Query Expansion Methods for Collaborative Information Retrieval

Armin Hust

Vom Fachbereich Informatik
der Technischen Universität Kaiserslautern
zur Verleihung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)
genehmigte Dissertation

Berichterstatter:	Prof. Dr. Michael M. Richter
	Prof. Dr.-Ing. Dr. h. c. Theo Härder
Dekan:	Prof. Dr. Hans Hagen

Datum der wissenschaftlichen Aussprache: 24. Mai 2004

D 386

Preface

The search for documents requires an increasing amount of time for employees of companies. This has become an important economical factor that cannot be neglected. The search takes place in the internet as well as in large internal document collections. A major problem is that the specific search problems are often not precisely specified. The reason is that the documents themselves are only vehicles for solving other, often complex, tasks. In this view the retrieved documents are more or less useful. In mathematical terms there is some utility function involved that is, however, difficult to formulate. It depends in particular on a certain person and a certain context in which this person is interested in the document. These two aspects are characterized by the key words *personalization* and *context* and play a major role in this thesis.

The retrieval mechanisms available today have a number of weaknesses. One is that for intranet applications the web browsers do not apply. From a structural point of view the retrieval engines are mainly based on syntactic issues and neglect semantic aspects what results in missing the intended utility.

In this context this thesis has a clear focus: To make use of past experiences in organizing the search by investigating the interactions of different users with the retrieval engine. The main methodology presented is *Collaborative Information Retrieval* (CIR). The past users provide information and opinions about the relevance of retrieved documents. This information is used to replace the original query by a new one, a process that is called *query expansion*. The expected effect is that the reformulated query is closer to relevant documents and more distant to non-relevant documents.

A central concept is concerned with various relations, between queries and documents, new queries, new and old queries and between documents. These relations are formulated in terms of similarity measures that have access to vector space representations of queries and documents. The application of a similarity measure results in a ranking of the documents and these rankings are significantly improved by using the query expansions. The thesis introduces into this area as much as necessary. On this basis several new algorithms are presented.

It turns out, not unexpectedly, that these methods are not fully satisfactory. An essential further improvement is obtained by applying machine learning methods. This means essentially to learn from the feedback given by previous users about the relevance's. Technically, the learning process results in improved similarity measures and more useful query expansions.

The results are tested systematically by comprehensive experiments. There are several lessons learned from these tests. A particular one is that the performance depends to a large degree on the existence of many queries in the past that have a sufficiently large overlap.

In summary, this thesis presents an integrated and new approach to the retrieval of text documents that is of foundational as well as of practical interest. It contains many new features and I am sure that this work will get increasing attention in the future.

Prof. Dr. Michael M. Richter

Zusammenfassung

In dieser Dissertation werden Methoden zur Anfrageerweiterung entwickelt, die im Kollaborativen Information Retrieval nützlich sind. Kollaboratives Information Retrieval ist die in dieser Arbeit entwickelte Methodik, bei der ein Information-Retrieval-System Informationen ausnutzt, die in früheren Suchprozessen von einem oder mehreren Benutzern gesammelt wurden. Mit diesen Informationen wird die Retrieval Performance für die aktuelle Anfrage verbessert. Wir zeigen, wie Kollaboration von einzelnen Benutzern die gesamte Retrieval Performance für alle Benutzer verbessern kann, wobei die Retrieval Performance anhand der Nützlichkeit der abgerufenen Information unabhängig von speziellen Benutzergruppen gemessen wird.

Zuerst werden die verschiedenen Information-Retrieval-Modelle beschrieben, die bereits seit mehreren Jahrzehnten benutzt werden. Wir vertiefen eines der meistgenutzten Modelle, das sogenannte Vector-Space-Modell, zeigen Verbesserungen in diesem Modell auf und zeigen die aktuellen Grenzen dieses Modells anhand der Retrieval Performance, die mit Recall und Precision gemessen wird.

Ein kollaboratives Szenario, in dem Benutzer ohne bewusste Beteiligung zusammenarbeiten, kann durch ein Kollaboratives Information-Retrieval-System selbst aufgebaut werden. Das System kann die notwendigen Daten selbst ohne weitere Benutzer-Interaktionen erfassen, speichern und verwalten. Die benötigten Daten können durch unaufdringliche Beobachtung der einzelnen Suchprozesse von verschiedenen Benutzern gewonnen werden, wobei jeder Suchprozess aus drei Schritten besteht: Anfrage stellen, Präsentation der Resultate durch das System in geordneter Reihenfolge und Betrachten der Resultate durch den Benutzer.

Wir entwickeln Algorithmen zur Anfrageerweiterung, analysieren und bewerten sie. Danach entwickeln, analysieren und bewerten wir Term-Gewichtungs-Algorithmen. Diese beiden Arten von Algorithmen können einfach in kollaborativen Information-Retrieval-Systemen integriert werden, solange Informationen über frühere Suchprozesse vorhanden sind. Wir zeigen, dass diese Algorithmen die Retrieval Performance insgesamt verbessern und damit die Qualität und Nützlichkeit der abgerufenen Information für den Benutzer verbessern. Dann entwickeln, analysieren und bewerten wir Algorithmen, mit denen Ähnlichkeitsfunktionen gelernt werden. Ähnlichkeitsfunktionen sind die Basis für die Darstellung der Resultate einer Anfrage in einer geordneten Reihenfolge. Wir integrieren diese gelernten Ähnlichkeitsfunktionen in die zuvor entwickelten Expansions- und Gewichtungs-Algorithmen und zeigen anhand relativer Verbesserungen und statistischer Aussagen, dass damit die Retrieval Performance weiter verbessert wird.

Schließlich fassen wir die Hauptbeiträge dieser Arbeit zusammen, ziehen Schlussfolgerungen und verweisen auf offene Fragestellungen für zukünftige Arbeiten.

Summary

This dissertation develops, analyzes, and evaluates query expansion methods to be used in collaborative information retrieval. We define collaborative information retrieval as a task, where an information retrieval system uses information gathered from previous search processes of one or several users to improve retrieval performance for the current user searching for information. We show how collaboration of individual users can improve overall information retrieval performance. Performance in this case is expressed in terms of quality and utility of the retrieved information regardless of specific user groups.

In this dissertation, we first analyze the various information retrieval models which now have been used for several decades. We then focus on one of the most popular models, the so-called vector space model, outline some improvements that have been applied to this model and show the current limitations of this model in terms of retrieval performance, which is measured in recall and precision values.

The collaboration scenario can be built by a collaborative information retrieval system automatically without conscious collaboration support by the users of the system. The system itself can acquire, store, and maintain the necessary data without the need for additional user interaction. The necessary data can be obtained by unobtrusive observation of the search processes from individual users, where each search process consists of three steps, i.e., entering the query into the system, presenting the query results by the system according to a ranking function, and viewing the query results by the user.

We develop algorithms to be used in collaborative information retrieval systems for query expansion procedures, analyze these algorithms, and evaluate them. Then we develop, analyze, and evaluate term reweighting algorithms. These two kinds of algorithms can easily be integrated into collaborative information retrieval systems as long as information from previous search processes are available in the systems. We show that these algorithms can improve overall system performance measured in terms of recall and precision, thus increasing the quality and utility of the retrieved information for the users. Then we develop, analyze, and evaluate learning algorithms for similarity functions to be used in collaborative information retrieval systems. Similarity functions are the basis of the ranking of query results for the presentation of the results. These learning algorithms are then integrated into the query expansion procedures developed beforehand. We show that, using these learning algorithms, we can improve information retrieval performance still further.

Finally, we summarize the main contributions of this thesis, draw some final conclusions, raise issues for future work, and make some final remarks.