Expert Systems with Applications 40 (2013) 2287–2296

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Improving large-scale search engines with semantic annotations

Damaris Fuentes-Lorenzo *, Norberto Fernández, Jesús A. Fisteus, Luis Sánchez

Carlos III University, Av. de la Universidad 30, 28911 Madrid, Spain

ARTICLE INFO

Keywords:
Semantic annotation
Semantic search
Wikipedia
Click-through data
Ranking algorithm
Collaborative tagging

ABSTRACT

Traditional search engines have become the most useful tools to search the World Wide Web. Even though they are good for certain search tasks, they may be less effective for others, such as satisfying ambiguous or synonym queries. In this paper, we propose an algorithm that, with the help of Wikipedia and collaborative semantic annotations, improves the quality of web search engines in the ranking of returned results. Our work is supported by (1) the logs generated after query searching, (2) semantic annotations of queries and (3) semantic annotations of web pages. The algorithm makes use of this information to elaborate an appropriate ranking. To validate our approach we have implemented a system that can apply the algorithm to a particular search engine. Evaluation results show that the number of relevant web resources obtained after executing a query with the algorithm is higher than the one obtained without it.

1. Introduction

Since its creation in 1989, the World Wide Web has become into one of the largest public information sources: some reports have pointed out that the indexable web contains more than 25.21 billion pages (Worldwidewebsize.com, 2012).

Though the large amount of information available on the Web is one of its main positive aspects, it also has a negative side: the vast number of pages makes difficult for users to find the information they are looking for (Bates and Anderson, 2002). Users need appropriate tools to help them in order to take full advantage of the information stored. Web search engines, such as Google or Yahoo, are well known examples of this kind of tools. The effectiveness perceived by users and their easiness of use have made engines to achieve positive results in the web market; however, current web search engines still have some limitations.

First, their retrieval model is mainly based on looking whether keywords in a user query match the content of web documents. For instance, the search engine may omit other documents referred to the same semantic information if these documents have not the same keywords of the query. Another case where the keyword-matching approach is problematic, is that of ambiguous queries; the shorter the queries, the smaller the context to disambiguate them. Taking into account that, according to Experian Hitwise (2011), the most frequent query lengths are 1 or 2 words, this problem can affect to a large number of queries.

In order to address this problem, semantic search (Baeza-Yates, Ciaramita, Mika, & Zaragoza, 2008; Fernández et al., 2011) has been proposed as an alternative to traditional keyword-based search, both in academia (Fernández et al., 2011; Kiryakov, Popov, Terziev, Manov, & Ognyanoff, 2004) and industry (Hakia, SenseBot). In general, these approaches design and build new systems from scratch; thus, they do not exploit the information indexed and functionalities already presented in traditional web search engines.

Second, navigating in a search for relevant information on the Web is one of the most lonely and time-consuming tasks (Jung, 2005). The performance of the overall searching process can be enhanced if users collaborate in this task. Given a query, previous users' opinions and interests about a similar query could improve the results of these algorithms.

In this paper, we develop a ranking algorithm to re-order the results obtained from a large-scale, traditional web search engine to obtain more relevant web pages on top of the rank. Our approach uses semantic search techniques but, instead of building a new information retrieval system, we elaborate a semantic layer which is set on top of current search engines.

To achieve its goal, the proposed algorithm relies on:

- Semantic annotations, to unambiguously tag queries and target documents. These annotations use Wikipedia as a reliable semi-structured encyclopedic source.
- The collaborative usage of information that users generate while searching, obtained through explicit relevance feedback techniques.

The rest of the paper is organized as follows. Section 2 elaborates a summary with the most important techniques related to

* Corresponding author. Tel.: +34 916248437.
E-mail addresses: dfuentes@it.uc3m.es (D. Fuentes-Lorenzo), berto@it.uc3m.es (N. Fernández), jaf@it.uc3m.es (J.A. Fisteus), luiss@it.uc3m.es (L. Sánchez).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.10.042

our solution and existing ranking algorithms in previous works. Section 3 explains the details of the ranking algorithm proposed in this paper. Section 4 introduces the implemented system and gives an overview of its general functionality flow. Section 5 discusses the experimental results taken and, finally, Section 6 closes this article with concluding remarks and future work lines.

2. Related work

In this section, a summary of the most relevant approaches related to the proposal in this paper is exposed. As indicated in the introductory section, the main goal of this work is to develop a ranking algorithm that exploits metadata, in the form of semantic annotations gathered from users, to re-order the results provided by large-scale, traditional web search engines. Taking this into account, the related work section has been structured into two main subsections: On the one hand, Section 2.1 is devoted to provide some background context on different techniques designed to acquire information from users to be exploited for information retrieval purposes. On the other hand, Section 2.2 is centered in analyzing other ranking algorithms and semantic search approaches already available in the state of the art.

2.1. Exploiting user information in search process

In last years, the idea of exploiting information obtained from users to improve results provided by search engines has been explored in different ways. The following subsections briefly describe some of the different available techniques.

2.1.1. Click-through data

Using the data that a user search session produces took relevance approximately one decade ago. Basically, the click-through data obtained from a search engine is composed of the queries users execute and the links users click on the ranked results presented, also called 'implicit feedback'.

In this area, Hansen and Shriver (2001) and Joachims (2002) are worth mentioning. The former proposed narrowing search results by observing the browsing patterns of users during search tasks. In the latter, Joachims used navigation data to improve the results in search engines by using classification techniques in conjunction with the click-through data of a meta-search engine. Outcomes showed that the results obtained improved retrieval quality with respect to using the engine alone.

However, this approach makes assumptions that may have a negative impact in the obtained results. For example, the approach considers that the mere selection of a result implies this result is relevant to the query, which may not be the case. Data collected with this technique should be pre-processed before using it directly in order to improve results in a trusty manner.

2.1.2. Collaborative filtering and tagging

Collaborative filtering is the process by which users help others to perform filtering tasks by annotating their reactions to the documents they read. For example, users can annotate whether they find a particular document interesting or not. Even though this task of scoring information has grown in popularity in the last years with the so-called web 2.0, there already exist collaborative filtering works dated in 1992, such as Tapestry (Goldberg, Nichols, Oki, & Terry, 1992), an email organizer system, or in 1994 with GroupLens (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994), a system for searching news articles.

On the other hand, collaborative tagging is the process by which many users add metadata in the form of keywords to organize their content. Some well-known applications which allow the use of this technique are del.icio.us (www.delicious.com), where the tagged resources are website bookmarks, or Flickr (http://www.flickr.com), where the target resources are photographs. Collaborative tagging can be seen as a subset of collaborative filtering, where the user reactions in this case are the words they relate to the documents.

Another approach to web resource tagging consists of exploiting user queries, obtained from search engine logs, as if they were textual tags. The terms used in a query can be considered as potential descriptions or tags of the URLs in the navigation data set obtained after the query execution. This is exactly the conclusion of several works, such as (Krause, Jäschke, Hotho, & Stumme, 2008), where it is shown that the clicking behavior of search engine users (the click-through data seen in the previous section), based on the presented search results, and the tagging behavior of social bookmarking users was driven by similar dynamics. Some of these works call the resulting network of query keywords a *logsonomy*.

One of the greatest benefits of tagging applications is the fact that there is not any predefined vocabulary for the tagging activity. Firstly, this provides users with freedom to choose any keyword to use. Secondly, no expert knowledge is needed to define a domain vocabulary.

However, as explained in many works such as Golder and Huberman (2006), Motta and Specia (2007) or Wu, Zhang, and Yu (2006), this apparent advantage leads to a number of limitations and weaknesses when using tags for information retrieval and search. Most of these problems can be reduced in the following ones:

- Ambiguity/Polysemy: A polysemous word has more than one meaning. When searching for documents with a word such as "play", related to a theatre piece, a search engine can return unrelated results such as, for example, a set of games for children.
- Lack of synonym relations: Words are synonymous if they have the same meaning. Words "irritated" and "annoyed" are very closely related; however, when searching for one of these words, found items will hardly contain the other word.
- Lack of consensus: The lack of consensus in the use of tags, especially as granularity is concerned, makes a traditional tagging system quite inefficient. To describe a particular item, different users may consider terms at different levels of generality/specificity. For example, a user can tag a photograph as "bird", and another user can tag the same photo as "eagle".

Some works, such as Heymann, Koutrika, and Garcia-Molina (2008), have already demonstrated, therefore, that social tagging does not improve web search.

The usage of formal annotation vocabularies, instead of plain text tags, may alleviate the aforementioned problems (Passant & Laublet, 2008). Ontologies are a possible type of formal vocabulary that may be exploited with this purpose. Appearing first in the Philosophy art, ontologies are grasped by the Artificial-Intelligence experts to represent needed parts of a particular domain (Borst, 1997; Gruber, 1993). Later on, the semantic-web community started to make use of them to formalize the concepts, relations and rules of a domain of knowledge (Berners-Lee, Hendler, & Lassila, 2001).

However, ontologies still lack of mass support, in contrast with the frequent use of tags in any web 2.0 applications. The development of any ontology is still an activity addressed to knowledge experts, whereas users with no expertise can be involved in the creation of sets of tags with no effort.

For these reasons, from several years up to now, Wikipedia is being presented as a good alternative to semantically annotate resources in applications where word sense disambiguation is