Anonymizing User Profiles for Personalized Web Search

Yun Zhu
Department of Math & CS
Emory University
Atlanta, GA
yzhu23@emory.edu

Li Xiong
Department of Math & CS
Emory University
Atlanta, GA
lxiong@emory.edu

Christopher Verdery
Department of Math & CS
Emory University
Atlanta, GA
cverder@alum.emory.edu

ABSTRACT

We study the problem of anonymizing user profiles so that user privacy is sufficiently protected while the anonymized profiles are still effective in enabling personalized web search. We propose a Bayes-optimal privacy notion to bound the prior and posterior probability of associating a user with an individual term in the anonymized user profile set. We also propose a novel bundling technique that clusters user profiles into groups by taking into account the semantic relationships between the terms while satisfying the privacy constraint. We evaluate our approach through a set of preliminary experiments using real data demonstrating its feasibility and effectiveness.

Categories and Subject Descriptors: H.2.7 [Database Administration]: Security, integrity, and protection; H.3.3. [Information Search and Retrieval]

General Terms: Design, Experimentation, Security

Keywords: Anonymization, privacy-preserving data publishing, personalized search

1. INTRODUCTION

Personalized web search is a promising technique to improve retrieval effectiveness. However, it often relies on personal user profiles which may reveal sensitive personal information. In this paper, we study the problem of grouping user profiles (represented as a weighted term list) so that user privacy is sufficiently protected while the grouped profiles are still effective in enabling personalized web search.

Existing Techniques. Most works on anonymization focus on relational data where every record has the same number of sensitive attributes. There are a few works taking the first step towards anonymizing set-valued or transactional data where sensitive items or values are not clearly defined [1]. While they could be potentially applied to user profiles, one main limitation is that they either assume a predefined set of sensitive items that need to be protected, which are hard to define in the web context in practice, or only guarantee the anonymity of a user but do not prevent the linking attack between a user and a potentially sensitive item. [6] proposed a technique for building user profiles with configurable levels of details. A few recent works specifically studied anonymizing query logs. Notably, [4, 2] have demonstrated the ineffectiveness or privacy risks of naive anonymization schemes. [3] studied anonymization techniques with differential privacy, however, the utility of the data is limited to statistical information and it is not clear how it can be used for personalized web search.

Contributions. In this paper, we define a Bayes-optimal privacy notion for user profiles represented as set-valued data. It does not require predefined quasi-identifying or sensitive terms nor does it require external knowledge database. Rather, it treats every term as potentially sensitive or identifying and bounds the difference between the prior and posterior probability of linking an individual to any term. In addition, we propose a novel bundling technique that clusters user profiles into user groups by taking into account the semantic relationships between the terms while satisfying the privacy constraint. Finally, we evaluate our approach through a set of preliminary experiments using real data, showing that our approach effectively enables personalized search with assured privacy.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

2. APPROACH

Problem Setting. We consider a set of user profiles for personalized web search. Each user profile is represented as a vector of tuples: $U = \{(t_1, w_1), (t_2, w_2), \ldots, (t_m, w_m)\}$, where t_i is a word or phrase representing a user's interest, and w_i quantifies the relative extent. Our goal is to cluster the user profiles into user groups so that the privacy of individual users is protected while the user groups are still useful for personalized web search.

Privacy Definitions. An adversary may link a user to a user group based on his background knowledge (e.g. certain terms that the user has searched for) and then identify additional terms in the user group that are contributed by the user. Our anonymization goal is to prevent such linking attacks that associate a user with an individual term in the anonymized user profile set. We adopt the Bayes-optimal privacy notion [1] to bound the difference between the prior and posterior beliefs of linking a user to a term in the user groups. Definitions such as l-diversity and t-closeness [1] can be adapted to bound the diversity of terms in each group or the difference between term distribution in each group and the global term distribution in the entire user profile set respectively. In this work, we also propose an instantiation of the privacy notion for the grouping approach, called p-linkability.

DEFINITION 1. *(p-**linkability**) A user profile grouping satisfies p-linkability if the probability of linking a user to an individual term in a user group does not exceed p.*

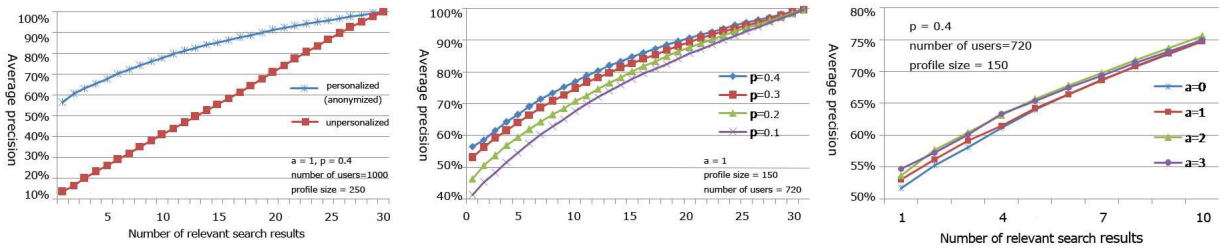

Figure 1: Search precision using anonymized user profiles

User profile grouping. We propose a bundling technique to group user profiles and use group representatives for personalized web search. Since the group representatives will be used for re-ranking the search results, we would like the users within each group to be similar to each other. Concretely, our goal is to perform similarity-based clustering while satisfying the privacy constraint. While traditional similarity metrics such as cosine similarity can be used, a main challenge is how to take into account the semantic similarity of two user profiles. For example, a user interested in *riding* and a user interested in *sports* should be similar to some extent as riding is one type of sports. To address this issue, we propose a user profile augmentation technique before clustering using term co-occurrence networks or term hierarchies. For this study, we use the WordNet (`http://wordnet.princeton.edu`) and employ the following augmentation steps. Table 1 shows the original profiles and the profiles after augmentation.

1. Synonym set replacement. It replaces every single term with its synonym set, a set of words and phrases including the term and all its synonyms.
2. Hypernym set augmentation. It augments every synonym set with its hypernym set, a set of words or phrases whose semantic range includes that of the synonym set and its hyponym. Only hypernym sets that could be reached within a given number of steps (configurable parameter a) from the current synonym set will be added into the user profile.

U_1	(kitten, 1), (riding, 0.8)
U_2	(pup, 0.6), (equitation,1)

U_1	({kitten, kitty}, 1), ({riding, horseback riding, equitation}, 0.8)}, ({young mammal}, 1), ({sport, athletics}, 0.8)
U_2	({pup, whelp}, 0.6), ({riding, horseback riding, equitation}, 1), ({young mammal}, 0.6), ({sport, athletics}, 1)

Table 1: The original and augmented user profile set

Once the profiles are augmented, the next step is to produce a clustered user profile set that satisfies p-linkability or other given privacy principles. We use a greedy algorithm as follows. In the beginning, a user profile is randomly selected as the seed of a new cluster. The closest user profile is continuously selected and combined with the seed until the cluster satisfies p-linkability. At next step, a user profile with the longest distance to the previous seed is selected as the seed of the new cluster. The process repeats until every user profile is clustered. The cluster centroid (based on the original user profiles not the augmented ones) is computed and used as the group representative.

3. PRELIMINARY RESULTS

We performed a set of preliminary experiments using a set of user profiles generated from the AOL search query log (`http://gregsadetsky.com/aol-data/`). We implemented a personalized search engine on top of Lucene and used the TIPSTER Information-Retrieval Text Research Collection (`http://www.ldc.upenn.edu/Catalog`) as our search corpus. When a result list is returned from Lucene, we re-rank them according to their similarity to the user's profile (both the original and anonymized profiles). Since our focus is to evaluate the effectiveness of anonymization, rather than the personalized search, we use the search result based on the original user profiles as the gold standard and measured the precision of the search results based on anonymized profiles. We use Average Precision [5] as our search quality metric.

Figure 1 compares the average precision of personalized search using anonymized profiles with the non-personalized search as well as the search quality with varying p and a. It shows that the search using anonymized user groups achieves good precision and provides significant improvement over non-personalized search. As expected, a lower value p provides stronger privacy guarantee at the cost of search precision. It is also verified, to some extent, that the higher a, the more semantic similarity between user profiles are taken into account and thus the better the search quality.

While these results demonstrated the feasibility of the approach, it certainly warrants further research. The current AOL dataset places many limitations for extracting users' specific interests. We plan to explore other options to collect or extract user profiles to further verify our approach. Moreover, we are interested in extending the work with similarity constraint in each group to provide certain utility guarantee. Finally, we are also exploring mechanisms for anonymizing user profiles with differential privacy.

Acknowledgement

The work is partially supported by a Career Enhancement Fellowship by Woodrow Wilson Foundation.

4. REFERENCES

[1] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey on recent developments. *ACM Computing Surveys*, 42(4), 2010.
[2] R. Jones, R. Kumar, B. Pang, and A. Tomkins. Vanity fair: privacy in querylog bundles. In *CIKM*, 2008.
[3] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas. Releasing search queries and clicks privately. In *WWW*, 2009.
[4] R. Kumar, J. Novak, B. Pang, and A. Tomkins. On anonymizing query logs via token-based hashing. In *WWW*, 2007.
[5] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *SIGIR*, 2006.
[6] Y. Xu, K. Wang, B. Zhang, and Z. Chen. Privacy-enhancing personalized web search. In *WWW*, 2007.