Automatic query expansion based on a similarity thesaurus

Author(s):
Qiu, Yonggang

Publication Date:
1995

Permanent Link:

https://doi.org/10.3929/ethz-a-001469785 →

Diss. ETH No. 11158

Automatic Query Expansion Based on a Similarity Thesaurus

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH
(ETH Zürich)

for the degree of
Doctor of Technical Sciences

presented by
Yonggang Qiu
B.Sc. and M.Sc., Southeast University, Nanjing
born December 3, 1962
citizen of the People's Republic of China

accepted on the recommendation of
Prof. Dr. H.P. Frei, examiner
Prof. Dr. P. Schäuble, co-examiner

1995

Aug. 7, 1995

Abstract

One method to improve the effectiveness of an information retrieval system, is to use an information structure for automatic query expansion. The objective of the method is to resolve the problem of vagueness of a query of the user by enhancing the query. However, most previous attempts in this direction have failed to improve the retrieval effectiveness.

This dissertation introduces a new concept-based query expansion model based on an information structure. The information structure, called a similarity thesaurus, is constructed automatically and consists of term-term similarities that are based on how the terms of a collection "are indexed" by the documents. It reflects domain knowledge about the collection and is used to select and weight additional search terms when expanding a query. A query is expanded by adding those terms that are most similar to the concept of the query, rather than by adding terms that are strongly related to one of the original query terms. The added terms are then weighted according to their similarity to the query concept.

We also present an extended version of the concept-based query expansion model that uses all the search terms to determine the query concept. The main idea of the extension is to represent the concept of a query only by good query terms appearing in the top ranked documents of the previous run of the query.

Because of the sheer size of a similarity thesaurus, it becomes impracticable when such a similarity thesaurus is constructed or used in connection with a commercial database containing millions of documents and terms. In this dissertation, we describe how to reduce this size and consequently bring down the cost of constructing, storing and accessing of a similarity thesaurus. In addition, novel algorithms are introduced for constructing and updating a similarity thesaurus.

Retrieval results of experiments on both small and large collections are presented. The results indicate that the retrieval effectiveness is considerably higher when the concept-based query expansion methods are applied than when using a reference method, and the extended model produces better retrieval results than the original expansion model.

For retrieving information from large commercial databases, we developed and implemented a retrieval system ISIR. ISIR is a very user-friendly front-end to the commercial Data-Star system. Data-Star, which only supported the classical Boolean retrieval, was enhanced by integrating real weighted retrieval as well as functions for supporting similarity thesauri. ISIR allows the user to express queries in natural language. These queries can then be expanded by applying our proposed approaches before being submitted to the enhanced Data-Star system. To help the user (re)formulate queries, two thesaurus browsers, one for manual thesauri and the other for similarity thesauri, are also integrated in ISIR.

Zusammenfassung

Um die Effektivität eines Information-Retrieval-Systems zu erhöhen, kann eine Informationsstruktur für automatische Anfrageerweiterung verwendt werden. Das Ziel dieser Methode ist es, das Problem der Vagheit einer Benutzerfrage zu lösen. Die meisten Versuche in dieser Richtung sind bisher leider gescheitert.

Wir stellen ein neues wissensbasiertes Information-Retrieval-Modell vor, das auf einer konzeptbasierten Anfrageerweiterung beruht. Das Modell benutzt als Informationsstruktur einen Ähnlichkeits-Thesaurus. Der Ähnlichkeits-Thesaurus wird automatisch von einer Kollektion konstruiert und besteht aus Termen und Ähnlichkeiten zwischen Termen. Die Ähnlichkeiten leiten sich davon ab, wie die Terme einer Kollektion durch die Dokumente "indexiert" sind. Der Ähnlichkeits-Thesaurus ist eine Wissensbasis der Kollektion und wird für die Erweiterung von Anfragen verwendet, um zusätzliche Suchterme auszuwählen und die Gewichte dieser Terme zu schätzen. Eine Anfrage wird unter Hinzufügung jener Terme erweitert, die zum Konzept der Anfrage am ähnlichsten sind.

Es wird zudem eine verfeinerte Version des konzeptbasierten Anfrageerweiterungsmodells vorgestellt. Die Verfeinerung besteht darin, dass das Konzept einer Anfrage nur durch gute Suchterme beschrieben wird, die in den hochrangierten Dokumenten vorhergehender Suchresultate vorkommen.

Allein wegen der Grösse eines Ähnlichkeits-Thesaurus ist es nicht möglich, solche Ähnlichkeits-Thesauri für grosse kommerzielle Datenbanken zu konstruieren. Es wird gezeigt, wie man diese Grösse reduzieren und infolgedessen die Kosten für den Gebrauch eines Ähnlichkeits-Thesaurus verringern kann. Ferner werden neue Algorithmen für die Konstruktion und Aktualisierung eines Ähnlichkeits-Thesaurus vorgestellt.

Es werden Retrieval-Experimente sowohl auf kleinen als auch auf grossen Kollektionen vorgestellt. Mit beiden Anfrageerweiterungsmethoden werden deutlich bessere Retrieval-Ergebnisse erzielt als mit einer Referenzmethode. Das verfeinerte Modell zeigt dabei wiederum leicht bessere Resultate als das ursprüngliche Modell.

Für die Informationssuche auf grossen kommerziellen Datenbanken haben wir ein Retrieval-System ISIR entwickelt und implementiert. ISIR ist eine benutzerfreundliche Schnittstelle zum kommerziellen Data-Star System. Data-Star wurde durch die Integration von echtem gewichtetem Retrieval und den Funktionen für die Unterstützung eines Ähnlichkeits-Thesaurus verbessert. Der Benutzer drückt seinen Informationswunsch in natürliche Sprache aus. Diese Anfragen können dann durch die vorgeschlagenen Methoden erweitert werden. Mit Hilfe zweier Thesaurus-Browser, die ebenfalls Teil von ISIR sind, wird die Formulierung einer Anfrage erleichtert.