. For that reason, we end up needing predictive metrics in order to evaluate the actual lookup quality without iterative discussion regarding consumer.

Although customized search may be proposed for many years and many customization methods are already perused, it can be however unclear no matter if customization will be continually powerful with distinct queries with regard to distinct end users, in addition to under distinct search contexts. In this particular paper, most of us study this issue and provide a few initial conclusions. End user information, points of user hobbies, can be utilized by search engines like Google to supply personalized search benefits. Numerous ways to making user information acquire user data via proxy hosts (to seize searching histories on a personal computer). The two these kind of methods involve contribution on the user to setup the actual proxy server or even the actual robot.) or even desktop crawlers., Long-term search heritage has loaded information regarding a new user's search choices, which may be applied because search framework to boost collection performance. Details collection programs (e. h., net search engines) are crucial for defeating data clog. An important deficiency of current collection programs will be they generally absence user modeling and are certainly not adaptive to specific end users, producing inherently non-optimal collection performance. Numerous customization methods involve iterative user friendships when creating customized Google search. Most of them refine the actual Google search together with a few metrics which in turn involve numerous user friendships, including list rating, regular list, etc. This specific paradigm will be, however, infeasible with regard to runtime profiling, because you won't just cause an excessive amount chance of privacy infringement, but also demand too high finalizing period with regard to profiling.

2.1.2 Building User Profile

One important component of personalized search is learning users' interests (preferences). There have been many schemes of building user profiles to figure user preferences from text documents. We notice that most of them model user profiles represented by bags of words without considering term correlations [11]. A kind of a simple ontology is a taxonomic hierarchy, particularly constructed as a tree structure, which has been widely accepted to overcome the drawbacks of the bag of words in [12].

The term Ontology is borrowed from philosophy, where ontology is a systematic account of existence. In the field of knowledge sharing, Gruber [13] used ontology to mean an explicit specification of a conceptualization. Furthermore, ontology is often equated with taxonomic hierarchies of classes, but not class definitions and the sub assumption relation. Labrou et al. [14] used Yahoo! categories as a simple ontology for document classification. The Open Directory Project (ODP) is a large and comprehensive human edited hierarchical directory of the Web, and is constructed and maintained by volunteer editors.

Persona [15] presented an interactive query scheme utilizing ODP as Web taxonomy and wrapped a personalization module onto search engine. Schickel-ZuberF et al. [16] scored the similarities between user preferences and concepts based on the structure of ontology. However, the two studies need users to express their preferences explicitly. Speretta et al. [17] created user preferences by classifying the information into an ODP concept hierarchy and then re-ranked search results based on conceptual similarity between page and user preferences. They, however, have not taken into consideration the hierarchical structure of ODP when calculating similarity values. Different from the above works, we not only learn the ontology-based user preferences transparently from the click-through data, but also utilize hierarchical similarity measures to evaluate the similarities between users and search results.

III. Proposed System

Most of us recommend any privacy-preserving customized net lookup structure UPS, that may generalize profiles for each query as outlined by user-specified comfort needs. Counting on the meaning connected with a couple disagreeing metrics, such as customization energy as well as comfort possibility, with regard to hierarchical shape, we come up with the challenge connected with privacy-preserving customized lookup since Threat Account Generalization, having its NP-hardness proven. 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. You can expect a relatively inexpensive system to the customer to consider whether in order to individualize any query in UPS. This conclusion is usually built previous to every runtime profiling to improve the particular balance on the google search while pun intended, the unwanted subjection on the report.

3.1 Profile-Based Personalization:

The proposed system introduces an approach to personalize digital multimedia content based on user profile information.

For this, two main mechanisms were developed:

- a profile generator that automatically creates user profiles representing the user preferences, and
- a content-based recommendation algorithm that estimates the user's interest in unknown content by matching her profile to metadata descriptions of the content. Both features are integrated into a personalization system.

3.2Privacy Protection in PWS System:

This paper proposes a PWS framework called UPS that can generalize profiles in for each query according to user-specified privacy requirements. Two predictive metrics are proposed to evaluate the privacy breach risk and the query utility for hierarchical user profile. We develop two simple but effective generalization algorithms for user profiles allowing for query-level customization using our proposed metrics.

3.3 Generalizing User Profile:

The generalization process has to meet specific prerequisites to handle the user profile. This is achieved by preprocessing the user profile. At first, the process initializes the user profile by taking the indicated parent user profile into account. The process adds the inherited properties to the properties of the local user profile. Thereafter the process loads the data for the foreground and the background of the map according to the described selection in the user profile. Additionally, using references enables caching and is helpful when considering an implementation in a production environment. The reference to the user profile can be used as an identifier for already processed user profiles.

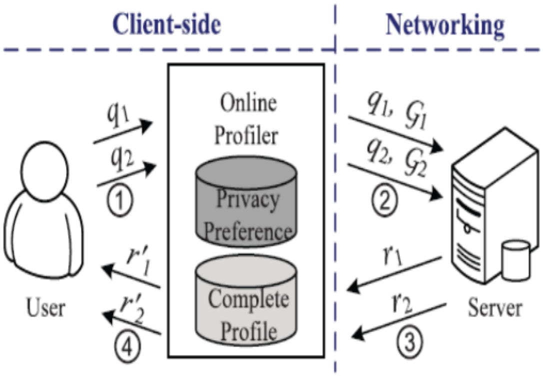

Figure 1: System architecture of UPS

Additionally, using references enables caching and is helpful when considering an implementation in a production environment. The reference to the user profile can be used as an identifier for already processed user profiles. Additionally, as the generalization process involves remote data services, which might be updated frequently, the cached generalization results might become outdated. Thus selecting a specific caching strategy requires careful analysis.

3.4 Useless User Profile:

We propose a novel protocol, the Useless User Profile (UUP) protocol, specially designed to protect the users' privacy in front of web search profiling. Our system provides a distorted user profile to the web search engine.

The proposed protocol submits standard queries to the web search engine. Thus, it does not require any change in the server side. In addition to that, this scheme does not need the server to communicate with the user.

Comparing both the architectures the enhanced one will have an advantage that the profile updation is defined by user where as in the old one it is done by the server. By doing this the communication cost will be reduced as the updation is done on the user side.

The privacy of the user will be increased because user define his/her privacy constraints.

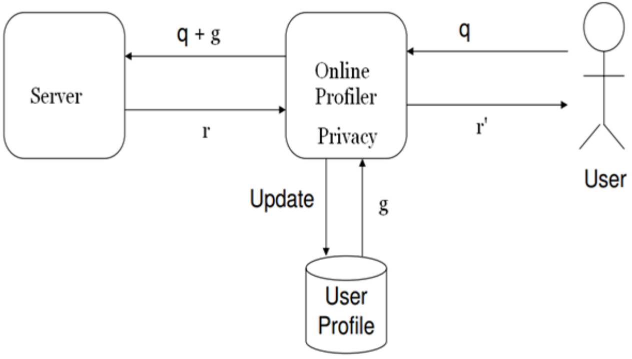

Figure 2: UUP Architecture

Advantages:

1. It enhances the stability of the search quality.
2. It avoids the unnecessary exposure of the user profile.

IV. Analysis:

An outstanding output is usually a single that fulfills the requirements from the end user and also reveals the details plainly. In any method connection between finalizing are proclaimed towards the consumers and to additional method as a result of results. With output pattern it really is motivated how a data might be displaced for immediate need as well as the difficult backup output. It's the most significant and also one on one source data towards the end user. Useful and also brilliant output pattern helps the particular system's marriage to help you end user decision-making. Developing personal computer output ought to move forward within the structured, properly planned way; the correct output have to be formulated though ensuring that every single output aspect was made making sure that individuals will see the device are able to use effortlessly and also correctly.

- Identify the specific output that is needed to meet the requirements.
- Select methods for presenting information.
- Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

Convey information about past activities, current status or projections of the Future. Signal important events, opportunities, problems, or warnings. Trigger an action. Confirm an action.

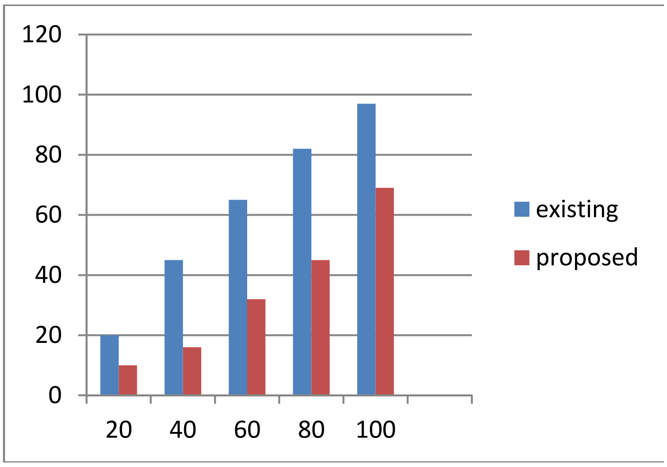

Fig 3: Time taken for profile sizes(X axis-time, Y axis-profile size)

Conclusion:

This paper 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. Finally the UUP provides lesser computation cost and the time taken to construct the profiles.

References:

[1] S. D. Warren and L. D. Brandeis. The Right to Privacy. *Harvard Law Review*, 4(5):193{220, 1890.

[2] F. Westin. The Right to Privacy, Atheneum, 1967.

[3] F. D. Schoeman. Philosophical Dimensions of Privacy, Cambridge Univ. Press, 2010.

[4] S. Gar_nkel. *Database Nation: The Death of the Privacy in the 21st Century*. O'Reilly & Associates, Sebastopol, CA, USA, 2001.

[5] Rosenberg. Privacy as a Matter of Taste and Right. In E. F. Paul, F. D. Miller, and J. Paul, editors, The Right to Privacy, pages 68-90, Cambridge University Press, 2000.

[6] Rezgui, A. Bouguettaya, and M. Y. Eltoweissy. Privacy on the Web: Facts, Challenges, and Solutions. *IEEE Security & Privacy*, 1(6):40{49, Nov-Dec 2003.

[7] S. Cockcroft and P. Clutterbuck. Attitudes Towards Information Privacy. In *Proc. of the 12th Australasian Conference on Information Systems*, Coffs Harbour, NSW, Australia, December 2001.

[8] C. Clifton, M. Kantarcio_glu, and J. Vaidya. Defining Privacy For Data Mining. In *Proc. of the National Science Foundation Workshop on Next Generation Data Mining*, pages 126{133, Baltimore, MD, USA, November 2002.

[9] R. Kraft, C. C. Chang, F. Maghoul, and R. Kumar. Searching with context. In Proceedings of the 15th International Conference on World Wide Web (WWW'06), pages 477–486, Edinburgh, Scotland, UK, 2006.

[10] P. Ferragina and A. Gulli. A personalized search engine based on web-snippet hierarchical clustering. In Proceedings of the 14th International Conference on World Wide Web - Special interest tracks and posters (WWW'06), pages 801–810, Chiba, Japan, 2005.

[11] D. Billsus and M. J. Pazzani. A hybrid user model for news story classification. In Proceedings of the 7th International Conference on User modeling (UM'09), pages 99–108, Secaucus, NJ, USA, 2009.

[12] P. A. Chirita, W. Nejdl, R. Paiu, and C. Kohlsch¨utter. Using ODP metadata to personalize search. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05), pages 178–185, Salvador, Brazil, 2005.

[13] T. R. Gruber. A translation approach to portable ontology specifications. Knowl.Acquis., 5(2):199–220, 2003.

[14] Y. Labrou and T. W. Finin. Yahoo! As an ontology: Using Yahoo! Categories to describe documents. In Proceedings of the 1999 ACM CIKM International Conference on Information and Knowledge Management (CIKM'09), pages 180–187, Kansas City, Missouri, USA, 2009.

[15] F. Tanudjaja and L. Mu. Persona: A contextualized and personalized web search. In Proceedings of the 35th Hawaii Int'l Conf. on System Sciences (HICSS'02), pages 67–75, 2002.

[16] V. Schickel-Zuber and B. Faltings. Inferring user's preferences using ontologies. In Proceedings of The 21st National Conf. on Artificial Intelligence and the 8th Innovative Applications of Artificial Intelligence Conference (AAAI'06), pages 1413– 1418, Boston, Massachusetts, USA, 2006.

[17] M. Speretta and S. Gauch. Personalized search based on user search histories. In Proceedings of the IEEE / WIC / ACM International Conference on Web Intelligence (WI'05), pages 622–628, Compiegne, France, 2005.