MASTER THESIS:

DEVELOPMENT OF A PRIVACY TOOL FOR THE PROTECTION OF PROFILES IN SEARCH ENGINES

Francisco Javier Rodrigo Ginés
Directed by Dr. Javier Parra Arnau

Master’s Degree in Computer Security and Artificial Intelligence
Index

1. Introduction
2. State of the Art
3. Contribution
4. Evaluation
5. Conclusions
6. Future work
   ■ Q & A
1. Introduction
1. Introduction

Fig. 1. Queries processed by Google per year.
1. Introduction

Fig. 2. The profile of a user is modelled in Google as a list of topic categories. The profile shown here reflects the user is interested in parenting-related topics, which might reveal she is pregnant.
1. Introduction

Fig. 3. Three-quarters of search users say collecting user information to personalize search results is not okay. Source: (Kristen Purcell, 2012) Survey, January 20-February 19, 2012. N=2,253 adults, age 18 and older, including 901 cell phone interviews. Interviews conducted in English and Spanish.
1. Introduction: Goal

- The aim of this work is to contribute to the development of privacy-enhancing technologies (PETs) that may attain a suitable trade-off between privacy and search accuracy.

- We have investigate a privacy mechanism that capitalizes on the generalization of queries, the replacement of specific and probably sensitive queries, into more general search terms (but semantically similar).

---

**Fig. 4.** Trade-off between privacy and search accuracy. Source: Javier Parra-Arnau: *Privacy Protection of User Profiles in Personalized Information Systems, 2013*
1. Introduction

- The designed tool, called *PrivacySearch*, is implemented as a Web-browser extension and allows users to generalize their search queries in an automated fashion, as they type the query, without consulting any external entity or database, and according to simple and intuitive privacy criteria.

Fig. 5. PrivacySearch privacy level selection menu.
2. State of the Art
2. State of the Art

Fig. 6. Schema representing privacy techniques in Web search. Orange boxes represent existing applications; green boxes represent proposals.
2. State of the Art: Background

- (A. Avi et al., 2013)* generates a set of more general, semantically-related queries that loosely correspond to that interest. Each of these queries are submitted independently to the WSE, and the level of privacy protection is determined by the least private term.

- (David Sánchez et al., 2013)** proposes replacing user queries with general terms. However, since the aim is not to protect each single search but the accumulated query profile, it may happen that certain individual queries are exposed to the service provider.

---


3. Contribution
3. Contribution

- Our approach is based on the principle of query generalization. Although there exist theoretical proposals relying on this same principle of information; there is no practical tool available to end-users.

- Our solution replaces the queries to be submitted by a user with generic terms, so that the search engine cannot find out the exact information they are looking for. How generic these terms are determined by the users themselves through appropriate and simple privacy configurations.
3. Contribution

- Specifically, our tool allows users to configure three levels of privacy: low, medium and high. The selected privacy level indicates how generic the query generated by PrivacySearch will be from the original query.

Fig. 7. Selection of the level of privacy in our tool.
3. Contribution

- Also, our technology is specifically designed to meet these two fundamental requirements:
  
  o **Real time.** At the time of sending a query, it must be replaced automatically by a term of a higher semantic category, without the user perceiving any degradation in the search engine’s response time.

  o **Local mode.** The generalization algorithm, and in general the query protection tool, must perform all operations on the user side, without the help of any type of infrastructure.
3. Contribution: System description

- For the development of PrivacySearch, several techniques of Natural Language Processing (NLP) have been used.
- To generalize user queries, we capitalize on WordNet as a categorizer. Since each WordNet entry stores its hyperonym and its hyponym, it can be construed as an ontology.

*Fig. 8. Example of hyperonymy in WordNet.*
3. Contribution: System description

- When our plug-in receives a query, it processes it in three different steps. First, PrivacySearch pre-process the text; subsequently it performs a linguistic disambiguation; and finally it concludes by making a categorization according to the level of privacy selected by the user.

Fig. 8. PrivacySearch processing steps.
3. Contribution: System description

- **Pre-processing Queries.** In this very first step, a series of tasks are conducted that aim to prepare the query for further processing by WordNet. This step is of special importance for the real-time requirement specified in the previous subsection. Essentially, queries are “simplified”, although the meaning is kept, in order to diminish the workload of the following two steps.
3. Contribution: System description

- To pre-process a query, we perform the following tasks:

  1. The query is converted to lowercase.
  2. The so-called “stop words” of the query are removed.
  3. The query is tokenized.
  4. Plural terms become singular.
  5. The atomic units or tokens obtained are lemmatized.
  6. We obtain the n-grams existing in the query. For reasons of efficiency, our categorization algorithm uses unigrams and bigrams.
  7. Finally, terms are eliminated in languages other than English, or that do not exist in WordNet, in order to reduce the computing time in the language disambiguation step.
3. Contribution: System description

- **Linguistic Disambiguation.** Once the query has been pre-processed, we must determine the correct meaning of each term. WordNet stores the different meanings that terms can have. For example, the word “bank” can refer to a pile or mass of some material, or a company dedicated to perform financial operations, among other meanings. The only way to ascertain the correct meaning is through linguistic disambiguation.
3. Contribution: System description

- The concrete implementation of the algorithm is shown below by means of an example:

1. Let $A \ B \ C$ be the input query.
2. Assume each term has a set of associated meanings, denoted as follows:
   
   $A \in \{a_1, a_2, a_3\}$, $B \in \{b_1\}$, $C \in \{c_1, c_2\}$.

3. All possible permutations of meanings are formed:
   
   $(a_1, b_1, c_1), (a_1, b_1, c_2), (a_2, b_1, c_1), \ldots, (a_3, b_1, c_2)$

4. A function $F(a, b)$ is defined that returns the distance between a pair of meanings:

$$F(a, b) = \sum_{i=0}^{H_a} \sum_{j=0}^{H_b} \frac{1}{H_a/2} + \frac{1}{H_b/2}$$

   being $H_a$ the sequence of existing hyperonyms between $a$ and entity, and $H_b$ the sequence of existing hyperonyms between $b$ and entity, where $H_{ai} \neq H_{ij}$

   1. $F$ is evaluated for each pair of permutation meanings, and the permutation with less distance is chosen.
3. Contribution: System description

- **Categorization.** Once the original query has been pre-processed and the correct meanings of each $n$-gram obtained, we proceed to conduct the categorization of the resulting terms. To carry out this step, we utilize WordNet as a categorizer. WordNet is a very popular lexical database of the English language. The proposed categorization algorithm is described in Algorithm 1.
3. Contribution: System description

Algorithm 1: Query categorization algorithm.

Input: \( Q \), a sequence of terms that represents a genuine preprocessed query.
Output: \( Q_{\text{Categorized}} \), a sequence of terms that represents a generalized query, built from the genuine query.

1. let \( PrivacyLevel \in \{\text{Low, Medium, High}\} \), level of privacy chosen by the user.
2. let \( entity \), root hypername of WordNet.
3. for each term \( t \) in \( C \) do
   4. \( H_t \leftarrow \) Sequence of direct hyperlinks from \( t \) until \( entity \).
   5. if \( PrivacyLevel \) is Low then
      6. \( Q_{\text{Categorized}} \leftarrow H_{t:1} \)
   7. else if \( PrivacyLevel \) is Medium then
      8. \( Q_{\text{Categorized}} \leftarrow H_{t(H_t|0.1)} \)
   9. else if \( PrivacyLevel \) is High then
      10. \( Q_{\text{Categorized}} \leftarrow H_{t(H_t|0.2)} \)
11. end
3. Contribution: Example
3. Contribution: Example
3. Contribution: Example
3. Contribution: Example

Affective Disorders: Types, Symptoms, and Causes

Affective disorders are a set of psychiatric disorders, all of which are characterized by mood changes. Depression, bipolar disorder, and other affective disorders are depression, bipolar disorder, and...
3. Contribution: Example
3. Contribution: Example
3. Contribution: Example
3. Contribution: Example
4. Evaluation
4. Evaluation

- Due to the relevance of showing recommended queries in real time, we have performed an analysis of the execution times obtained with selected real queries from a database in order to evaluate our tool in terms of computational efficiency and performance.
4. Evaluation

- The database of queries we have employed in our experiments was published by the AOL search engine in 2006. This database contains about 37 million queries of 657,000 unique users, obtained during a period of three months (from 1 March 2006 to May 31, 2006).

Fig. x. AOL queries database categories.
Source: (Steven Beitzel, et al., 2005)*

4. Evaluation

<table>
<thead>
<tr>
<th>Terms in Query</th>
<th>Max.</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>29502</td>
<td>21731</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>14939</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6082</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1925</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>962</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Some statistics of the first 100,000 queries of the AOL query database.

<table>
<thead>
<tr>
<th>Terms in Query</th>
<th>Max.</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>26149</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>18750</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>11195</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6415</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6888</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Some statistics of the first 100,000 queries, after pre-processing.
4. Evaluation

**Fig. 9.** Average execution time based on the number of terms.
4. Evaluation

Fig. 10. Average execution time.
5. Conclusions
5. Conclusions

- The use of personalization techniques by WSEs is a promising way to improve the quality of searches, but pose serious concerns to user privacy.

- Although there exist few proposals based on query generalization, no solution has been designed nor developed that brings this principle into practice and is intended for end-users.

- Our tool protects each individual query independently and, as such, does not make any assumption on the ability of the WSE to track all them.

- Experimental results show that our tool is able to categorize complex searches in real time while users types their queries, without affecting the performance of the system.
6. Future work
6. Future work

- Evaluate the proposed tool further with real users.
- Attempt to determine the utility loss incurred by the generalized queries.
- Extend the tool to more browsers.
- Use multilingual databases instead of WordNet.
Thanks for your attention