Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 44 (2008) 781–789
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Searching in MEDLINE: Query expansion and manual indexing evaluation

Samir Abdou, Jacques Savoy *

Computer Science Department, University of Neuchatel, 2009 Neuchâtel, Switzerland

Received 19 December 2006; received in revised form 16 March 2007; accepted 17 March 2007
Available online 23 May 2007

Abstract

Based on a relatively large subset representing one third of the MEDLINE collection, this paper evaluates ten different IR models, including recent developments in both probabilistic and language models. We show that the best performing IR models is a probabilistic model developed within the *Divergence from Randomness* framework [Amati, G., & van Rijsbergen, C.J. (2002) Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM-Transactions on Information Systems 20*(4), 357–389], which result in 170% enhancements in mean average precision when compared to the classical *tf idf* vector-space model. This paper also reports on our impact evaluations on the retrieval effectiveness of manually assigned descriptors (MeSH or Medical Subject Headings), showing that by including these terms retrieval performance can improve from 2.4% to 13.5%, depending on the underling IR model. Finally, we design a new general blind-query expansion approach showing improved retrieval performances compared to those obtained using the Rocchio approach.

Keywords: Manual indexing; Blind query expansion; Medline; MeSH; Genomics TREC; Probabilistic model; Language model; Rocchio query expansion; Evaluation

1. Introduction

MEDLINE is a well-known premier bibliographic collection that contains references to articles contained in journals on life sciences. The Genomics TREC 2004 evaluation campaign provides access to one third of this large corpus together with fifty real information need descriptions. Within this realistic context, our first goal is to evaluate the retrieval performance of various IR models, including recent developments in probabilistic and language models, and also vector-space schemes.

Second, we accept the fact that manually assigned descriptors should increase the probability of retrieving more pertinent documents, as compared to those searches based only on scientific article titles and abstracts.

* Corresponding author. Tel.: +41 032 718 1375; fax: +41 032 718 2700.
E-mail addresses: Samir.Abdou@unine.ch (S. Abdou), Jacques.Savoy@unine.ch (J. Savoy).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2007.03.013

Manual indexing, usually based on controlled vocabularies, should prescribe a uniform and invariable choice of indexing descriptors and thus normalize orthographic variations (e.g., "database" or "data base"), lexical variants (e.g., "analyzing", "analysis") or any other expressions having similar meanings (e.g., "computer science", "informatics").

A third issue concerns information submitted in queries to express user needs. As is commonly recognized, users do not supply all details and thus there is a lack of certain synonyms or related terms. To partially resolve this problem, a query expansion technique should take different term-term relationships into account and expand the original query. As seen from various empirical studies, this usually results in better retrieval performance.

The rest of this paper is organized as follows. Section 2 describes related works in the two different subdomains presented in this paper: manual and automatic indexing, and automatic query expansion approaches. Section 3 depicts the main characteristics of our test-collection, while Section 4 briefly describes the IR models applied during our experiments. Section 5 explains our new query expansion model, and Section 6 evaluates the performance of various IR models, in addition to two query expansion approaches. The main findings of this paper are presented in Section 7.

2. Related work

2.1. Manual & automatic indexing

Only a few studies have undertaken to directly compare the performance of manual vs. automatic indexing methods. The well-known Cranfield experiments for example studied and evaluated the retrieval impact of various manual-indexing strategies. For example, Cleverdon (1967) reported that single-word indexing was more effective than extracted terms from a controlled vocabulary, where both indexing schemes were compiled by human beings (1400 documents, 221 queries).

In order to evaluate the importance of manually assigned descriptors, Hersh, Buckley, Leone, and Hickam (1994) investigated search performance differences resulting from input provided by users from different backgrounds (physicians or librarians, novices or expert users) when searching OHSUMED (a subset of the MEDLINE collection). Overall, performance differences were small and statistically insignificant, thus illustrating that the MeSH descriptors were not really advantageous. In an opposing viewpoint, Srinivasan (1996) reported that MeSH may in some cases help retrieving information in MEDLINE.

Based on the Amaryllis database, containing a French bibliographic collection (148,688 records and 25 queries), Savoy (2005) demonstrated that the inclusion of manually assigned descriptors could significantly enhance mean average precision by about 35% based on title-only queries, compared to an approach that ignored these additional descriptors. The question then arises: "Does the inclusion of MeSH headings improve mean average precision within the MEDLINE corpus?" Then, if the answer is positive: "What percentage improvement could we expect when such manually assigned descriptors are taken into account?"

2.2. Query expansion

To provide a better match between user information needs and documents, various query expansion techniques have been suggested. The general principle is to expand the query using words or phrases having meanings similar to or related to those appearing in the original request. To achieve this, query expansion approaches consider various relationships between these words, as well as term selection mechanisms and term weighting schemes. The specific answers to these three questions may vary, leading to a variety of query expansion approaches (Efthimiadis, 1996).

In a first attempt to find related search terms, we might ask the user to select additional terms to be included in an expanded query (e.g., (Vélez, Weiss, Sheldon, & Gifford, 1997)). This could be handled interactively through displaying a ranked list of retrieved items returned by the first query. Using the WordNet thesaurus, Voorhees (1994) demonstrated that terms having a lexical-semantic relation with original query words (extracted because of synonym relationship) provided very little improvement (around 1% compared to the original query).