MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Use of the Vector Space Model for Expansion of Medical Queries

Wallnöfer R[1], Rammer Th[1], Schabetsberger Th[2], Pfeiffer KP[1], Göbel G[1]

[1]Department for Medical Statistics, Informatics and Health Economics,Medical University of Innsbruck, Austria
[2]University for Health Sciences, Medical Informatics and Technology, Innsbruck - Austria

Abstract and objective

We present an application of the Vector Space Model in connection with Medical Thesauri. The general idea is to use the Vector Space Model for a weighted expansion of medical queries. The application is based on the standardized ontology format OWL and – as first step – the concerning MeSH thesaurus. This includes an more generic approach than others. The application is implemented as prototyp web-service.

Keywords:

internet, information retrieval, Vector space model , MeSH

Introduction

Information Retrieval (IR) has become an important practice for millions of people in nearly every area. Especially in the area of Health Information Retrieval there is a lack of tools which support customers (patients, citizens) to find relevant and satisfying information. Most systems are based on full-text search. One example is Apache Lucene [1]. If you formulate a query you often get a lot of non relevant results because the system doesn't use any semantic relations. Another approach is used by Zeng et al [9]. With this tool the user is supported by a Query Assistant which assists the user to formulate his\her query.

This work focuses on the application of the vector space model on medical domain knowledge. Additionally to full-text search techniques the semantic relations between terms from a given domain should be considered. A simple and very common form of Information Retrieval is the Boolean model [2, 3]. This model is binary in other words documents are considered depending upon the presence or absence of the query term. Further, the queries are build by combining terms with Boolean operators (AND, OR, NOT). This model does not use weighting of terms. In this paper individual weighting of terms shall be reached with a variant of the vector space model [4, 5, 15]. The semantic relation between each search term shall become extracted with the Main Headings from the medical thesaurus "Medical Subject Headings" (MeSH) [6] which was published by the National Library of Medicine. With input from the desired search term an acyclic graph (represented by an n x n adjacent matrix) including all related terms (from the thesaurus) is generated. Furthermore the distance to each received term is calculated from the matrix by using the Floyd-Warshall Algorithm [7]. Finally, the calculation of the similarity between the search term and the existing documents is done regarding the vector space model.

Methods

The following components are used and form the base for this work:

- Medical Subject Headings (MeSH) thesaurus [6]
- Using vector space model for medical domain knowledge [8]
- Web Services, Apache Web Services [10,11]
- RDF/OWL [12,13]
- Converting Thesauri to RDF/OWL [14]

Based on [8] we develop a tool which makes a weighted query-expansion based on MeSH terms available via a web service.

Results

The results are implemented as web-service based on the vector space model using state of the art technologies. As the software will be realized as a web service it will be available on a server and can be used by a client from each arbitrary location through the Standard Internet Protocol (HTTP). The medical thesaurus MeSH is used as domain knowledge, as already stated above. MeSH is mainly used to index and to search for articles in large databases (e.g.: MEDLINE/PubMED). However this thesaurus is only available in a proprietary XML format. To be compatible with the semantic web standard format RDF(S) the MeSH thesaurus shall be used in RDF/OWL format. A description for converting thesauri to RDF/OWL is available at [14].

If a user is searching for information in a distinct subject area, conventional full-text based search engines often deliver a lot of non relevant responses. With this work it will be much easier for a user to find similar documents, related to his request, in a defined document pool. The user submits a suited search term (MeSH term) to the system and gets as a result the documents which will be most similar to the given search term. The calculation of the similar documents is done with the additional received Main

Headings from the MeSH thesaurus on the bases of the vector space model in combination with the Floyd-Warshall Algorithm.

Conclusion

By using domain knowledge in a standardized format in the future it is possible to use other medical knowledge domains like ICD-10 or SNOMED for searching similar documents. But it is important that these domains are also available in the same standardized format. In any case Information Retrieval will be much easier with this web service and vector space model based tool. Especially in the health Information Retrieval area customers will profit from this approach that take as a first step the MeSH thesaurus as a knowledge domain.

References

[1] Apache Lucene http://lucene.apache.org/java/docs/index.html
[2] Information Search and Retrieval http://www.iicm.tugraz.at/cguetl/education/isr/vo/inhalte/block02/Zusammenfassung%20VO2.html
[3] Summary of Search Engine Models http://mathdl.maa.org/mathDL/4/?pa=content&sa=viewDocument&nodeId=636&bodyId=1033
[4] Vector space model http://de.wikipedia.org/wiki/Vektorraum_Retrieval
[5] The Classic Vector Space Model http://www.miislita.com/term-vector/term-vector-3.html
[6] Medical Subject Headings http://www.nlm.nih.gov/mesh/
[7] Floyd-Warshall algorithm http://de.wikipedia.org/wiki/Algorithmus_von_Floyd_und_Warshall
[8] Goebel G, Andreatta S, Masser J, Pfeiffer KP. “A MeSH based intelligent search intermediary for Consumer Health Information Systems” Int. J. Med. Inform. 2001; 64:241-51
[9] Zeng QT, Crowell J, Plovnick RM, Kim E, Ngo L, Dibble E. “Assisting consumer health information retrieval with query recommendations” J. Am. Med. Inform. Assoc. 2006 Jan-Feb; 13(1):80-90.
[10] Web Services Activity http://www.w3.org/2002/ws/
[11] Apache Web Service Project http://ws.apache.org/
[12] Resource Description Framework (RDF) http://www.w3.org/RDF/
[13] OWL Web Ontology Language Overview http://www.w3.org/TR/owl-features/
[14] M. van Assem, M. R. Menken, G. Schreiber, J. Wielemaker, and B. J. Wielinga, "A Method for Converting Thesauri to RDF/OWL", presented at ISWC'04, Hiroshima, Japan, 2004 http://www.cs.vu.nl/~mrmenken/pubs/ISWC041.pdf
[15] F. Wiesman, Arie Hasman, H.J. van den Herik, “Information retrieval: An overview of system characteristics” Int. J Med Inform. 47 (1997) 5–26