1-7-2005

Ontology-based resource descriptions for distributed information sources

Hui Yang
University of Wollongong

Minjie Zhang
University of Wollongong, minjie@uow.edu.au

Ontology-based resource descriptions for distributed information sources

Abstract

Content-based resource description is the key to find appropriate information sources that are most likely to contain the relevant documents for a given user query. However, semantic heterogeneity makes it difficult to acquire accurate and meaningful resource descriptions from distributed, heterogeneous information sources. To address this problem, we describe an ontology-based approach which uses domain-specific ontologies to extract content-related information from information sources, and to generate ontology-based resource descriptions. The preliminary experimental results demonstrate that our ontology-based approach could improve selection accuracy.

Keywords

content-based retrieval, information resources, ontologies (artificial intelligence)

Disciplines

Physical Sciences and Mathematics

Publication Details

This paper originally appeared as: Yang, H & Zhang, M, Ontology-based resource descriptions for distributed information sources, Third International Conference on Information Technology and Applications, 4-7 July 2005, 1, 143-148.

Ontology-based Resource Descriptions for Distributed Information Sources

Hui Yang and Minjie Zhang
School of Information Technology and Computer Science
University of Wollongong
Wollongong, 2500, Australia
{hy92, minjie}@uow.edu.au

Abstract

Content-based resource description is the key to find appropriate information sources that are most likely to contain the relevant documents for a given user query. However, semantic heterogeneity makes it difficult to acquire accurate and meaningful resource descriptions from distributed, heterogeneous information sources. To address this problem, we describe an ontology-based approach which uses domain-specific ontologies to extract content-related information from information sources, and to generate ontology-based resource descriptions. The preliminary experimental results demonstrate that our ontology-based approach could improve selection accuracy.

1. Introduction

Nowadays, the World Wide Web has achieved an impressive success with over 8 billion pages available on the web, but it has left web users the heavy burden of accessing to and searching huge amounts of information. To help users find useful information, some information retrieval (IR) tools (e.g., Search Engines) are developed to support effective search and retrieval for the information of interest. Content-related resource description plays an important role in intelligent information retrieval. Especially on the World Wide Web, meaningful resource descriptions representing the contents of distributed information sources are the key to locate the potential useful information sources that might contain relevant documents with respect to a user's query. This is because the selection of suitable information sources is based on the relevance degree of resource descriptions to the query. However, due to the inherent semantic heterogeneity in information sources, the acquisition of appropriate and accurate resource descriptions remains a problem.

The problem of semantic heterogeneity is always well known in a distributed, heterogeneous information environment [4]. It occurs when the contexts of information sources do not use the same interpretation of the information (e.g., the use of different terms to refer to the same concept). Hence, in order to select appropriate information sources, semantic interoperability is required so that the meaning of the information that is required by the user can be understood across information sources. A domain-specific ontology is a shared and common understanding of a specific domain that can be communicated across people and systems. It can be defined as a formal, explicit specification of a shared conceptualization [3]. The interoperability feature of ontologies provides a possible solution to overcome the problem of semantic heterogeneity. In this paper, we have developed an ontology-based model that uses concepts and their semantic relationships in domain-specific ontologies to extract content-related information from information sources, and to generate resource descriptions in terms of ontologies.

2. Domain-specific ontologies

In this paper, our work focuses on the use of domain-specific ontologies for resource descriptions. The basic idea behind this method is that ontologies serve as a means for establishing a conceptually concise basis for communicating knowledge. A domain-specific ontology is a shared and common understanding of a particular domain. It includes a representational vocabulary of terms that are precisely defined, and specified with relationships between terms. These terms may be considered as semantically rich metadata to capture the information contents of the underlying information sources. The use of ontologies with these semantically rich descriptions offer a promising way to deal with semantic heterogeneity in information sources mentioned in the introduction.

Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05)
0-7695-2316-1/05 $20.00 © 2005 IEEE

For the purpose of this paper, we will first introduce the most important components in a domain-specific ontology. Figure 1 shows a simple example of a '*University Department*' ontology. Concepts are linked by lines with different shapes that denote various kinds of relationships.

A domain-specific ontology specifies a conceptualisation of a domain in terms of concepts. Each *concept* represents a class for a specific set of entities. It is characterised by a unique label name in the ontology, and is usually expressed as a combination of synonymous words. For example, the concept '*Research Centre*' has a synonymous list which consists of '*Research Group'*, '*Research Unit'*, and '*Research Project'*.

The concepts are typically organised into a *taxonomy tree*, where each node represents a concept. Concepts are linked together by means of their semantic relationships. The set of concepts together with their links form a semantic network. Various kinds of semantic relationships are maintained between the concepts. Among these, the most relevant for our purposes is the *Part-Of* (S*ubsumption*) relationship, which allows a set of concepts to be organised according to a generalization hierarchy. For instance, the concept '*People'* is more general than its subclass concept '*Staff*'. In addition, in the hierarchical mechanism, there is the *context-related* relationship which links a set of non-hierarchical concepts together. These concepts are semantically related in a certain context. For example, the concepts '*Research Centre*', and '*Academic Staff*' are semantically related in the research activities of the school.

Another important relationship associated with concepts is the *Instance-Of* relationship, which denotes the concrete occurrence of abstract concepts. For example, the concept '*Research Area*' is associated with a set of concept instances such as '*Network Security'* and '*Machine Learning*'.

3. Concept-based resource descriptions for information sources

Our approach to generate resource descriptions which capture meaningful information in information sources, is to use concepts from domain-specific ontologies as the vocabulary to characterize the information. According to the concepts in the ontologies, the meta-information extracted from web documents in an information source is used to classify the information source into one or more topic domains. In each topic domain, relevant concepts are identified and stored in the resource description as well as their semantic relations.

3.1. Content-related metadata extraction of web resource

Content-related metadata plays an important role in information retrieval systems. Meaningful metadata describing the contents of web resources is the key to effective search and retrieval of information. Our metadata extraction method is text-based, which

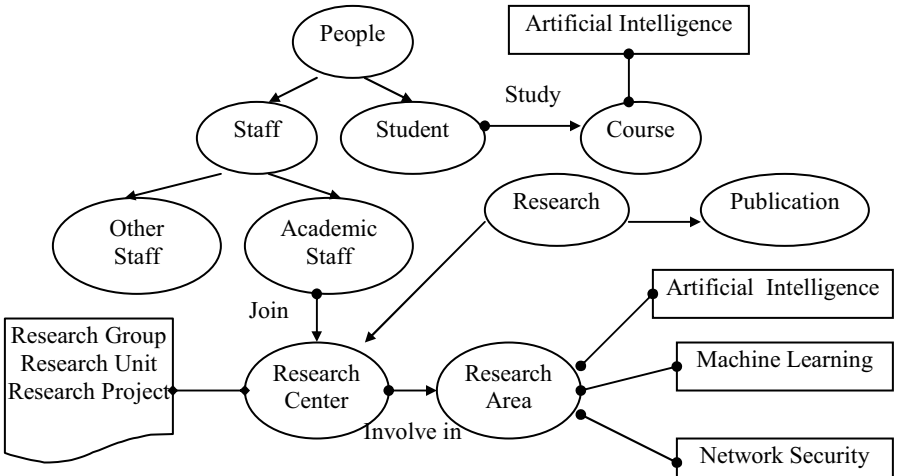

Fig. 1 A small part of a '*University Department*' ontology

mainly focuses on the content-related information found in HTML tags such as the title or a heading element, and metatags for keywords and descriptions. They are always the primary source of text features. In addition, the hyperlink structure of the web can also be exploited by using the anchor text and the metatag contents from linking documents as another source of text features. However, the importance degrees of these text features are different in the resource description. All extracted text features together with different weights of importance degree are concatenated into a

single representative document as the meta-information of the information source.

3.2 The generation of ontology-based resource descriptions

In our approach, the meta-information of each information source is structured, and domain-specific ontologies are used to describe the semantics of the meta-information of the information source. In fact, the generation of resource descriptions might be divided into two stages: First, at the domain-level stage, the system identifies suitable topic domains which might cover the subject content of the information source; Second, at the concept-level stage, for each topic domain, the system maps the meta-information to the concepts and semantic relationships in the corresponding domain ontologies.

3.2.1. The selection of suitable domain topics

At the domain-level stage, the first issue that we address is to classify an information source into one or more related topic domains. The classification method we use in this paper is based on Naive Bayes leaning technique [5], one of the most popular and effective text classification methods. In order to distinguish the appropriate topic domain from a set of domains in the classification schema, a set of features that have enough distinguishing power (i.e., in text classification, the features are the words that are strongly associated with one specific category) are needed for the classifier. In this paper, the acquisition of these features related to a specific domain is accomplished through the textual content of the corresponding domain ontology since an ontology represents a collection of common terms that are particularly useful to conceptualize a knowledge domain. As explained previously, a concept typically has a label name, a list of synonymies, and a possible set of associated concept instances. We treat the label name, the synonymous list, and the concept instance set as the *textual content* of this concept. Consequently, the textual content of an ontology, in practice, is the combination of textual contents of all the concepts contained in this ontology.

In the classification schema, each topic domain is associated with a feature space F, $F = \{f_1, f_2, \cdots, f_m\}$, which is used to construct the Naive Bayes classifier. The probability $P(f_i \mid T)$ of a feature f_i (word) in a topic domain T is estimated by exploiting the frequency of the feature that occurs in the textual content of the domain ontology.

Given a set of topic domains with the Naive Bayes classifier, the similarity of a topic domain T_i to an information source S is the posterior probability $P(T_i \mid S)$. Using Bayes's theorem, the posterior probability $P(T_i \mid S)$ can be denoted as

$$P(T_i \mid S) = \frac{P(S \mid T_i)P(T_i)}{P(S)} \propto P(S \mid T_i)P(T_i) \qquad (1)$$

where $P(S)$ can be ignored because it is just a normalizing constant. $P(T_i)$ is the prior probability that the topic is relevant. Here, we made the simplifying assumption that the prior probability of relevance $P(T_i)$ is a constant for all topic domains. As a consequence, we focus our attention on the remaining term $P(S \mid T_i)$.

Let $S = \{d_1, d_2, \cdots, d_n\}$ be the text features extracted from the information source S (recall Section 3.1), where each textual feature $d_i (1 \le i \le n)$ is associated with a weight of importance degree $w(d_i)$. So with Naive Bayes assumption that the probability of each word in a domain is independent of the word's context and position in the domain, the posterior probability $P(T_i \mid S)$ can be described as

$$P(T_i \mid S) = \prod_j P(d_j \mid T_i) w(d_j) \qquad (2)$$

where $P(d_j \mid T_i)$ can be obtained from the feature space F associated with the domain.

Once the similarities of topic domains to the information source are acquired, a *k*-nearest neighbor window method is used to assign relevant topic domains to the information source. Consider such a scenario where some large-scale information sources contain the documents of one or more topic domains. We use a window to capture the topic domain as many as possible. The window is defined as follows:

$$\frac{P_{\max}(T \mid S)}{P(T_i \mid S)} \le 1 + \varepsilon \qquad (3)$$

$$\frac{P(T_i \mid S)}{P_{\max}(T \mid S)} \ge 1 - \varepsilon \qquad (4)$$

where $P_{\max}(T \mid S)$ is the maximum of the posterior probabilities of all the topic domains, and ε is the parameter of window size. As long as the posterior probability of topic domain T_i satisfies all of the above conditions, topic domain T_i will be chosen as an appropriate topic domain for the database *S*.

3.2.2. The generation of an ontology-based resource description

Once suitable topic domains are chosen, the next step is to map the meta-information of the information source to the relevant concepts and semantic relationships in the corresponding domain ontologies. In this paper, our approach to the creation of an

ontology-based resource description can be decomposed into the following three steps:

- *STEP 1*: to classify web pages into ontology concepts based on the contents of web pages.
- *STEP 2*: to determine the semantic relationships between the discovered concepts by using the hyperlink structure between the involved web pages.
- *STEP 3*: to add concept instances to the corresponding concepts that have been detected in Step 1 by performing a full-text search in the text body of web pages.

The mapping of concepts: to map web pages to the appropriate concept nodes in the ontology taxonomy, we use metadata (e.g., the title and keywords) of the web page to match textual contents of concepts in the ontology (recall Subsection 3.2.1). Similarity measurement between the metadata X of a web page and the textual content Y of a concept is calculated using the Dice Coefficient:

$$Simi(X,Y) = 2\frac{X \cap Y}{X \cup Y} \quad (5)$$

The more the words in the metadata of the web page occur in the textual content of the concept, the greater the similarity score will become. We assign the concept with the biggest similarity score to the web page.

The mapping of semantic relationships: once suitable concepts that the web pages are related to are detected, the following work is to find the actual relationships between these concepts in that it is likely that only part of the concepts in the ontology are reflected in the documents of the information source. As we known, in the ontology, concepts are linked together by means of their semantic relationships. Therefore, one efficient way to locate relationships between detected concepts in the resource description is to take advantage of the semantic relationships between linked web pages in the information source. We exploit the information source's implicit semantic structure through following a set of hyperlinks. The hyperlink structure and the anchor texts contained in the hyperlinks are useful for analyzing semantic relationships between the concepts that the linked web pages belong to.

To identify the proper relationships between the concepts in the resource description, we made some assumptions on the basis of the semantic relationships in the ontology taxonomy. These assumptions are expressed with sufficient information which makes it possible to perform the inference on the relationships between concepts.

Example 1: Assume that document A matches concept X and document B matches concept Y. If document A is linked with document B by a hyperlink, and there exists a relationship (e.g., *Part-of* or *Context-related* relationship) between concept X and concept Y in the ontology taxonomy, then there is also the same semantic relationship between concept X and concept Y in the resource description.

Example 2: Assume that document A matches concept X and document B matches concept Y. If document A is linked with document B by a hyperlink, and concept X is the ancestor of concept Y in the ontology taxonomy, then the relationship between concept X and Y will retain *Part-Of* relationship in the resource description.

The mapping of concept instances: there are special cases in the information source where some web pages contain the information about data instances associated with the concepts that we are matching. We note that many real-world ontologies have been built with associated concept instances. The reason for this is that some well-known instances constitute an important part of a common vocabulary in a specific domain. For example, in Figure 1, instances '*Network Security*' and '*Intelligent Systems*' enrich the content of abstract concept '*Research Area*'. A moderate number of concept instances in the conceptual model of the resource description is necessary to obtain good matching accuracy to the query. Therefore, we create some concept instances by analyzing the body content of web pages, and assign them to the corresponding concepts. As a result, the conceptual model in the resource description, in fact, comprises concepts and their semantic relationships as well as the associated concept instances.

4. The selection of relevant information sources

In order to select appropriate information sources, it is necessary to find relevant concepts in the resource description that match the query terms in the query. Assume that a user query Q consists of a set of query terms, $Q = \{q_1, q_2, \cdots, q_t\}$. To overcome semantic heterogeneity (e.g., using different names to express the same intended meaning), the text content of a concept in the resource description includes a label name and a complemental synonymous list. In addition, a set of possible concept instances associated with this concept will be additional information for considering. Since a concept instance is only an example of concept specialization, the terms in the instance set are far less important than ones in the label name or the synonymous list during query matching. Therefore, we assign lower weights to the terms in the instance set. Then, the text content of a concept c can be described as

$$c = \{t_1 w_1, t_2 w_2, \cdots, t_u w_u\} \quad (6)$$

where term $t_j (1 \le j \le u)$ is a word occurs in the text content of the concept *c*, and w_j is the relevant weight associated with the term t_j. Note that w_j is normalized and $\sum_j w_j = 1$

So the relevance score of a concept *c* to a query *Q* can be calculated as

$$relevance\_score(Q \mid c) = \sum_{i=1}^{i=t} q_i w_i \quad (7)$$

where w_i is the weight associated with the query term q_i which occurs in the text content of the concept *c*. If the relevance score is greater than a relevance threshold τ, this concept *c* will be selected as a query concept with respect to the query *Q*.

Once the relevant query concepts in the resource description corresponding to each information source have been identified, the selection of appropriate information sources will be based on the number of query concepts in the resource description that are matched with the query. Each information source *S* contains a concept-match score which can be estimated by the following formula

$$Concept\_Match(C_Q \mid S) = \frac{\text{the number of query concepts matched}}{\text{the total number of concepts in the resource description}}$$

where C_Q be a set of matched query concepts in the resource description. The denominator is used to nullify the effect of the broadness of subject content of the information source. Considering search efficiency, this formula ensures that a specific-purpose information source which focuses on documents in confined subject domains is assigned with higher concept-match score than a large-scale general-purpose information source when these two information sources have the same number of matched query concepts in their resource descriptions.

Finally, information sources are ranked by the concept-match score, and those top-ranking ones will be chosen as relevant to the query.

5. Experiments

5.1. Experimental Setup

In this section, we present our experimental setup, which includes the construction of test data, experimental baseline and evaluation metrics.

We have evaluated our ontology-based search approach on three real-world domains – *University Department*, *Travel Agent* and *Hotel*. To collect experimental data, we used our developed spider to crawl relevant Web sites and fetch Web pages of these three topic domains. In each domain, we downloaded 40 Web databases from real Web sites. For each Web database, we downloaded the snapshot of the entire set of Web pages. The number of documents in these databases varies from 50 to 500. Among them, 15 databases out of the 40 are treated as training data to construct the domain-specific ontology using the method outlined in Section 3. The rest of the databases are used as testing data to verify the effectiveness of our proposed approach. Table 1 shows the characteristics of ontology taxonomies in these three domains.

To compare the selection performance of our proposed content-based search approach, we provide a widely-used keyword-based technique – the CORI database selection algorithm [1] as the experimental baseline. The CORI algorithm uses a variant of $tf \cdot idf$ adapted for ranking databases.

Instead of using popular IR measures – *Precision* and *Recall*, in this paper, we use a more reasonable method – the $\hat{R}_k(E,B)$ metric to evaluate the performance of resource discovery [2]. For each query, two database ranks are provided: one is a *baseline or desired rank B* in which databases are ranked by their $r(Q,S)$ value, where $r(Q,S)$ is the number of documents contained in database *S* which are relevant to the query *Q*; the other is a *estimated rank E* which is ranked by the relevance score calculated by the database selection algorithm. The $\hat{R}_k(E,B)$ metric measures the percentage of relevant documents contained in the *k* top-ranked database, which is defined as

$$\hat{R}_k(E,B) = \frac{\sum_{S_i \in E_k} r(Q,S_i)}{\sum_{S_i \in B_{k^*}} r(Q,S_i)} \quad (12)$$

where E_k is the *estimated rank* of the *k* top-ranked database, and B_{k^*} is the *baseline rank* of all the databases that are useful for the query. The primary objective of database selection is to select a small set of databases that cover as many relevant documents as possible. This means that the higher the $\hat{R}_k(E,B)$ value, the better the database selection algorithm.

5.2. Preliminary experimental results

We carried out the evaluation with 75 test Web databases in three topic domains, which we believe are enough to do statistically comparative studies. Here, we provide an analysis with respect to the performance

of our concept-based approach comparing that of the keyword-based approach.

We now turn to report the results of the experimental comparison of three database selection approaches, namely, CORI, the ontology-based approach in the Web database set. Figure 2 shows the statistical $\hat{R}_k(E,B)$ metric value for answering 80 queries. Focusing on the accuracy lines in Figure 2, we can draw the following preliminary conclusions.

As we expected, compared with the keyword-based selection approach-CORI, the concept-based approaches achieves high selection accuracy with performance improvement of 38.4% on average We noted that the biggest improvement takes places at the point of Top 10 database with the $\hat{R}_k(E,B)$ metric value being 0.873 in ONTO against 0.601 in the CORI.

One possible explanation for the improvement is that domain-specific ontologies with a vocabulary of concepts and associated relationships provides an effective tool to describe and represent the subject contents of Web databases. Since the ontology-based approach selects the databases based on the context or meaning instead of keywords, it might better capture the contents of web databases.

6. Conclusions

Accurate and meaningful resource descriptions that describe and represent the contents of web-based information sources are an important part in information source selection. However, the acquisition of appropriate resource descriptions is hindered by semantic heterogeneity in information sources. In this paper, we discussed how domain-specific ontologies could be used to overcome the problem of semantic heterogeneity. We described an ontology-based approach to generate conceptual models in resource descriptions by the use of concepts and their semantic relationships in domain-specific ontologies.

References

[1] J. Callan, "Distributed information retrieval," *Advances in Information Retrieval*, W. B. Croft, Ed.: Kluwer Academic Publishers, 2000, pp. 127-150.

[2] J. C. French and A. L. Powell, "Metrics for evaluating database selection techniques," *World Wide Web*, vol. 3, pp. 153-163, 2000.

[3] T. R. Gruber, "Towards Principles for the Design of Ontologies Used for Knowledge Sharing," in *Formal Ontology in Conceptual Analysis and Knowledge Representation*, N. G. a. R. Poli, Ed. Deventer, The Netherlands: Kluwer Academic Publishers, 1993.

[4] V. Kashyap and A. Sheth, "Semantic Heterogeneity in Global Information Systems: The Role of Metadata, Context and Ontologies," in *Cooperative Information Systems*: Academic Press, 1998, pp. 139-178.

[5] D. D. Lewis, "Naive Bayes at Forty: The Independence Assumption in Information Retrieval," in *EDML-98*. Chemnitz, Germany, 1998.

[6] C. J. v. Rijsbergen, *Information Retrieval*, Second Edition ed, 1991.

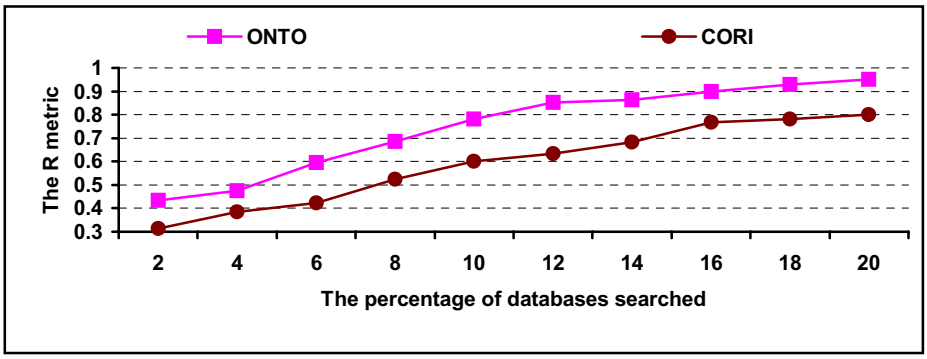

Fig. 2 The comparison of different search mechanisms on selection performance on Web databases

Taxonomies	Concepts	Non-leaf concepts	Depth	Instances in taxonomy	Axiom in taxonomy
Department	*258*	*35*	*4*	*712*	*26*
Travel Agent	*327*	*42*	*4*	*876*	*38*
Hotel	*176*	*23*	*3*	*475*	*20*

Table 1. Domains and taxonomies for our experiments