Information Sciences 218 (2013) 17–30

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Knowledge-based scheme to create privacy-preserving but semantically-related queries for web search engines

David Sánchez, Jordi Castellà-Roca, Alexandre Viejo *

Departament d'Enginyeria Informàtica i Matemàtiques, UNESCO Chair in Data Privacy, Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Spain

ARTICLE INFO

Article history:
Received 22 March 2011
Received in revised form 21 May 2012
Accepted 23 June 2012
Available online 1 July 2012

Keywords:
Complex query
Ontologies
Privacy
Private information retrieval
Semantic analysis
Web search

ABSTRACT

Web search engines (WSEs) are basic tools for finding and accessing data in the Internet. However, they also put the privacy of their users at risk. This happens because users frequently reveal private information in their queries. WSEs gather this personal data and build user profiles which are used to provide personalized search (PS). PS improves the users' search results and, hence, it is a key element for the successfulness of WSEs: the entity that offers the best searching experience should attract more users. Nevertheless, profiles can also be used in an improper way by WSEs or they can be stolen by attackers. This situation requires privacy-preserving schemes able to handle from simple queries (one single term) to complex queries (several words with or without relation). Generally, these systems generate and submit inaccurate queries in order to provide privacy, but these queries must be carefully built in order to keep the usefulness of the user profiles. Current literature does not address the generation of privacy-preserving and useful complex queries. Therefore, this paper presents a new scheme that generates distorted user queries from a semantic point of view in order to preserve the usefulness of user profiles. Besides, linguistic analysis techniques are used to properly interpret complex queries performed by users and generate new semantically-related ones accordingly. The performance of the new scheme is evaluated in terms of semantic preservation of new queries, privacy level and runtime. A set of query logs taken from real users and compiled by AOL is used as test data.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

Web search engines (WSEs) – e.g. Google or Bing among others – are essential tools in order to find specific data on the Internet. Everybody is aware of the vast volume of information which is hold by the virtual world and how fast it grows.

WSEs present their search results using several result pages (web pages containing links to the resulting data). Studies show that 68% of the users click a search result within the first page of results and 92% of the users click a result within the first three pages of search results [16]. As a consequence of that, WSEs are interested in ordering the search results (specifically, the links connected to them). This is done according to two factors:

- *Purchased positions*. Companies looking for visibility (i.e. advertising for the goods or services that they offer) can purchase a better position in the ranked result pages as a service offered by WSEs.

* Corresponding author.
E-mail addresses: david.sanchez@urv.cat (D. Sánchez), jordi.castella@urv.cat (J. Castellà-Roca), alexandre.viejo@urv.cat (A. Viejo).

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2012.06.025

- *Better user experience*. WSEs put in the first result pages the links which are more interesting for the users. This is an indirect source of revenues: the WSE that offers the best experience should attract more users and companies are expected to choose the most popular WSE to place their advertisements.

An example of the first factor (purchased positions) is *Google AdWords* [12]. In this scheme, advertisers select the words that should trigger their advertisements. When a user searches on Google, advertisements for relevant words are shown as "sponsored links" in the search results page. This process is quite simple and it can be straightforwardly deployed.

The second factor is more complicated: it is not easy to know the interests of the users. The word "Mercury" is an example of this situation: this term can refer to the planet Mercury or to an element in the periodic table. The concept *disambiguation* represents the process of identifying the correct sense when a certain word has different ones. Personalized search (PS) [33,40,45,60] uses it in order to provide personalized results to users.

The disambiguation process requires knowledge of: (i) the interests of the user or (ii) the query context. Both can be obtained from the *user profile*. For example, if the user profile contains "Astronomy" among the interests, the WSE will assume that "Mercury" refers to the planet Mercury and not to the element in the periodic table.

User profiles can be build using several tools: [48] proposes the use of the browsing history. In [36], the authors use clickthough data. Two schemes presented in [46,38] consider the use of web communities for this purpose. A client side application which stores users' interests is presented in [52]. Finally, some schemes [46,14] use the search queries which have been previously submitted by the users. WSEs generally use this last approach because it is very effective and it profiles the users without their collaboration.

The use of profiles enables the WSEs to offer a better user experience. Nevertheless, this interesting feature is not provided without cost: profiles built from search queries contain personal information which can univocally identify their owners. In this way, the authors in [19] study the application of simple classifiers to map a sequence of queries into the gender, age, and location of the user issuing the queries. Then, they show how these classifiers may be carefully combined at multiple granularities to map a sequence of queries into a set of candidate users that is 300–600 times smaller than random chance would allow. This paper shows that the proposed approach remains surprisingly accurate even after removing personally identifiable information such as names, digits and places or limiting the size of the query log. The results of this paper explain the privacy concerns raised by the AOL scandal [2]: in this incident, Thelma Arnold, user of the AOL's WSE, was identified by her searches submitted over a 3-month period.

WSEs store complete profiles that contain sensible information about their users. Nevertheless, users are generally not aware of this behavior. This improper acquisition of personal information affects them in two different ways:

- The big Internet companies sell user profiles to law enforcement agencies. For example, AOL handles nearly 1000 requests each month for information in criminal and civil cases [15]. Facebook receives between 10 and 20 requests for this kind of information each day [49]. Recently, the Yahoo Compliance Guide for Law Enforcement was disclosed. This document specifies that Yahoo charges the government about 30–40$ for the contents, including e-mail, of a subscriber's account and 40–80$ for the contents of a Yahoo group [61]. Therefore, user profiles represent a source of revenues for WSEs. Nevertheless, the users (the real owners of that information) usually get no income from this item.
- Profiles provide valuable information, hence, attackers are motivated to steal them and get diverse benefits. Therefore, WSEs are responsible for storing user profiles in a secure place and applying access control measures to that content. The AOL scandal [2], which have been explained above, proves that WSEs do not always fulfill these requirements. In that particular case, 20 million queries made by 658,000 users were publicly disclosed by employees of the company itself.

Both points show that users should prevent the WSEs from storing and using their profiles in an uncontrolled way. However, profiles are needed in order to provide an efficient service to the users. Thus, allowing the WSEs to only profile them in an *inaccurate* way may be a proper solution that addresses these two issues. This implies that the resulting profile might be detailed enough to allow the personalized search feature but inaccurate enough to avoid the disclosure of personal information considered risky.

As explained above, in the WSE scenario, users are generally profiled based on their search queries. Therefore, a straightforward way of obtaining inaccurate profiles is to submit imprecise queries to the WSE. The level of inaccuracy of these queries directly affects the resulting profile. In this way, queries about topics which are far away from the real interests of the user should generate a heavily distorted profile. On the other hand, queries about general contents which are related to her true interests should produce a more general but still useful profile.

In addition to the trade-off between quality of service and privacy, the kind of queries which are submitted by the users have to be considered too. Generally, a query can contain from one single word up to several words. The latter is the worst case because there is no pattern that defines the structure of a query: sometimes it can be a quite accurate sentence but in other cases it can be some words together which make no sense at all. We name this kind of queries as *complex queries*. Privacy-preserving mechanisms should be able to deal with complex queries in order to be usable in a real scenario because they are usually used by the users of WSEs [1].

Finally, it is worth to mention that the privacy issues related to the users of WSEs also affect companies and their employees. In this way, the work presented in [34] introduces the concern of confidentiality protection of business information for the publication of search engine query logs and derived data.

Download English Version:

https://daneshyari.com/en/article/392773

Download Persian Version:

https://daneshyari.com/article/392773

Daneshyari.com