Query Reformulation for Task-Oriented Web Searches

Nikos Kirtsis[*], Sofia Stamou[*+]
[*]*Computer Engineering and Informatics Department, Patras University, 26500, GREECE*
[+]*Department of Archives and Library Science, Ionian University, 49100, GREECE*
{kirtsis, stamou}@ceid.upatras.gr

Abstract

Web searches are driven by information needs and intend the accomplishment of specific tasks. Information needs are determined by the topical subject of queries, i.e. what we search for, while tasks are determined by the user motives that induce the submission of queries, i.e. why we search. Though there exist numerous studies on how to assist searchers specify queries that are expressive of their underlying information needs, little has been done to help searchers specify queries that describe the tasks they pursue via their searches. In this paper we propose a query reformulation method to empower task-oriented web searches. Given a query, our method starts with the identification of terms that could serve as descriptors of the potential search tasks the query represents. Based on the identified search task descriptive terms, it generates query reformulations that explicitly verbalize the possible query tasks. Query reformulations are presented to the user in order to select the one that best suits her search intention.

1. Introduction

With the proliferation of both online content and information seekers, there has been a shift of interest from the retrieval of query-relevant documents to the retrieval of information that relates to the task the user is trying to accomplish via search [11]. This is because the tasks that lead people to engage in information seeking behavior affect their judgments of usefulness of the retrieved results. However, most web queries do not explicitly verbalize the user intended search tasks. To that end, a number of search engines have recently been equipped with enhanced search services[1] to assist users formulate better queries. Existing query reformulation methods mainly focus on the user information needs hidden behind queries and are less concerned about the tasks pursued via search. Therefore, the query reformulations they suggest improve the vocabulary and the syntax of the initial query by enriching it with terms that mainly relate to the query topic and not the search task.

In this paper, we focus on the problem of query reformulation for task-oriented searches. We base reformulations on an internal repository of search task descriptive terms that we harvested from the subtopics of the TREC 2009 web track topics [3]. The search tasks examined in our study are *navigational* and *informational* queries as these are defined in [2] and [8]. We focus on these categories of search tasks firstly because every subtopic description in TREC is categorized as being either navigational (*nav*) or informational (*inf*) and secondly because there is a lack of consensus on the types of search tasks other categories pursue.

Given an input query *q*, our objective is to generate a reformulated query *q*' for each of the considered search tasks, i.e. *q*'(*nav*) to denote that the user task behind q is to reach a particular site and *q*'(*inf*) to denote that the user task behind *q* is to gather information from one or more pages. Query reformulations are generated based on a repository of search task descriptive keywords, from which we extract the terms to be added to the input query so as to make it expressive of its underlying search tasks. The generated query reformulations are displayed to the user in order to select the reformulation that better expresses her underlying search intention.

The remainder of the paper is organized as follows. We begin our discussion with a brief overview of previous research that addresses the challenge of empowering task-oriented web searches. In Section 3, we introduce our query reformulation method, emphasizing on the construction of terminological repositories of task-descriptive search terms. In Section 4, we present the results of a preliminary experimental study we carried out in order to evaluate the retrieval performance of the reformulated queries. In Section 5, we conclude the paper and outline our plans for future work.

[1] As examples consider Google Instant (http://www.google.com/instant/) and the Related Searches option of Bing (http://www.bing.com/).

2. Related Work

Previous research falls in two main categories: patters of search analysis and query refinement. With respect to the former, a number of researchers proposed methods for classifying search queries according to the type of interaction they intend on the retrieved results. Based on the initial classification of queries as *navigational*, *informational* and *transactional* proposed in [2], researchers studied ways to automatically identify the user intend hidden behind queries [5] [7] [8]. Most of the studies concentrate on the analysis of web transaction logs for associating user queries with particular search intentions and information needs as well as for inferring the user satisfaction from retrieval performance. The commonality in the above approaches is that they try to encapsulate the user intentions into retrieval ranking algorithms in an attempt to personalize search results according to the user needs. Although useful, the above attempts are primarily oriented towards improving retrieval performance and are less concerned with assisting the users improve their querying skills.

To assist users specify queries that are expressive of their underlying information needs, researchers have suggested various query refinement methods. Existing approaches to query refinement span two main categories, namely query expansion with semantically related or equivalent terms [6] and query reformulation based on implicit user feedback [1] [12]. Both approaches rely on the semantics of the user typed queries to come up with terms that once employed as (additional) query terms they would help the user find the information sought easier. Again, most of the works in this direction aim at improving retrieval performance by suggesting queries that are indicative of the user search needs but not necessarily of the user search tasks.

Our work conflates the objectives of search pattern analysis and query improvement into a single goal; that of reformulating web queries into task-descriptive web searches. The motive for our work is to assist information seekers specify queries that reflect both the topic and the intention of their searches in the hope of improving their querying skills and consequently their satisfaction from retrieval performance. To fulfill our study objective, we propose a method for building a repository of task descriptive search terms that could be utilized for reformulating web queries. In the following paragraphs, we discuss the details of the term selection process and we demonstrate via a preliminary experiment the effectiveness of our method in improving the users' search experience.

32. Query Reformulation

3.1. Identifying Search Task-Descriptive Terms

To build a repository of search task descriptive terms, we explored the TREC 2009 web track topics. Every topic in the collection has a main *theme* that describes what the search is about and is verbalized via a <query> element. Moreover, every topic has a set of subtopics to cover the different aspects of the query and every subtopic is categorized as being either navigational (*nav*) or informational (*inf*). An example of a 2009 web track query is:

```
<topic number="1" type="faceted">
  <query>obama family tree</query>
  <description> Find information on President Barack Obama's family
  history, including genealogy, national origins, places and dates of birth, etc.
  </description>
  <subtopic number="1" type="nav">
      Find the TIME magazine photo essay "Barack Obama's Family Tree".
  </subtopic>
  <subtopic number="2" type="inf">
      Where did Barack Obama's parents and grandparents come from?
  </subtopic>
  <subtopic number="3" type="inf">
      Find biographical information on Barack Obama's mother.
      </subtopic>
</topic>
```

We relied on the subtopics of these collection topics and treated them as queries that explicitly verbalize the search tasks they pursue, which we call **explicit task-oriented searches**. Note that their pursued search tasks are indicated via <*inf*> or <*nav*> labels. We grouped these explicit task-oriented searches into two clusters based on their pursued search tasks. The first cluster contains the **explicit informational queries** and the second cluster contains the **explicit navigational queries**.

We processed the queries in every cluster in order to identify keywords that can serve as descriptors of the search task represented in the cluster elements. Query processing involved applying tokenization, part-of-speech tagging, stopword removal and lemmatization to the query elements. Search task descriptive keywords' selection involved counting query term frequencies and making binary human judgments as to whether or not each of the query terms explicitly verbalize the search task associated with the cluster elements. For the explicit navigational queries we identified a set of frequently occurring terms, such as *website*, *homepage*, *webpage*, *blog*, *forum*, which verbalize their associated search task. We manually extracted those terms and stored them in a repository of **navigational search keywords.**

For the explicit informational queries, though, we were unable to identify terms that explicitly verbalize their associated search task, since the goal of those

queries can be inferred based on the semantics of their contextual elements. As example, consider the term *find*, which is the most frequent term within the explicit informational queries. Evidently, the goal of *find* can be either navigational (e.g. <*go to the find.com website*>) or informational (e.g. <*find famous people*>) depending on its contextual terms. To overcome this bottleneck, we tried to capture the semantic relations that might hold between the query terms and the terms that signal the query informational tasks.

To achieve that, we relied on WordNet[2] and examined whether the terms in the TREC topics are semantically related to the terms contained in their informational subtopics. For our examination, we firstly lemmatized the topic terms and we generated bigram representations for topics longer than two words. Then, we matched the lemma or bigram of every topic as well as the terms in the topic's informational subtopics against WordNet nodes and upon their detection we extracted the semantic relations (if any) between the topic terms and the terms in the topic's informational subtopics. After applying the above process to our dataset, we obtained the following results. In 70% of the TREC topical queries, the terms used to explicitly verbalize their informational task were semantically related to the query terms. Out of all the TREC topic terms, 60.87% exhibit some semantic relation to at least one term in their informational descriptions. Out of all the identified semantic relations, hyponymy holds most of the times (77.18%) between the query terms and their corresponding informational oriented terms, followed by meronymy (9.88%), holonymy (6.93%) and hypernymy (6.01%) in that order.

Obtained results imply that the terms people use to explicitly verbalize the informational tasks of their searches are mainly semantic specializations of the query terms. Based on the above, we extracted the terms that are specializations (i.e. hyponyms) of the TREC topic queries, we associated every query with its corresponding hyponyms (<*query, query specialized terms*> and stored them in a repository of **informational search keywords** for the respective queries. Having described our approach toward building a repository of navigational and informational search keywords respectively, we proceed with the presentation of the query reformulation process.

Before that, we should note that although the navigational search keywords our method identifies are generic and can be added to any query for reformulating it into a navigational search; this is not the case for the informational search keywords since as already noted these are query-dependent. To build a global repository of informational search keywords, one could rely on a large set of queries mined from the log files of commercial search engines, extract their corresponding WordNet hyponyms and store them together into a local index of candidate terms for generating informational reformulations for the respective queries.

3.2. Generating Query Reformulations

Given an input query q and a repository of search goal descriptive terms, built as previously described, we seek to generate an informational and a navigational reformulation of q. For reformulating q into an explicit navigational search, we firstly compare the textual tokens of q to the navigational search keywords in our repository and we compute their Overlap Similarity (*OSim*) [5]. If *OSim*> 0, we conclude that q already represents a navigational term, so we do not reformulate it. Conversely, if *OSim*= 0, we expand q with terms extracted from our repository of navigational search keywords. Similarly for reformulating q into an explicit informational search, we compare the textual tokens of q to the informational search keywords in the repository and again we compute their *OSim* value. If *OSim*> 0, we expand q with specialized terms associated with the overlapping q terms in the repository of informational search keywords. If *OSim*= 0 we firstly capture the semantics of q and then we extract its corresponding WordNet hyponyms, which we add to the input query for reformulating it into an explicit informational search.

To capture the semantics of an input query not encountered in our repository of search task descriptive terms, we rely on the method proposed in [9], which uses a topical ontology for capturing the query semantics within individual search sessions. Having disambiguated the input query, we enrich it with its corresponding WordNet hyponyms and deliver its explicit informational reformulation. Both query reformulations, i.e. q'(*nav*) and q'(i*nf)* are displayed to the users in order to select the one that suits their search intentions.

4. Experiment

To assess the performance of our query reformulation method, we collected the queries issued by 5 volunteers on a single day and for every query we generated an informational and a navigational reformulation. After manually inspecting the 108 queries recorded in our search trace, we found out that 93 of them did not explicitly verbalize their pursued search tasks. We relied on those 93 queries to reformulate them with informational and navigational keywords respectively.

[2] http://wordnet.princeton.edu/

That is, we generated for every query one informational reformulation by adding to the query specialized keywords and one navigational reformulation by adding to the query navigational keywords extracted from our corresponding repository of task descriptive terms. The amount of terms added to each reformulated query equaled the number of terms it initially contained, i.e. a two-word query was reformulated as informational by enriching it with two specialized terms (one specialization per query term) and it was reformulated as navigational by enriching it with two navigational terms that were randomly extracted from our navigational search keywords repository.

We showed the reformulated queries to our study participants and asked them to evaluate their accuracy in verbalizing their associated search tasks by assigning to every reformulation a binary score (1 or 0). The reformulations that accurately verbalized their corresponding search tasks assigned a score of 1 whereas the reformulations that did not verbalize their corresponding search tasks assigned a score of 0. The evaluation metrics we used were precision (% of accurate reformulations) and query coverage (% of queries for which our method delivers at least one informational and one navigational reformulation). Results show that our method generated a navigational reformulation for all 93 test queries (100% coverage) with an average precision of 97.84%, whereas it generated informational reformulations for 86 queries (92.47% coverage) with an average precision of 91.39%.

5. Conclusions

We have addressed the problem of query reformulation for task-oriented web searches. We propose a method for identifying terms, which when added to a user query, they would make explicit its underlying search intention. The work reported in [10] generates intentional query suggestions by adding verbs to the input queries, but without discriminating between the search tasks every suggestion might accomplish. Our work improves existing research by offering query reformulations that explicitly verbalize the user goals hidden behind search requests. Among the benefits to understanding the tasks pursued via web searches is being able to automatically suggest query reformulations that encapsulate the query intentions. Thus, if many users searching for the same query have a common task in mind, the search engine can be proactive and suggest query alternatives that retrieve results, which not only relate to the query topic but they are also useful for accomplishing the query task.

In the future, we plan to apply our query reformulation method to a large dataset of real web searches and estimate retrieval performance for the improved queries by measuring user satisfaction from both search results and suggested queries. Currently, we are working on the design of a query reformulation add-on module that once employed by searchers it would automatically suggest query-specific task-descriptive terms to be employed as additional query keywords.

6. References

[1] Baeza-Yates R., Hurtado C., Mendoza M. 2004. Query Recommendation using Query Logs. In *Proceedings of the 9th International Conference on Extending Database Technology*, pp. 588-596.

[2] Broder A. 2002. A Taxonomy of Web Search. In *SIGIR Forum*, 36(2).

[3] Clarke C.L.A., Craswell N., Soboroff I. 2009. Overview of the TREC 2009 Web Track. In *Proceedings of the 18th Text Retrieval and Evaluation Conference.*

[4] Jansen B., Booth D., Spink A. 2007. Determining the User Intent of Web Search Engine Queries. In P*roceedings of the 16th International World Wide Web Conference*, pp. 1149-1150.

[5] Haveliwala T. 2002. Topic Sensitive PageRank. In *Proceedings of the 11th International World Wide Web Conference*, pp.517-526.

[6] Guo J., Xu G., Li H., Cheng X. 2008. A Unified and Discriminative Model for Query Refinement. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 379-386.

[7] Lee U., Liu Z., Cho J. 2005. Automatic Identification of User Goals in Web Search. In *Proceedings of the 14th International World Wide Web Conference*, pp. 391-401.

[8] Rose D.E., Levinson D. 2004. Understanding User Goals in Web Search. In *Proceedings of the 13th International World Wide Web Conference.*

[9] Stamou S., Ntoulas A, 2009. Search Personalization through Query and Page Topical Analysis. *Journal of User Modeling and User Adapted Interaction*, vol. 19, no. 1-3, pp. 5-33.

[10] Strohmaier M., Kroll M., Korner Ch. 2009. Intentional Query Suggestion: Making User Goals More Explicit During Search. In *Proceedings of the Web Search and Click Data Workshop*.

[11] Taksa, I., Spink, A. and Goldberg, R. 2008. A Task-Oriented Approach to Search Engine Usability Studies. *Journal of Software*, vol. 3, no. 1, pp. 63-73.

[12] Zhang Z., Nasraoui O. 2007. Efficient Hybrid Web Recommendations based on Markov Click Stream Models and Implicit Search. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp. 621-627.