Information Sciences 325 (2015) 215–226

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Constructing plausible innocuous pseudo queries to protect user query intention

Zongda Wu [a,b], Jie Shi [c], Chenglang Lu [a], Enhong Chen [b], Guandong Xu [d], Guiling Li [e,f,*], Sihong Xie [g], Philip S. Yu [g]

[a] Oujiang College, Wenzhou University, Wenzhou, China
[b] School of Computer Science, University of Science and Technology of China, Hefei, China
[c] School of Information Systems, Singapore Management University, Singapore
[d] Faculty of Engineering and IT, University of Technology, Sydney, Australia
[e] State Key Laboratory of Biogeology and Environmental Geology, China University of Geosciences, Wuhan, China
[f] School of Computer Science, China University of Geosciences, Wuhan, China
[g] Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA

ARTICLE INFO

Article history:
Received 30 May 2014
Revised 15 May 2015
Accepted 4 July 2015
Available online 13 July 2015

Keywords:
Knowledge
Privacy model
User intention
Query protection

ABSTRACT

Users of web search engines are increasingly worried that their query activities may expose what topics they are interested in, and in turn, compromise their privacy. It would be desirable for a search engine to protect the true query intention for users without compromising the precision-recall performance. In this paper, we propose a client-based approach to address this problem. The basic idea is to issue plausible but innocuous pseudo queries together with a user query, so as to mask the user intention. First, we present a privacy model which formulates plausibility and innocuousness, and then the requirements which should be satisfied to ensure that the user intention is protected against a search engine effectively. Second, based on a semantic reference space derived from Wikipedia, we propose an approach to construct a group of pseudo queries that exhibit similar characteristic distribution as a given user query, but point to irrelevant topics, so as to meet the security requirements defined by the privacy model. Finally, we conduct extensive experimental evaluations to demonstrate the practicality and effectiveness of our approach.

1. Introduction

Web search engines such as Google, Yahoo! and Microsoft Bing are becoming increasingly important in people's daily activities. As pointed out in [1–3], while search engines enable users to retrieve information from the Internet intuitively and effectively, the queries issued by these users can potentially compromise their privacy, i.e., the queries themselves can lead to an undesirable disclosure of user activities and topics of interest, and even confidential personal or business profiles.

It has been pointed out in [2,4] that the problem of disclosing user query intentions cannot be solved by using an anonymization scheme (e.g., those in [5,6]) to process a query log. For example, in 2006, AOL released an anonymized query log of around hundreds of thousands of randomly selected users [1,7]. The log data had been anonymized by removing individual

* Corresponding author. Tel.: +86 5513601551.
E-mail addresses: zongda1983@163.com, wuzongda@ustc.edu.cn (Z. Wu), shijie1123@gmail.com (J. Shi), chenglang.lu@qq.com (C. Lu), enhc@ustc.edu.cn (E. Chen), guandong.xu@uts.edu.au (G. Xu), guiling@cug.edu.cn (G. Li), xiesihong1@gmail.com (S. Xie), psyu@uic.edu (P.S. Yu).

http://dx.doi.org/10.1016/j.ins.2015.07.010
0020-0255/

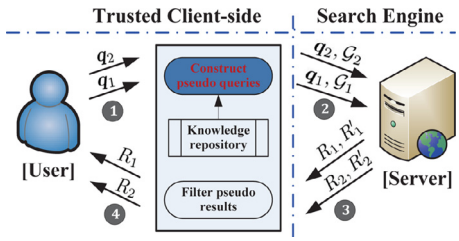

Fig. 1. The system model that we use, where the part "construct pseudo queries" is the key point.

identification information (e.g., IP address, username) associated with each user, while only keeping actual query text, timestamp, etc. However, such simple anonymization was proved ineffective, because user queries themselves still contained identification information [1]. It was shown that detailed user profiles (e.g., age, gender, location) could be constructed from the anonymized log data [8].

In addition, the problem also cannot be solved by other well-known solutions with regard to user privacy protection, such as cryptographic protocols [9–12], and Private Information Retrieval (PIR) [13,14]. As pointed out in [2,4], cryptographic protocols such as searchable encryption protocols [11] are not applicable to modern text search engines, because they cannot support similarity retrieval; moreover, PIR is also not practical, because it not only has high performance overheads, but also requires changes to existing search engines.

Recently, a system model is proposed to protect user privacy by masking the user query topics [1,2,15]. Its basic idea is to hide each user query among some pseudo queries, without any change to existing search engines. However, the system model seems to lack a practical implementation: for the approach proposed in [15] or [2], the generated pseudo queries are not meaningful, thus can be easily ruled out; for the approach proposed in [1], the results returned by the generated queries are not a superset of the genuine results, i.e., it is required to compromise the precision-recall performance (see Section 2 for detail).

In this work, we aim to prevent a search engine from identifying users' topics of interest (also called user intention) according to search terms, under the constraints of not compromising the precision-recall performance and not changing the search engine. To this end, we adopt the system model proposed in [1,2,15], that is, we attempt to mask the intention hidden in a user query by using well-designed pseudo queries. Fig. 1 presents the system model, which consists of an untrusted search engine and a number of clients (users). Each client accessing the search service trusts no one but himself/herself. As shown in Fig. 1, the pseudo queries $\mathcal{G}_i$ are constructed in a trusted client, and submitted together with the user query $\boldsymbol{q}_i$ to the search engine. Then, the search results R_i' that correspond to the pseudo queries $\mathcal{G}_i$ are discarded by the client, so only the search result R_i that corresponds to the user query $\boldsymbol{q}_i$ is returned to the user. It can be seen that the system model is transparent to both the search engine and the client, i.e., it requires no change to existing search engines; moreover, the result returned from the search engine is certainly a superset of the genuine result of a user query, thus it requires no compromise to the precision-recall performance.

However, it can also be seen that the quality of the pseudo queries generated by the client is very important in the system model, e.g., randomly constructed pseudo queries are often easy to be detected by the untrusted search engine, thus failing to hide the user query intention. To this end, given any user query, we aim to construct a group of pseudo queries that satisfy the following two requirements:

- **Plausibility**, i.e., the pseudo queries should exhibit similar characteristic distribution as the genuine user query. A user query is likely to include characteristic terms, e.g., synonymy, polysemy and high-specificity, thus, making it easy to be detected. For example, given two queries "X86 SSE4" and "puma cougar", where the first contains two terms of high-specificity, and the other contains two synonymous terms, such a characteristic distribution makes them unlikely to be randomly generated, so they are probably genuine.
- **Innocuousness**, i.e., the topics of the pseudo queries should be semantically-irrelevant to those of the user query, so that they are innocuous to the genuine user intention. For example, given a user query "Nike sneaker", an ideal pseudo query could be "Intel processor", because it has characteristics similar to the user query (plausible), but points to other irrelevant topics (innocuous).

The above requirements entail the following three challenges: (1) identifying the true intention for a user query; (2) capturing key characteristics inherent in the user query; and (3) constructing pseudo queries that have similar characteristics to the user query but point to innocuous topics.

It should be pointed out that in this work we mainly focus on protecting against a search engine. A search engine is deemed to be the most powerful potential adversary, because it possesses the most information, e.g., it hosts the plaintext corpus and executes the query processing algorithms [2]. However, we exclude tampering concerns posed by active adversaries, which have been addressed extensively in the context of query result authentication [16]. This work is also orthogonal to the privacy of user identity, which may be mitigated by query log anonymization, or by letting users connect to the search engine using an anonymous network [6].