The Journal of Systems and Software 85 (2012) 1577–1587

Contents lists available at SciVerse ScienceDirect

The Journal of Systems and Software

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

Dynamic refinement of search engines results utilizing the user intervention

Dimitris Antoniou*, Yannis Plegas, Athanasios Tsakalidis, Giannis Tzimas, Emmanouil Viennas

University of Patras, Computer Engineering and Informatics Department, 26500 Patras, Greece

ARTICLE INFO

Article history:
Received 28 March 2011
Received in revised form 27 January 2012
Accepted 28 January 2012
Available online 25 February 2012

Keywords:
Search engines
Post-ranking
Semantic matching
Personalization
Relevant feedback
Dynamic refinement

ABSTRACT

Nowadays, modern search engines quite satisfactorily answer users' queries, but the top results returned are not always relevant to the data the user is actually looking for. Hence, considerable efforts are made by search engines in order to rank the most relevant to the query results at the top. This work addresses the above problem and improves the performance of a search engine, especially when it comes to queries which have for example twofold meanings. The matter which the user is interested in is identified based on the results that he/she chooses, and then the most relevant ones are ranked higher. In addition, the results are recognized not only as text but also as semantic entities, which contain various semantic features. The semantic relation between results and text coverage are used as the main tool to achieve an optimized ranking, as opposed to other research papers so far. As a result, a new meta search application is developed, which, given a set of terms, combines Google results and then reorganizes (re-ranks) them based on the disambiguation offered by user clicks. In particular, after a ranking is achieved, the user makes a choice (click), the ranking is updated and the process is repeated. In order to prove our claims, apart from the description of the algorithm for refining the ranking of results, a web application has been developed, which was used to test the effectiveness of the system proposed.

1. Introduction

The tremendous growth of the World Wide Web the last decades has promoted searching as one of the most prominent issues in the field of web research. In this work, a real-time improvement on a search engine results page (SERP) is attempted. SERP is the listing of web pages returned by a search engine in response to a keyword query. The results normally include a list of web pages with titles, a link to the page, and web page snippets, a short description showing where the keywords have matched content within the page. A SERP may refer to a single page of links returned, or to the set of all links returned for a search query.

Till now, in general, search engines involve the zoning information and frequency of keywords on a web page, in order to rank the search results. Pages with the search terms appearing in the HTML title tag are often assumed to be more relevant than others to the topic. Search engines will also check to see if the search keywords appear near the top of a web page, such as in the headline or in the first few paragraphs of text. They assume that any page relevant to the topic will mention those words right from the beginning. In general, zoning is the procedure to parse an item into logical subdivisions that have meaning to the user. It is used to increase the precision of a search and optimize the display by allowing searches to be restricted to a specific zone.

Frequency is the other major factor in how search engines determine relevancy. A search engine will analyze how often keywords appear in relation to other words in a web page. Those with a higher frequency are often deemed more relevant than other web pages. Link analysis is also used by several engines as part of their ranking algorithm, most notably Google. The belief is that link analysis gives search engines a useful and unadulterated method to determine which pages are good for particular topics. User logs are used as well in order to make use of cross-references between the users' queries and the documents that the users have chosen to read. Conclusively, a key factor in the success of Web search engines is their ability to rapidly find good quality results to queries.

Our work suggests a new technique for the refinement of search results against the results that the user chooses. In cases of polysemy, the keyword (or keywords) used do not have a clear but a general meaning and do not specify the matter which the user searches for effectively. For example, in the case of a query for the word "rockets", the user could be interested in the basketball team "rockets" and not missiles with their actual meaning. Apparently in the first ten results irrelevant results will also appear. By selecting the first result, the user offers the semantic information according to which, he/she is interested in the team "rockets". We extend the

* Corresponding author.
E-mail addresses: antonid@ceid.upatras.gr (D. Antoniou), plegas@ceid.upatras.gr (Y. Plegas), tsak@cti.gr (A. Tsakalidis), tzimas@cti.gr (G. Tzimas), biennas@ceid.upatras.gr (E. Viennas).

0164-1212/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jss.2012.01.049

information kept per user from a simple set of keywords, to a set of information that contains semantic information about a page that relates to what the user is looking for. This enriched set of information is used in refining the ranking of the results returned by the search engine, in order to rank higher the results that really interest the user.

Initially, the information the user is searching for is determined only by keywords entered into the search engine. The basic idea of our algorithm is to extend this information set by exploiting the user choice of result in order to gather additional information about the topic he/she is looking for. For the implementation of this idea, there is no need for information gathered from previous queries or importing additional data. This fact makes the enhancement of this idea in search engines simple and efficient. The information set consists of the initial keywords and parts of the chosen results. Each result is analyzed semantically as well, and its similarity with the other results is the criterion that determines the position it receives at the final ranking. Using semantic similarity and coverage of text between the results that users select and the remaining results, a high rank for the results that best match the subject that user is interested in is achieved.

The rest of the paper is organized as follows: Section 2 presents the related work, while Section 3 presents the architecture of the system. Next, in Section 4 we present the proposed algorithm, as well as the algorithms used to implement semantic similarity and text coverage. In Section 5, an experimental paradigm of our system is presented and Section 6 shows the results of our experiments. Finally, Section 7 concludes the paper and Section 8 provides future steps.

2. Related work

In the last decade many search techniques have been proposed (Henzinger, 2003; Ozsoyoglu, 2003) with a large number of variants. Two of the most comprehensive and theoretically well-established are those presented in Brin and Page (1998) and Kleinberg (1999). In the most popular search engines (Google and Yahoo), to each web page a category is assigned. According to Boyan et al. (1996) this classification facilitates searches, since most users have a propensity to follow pages with a high score even if they are not related to the search topic.

Several techniques have been proposed and used in order to improve the results returned by search engines. Focused crawling is a technique for filtering non-relevant documents. Focused crawlers (Diligenti et al., 2000; Chakrabarti et al., 1998, 1999, 2002; De Bra et al., 1994; Herscovici et al., 1998) overlook pieces of the Web graph and are assembled only in documents that are considered relative according to a predetermined hierarchy of categories. The relevance is achieved by using text categorization strategies. Special attention should be given to Diligenti et al. (2000) and Chakrabarti et al. (1998, 1999), since some of the ideas used in techniques for the construction of crawlers are directed in thematic categories.

In Chakrabarti et al. (1999) a focused crawler is presented that selectively chooses pages (documents) relevant to a predefined set of topics. The topics are not specified using keywords, but using exemplary documents found in a thematic classification. The human input is provided either as a specific thematic node in a standard classification (such as Yahoo and Open Directory Project), or as a set of page addresses to be used as home pages for the search. The proposed development is discussed in Herscovici et al. (1998), where the basic idea is to use two (instead of one) classifiers: one classifier as a report point and a new one, which calculates a page's relevance to a range of subjects, using DOM features pages associated with it. In Diligenti et al. (2000) the idea of using the content in which lies the information the user searches for, is explored. The proposed crawler exploits the ability of search engines return pages that link to a document, which are set to a certain distance from the original document. This representation is used then to guide a set of classifiers to identify and assign documents into different categories based on the expected distance selected from the original document. During the navigation the classifiers estimate how close the target document (original) is to the retrieved document.

In general, refining the results returned by a search engine is a topic where several techniques have been proposed in the past recent years, in order to improve the user's experience. Employing clustering is one way to address the problem. In Zhang and Liu (2004) a new clustering algorithm is proposed that classifies the results of a query from a search engine into subgroups and assigns each group a short series of keywords together with some statistics data. Then, the user may look into the group with the keywords that he/she finds interesting. Compared to the approaches available in the literature, this algorithm does not require the number of groups as a prior knowledge.

There are very few studies investigating the nature of interaction using real world clustering search engines. Koshman et al. (2006) reports on results from a transaction log analysis of Vivisimo.com, which is a Web meta-search engine that dynamically clusters users' search results. The focus of Vivisimo's research thus far has been the concept of clustering search results based on topic: for example, dividing the results of a search for "cell" into groups like "biology" battery, and "prison".

Another algorithm for clustering search results is proposed in Mecca et al. (2007). Differently from many other clustering systems that have been recently proposed as a post-processing step for Web search engines, this system is not based on phrase analysis inside snippets. Instead it uses Latent Semantic Indexing on the whole document content. A main contribution of the paper is a novel strategy – called Dynamic SVD Clustering – to discover the optimal number of singular values to be used for clustering purposes.

The problem of clustering the refinements of a user search query is addressed also in the recent work presented in Sadikov et al. (2010). Here, the clusters computed by the proposed algorithm can be used to improve the selection and placement of the query suggestions proposed by a search engine, and can also serve to summarize the different aspects of information relevant to the original user query. The algorithm clusters refinements based on their likely underlying user intents by combining document click and session co-occurrence information.

Some of the most important issues in web search engine optimization through past query mining are covered in Silvestri (2010). The primary focus of this survey is on introducing to the discipline of query mining by analyzing the basic algorithms and techniques that are used to extract useful knowledge. It is show how search applications may benefit from this kind of analysis by analyzing popular applications of query log mining and their influence on user experience.

Another way to address the problem is to exploit past user queries and preferences. Environments that use meta-search, implement strategies that match users' queries to collections results (Meng et al., 2002; Howe and Dreilinger, 1997). In such transactions, user profiles are created, either during the search process or in search environments such as Letizia (Lieberman, 1995) and CiteSeer (Bollacker et al., 1999). These profiles help users by offering the results ranked according to their previous behavior (pages already navigated). In Zhuang and Cucerzan (2006), the authors address the problem of frequently occurring with underspecified user queries: the top-ranked results for such queries may not contain documents relevant to the user's search intent, and also deal with the fact that fresh and relevant pages may not get