Towards Privacy-Preserving Query Log Publishing

Li Xiong Eugene Agichtein
Mathematics and Computer Science Department
Emory University
{lxiong,eugene}@mathcs.emory.edu

1. INTRODUCTION

It's an open secret that search engines collect detailed query logs[1], and sometimes release these data to third parties. While making this wealth of information available provides enormous opportunities for information retrieval and web mining research, it also raises serious concerns about the privacy of individuals. We strongly believe that this data *should* be published to allow researchers to develop new information access algorithms, however, it is desirable to *anonymize* these logs, so that they are still usable for research but do not contain sensitive information.

The most important need is to define in a principled way the notion of privacy for query logs. This paper attempts to lay out some dimensions for defining privacy guidelines for query log publishing. We focus on the central issue of how to strike a balance between protecting the sensitive information and maintaining useful data for analysis. This work is within the overall vision of developing anonymization techniques to allow construction of IR algorithms (e.g., spelling correction) that maintain state-of-the-art performance over the anonymized data.

We first describe some important applications of query log analysis and discuss their requirements on the degree of granularity of query logs. We then analyze the sensitive information in query logs and classify them from the privacy perspective. We lay out two orthogonal dimensions for anonymizing query logs and present a spectrum of approaches along those dimensions. We discuss whether existing privacy guidelines such as HIPAA[2] can apply to query logs directly, or whether these guidelines require significant adaptation. For each of the approaches, we discuss the implications on query log utility regarding the important applications as well as the privacy of the anonymized query logs. More generally, our goal is to bring up questions and suggest challenges for privacy-preserving query log publishing.

2. LOG ANALYSIS APPLICATIONS

A useful query log contains a user ID or user session ID, query terms, time stamp, and possibly a URL of a clicked result and the result position. We briefly describe some important applications which rely on query log analysis and discuss their requirements on degree of granularity of query logs. Recent increase in academic research activity in these areas is entirely due to availability of, and access to, published query logs. In effect, availability of these data levels the playing field and allows academic researchers to perform relevant and realistic information retrieval experiments. We primarily focus on search-related applications, though query logs are certainly also valuable for many other web mining tasks. In particular, development and evaluation of algorithms for ranking, query suggestion, implicit feedback, relevance monitoring, and personalization, just to mention a few areas, depend on making use of query log data, as described below.

Implicit feedback for web search ranking: Implicit feedback such as result clicks has long been used in the IR community for improving ranking algorithms. This area has experienced a renaissance as click logs became increasingly available. For example, Joachims [12] introduced a ranking method based entirely on clickthrough data. More recently, Agichtein et al. [1] developed rich and robust interaction models to improve the ranking accuracy further. These approaches require, at a minimum, the session level information (i.e., the sequence of queries and clicks for the same user within the same query session), but could be also used on aggregate or probabilistic level, i.e., information about individual user actions is not required. However, the accuracy could be improved if the system was personalized for a particular user or group of users, which does require a user id to persist across sessions.

Query spelling correction: Query logs to an internet search engine provide a large amount of implicit and explicit information about language. As a representative application, Cucerzan and Brill [5] investigate using the query logs in spelling correction of search queries, a task which poses many additional challenges beyond the traditional spelling correction problem. The key idea is to exploit the actual errors made by users to collect accurate likelihood statistics. The approach relies solely on aggregated query frequency statistics and does not require any identity information (e.g., that any particular queries were issued by the same user).

Query suggestion and refinement: Query expansion can potentially resolve the short query and word mismatching problems in web search. Recently, it has been shown (e.g., [6]) that query logs can be successfully exploited to suggest queries, both to better specify the query and to suggest alternatives for exploration. As an extreme form of query suggestion query logs could be exploited to proactively suggest search results (e.g., via the "alerts" mechanism) as

[1]We will use the term query logs for the combination of query and result click logs.
[2]http://www.hhs.gov/ocr/hipaa/

WWW2007, May 8–12, 2007, Banff, Canada.

recently presented by Yang and Jeh [23]. These approaches require session level information as described above.

Automatic monitoring and evaluation: Query logs have long been used for monitoring the performance of search engines and other web services. Not surprisingly, there is a plethora of methods mining these data for automatic monitoring. For example, Fox et al. [8] explored the relationship between implicit and explicit measures computed using web search logs. As another example, Agichtein et al. [2] extended this work with richer features and more robust models to better predict preferences of users for search results. Most recently, Downey et al. [7] presented predictive models of user actions (e.g., query reformulation) that can also be derived from query log analysis. All these predictions can be used to detect search abuse, web spam, and for automatic search engine evaluation and "system health" monitoring. These approaches at a minimum require session-level information described above, but can benefit from longer user history, requiring a user id that persists across query sessions.

Web search personalization: Web search personalization has been a hot area of research. Ad placement and ranking accuracy is also an inherently personalized application (e.g., by location, user id or user profile). To learn personalized models, a persistent user id across query sessions is required.

3. QUERY LOG PRIVACY

We now consider some approaches for sharing the query log to enable research in the areas above while limiting potential compromises of user privacy. Before we propose the anonymization approaches, we analyze the potential sensitive information in query logs and classify them along a few dimensions. We present a few scenarios for potential privacy breaches and motivate our anonymization approach.

3.1 Sensitive Information in Query Terms

A useful query log contains a user ID or user session ID, query terms, time stamp, and possibly a URL of a clicked result and the result position. The query terms contain a wealth of information that potentially include sensitive information. We classify them along a number of dimensions: information ownership, subject entity, and type of sensitive information.

Information ownership: The sensitive information contained in query terms can be either personal information or business information. *Personal information* belongs to a private individual and its disclosure might not be in the individual's best interests. It may include a person's identity information, financial information, genetic information, and medical information. *Business information* belongs to a business whose disclosure may harm the business. It may include sales and marketing plans, new product plans, and notes associated with patentable inventions. Without loss of generality, we mainly focus on personal information in this paper and classify the personal information based on subject entity and type of sensitive information below.

Subject entity: The personal information contained in query terms refers to a specific individual and we refer to this individual the subject entity. Subject entity of the query terms could be the *query user* who issued the query itself. For instance, when a person performs a vanity search by typing in his own name and other personal information, the person itself is the subject entity. More often, the subject entity of the query terms could be a *third-party individual*. For instance, if a user searches for a private or unpublished phone number he received a call from to figure out if he should call back or not, the owner of the phone number is the subject entity whose personal information is at stake.

Type of sensitive information: Personal information should be protected if their disclosure will cause a person's damage. This include identifying information that could lead to identity theft and a range of other information that could be exploited for discrimination and fraud.

A person's *identifying information* include the direct identifiers such as name and SSN as well as other indirect ones such as addresses, telephone numbers, e-mail addresses, and so on. The HIPAA standard states that identifiable information refers to data explicitly linked to a particular individual as well as data that could enable individual identification. Adopting the HIPAA guidelines, we could define a list of safe harbor identifiers for query terms in addition to name and SSN including: all geographic subdivisions smaller than a state, including street address, city, county, zip code, and their equivalent geocodes; all elements of dates (except year) directly related to an individual; voice and fax telephone numbers; email addresses; certificate/license numbers; vehicle identifiers and serial numbers, including license plate numbers; and any other unique identifying number, characteristic, or code.

In addition to identifying information, *financial information* such as credit card numbers is considered sensitive since the disclosure may lead to fraud. *Health information* is kept confidential to the patient governed by HIPAA to prevent possible discrimination against people with a certain medical condition. Political people may wish to keep their *political viewpoints* secret as they may be used by political groups to punish those who disagree with them.

3.2 Privacy Breach of Query Logs

Consider an adversary who gets access to query logs and wants to mine it for personal information, two steps are likely involved. The first step is to link all relevant query terms to a single entity. If the query terms contain personally identifiable information and other sensitive information, then the privacy for that individual is breached.

Query terms in one single query or even one single session are likely to be related to one another. So any identifying information contained in one query or one session are likely to be related to one single identity, either the user who issued the query, or a third-party entity. However, if there is identifying information contained in one session and there is identifying information contained in another session, they are less likely to be related to the same entity than if they were found together in the same session. So without persistent user ID or session ID that links the relevant query terms, the adversary will have to work harder to piece together the information for a single identity.

4. QUERY LOG ANONYMIZATION

With the above discussion in mind, we propose two orthogonal dimensions for anonymizing query logs for publishing and analysis. These dimensions are designed considering the important query log analysis applications discussed in

Section 2. For each of the approaches along the dimensions, we discuss their implication on privacy as well as data utility for the common applications.

4.1 Query Log Grouping

We first consider the following degrees of granularity for query logs with regard to the query entry information from both application perspective (improving web search) and privacy perspective (minimizing privacy breach).

1. *User*: At one end of the spectrum, we could keep the user ID as well as session ID for query logs so the logs can be grouped by users across all sessions. The availability of the persistent user ID as well as session ID will maximize the utility of the data and enable many applications that are otherwise not possible. In particular, web search personalization and automatic relevance monitoring and evaluation both require or benefit from user level information. On the other hand, as query terms are grouped for each user, it is likely they are associated with the user itself or a small set of third-party entities. Thus there is a higher chance for the adversary to link the data to an identifiable entity.

2. *Session*: Moving down the spectrum, we could remove user ID but group all the query entries by a unique session ID so that user is not tracked across sessions. As many of the applications except web search personalization require only session level information, this approach provides good data utility. Regarding privacy, it is likely the query terms in each query session are associated with a single subject entity, the attacker would still have some chance linking the data to an identifiable entity but less chance linking data to the query user itself. It is worth pointing out that an attacker would still be able to link some sessions together based on similarities of query patterns.

3. *Query Session*: Further down the spectrum, we could remove both user ID and session ID, but combine a set of actions (clicks) and a follow-up query, if any, for an individual user. The data could be still useful for those applications requiring session level information as it preserves a *mini* session. The privacy improves substantially as the attacker has very little data from a session to link to a single entity with high probability. The key here is that the query session is very small and close to a single query.

4. *Query*: In this approach, only individual queries are preserved without any user ID, session ID, or sequence information. Our belief is that it will provide a similar level of privacy as the query session approach above, however, it will sacrifice considerably for data utility due to the lack of any query sequence or query chain.

5. *Aggregate*: At the other end of the spectrum, only the aggregate information (e.g., total number of clicks, total number of users issued a particular query reformulation, etc.,) is released (no individual actions or query sessions). While only enabling those applications that do not require session level and user level information, this will offer the maximum privacy as only statistics information are disclosed.

4.2 Query De-Identification

An orthogonal dimension is the private information revealed in query terms directly. It is important to note that the sensitive information, such as financial and medical information, by themselves are usually not sensitive unless they are associated with an identifiable entity. For instance, nobody will care if a disease name is disclosed in the query terms if it is not associated with any known entity. Thus our approach is to de-identify query terms by removing or generalizing the identifying information.

Given the safe harbor identifiers we discussed in Section 3, we believe there is a spectrum of anonymity strictness that can be applied to the query terms. We discuss each of them and their implications on data privacy as well as data utility.

1. *Full De-identification.* At one end of the spectrum, query terms are considered fully de-identified if all of the above identifiers have been removed, and there is no reasonable basis to believe that the remaining information could be used to identify a person. This offers the maximum privacy as we assume that once the individuals cannot be identified, other private information such as political viewpoints would not cause any damage even disclosed. However, this is also a theoretical position as no one will be able to guarantee with absolute certainty that a query log is fully de-identified. On the other hand, given such a full de-identification, the resulting query log will have a minimum coverage as the query entries will be removed as long as they contain any identifying query terms. It is worth noting that with full de-identification, user ID can be preserved as the attacker will not have any basis for identifying an individual even if it can link all the information to an individual based on the user ID. However, the query chain may be broken because of the removed query entries.

2. *Partial De-identification.* As an alternative to full de-identification, we can remove direct identifiers (such as name and SSN) only but not indirect ones (such as age). This approach provides a better data coverage as less query entries will be removed. However, it is prone to data linkage attacks [19] that combine subject terms with other publicly available information to re-identify represented individuals. For example, an individual who is the only Caucasian male born in 1925 living in a sparsely populated area could have his age, race, gender, and zip code joined with a voter registry from the area to obtain his name and mailing address.

3. *Statistical De-identification.* As a tradeoff, we could have a more intelligent and statistically acceptable de-identification that maintains as much "useful" data as possible while guaranteeing data privacy. Many approaches have been proposed recently for privacy preserving data publishing focused on structured data [10, 11, 20, 4, 9]. k-anonymity [19] with extensions l-diversity [17] and t-closeness [16] have been proposed and widely recognized as a privacy model by requiring a set of k entities to be indistinguishable from each other based on a quasi-identifier set. Multiple schemes have been proposed to transform a structured dataset to one that meets the k-anonymity requirement [19, 22,

3, 8, 24, 13, 14, 15, 21]. An interesting research direction is to apply such models to query logs. It presents interesting research challenges due to the difficulty of mapping the relevant query terms to each subject entity. Also it is an open question how the generalization techniques apply to query logs as a generalized query term will most likely lose its semantics for many query log analysis applications.

4. *No De-identification.* At the other end of the spectrum, all query terms can be preserved without any de-identification. It maximizes the query log coverage but also sacrifices the privacy. If combined with user ID in the query log, there is a high chance the privacy will be breached for individual users.

5. RESEARCH DIRECTIONS

The approach we have sketched in the previous section requires addressing several interesting challenges.

Detecting Identifying Information. The first step of removing identifying information such as names is to extract them from the text of the queries. This alone is a challenging issue. While entity tagging techniques are mature for natural language text [18], queries typically consist of just a few terms, which makes information extraction challenging. While some entities such as phone numbers are easily extractable using pattern matching, person, location, and organization names can sometimes only be identified based on the context, which is often absent in queries. Dictionaries and other simple approaches will fail for rare and foreign names, which also happen to be the most "useful" personal identifiers. Also, the named entities in query logs are often misspelled, making matters even worse.

Entity Mapping. In order to apply recent k-anonymity research results to query log anonymization, one main challenge lies in how to map the relevant query terms to the same subject entity. The query terms in a single query are likely referring to a single subject entity. The probability decreases when they are across different queries, different sessions, and different users. It is an open question how to map these query terms to the same entity in a principled way.

Metrics for Privacy and Utility. Our ultimate goal for query log anonymization is to maximize the utility of the published query logs while preserving individual privacy. A key question is: how can we determine whether or not a certain approach provides a sufficient level of privacy and usability? Our primary viewpoint is that the data usability has to be measured with the targeted applications of the query logs, as they pose different requirements of the data.

6. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proc. of the 29th SIGIR conference on Research and development in information retrieval*, 2006.

[2] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proc. of the 29th SIGIR conference on Research and development in information retrieval*, 2006.

[3] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE '05: Proceedings of the 21st International Conference on Data Engineering (ICDE'05)*, pages 217–228, Washington, DC, USA, 2005. IEEE Computer Society.

[4] E. Bertino, B. Ooi, Y. Yang, and R. H. Deng. Privacy and ownership preserving of outsourced medical data. In *ICDE*, 2005.

[5] S. Cucerzan and E. Brill. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of EMNLP 2004*, 2004.

[6] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Probabilistic query expansion using query logs. In *Proc. of the 11th international conference on World Wide Web*, 2002.

[7] D. Downey, S. Dumais, and E. Horvitz. Models of searching and browsing: Languages, studies, and applications. In *Proc. of IJCAI*, 2007.

[8] B. C. M. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In *Proc. of the 21st IEEE International Conference on Data Engineering (ICDE 2005)*, pages 205–216, Tokyo, Japan, April 2005.

[9] J. Gehrke. Models and methods for privacy-preserving data analysis and publishing. In *ICDE*, 2006.

[10] H. Hacigumus, B. Iyer, C. Li, and S. Mehrotra. Executing sql over encrypted data in the database service provider model. In *SIGMOD*, 2002.

[11] B. Hore, S. Mehrotra, and G. Tsudik. A privacy-preserving index for range queries. In *ACM Symposium on Principles of Distributed Computing*, 1997.

[12] T. Joachims. Optimizing search engines using clickthrough data. In *Proc. of the eighth SIGKDD international conference on Knowledge discovery and data mining*, 2002.

[13] K. LeFevre, D. Dewitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *ACM SIGMOD International Conference on Management of Data*, 2005.

[14] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *IEEE ICDE*, 2006.

[15] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Workload-aware anonymization. In *SIGKDD*, 2006.

[16] N. Li and T. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *To appear in International Conference on Data Engineering (ICDE)*, 2007.

[17] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, page 24, 2006.

[18] C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[19] L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, 2002.

[20] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis. State-of-the-art in privacy preserving data mining. *ACM SIGMOD Record*, 33(1), 2004.

[21] K. Wang and B. C. M. Fung. Anonymizing sequential releases. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, pages 414–423, Philadelphia, PA, August 2006.

[22] K. Wang, P. S. Yu, and S. Chakraborty. Bottom-up generalization: a data mining solution to privacy protection. In *Proc. of the 4th IEEE International Conference on Data Mining (ICDM 2004)*, November 2004.

[23] B. Yang and G. Jeh. Retroactive answering of search queries. In *Proc. of the 15th international conference on World Wide Web*, 2006.

[24] S. Zhong, Z. Yang, and R. N. Wright. Privacy-enhancing k-anonymization of customer data. In *PODS '05*, 2005.