Chapter 57
On The Reuse of Past Searches in Information Retrieval: Study of Two Probabilistic Algorithms

Claudio Gutiérrez-Soto
Université de Toulouse, France & Universidad del Bío, Chile

Gilles Hubert
Université de Toulouse, France

ABSTRACT

When using information retrieval systems, information related to searches is typically stored in files, which are well known as log files. By contrast, past search results of previously submitted queries are ignored most of the time. Nevertheless, past search results can be profitable for new searches. Some approaches in Information Retrieval exploit the previous searches in a customizable way for a single user. On the contrary, approaches that deal with past searches collectively are less common. This paper deals with such an approach, by using past results of similar past queries submitted by other users, to build the answers for new submitted queries. It proposes two Monte Carlo algorithms to build the result for a new query by selecting relevant documents associated to the most similar past query. Experiments were carried out to evaluate the effectiveness of the proposed algorithms using several dataset variants. These algorithms were also compared with the baseline approach based on the cosine measure, from which they reuse past results. Simulated datasets were designed for the experiments, following the Cranfield paradigm, well established in the Information Retrieval domain. The empirical results show the interest of our approach.

INTRODUCTION

A wide range of research lines provide support to Information Retrieval (IR), such as indexing techniques, weighting schemes, matching functions, formal models, and relevance feedback. Over a considerable period of time the retrieval process was applied in a standardized manner for all users, who used a given IR system (IRS). Therefore, given a common query for different users, the final result of this query was the same for each user. Then, new subfields have appeared

DOI: 10.4018/978-1-4666-9562-7.ch057

such as personalized IR, contextual IR, and collaborative IR. These subfields aim at exploring new approaches that adapt the returned results according additional aspects such as the profile of the user who submitted the query, the context (e.g., location) in which the query was submitted, and other users having the same information need.

Nowadays, most of IRSs store data that are associated with the queries. Nonetheless, few approaches take advantages from past search results. Our main assumption is that a set of past search results can be useful to answer new queries. Some approaches exploit the previous searches in a customizable way for a single user. Nevertheless, the use of previous searches is based mainly on either repeated queries or query reformulations inside search sessions. Approaches that address past searches collectively are less common. This type of approaches is catalogued in collaborative information retrieval. An example of such research is provided by (Hust, 2004), which proposes query expansion based on past searches for new submitted queries. The assumption is that a user may benefit from search experiences of other users who share the same information need, to reformulate her/his query.

The approach proposed in this paper leverages past searches by using past results to build the results returned for new submitted queries. The proposed approach does not aim at improving queries previously submitted by a user (i.e., repeated queries). On the contrary, the approach concerns new queries submitted by a user (i.e., not yet submitted by the user in the past) that were submitted by other users in the past. It aims at introducing a collaborative aspect by taking advantage of previous searches of other users. The approach is based on two separated aspects: a storage aspect and a retrieval aspect. This paper focuses on the retrieval aspect, where relevant documents are selected from the result list of the most similar query to build the response for a new query. It presents, among others, a study of two possible probabilistic algorithms for the document selection.

The IR literature holds a lot of probabilistic approaches, some of them proposed a long time ago. The probabilistic algorithms proposed in this paper correspond to the Monte Carlo category (Burgin, 1999). These algorithms assign the highest likelihood to top-ranked documents in past result lists. The assigned likelihood decreases according to descending ranks of documents in past result lists. The major advantages of these algorithms are the next: they are quite easy to implement and they do not require learning step.

Validation of our approach implies experiments with multiple objectives. A first objective is to compare our approach based on the reuse of past queries with the baseline approach of information retrieval from which it reuses past results. A second objective is to study the two proposed algorithms, which can be used in our approach for the construction of new result lists from past results, in different environments.

Traditionally, a significant portion of research in IR has been related to system evaluation. Evaluation is devoted essentially to effectiveness (i.e., result accuracy), which is obtained by measuring precision (i.e., fraction of relevant documents in the retrieved documents). Although a large amount of IR collections can be found currently, it is a difficult task to find ad-hoc collections to evaluate approaches based on past queries. Most existing IR collections, such as the well-known TREC collections, are composed of very dissimilar queries. With such collections, evaluating an approach looking for similar past queries has no sense. Some collections using query logs of search engines contain similar queries, but these similar queries are mostly repeated queries or query reformulations submitted by the same user. These collections are designed to particular IR tasks such as session detection (Kanoulas et al.,