

Exploring the Use of Ontological Relations in Information Retrieval

Miao Chen

School of Information Studies

Syracuse University, Syracuse NY 13210

mchen14@syr.edu

ABSTRACT

The paper tries to explore how ontologies can contribute to information retrieval (IR) systems. Concepts and hierarchical relations of ontologies have been frequently used to expand concepts in queries, while non-hierarchical relations are seldom used in IR systems. We propose frameworks of integrating ontological relations in two parts of an IR system, query expansion and retrieved document organization. The effect of relations within the two frameworks will be examined and the methods are discussed as well.

Keywords

Ontological relations, information retrieval, query expansion, retrieved document organization.

1. INTRODUCTION

Ontology is “an explicit specification of a conceptualization” (Gruber, 1993), and it can be used as an approach of knowledge representation and organization. Information retrieval aims to find relevant resources that meet information need of users. Though having different goals, ontology and information retrieval have the same point of helping users finish their information tasks. According to Belkin (1993), information retrieval is one part of the information seeking process, which involves other parts supporting the process as well. Ontology, which explicitly represents domain knowledge, can serve as the system preparation part of information seeking. Therefore we can see the connection of ontology and information retrieval: it is that ontology represents and stores knowledge in a computable way and then information retrieval makes use of the computable knowledge to facilitate user information seeking.

Therefore it comes naturally that how we should use ontologies in information retrieval. Theoretically, ontologies are built to represent modeling of some domain; and practically, they are formalized by five kinds of components for computational purpose: concepts, relations, functions, axioms, and instances (Gruber, 1993). The five components, which together construct a domain knowledge base, can transmit knowledge to information retrieval systems thus the combination of the two is called “ontology-based information retrieval systems”.

Examining the use of ontology components in previous studies of ontology-based information retrieval systems, we find that the concepts and hierarchical relations are frequently used while other components are rarely applied in IR systems. The gap is reflected both in the query processing and retrieved document organization parts of retrieval. For the query processing part, in many times ontologies are used to expand queries, i.e. studies of Hersh (1995), Aronson (1997), and Wollersheim & Rahayu (2005). Correspondingly strategies and frameworks of selecting

appropriate terms to expand queries are provided in the studies. However, ontologies contain components representing knowledge beyond concepts and hierarchical relations. Relations, especially non-hierarchical relations, are not well explored so far and their potential use needs to be investigated.

In the retrieved document organization part, ontological relations are even more hardly involved. Retrieved documents are primarily organized in two ways: relevance ranking and similarity (or distance) based clustering (Pratt et al., 1999). Relevance ranking displays retrieved documents based on their ranking scores, but this representation is hard for users to find documents they need (Zamir, 1999; Pratt et al. 1999). As an alternative way of relevance ranking, clustering groups similar retrieved documents according to some criteria, such as relations between documents and query terms, predefined document attributes, and user-specified attributes (Zamir, 1999). There have been trials in representing documents by knowledge-based approaches, i.e. organizing documents by concepts and hierarchical structure of ontologies (Pratt et al. 1999; Chen & Dumais, 2000). But again, other types of relations are seldom taken into consideration in this part of retrieval. To the best of our knowledge, there has been no study on organizing retrieved documents based on relations (both hierarchical and non-hierarchical) between query terms.

This study is motivated by the two gaps mentioned above. We will propose frameworks of applying ontological relations (both hierarchical and non-hierarchical) in query processing and retrieved document organization responsively. It will show how ontological relations can be integrated to IR process and their effect will be examined as well. We choose to do experiments in medical domain because of convenience of ontologies and corpus.

2. IDENTIFICATION OF PROBLEMS

In some queries, relations between query terms are not specified, maybe due to users' search habit or their unknowing of relations between the concepts. Ontological relations can be used to clarify relations between query terms and thus reducing ambiguity of the query. On the other hand, ontological relations have structure for connecting relations, i.e. semantic relations in the UMLS¹ ontology is in hierarchical structure. It allows queries to be enriched by expanding specified relations. Therefore our first problem is to design a framework to take advantage of relations for query processing.

Secondly, we will design a framework of organizing retrieved documents by query term relations. On the one hand, there might be multiple relations between query terms. For example, food can

¹ UMLS is the abbreviation for Unified Medical Language System.

affect “disease”, and “food” may also *cause* “disease”. Retrieved documents can be organized based upon these two relations. On the other hand, the structure of ontological relations can be used to further group documents. A framework of retrieved document clustering will be described as solution to our second problem.

3. FRAMEWORK DESIGN

3.1 Semantic relation expansion

As abovementioned, users frequently input isolated concepts in the query without specifying relations between them. In this case, our first step is to identify candidate relations and expand queries based on the relations. For instance, in medical retrieval systems, users may input queries such as “liver cancer, food”. There are two concepts in the query “liver cancer” and “food” while their relation is missing. We will follow our framework and expand the query into an XML semantic query using a biomedical ontology UMLS.

The example query above is used to illustrate our framework: 1) two candidate relations are identified in the ontology, “cause” and “affect”. That is, food can *cause* cancer or food can *affect* cancer; 2) we could find the semantic family of the candidate relations from UMLS. For instance, the *affect* relation has one hyponym as *treat*, which means *treat* is a kind of *affect*. Thus we could say food can *treat* liver cancer; 3) using the family members of the candidate relations, such as hypernym, hyponym, and synonym, we can do relation-based query expansion. For example, *treat* as a hyponym of *affect* is added to original query. 4) a semantic tree for relations is built in XML format (as shown in Figure 1). It is annotated based on an ontology and the resulted query is called “ontology-annotated query. 5) therefore instead of using just the original query, now we have an array of queries, which is written as “Concept_1 Candidate_relation[] concept_2”.

```

<Query>
  <Concept_1>
    <Category>neoplastic process</Category>
    liver cancer
  </Concept_1>
  <Candidate Relations>
    <Concept_2_Concept_1>
      <Relation 1><b>cause</b></Relation 1>
      <Relation 2><b>affect</b>
        <Hyponym><b>treat</b></Hyponym >
      </Relation 2>
    </Concept_2_Concept_1>
  </Candidate Relations>
  <Concept_2>
    <Category>food</Category>
    food
  </Concept_2>
</Query>

```

Figure 1. Ontology-annotated Query in XML

In the expanded query, there are four type of information included:

1. Concept boundary identification: identify boundary of concepts in the query, to break query terms into UMLS concepts.
2. Category information: find semantic types of the concepts from the UMLS ontology.

3. Candidate relation information: For the targeted concepts pairs, find candidate relations between them based on their semantic types and then tag them in XML.
4. Candidate relation family information: Relations in UMLS ontology are arranged in hierarchical style, and related relations such as hyponyms and synonyms of the candidate relation are selected for relation expansion.

3.2 Relation-based organization of retrieved documents

In search results, we will implement real-time parsing based on the concepts and relations from the query XML file for each retrieved document. The major steps are: 1) we identify candidate sentences which contain related concepts and relations of the expanded query. The candidate sentences may have both candidate concepts and at least one candidate relation term; 2) syntactic level natural language processing (NLP) algorithm will be used to parse each candidate sentence in the documents; 3) by combining the result of steps 1 and 2, we derive ontology-annotated documents from search results.

Based on the document parsing result, we can group similar search results together. Retrieved documents with the same relations between concepts are organized together. The search results are organized based on candidate relations as well as related relations. In this example, retrieved documents are categorized into two sets, the “*cause*” set and the “*affect*” set. Under the two categories, documents are further divided according to the UMLS relation structure. For example, the “*treat*” set is arranged under the “*affect*” set, the same way as in the ontological relation structure. Documents of the same category do not necessarily share the similar word distribution, but they will have the same concept relation.

4. EXPERIMENTS AND EVALUATION

To examine the effectiveness of our first framework, we will conduct experiments in three IR systems. The three systems are: 1) a baseline IR system with no query expansion; 2) the baseline IR system with concepts expanded in queries; 3) the baseline IR system with both concepts and relations expanded. The measurement of performance adopts the Cranfield model, including recall and precision rates (Cleverdon & Mills, 1963).

We will use the OSHUMED set (Liu et al 2004), which includes corpus and queries in the medical domain, for the IR systems. It contains documents from Medline, user-specified queries, and relevance judgment by domain experts. Recall and precision rates will be compared through statistical analysis to examine whether ontological relations makes a significant different in retrieval.

For the second framework, we examine it through interviewing user opinions. We use the third IR system above and compare two cases: 1) retrieved results organized in relevance ranking; 2) retrieved results organized based on relations. Users will be asked to use both two organization methods and provide their experience and feelings towards them. Content analysis will be conducted on the interview data to understand the strengths and weaknesses of relation-based organization from user perspective.

5. REFERENCES

- [1] Aronson, A.R., & Rindflesch, T.C. (1997). Query expansion using the UMLS metathesaurus. Proceedings of AMIA Annual Fall Symp, 485-9.
- [2] Cleverdon, C.W., & Mills, J. (1963). The Testing of Index Language. Devices. Aslib Proceedings, 15(4), 106-120.
- [3] Gruber, A. (1993). A translation approach to portable ontology specification. Knowledge Acquisition, 5(2), 199-220.
- [4] Hersh, W., S. Price, Donohoe, L. (2000). Assessing thesaurus-based query expansion using the UMLS metathesaurus. Proceedings of American Medical Informatics Association Symposium.
- [5] Liu, Y., Loh, H. T., & Tor, S. B. (2004) Proceedings of the Singapore MIT Alliance Symposium
- [6] Pratt, W., Hearst, M.A., Fagan, L.M. (1999). A knowledge-based approach to organizing retrieved documents. Proceedings of AAAI 1999.
- [7] Wollersheim, D., & Rahayu, W.J. (2005). Ontology based query expansion framework for use in medical information systems. Journal of Web Information Systems, 1(1), 1-17.
- [8] Zamir, O., & Etzioni, O. (1999). Grouper: A dynamic clustering interface to web search results. Proceedings of the Eighth International WWW Conference.
- [9] UMLS: <http://www.nlm.nih.gov/research/umls/>