Expert Systems With Applications 64 (2016) 523–535

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Working at the web search engine side to generate privacy-preserving user profiles

David Pàmies-Estrems[a], Jordi Castellà-Roca[b], Alexandre Viejo[b,*]

[a] Aneris, Raval Sta. Anna 43, 2n 2a, E-43201 Reus, Spain
[b] Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, UNESCO Chair in Data Privacy, Av. Països Catalans 26, E-43007 Tarragona, Spain

ARTICLE INFO

Article history:
Received 7 May 2016
Revised 22 June 2016
Accepted 4 August 2016
Available online 5 August 2016

Keywords:
Web search engines
Privacy
Query logs
Data stream
Data monetization

ABSTRACT

The popularity of Web Search Engines (WSEs) enables them to generate a lot of data in form of query logs. These files contain all search queries submitted by users. Economical benefits could be earned by means of selling or releasing those logs to third parties. Nevertheless, this data potentially expose sensitive user information. Removing direct identifiers is not sufficient to preserve the privacy of the users. Some existing privacy-preserving approaches use log batch processing but, as logs are generated and consumed in a real-time environment, a continuous anonymization process would be more convenient. In this way, in this paper we propose: (i) a new method to anonymize query logs, based on k-anonymity; and (ii) some de-anonymization tools to determine possible privacy problems, in case that an attacker gains access to the anonymized query logs. This approach preserves the original user interests, but spreads possible semi-identifier information over many users, preventing linkage attacks. To assess its performance, all the proposed algorithms are implemented and an extensive set of experiments are conducted using real data.

1. Introduction

Nowadays, surfing the Web usually requires using a Web Search Engine (WSE) as main entry point. Individuals interact with WSEs by submitting search queries that contain certain keywords related to the topics they are looking for. WSEs, then, retrieve and present, with a minimal response time, an accurate list of web sites that tackle those topics. The popularity of WSEs has grown with the number of websites present on the Internet. According to a recent study (Netcraft, 2016), the number of websites is nearly doubling every three year period; therefore, it can be assumed that WSEs such as Google or Bing will continue to be essential players in the next years.

When a WSE looks for the requested information among its indexed web pages, it also stores the submitted query (i.e., the keywords) and some metadata (e.g., date of the query, some identifiers of the query issuer, which specific search result was selected by the issuer, etc.). The files that contain all the recorded queries and their related metadata are generally referred as *query logs* (Chau, Fang, & Liu Sheng, 2005). As a result, everyone who uses a WSE is disclosing personal data, such as personal characteristics and preferences, and enabling WSEs to compile those query logs, which are part of the huge bunch of data that has been recently designated as *Big Data*.

Turning this data into knowledge is a main interest for many companies. Knowledge is gathered using several techniques related to data mining and machine learning. The result of these processes is the extraction of valuable information that enables those companies to obtain economic benefits. This process is known as *data monetization* (Saagar, 2014) and it was recently put on the spotlight by the U.S. Federal Trade Commission when this organization published a report about data collection and use practices of the most relevant Data Brokers (U.S. Federal Trade Commission, 2014).

Once the query logs are generated, they are processed and analyzed in order to build user profiles. In the literature, a user profile is generally considered a set of well-defined categories of interests (e.g., science, health, society, sports, etc.) with a certain weight assigned to each one according to the evidences generated by the corresponding user and how they have been classified under each category (Viejo, Sánchez, & Castellà-Roca, 2012). When focusing on WSEs, search queries are the evidences, and the amount of queries from each user classified under each category reflects the related weight.

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

http://dx.doi.org/10.1016/j.eswa.2016.08.033
0957-4174/

WSEs process and analyze query logs and related user profiles in order to perform the following services:

- *Personalization*: Providing relevant results to their users' needs, where relevant links are placed in the first positions of the returned results. This is achieved by analyzing the past queries submitted by users; this knowledge allows WSEs to contextualize and disambiguate next queries (Shen, Tan, & Zhai, 2007). In this way, if the user searches for "Mercury" and her profile indicates that she is interested in "Astronomy", the WSE will put the results that correspond to the planet Mercury in the first pages (instead of the chemical element).
- *Improved web search*: By knowing the frequencies of most formulated queries and most selected results, WSEs are able to improve the ranking algorithms (Agichtein, Brill, & Dumais, 2006) and to suggest reformulated queries that can add specificity to the user's initial query (Jones, Rey, Madani, & Greiner, 2006). Suggestions can be offered while the user is typing her query or also after retrieving poor search results for a query submission (Cooper, 2008). Following with the example of the term "Mercury", if a user inputs just this term in the text-field of Google's WSE, it will suggest as better alternative queries: "Mercury - Planet", "Mercury - Element", "Mercury poisoning" and "Mercury marine" among others. These alternative keywords are expected to retrieve more accurate results to the user.
- *Marketing*: Characterizing general user profiles, user behavior and user search habits, it's possible to improve keyword advertising campaigns and extract market tendencies among others (Brenes & Gayo-Avello, 2009; Korolova, Kenthapadi, Mishra, & Ntoulas, 2009; Poblete, Spiliopoulou, & Baeza-Yates, 2007). More concretely, search analytics is one of the cornerstones of the so-called *Search Engine Marketing* and it is in charge of using search data to investigate particular interactions among Web searchers, the search engine, or the content during searching episodes (Jansen, 2006). For example, Google Trends[1] is a service that exploits this kind of data. In particular, Google Trends shows how often a particular term is searched according to the total search-volume across various regions of the world.
- *Research*: As stated in Richardson (2008), query logs contain information that would never be available to researchers using conventional data-collection techniques. For example, a medical researcher might discover that people with asthma tend to wear wool, or live in areas with coal power plants; a sociologist may study how ideas spread from one person to an entire community, or may investigate the differences between the interests that individuals claim to have during face-to-face interviews and the real interests that their search queries reveal (Tancer, 2008); a political scientist might learn about democracy by studying the evolution of political searches by users in a developing country; and a computer scientist may study and analyze new Information Retrieval (IR) algorithms via a common benchmark query log (Bar-Ilan, 2007). Last but not least, query logs also enable researchers to ask questions that would normally require going backward in time. For example, a medical researcher might study people diagnosed with diabetes today to find out what their primary symptoms were six months ago. Asking them directly, once they learn they have diabetes, may result in subjective bias (Richardson, 2008).

In addition to those benefits, building and exploiting query logs may lead to serious privacy problems as well. More specifically, the keywords of each query and the related metadata may provide to anyone who has access to the logs with sensitive information from the users such as behaviors, habits, interests, religious views, sexual orientation, etc. Even worse than that, some query contents may contain identifiers and quasi-identifiers which may allow to link a certain query with a real person. An example of this situation happens with the so-called vanity searches (Soghoian, 2007) in which an individual looks for its own name on the Internet.

Although query logs could be protected prior to their publication, there is no absolute guarantee of anonymity, as the combination of modified data may disclose enough information to re-identify some users (Jones, Kumar, Pang, & Tomkins, 2007; Poblete et al., 2007). As an example, there is one well-known case, the AOL scandal, in which around 36 million queries performed by AOL's costumers were publicly disclosed. Although records were previously de-identified, it was possible to identify some users from the disclosed query logs and other sensitive information was exposed (Barbaro & Zeller, 2006). This case ended up with an important damage to AOL users privacy and to AOL itself, with several lawsuits against the company (Mills, 2006).

Therefore, in order to get viable data monetization, better tools capable of modifying query logs by limiting the privacy disclosure risk but preserving as much data utility as possible should be provided. More specifically, in this paper, we propose using a server side software capable of processing queries in real time and building anonymized query logs that still retain enough data utility to allow its monetization. As a result, WSEs may then offer the protected query logs to external organizations for data monetization purposes while keeping the real query logs in a safe place; otherwise, WSEs may also decide to only keep protected logs, getting in turn a lower risk of information disclosure in case of a direct attack.

The rest of this paper is organized as follows: in Section 2, we summarize the most relevant research published up to date. In Section 3, we define the method used to implement the system. In Section 4, we analyze the results. In Section 5, we conclude and point out some future lines of research.

2. Previous work

To avoid record-linkage attacks, the concept of k-anonymity is proposed in Samarati and Sweeney (1998). A k-anonymized dataset has the property that each record is indistinguishable from at least $k-1$ other records, no individual can be re-identified with probability exceeding $1/k$ through linking attacks alone. Currently there are two main approaches to protect users' records: batch or real-time processing.

2.1. Batch processing

Historically, these were the first proposed methods. Because all the logs are stored, it's possible to remove the queries that disclose the identities of the individuals by means of basic statistic or semantic methods. We next detail the different methods that belong to this category considering two main subcategories: (i) query removal; and (ii) statistical disclosure control (SDC) and semantics.

2.1.1. Query removal

As shown in Cooper (2008), several approaches dealing with the query log anonymization problem based on query removal or hashing can be found in the literature. More concretely, some protocols and methodologies simply remove old query sets assuming that query logs will not be large enough to enable identity disclosure, however, this assumption does not take into account the existence of highly identifying queries (Barbaro & Zeller, 2006).

A more appropriate approach suggests deleting only infrequent queries (Adar, 2007), assuming that those are more likely to refer to identifying or quasi-identifying information. However, this is

[1] Google Trends. https://www.google.com/trends/