Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

Preventing automatic user profiling in Web 2.0 applications

Alexandre Viejo, David Sánchez *, Jordi Castellà-Roca

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 20 January 2012
Received in revised form 7 June 2012
Accepted 5 July 2012
Available online 16 July 2012

Keywords:
Profiling
Privacy
Knowledge-based system
Web 2.0
Twitter

ABSTRACT

The rise of the Internet and Web 2.0 platforms have brought very accessible publishing techniques that have transformed users' role from mere *content consumers* to fully *content consumers–producers*. Previous works have shown that user-generated content can be automatically analyzed to extract useful information for the society. Nevertheless, researchers have also shown that it is possible to build individual user profiles automatically. This situation may provoke concerns to the users worried about their privacy. In this paper, we present a new scheme that effectively obfuscates the real user's profile in front of automatic profiling systems, while maintaining her publications intact in order to interfere the least with her readers. The proposed system generates and publishes fake messages with terms semantically correlated with user posts to distort and, hence, hide the real profile. Our method has been tested using Twitter, a very well-known Web 2.0 microblogging platform. Evaluation results show that this new scheme effectively distorts user profiles, producing uniform (i.e. balanced) profiles that hardly characterize users and outperforming simpler methods based on random distortions. In addition to that, the presented system is adaptive, capable of profiling and anonymizing users with a quite limited number of publications and it reacts quickly to any variation in their interests.

1. Introduction

The Internet is continuously evolving and creating new possibilities of interaction for its users. Nowadays, the rise of the Web 2.0 and its related technologies have transformed *passive users* into *active* ones. The term Web 2.0 is associated with web applications (*e.g.*; social networking sites, blogs, wikis, video sharing sites, etc.) that allow users to interact and collaborate between them in a social dialogue using highly accessible publishing techniques. These new technologies enable users to both generate and consume content. As a result, they have evolved from *content consumers* to *content consumers–producers*.

From the different publishing methods which are offered in any Web 2.0 application, this paper focuses on the text-based posts. This service is known as *blogging* when there is no limitation in the length of the posts or *microblogging* when they are limited to a certain (and usually small) number of characters.

Users publish text-based posts containing opinions and information about a broad range of topics. These topics can be general (*e.g.*, films, music, sports, etc.) but also personal (*e.g.*, current activity, current localization, current feeling, etc.). Regarding what motivates people to share information on the Internet [30], states that people contribute in the Web to seek attention from others. By getting attention, they obtain publicity, vanity or ego gratification from peer recognition.

Researchers have shown that the information shared by the users of Web 2.0 technologies can be automatically analyzed and relevant data can be retrieved. For example, user-generated content has been used to forecast flu trends [3], develop earthquake warning systems [32], compile product recommendations [15], build interest-based recommendation systems [20] or help business to predict future sales [31] among others.

Tools which are used to automatically analyze user-generated content can extract useful knowledge in an *aggregated* way or in an *individual* way. The former groups the information gathered from several users and, hence, it weakens the link between the data and the person who has generated it. On the contrary, the latter approach builds a *complete profile* of each analyzed user, explicitly linking all the user characterization with her identity.

Since some of the topics included in the gathered data may refer to *personal* data, user profiling clearly puts at risk the privacy of the users of the Web 2.0. More specifically, the *Consumer Reports'2010 State of the Net analysis* [9] states that more than half of users of social networks and similar applications share private information about themselves online. As explained in [46], user characterization and leakage of personal data may invite malicious attacks from the cyberspace (*e.g.*; personalized spamming, phishing, etc.) and even from the real world.

In order to alleviate these problems, Web 2.0 applications offer *privacy settings* to allow users to control who has access to certain

* Corresponding author. Tel.: +034 977 556563; fax: +034 977 559710.
E-mail address: david.sanchez@urv.cat (D. Sánchez).

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.07.001

contents generated by them. However, this basic privacy-preserving tool suffers from the following problems:

- These privacy settings are generally not sufficiently understood by the average user who seldom changes the default configuration which is provided by the company that owns the web application [43]. To make it worse, these companies offer privacy configurations that make most of their information public by default and they require the users to make a choice if they wish to keep information private [5].
- A recent study from Barracuda Labs (a security company) states that 30% per cent of Twitter users are worried about how this company protects their privacy [45]. Specifically, Twitter offers access to its archive of billions of posted user messages dating back to January 2010 [7,10,19] and it has been reported that individuals are concerned about their data being exploited by advertisers in order to target them [10]. It can be assumed that these privacy concerns can be extended to other similar Web 2.0 platforms.
- A privacy-preserving method that is based on restricting the visibility of the user-generated content compromises as well the capability of the users to gain attention from others. This situation can represent a strong limitation to the use of those privacy settings because, as explained above, seeking attention is the main motivation of the users of Web 2.0 applications.

As a result, users who are willing to decrease the chances of privacy breaches should use privacy-preserving mechanisms deployed and managed by themselves to prevent data extraction techniques from profiling them in an accurate way. In addition to that, these methods must protect the privacy without limiting the visibility of the user-generated data.

1.1. Contribution and plan of this paper

Text-based posts published on Web 2.0 sites can contain relevant and useful information for the society. Therefore, developing text-based data extraction methods for gathering this data in an effective way is an important field of research. However, those techniques can be also used to automatically profile content generators and jeopardize their privacy. As a consequence, measures to prevent automatic user profiling in Web 2.0 applications should be proposed.

In this paper, we offer three contributions to this research field:

1. We propose a new knowledge-based profiling approach grounded in the Information Theory that dynamically quantifies the amount of information provided by terms contained in the text messages published by users to accurately characterize their profile according to a set of categories.
2. We present an adaptive method that, according to the characterized profile, distorts it to prevent automatic data analysis techniques from profiling the user, while maintaining her published data intact. Current profilers mainly characterize users according to the distribution of terms appearing in their messages. Those general profilers cannot detect the fake queries generated by our proposal.
3. We use Twitter (the leading social network based on text-based messages [25]) to test both the profiling and profile obfuscation proposals. Results show that our profiler characterizes users faster and more accurately than methods based solely on absolute term frequencies, whereas our profile distortion method effectively balances user profiles in front of automatic profilers based on information distribution.

Section 2 introduces the state of the art related to text-based profiling techniques used in the Web 2.0 environment; it also reviews privacy-preserving approaches found in the field. Section 3 details our new proposals. Section 4 evaluates the accuracy of our profiling method and the obfuscation level provided by our masking scheme. Finally, Section 5 reports some concluding remarks.

2. State of the art

As explained above, this section first covers automatic profiling techniques applied in Web 2.0 applications. Then, it reviews methods which can be used to prevent automatic user profiling from textual data.

2.1. Text-based automatic profiling methods

In the work presented by Ebner et al. [12], the authors try to categorize different users in an unsupervised manner according to the overlapping keywords found in their published messages. In order to extract the keywords of the messages, the *Yahoo Term Extraction Web Service* is suggested. The authors test this approach using the messages which were published on Twitter by the attendants to a Conference. The results of their evaluations show that user posts present high diversity and nearly no overlapping keywords that prevents from achieving an accurate profiling. Authors then argue that a knowledge-based semantic analysis is needed to deal with the high keyword diversity. Authors propose to manually linking each keyword with its related category.

In [47] Zoltan and Johann present a knowledge-based framework that builds user profiles from text messages shared in social platforms. To do so, authors extract Named Entities and keywords and match them to categories found in a knowledge base. Specifically, they exploit Linked Data vocabularies (such as DBpedia [11]) as knowledge base. Authors leverage the contribution of extracted information to the user profile according to their degree of occurrence (*i.e.,* TF-IDF [33]) with respect to the linked categories. As a result, user profiles are characterized according to a set of weighted categories.

In [26] Michelson and Macskassy present a similar approach, discovering users' topics of interest by examining the Named Entities they mention in their posts in Twitter. First, the entities in each message are found by using simple capitalization heuristics. The author argues that this can be challenging because tweets are generally ungrammatical and noisy. Then, each entity is disambiguated and categorized using Wikipedia as a knowledge base. The process is as follows: the terms contained in the publication are considered to be the context for that entity. Then, this context is compared to the text of each candidate entity's page from Wikipedia. The entity from Wikipedia which have more term occurrences is selected. Finally, the tree of Wikipedia categories related to the selected entity is retrieved. Due to the complexity of Wikipedia category trees, in the last step, the proposed scheme filters them by selecting the categories (nodes in the trees) that occur frequently to generate useful topic profiles.

In [4] Bernstein et al. discuss the problems of relying on term occurrence and co-occurrence to identify topics of messages published in a social network like Twitter. They argue that the current best practices for topic identification assume that user posts are of a decent length. Since, messages in Twitter (and other Web 2.0 applications) are at most 140 characters long, the authors explain that this assumption fails almost by definition. Additionally, users usually compress similar words to gain space in order to insert their opinions. To tackle these problems, [4] present a novel approach based on transforming noun phrases found in each user