Chapter XIII
Query Expansion by Taxonomy

Troels Andreasen
Roskilde University, Denmark

Henrik Bulskov
Roskilde University, Denmark

ABSTRACT

The use of taxonomies and ontologies as a foundation for enhancing textual information base access has recently gained increased attention in the field of information retrieval. The objective is to provide a domain model of an application domain where key concepts are organized and related. If queries and information base objects can be mapped to this, then the ontology may provide a valuable basis for a means of query evaluation that matches conceptual content rather than just strings, words, and numbers. This chapter presents an overview of the use of taxonomies and ontologies in querying with a special emphasis on similarity derived from the ontology. The notion of ontology is briefly introduced and similarity is surveyed. The former can be considered a generalization of taxonomy, while the latter can be seen as an interpretation where aspects of formal reasoning are ignored and replaced by measures reflecting how close concepts are connected, thereby significantly enhancing performance. In turn, similarity measures can be used in conceptual querying. Queries can be expanded with similar concepts, thereby causing query evaluation to be based on concepts from the domain model rather than on words in the query.

INTRODUCTION

Information retrieval (IR) deals with access to information as well as its representation, storage, and organization. The overall goal of an IR process is to retrieve the information relevant to a given request. The criteria for complete success are the retrieval of all the relevant documents stored in a given system and the rejection of all the non-relevant ones. Thus, the notion of relevance is at the heart of IR, and the retrieved documents are those found to be most relevant to the given request under the conditions available (representation, search strategy, and query). The set of relevant documents includes the documents that are likely to contain the information desired by the user and the selection of these is typically based merely on bag-of-words descriptions of the documents. Thus, finding relevant documents in an information retrieval system (IRS) obviously involves uncertainty, not only with regard to the interpretations that document descriptions represent but also due to possible interpretations of user requests.
A number of different approaches have been introduced over the years in order to handle the problem of uncertainty in IR. Among these are variants of the vector retrieval model, the probabilistic retrieval model, and the extended Boolean retrieval model. On important and very promising extension of the Boolean model is the fuzzy retrieval model. In this model, documents may be more or less relevant, given a query, and “uncertainty is an inherent part of the decision problem since the criteria for determining the answer are not altogether clear” (Lucarella, 1990).

To a large extent, every day users, as well as researchers and developers, have recognized the limitations of standard keyword-based IRSs, for example, search engines on the Internet. Even when advanced search options designed to facilitate increased precision and recall are available, they are not of much help to average users, who generally avoid them due to poor usability and high perceived difficulty of use (Bandos & Resnick, 2002).

One obvious alternative to standard keyword-based search is a less rigid natural language interpretation of queries, an idea that goes almost as far back as the idea of natural language processing does. Many natural language query approaches have been presented and it appears that recent approaches applying shallow parsing, for example, Penev and Wong (2006), might contribute to improved search.

Another important direction concerns domain knowledge processing. Most prominent among knowledge-based approaches is probably the use of taxonomies and ontologies, which recently has gained increased attention in the field of IR. Taxonomies and ontologies organize key concepts of an application domain and provide semantics through relations connecting concepts. Taxonomies, which can be considered controlled vocabularies that describe objects and relations between objects, are special cases of more general ontologies that include richer semantic relationships as well as rules for specification and formation.

The main focus in this chapter is query evaluation based on domain knowledge captured by taxonomies and ontologies. The general idea is to provide a mapping of concepts extracted from queries and documents into an ontology and to utilize this during query evaluation to obtain a matching on conceptual content rather than on just strings, words, and numbers. The emphasis in this chapter is on how queries are interpreted and evaluated and less on how queries are expressed. Query expressions may simply be a set of keywords or concepts, may apply some more advanced operators, or may consist of controlled natural language. The important issues are the identification of concepts in queries and documents as well as the conceptual level at which they are compared. Thus, to provide ontology-based querying, we need a methodology where we can extract key concepts from queries and documents as well as a technique to compare descriptions to evaluate the degree to which they match. The objective of the former is to provide conceptual descriptions of queries and documents and the purpose of the latter is to permit query evaluation based on these conceptual descriptions. As regards extraction of key concepts from queries and documents as descriptions that indicate semantic content, it should be noted that when refraining from full semantic analysis, it is possible to produce parsers that can perform efficiently on large volumes of data. A very simplified two-phase processing principle, for example, was implemented in the OntoQuery project (Andreasen, Jensen, Nilsson, Paggio, Pedersen, & Thomsen, 2002) where the first phase was noun phrase bracketing and the second phase an extract of concepts from the individual noun phrases. A naive but powerful second phase is to extract nouns and adjectives only and combine them into “noun CHR adjective” pattern concepts (where CHR represents a “characterized by” relation). Thus, for instance, for the phrase “the black dog,” the parser may produce the following concept: “dog CHR black.” We will not go into detail on key concept extraction in this chapter but rather turn our focus to evaluation with special attention to comparison of descriptions.

In order to compare conceptual descriptions, a means for measuring the extent of concept cor-