Knowledge-Based Scheme to Create Privacy-Preserving but Semantically-Related Queries for Web Search Engines

David Sánchez, Jordi Castellà-Roca, Alexandre Viejo

Departament d'Enginyeria Informàtica i Matemàtiques,
UNESCO Chair in Data Privacy, Universitat Rovira i Virgili,
Av. Països Catalans 26, E-43007 Tarragona, Spain
E-mail: {david.sanchez,jordi.castella,alexandre.viejo}@urv.cat

Abstract

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

Key words: Complex queries, Ontologies, Privacy, Private information retrieval, Semantic analysis, Web search

1 Introduction

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

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

- *Purchased positions.* Companies looking for visibility (*i.e.* advertising for the goods or services that they offer) can purchase a better position in the ranked result pages as a service offered by WSEs.
- *Better user experience.* WSEs put in the first result pages the links which are more interesting for the users. This is an indirect source of revenues: the WSE that offers the best experience should attract more users and companies are expected to choose the most popular WSE to place their advertisements.

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

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

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

User profiles can be build using several tools: [48] proposes the use of the browsing history. In [36], the authors use click-though data. Two schemes presented in [46,38] consider the use of web communities for this purpose.

A client side application which stores users' interests is presented in [52]. Finally, some schemes [46,14] use the search queries which have been previously submitted by the users. WSEs generally use this last approach because it is very effective and it profiles the users without their collaboration.

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

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

- The big Internet companies sell user profiles to law enforcement agencies. For example, AOL handles nearly 1000 requests each month for information in criminal and civil cases [15]. Facebook receives between 10 and 20 requests for this kind of information each day [49]. Recently, the Yahoo Compliance Guide for Law Enforcement was disclosed. This document specifies that Yahoo charges the government about 30$ to 40$ for the contents, including e-mail, of a subscriber's account and 40$ to 80$ for the contents of a Yahoo group [61]. Therefore, user profiles represent a source of revenues for WSEs. Nevertheless, the users (the real owners of that information) usually get no income from this item.
- Profiles provide valuable information, hence, attackers are motivated to steal them and get diverse benefits. Therefore, WSEs are responsible for storing user profiles in a secure place and applying access control measures to that content. The AOL scandal [2], which have been explained above, proves that WSEs do not always fulfill these requirements. In that particular case, 20 million queries made by 658000 users were publicly disclosed by employees of the company itself.

Both points show that users should prevent the WSEs from storing and using their profiles in an uncontrolled way. However, profiles are needed in order to provide an efficient service to the users. Thus, allowing the WSEs to only

profile them in an *inaccurate* way may be a proper solution that addresses these two issues. This implies that the resulting profile might be detailed enough to allow the personalized search feature but inaccurate enough to avoid the disclosure of personal information considered risky.

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

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

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

1.1 Previous work

There are several schemes in the literature which try to hide the personal information of the users who submit queries to a WSE. Generally, all these methods provide a distorted profile to the WSE by submitting queries which contain *inaccurate* interests. Depending on the scheme, these inaccurate interests can be practically random or they can be similar to the real ones. The level of inaccuracy directly affects the usefulness of the personalized search: a certain profile which is completely distorted with random interests will make the personalized search totally useless.

In the following, we focus in three main factors that determine the deployability of the different proposals in a real environment:

- *Response time*. This is the time needed to submit a query to the WSE and

receive the answer. Note that a direct query (without any privacy-preserving mechanism) to Google has a response time of 300 ms [5].

- *Personalized search usefulness.* This factor refers to the level of distortion of the resulting user profile. A heavily distorted profile is useless for providing personalized search.
- *Query syntax.* This factor refers to the level of complexity of the queries supported by the proposals. In this paper, a *simple query* contains a single term and a *complex query* contains sentences, parts of sentences or even unconnected noun phrases or words.

Recently, Google presented a new feature which is named *Google Instant* [13]. This innovation places emphasis on making the search experience fast and shows that users are mainly interested in the *response time*. Therefore, current proposals are first classified into two categories which are directly related to this factor: *distributed schemes* and *stand-alone schemes*. Inside each category, all the schemes are analyzed according to the other two factors.

1.1.1 Distributed schemes

Proposals that fall in this section require the collaboration of external entities. These schemes suffer from a significant communication overhead that increases the response time. Nevertheless, the use of external elements allows these proposals to work with real queries which are generated by real users. Therefore, all the schemes in this category submit queries which are complex and make sense. This is very convenient in order to implement these systems in a real environment and preserve the privacy of the users.

Distributed proposals use two main types of external entities:

- *Proxies.* The use of an anonymizing proxy is a simple way of concealing the source of a query. There are several public proxy servers available on the Internet. This solution works as follows: the user first sends her query to the proxy, then the proxy submits the query to the WSE and sends the answer back to the user. The whole process is done anonymously, *i.e.* the true source of the query remains hidden. However, since all the queries are sent through the same proxy, they can be linked together easily by the proxy itself. An adversary with access to the logs of the proxy can identify the true source of the queries.

  Chaum proposed in [6] the use of a group of proxies instead of only one in order to address this problem. The proxies form a network where they act as anonymizers. The input and the output proxies in the network are different at each protocol execution. As a result, an adversary trying to identify the true source of a certain query needs to know the logs of all the proxies of the network. The *Tor Project* [53] is the most renowned implementation of

this approach.

- *Users*. There are some proposals that follow this approach [9,5,38,58]. The main idea beneath all these schemes is that each user submits queries generated by other users. As a result, individual users are hidden inside *group identities*. These schemes mainly differ in the way the users are grouped together.

  The system proposed in [9] uses memory sectors which are shared by a group of users. These users use the shared memory to store and read queries and their answers, hence, there is no direct connection between them.

  The authors in [5] present a scheme which groups users dynamically: first, a central node groups the users who request that; then, the group members establish network connections between them and start to communicate without the interaction of the central node. Users join a new group each time that they want to submit a query.

  Finally, the schemes presented in [38] and [58] use a static network of users. When a user wants to submit a query to the WSE, she forwards it to a randomly selected user of the network. The user that receives the query can either submit or reforward it to another neighbour with a certain probability. In this way, some user eventually submits the query to the WSE. Then, the results are forwarded towards the original user following the reversed path that the query used. In [38], the users' network topology is a complete graph which is constructed and managed by a central entity. On the other hand, [58] can be applied to already developed social networks. (*e.g.* Facebook, Windows Live Messenger, etc).

Distributed schemes share certain shortcomings which are next summarized:

- *Slow response time*. The use of external entities which act as intermediaries implies that the process of submitting a query to the WSE and receiving the answer is very time-consuming. For example, the authors in [40] analyze the anonymous network Tor with paths of length two (note that the default length is three). The results show that submitting a query costs on average 10 seconds. This response time is 25 times slower than the cost of performing a direct search. Regarding the user-based schemes, [5] has been tested in real conditions (*i.e.* users connected through the Internet) and it provides a response time of 5.2 seconds with groups of three users.
- *System availability*. These schemes depend on external entities to submit queries to the WSE, hence, their availability also depends on the availability of these entities. Proxy-based proposals require them to be up and running. Regarding the user-based schemes, they require a large number of users to work properly [5]. In addition to that, the computers of the users are expected to be permanently connected to the network [58].
- *Liability for query content*. This problem is only related to user-based proposals. In these schemes, users submit queries which have been generated by others. Therefore, some participants might be uncomfortable with cer-

tain query contents. Even worse, dishonest users can exploit this approach to send queries about illegal contents. These protocols work without user interaction, they are executed by the software running in the computers of the users. Thus, it can be difficult for that software to identify a legal query from an illegal one. As a result, a certain user can finally submit a illegal query without notice its content. Such a situation is really dangerous and must be addressed.

1.1.2 Stand-alone schemes

This kind of schemes work directly in the computer of the user to be protected. They do not require the collaboration of any external entity. Therefore, these proposals achieve the best response time when submitting queries to the WSE.

A trivial and stand-alone way to provide anonymity to the users who use a WSE is to use dynamic IPs and a web browser without cookies. This approach only hides the real identity of the user. The user queries remain the same. As we have explained above, this solution prevents personalized search from working properly. Besides, this proposal has some additional drawbacks:

- The renewal policy of dynamic IP addresses is not controlled by users but the network operator. This operator can always give the same IP address to the same Media Access Control (MAC) address. In this case, the anonymity of the users will not be protected. In addition to that, certain users require static IP addresses, hence this approach is not suitable for them.
- A browser without cookies losses its usability in a high number of web applications. This situation may not be affordable for certain users.

The authors in [60] present a client application that provides a scalable way for users to automatically build their own user profiles. These profiles summarize the user's interests into a hierarchical organization according to specific interests. Two parameters for specifying privacy requirements are proposed to help the user to choose the content and degree of detail of the profile information which is exposed to the WSE. The queries which are submitted to the WSE are modified according to those factors. This proposal offers to the user a trade-off between the privacy level achieved and the quality of the service. Nevertheless, this scheme requires deep modifications at the server side. Specifically, a search engine wrapper is developed at the server side to incorporate a partial user profile with the results returned from the search engine. Results from both sources are combined and the customized results are delivered to the user by the wrapper. This proposal has two main shortcomings: (i) it is not realistic to assume that well-known WSEs will undergo these kind of internal modifications to preserve the privacy of the users; and (ii) complex queries are not addressed by the authors, hence this proposal only works with

single-word queries at the present time.

There are two works in the literature which are based on submitting random queries together with real ones. Both systems have been implemented in the Firefox web browser: GooPIR [10] and TrackMeNot [54,25]. Their target is to misrepresent the profiles of the users. The authors in [45] state that this approach can be considered as a way to get k-anonymity (see [50] for a definition of k-anonymity). However, the use of random queries damages the personalized search usefulness and, hence, users may get a low quality of service.

GooPIR uses a Thesaurus to obtain the words which are mixed with the real queries. Consequently, the fake queries are single words, while full sentences are not addressed (sentences cannot be formed by random words).

On the other hand, TrackMeNot generates dynamic queries using RSS feedback. These queries are periodically submitted to the WSE and they can be formed by words or sentences. Therefore, this scheme uses complex queries. Nevertheless, the work presented in [31] proves that it is possible to distinguish real queries from TrackMeNot queries. More specifically, [31] develops a classifier which is very accurate in identifying TrackMeNot queries, with a mean of misclassification around 0.02%.

In conclusion, stand-alone schemes are generally affected by two main issues which are next summarized:

- *Personalized search usefulness.* These schemes are usually based on modifying the user queries in order to distort the user profile. Queries can be altered by several different mechanisms, but, in general, it can be difficult for a computer to generate fake queries that look like valid ones (let us assume that a valid query is a query that makes sense) and that contain distorted interests which are still usable by the personalized search process.

  On the other hand, there are some approaches that only focus on hiding the identity of the user while leaving unaltered her queries. If the WSE cannot link different queries which have been submitted by the same anonymous user, then the WSE will not be able to build a profile and the personalized search process will be totally useless.
- *Use of complex queries.* As explained above, generating fake queries that make sense in a fully automated way is an important issue for stand-alone schemes. This situation becomes worse when considering queries which are made of sentences or parts of sentences (these are complex queries). In this case, each query contains different terms which are connected between them and the modification process should consider this. The use of complex queries has not been deeply addressed in the literature, however, [1] shows that they are usually used by the users of WSEs.

In this paper, we propose a new method to hide user's information submitted to WSEs by creating *new* distorted but semantically-related queries. Our proposal only addresses how to generate privacy-preserving and still useful fake queries and it is not linked to a particular protocol for submitting them. Therefore, our approach can be applied to already existing systems. Notice that, given the characteristics of stand-alone schemes according to response time and lack of dependency on external entities, our new proposal suits best with schemes like GooPIR [10] or TrackMeNot [54,25]. In turn, our proposal can provide certain enhancements to these protocols. These improvements are next summarized:

- Analysis and construction of fake queries are handled from a semantic point-of-view. In this manner, user queries can be interpreted according to their underlying semantics and fake ones can be constructed in a way that the utility of distorted query logs can be maintained up to a degree.
- Support of *complex queries* made of sentences, parts of sentences or even unconnected noun phrases or words is provided. Several linguistic analysis techniques, in addition to the basis of the information theory, are used to properly interpret complex queries performed by users and generate new semantically-related ones accordingly.
- Trade-off between quality of service and privacy is properly addressed due to the use of certain configuration parameters. The degree of distortion applied by the proposed system can be configured by varying the semantic distance between user queries and fake ones and/or by stating the amount of fake queries added for each original one.
- As explained above, due to its general design, the proposed method can be incorporated to already existing query log distortion schemes. The overhead added by the lexico-syntactic and semantic analysis is reasonable and assumable in real environments.

The rest of this paper is organized as follows: Section 2 introduces the background needed and describes the new scheme. Section 3 evaluates the proposal in terms of semantic preservation of new queries, the privacy level achieved by the users and its execution time. Finally, Section 4 reports some concluding remarks.

2 Our proposal

In this section, the proposed method is described in detail. Being a stand-alone approach, it does not rely on queries performed by third parties. How-

ever, queries already submitted by the user are analyzed to generate new semantically-related ones in order to preserve the user's profile. Queries are composed by words and these refer to concepts that define their semantics. Therefore, a key to preserve the user's profile while hiding her personal information is to maintain, as much as possible, the semantics of the original queries when creating new ones. However, on the contrary to numerical data, which can be easily manipulated by means of mathematical operators, the processing of textual data and the interpretation of their semantics is a changeling task. In fact, information semantics are inherently human features, defined according to a social consensus [21]. In consequence, semantic interpretation of textual data relies on evidences found in one or several manually constructed knowledge sources. The idea is to mimic human reasoning by using implicit or explicit knowledge.

2.1 Preliminaries

The proposed method relies on structured knowledge modeled in the form of taxonomies or, more generally, ontologies. Ontologies are formal and machine-readable structures of shared conceptualizations of knowledge domains, expressed by means of (mainly taxonomic) semantic relationships [7]. They have been successfully applied in many areas that deal with textual resources like information extraction [41] and knowledge management [57]. Thanks to initiatives such as the Semantic Web [3], many ontologies have been created in the last few years, going from general purpose ones to domain-specific repositories.

Using these structured knowledge sources, words found in user's queries (*e.g.* "water sports") can be mapped to ontological concepts (*e.g.* water sports: "sports that involve bodies of water"), by simply matching their textual labels. Exploring the net of semantic pointers found in the ontology, starting from the matched concept (*i.e.* going upwards or downwards in the taxonomic tree), it is possible to retrieve new semantically related concepts (*i.e.* a taxonomic ancestor or an specialization). Textual labels of these concepts can be used to create semantically related – but different – queries, which can be used to hide the original ones, while retaining their meaning up to a certain degree. For example, if a user queries for *"water sports"*, it is possible to create new related queries like *"swimming"* or *"sailing"* (*i.e.* taxonomic specializations), but also *"sports"* or *"recreation"* (*i.e.* taxonomic subsumers) and *"skiing"* or *"cycling"* (*i.e.* taxonomic siblings).

It is important to note that concepts are referred in a textual query by means of *nouns*, which are considered the minimal *semantic units* (SUs) found in a discourse. These can be recursively refined by adding new nouns or adjectives (*e.g. sport* $\rightarrow$ *water sport*), creating noun phrases (NP). These can be linked

to other ones by means of verbs or verb phrases (VP) (*e.g. this facility offers water sports*) and general connectors (such as conjunctions, prepositions, determinants, etc.) creating sentences. This paper focus the query analysis on nouns and, more generally, noun phrases.

A problem typically faced when mapping input text to ontological concepts appears when text is formed by sentences rather than individual words or simple noun phrases. A sentence can be seen as an ordered set of units, each one being a noun or verb phrase (with one or several words). In these cases, an individual analysis and transformation of each word to construct related but different queries can easily result in a rigmarole. The syntactical interrelation between terms must be maintained when creating a new related syntactical construction. The case of WSE queries is even more difficult because, in many cases, multiple-word queries are not well formed sentences (*e.g. "where to practice water sports in the Mediterranean coast"*). Instead of that, they can be schematic or even a list of unconnected terms (*e.g. "beach water sports Mediterranean"*). We name these kind of queries as *complex queries* that, due to their challenging analysis, are typically omitted by related works [60,10,54]. This paper proposes the use of several linguistic analysis techniques to coherently interpret the original query and enable a meaningful construction of new related queries.

Finally, when dealing with *complex queries* composed by several *semantic units*, a new problem arises. As stated above, a query may include several SUs (*i.e.* individual noun phrases) syntactically interrelated in a sentence or simply on a raw list. When substituting each unit by a new one found in a knowledge base, the implicit semantic interrelation between SUs should be maintained. For example, for the query *"beach water sports Mediterranean"*, *beach*, *water sports* and *Mediterranean* can be identified as the SUs via syntactical analysis. By looking for unbounded transformations of each WSE individually in a knowledge base, it is possible to construct a query like *"cave skiing Pacific"* which is a non-coherent (and probably fake) query. The problem is to detect the main topic of the query (*i.e. sports* and *waters sports* in particular) and retain the implicit semantic relation that states that some types of sports can only be practiced in certain places. This paper proposes focusing the analysis and transformation of the query on the WSE that better represents the *topic* of the query and, in consequence, the user's preferences. It is assumed that the WSE that better represents the query topic is the one that provides *more information* (*i.e.* better describes and identifies the query). Our proposal relies on the *information theory* to quantify the *amount of information of terms* and to guide the query construction toward the transformation that better retains the user's profile. For example, by identifying *water sports* as the WSE which provides the highest amount of information, it can be deduced that the user is interested in *sports*, proposing new coherent queries (*e.g. "sailing"*, *"skiing"*, *"outdoor sports"*).

In the next three subsections each step of the query analysis and construction is described in detail. More specifically:

- The first one performs a morpho-syntactical analysis over each *complex query* in order to obtain the basic SUs that compose it. See section 2.2.
- In the second step, the *information content* of each SU is calculated in order to detect the one(s) that better represents the main topic of the query. See section 2.3.
- The final step semantically analyzes the result by using a knowledge base. It proposes several transformations according to the desired degree of privacy and generates new semantically-related queries to be submitted to the WSE. By varying the semantic distance between user queries and fake ones and/or by stating the amount of fake queries added for each original one, results can be adapted to environments where privacy is crucial or scenarios where query usefulness is required. See section 2.4.

2.2 Analysis of user's queries

The proposed system receives as input each query (q_i) sent by the user in the past. In the most general case, q_i will be a string composed by several words (*i.e.* a *complex query*).

In the first step, a morpho-syntactical analysis is applied to each q_i to extract its SUs. As stated above, these units are noun phrases (NPs) that consist of sets of words in which at least one (*i.e.* the one most on the right) must be a noun. That noun and its possible specializations defined by the other nouns or adjectives attached to it in the NP (*e.g. exciting water sports*) state the semantics which have to be properly interpreted in order to preserve the user's profile.

Several *natural language processing* techniques (NLP) based on *maximum entropy models* are used to syntactically analyze queries. Maximum entropy modeling is a framework for integrating information from many heterogeneous sources for classification [24]. In this paper, the classification task consists in detecting part-of-speech (POS) tags and syntactic units by exploiting a large set of pre-tagged textual corpora as data sources. It has been studied that programs using maximum entropy models have reached state of the art on linguistic analysis [24]. The following tasks are preformed (see an example in Figure 1):

- *Sentence detection.* Individual sentences are detected in case of very long queries.
- *Tokenization.* Atomic parts (words) of the sentence are detected. Shortened forms like "don't" are separated to "do" and "not" for a proper analysis.

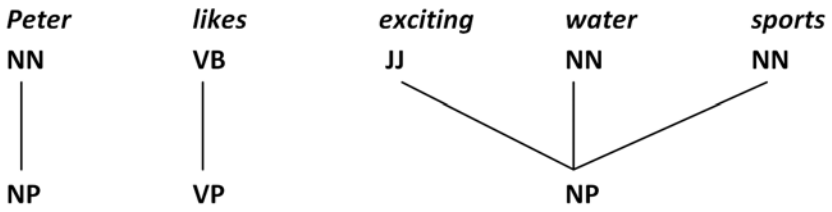

Fig. 1. Example of NL analysis

- *Part-of-speech tagging.* It is the process of marking up the words in a text as corresponding to a particular part of speech (POS) such as nouns (NN), verbs (VB), adjectives (JJ), etc.
- *Syntactic parsing.* Individual words are grouped according to their role in the sentence. As a result, units of meaning represented as noun phrases (NP) are obtained. Other units like verb (VP) or prepositional (PP) phrases are also gathered. This process is often called "chunking".

The OpenNLP Maxent package [30] has been used to perform this analysis. It is a mature Java package for training and using maximum entropy models. It also provides a large collection of models trained for several components of syntactical analyses. These models are language-dependent because syntactic rules may change from one language to another. Several languages are currently supported, including English, Spanish, Portuguese, German or Danish.

As a result of this process, for each q_i, the set of its semantic units $SU_i = \{su_{i,1}, su_{i,2}, ..., su_{i,m}\}$ is obtained, where each $su_{i,j}$ is a noun phrase referring to a distinct concept (*e.g. water sports*). However, several morphological variations or lexicalizations of terms may refer to the same concept (*e.g. water sports*, *water sport*, *some water sports*, etc). To detect equivalent formulations referring to the same concept, an additional linguistic analysis is performed. It consists of:

- *Stop word removal.* Stop words are a finite list of domain independent terms with a very general meaning that can be omitted during a semantic analysis. Determinants, prepositions or adverbs can be removed from each $su_{i,j}$ without altering its meaning (*e.g. some water sports* $\rightarrow$ *water sports*).
- *Stemming.* It removes derivational affixes of words that define, for example, plural forms (*e.g. water sports* $\rightarrow$ *water sport*). It enables the identifications of morphological variations of the same term. The Porter stemming algorithm [35], one of the most common ones applied in Information Extraction, is used.

Again, these analyses are language-dependent. Even though, stop word lists have been compiled and rules for stemming algorithms have been formulated for many languages.

At the end of this step, the set of SUs for each q_i expressed in their simplest

forms are obtained. This will ease the knowledge retrieval from ontologies in a later stage.

2.3 Assessment of query topic

If several SUs have been extracted from a query, the next step consists in ranking them according to the *amount of information* they provide. As stated above, the idea is to focus the semantic analysis and query construction on the SU that better represents the main topic of the query in order to preserve, as much as possible, the profile of the user.

In terms of *information theory* applied to textual data, the Information Content (IC) of a term states the amount of information that it provides when it appears in a certain context (in our case, the query). The basic idea is that general and abstract terms (*e.g. activity*) provide less information than more concrete and specialized ones (*e.g. sailing*) when they appear in a discourse. As a consequence of that, the former presents less IC than the latter. A proper quantification of the IC of terms improves text understanding and it has been applied in the past in the field of computational linguistics [39,18,23,32].

Classical IC-based approaches [39,18] obtain the IC of a term *a* by computing the inverse of its appearance probability in a corpus (*p(a)*) (see expression (1)). In this manner, infrequent terms are considered more informative than common ones.

$$IC(a) = -log(p(a)) \tag{1}$$

Ideally, if the corpus is large and heterogeneous enough to represent term usage at a social scale, *p(a)* will enable an accurate computation of the IC of *a*. However, textual ambiguity and data sparseness severely hamper *p(a)* calculus [42]. First, it is necessary to disambiguate term appearances, identifying word senses by means of tedious and time-consuming manual tagging. Moreover, it is unlikely to have such large and representative tagged corpora, especially for specific domains; this may lead to data sparseness due to the fact that available corpora is not large enough to truly represent term semantics.

To overcome these problems, some authors have proposed estimating *p(a)* intrinsically from the knowledge modeled in an ontology/taxonomy [44,62,43]. These models assume that the taxonomic structure of ontologies is organized in a meaningful way, according to the principles of cognitive saliency [4]: concepts are specialized when they must be differentiated from existing ones. As a result, concepts located at a higher level in a taxonomy and presenting many hyponyms (*i.e.* specializations) would have more probability to occur in a

corpus (*i.e.* higher *p(a)*) than highly specialized concepts located at the leaves of the hierarchy. As a consequence of that, the former will present less IC than the latter. In the most recent approach [43], *p(a)* is estimated from the taxonomic structure of an ontology according to the number of the concept's hyponyms and subsumers (2).

$$IC(a) \cong -log\left(\frac{\frac{|leaves(a)|}{|subsumers(a)|}+1}{max\_leaves+1}\right) \quad (2)$$

Where *leaves(a)* is the set of concepts found at the end of the taxonomic tree under *a* and *subsumers(a)* is the complete set of taxonomic ancestors of *a* including itself. The ratio is normalized by the least informative concept (*i.e.* the root of the taxonomy), for which the number of leaves is the total amount of leaves in the taxonomy (*max_leaves*) and the number of subsumers including itself is 1.

Compared to corpora-based approaches, intrinsic IC calculation models are efficient and scalable and, if large ontologies are available, they are less affected by data sparseness [43]. Due to these reasons, these models (and concretely expression (2)) are used in our proposal to compute the IC of each SU extracted from the original query. The ontology/knowledge repositories used to compute this value are detailed in section 2.6.

At the end of this stage, all $su_{i,j}$ extracted from each query q_i are rated and sorted, starting by the most informative one (mi_su_i). It is assumed that this term is the one that better represents the query topic and, in consequence, the one in which the semantic analysis for query construction/transformation should be focused.

2.4 Construction of new queries

The last step consists in creating new queries which are semantically related to the main topics referred by the original ones. These new queries will be submitted to the WSE in order to hide the original user's queries while maintaining, as much as possible, her personal profile.

First, the system looks for the SU with the highest IC for q_i (mi_su_i) in an ontology. As a result, it retrieves the associated concept (c_i). If mi_su_i is not found in the ontology, the scheme tries to look for simpler forms of the SU by removing adjectives/nouns starting from the one most on the left (*e.g. exciting water sports* $\rightarrow$ *water sports*). The fact that queries incorporate qualifiers is quite common but these can be hardly covered by ontologies, which try to model concepts in a general way. Once (c_i) is found, the system explores the

net of semantic links starting from it in the ontology, in order to retrieve *new* terms that will be used to construct the new queries. Only taxonomic pointers are considered in our approach because, as stated above, taxonomic knowledge is the most structuring and commonly available one in ontologies [8].

The number of semantic links needed to go from a concept to another in an ontology can be interpreted as a function of their semantic distance [37]. This idea has been used in the past to define general-purpose *semantic similarity measures* that are able to quantify the alikeness between terms [37,22,59]. In these approaches, the semantic distance between two concepts found in an ontology (c_1 and c_2) is computed as a function of the length of the minimum taxonomic path connecting them (see expression (3)).

$$distance(c_1, c_2) = |min\_path(c_1, c_2)| \tag{3}$$

In our case, the assessment of the semantic distance between a concept referred in a user's query and a new one retrieved from an ontology, enables a quantification of the semantic differences introduced in the user profile due to the new queries executed. Obviously, there is a trade-off between the preservation of user's profile semantics and the hiding of the user's queries. The more semantically similar the new queries are, the less distortion will be introduced in the user profile but also, the more evident the user's information will remain. Contrarily, the more semantically distant the new queries are, the less clear (and privacy preserved) the user's profile will be. As a consequence, the semantic distance between c_i and those concepts used to construct new queries can be used to balance the degree of desired privacy and the level of semantic distortion introduced in user's profile. This is a configurable parameter (*sem_dis*) of our method.

The proposed scheme explores the ontology starting from c_i going through *sem_dis* taxonomic links (upwards – generalizations – and downwards – specializations – in the hierarchy). All the possible taxonomic paths with a length equal to *sem_dis* are considered, retrieving a set of new concepts $NC_i = \{nc_{i,1}, nc_{i,2}, ..., nc_{i,l}\}$ found at the end of these paths. The fact that subsumers, specializations and even taxonomic relatives are retrieved going upwards ($\uparrow$) and downwards ($\downarrow$) in the taxonomy is desirable, as it avoids constructing queries following a fixed pattern. For example, expressions constructed following only taxonomic generalizations (*e.g. swimming* $\uparrow$ *water sports*; *water sports* $\uparrow$ *sports*; *sports* $\uparrow$ *activities*) will result in general queries created with abstract terms which are rarely performed by WSE users [47]. On the contrary, our method retrieves new concepts that can be subsumers (*e.g. water sports* $\uparrow$ *sports* $\uparrow$ *activity*), specializations (*e.g. water sports* $\downarrow$ *swimming*) and relatives (*e.g. water sports* $\uparrow$ *sports* $\downarrow$ *mountain sports* $\downarrow$ *skiing*).

Once the set of new concepts (NC_i) is retrieved, the label of each $nc_{i,j}$ is used

to construct each new query (nq_i), which will be semantically related to the original one according to the *sem_dis* parameter.

From all the $nc_{i,j}$ retrieved from the ontology, a random set of K of them (if available) is selected, constructing K new queries. K is another predefined parameter that influences how the user's profile is hidden. It states the number of new queries submitted to the WSE for each user's query. The higher the value, the more hidden the original query will be (and also the more distorted the user's profile).

As a final result, for each original query q_i, our proposal creates a set of K new queries $NQ_i = \{nq_{i,1}, nc_{i,2}, ..., nc_{i,K}\}$.

2.5 A note on semantic ambiguity

Language ambiguity may appear when matching mi_su_i to ontological concepts. If mi_su_i is polysemic (*e.g. virus*), the matching process will retrieve a set of ontological concepts $C_i = \{c_{i,1}, c_{i,2}, ..., c_{i,x}\}$, rather than an unique one (referred above as c_i). Each $c_{i,m}$ corresponds to a sense of mi_su_i (*e.g.* a concept referring to *a malicious computer program* and another one referring to *a harmful biological entity*). Each $c_{i,m}$ will belong to a different taxonomic branch and, hence, different subsumers (*e.g. microorganism* or *malevolent program*, respectively) and specializations (*e.g. bacteriophage* or *blaster*, respectively) will be retrieved when exploring the ontology for each one. The selection of the adequate concept according to the sense in which mi_su_i is used in q_i is necessary to avoid ambiguity and to accurately preserve the user's profile.

A way to decide which is the appropriate c_i for mi_su_i is to analyze the *context* of mi_su_i in q_i. The context of a word in a text bounds the sense in which the word is used, because each additional term in the context constrains the domain of the discourse. The context of a term in a query is, in our case, the other terms found in the same query (in case of complex queries with several semantic units, for example *"virus removal windows xp"*) and/or previous queries (in case of simple ones, such as *"antivirus"*, *"mcafee"*). The appropriate concept for mi_su_i will be the one that is the most *semantically similar* to all the other terms found in q_i (*i.e.* its context). This assessment is performed using the notion of *semantic similarity*. Formally, the system computes the semantic distance (3) between each $c_{i,m}$ retrieved for mi_su_i and each other SUs found in the same or previous queries. The $c_{i,m}$ that, in average, results in the *minimum semantic distance* to all of SUs considered is selected as the most adequate one (c_i) for mi_su_i.

The methodology presented above has been designed in a generic way, hence, it can be applied to any kind of textual query and it can use any structured knowledge base including, at least, a taxonomic backbone. Obviously, the larger and more detailed the knowledge base is, the more accurate the semantic analysis will be. Moreover, a large ontology covering the topics of the input queries is needed in order to map query terms to concepts.

Nowadays, many general-purpose or domain-specific ontologies or structured thesauri have been created. A paradigmatic example is WordNet [11], a domain-independent knowledge source that describes and organizes more than 100000 general English concepts, which are semantically structured in an ontological fashion. It contains words that are linked to sets of cognitive synonyms (*synsets*), each one expressing a distinct concept (*i.e.* a word sense). As a matter of fact, polysemic words correspond to an average of 2.77 concepts. Synsets are linked by means of semantic relations such as *synonymy*, *hypernymy* (is-a), six types of *meronymy* (part-of), *antonymy*, etc. The backbone of the network is the subsumption hierarchy which accounts more than an 80% of all the modelled semantic links, with a maximum depth of 16 levels. The result is a network of meaningfully related terms, where the graph model can be exploited to interpret the meaning of the concept. WordNet can be downloaded and consulted off-line by means of several APIs available for different programming languages.

WordNet offers a detailed and coherent knowledge structure for general concepts, as it have been manually and carefully developed by knowledge engineers [11]. However, its coverage for *named entities* (*e.g.* proper nouns, place names, brand names, etc.) is very reduced. Named entities configure a very wide and dynamic domain which can be hardly covered by "static" repositories like WordNet. At the same time, named entities commonly appear in WSE queries. In these cases, other repositories like the Open Directory Project (ODP) should be used.

ODP is the largest, most comprehensive human-edited directory of the Web [29]. It is constructed and maintained by a vast, global community of volunteer editors. The purpose of ODP is to list and categorize web sites. Manually created *categories* are taxonomically structured and associated with related web resources. This structure can be used in the same manner as WordNet to analyze and create new queries. The advantage is its large size and high recall, with more than 1 million categories covering up-to-date named entities [29]. Directory data files can be downloaded in SQL format and categories can be consulted off-line efficiently. As a drawback, the fact that the repository is cooperatively edited by a large amount of uncoordinated end-users (more

than 88000 [29]) causes that the taxonomic structures are not as coherent and meaningful as those offered by WordNet for the same topics.

Our proposal uses both WordNet and ODP to assess query topics (second step, see section 2.3) and create new queries (third step, see section 2.4) from a domain independent perspective. WordNet is the preferred repository in cases in which the topic is found in both sources. ODP is only used when the term is not found in WordNet.

It is important to note that, in cases in which queries are domain specific, domain ontologies can be used instead. For example, queries formulated to search engines of specialized repositories such as medical digital libraries like PubMed [56] can be processed by means of large biomedical ontologies as SNOMED-CT [51] or MeSH [26]. These ontologies are comparable in terms of size and detail to the above-mentioned general-purpose knowledge bases. The same can be applied to other domains such as chemistry [27].

3 Evaluation of the proposed scheme

As stated in Section 2.1, the proposed method aims to hide user's queries while maintaining, up to a certain degree, the semantic of her profile. The first aspect can be understood as the difficulty to distinguish between queries performed by the user and fakes ones created by the proposed method. The second aspect influences the utility of the masked query set from the data exploitation point of view.

In this section, these two opposed dimensions are evaluated. As test data, a set of query logs taken from real users and compiled by AOL during three months in 2006 [1] has been used. Specifically, a random subset of 1000 queries executed by 26 users have been selected. Being real web queries, it is very common to deal with *complex queries* made of several words/NPs. As a matter of fact, around 60% of the queries evaluated contained more than one word. As a result, it configures a good and realistic testbed to evaluate the behavior of our proposal.

Two parameters influence the results: K states the number of fake queries introduced by our scheme for each original one; *sem_dist* defines the semantic distance between the original query and the new ones, according to the background ontologies. Intuitively, the higher these parameter values are, the more difficult the identification of the original queries in the masked set will be. On the contrary, the lower the values, the more preserved the user's profile would remain.

In addition to the latter two aspects, Section 1.1 states the interest of the users in the *response time* of the provided schemes. Accordingly, we also discuss in this section the total overhead introduced by our proposal to any mechanism used to submit the generated queries.

As a result, three evaluations have been performed: (i) preservation of semantic of new queries with respect to the original ones; (ii) privacy of masked results; and (iii) runtime of the proposed method.

3.1 Semantic preservation evaluation

In order to quantify the degree of preservation of semantics of the user's profile, we computed the difference of Information Content (IC) between each of the original 1000 queries considered in the evaluation and the new queries constructed when $K = 1$. The idea is that, if an original query has a certain degree of concreteness (represented by its IC), each new query created from it should maintain, as much as possible, that concreteness and, hence, present a similar IC value.

To compute the IC of original and new queries in an objective manner, we used the *web hit count* provided by a web search engine when querying them. In this manner, we compute term's IC from its appearance frequencies in the largest corpus available: the Web [42].

Hit counts, even being a rough estimation of real appearance frequencies of queried terms in the Web [20], have some advantages. First, it is possible to obtain hit counts from almost any query, regardless its complexity and syntactic construction. On the contrary, IC values accurately computed from reliable sources like ontologies can only be obtained for simple terms (*i.e.* words or individual noun phrases). Moreover, web hit counts are not biased by the background ontology used to create new queries and, hence, they offer an objective comparative measure of query concreteness. The IC of a query can be computed from the Web as follows (4):

$$IC(a) = -log(p(a)) = -log\left(\frac{hit\_count(a)}{total\_webs}\right) \tag{4}$$

Where *hit_count* is the number of results returned by a WSE (Bing, in our case) and *total_webs* is the amount of web resources indexed by that search engine. 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 [55]. As a result, the IC of a term can be directly calculated using *hit_count*.

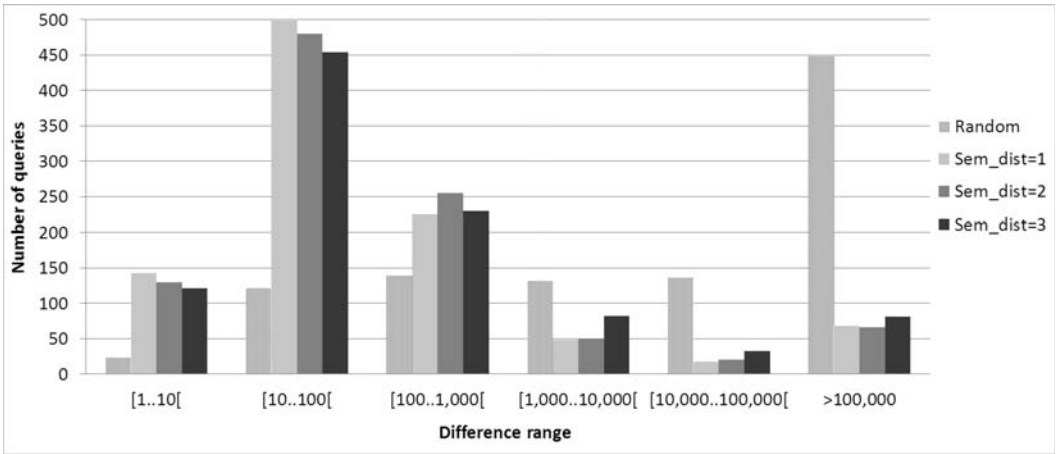

Fig. 2. Difference between the IC of original and new queries

In Figure 2, the histogram grouping queries according to their degree of IC preservation is shown. For each original query q_i, the *hit_count* of its most informative SU (*hit_count(mi_su$_i$)*) was computed. Then, for the resulting new query nq_i, its *hit_count(nq_i)* was also calculated. The degree of IC preservation between them is evaluated as the ratio between the query with the highest *hit_count* with respect to the other expression (5). The result is the absolute difference (in orders of magnitude) between queries' IC.

$$IC\_difference(a,b) = \frac{Max(hit\_count(a), hit\_count(b))}{Min(hit\_count(a), hit\_count(b))} \tag{5}$$

Notice that, due to the enormous size of the Web, *hit_count* may range from a few dozens to several billions. In consequence, the X-axis of Figure 2 groups queries following a logarithmic scale. Y-axis shows the number of queries for each group. Several tests have been performed varying the *sem_dist* of new queries from 1 to 3 (according to expression (3)). As a comparative baseline, we also include the evaluation results when using a naive query construction: for each q_i, we created a new one by taking a *random* concept extracted from input ontologies and compared them using expression (5).

Figure 2 shows a clear difference between the proposed semantically-grounded query construction in comparison to the random approach. Our method creates new queries that, in most cases, differ among 10-100 times with respect to the original query topic. As stated above, considering the size of the Web and the range in which *hit_count* may vary, this is a quite constrained result. The non-semantic query construction, on the contrary, results in queries that, in most cases differ more than 100000 times. This shows the convenience of exploiting background knowledge to create semantically related queries and preserve, up to a degree, the user's profile.

Differences according to *sem_dist* values are comparatively low. Figure 2 shows

that differences tend to be lower if *sem_dist* values are low. This is maintained up to the [100-1000[range, in which the tendency is inversed. Concretely, new queries with *sem_dist=1* are most in the range [1..100[, whereas queries with *sem_dist=2* are most in the range [100..1000[and queries with *sem_dist=3* are most in the range above 1000. This shows an almost proportional growth in queries' IC difference as the semantic distance between original and new increases. Obviously, by rising the K parameter, the observed differences would linearly grow. Summarizing, user's profile semantics would be less preserved (but better masked) as more semantically distant queries are added to her query log.

3.2 Privacy evaluation

The privacy level provided by the new proposal has been evaluated using the Profile Exposure Level (PEL). This is a privacy metric which was introduced and successfully applied in [28].

The next subsection explains this metric. After that, the results of the privacy analysis are presented and discussed.

3.2.1 Profile Exposure Level (PEL)

Let X be the original set of queries of user id, and Y be the protected ones. X and Y can be seen as random variables, which can take so many values as different queries they have and with probability proportional to the number of repetitions. Then, the Profile Exposure Level (PEL) [28] is defined as follows:

$$PEL_{id} = \frac{I(X,Y)}{H(X)} \cdot 100$$

where H(X) is the entropy of the original set of queries and I(X,Y) is the mutual information between X and Y.

The mutual information between two random variables X and Y is given by:

$$I(X,Y) = H(X) - H(X/Y)$$

Regarding the information entropy of variable X and X/Y, they are expressed in terms of a discrete set of probabilities p_i:

$$H(X) = -\sum_{x} p(x) \cdot \log_2 p(x)$$

$$H(X/Y) = -\sum_{x,y} p(x,y) \cdot \log_2 p(x/y)$$

Notice that $H(X/Y)$ is the conditional entropy of the variable X given the variable Y. This is the uncertainty about X which still remains after Y is known.

PEL measures the percentage of user information that is exposed when Y is disclosed. Thus, the user information is calculated as the entropy of X, and the mutual information gives a measure of the information that Y provides about X (*i.e.* if Y is known, how much does this reduce the uncertainty about X?).

3.2.2 Results and discussion of the privacy analysis

In order to evaluate the privacy level, we have used the same subset of queries which have been introduced before (see Section 3). This subset is formed by 1000 queries from the AOL files which were generated by 26 users. Each user submitted a different number of queries to the WSE: some of them sent about 50-200 queries while others submitted approximately 5-10 queries.

The PEL metric has been applied comparing the set of real queries with the protected ones. We have tested the proposed scheme for values of K from 1 to 5 (this is the number of fake queries introduced by the method for each original one) and semantic distances (value *sem_dist*) from 1 to 3.

Figure 3 shows the results according to different values of K and *sem_dist*. As expected, the most relevant factor from the privacy protection point of view is K: as this value grows, the percentage of user information that is exposed decreases (for $K = 1$ there is about 65% of exposure and for $K = 5$ there is about 35%). As expected, the semantic distance has no significant incidence in the privacy level achieved by the users, because the latter is based on query-matching. Semantic distance is however related to the usefulness of the resulting profiles.

Considering only the privacy point of view, we argue that 40% of exposure is good enough ([28] considers that a user profile has enough protection when it achieves a PEL of 40% or less). According to our results, a PEL result of 40% is achieved with $K \geq 4$.

As introduced in Section 3.1, by rising the K parameter, the semantics of

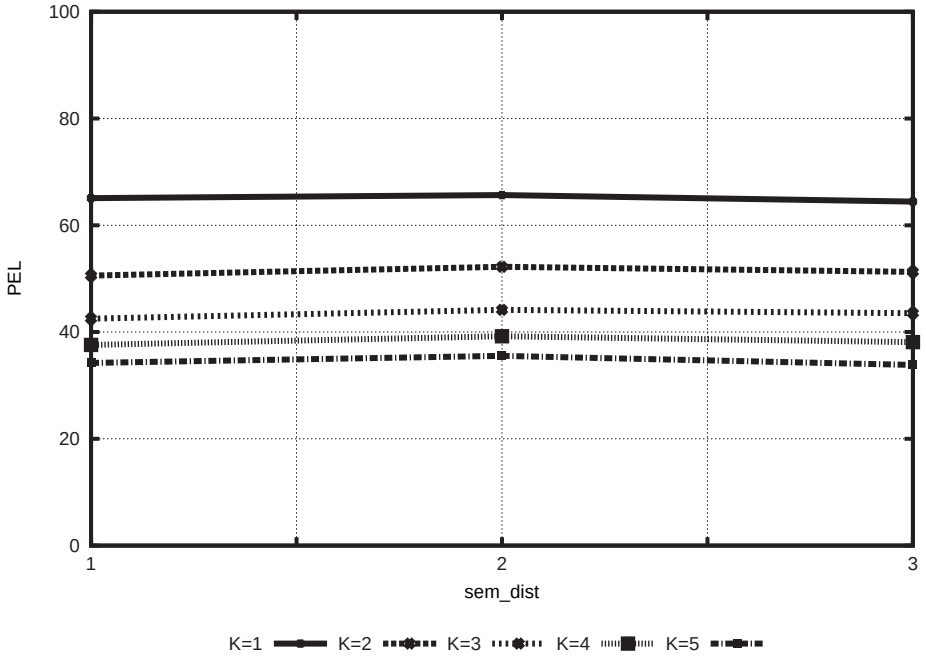

Fig. 3. PEL results for different semantic distances (*sem_dist*) and different K values

user profiles would be less preserved as more queries with a certain semantic distance are added to their query log. Therefore, K and *sem_dist* represent a trade-off between privacy level and profile usefulness. Each user should select the configuration that best fits her needs.

3.3 Runtime evaluation

Regarding the expected runtime of the proposed method, by far, the most CPU-demanding task is the retrieval of concepts (for IC calculus or query construction) from the background ontologies. Considering that queries are typically very short texts, the runtime required by the morpho-syntactical analysis is almost negligible. Background ontologies, on the other hand, contain thousand or millions of terms, and their retrieval introduces a noticeable overhead.

This overhead depends on the method used to access knowledge. WordNet's ontological structure can be completely pre-loaded in memory and accessed efficiently via Java-based API calls. As a result, the complete analysis of a query found in WordNet takes an average of 30 ms. ODP, on the contrary, must be accessed by SQL calls to a local database. Due to the lower-efficient access method and the larger size of the repository, the analysis of a query not found in WordNet (in the first try) but retrieved from ODP (in a second chance) takes an average of 1500 ms. These runtimes correspond to an Intel

Core 2.6Ghz CPU.

In any case, considering the volume of queries that a typical user may perform to a WSE (according to [17], in 2005 each user submitted 971 queries on average), the overhead introduced in her system by the proposed method should not be noticeable.

4 Concluding remarks

User profiles enable the WSEs to offer a better user experience. Nevertheless, profiles built from search queries contain personal information which can univocally identify their owners. This personal information can represent a privacy threat for the users and new schemes which address the trade-off between quality of service and privacy should be proposed.

Besides, the kind of queries which are submitted by the users must be considered too. Generally, a query can contain from one single word up to several words. The latter is the worst case because there is no pattern that defines the structure of a query. We have named this kind of queries as *complex queries*.

In this paper, a new method to hide user's information submitted to WSEs has been proposed. This new scheme works by creating *new* distorted but semantically-related queries and it is not linked to a particular protocol for submitting them. The use of stand-alone protocols for this purpose has been suggested. More specifically, mechanisms like GooPIR or TrackMeNot can exploit its benefits.

A first contribution of this paper is related to the construction of new queries. This process is handled from a semantic point of view in order to preserve the user's profile. The new queries, created by analyzing those already performed by the user, introduce a configurable degree of information distortion proportional to the desired level of privacy.

A second contribution is the support of *complex queries*. Several linguistic analysis techniques, in addition to the basis of the information theory, are used to properly interpret complex queries performed by the user and generate new semantically-related ones accordingly.

Finally, the proposed method offers a high flexibility due to its configuration parameters. The degree of distortion can be configured by varying the semantic distance between user queries and fake ones and/or by stating the amount of fake queries added for each original one. By varying these parameters, results can be adapted to environments where privacy is crucial (*e.g.,* stating a high

semantic distance) or scenarios where query usefulness is preferred.

The performance of the new scheme has been evaluated in terms of semantic preservation of new queries and the privacy level achieved by the users. The first test shows that the proposed scheme provides a fairly better degree of preservation of semantics than a random approach (this approach is likely to be applied by systems that do not consider the semantic distance between real and fake queries). Regarding the privacy evaluation, the results of the test offer relevant information about the K value (the number of fake queries introduced by the proposed method for each original one) which is needed in order to achieve a certain privacy level. Notice that the K value represents a trade-off between privacy level and profile usefulness.

Finally, the runtime overhead introduced by the new proposal has been discussed. Considering the volume of queries that a typical user may perform to a WSE, it should not be significant.

Disclaimer and acknowledgments

The authors are with the UNESCO Chair in Data Privacy, but they 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 Spanish Ministry of Science and Innovation through projects TSI2007-65406-C03-01 "E-AEGIS", CONSOLIDER CSD2007-00004 "ARES", PT-430000-2010-31 "Audit Transparency Voting Process", by the Spanish Ministry of Industry, Commerce and Tourism through projects TSI-020100-2009-720 "eVerification", TSI-020302-2010-153 "SeCloud", and by the Government of Catalonia under grant 2009 SGR 1135.

References

[1] M. Arrington, "AOL proudly releases massive amounts of private data", *TechCrunch*, August 2006.

[2] M. Barbaro, T. Zeller, "A face is exposed for AOL searcher No 4417749", *New York Times*, August 2006.

[3] T. Berners-Lee, J. Hendler, O. Lassila, "The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities", *Scientific American*, vol. 284, no. 5, pp. 34-43, 2001.

[4] A. Blank, "Words and Concepts in Time: Towards Diachronic Cognitive Onomasiology", *Words and Concepts in Time: towards Diachronic Cognitive*

Onomasiology, pp. 37-36, 2003.

[5] J. Castellà-Roca, A. Viejo, J. Herrera-Joancomartí, "Preserving user's privacy in web search engines", *Computer Communications*, vol. 32, no. 13–14, pp. 1541–1551, 2009.

[6] D.L. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms", *Communications ACM*, vol. 24, no. 2, pp. 84–90, 1981.

[7] P. Cimiano, "Ontology Learning and Population from Text: Algorithms, Evaluation and Applications", *Springer-Verlag*, 2006.

[8] L. Ding, T. Finin, A. Joshi, R. Pan, R.S. Cost, Y. Peng, P. Reddivari, V. Doshi, J. Sachs, "Swoogle: A Search and Metadata Engine for the Semantic Web", *Proc. of the thirteenth ACM international conference on Information and knowledge management*, pp. 652–659, 2004.

[9] J. Domingo-Ferrer, M. Bras-Amorós, Q. Wu, J. Manjón, "User-Private Information Retrieval Based on a Peer-to-Peer Community", *Data and Knowledge Engineering*, vol. 68, no. 11, pp. 1237–1252, 2009.

[10] J. Domingo-Ferrer, A. Solanas, J. Castellà-Roca, "$h(k)$-Private Information Retrieval from Privacy-Uncooperative Queryable Databases", *Journal of Online Information Review*, vol. 33, no. 4, pp. 1468-4527, 2009.

[11] C. Fellbaum, "WordNet: An Electronic Lexical Database", *MIT Press*, 1998.

[12] Google AdWords, 2011. `http://adwords.google.com`

[13] Google instant, 2010. `http://www.google.com/instant/`

[14] Google personalized search, 2010. `http://www.google.com/psearch`

[15] S. Hansell, "Increasingly, Internet's Data Trail Leads to Court", *New York Times*, February 2006.

[16] iProspect.com, Inc,. "iProspect Blended Search Results Study", 2008. `http://www.iprospect.com`

[17] Internet World Stats, 2008. `http://www.internetworldstats.com`

[18] J.J. Jiang, D.W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", *Proc. of the 10th International Conference on Research in Computational Linguistics*, pp. 19–33, 1997.

[19] R. Jones, R. Kumar, B. Pang, A. Tomkins, "I know what you did last summer: query logs and user privacy", *Proc. of the 16th ACM Conference on Information and Knowledge Management –CIKM'07*, 2005.

[20] A. Kilgarriff, "Googleology is Bad Science", *Journal of Computational Linguistics*, vol. 33, no. 1, pp. 147–151, 2007.

[21] T. Landauer, S. Dumais (1997), "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge" *Psychological Review*, vol. 104, pp. 211–240, 1997.

[22] C. Leacock, M. Chodorow, "Combining local context and WordNet similarity for word sense identification", *WordNet: A Lexical Reference System and its Application* , pp. 265–283, 1998.

[23] D. Lin, "An Information-Theoretic Definition of Similarity", *Proc. of the Fifteenth International Conference on Machine Learning*, pp. 296–304, 1998.

[24] C. Manning, H. Schutze, "Foundations of Statistical Natural Language Processing", *MIT Press*, 1999.

[25] G. T. Marx, "A Tack in the Shoe: Neutralizing and Resisting the New Surveillance" *Journal of Social Issues*, vol. 59, no. 2, pp. 369–390, 2003.

[26] Medical Subject Headings, 2010. `http://www.ncbi.nlm.nih.gov/mesh`

[27] J. Morbach, A. Yang, W. Marquardt, "OntoCAPE-A large-scale ontology for chemical process engineering", *Engineering Applications of Artificial Intelligence*, vol. 20, pp. 147–161, 2007.

[28] G. Navarro-Arribas, V. Torra, A. Erola and J. Castellà-Roca, "User k-anonymity for privacy preserving data mining of query logs", *Information Processing & Management*, in press.

[29] Open Directory Project, 2010. `http://www.dmoz.org/docs/en/about.html`

[30] OpenNLP Maxent Package, 2010.
`http://maxent.sourceforge.net/about.html`

[31] S. T. Peddinti, N. Saxena, "On the privacy of web search based on query obfuscation: a case study of TrackMeNot", *Proceedings of the 10th international conference on Privacy enhancing technologies – PETS'10*, pp. 19–37, 2010.

[32] G. Pirro, "A semantic similarity metric combining features and intrinsic information content", *Data and Knowledge engineering*, vol. 68, no. 11, pp. 1289-1308, 2009.

[33] J. Pitkow, H. Schuetze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, T. Breuel, "Personalized search", *Communications of the ACM*, vol. 45, no. 9, pp. 50–55, 2002.

[34] B. Poblete, M. Spiliopoulou, R. Baeza-Yates, "Privacy-preserving Query Log Mining for Business Confidentiality Protection", *ACM Transactions on the Web*, vol. 4, no. 3, 2010.

[35] M.F. Porter, "An algorithm for suffix stripping", *Readings in Information Retrieval*, pp. 313-316, 1997.

[36] F. Qiu, J. Cho, "Automatic identification of user interest for personalized search", *Proc. of the 12th International World Wide Web Conference*, 2006.

[37] R. Rada, H. Mili, E. Bichnell, M. Blettner, "Development and application of a metric on semantic nets", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 17–30, 1989.

[38] M.K. Reiter, A.D. Rubin, "Crowds: anonymity for Web transactions", *ACM Transactions on Information and System Security*, vol. 1, no. 1, pp. 66–92, 1998.

[39] P. Resnik, "Using Information Content to Evalutate Semantic Similarity in a Taxonomy", *Proc. of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–453, 1995.

[40] F. Saint-Jean, A. Johnson, D. Boneh, J. Feigenbaum, "Private Web Search", *Proc. of the ACM workshop on Privacy in electronic society – WPES'07*, pp. 84–90, 2007.

[41] D. Sánchez, D. Isern, M. Millán "Content Annotation for the Semantic Web: an Automatic Web-based Approach", *Knowledge and Information Systems*, in press, 2010.

[42] D. Sánchez, M. Batet, A. Valls, K. Gibert, "Ontology-driven web-based semantic similarity", *Journal of Intelligent Information Systems*, vol. 35, no. 3, pp. 383–413, 2010.

[43] D. Sánchez, M. Batet, D. Isern, "Ontology-based information content computation", *Knowledge-Based Systems*, vol. 24, no. 2, pp. 297-303, 2011.

[44] N. Seco, T. Veale, J. Hayes, "An Intrinsic Information Content Metric for Semantic Similarity in WordNet", *Proc. of the 16th European Conference on Artificial Intelligence, ECAI 2004, including Prestigious Applicants of Intelligent Systems*, pp. 1089-1090, 2004.

[45] X. Shen, B. Tan, C.X. Zhai, "Privacy Protection in Personalized Search", *ACM SIGIR Forum*, vol. 41, no. 1, pp. 4–17, 2007.

[46] M. Speretta, S. Gauch, "Personalizing search based on user search history", *Proc. of the 13th ACM Conference on Information Knowledge Management –CIKM'04*, 2004.

[47] A. Spink, D. Wolfram, B.J. Jansen and T. Saracevic, "Searching the Web: The Public and Their Queries", *Journal of the American Society for Information Science*, vol. 52, no. 3, pp. 226–234, 2001.

[48] K. Sugiyama, K. Hatano, M. Yoshikawa, "Adaptive Web Search based on user profile constructed without any effort from users", *Proc. of the 13th International World Wide Web Conference*, 2004.

[49] N. Summers, "Walking the Cyberbeat", *Newsweek*, May 2009.

[50] L. Sweeney, "k-anonymity: a model for protecting privacy", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 557-570, 2002.

[51] Systematized Nomenclature of Medicine, 2010. `http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html`

[52] J. Teevan, S. T. Dumais, E. Horvitz, "Personalizing search via automated analysis of interests and activities", *Proc. of 28th Annual International ACM Conference on Research and Development in Information Retrieval –SIGIR'05*, 2005.

[53] Tor Project, 2010. `http://www.torproject.org`

[54] TrackMeNot, 2010. `http://mrl.nyu.edu/ dhowe/trackmenot`

[55] P. Turney, "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL", *Procs. of the 20th European Conference on Machine Learning*, pp. 491–502, 2001.

[56] U.S. National Library of Medicine, National Institues of Health, 2010. `http://www.ncbi.nlm.nih.gov/pubmed`

[57] A. Valls, K. Gibert, D. Sánchez, M. Batet, "Using ontologies for structuring organizational knowledge in Home Care assistance", *International Journal of Medical Informatics*, vol. 79, no. 5, pp. 370–387, 2010.

[58] A. Viejo, J. Castellà-Roca, "Using Social Networks to Distort Users' Profiles Generated by Web Search Engines", *Computer Networks*, vol. 54, no. 9, pp. 1343–1357, 2010.

[59] Z. Wu, M. Palmer, "Verb semantics and lexical selection", *Procs. of the 32nd annual Meeting of the Association for Computational Linguistics*, pp. 133–138, 1994.

[60] Y. Xu, B. Zhang, Z. Chen, K. Wang, "Privacy-Enhancing Personalized Web Search", *Proc. of the 16th international conference on World Wide Web*, pp. 591–600, 2007.

[61] K. Zetter, "Yahoo Issues Takedown Notice for Spying Price List", *Wired*, December 2009.

[62] Z. Zhou, Y. Wang, J. Gu, "A New Model of Information Content for Semantic Similarity in WordNet", *Proc. of the Second International Conference on Future Generation Communication and Networking Symposia*, pp. 85–89, 2008.