Searching biomedical ontologies based on content

Searching Biomedical Ontologies Based on Content

Harith Alani[1], Natasha Noy[2], Nigam Shah[2], Nigel Shadbolt[1], and Mark Musen[2]

School of Electronics and Computer Science, University of Southampton, Southampton, United Kingdom
h.alani, kmo, nrs@ecs.soton.ac.uk

Stanford Medical Informatics, Stanford University, Stanford, CA, USA
noy, nigam, musen@stanford.edu

Abstract. As more ontologies become publicly available, finding the "right" ontologies becomes much harder. In this paper, we introduce a new ontology search technique which is based on corpus analysis. In particular, we look at the case when users search for ontologies relevant to a particular *topic* (e.g., an ontology about anatomy). Our experiments demonstrate that using our method for query expansion improves retrieval results by a 113%, compared to the tools that search only for the user query terms and consider only class and property names.

1 Introduction

Ontologies are the key component of the Semantic Web. Today, an ever growing number of ontologies in various domains is becoming available. However, the more ontologies are available, the harder it is for users to find ontologies relevant to their domain of interest. Swoogle [1] is currently the dominant engine for searching ontologies. It searches a large index of ontologies crawled off the Web for classes and properties with labels containing the keywords entered by the user.

One would expect that typing "anatomy" in Swoogle or similar ontology search engines should get the user the Foundational Model of Anatomy (FMA)[2] as one of the top results. Yet, the FMA does not have a single class with the word "anatomy" in its name. Therefore, a keyword search on Swoogle on for "anatomy" will not actually return the FMA.

To understand better how users tend to search for ontologies, we monitored the user mailing lists of Protégé,[1]. Protégé mailing lists often receive requests from users seeking ontologies for particular domains. We observed that almost all such user requests name the domain (e.g. History, Economy, Algebra), but not the representative terms for the domain. Search engines usually return only the ontologies that have the query term itself in their class or property names, rather than searching for the ontologies that cover the *domain* described by the query term.

[1] http://protege.stanford.edu

In this paper, we present a new mechanism for ontology search that adds the element of domain knowledge to the process. We use the Web itself to expand the user query with terms that are representative of the topic. We collect these terms from the Web pages returned by a Web search with the user query. Thus, when looking for pages on "anatomy", we find that the following terms are representative of this topic: *body*, *brain*, *skin*, *bone*, *eye*, *neck*, and so on. Then we use these terms to query the ontology repository.

2 Searching Ontologies with Corpus Analysis

Once the user submits a query, we locate a number of relevant Web pages to use as a corpus that describes the query domain. To locate the relevant Web pages, we perform a Google search on Wikipedia pages using the query entered by the user. After retrieving the corpus, we calculate the frequency of all the terms that appear in the corpus (excluding stop words) using a simple TF/IDF algorithm [3], and select the top 50 terms to be used as the new user query. We search the ontologies for all the terms in the expanded query.

To determine if an ontology in our repository is relevant to the expanded set of query terms we determine how many times each query term appears in the labels of classes, labels of properties, and in property values for datatype properties (e.g., string-valued properties). We normalize the number of occurrences by the TF/IDF frequency of the term in our corpus. For each ontology, we also remove the term that appears most often in (the outlier term) from determining the score for. The latter step of removing the outliers allows us to account for cases where a common term appears hundreds of times in the ontology, but no other terms from the query do.

3 Experiment Setup

We chose the domain of biomedical ontologies for our empirical evaluation of the algorithm. For our **ontology repository** R we chose the Open Biomedical Ontologies available through the BioPortal of the National Center for Biomedical Ontologies.[2] At the time that we performed the experiments, the repository consisted of 55 ontologies representing various biomedical areas. We used the following four **queries** in our experiments: (1) *anatomy*; (2) *pathology*; (3) *physiological process*; (4) *histology*.

We asked 5 experts in the domain of biomedical ontologies to identify the ontologies from our repository that they considered relevant to the four queries above. If at least one of the experts considered an ontology that our algorithm returned to be relevant for the query, we considered this ontology a hit. The returned ontology was a miss otherwise.

We also created 3 baseline cases with which we compared our results. These cases correspond to basic searches, with no query expansion: query terms in

[2] `http://www.bioontology.org/ncbo/faces/index.xhtml`

labels (L): search only class and property labels for the terms in the (non-expanded) query; query terms in labels and property values (LV): search not only labels but also property values, such as synonyms and comments, for the terms in the non-expanded query; all ontologies ($NULL$): return all ontologies in the repository. The null case obviously provides 100% recall but variable precision, depending on the query. We used the BioPortal search engine to get the data for the L and LV cases.

4 Experiment Results

Table 1 shows the number of ontologies that at least one expert identified as relevant for each of the queries. More interesting, Table 2 shows the low level of expert agreement in the ontology relevancy.

query	number of ontologies
anatomy	21
physiological process	15
pathology	6
histology	21
total	63

Table 1. The number of ontologies identified by experts as relevant to each query (out of 55 ontologies in the repository)

number of experts agreeing	answers in agreement (number)	answers in agreement (%)
1 expert	39	62%
2 experts	5	8%
3 experts	1	2%
4 experts	3	5%
5 experts	15	24%

Table 2. Inter-expert agreement for ontologies marked as relevant.

Figure 1 and Table 3 show the average precision, recall, and f-measure values for all the cases above. We show the results for using a different number of pages to create the corpus (2, 5, 10, 20, and 50 pages). We get the best retrieval performance with a corpus of 2 pages only, which gives an average precision, recall, and f-measure of 54%, 63%, and 58% respectively.

Note that the retrieval results of looking for the original query terms only in labels (the L case) is extremely low: the average f-measure is 27% (precision is 64% and recall is 13%). When we add property values into consideration, the f-measure becomes 40%, which constitutes a 48% improvement over using only labels. If we compare the f-measure for our ontology-search method based on query-expansion with that of the search for the original query term (LV), we get a 43% improvement. Comparing to the traditional ontology search, where the search engine uses only labels and only the original query terms (e.g., Swoogle), our method provides improvement of 113%.

5 Summary and future work

We have discovered several unexpected facts related to the problem. First, inter-expert agreement in determining ontologies relevant to a user query is extremely

low: only in 24% of the cases all 5 of our experts agreed on an answer. Second, using only the query term and searching only labels of classes and properties provides extremely poor retrieval performance (f-measure of 27%). Using property values in the search in addition to labels improves the results by 48%. Using the query-expansion method that we discussed and searching in labels and property values improves the result by 43%.

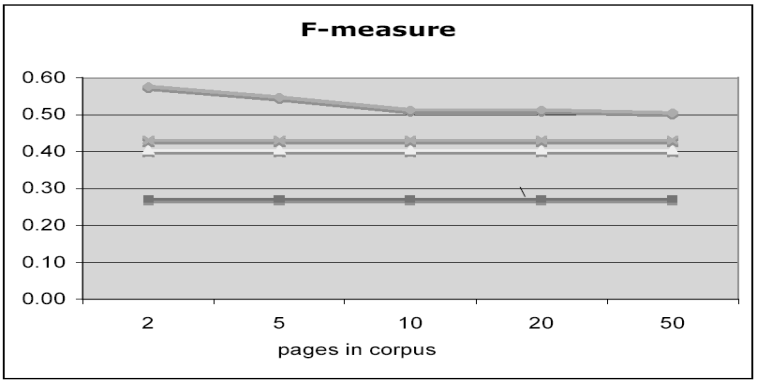

Fig. 1. The average f-measure value for different numbers of pages in the corpus.

	Precision	Recall	F-measure
Query expansion	54%	63%	58%
Null case ($NULL$)	29%	100%	43%
Labels+values (LV)	65%	27%	40%
Labels only (L)	64%	13%	27%

Table 3. Average values for precision, recall, and f-measure for the 4 queries in different cases, using a 2-document corpus for query expansion.

Acknowledgement

We are especially grateful to Mathew Jones for implementing some of the code used in this system, and to Daniel Rubin in helping to collect results from experts. Special thanks to the researchers in biomedical ontologies who provided expert results.

References

1. L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. C. Doshi, and J. Sachs. Swoogle: A semantic web search and metadata engine. In *Proc. 13th ACM Conf. on Information and Knowledge Management*, Nov. 2004.
2. C. Rosse and J. L. V. Mejino. A reference ontology for bioinformatics: The foundational model of anatomy. *Journal of Biomedical Informatics.*, 2004.
3. G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw- Hill, 1983.