Knowledge Acquisition of Generic Queries for Information Retrieval

Yoon-Ho Seol, MPhil, Stephen B. Johnson, PhD and James J. Cimino, MD

Department of Medical Informatics,
Columbia University College of Physicians and Surgeons, New York, New York

Several studies have identified clinical questions posed by health care professionals to understand the nature of information needs during clinical practice. To support access to digital information sources, it is necessary to integrate the information needs with a computer system. We have developed a conceptual guidance approach in information retrieval, based on a knowledge base that contains the patterns of information needs. The knowledge base uses a formal representation of clinical questions based on the UMLS knowledge sources, called the Generic Query model. To improve the coverage of the knowledge base, we investigated a method for extracting plausible clinical questions from the medical literature. This poster presents the Generic Query model, shows how it is used to represent the patterns of clinical questions, and describes the framework used to extract knowledge from the medical literature.

PROPOSED APPROACH

Our work is motivated by the hypothesis that users' information needs in medicine can be modeled and represented in a formal way, and that a knowledge base containing the patterns of users' information needs (generic queries) can facilitate query construction during the information retrieval process.

A generic query describes semantic relationships between concepts explicitly by using the UMLS knowledge sources. We make use of the abstraction of concepts and interconcept relationships that the UMLS semantic types and semantic relationships provide.

GENERIC QUERY MODEL

In order to capture the meaning of users' queries, it is important to identify semantic relationships among concepts explicitly. We use a graph that consists of concept nodes and relation edges. Concept nodes contain a UMLS semantic type and an instance of that semantic type (a medical term) if it is available. Interconcept relations are directed edges in the graph and are labeled with UMLS semantic relationships.

Our generic query model has been applied to clinical questions from published user studies to populate a knowledge base for our current prototype. XML is employed to represent generic queries, as XML is an emerging standard and supports graph-based data representation.

ACQUISITION OF GENERIC QUERIES

We developed a method to extract generic queries from the medical literature using a semantically enriched representation of documents in which the relationships of terms are identified explicitly. An existing concept or keyword-based indexing of a document is extended with the UMLS semantic types and semantic relationships.

The semantically enriched document is represented with XML. Parts of graphs (subgraphs) are extracted based on their frequency of occurrence in the documents. We tested our approach with a collection of abstracts from MEDLINE which focused on the treatment and diagnosis of cardiovascular disease.

This approach is motivated by the idea that a high frequency subgraph may be an interesting structure for representing users' information needs and may be a potential generic query.

DISCUSSION

A graph model with the UMLS knowledge sources is a useful tool for representing patterns of clinical questions. To improve the coverage of our knowledge base, we extracted patterns from citations to form candidate generic queries. A comparison of the generic queries from the literature analysis against the ones from published user studies reveals similar patterns, suggesting that this may be an alternative method for establishing generic queries. A formal evaluation with a larger document collection is needed to validate our approach.