Query Disambiguation Based on Novelty and Similarity
User’s Feedback

G. Bordogna\textsuperscript{1}, A. Campi\textsuperscript{2}, G. Psaila\textsuperscript{3}, and S. Ronchi\textsuperscript{2}

\textsuperscript{1} CNR-IDPA, Via Pasubio 3, 24044 Dalmine (BG), Italy
\textsuperscript{2} Politecnico di Milano, DEI, Piazza L. da Vinci 32, 20133 Milano (MI), Italy
\textsuperscript{3} Università degli Studi di Bergamo, Viale Marconi 5, 24044 Dalmine (BG), Italy

Abstract. In this paper we propose a query disambiguation mechanism for query context focalization in a meta-search environment. Our methods start from a set of documents retrieved executing a query over a search engine and applies clustering in order to generate distinct homogeneous groups. Then, the following step is to compute for each cluster a disambiguated query that highlights its main contents. The disambiguated queries are suggestions for possible new focalized searches. The ranking of the clusters from which the queries are derived is provided based on a balance of the novelty of cluster contents, and their overall similarity with respect to the query.

1 Introduction

This paper focuses on two well known problems of web searches. The first one faces the well-known ambiguity of the semantics of commonly very short user queries to Web search engines. With this rough and ambiguous description of users’ needs, the search engines can hardly guess what users are really searching for. Thus search engines retrieve documents regardless of the possible multiple semantics of the query terms. The second problem is related with the way users search on the Web by iterative trial and error cycles of query reformulation: in each step, users submit slight variations of the original query, in order to better express their needs. Further, at each step, users only analyze the first few pages of the results, that in many cases contain almost the same documents. However, for some searches, such as for bibliographic searches, the user objective is to obtain new documents, not already seen, and covering the topics of interest in the first pages of results list. The issues described above have already been faced in the literature distinctly. Several techniques for web search result disambiguation were proposed, such as those based on clustering \cite{1}, or those exploiting contextual information of the user (for example the techniques based on the history of searches previously performed by the user, or those based on collaborative query log analysis \cite{2,3}). The evaluation of the novelty of retrieved documents with respect to past results, is mainly used for news ranking in the routing systems \cite{4}, and for relevance feedback (see \cite{5,6}). But these two issues have never been faced together in the same context. This paper is a first attempt toward the generation of disambiguated query formulations focalizing the query context that potentially may retrieve both new and relevant documents. The core of our proposal is the generation of disambiguated queries, that provide suggestions to
the user to execute new searches, in order to deepen the retrieval of focalized contents in a specific context. This is done in order to overcome the problem that is common to many relevance feedback mechanisms based solely on similarity ranking: they tend to retrieve the same documents several times at each cycle. In our proposal, at each iteration of the search process, a set of disambiguated queries is generated from the retrieved documents. The set of retrieved documents is divided into groups by dynamic clustering based on the titles and the snippets of the documents. To each cluster, we associate a rank based on a balance of two criteria: the novelty of contents of the groups with respect to past results, and their overall similarity with respect to the original query. A disambiguated query is then generated from each group: this query highlights the main contents of the group. At this point, in order to retrieve novel documents about contents focalized in a context of interest, the user can select one of the disambiguated queries and submit it to the search engine. Together with the query the user can possibly provide a personal preference degree for the novelty with respect to the similarity ranking criteria: on one side, the user can adapt the query generation mechanism to formulate disambiguated queries that are either more keen to potentially retrieve new documents previously unseen; on the other side, the user can obtain more precision in expressing the focalized contents of interest. The paper is organized as follows: in Section 2 the related works are presented. In section 3 the model of the disambiguation process is formalized. Some examples of its application are discussed in section 4. Finally, some conclusions are given.

2 Related Works

It is well known that experts in many fields are troubled in finding the web pages that suit their purposes. This because they generally formulate queries consisting of two or three terms that are often ambiguous to gather appropriate Web pages (Sanderson reports that from 7% up to 23% of web searches in query logs are ambiguous [8]). Several papers have been published on the topics of query disambiguation based on extracting terms from searched documents: the idea is to replace user-provided query terms with more specific ones, to the aim of narrowing the search to the context the users have in mind. In order to obtain these terms, some approaches exploit external knowledge, such as Wordnet, to normalize the meaning of search terms in distinct contexts [9][10]. These approaches revealed that they expand the search terms with too fine grained words [11]. Some other approaches are based on long query log analysis over long periods of time on the server side [2][3], which, however, have the drawback of raising privacy issues. In an attempt to skip this privacy concern, recent approaches determine user intention based on the analysis of short query session logs [12], thus confuting the idea that session-only information is too sparse to allow for any meaningful prediction [13]. Some approaches apply collaborative techniques such as the similarity between query logs of distinct users and corresponding appreciated results [14][15]. However, these approaches suffer from the start up problem, when no recommendations are available. Some other approaches generate groups of terms based either on terms co-occurrence analysis [16], or dynamic clustering of query results [1] to identify the contexts of terms. In our proposal, we also apply dynamic clustering of search results to