Suggesting Terms for Query Expansion in a Medical Information Retrieval System

Morris Hirsch, MS, David Aronow, MD, MPH

Center for Intelligent Information Retrieval

University of Massachusetts at Amherst

{hirsch, aronow}@cs.umass.edu

The performance of Information Retrieval (IR) systems is critically dependent upon the quality of the queries submitted by the user. Queries may often be improved by "expanding" them with additional terms, either automatically or chosen in cooperation with the user. Typical expansion strategies include:

- selecting the most-significant words in documents judged relevant by the user after retrieval by the original query.
- term manipulations in controlled vocabulary systems, such as MeSH and ICD-9-CM, through explosion to include more specific terms, and truncation to include more general terms.
- inclusion of terms related to the original query by a thesaurus, which may be either manually prepared or computer-generated based on co-occurrences in the corpus.

The work reported here is part of a larger effort to improve access to the MEDLINE database, using the INQUERY system developed at the Center for Intelligent Information Retrieval of the University of Massachusetts at Amherst. INQUERY supports full-text retrieval, in which all text words are available as indexing terms.

Although INQUERY supports all these methods, this report is concerned only with thesaurus-based query expansion.

A well-designed manual thesaurus is an effective means of query expansion, but maintaining one requires large amounts of time and effort on the part of domain experts. We seek to avoid this effort by automatic construction.

We generate a thesaurus by scanning the corpus for noun phrases. Each occurrence is noted along with the surrounding context, after which they are sorted to bring all contexts of each phrase together. Each context group is treated as a "pseudo-document" with the noun phrase as the document title. The set of pseudo-documents is indexed as a thesaurus database that accompanies the original collection. A query that matches some words of context in a pseudo-document causes that document to be selected, and its title to be listed as a possible expansion term.

Context may be saved as either surrounding text or other nearby noun phrases. In the following example, based on healthcare policy documents, this list of terms results from the query "child"

- younger child
- while testifying
- well child visit
- warfarin-resistant strain
- valproic acid syrup

If the user asks the system to explain the "while testifying" choice, they are shown the matched context terms

- **child** abuse
- **child** abuse case
- **child**hood fracture
- medical record
- pediatric radiologist
- pediatric service

The user may select any of the terms shown as additions to their original query.

This material is based on work supported in part by NRaD Contract Number N66001-94-D-6054. This material is also based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.