Hierarchical Expansion for Concept-Based Search

Roee Sa'adon[a], Robert Moskovitch[a] and Yuval Shahar[a]

[a] *Medical Informatics Research Center, Ben-Gurion University, Beer Sheva, Israel*

Abstract

Although many digital libraries are indexed using a hierarchical conceptual structure, improving the performance of traditional free text search is not trivial. We present a preliminary evaluation of a novel method of a hierarchical concept based search, as implemented in the Vaidurya search engine, with the innovation of using automated query mapping. Using MetaMap, textual query is mapped into MeSH concepts, which are later abstracted. Preliminary encouraging evaluation results on the TREC Genomics 2004 test collection are presented.

Keywords:

Information Storage and Retrieval, Concept Based Search

Introduction

While the idea of concept based search is researched for several decades, not too many studies shown that it outperforms traditional free text search. Recently we introduced Vaidurya [jamia], a search and retrieval tool developed originally within the DeGeL library, which in a rigorous evaluation a significant improvement was observed when concept based search was used. In this study we propose a novel approach, in which the user has to enter only a textual search phrase, which is converted into a concepts using *MetaMap*.

Vaidurya

Vaidurya implements a free text search (FTS) and concept based search (CBS). For the FTS, documents are represented and indexed using the *vector space model* introduced by Salton [14]. The CBS in Vaidurya uses the logical operators *conjunction* (AND) and *disjunction* (OR) at two step of retrieval, called *inner-op* and *outer-op*. All the documents classified along a queried concept and its descendent are retrieved. Based on the *inner* the documents sets are intersected of unified to represent each tree, which then processed similarly based on the *outer* operator.

Methods

Based on the previous evaluation, we concluded that adding a query for concepts at the top levels will enhance FTS results, since the relevant documents commonly share abstract concepts. Our goal was to avoid the need in specifying explicitly the concepts, thus, enabling querying using a normal textual query, while exploiting the CBS. Given a set of *query terms*, a set of concepts at *varying* levels of the hierarchy is extracted using *MetaMap*. Then, the concepts are abstracted to their ancestors at a required level *k*, by climbing up through the MeSH hierarchy. In order to evaluate our method we used the TREC 2004 Genomics Track test collection, which offers 50 queries, and in which the documents are classified along MeSH concepts. As evaluation measures we refer to the precision measured at the top *5, 10 and 50* retrieved documents, and mean average precision (MAP). We report the preliminary evaluation of three important variables: (1) *CBS method* (2) *abstraction level* (3) *CBS weight.*

Results

Table 1 shows the best runs compared to the baseline.

Table 1 – The best results compared to the baseline, the algorithm is specified by inner^{op}-outer^{op}/level/weight,

Algorithm	MAP	P@5	P@10	P@50
OR-OR/5/0.2	**0.336**	**0.512**	**0.494**	**0.343**
AND-OR/5/0.5	**0.345**	0.484	0.478	0.327
OR-AND/4/0.1	**0.327**	**0.508**	**0.500**	0.342
Free Text Baseline	**0.326**	**0.488**	**0.484**	**0.342**

We also measured the CBS contribution over the queries which had the worst FTS performance. With algorithm AND-OR/4/0.2, for the worst 10 queries, we discovered an improvement of 150%, 200% and 50% for P@5, P@10 and P@50, respectively. For the worst 15, a relative improvement of 83%, 33% and 29% was found, for the same measures.

Discussion

We introduced an extension of Vaidurya which enables CBS through an automatic conversion of textual to conceptual queries using MetaMap. The evaluation done with the TREC-G showed a slight improvement. In the 10 and 15 queries, having the worst FTS performance, a significant relative improvement was observed. These are encouraging results for further development of the method**.**

Address for correspondence

roeesa@bgu.ac.il
Department of Information Engineering,
Ben Gurion University of the Negev, Israel .
P.O.B. 653, Beer Sheva 84105, Israel.