Combining Statistical and Lexical Processing for Query Refinement

Paraskevi Tzekou, Lefteris Kozanidis, Dimitris Christodoulakis, Sofia Stamou

Computer Engineering and Informatics Department, Patras University, 26500, GREECE

ABSTRACT

The most popular way for finding information on the web is go to a search engine, type a natural language query and obtain a set of retrieved results in response to that query. Despite the simplicity of web searching and the numerous data sources that exist on the web, information seekers do not always find what they are looking for. This is basically because users fail to come up with a query that is both representative of their information need and understandable by the engine. One way to overcome this problem is to equip a search engine with a query refinement module, which helps the user issue more comprehensive queries by suggesting alternative terms in the query formulation process. Query refinement has long been addressed in the literature as a technique for improving retrieval performance. In this paper, we address the query improvement process from a personalized search perspective and we build upon previous query refinement techniques. In particular, we propose a query expansion module which performs query refinement based on a combined analysis of the user's past searches and the semantics of typed queries. Considering both statistical and semantic data in the information selection process significantly improves retrieval results as our experimental study demonstrates.

Introduction

Query expansion has long been accounted as a retrieval improvement technique. In brief, query expansion is the process of appending relevant terms to an issued query with the expectation that those terms will contribute to the retrieval of better search results. Query expansion, although perceived as a straightforward process, it has inherent difficulties most of which are introduced by query ambiguities. In other words, it is essential that before a query expansion module actually enriches a query it has successfully resolved any query term ambiguities. Query disambiguation is a vital prerequisite in the expansion process. For query sense disambiguation many approaches have been proposed in the literature, a great number of which attempt query sense resolution based on the users' search interests. In particular, these approaches suggest the exploitation of the user's search profile for query disambiguation and the subsequent refinement of the query with terms that match the given profile. User profile construction relies on either explicit information supplied by the user himself, or on implicit data collected from the user's previous searches. The above techniques are widely known as explicit (Chen, 2001) and implicit user feedback (Joachims, 2002) respectively.

In this paper, we explore the implicit user feedback approach and we propose a novel query refinement technique which on the one hand identifies a user's search profile through the analysis of his past clickthrough data and on the other it associates the semantics of the documents viewed by the user in response to previous searches and the information encoded in the Greek WordNet in order to select a set of candidate terms as alternative formulations to future queries of the user. More specifically, our technique incorporates three distinct yet complementary components each of which contributes to the query refinement process. These components are:

- A mechanism that monitors the user's behavior while interacting with a search engine. This mechanism collects data about the queries that a user submits, the pages he receives in response to those queries, the pages he views out of the total retrieved data, the time he spends on each page as well as the frequency in which he revisits those pages. The above information is processed and analyzed for concluding on the pages that the user deems relevant to his search request.
- A search session recording module, which groups the queries submitted by the same user into time clusters in order to record whether the user's interests change over time as he interacts with the search engine and how these changes affect the user's initial search intentions. The data collected by our session monitoring module is explored for detecting shifting in the user's search intentions.
- A semantic similarity formula, which examines the semantic associations between the user's query terms and the relevant documents' indexing keywords in order to compute relevance scores between the issued queries and the documents viewed for those queries. Query-document relevance is a valuable indicator for inferring the user's interests. Note that semantic associations between query and document terms are determined on the basis of the information encoded in Greek WordNet.

Based on a combined analysis of the data collected by the above modules, our system performs query refinement by suggesting to the user a set of alternative query terms which not only relate to the query at hand but which are also relevant to the user's search profile. Alternative query terms are borrowed from the indexing keywords of the documents that are perceived by the user to be relevant to his search profile. However, to keep the number of alternative query wordings balanced, we filter the keywords of the relevant documents through the use of Greek WordNet and we enforce only semantically related terms participate in the refinement process.

To demonstrate the potential that our query refinement technique has in a real world setting we implemented a prototype query refinement system which incorporates our modules and performs query improvement in the way described above. Preliminary experimental results indicate that our approach has a significant potential in selecting alternative query formulations; an argument that is implied by the improved retrieval results of our system. The remainder of the paper is organized as follows. We first review related work in the area of both query refinement and user relevance feedback and we see how the one complements the other in the retrieval enhancement process. In section 3, we present our query refinement module and we give detailed descriptions of the data that our module explores before supplying the user with alternative query terms. In Section 4, we present our experimental study and we discuss obtained results. Finally, we summarize the contribution of our approach in Section 5, which concludes the paper.

Related Work

Query refinement is a broad field of active research. There are three predominant approaches associated with the query enhancement process, namely (i) explicit user feedback (Pazzani et al., 1996), (ii) implicit user feedback (Joachims, 2002; Sehn et al., 2005), and (iii) thesaurus-based query expansion (Mandala et al., 1999). In detail, the query expansion approach normally relies on the use of thesauri for selecting terms that are synonyms to the query terms. Query synonymous terms are appended to the initial query and a new search for the expanded query is performed. On the other hand, in the explicit feedback approach, query refinement relies on the information that is explicitly provided by the user and which describes his information needs, e.g. in the commercial system MyYahoo. The main constraint of this approach lies on the users' reluctance to inform a retrieval system about their interests. To overcome this difficulty, many of the query refinement techniques monitor the user's search behavior through the examination of search sessions, proxy logs, clickthrough history, etc. and based on the analysis of these data they create implicit user search profiles. Inferred user profiles are employed by they retrieval system, which refines the user formulated queries by associating them with information about the user's search profile. A more recent approach in the implicit feedback method concerns the examination of the user's previous searches in terms of entire search sessions rather that isolated queries, in an attempt to learn a generic user profile that is subsequently applied for refining future searches (Chen, 2001). In this paper we build upon this final approach and we propose an efficient way for inferring the user's profile through the analysis of his past search sessions. The learned profile is then supplied to a query refinement mechanism, described next.

Query Refinement: Our Approach

Our query refinement method operates on the basis of the knowledge of a user's search profile and the association between the user's interests and the documents that he liked in previous searches. To obtain the user profile knowledge we employ the implicit user feedback paradigm and we identify from a user's past search activity, the documents that the user deems related to his information interests. Relevant documents participate in the refinement process by supplying to our module their indexing keywords, as the later have been determined through the *tf*idf* formula (Salton and Buckley, 1988). Terms used to describe the inferred profiles as well as the indexing keywords of the relevant documents are mapped to the Greek WordNet hierarchies, where we compute their semantic similarity values. Semantic similarity is determined by the distance of the above concepts in the Greek WordNet graph, following the implementation of (Dao, 2005). Document keywords that associate to the user profile are selected as candidate terms for alternative query formulation. In the following paragraphs we elaborate on the main components of our query refinement technique, illustrated in Figure 1.

Inferring the User's Interests

The first step towards the identification of a user's profile is to detect the documents that the user deems related to his information needs. To achieve that, we have built a mechanism that monitors the user's activity while interacting with a search engine, as well as a module that identifies the search sessions initiated by that user. Based on a combined analysis of the data supplied by these modules, we can efficiently track which documents are of interest to the user for his specific search requests. Interesting documents are determined on the basis of the time the user spent reading them, the number of

times we revisited them, their inclusion or not to his favorites, and so forth. Due to space constraints, we retain the detailed description of the user interests' inference process for an extended version of this paper. Having identified interesting documents, we explore the data collected from the user's search sessions and we associate the indexing keywords of the interesting documents to the query terms appearing within the respective search session. Relevant documents' keywords that highly correlate to the user issued keywords are employed by our query refinement framework, which offers them as alternative formulations for the user's future queries. The details of the query refinement process are given next.

Improving the User Queries

As mentioned earlier, our query refinement module selects the terms to suggest the user for improved query formulations based on the content of the documents that the user found interesting for the searches he had previously conducted on a similar information interest. Having determined a set of candidate terms for query reformulation, our technique relies on the information encoded in Greek WordNet for filtering the list of candidate wordings for improving the query. Filtering accounts to the elimination of the terms (from the list of candidate query alternatives) that exhibit moderate or low semantic similarity to the initial query terms in the WordNet graph. The degree of semantic similarity (low, moderate or high) has been empirically fixed after experimenting with various similarity thresholds. At the end of the filtering process, our technique takes the relevant document's indexing keywords that are highly correlated to the issued queries and suggests them to the user for improving his future queries on a related subject area. In the following section we experimentally study the effectiveness of our query refinement technique in assisting the user find the information sought.

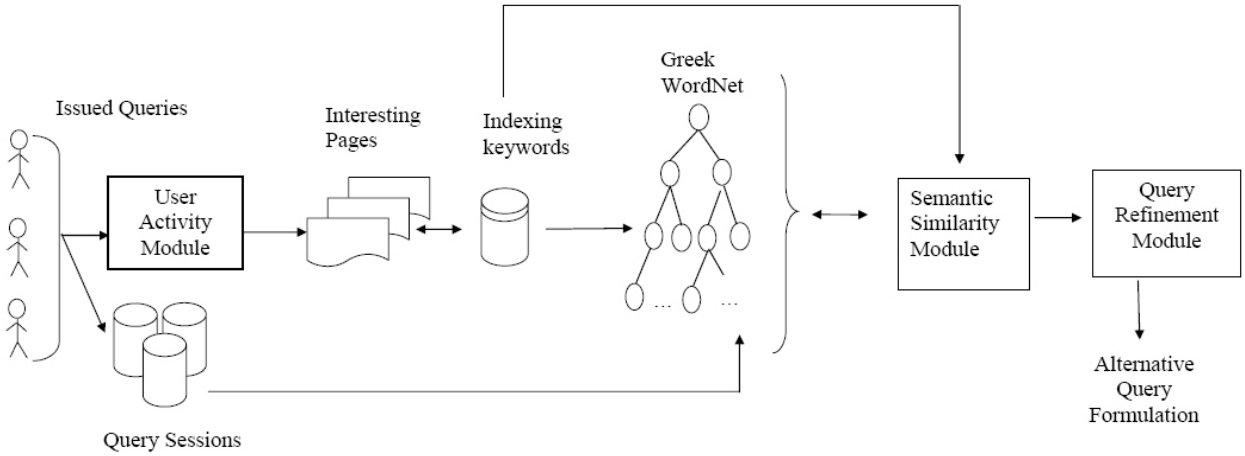

Figure 1: The Query Refinement Process

Experimental Setup

To evaluate the effectiveness of our query refinement technique in selecting improved query formulations, we experimentally studied the effect that our refined queries have on retrieval performance. For our study we implemented a prototype query refinement system and we used a number of seven different query sessions for borrowing a set of our experimental queries. Query sessions were gathered from real data that our query session module recorded. Thereafter, we turned to the user activity monitoring mechanism for detecting which of the documents retrieved for those queries were actually viewed. Based on the user's clickthrough activity we computed a query relevance score for each of the documents viewed. Thereafter, we supplied our refinement system with the above data and based on the information encoded in Greek WordNet and by following the process described above our system returned a list of alternative formulations for each of the queries participating in every search session. Finally, we executed the improved queries that were suggested by our system and we monitored the user's clickthrough activity on the new data retrieved. Through the analysis of the user's activity, we computed a new query relevance score for each of the documents that were viewed by the user in response to the improved queries. At the end, we compared average relevance scores between the issued queries (simple and refined) and their viewed documents in order to gain perceptible evidence of our system's performance. Experimental results are given in the next section.

Experimental Results
Obtained results demonstrate the potential our approach has in assisting the user find useful information. Specifically, Table 1 summarizes average relevance scores for each of the query sessions before and after query refinement.

Sessions	Average relevance before refinement	Average relevance after refinement
S_1	0,643	0,733
S_2	0,587	0,814
S_3	0,601	0,797
S_4	0,608	0,759
S_5	0,573	0,687
S_6	0,634	0,774
S_7	0,593	0,730

Table 1: Average query-document relevance scores

As we can see from the figures in Table 1, our query refinement method yields significant improvements to the relevance of the retrieved results. Note that by relevant results we refer to the documents that the user spent some time reading, the documents that the user bookmarked and/or the documents that the user revisited in response to future similar information requests. Figure 2 gives a clearer view of the obtained results, which demonstrate that our query refinement approach has a great potential in helping users formulate queries that are both expressive of their information needs and understandable by the search engine's mechanisms.

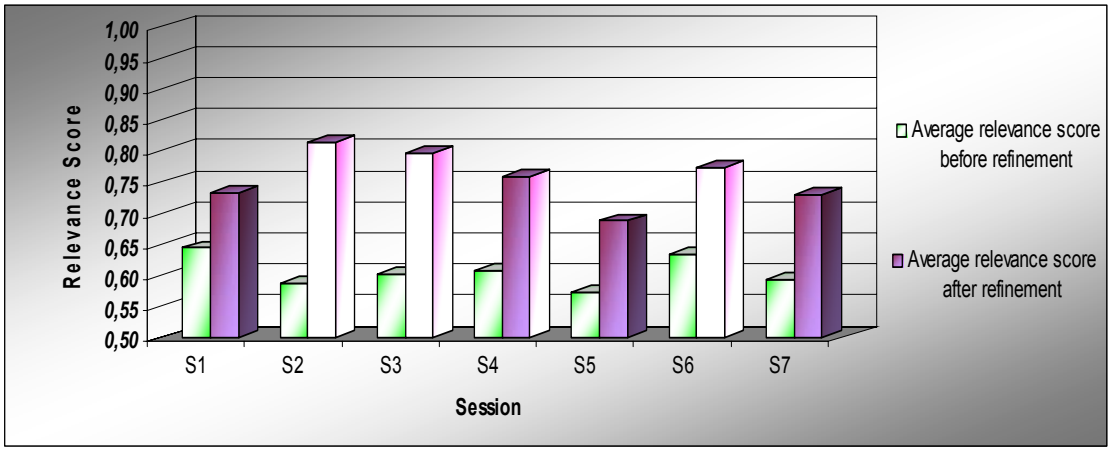

Figure 2: Average query-document relevance before and after query refinement

Concluding Remarks
In this paper we addressed the query refinement challenge and we proposed a novel technique for selecting alternative expressions for query representation. In particular, we proposed the exploitation of the user's past searches and the relevant document semantics for determining which terms within the relevant documents are informative of the user's initial search intention. Informative terms are subsequently employed by our expansion module which computes their semantic closeness in the Greek WordNet and determines which of those terms are good candidates for reformulating the initial query.
Preliminary experimental evaluation of our technique demonstrates that our query refinement method has a significant potential in improving the user search experience. Although, further experimentation is needed before we generalize our findings, nevertheless we believe that our approach can pave the ground for more elaborate approaches in the query

refinement process, especially when it comes to less studies languages.

References

Chen, D.H. 2001. Personalized Spiders for Web Search and Analysis. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*.

Dao, T. 2005. An Improvement on Capturing Similarity between Strings. Available at http://www.codeproject.com/cs/algorithms/improvestringsimilarity.asp

Joachims, T. 2002. Optimizing Search Engines Using Clickthrough Data. In *Proceedings of he SIGKDD Conference*.

Mandala, R., Takenobu, T. and Tanaka, H. 1999. Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion. In *Proceedings of the 22nd ACM SIGIR Conference*.

Pazzani, M., Muramatsu, J. and Billsus, D. 1996. Syskill & Webert: Identifying Interesting Web Sites. In *Proceedings of the 13th International Conference on Artificial Intelligence*, pp. 54-61.

Salton, G. and Buckley, C. 1988. Term Weighting Approaches in Automatic Text Retrieval. In Information Processing Management, Vol. 24(5), pp. 513-515.

Shen, X., Tan, B. and Zhai, C. 2005. ContextSensitive Information Retrieval Using Implicit Feedback. In *Proceedings of the 28th ACM SIGIR Conference*.