Role of Ontology in Information Retrieval

WU Dan¹, WANG Hui-lin²

¹. Department of Information Management, Peking University  Beijing  100871  China
². Training Center for Post-graduates, Institute of Scientific & Technical Information of China  Beijing  100039  China

Abstract  Based on the comparison between ontology and thesaurus, and the analysis of an ontology-based Information Retrieval (IR) model, the potential advantages that ontology may contribute to IR are analyzed. Then a general architecture of ontology-based Information Retrieval System (IRS) and the approach of constructing it are presented. Based on the researches, the role of ontology in IR is summarized from four aspects and a typical system called Textpresso is analyzed. Finally, a conclusion is drawn that utilizing ontology is the trend of IR and can really improve the IRS.

Key words  ontology;  Information Retrieval (IR);  Information Retrieval System (IRS)

The growing digital contexts generate huge amounts of information. The potential for exchange and retrieval of information is vast and daunting. When talking about Information Retrieval (IR), users are now experiencing huge difficulties in finding precisely what they are looking for, among the tons of documents available. The key problem in achieving efficient and user-satisfactory retrieval is the development of a search mechanism to guarantee delivery of minimal irrelevant information (high precision) while insuring relevant information is not overlooked (high recall). The traditional solution employs keyword-based search that returns results based on user specified keywords matches without any regard for the concepts in which the user is interested. But many documents convey desired semantic information without containing these keywords. So new retrieval technology must be reconstructed or introduced into existing information retrieval, namely more exact and faster retrieval tools should be developed.

One can overcome the above problem by indexing documents according to meanings rather than words, although this will entail a way of converting words to meanings. Ontology is such a tool that can describe word meanings and relations. Nowadays, ontology is introduced in Information Retrieval System (IRS) for the purpose of solving the problem of semantic understanding. An ontology is “a formal, explicit specification of a shared conceptualization”[1]. When used in the Artificial Intelligence or Knowledge Representation Community, the word ‘ontology’ has been used to refer to things that have a rich and formal logic-based language for specifying meaning of the terms. Therefore, we may understand ontology as a domain information model, comprehensible both by humans and computers. With regard to the converting words to meanings, the key issue is to identify appropriate concepts that both describe and identify documents, as well as language employed in user queries.

1 Using Ontology for IR

1.1 Relations between Ontology and Thesaurus

The traditional information retrieval is based on thesaurus that represents a field of specific knowledge through its conceptual structure. This conceptual structure provides a semantic organization by making explicit the conceptual relations and restricting the meaning of the terms that represent them. The field of knowledge is structured based on hierarchical, associative equivalence-based conceptual relations. When used during retrieval and searching, thesauri are useful in bridging the gap that exists between the metadata provided by the indexer and the concepts presented by a searcher[2]. Functions of the thesaurus in IR can be summarized as standardizing user’s query expression, and adjusting the scope of IR automatically.

Compared with thesaurus, ontology has a rich and formal logic-based language for specifying meaning of
terms. The conceptual relations represented in an ontology are extremely varied and they depend on the field of knowledge to be structured. There is a growing number of voices claiming that the use of ontology for IR is an efficient method that can be superior to others in both precision and relevance. We can see that from the comparison between ontology and thesaurus.

The relations between thesaurus and ontology in an IRS can be described as the following formula:

\[ \text{Ontology} = \text{Thesaurus} + \text{Instances} + \text{Rules} \]

In an IRS, a thesaurus permits to organize all the concepts a user may use within an application and to build generalization hierarchies from the gathered concepts. Users’ requests become searches by concepts, whatever terminology they used. Ontology is about organizing, collecting and managing knowledge in an IRS. They have many similarities that they may both describe domain specific knowledge and contain terms and relations, however, building ontology for large domains is a costly affair and in many domains thesauri have been built. So thesaurus is a basis for the construction of an ontology. The main differences between ontology and thesaurus are instances and rules.

On the one hand, the method of building an ontology has two ways that are top-down and bottom-up. Top-down method organizes concepts from upper ontology such as SUMO, SENSUS, etc. Bottom-up method extracts instances from documents. The ‘instances’ in the ontology provide ‘solutions’ to the users, solutions made of explanatory texts and transaction possibilities. Each instance is indexed using a combination of concepts chosen in the hierarchies from the thesaurus.

On the other hand, most thesauri have limited capability of knowledge representation, for there is no relevant software to realize the function. While ontology provides such strong knowledge representation language as XML, RDF, OIL, OWL, etc. and on the basis of the languages, some ontology learning tools such as Text-To-Onto, OntoLearn, Hasti have been put forward to make reasoning on new knowledge for the purpose of expressing abundant concepts and making good human-machine interaction with users\(^3\text{-}^5\). From there, rules are added in the IRS to model in depth the details of a search. Rules would define search’s actions based on information from the database and about the user.

### 1.2 Potential Advantages of Ontology in IR

An ontology is constructed with the aim of sharing and reusing stored information, which, having been formalized, can be interpreted by both persons and computers.

![Ontology-based IR model](image)

Fig.1 depicts the ontology-based IR model. On the one hand, user’s information need turns into query expression that computer can read through the shared ontologies. Then in order to improve the recall we should expand the queries. This process is also based on the relevant concepts in the shared ontologies. The query formulation can be supported by interacting with one or several ontologies, initially selected according to the domain of the information need – an important element in Bates’ model\(^6\). On the other hand, the documents should be annotated and indexed through the same ontologies. The representation of the documents includes logical assertions that make it an integrated part of the ontological structure. In this way, the matching process can be generalized to an exploration process that can be implemented in a number of different ways, depending on the form and logical interpretation of the query.

Then, we summarize the potential advantages that ontology may contribute to IR as follows:

1) Exactly comprehend user’s information need: on the basis of ontology, computer can know the exact information that users want. Ontology is constructed to share and reuse stored information. It contributes to avoiding ambiguities and standardizing user’s query expression. This will greatly improve the IR precision.
2) Exactly understand documents’ meanings: with ontology, the semantic understanding of documents can be realized. When talking about the organization of documents, based on ontology, semantic annotation and comprehension may be carried out for the purpose of improving the IR precision.

3) Knowledge representation: ontology may provide a general knowledge representation to users. When searching for data, users may also realize knowledge-based retrieval.

4) Strong reasoning capability: due to the use of Artificial Intelligence Technology, the search process of ontology-based IR has excellent reasoning capability which make the whole search process intelligent, such as expanding relevant concepts and filtering irrelevant concepts etc.

5) Heterogeneous knowledge integration: inconsistent representation of information is the main problem of heterogeneous data sources, including name conflict and structure conflict etc. Ontology can establish a set of shared representation structure of terms and information. Heterogeneous information turns into isomorphic information on the basis of this structure and then realizes the heterogeneous knowledge integration and retrieval.

2 Architecture of Ontology-based IRS

Nowadays many researches on ontology applied to IRS are becoming more and more popular. However, such IRS is almost being experimented in the laboratory and few ontology-based IRS is launched into commercial application. Among them, in order to make use of expression and inference of ontologies, Ontobroker system stresses that semantics of Web context knowledge should be described by formal ontology\(^7\). By integrating ontologies with lexicon, and regarding ontology as domain vocabulary with semantics, OntoSeek system attempts to provide users with interactive semantics query interface\(^8\). An information retrieval server based on ontology and multi-agents integrates several kinds of agents, such as information processing agent with mobile ability, and uses ontologies to classify the domains of documents and assist user to normalize their queries\(^9\). On the basis of the existing researches, we present the general architecture of ontology-based IRS and the approach of constructing it.

2.1 Design for Ontology-based IRS

Ontology has good conceptual structure and is especially used in knowledge-based retrieval. The basic design idea of ontology-based IRS can be summarized as follows:

1) Constructing domain ontology under the help of domain expert. The construction of ontology is the first step of IR. In view of the current researches, the approaches of constructing ontology are manual work, machine learning, and semi-automatic construction, etc. But whatever the approach is, domain experts’ participation is indispensable.

2) Selecting data from information resources, and storing data into databases including relation database and knowledge base etc., in accordance with the domain ontology. This process is the classification on resources on the basis of ontology, including the construction of database and knowledge base.

3) Converting queries from interface into prescriptive formats according to the domain ontology. With the help of ontology the IRS matches queries in accordance with data from database. This friendly user-machine interface may well understand users’ requirements and carry on precise search.

4) Returning search results to users after customizing disposal. During this process, we can filter some results according to users’ personalized requirements and only display the information that interests users.

2.2 Architecture of Ontology-based IRS

Fig.2 shows the general architecture of ontology-based IRS from which we can see that ontology plays an essential role in the IRS.

![Fig.2 Ontology-based IRS](image-url)
judgement on retrieval from ontology, judging to which fields the keywords belong and then returns some information, such as relevant concepts or child concepts, to users for selection. Then the IRS makes exact search.

2) In classifier: traditional classification is a catalogue as used in Yahoo. This manual work can’t deal with mass information so Natural Language Processing (NLP) technology and ontology are introduced to IRS. We hope to organize information around ontology, that is to say, we may classify documents by ontology because ontology provides a standard classification that helps us to know to which field or concept the documents belong.

3) In filter: personalized IR becomes more and more popular nowadays, search engine as google has provided Personalized IR to users. Under the help of ontology, the IRS can learn user’s interest and filter out the results that are not consistent with user’s interest. There are three methods of ontology-based filter technology. Firstly, ontology divides different domains and exactly expresses domain words, this may help users to select domain ontology and set up their own user model on the basis of their interests. Secondly, the IRS uses XML to describe user model under the ontology that makes user model more semantic and to be better processed by computer. Thirdly, the IRS adopts machine learning to mine and find user’s interest, and then mappings it to ontology, this may help the IRS to know more about user’s behavior and dynamically update user model.

3 Functions of Ontology in IR

From all the analysis of the above three sections, we may summarize how ontology contributes to IR from four aspects.

3.1 Query Expansion

We may take a user query as our starting point for detecting which lies at the heart of the query (the information need of the user), and subsequently expand on it with variations that allow us to find documents that contain the most relevant information. However, usually the query does not supply sufficient data to distinguish between relevant information and that is thematically related to it\[^{10}\].

The IRS making use of ontology and alternatively of thesaurus is based on term expansion. The lexis identified as relevant in a document is used as a starting point for establishing correlations with concepts or other lexical units that represent these concepts or related concepts. From a linguistic point of view, we could say that from one word, syntax unit or set of words from the query, the system would search in the ontology for other words or phrasal units expressing related concepts. This proximity would correspond in linguistics (and as a general rule) to synonyms or variations, with hyperonyms or concept classes, and with cotypeonims or concepts that belong to a same class.

WordNet is one of the most popular resources used as ontology in query expansion based IR. The way lexical hierarchy structures the lexis of languages is based on the concept of synset or synonym set. For each query vector the IRS automatically associates sets of synonyms\[^{11}\]. The only alternatives to generalist ontologies such as WordNet that are capable of producing better results in terms of precision and more importantly in terms of relevance are based on confining the scope of the ontology or lexical structure. There are cases of ontologies generated from words and from the concepts used within an organization (company, public sector body, work network). These experiments that work with controlled lexis and little data are not easily exported to other organizations, languages or open IR on the Internet\[^{12}\].

On the other hand, some IRS confines its reach to a specialized field. These fields generally have a consolidated tradition within the scientific community, and therefore an ample consensus on concepts, conceptual relations and standardized terminology tends to exist within them. A query expansion project based on association to concepts in the UMLS is the case\[^{13}\].

3.2 Information Abstraction

Ontology permits multiple inheritances, in other words, each concept is able to receive properties and conceptual relations from more than one concept that is higher up in the hierarchical scale. In addition, if ontology is associated with a terminological database, we would obtain information on linguistic variants and equivalents for the terms related to each concept. Furthermore, all these information parameters specified (definition, categorization, hierarchy, properties and inheritance) are coded formally and explicitly; in another way, it will be possible for them to be understood by humans and be interpreted or decoded by computers. Therefore, the efficiency of
information abstraction will be improved greatly by applying ontology to information abstraction tools in IR.

3.3 Semantic Formalization

Semantic formalization makes it possible to obtain a logical, coherent representation of the conceptual structure, thereby generating a computer interpretation and facilitating the meaning of the type of relation existing between two or more concepts. In the process of constructing IR tools, ontology contributes an explicit declaration of the conceptual relations within a particular field through the semantic formalization of the conceptual structure.

The confluence of both characteristics in constructing the same facet poses problems in terms of structural coherence, since in theory, a facet must be constructed based on the application of a single subdivision criterion applied to a specific concept. However, this not only affects the ‘purist’ construction of thesauruses, but also, in more complex cases, it may be an erroneous guide in the IR by the end-user.

For this reason, the semantic formalization obtained through ontologies has a clear use in verifying the coherence of conceptual relations. It is even worthwhile asking oneself whether ‘ontology conceptual relations’ within the context of documents would also make it possible for the labeling of those relations to be transferred to the user interface, thereby making the conceptual structure more transparent when carrying out the IR.

The ontology description languages such as Ontolingua, CycL, Loom which come from AI, and XML, RDF, OIL, etc. which come from Internet, all provide uniform semantic formalization and have strong reasoning mechanism to represent knowledge. In short, ontologies permit a deeper semantic development to be obtained, as they provide a logical, formal description of the information they store. And that can be interpreted by both humans and computers.

3.4 Natural Language Understanding

Ontology helps IRS to understand user’s requirement and mapping it to information resources. It reasons the meaning of a concept via comparing concepts’ logic structure. For example, if concept B is one of the requirements of being concept A, ontology can make reasoning that concept B is an instance of concept A and then concept B is automatically put under concept A. When user’s query is composed of natural language, along with Natural Language Processing (NLP) technology, the IRS separates the query into terms that appear in the domain ontology. Under this circumstance IR may realize natural language understanding. On the other hand, ontology also makes documents clustering by their semantic information to form some semantic clusters, so that IR may conduct precise matching. The construction of semantic clusters is based on NLP technology, but domain ontology may optimize such processes as lexical analysis, annotation and parsing of the NLP. For example, concepts in the domain ontology help to recognize person’s name, names of organizations, technical terms, etc. This is the basis of lexical analysis. Moreover the information and relations about concepts in the domain ontology also help to confirm how to make documents clustering.

4 A Case Study

4.1 Textpresso Ontology

Textpresso is an ontology-based information retrieval and extraction system for biological literature developed by the California Institute of Technology. Textpresso’s two major elements are a collection of the full text of scientific articles splitting into individual sentences, and the implementation of categories of terms for which a database of articles and individual sentences can be searched. The categories are classes of biological concepts (e.g., gene, allele, cell or cell group, phenotype, etc.) and classes that relate two objects (e.g., association, regulation, etc.) or describe one (e.g., biological process, etc.). Together they form a catalog of types of objects and concepts called an ontology.

After this ontology is populated with terms, the whole corpus of articles and abstracts is marked up to identify terms of these categories. There are three corresponding categories in Textpresso ontology. The first group of categories consists of biological entities; the second group of categories comprises terms that relate two objects (e.g., association, regulation, etc.) or describe one (e.g., biological process, etc.). Together they form a catalog of types of objects and concepts called an ontology.

The Textpresso ontology is organized into a shallow hierarchy with 33 parent categories. The parent categories may have one or more subcategories, which are specializations of the parent category.

4.2 Functions of Ontology in Textpresso

4.2.1 Query Expansion

Textpresso ontology is organized into categories
that facilitate broader searches of biological entities. To be useful, it should also contain other categories that are not composed of biological entities, but describe relationships between entities. Textpresso offers the user an opportunity to query the literature in the framework of the ontology such that it returns sentences for inspection by the user. Searching the corpus of text with a combination of categories of an ontology could facilitate a query that contains the meaning of a question in a much better way than with keywords alone. For example, there is a ‘gene’ category containing all gene names and a ‘regulation’ category including all terms (nouns, verbs, adjectives, etc.) describing regulation. To search for two instances of the category gene and one instance of the category regulation in a sentence, the search engine will probably return a sentence describing a gene regulation. The search could then be limited by using a particular gene name as a keyword to get a list of genes that regulate or are regulated by that particular gene.

4.2.2 Information Abstraction

Distilling information from scientific papers manually is expensive and slow, if the full text is available to the researcher at all. Textpresso allows users to locate efficiently information of their interests. Users can determine whether a query is to be met in the whole publication or in a sentence. These options make the IRS powerful; for example, if a query is met in the whole article, the search has the function of text categorization, while meeting it in a sentence aims at extracting facts, which can be viewed in the context of a paragraph. The scope of a search can be confined to full text, abstract, title, author, year, or any combination thereof, for document searches as well as sentence searches. Ontology allows word meaning to be queried, it is possible to formulate semantic queries.

4.2.3 Semantic Formalization

Textpresso splits papers into sentences, and sentences into words or phrases. Each word or phrase is then labeled using the XML according to the lexicon of the ontology. The labels fall into 33 categories that comprise the Textpresso ontology.

For example, to annotate a sentence “In par-1, par-4 and par-3 mutant four-cell embryos, MEX-3 is present at high levels in all cells, indicating that activity of these par genes is required to restrict MEX-3 to the anterior.” The computer identifies terms by matching them against regular expressions and encloses them with XML tags. The tag <text> serves as a containment of terms not semantically marked up. These tags will be used for a repeated reevaluation of the lexicon, as these terms can be easily pulled out and analyzed. A list of the most frequently missed terms is then produced and included in the lexicon for the next markup.

Textpresso then indexes all sentences with respect to labels and words to allow a rapid search for sentences that have a desired label and/or keyword. The marked-up text is stored in a database and can be queried.

The above sentence is annotated as follows:

```
<sentence id='s38'>
  <preposition type='in'>in</preposition>
  <gene>par-1</gene>
  <punctuation type='comma'>,</punctuation>
  <gene>par-4</gene>
  <conjunction type='coordinating'>and</conjunction>
  <gene>par-3</gene>
  ……
  <spatial_relation type='unknown'>anterior</spatial_relation>
  <text>.</text>
</sentence>
```

4.2.4 Natural Language Understanding

With the help of ontology, Textpresso can deal with some semantic information as well as experts do, which greatly improve search efficiency. For example, extraction of particular biological facts, such as gene-gene interactions, can be accelerated significantly by ontologies, with Textpresso automatically performing nearly as well as expert curators to identify sentences; in searches for two uniquely named genes and an interaction term, the ontology confers a 3-fold increase of search efficiency.

4.3 Evaluation of Textpresso

In Ref.[14], to directly assess how Textpresso performs, the authors tested the accuracy of a search combining word categories and keywords to retrieve sentences containing genetic interaction data. They formulated a Textpresso query that searched for the presence of at least two genes mentioned by name and at least one term belonging to the ‘regulation’ or ‘association’ word categories. A total of 178 sentences were matched for this query in the eight journal articles. A human expert assessed the returned sentences and
determined that 63 sentences contained gene-gene interaction data. The same set of journal articles had been independently manually evaluated for their description of genetic interactions, and 73 true sentences were identified.

So we can say that textpresso is a system with high recall and as high precision as possible.

5 Conclusions

In this paper, we mainly analyze how the advantage of ontology is embodied in IR in theory. We present the general architecture of ontology-based IRS and the basic design idea for constructing it. We summarize the role of ontology in IR as query expansion, information abstraction, semantic formalization and natural language understanding. The Textpresso system also helps us to verify the four functions that ontology contributes to IR. Therefore utilizing ontology is the trend of IR and can really improve the IRS. However, a lot of application work is left to future researches. On the basis of the analysis in this paper, further researches on ontology-based IRS may have emphases.

References


Brief Introduction to Author(s)

WU Dan (吴 丹) was born in Hubei Province, China, in 1978. She got her Bachelor’s degree in information science and Master’s degree in computer application both from Huazhong Normal University in 2001 and 2004, respectively. She is now pursuing her Ph.D. degree in Peking University. Her research activities have been concerned with information retrieval and knowledge engineering.

WANG Hui-lin (王惠临) was born in Beijing, China, in 1948. He got his Ph.D. degree from Université de Nancy II, France, in 1984. He is now a doctoral supervisor of Peking University and senior research fellow of Institute of Scientific & Technical Information of China. Since 1984, his research activities have been concerned with machine translation, natural language processing.