Profiling Social Networks to Provide Useful and Privacy-Preserving Web Search

Alexandre Viejo, David Sánchez

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

Corresponding author: David Sánchez; E-mail: david.sanchez@urv.cat

Abstract

Web search engines (WSEs) use search queries submitted by users to profile them and to provide personalized services such as query disambiguation or query refinement. On the one hand, these services are valuable for the users because they get an enhanced web search experience. On the other hand, the compiled user profiles may contain sensitive information which might represent a serious privacy threat. This privacy issue should be addressed, but it must be done in a way that it also preserves the utility of the profile with regard to web search services. State-of-the-art approaches tackle these issues by generating and submitting synthetic queries that are fake but related to the real general interests of the user. This technique allows the WSE to only know general (and useful) data while the detailed (and potentially private) data is obfuscated. To build fake queries, these proposals rely on past user queries to obtain the interests of each user. However, we argue that this is not always the best strategy and, in this paper, we study the use of social networks to gather more accurate user profiles that enable a better personalized service while offering a similar, or even better, practical privacy. These hypotheses are empirically supported by evaluating the performance of the proposed system using real profiles gathered from Twitter and a set of search queries extracted from the AOL's files.

Keywords: privacy, data utility, web information retrieval, profiling, social networks

Profiling Social Networks to Provide Useful and Privacy-Preserving Web Search

Introduction

Web Information Retrieval is the process where users submit queries to a *Web Search Engine (WSE)* and get related search results. During this process, WSEs store in their databases the list of queries sent by each user. These collections or data are named *query logs*. The identification of the source of each query is essential to generate valid query logs. This process can be performed by means of stored browser cookies (Collusion mozilla add-on, 2013), identical IP addresses or identical browser configurations (Eckersley, 2010) among others.

WSEs use query logs to improve the accuracy and personalization of web information retrieval (Cooper, 2008). Among the different processes where they are involved, we stress their key role in profiling users (i.e., building *user profiles* that contain their interests). For example, if a certain user has a list of past searched terms such as “solar system”, “super nova”, “galaxy”, etc. It can be inferred that she is interested in “Astronomy” and, hence, her user profile will reflect this interest.

User profiles are essential to provide *personalized web search* services (Shapira & Zabar, 2011). These can offer an enhanced web search experience to the users by means of *query disambiguation* (Makris, Plegas, & Stamou, 2012) or *query suggestion/refinement* (Shi & Yang, 2007), among others. *Query disambiguation* is used to sort search results which are shown to a certain user. 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). Regarding *query suggestion/refinement*, this mechanism focuses on suggesting new queries or refining already queried ones with additional

terms that can provide more accurate search results according to user's interests. Suggestions can be offered while the user is typing her query or also after retrieving poor search results for a certain query submission (Cooper, 2008).

The motivation behind the application of *personalized web search* is twofold: on one hand, WSEs use it to provide an enhanced web search service; on the other hand, WSEs may exploit *user profiles* to get important economical revenues. More specifically, they can be used to sell personalized advertising (Hansell, 2006; Ortiz-Cordova & Jansen, 2012) or they can also be directly sold to law enforcement agencies (e.g., AOL and Facebook have been reported to handle several requests for information in criminal and civil cases (Hansell, 2006; Summers, 2009)).

The possibility that those generated profiles might be linked with real users implies a significant privacy threat that is generally neglected: along with probably innocuous data such as general interests, these profiles may also contain unequivocal sensitive information regarding diseases, sexual tendencies, economical status, etc.

Some situations reported in the literature show that the re-identification of users and the subsequent disclose of sensitive data is quite real. More specifically, (Barbaro & Zeller, 2006) were able to identify a certain user using a query log of 20 million queries made by 658000 users which was publicly disclosed by AOL. In fact, some researchers (Ye, Wu, Pandey, & Chen, 2008) have argued that, even though the identifying elements used to group the queries sent by the same source in the query logs (e.g., IP addresses, browser cookies, etc) do not directly identify users, they can be a very effective tool to achieve that. For example, an *Internet Service Provider* can connect the IP address linked to a group of queries with the name of the user who submitted them. The WSE alone can do the same if the user is submitting queries while logged in

her account of that WSE (Google History, 2013). Moreover, it is worth to mention that the users' behavior may also enable their identification. For example, users might submit queries containing personal information such as their own name, national ID, etc (Jones, Kumar, Pang, & Tomkins, 2008).

According to all the explained above, new tools to enable privacy-preserving web search should be designed to enable the users to restrict the amount of unequivocal sensitive information that appears in their profiles built by the WSEs. Nevertheless, it must be stressed that these profiles are needed to provide an efficient personalized service to users, hence, any provided solution must address both privacy and usability requirements. Last but not least, these new tools should not require any collaboration from the WSEs that, being stakeholders, depict lack of motivation in preserving the privacy of their users.

Previous work

In general, all the schemes that try to provide privacy-preserving web search follow two main approaches:

- *Conceal the real identity of the user in front of the WSE*. Using dynamic IPs and a plain web browser without cookies is a simple example of this. Other methods include the use of anonymizing proxies (e.g., TOR (Dingledine, Mathewson, & Syverson, 2004)) or similar. As a result, queries cannot be linked to users and, hence, WSEs are unable to build profiles. This approach pursuits user anonymity.
- *Distort the user profile by submitting fake queries to the WSE*. This privacy-preserving scheme submits a certain number of *fake* queries (they can be synthetically generated or even gathered from other users) to the WSE on behalf of the user together with the user's *legitimate* queries. With this approach, WSEs are capable of

linking users with queries and building their profiles. However, query logs will contain a mix of *legitimate* and *fake* queries (user interests), so that *legitimate* sensitive information cannot be unequivocally identified. This approach preserves data confidentiality.

We have previously stressed the interest in providing privacy together with personalized search. Therefore, schemes based anonymity, which prevent WSEs from building profiles, are not suitable. As a result, in the rest of this document, we focus on those proposals aiming at confidentiality, which just *distort* or *obfuscate* user profiles while keeping a degree of query utility for personalized search services. Schemes that follow this approach can be divided into two main categories: *multi-party* and *single-party*.

Multi-party protocols require the collaboration of external entities such as other human users, central servers, etc. In all the cases, these methods suffer from *slow response time* (due to the interaction required with other parties) and *system availability* (external entities must be online and available). Both are major issues when considering that users are used to a response time of 300 ms (Castellà-Roca, Viejo, & Herrera-Joancomartí, 2009) for Google queries, while multi-party schemes such as (A Viejo & Castellà-Roca, 2010), (Romero-Tris, Viejo, & Castellà-Roca, 2011) or (Castellà-Roca et al., 2009) report response times between 3 and 6 seconds.

Regarding the *single-party protocols*, these systems work directly in the computer of the user and they do not require a direct interaction with any external party. This gives these schemes two important advantages: (i) they are suited to provide fast response times; and (ii) they enable users to control the content of the queries which are submitted to the WSE and, hence, the level of detail of the profile that is being built by this entity. Considering to these benefits, this paper focuses on single-party schemes.

TrackMeNot (Howe & Nissenbaum, 2009) and GooPIR (Domingo-Ferrer, Solanas, & Castellà-Roca, 2009) are well-known single-party schemes that are based on submitting random queries to the WSE. Specifically, TrackMeNot periodically submits *fake* queries that are randomly gathered from blog entries and news headlines among others. Regarding GooPIR, this system submits a unique query to the WSE that contains *fake* terms together with the *legitimate* ones. *Fake* terms are obtained from a Thesaurus without considering their similarity to the *legitimate* contents.

Both schemes use different strategies to generate *fake* queries but, in both cases, *fake* contents are mainly random. In a scenario where a certain user employs any of these two mechanisms against an unaware WSE, her resulting profile will contain a mix of authentic and random interests. As a result, the WSE will be unable to provide a satisfactory personalized service, since the analysis of random *fake* queries may provide unexpected or even embarrassing results. Moreover, both methods have been found to be vulnerable in front of an aware WSE capable of performing attacks based on exploiting the semantics or the grammatical construction of the *fake* queries (Balsa, Troncoso, & Diaz, 2012).

In contrast, some other works have acknowledged the importance of distorting the user profiles to provide privacy but also ensuring the quality of the personalized search service. In order to achieve that, these proposals consider the semantic distance between *legitimate* and *fake* queries. Each *fake* query is synthetically generated according to the profile of the *legitimate* queries. The goal is to force the WSE to build a more or less accurate profile that preserves the privacy of the user while being still useful.

(Shapira, Elovici, Meshiach, & Kuflik, 2005) and (Sánchez, Castellà-Roca, & Viejo, 2013) are two proposals that try to preserve the utility of search queries. More concretely,

(Shapira et al., 2005) generates *fake* queries using a mix of terms from user queries and from a local database that contains general concepts related to the user's interests. Nevertheless, the authors do not specify how this database is obtained or how each term is selected. Regarding (Sánchez et al., 2013), this proposal first semantically interprets the contents of the legitimate queries and, after that, generates *fake* queries containing interests that are semantically similar to the *legitimate* ones. The maximum semantic distance between *fake* and *legitimate* interests can be fixed by the user and it is computed using knowledge bases like WordNet (Fellbaum, 1998) and the Open Directory Project (ODP) (Open Directory Project, 2013).

The main hypothesis of those works is that past queries effectively reflect the real interests of the user and, hence, they can be used to build suitable user profiles and to create utility-preserving (with regards to user interests) fake queries. However, we believe that this assumption does not always hold due to the following points that may introduce bias and noise to the resulting user profiles:

- Due to external and circumstantial needs (e.g., a student doing her homework), users may submit queries related to certain topics which are quite far from their real interests.
- Many users, specially those that are not very familiar with the internals of Web information retrieval or WSE query languages, may submit quite inaccurate queries to the WSE in order to retrieve more suitable suggestions (i.e., Query refinement (Shi & Yang, 2007)), or may be forced to reformulate their queries several times in order to retrieve more appropriate results (Lin & Xie, 2013). These tryouts or trial-and-error interactions add an undesirable bias to the generated profiles, which may artificially favor the topics associated to these recurrent queries.

- It might happen that several different users share the same computer, IP or web browser to submit their queries to the WSE. As a result, the profile that the WSE may build would merge the interests of all of them.

In addition to the above issues, from a linguistic perspective, queries may be quite variable and unstructured: they may contain from a single term to several ones and, in this last case (also called *complex queries* (Sánchez et al., 2013), queries quite usually correspond to schematic sentences or even lists of unconnected terms. The lack of a proper grammatical structure hampers naive analyses that are usually applied over queries to detect topics or interests, resulting in the rejection of queries (due to the impossibility of properly interpreting them) or in profile inaccuracies (due to the improper interpretation of query semantics) (Domingo-Ferrer et al., 2009; TrackMeNot, 2013; Xu, Zhang, Chen, & Wang, 2007). Due to the above difficulties, a large number of queries might be required to build a query-based user profile, and thus, to construct appropriate *fake* queries.

Considering the above points, we argue that the utility of privacy-preserving procedures with regard to the personalized services offered by WSEs (which rely on past queries to construct or evaluate user profiles), might be unsatisfactory. Therefore, we stress that other sources of data more suited to build user profiles and to ensure the utility of privacy-preserved queries should be investigated.

Contribution and plan of this paper

In this paper, we address the issues stressed above by proposing a new privacy-preserving scheme based on submitting *fake* queries to the WSE that contribute to hide concrete and confidential user information. A novelty of the proposed method comes from the fact that it generates the *fake* queries according to local profiles that are built from the social network

accounts of the users (e.g., Twitter) rather than from past queries or random-based approaches. Clearly, with this approach, our scheme is capable of providing a more satisfactory personalized service than a scheme that submits random *fake* queries. This is normal because such a scheme is very likely to generate partially random user profiles at the WSE side. Moreover, it has been proved that the random-based mechanisms can be easily detected by aware WSEs (Balsa et al., 2012).

Regarding the schemes that build profiles from past queries, the resulting local user profile generated by our proposal is expected to be more accurate than a query-based one due to the following reasons:

- The interests reflected by a user in her social network are assumed to be more stable and accurate than those related to web queries, which quite usually reflect circumstantial, partial or biased topics.
- In general, a social network account can be assumed to be uniquely managed by the owner herself. Hence, the possibility of having mixed interests from different individuals can be practically neglected.
- Even though messages posted on social networks may also be quite schematic (e.g., Twitter), they can be assumed to be more structured than Web queries and, hence, they are likely to offer a better interpretation of their semantics with regard to automatic profiling. In fact, some profilers for social networks such as Twitter can be found in the literature (A. Viejo, Sánchez, & Castellà-Roca, 2012a, 2012b), which have shown a good performance gathering the real interests of the users with a small amount of inputs. For example, empirical tests carried out in (A. Viejo et al., 2012a)

showed that less than 20 publications were needed to profile users with a clear predominant interest.

Hence, a profiler system based on publications in a social network is expected to build user profiles faster and more accurately than those based on Web queries.

The proposed method builds on the above premise in order to: (i) generate and submit *fake* queries to the WSE to help hiding concrete user information contained in her legitimate Web queries, and (ii) introduce a *positive* distortion in the user profile that a WSE may build from the queries received from, apparently, an individual (i.e., both legitimate and fake ones); the latter is possible thanks to the fact that *fake* queries are built according to the interests gathered from the local profile, which is built from the -reliable and accurate- user social network account. As a result, the adequacy of the personalized services offered by the WSE can be retained or even improved while hiding specific user information.

In other words, our main goal is to force the WSE to build an accurate user profile with regards to user *macro-interests* but hiding *micro-details* related to the user that may compromise her privacy. For example, using our proposal, a WSE will be able to know that a certain user is interested in "sports" but it will not be able to unequivocally discern her favorite football team (information that may help to disclose her approximate address, for example). This happens because the *legitimate* queries performed by the user (e.g., football player of her favorite team) are mixed with *fake* but plausible queries (e.g., other football players from other teams or even players of other sports) that are indistinguishable for the WSE. In this way, the proposed method exploits the fact that *macro-interests* can be useful to provide personalized web search services but they do not disclose enough information to represent a major privacy concern.

The proposed method incorporates several linguistic techniques to enable a coherent analysis of user publications in her social network account and of her queries submitted to the WSE, regardless their complexity and syntactical structure. Moreover, it exploits a knowledge base to provide a semantically-coherent interpretation of textual inputs, so that the utility of privacy-preserving queries can be guaranteed from a semantic point-of-view. Finally, it grounds on the basis on the Information Theory to properly quantify the informativeness of textual extractions and to build accurate and generalizable user profiles. Evaluations are performed using real data extracted from Twitter and real query logs obtained from the AOL files. The new scheme is analyzed in terms of *profile utility*, *user privacy* and *fake query feasibility* and it is compared with the other proposals available the literature.

The results in this paper are an extension of the preliminary research described in (A. Viejo & Sánchez, 2013). More specifically, the proposed method has been improved to exploit the Information Theory and the concept of Information Content (IC) to accurately quantify extracted evidences and to better characterize the user profiles; accordingly, the section that describes in detail the proposed scheme has been deeply rewritten and extended. New and more exhaustive evaluations have also been added to analyze the new method and to compare it with other proposals available in the literature. Indeed, the evaluation section is almost completely new. Finally, Introduction and Previous work sections have been thoroughly extended as well.

The rest of this paper is organized as follows. First, we present and formalize our proposal. Then we detail the evaluation of our approach on real data and compare its behavior with related works. Finally, we give some concluding remarks and propose some lines of future research.

Proposed method

The proposed method consists of three main modules: *construction of the user local profile*, *assessment of the WSE's user profile* and *creation of fake queries*. The first one profiles the user according to her publications in her social network account. Since this profiling process is based on evaluating unstructured textual messages, any social network providing this kind of (micro)-blogging user-related information (e.g., Twitter) is well-suited for this purpose. Regarding the second module, it assesses the user profile that is being built by the WSE according to the already performed *legitimate* queries. The last module compares the two profiles to evaluate which are the dominant and dominated topics in the user *local profile* and which should be the contents of the *fake* queries to be submitted to the WSE. Those fake queries are aimed to fulfill two requirements: (i) they should be plausible in order to add uncertainly and effectively hide the *legitimate* queries that may reveal user-specific information, and (ii) they should force the WSE to build a profile that approximates the *local profile* built from the user's social network account.

In the following, these modules are described in detail. After that, the workflow stating how the whole proposal works is detailed.

Constructing the local profile

In the literature, there are several automatic profiling techniques that can be applied to social environments such as social networks. As stated in (A. Viejo et al., 2012a), the most basic methods mainly rely on term occurrences/co-occurrences to quantify their contribution to the user profile. These approaches consider that all the terms are equally important in order to build a certain profile. Nevertheless, this assumption is not always true. In fact, it is coherent to consider some terms (e.g., iPhone) more informative than others (e.g., cell-phone) due to their

different degrees of specificity. In the same way, these approaches do not consider that some potential profiling categories (e.g., cell-phones) may be more specific than others (e.g., electronics). As a result, resulting profiles may be biased towards the most specific categories and terms and, thus, generate profiles that would be hardly generalizable and comparable against others (in this case, the one assumed by the WSE). In order to prevent this issue, the authors in (A. Viejo et al., 2012a) proposed a knowledge-based profiling approach grounded in the Information Theory that dynamically quantifies the amount of information provided by the terms contained in the publications of social applications such as social networks. The authors of that work showed that their method generates more general and accurate profiles.

According to that, in order to build the local profile, we rely on the profiling method presented in (A. Viejo et al., 2012a). This scheme is suited for any (micro)-blogging social application based on textual publications (e.g., Twitter). In order to accurately characterize user interests from textual entries, the profiling method relies on: (i) a set of *linguistic tools* to extract textual units with rich semantics from user messages; (ii) a knowledge-base to enable a *semantic interpretation* of the extracted units; and (iii) the foundations of the Information Theory to accurately quantify the *amount of information* that each unit is contributing to the user profile.

Let us consider that a *profile* is formalized as a set of well-defined categories $C = \{c_1, \dots, c_k\}$) (e.g., science, health, society, sports, etc), for which their relative weights (v_i) are computed according to the amount and informativeness of the extracted textual units for each category. As a result, the local profile of a certain user (LP) can be characterized according to the set of weighted categories obtained from the analysis of her publications in her social network account: $LP = \{< c_1, v_1 >, \dots, < c_k, v_k >\}$.

In order to calculate the corresponding weight v_i for each category c_i, three steps are performed:

1. *Extract semantic evidences of user interests*. The profiler focuses on extracting noun phrases (NPs) from the set of user publications, which correspond to pieces of text with rich semantic content that are commonly used in the literature to build profiles (Michelson & Macskassy, 2010; Zoltan & Johann, 2011). To coherently extract them, we rely on several natural language processing tools (OpenNLP Maxent Package, 2013) aimed at detecting sentences, identifying tokens (i.e., individual words), performing part-of-speech tagging of each word and, finally, syntactically parsing text to detect phrases and, particularly, noun-phrases. As a result, given a set of user publications in the context of her social network account, the set $\{< NP_1, w_1 >, \dots, < NP_p, v_p >\}$ is obtained, being w_i the number of appearances of each noun-phrase NP_i.
2. *Classify extractions*. The next step consists in semantically analyzing extractions in order to classify them in any of the profile categories. A knowledge base is used to link each NP with its conceptual abstraction and to discover the category to which it belongs. Specifically, we use the Open Directory Project (ODP) (Open Directory Project, 2013), which offers more than 1 million categories (including named entities and recently coined terms) organized in a taxonomic way. Moreover, directory data files can be downloaded in SQL format and categories can be efficiently obtained off-line. To semantically classify NPs, the system looks for each extracted NP_i in ODP. If found, ODP provides the hierarchy H_i of categories ($H_i = h_{i,1} \rightarrow \cdots \rightarrow h_{i,l}$) to which NP_i belongs. For example, if the system looks for *"MacBook"*, ODP provides the following hierarchy: *MacBook* → *Portables* → *Hardware* → *Macintosh* → *Apple* →

Systems → *Computers*. In order to improve the retrieval recall, stemming is applied to detect equivalent morphological constructions (e.g., “cell phones” = “cell phone”) and to omit punctuation marks. Moreover, since NPs may incorporate circumstantial qualifiers (e.g. ``new MacBook'') which would be unlikely covered by these kind of general knowledge structures, if the NP_i is not found “as is” in ODP, we iteratively look for simpler forms obtained by removing nouns or adjectives on the left side of the phrase (e.g. “new MacBook” → ”MacBook”). In this manner, the recall of the semantic classification can be improved while retaining the core semantics of the extraction (Vicient, Sánchez, & Moreno, 2013). If after all this process, the NP_i is still not found, it is discarded because no semantics can be incorporated to the user profile.

3. *Update LP category weights*. In this last stage, for each NP_i , the system evaluates its degree of informativeness with regard to the profile category to which it belongs (according to the hierarchy H_i obtained from ODP).The system checks if any of the profile categories c_i is included in H_i. If this is the case, the contribution of NP_i is added to c_i by measuring the *amount of information* that NP_i provides and also considering its number of repetitions (i.e., w_i). In this manner, frequently appearing NPs and/or highly informative ones will have a larger influence in the user profile (i.e., category weights). The fact that extractions are evaluated according to their informativeness rather than their absolute number of occurrences (as done in some works such as (Abel, Gao, Houben, & Tao, 2011; Ebner et al., 2010; Zoltan & Johann, 2011)) tends to build more accurate profile characterizations (A. Viejo et al., 2012a). To measure the informativeness of NPs, we rely on the notion of the

Information Content of a NP, which is computed as the inverse of its probability of occurrence in a corpus (Ross, 1976).

$$IC(NP_i) = -\log\ p(NP_i)$$

In this manner, specific terms that appear more rarely in corpora are considered more informative than commonly used ones. IC requires from robust probabilities compiled from large and heterogeneous sources. To do so, we rely again on ODP since it indexes and classifies around 5 millions web documents that can be used as corpora. Specifically, the web count provided by ODP for a given term is used to compute the probability of NP_i (where *total_webs* stands for the total amount of web sites indexed in ODP):

$$p(NP_i) = \frac{web\_count(NP_i)}{total\_webs}$$

Then, we compute the contribution φ of each NP_i to the profile category c_j to which it belongs (according to the ODP hierarchy H_i) as the product of its number of repetitions w_i and its IC. This contribution is added to the weight of the corresponding category c_j for each NP_i.

$$v_j = \sum_{\forall\ NP_i\ is\text{-}a\ c_j} (w_i\ \times\ IC(NP_i))$$

Once all user messages are analyzed, her *local profile* is defined as a ranked list of categories according to their computed weights. In order to make this profile characterization general and more independent on the predefined list of categories (which may vary from one profiler to another), we normalize category weights according to the informativeness of each category. This is motivated by the fact that categories may have different degrees of generality (i.e., Information Content). Hence, for general categories (e.g., society) the chance to extract a

NP belonging to it in an uniformly distributed text would be higher than for a more specific one (e.g., computers). To compensate these imbalances and to make the profile characterization independent on the concrete list of categories, we scale each category weight (v_j) according to the IC of the category itself ($IC(c_j)$):

$$v_j = v_j \times IC(c_j)$$

Hence, if general categories are included in the profile, a higher number of more informative extractions will be needed to achieve the same weight in comparison with other more concrete categories.

Assessment of WSE's profile

Once the *local profile* of the user is built, the purpose of our method is to generate *fake* queries and submit them to the WSE. These *fake* queries together with the *legitimate* ones performed by the user herself will force the WSE to build a user profile similar to *LP*. Note that, as explained previously, we assume that *LP* is a realistic representation of the interests of the user. Moreover, since *fake* queries will contain plausible contents, they will effectively hide the real *micro-details* of the user by adding a degree of uncertainty caused by the fact that *legitimate* and *fake* queries are indistinguishable by the WSE. In this way, we retain the utility of the query logs generated by the WSE, so that the latter will be able to provide suitable personalized services, while improving the privacy of user queries.

To do so, in addition to the *local profile* (*LP*) already built, we require a representation of the user profile that, presumably, is being built by the WSE. Let us name it as *public profile* (*PP*). This profile is created according to the queries submitted to the WSE in a similar way that the WSE is assumed to build its own user profile according to the queries received from a user/IP. The idea is that, by comparing *LP* and *PP* distribution of category weights, we can

detect the differences between the profile that is being built by the WSE (which may be hampered by the issues discussed in the previous work section), and the accurate user profile built from her social network account. As a result, new *fake* queries can be constructed and submitted to the WSE on behalf of the user so that *PP* can be *positively distorted* towards the distribution of category weights of *LP*, while contributing to hide the *legitimate* user queries. For example, if the user's *LP* shows a dominance of *computer* and *sport* categories, but user queries have only covered sport-related topics, our system may create computer-related *fake* queries such as "data recovery" or "MacBook" to be sent the WSE on behalf of the user.

PP is defined in the same manner as *LP*. Therefore, we consider that $PP = \{< c_1, v_1 >, \dots, < c_k, v_k >\}$. The difference is that, on the contrary to *LP*, which represents a more or less "static" profile, *PP* is dynamically built as new queries (*legitimate* or *fake* ones) are submitted to the WSE. Hence, at the beginning of its execution, *PP* is initialized as $PP = \{< c_1, 0 >, \dots, < c_k, 0 >\}$.

Then, to calculate the corresponding weight v_i for each category c_i in *PP*, we apply the same profiling that has been already used to build *LP*. Nevertheless, in this case, the inputs are web queries instead of user publications in her social network. Therefore, each time a query is submitted to the WSE, the profiler analyzes the query, extracts its NPs and semantically matches those to the profile categories, whose weights are updated according to the informativeness of each NP. Thanks to the application of the linguistic analysis, query semantics can be more accurately interpreted than in most related works that focus on query analysis (Domingo-Ferrer et al., 2009; TrackMeNot, 2013; Xu et al., 2007). The latter either omit *complex queries* (i.e., queries consisting on several words) that cannot be found "as is" in the knowledge base,

neglecting a number of semantic evidences (Sánchez et al., 2013), or evaluate query words independently, which may result in semantic incoherencies (Sánchez et al., 2013).

Note that, in our proposal, both *legitimate* and *fake* queries (these are the queries generated in this step) contribute to build and update the *PP*. The final category weights are also scaled according to the informativeness of each category. The information-based evaluation of queries and categories contributes to produce more general characterizations of user profiles, with independence on the specific set of profile categories. Moreover, it has been proven to produce more accurate profiles from a lower amount of evidences than other approaches based on counting category occurrences (A. Viejo et al., 2012a).

Creation of fake queries

Next, we focus on the steps required to generate *fake* queries given: (i) the "ideal" *LP*; and (ii) the current *PP* built from the queries already submitted to the WSE up to this moment:

1. *Normalize category weights for LP and PP*. Even though both profiles *LP* and *PP* are defined according to the same set of categories, in order to properly compare them, category weights should be normalized into a common scale. To perform this normalization, we translate absolute category weights to percentages λ_i that reflect the relative level of dominance of each category c_i in a certain profile. This is done by dividing the weight of each category by the sum of weights of all categories. Formally, let us define LP' and PP' as the normalized versions of *LP* and *PP*, respectively. According to that, LP' and PP' are defined as $\{< c_1, \lambda_1 >, \ldots, < c_k, \lambda_k >\}$, where:

$$\lambda_i = \left(v_i / \sum_{j=1}^{k} v_j \right) \times 100$$

2. *Select the category of the new fake query.* As discussed above, since LP' is assumed to be the *ideal* user profile, it should reflect the *optimal weight* for each category with regard to the user interests. On the other hand, PP' is assumed to reflect the current weight disclosed to the WSE. Therefore, our goal is to force PP' to approximate LP' by submitting new *fake* queries. The topic to which the new *fake* query should be related to is decided by detecting which category from PP' is currently the most distant from its *optimal weight* that is stated by LP'. To do that, for each category c_i, the system computes the difference between its optimal weight and its current weight. Let us denote this as:

$$\Delta_{c_i} = (\lambda_i; LP') - (\lambda_i; PP')$$

This difference (expressed as a relative percentage of category dominance) is a direct function of the amount of information that is needed to balance each category in PP' towards LP'. Positive Δ values indicate that the weight of c_i in PP' is lower than optimal and, hence, additional information for this category c_i should be added to PP (i.e., additional queries related to c_i should be submitted to the WSE). On the contrary, negative values state that c_i weighs too much in PP' and, hence, queries related to other non-overweight categories should be submitted. Finally, values close to zero correspond to categories that are well matched between global and local profiles. Our goal here is to achieve this close to zero Δ values for *all* categories. In order to correct category imbalances, we select the category c_{max} with the highest positive Δ value, since it is the one requiring to add the largest amount of *fake* queries.

$$c_{max} = argmax_{\forall c_i \in PP'}(\Delta_{c_i})$$

3. *Construct a new fake query*. At this point, the system creates a new plausible query (according to user's LP'). To do that, it exploits ODP again to retrieve a random term that is a specialization of the profile category that should be balanced first (i.e., c_{max}). This term, which may be either a single word (e.g., "MacBook'"), a noun phrase (e.g., "digital portable player") or an expression (e.g., "technical evaluation and product reviews"), constitutes the new *fake* query to be submitted to the WSE on behalf of the user. On the one hand, due to the fact that this term is related to the category to be balanced, it is assumed to contribute approximating *PP* to *LP*. On the other hand, its semantic coherence (with regards to LP' and ODP hierarchies) and its randomness are assumed to add uncertainty that helps hiding the user *micro-details* provided by her *legitimate* queries among *fake* but plausible queries.

System's workflow

This section details how the above-described modules are put together and how the system operates as new queries are submitted by the user. Input parameters are: (i) the user account from which the *LP* shall be built; and (ii) a numerical value *t* stating the number of *fake* queries to be created and submitted for each *legitimate* one. The last parameter controls the system behavior and configures the expected degree of privacy/profile balancing.

1. The first step consists in *constructing the local profile*, which generates *LP*. As discussed above, *LP* is considered to be a quite *static* profile that could remain unaltered for long periods of time. However, if changes are introduced in the user's social network account that may significantly affect the characterization of user interests, *LP* could be updated online at any point of the system's life cycle by requesting the execution of the profile construction module. The decision to update

the profile is left for the final user; however, this process is computationally lightweight and, hence, it is possible to automatically perform it after each modification introduced in the social network. Regarding the number of publications in the social network that are required to build an accurate local profile, the authors in (A. Viejo et al., 2012a) state that a value between 10 and 20 publications is enough for users with clearly defined interests. In the case of heterogeneous users, up to 40-50 publications might be required. In any case, these values are suggested to get a very accurate local profile. If the users have less publications, the resulting local profile will be less representative but still usable. Moreover, thanks to the on-line behavior of the proposed scheme, profiles can be dynamically improved as new publications are added.

2. The system waits for the user to submit a *legitimate* query to the WSE. As a result, *PP* is updated with the contents of her query.
3. Then, the system executes the *creation of a fake query* module to build the first *fake* query (of *t*). After that, *PP* is updated according to the informativeness of the contents of that new query, so that the second *fake* query (if $t > 1$) will be created according to an up-to-date characterization of *PP* that considers all legitimate and *fake* queries up to that moment. The process is repeated until *t fake* queries have been created. The higher *t* is, the more *fake* queries will be created and, hence, the faster *PP* would likely converge to *LP*. The influence of this parameter in the system behavior and in the profile adaptation will be tested in evaluation section.
4. The system goes back to step-2 and waits for a new *legitimate* query.

It is important to note that the creation of the *t fake* queries per each *legitimate* one is always executed regardless the fact that *PP* has effectively converged to *LP* or not. In this manner the same degree of privacy can be guaranteed through the whole system's life cycle. Notice that, if *PP* has converged to *LP*, *fake* queries will tend to follow the category distribution of *LP*, so that *PP* will remain the same as *LP*.

The above process addresses how to create privacy-preserving but useful *fake* queries but it is not linked to any particular protocol for actually submitting them to the WSE. In fact, *fake* queries should not be submitted as they are created, since this may produce a repetitive pattern in query logs (i.e., each *legitimate* query is immediately followed by *t fake* ones) that would be easy to detect. Therefore, a protocol developed to mimic human querying behavior is assumed to be the best option to submit the *fake* queries generated by our new method. Nevertheless, it is worth to mention that, given its general design, our method can be applied to any other already existent system designed for that purpose, independently of its *fake query-submission* strategy.

Evaluation

In this section we evaluate the performance of the proposed system in terms of *profile utility*, *user privacy* and *feasibility of fake queries*. The first aspect is related to the performance of the system in approximating *PP* to *LP*. The second aspect reflects the amount of *legitimate* information that is revealed to the WSE considering both *legitimate* and *fake* queries submitted by the system. Finally, the third aspect can be understood as the difficulty to distinguish between queries performed by the user and *fake* ones created by the proposed scheme.

By profiling categories in *C* we defined eight well-differentiated general ones corresponding to root categories in the ODP hierarchy with the minimum overlap. These are:

Arts, *Health*, *Shopping*, *Science*, *Computers*, *Sports*, *Society* and *Business*. This configuration has been previously used in (A. Viejo et al., 2012b).

As test data, we considered a set ψ of user search queries that were used as *legitimate queries* in our evaluations. More specifically, the set ψ contained 100 queries randomly selected from the query logs taken from real users and compiled by AOL during 3 months in 2006 (AOL Search Data Mirrors, 2006). Queries in ψ were profiled using the method introduced above. Using ψ as input, the corresponding normalized weights related to each considered category are shown in Table 1. This represent the PP' that the WSE would build if a certain user submits the whole set ψ (and *only* this set).

Insert Table 1 here

Table 1 shows that the most representative category in the AOL dataset ψ is *Arts*, while the least representative one is *Science*.

In order to evaluate the performance of our scheme in well-differentiated scenarios, we configured the best and worst settings with regard to the set of queries: two Twitter users whose normalized local profiles LP' presented a very predominant focus on *Arts* (best setting, in which the convergence between the real user interests match with those inferred from her web queries) and another two users whose main category is *Science* (worst setting, in which there is a large divergence between the profile and the actual web queries). The first two users correspond to @johnmaeda (President of the Rhode Island School of Design) and @MuseumModernArt (Art Museum in New York City). The last two users are @ReutersScience (the Science Team of Reuters.com) and @CERN (the European Organization for Nuclear Research). These concrete

users were selected by looking at the *WhoToFollow search engine* (Twitter - WhoToFollow, 2013) provided by Twitter that offers a list of the most relevant Twitter users according to each specific topic. From each Twitter user, 100 tweets were extracted and profiled using the method introduced above. The normalized local profiles LP' built for each user are shown in Table 2.

Insert Table 2 here

On one hand, users @johnmaeda and @ReutersScience simulate the two extremes of the possible situations faced by our system (i.e., @johnmaeda's profile is very close to the profile represented by the set of AOL queries and @ReutersScience provides a local profile that largely diverges from that set of queries). On the other hand, @MuseumModernArt's profile is relative close to @johnmaeda but it has much more weight in the *Arts* category. Also, @CERN provides a profile that is even more extreme than the one gathered from @ReutersScience. As a result, other potential Twitter users (and their profiles) are expected to provide results in between of those reported in this section. Since the analysis of these intermediate situations would not add new insights on the system's performance, for clarity and conciseness, we did not include them in this evaluation report.

Profile utility evaluation

This subsection evaluates the performance of the proposed system in approximating *PP* to *LP*, that is, in making the profile gathered by the WSE more useful and suited for personalized web search services. As previously explained, *LP* is assumed to reflect the optimal weight for each category while *PP* is assumed to reflect the current disclosed weights in front of the WSE.

Considering that PP' and LP' are defined as $\{< c_1, \lambda_1 >, \ldots, < c_k, \lambda_k >\}$ where λ_i is a percentage that represents the corresponding weight, in order to numerically quantify the divergence between both profiles after each submitted query (i.e., $D(LP', PP')$), we compute their distance as follows:

$$D(LP', PP') = \frac{\sum_{i=1}^{k} |(\lambda_i; LP') - (\lambda_i; PP')|}{k}$$

When D is close to 0, it means that there is no difference between LP' and PP' (i.e., the optimal situation). On the other hand, the larger D is, the bigger the difference between LP' and PP' becomes.

The proposed method is compared with the three different approaches:

1. The four tested users (@johnmaeda, @ReutersScience, @MuseumModernArt and @CERN) submit only *legitimate* queries (i.e., no *fake* queries are added to ψ).
2. The four users use a naive privacy-preserving method that adds *t* random *fake* queries extracted from ODP per each *legitimate* user query. This strategy tries to emulate any scheme in the literature that hides *legitimate* queries among random queries. For example, (Domingo-Ferrer et al., 2009) and (TrackMeNot, 2013).
3. The four users use a scheme that builds *LP* using past *legitimate* queries. More specifically, the proposal presented in (Sánchez et al., 2013) is used in this scenario. As explained in the previous work section, (Sánchez et al., 2013) semantically interprets the contents of *legitimate* past queries and generates *t fake* queries containing interests close to the *legitimate* ones. The maximum distance *d* between *fake* and *legitimate* interests can be fixed by the user. More specifically, this parameter corresponds to the length of the taxonomic path connecting the *legitimate*

and *fake* interests in a knowledge base such as ODP or Wordnet. As suggested in (Sánchez et al., 2013), in our simulations, we have used $d = 3$.

The last two approaches and also the proposed scheme have been also tested for *t* values (i.e., number of *fake* queries) among 1 and 8. Obviously, large values of *t* allow the system to modify PP' faster. Also, from the privacy point of view, it is better to use a large *t*, because, in a scenario where *legitimate* and *fake* queries are indistinguishable, the probability of randomly guessing the appropriate query is $1/t + 1$. Nevertheless, a large *t* also implies more bandwidth overhead.

Figures 1 and 2 show, for @johnmaeda and @ReutersScience respectively, how the distance between LP' and PP' evolves after submitting each *legitimate* query and the corresponding set of *t fake* queries.

Insert Figure 1 here

Insert Figure 2 here

First, it should be noted that the reported figures are statistically significant. Indeed, the *p-value* of the average correlation between the local and public profiles computed after each new query was below 0.05 in all cases. This suggests that reported figures are highly unlikely to be the result of a random chance.

In general, it can be clearly seen that our proposal achieves the best matching between profiles in both scenarios. This was expected since our scheme is specifically designed to approximate PP' to LP' while the other tested methods are not. The random approach and

(Sánchez et al., 2013) perform in a very similar way in the evaluations related to @johnmaeda due to the fact that, in this case, PP' and LP' are quite analogous and these two proposals do not significantly alter the distribution of PP'. Regarding @ReutersScience, in this case, PP' is really far away from LP'; therefore, (Sánchez et al., 2013) obtains worse results than its random counterpart because the former generates fake queries related to PP' and, hence, far from LP'. On the other hand, the random method introduces random noise that uniformly affects all the categories and, hence, the long distance between PP' and LP' becomes a certain advantage.

Analyzing the specific results for each user, @johnmaeda has a LP' closer to the interests reflected by the set ψ; Therefore, (Sánchez et al., 2013) and the random approach get a distance around 5% for $t = 8$. However, even in this favorable situation for both systems, our scheme, for all the tested *t* values, obtains better results. It is worth to mention that our system gets a distance around 3% for $t = 1$ and a result very close to 0% for $t = 8$.

In the case of @ReutersScience, both (Sánchez et al., 2013) and the random approach show their weaknesses and the distance grows to a value around 11% for (Sánchez et al., 2013) and 8% for the random method. Regarding our proposal, it clearly outperforms them even when our system generates only one *fake* query per each *legitimate* one ($t = 1$) and the other tested schemes generate eight *fake* queries ($t = 8$). It is interesting to see that, with our scheme working with $t = 8$, the distance between profiles becomes almost 0% at the 20th *legitimate* query submitted to the WSE.

Insert Figure 3 here

Insert Figure 4 here

Regarding the users @MuseumModernArt and @CERN, Figures 3 and 4 respectively, provide some evaluation results that strengthen the conclusions stated above. More specifically, in the case of @MuseumModernArt all the schemes behave worse than in the case of @johnmaeda but they keep the proportionality and, hence, our proposal obtains the best results. Note that, even being similar profiles, @MuseumModernArt has more weight on *Arts* than @johnmaeda (45.7 % and 27.7 % respectively), this fact makes harder to reduce the distance between LP' and PP'. A similar behavior can be seen for @CERN, due to the fact that this user has a very extreme profile with a lot of weight on *Science* (49.9 %).

Finally, it should be mentioned that a realistic situation is expected to be more close to the @johnmaeda scenario than to the other tested users. The reason is that, in general, users are expected to submit queries close to their local profiles. However, in any case, our scheme has shown that it can effectively approximate PP' to a given LP' in any situation.

User privacy evaluation

User privacy evaluations considering the same set of 100 *legitimate* queries have been also performed with the different approaches detailed in the previous section: our proposal, the random approach and (Sánchez et al., 2013) (with $d = 3$) to generate *fake* queries considering different values for *t*.

The privacy level achieved by each scheme is measured by computing the percentage of *legitimate* micro-interests (i.e., the detailed interests of the user extracted from *legitimate* queries) that can be found also in the whole set of queries submitted to the WSE (this includes both *legitimate* and *fake* queries). On the contrary to macro-interests, which correspond to the main categories considered in the user profiles, micro-interests are the most concrete categories

extracted from ODP that correspond to each of the submitted queries. For example, the query "iphone 5" provides the following hierarchy from ODP: *Computers* → *Systems* → *Handhelds* → *Smartphones* → *iPhone*. Here, the macro-interest (which provides profile utility) is *Computers*, while the micro-interest (which might represent a privacy threat) is *iPhone*.

Insert Figure 5 here

Insert Figure 6 here

Figures 5 and 6 show that, as expected, the most relevant factor from the privacy protection point of view is *t* (the number of fake queries): as this value grows, the percentage of user information that is disclosed decreases.

The random method adds completely random micro-interests to the user profile built by the WSE and, hence, this is the optimal approach from the privacy point of view. As a consequence, it is very interesting to see that this optimal approach gets only slightly better privacy results than our proposal (they are practically equivalent in all the cases). It is worth to mention that, as shown in the previous subsection, the addition of random queries is good for the privacy but generate quite inaccurate profiles from the utility perspective. On the other hand, our proposal achieves almost the same privacy as the random technique while providing much more accurate profiles.

Regarding (Sánchez et al., 2013), as explained previously, this method generates *fake* queries related to the interests of past queries and, due to the fact that the fake} queries will be quite similar to the *legitimate* ones, this scheme is less effective from the privacy point of view

than the other two proposals. Note that the results shown for (Sánchez et al., 2013) are achieved with $d = 3$; For $d < 3$, the micro-interests of legitimate and fake queries are expected to be even closer and, hence, the privacy results are assumed to be worse.

Fake query feasibility evaluation

The use of *fake* queries is a well-known approach to preserve the privacy of the users. However, to be really effective, these *fake* queries must be indistinguishable from the *legitimate* ones. If this is not the case, an aware WSE may be able to detect and discard them. Therefore, in this section we evaluate the feasibility of the generated queries considering that a *feasible* query is a query that *looks* authentic.

In order to quantify the feasibility of the fake queries generated by the proposed system, we computed the difference of Information Content (IC) between each legitimate query (the 100 legitimate queries) and the new fake queries built by our approach for the largest test: $t = 8$, which represents 800 fake queries. The idea is that, if a legitimate query has a certain degree of concreteness (represented by its IC), in order to be feasible (or realistic) and hardly distinguishable from a legitimate one, each fake query should maintain, as much as possible, the informativeness of the legitimate query, thus presenting a similar IC value. This approach was previously used in (Sánchez et al., 2013).

To compute the IC of legitimate and fake queries in an objective manner, the *web count* provided by the *Bing Web Search Engine* when querying them was used. More specifically, the IC of a query *a* can be computed from the Web as follows:

$$IC(a) = -\log(p(a)) = -\log\left(\frac{web\_count(a)}{total\_webs}\right)$$

Due to the fact that *total_webs* is a common factor to all the evaluated queries and $-\log()$ is a monotonic function that does not alter the relative order between queries, they can

be dropped from the equation (Turney, 2001). As a result, the IC of a term can be directly estimated using *web_count*.

Figure 7 shows the histogram of *web_count* differences, grouping queries in ranges of orders of magnitude. Our scheme is compared with a standard method that generates completely random *fake* queries. This method does not apply any syntactic analysis on the queries and it only substitutes each term from the legitimate query for a random term.

Insert Figure 7 here

The degree of IC preservation between legitimate and fake queries is evaluated as the difference (in orders of magnitude) between their *web_count* values. Notice that, due to the enormous size of the Web, *web_count* may range from a few dozens to several billions. In consequence, the X-axis of Figure 7 groups queries following a logarithmic scale. Y-axis shows the number of queries for each group. *Fake* queries that obtain a similar *web_count* to their *legitimate* counterparts are assumed to *look* feasible (or authentic). Therefore, the ideal situation is to put almost all the generated fake queries in the lowest intervals (i.e., [1 ... 100[ ).

Evaluation results show a clear difference between our semantically-grounded proposal and the random approach, which neglects syntactic and semantic analysis. More specifically, our method creates *fake* queries that, in most cases, differ among 1-100 times with respect to the original query *web_count*. Considering the size of the Web and the range in which *web_count* may vary, this is a quite constrained result.

In contrast, the random method generates *fake* queries that, in most cases, differ more than 100,000 times. This is normal because a *fake* query which is build like a bag-of-words will

probably lack any coherence and will obtain very few *web_counts* (or even none) from a WSE. These results show that, for an aware WSE, it is easier to detect the *fake* queries generated by randomly than those built by our proposal.

Conclusions and future work

In this paper, we have proposed a new scheme that *positively* distorts user profiles by submitting *fake* queries to the WSE with the aim of providing improved privacy. Our system locally profiles users according to their social network accounts (e.g., Twitter) and relies on the notion of Information Content (IC) to accurately quantify the extracted semantic evidences.

We argue that, using this strategy, our scheme enables a better profile than other schemes in the literature that rely on past queries (e.g., (Shapira et al., 2005) and (Sánchez et al., 2013)) or proposals that simply submit random *fake* queries (e.g., (Howe & Nissenbaum, 2009) and (Domingo-Ferrer et al., 2009)). More specifically, the interests reflected by a user in her social network are assumed to be more accurate and stable than those that can be inferred from past search queries, which may be more circumstantial. Regarding random-based methods, they produce noisy profiles that clearly disrupt any possible personalized service. Worse than that, this kind of schemes might introduce fake contents completely unexpected and even uncomfortable for the users.

In terms of privacy, our proposal is based on building *fake* queries that keep the *macro-interests* of the user (which provide usability in front of the WSE) while contributing to hide or add ambiguity to her *micro-interests* (which may disclose specific confidential information).

The evaluation results show that: (i) the proposed system effectively approximates *PP* to *LP*; (ii) a small number of *legitimate* queries are required to achieve this (between 20 and 25 queries are enough); (iii) it is clearly more effective than systems based on submitting random

fake queries and schemes that rely on past queries to perform the profiling process; and (iv) it achieves almost the same privacy level as the random method (which is considered the optimal approach from the privacy point of view) while providing much more accurate profiles.

As future work, it would be interesting to study the suitability of additional sources of social data to enhance the accuracy of the local profiles built by our proposal. More specifically, the use of multi-layered/multiplex social networks (Jung, Juszczyszyn, & Nguyen, 2007), that is, the exploitation of profiles gathered from the several accounts of the same user in different social networks (e.g. Twitter and Facebook), may significantly improve this aspect.

Moreover, it is worth to mention that, currently, the *public profile* (*PP*) used by our proposal is created in the way that the WSE is *assumed* to build its own user profile, which is assumed to follow the usual premises found in most of the profiling literature (A. Viejo et al., 2012a). In any case, since it is very unlikely that the WSEs will provide the concrete details about their profiling schemes, it can be interesting to analyze how different profiling strategies affect the search results generated by different WSEs and, hence, try to ascertain to a certain extent, the concrete details on how each WSE build its profiles.

Disclaimer and acknowledgements

Authors are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization. This work was partly supported by the European Commission under FP7 project Inter-Trust, by the Spanish Ministry of Science and Innovation (through projects eAEGIS TSI2007-65406-C03-01, CO-PRIVACY TIN2011-27076-C03-01, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004, Audit Transparency Voting Process IPT-430000-2010-31, ICWT TIN2012-32757 and BallotNext IPT-2012-0603-430000) and by the Government of Catalonia (under grant 2009 SGR 1135).

References

Abel, F., Gao, Q., Houben, G.-J., & Tao, K. (2011). *Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web.* Paper presented at the Proceedings of the 8th Extended Semantic Web Conference on The Semanic Web: Research and Applications (ESWC'11).

AOL Search Data Mirrors. (2006). *http://gregsadetsky.com/aol-data/(last accessed: 27/05/2013)*

Balsa, E., Troncoso, C., & Diaz, C. (2012). *OB-PWS: Obfuscation-Based Private Web Search.* Paper presented at the Proceedings of the IEEE Symposium on Security and Privacy (SP'12).

Barbaro, M., & Zeller, T. (2006). A Face Is Exposed for AOL Searcher No. 4417749. *The New York Times*.

Castellà-Roca, J., Viejo, A., & Herrera-Joancomartí, J. (2009). Preserving user's privacy in web search engines. *Computer Communications, 32*(13-14), 1541-1551.

Collusion mozilla add-on. (2013). *http://www.mozilla.org/en-US/collusion (last accessed: 27/05/2013)*

Cooper, A. (2008). A survey of query log privacy-enhancing techniques from a policy perspective. *ACM Transactions on the Web, 2*(4), 1-27.

Dingledine, R., Mathewson, N., & Syverson, P. F. (2004). *Tor: The second-generation onion router.* Paper presented at the Proceedings of the 13th USENIX Security Symposium (USENIX'04).

Domingo-Ferrer, J., Solanas, A., & Castellà-Roca, J. (2009). h(k)-Private Information Retrieval from Privacy-Uncooperative Queryable Databases. *Journal of Online Information Review, 33*(4), 1468-4527.

Ebner, M., Mühlburger, H., Schaffert, S., Schiefner, M., Reinhardt, W., & Wheeler, S. (2010). Getting Granular on Twitter: Tweets from a Conference and Their Limited Usefulness for Non-participants. *Key Competencies in the Knowledge Society, 324*, 102-113.

Eckersley, P. (2010). *How unique is your web browser?* Paper presented at the Proceedings of the 10th International Conference on Privacy Enhancing Technologies (PETS'10).

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*: MIT Press.

Google History. (2013). *http://www.google.com/history (last accessed: 27/05/2013)*

Hansell, S. (2006). Increasingly, Internet's Data Trail Leads to Court. *The New York Times*.

Howe, D. C., & Nissenbaum, H. (2009). TrackMeNot: Resisting surveillance in web search. *Lessons from the Identity Trail: Anonymity, Privacy, and Identity in a Networked Society, 23*, 417-436.

Jones, R., Kumar, R., Pang, B., & Tomkins, A. (2008). *Vanity fair: Privacy in query log bundles.* Paper presented at the Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM'08).

Jung, J. J., Juszczyszyn, K., & Nguyen, N. T. (2007). Centrality measurement on semantically multiplex social networks: divide-and-conquer approach. *International Journal of Intelligent Information and Database Systems, 1*(3/4), 277-292.

Lin, S., & Xie, I. (2013). Behavioral Changes in Transmuting Multisession Successive Searches Over the Web. *Journal of the American Society for Information Science and Technology, 64*(6), 1259-1283.

Makris, C., Plegas, Y., & Stamou, S. (2012). Web Query Disambiguation Using PageRank. *Journal of the American Society for Information Science and Technology, 63*(8), 1581-1592.

Michelson, M., & Macskassy, S. A. (2010). *Discovering Users' Topics of Interest on Twitter: a First Look.* Paper presented at the Proceedings of the 4th Workshop on Analytics for Noisy Unstructured Text Data.

Open Directory Project. (2013). *http://www.dmoz.org (last accessed: 27/05/2013)*

OpenNLP Maxent Package. (2013). *http://maxent.sourceforge.net/about.html (last accessed: 27/05/2013)*

Ortiz-Cordova, A., & Jansen, B. J. (2012). Classifying Web Search Queries to Identify High Revenue Generating Customers. *Journal of the American Society for Information Science and Technology, 63*(7), 1426-1441.

Romero-Tris, C., Viejo, A., & Castellà-Roca, J. (2011). *Improving query delay in private web search.* Paper presented at the Proceedings of the International Workshop on Securing Information in Distributed Environments and Ubiquitous Systems (SIDEUS'11).

Ross, S. (1976). *A First Course in Probability*: Macmillan.

Sánchez, D., Castellà-Roca, J., & Viejo, A. (2013). Knowledge-Based Scheme to Create Privacy-Preserving but Semantically-Related Queries for Web Search Engines. *Information Sciences, 218,* 17-30.

Shapira, B., Elovici, Y., Meshiach, A., & Kuflik, T. (2005). PRAW - A PRivAcy model for the Web. *Journal of the American Society for Information Science and Technology, 56*(2), 159-172.

Shapira, B., & Zabar, B. (2011). Personalized Search: Integrating Collaboration and Social Networks. *Journal of the American Society for Information Science and Technology, 62*(1), 146-160.

Shi, X., & Yang, C. C. (2007). Mining Related Queries from Web Search Engine Query Logs Using an Improved Association Rule Mining Model. *Journal of the American Society for Information Science and Technology, 58*(12), 1871-1883.

Summers, N. (2009). Walking the Cyberbeat. *Newsweek.*

TrackMeNot. (2013). *http://mrl.nyu.edu1 dhowe/trackmenot (last accessed: 27/05/2013)*

Turney, P. D. (2001). *Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL.* Paper presented at the Proceedings of the 12th European Conference on Machine Learning (ECML'01).

Twitter - WhoToFollow. (2013). *http://twitter.com/\#!/who_to_follow (last accessed: 27/05/2013)*

Vicient, C., Sánchez, D., & Moreno, A. (2013). An automatic approach for ontology-based feature extraction from heterogenous textual resources. *Engineering Applications of Artificial Intelligence, 26*(3), 1092-1106.

Viejo, A., & Castellà-Roca, J. (2010). Using Social Networks to Distort Users' Profiles Generated by Web Search Engines. *Computer Networks, 54*(9), 1343-1357.

Viejo, A., & Sánchez, D. (2013). *Providing Useful and Private Web Search by Means of Social Network Profiling.* Paper presented at the Proceedings of the 11th Annual Conference on Privacy, Security and Trust (PST'13).

Viejo, A., Sánchez, D., & Castellà-Roca, J. (2012a). Preventing Automatic User Profiling in Web 2.0 Applications. *Knowledge-Based Systems, 36,* 191-205.

Viejo, A., Sánchez, D., & Castellà-Roca, J. (2012b). *Using Profiling Techniques to Protect the User Privacy in Twitter.* Paper presented at the Proceedings of the 9th International Conference on Modeling Decisions for Artificial Intelligence (MDAI'12).

Xu, Y., Zhang, B., Chen, Z., & Wang, K. (2007). *Privacy-Enhancing Personalized Web Search.* Paper presented at the Proceedings of the 16th International Conference on World Wide Web.

Ye, S., Wu, F., Pandey, R., & Chen, H. (2008). *Noise Injection for Search Privacy Protection* (Technical Report): University of California.

Zoltan, K., & Johann, S. (2011). *Semantic Analysis of Microposts for Efficient People to People Interactions.* Paper presented at the Proceedings of the Roedunet International Conference (RoEduNet'11).