Content-based Ontology Ranking

Matthew Jones
Dept. of Electronics & Computer Science
University of Southampton
msj103@ecs.soton.ac.uk

Harith Alani
Dept. of Electronics & Computer Science
University of Southampton
h.alani@ecs.soton.ac.uk

ABSTRACT

Techniques to rank ontologies are crucial to aid and encourage the re-use of publicly available ontologies. This paper presents a system that obtains a list of ontologies from a search engine that contain the terms provided by a knowledge engineer and ranks them. The ranking of these ontologies will be done according to how many of the concept labels in those ontologies match a set of terms extracted from a corpus of documents related to the domain of knowledge identified by the knowledge engineer's original search terms.

1 INTRODUCTION

Ontologies can be very time-consuming and expensive to construct. As the use of ontologies for the representation of domain knowledge increases, so will the need for an effective set of tools to aid the discovery and re-use of existing knowledge representations. This is because a major advantage of ontologies is their ability to be re-used as well as easily adapted to work with new knowledge-based applications.

Recently, a small number of search-engines to aid in the discovery of ontologies have been developed, but the techniques for ranking the results of these search engines are still in their early stages. The ontology search engines; Swoogle [3] ranks its results using an adaptation of Google's PageRank [4] scoring system. A major downside with this method is that many ontologies are poorly inter-referenced, which does not necessarily reflect in the quality of the ontologies. AKTiveRank[1] is an ontology ranking method that applies a number of graph analysis measures to estimate how well does each potential ontology represent the classes of interest. This ranking method is purely dependent on the terms given by users when searching for ontologies.

When developing ontology ranking techniques, it is important to also consider how users perform ontology searches in the first place. The ontology search engine mentioned above allows searching for specific terms, which has to exist in the ontology (e.g. part of a class or a property name) for that ontology to be retrieved. However, when analysing queries sent by ontology seekers to the Protégé mailing lists, we found that they all describe a domain (e.g. History, Economy, Algebra), rather than specific entity labels. In this paper we introduce a ranking method that is based on the content similarity of an ontology to a corpus that is selected for the given search terms.

2 CONTENT-BASED RANKING

In order to rank ontologies, our system will attempt to find a corpus that relates to the domain that the user requires an ontology to represent. This method is inspired by [2], but differs from it in that the corpus is selected based on the user query, rather than the ontology itself. The corpus will then be analysed to identify domain-related terms to use for evaluating the existing ontologies in terms of how well they cover the domain of interest. Using a representative corpus allows terms to be extracted using term frequency measures (tf-idf [5]). The terms which get the highest Tf-idf score from this corpus can then be considered as potential concept labels. This system uses the top 50 words of such an analysis. An ontology which has more class labels that match these words is deemed more suitable by the system and is therefore ranked higher than others. The following sections demonstrate our ranking method.

2.1 Obtaining a Corpus

To obtain a set of documents relevant to a user query, this system uses a Google search, and takes the first 100 pages as its corpus. Initially we thought that using the same search terms provided by the user would be enough to get a set of documents rich in domain-related information. However, many of the documents returned in such cases where too general (e.g. charity sites and general organisations' web pages when searching for 'Cancer'). As a remedy to this problem, WordNet was used to expand user search terms to make the search for pages more specific to the domain of knowledge required.

2.2 Adding WordNet

For more specific queries in Google, more terms need to be added to the Google query string, other than those given by users when searching for ontologies. These extra query terms can be obtained from WordNet. The use of WordNet has two benefits; while specifying a more specific query to Google; it also allows the system to disambiguate any terms provided by the user which may have more than one meaning (e.g. Cancer as a disease rather than a zodiac sign).

Table 1 shows the top 50 terms used in a corpus obtained from Google for the term ‘Cancer’, compared with those obtained using a query expanded with WordNet, by specifying the disease sense of the word cancer. The words added to the Google query are simply synonyms, hyponyms and meronyms of the original query terms. The addition of these extra words is simply to obtain a more specific query for Google. The improvement in the selection of potential concept labels in column (b) is quite apparent when compared with column (a).

(a) Using Basic Google Search						(b) Using WordNet Expanded Google Search					
1.	cancer	18.	loss	35.	cliphead	1.	cancer	18.	thymoma	35.	studies
2.	cell	19.	dine	36.	apologize	2.	cell	19.	malignant	36.	ovarian
3.	breast	20.	mine	37.	changed	3.	tumor	20.	gene	37.	information
4.	research	21.	dinner	38.	unavailable	4.	patient	21.	clinical	38.	research
5.	treatment	22.	cup	39.	typed	5.	document	22.	neoplasm	39.	drug
6.	tumor	23.	strikes	40.	bar	6.	carcinoma	23.	pancreatic	40.	related
7.	information	24.	heard	41.	spelled	7.	lymphoma	24.	Tissue	41.	associated
8.	color	25.	signposts	42.	correctly	8.	disease	25.	therapy	42.	neoplastic
9.	patient	26.	teddy	43.	typing	9.	access	26.	lesion	43.	oral
10.	health	27.	bobby	44.	narrow	10.	treatment	27.	blood	44.	bone
11.	support	28.	betrayal	45.	entered	11.	skin	28.	study	45.	chemotherapy
12.	news	29.	portfolio	46.	refine	12.	liver	29.	thyroid	46.	body
13.	care	30.	lincoln	47.	referenced	13.	leukemia	30.	smoking	47.	oncology
14.	wealth	31.	inn	48.	recreated	14.	risk	31.	polyp	48.	growth
15.	tomorrow	32.	endtop	49.	delete	15.	breast	32.	human	49.	medical
16.	entering	33.	menuitem	50.	bugfixes	16.	genetic	33.	health	50.	lung
17.	writing	34.	globalnav			17.	tobacco	34.	exposure		

Table 1: Comparison of tf-idf results from a corpus of 100 documents obtained from Google for the term ‘cancer’ (column a) and terms expanded using WordNet (column b).

2.3 Calculating Ontology Score

Each ontology is then ranked according to how many of these new terms match class labels within them; the class match score (CMS).

Definition: Let O be the set of ontologies to be ranked and P be the set of potential class labels obtained from the corpus. And n is the number of terms collected from the corpus.

$$CMS[o \in O] = \sum_{i=1}^{n} I\left(P_i, o\right) \times 5 \log (n + 2 - i)$$

$$I\left(P_i, o\right) = \begin{cases} 1 & : \textit{if } o \textit{ contains a class with label matching } P_i \\ 0.4 & : \textit{if } o \textit{ contains a class with label which contains } P_i \\ 0 & : \textit{if } P_i \textit{ does not appear in any of } o\textit{'s class labels} \end{cases}$$

The values 1 & 0.4 can be adjusted according to how much emphasis is put on a complete class label match compared with a partial one. The ontologies are also analysed to see if any literal text, e.g. comments, matches the potential class labels. The literal text match score (LMS) is the same as CMS, except that $I(P_i, o)$ is now 1 if the ontology (o) contains text that matched a given terms (P_i), and 0 otherwise.

The total score for each ontology is a combination of these scores, which are weighted, to emphasise the importance of one over the other; $Total = \alpha\ CMS + \beta\ LMS$. Where α & β are weights, which the experiment in section 4.1 is concerned with manipulating.

3 EXPERIMENTS

In this section two experiments are presented that show how manipulations of how the system ranks the ontologies affect the ranking order. The example used here is a search for ontologies for ‘Cancer’. The results from the experiments will then be compared and evaluated in section 5. The set of ontologies to be ranked in these experiments appear in table 2, and were chosen carefully from Google results, after throwing away duplications and broken ontologies.

3.1 Experiment 1

This experiment looks at how changing the significance of the class match and literal text match score affects ontology's ranking. This is done by changing the α & β values described in section 2.3. For this experiment two attempts to rank the ontologies are made. Experiment 1(a) uses 0.8 and 0.2 for α & β respectively (a class match being considered more important). For 1(b), both α & β take the value 0.5 (both being assumed to have the same importance for ranking purposes).

3.2 Experiment 2

This experiment looks at the effects of restricting the corpus to being comprised solely of Wikipedia pages. This experiment is repeated twice using two sets of weights for α & β as done in experiment 1.

ID	Ontology URL
1	http://semweb.mcdonaldbradley.com/OWL/Cyc/FreeToGov/60704/FreeToGovCyc.owl
2	http://www.inf.fu-berlin.de/inst/agnbi/research/swpatho/owldata/swpatho1/swpatho1.owl
3	http://www.mindswap.org/2003/CancerOntology/nciOncology.owl
4	http://sweet.jpl.nasa.gov/ontology/data_center.owl
5	http://compbio.uchsc.edu/Hunter_lab/McGoldrick/DataFed_OWL.owl
6	http://www.cs.umbc.edu/~aks1/ontosem.owl
7	http://homepages.cs.ncl.ac.uk/phillip.lord/download/knowledge/ontologyontology.owl
8	http://www.daml.org/2004/05/unspsc/unspsc.owl
9	http://envgen.nox.ac.uk/miame/MGEDOntology_env_final.owl
10	http://www.fruitfly.org/%7Ecjm/obo-download/obo-all/mesh/mesh.owl

Table 2: URLs of ontologies to be used in all experiments.

4 EVALUATION

In order to evaluate our results, it was necessary to compare the system's ranking with those produced by humans. Two medical students were asked to rank each of our selected ontologies according to how well the ontologies cover the concept of cancer. These comparisons are shown in Table 3.

The results obtained from the system are promising, but obviously non-conclusive due to the very small size of the experiment. Our system agreed with our experts by ranking the NCI ontology first. The comparison shows that while some of the ontologies are ranked similarly by the system, there are still a few of the ontologies that seem to be out of place. Notably, a number of the larger, more general ontologies (e.g. ID 1) are given lower scores by our experts. This is possibly due to the fact that they considered the ontologies 'too general'. This indicates that perhaps there is a need for checking the specificity of an ontology when evaluating its relevance to a search query.

The results of experiment 2(b) turned out to be the most similar to the ranks produced by our experts, scoring 0.693 in Pearson's Correlation Coefficient (PCC), where a value of 1 is a perfect match, 0 is total randomness, and -1 is an inverse match (table 4). This similarity drops to 0.492 when the user search terms are *not* expanded with WordNet. Note that the average PCC value between the ranks of our two experts is 0.92, indicating high agreement.

Ontology ID	Human Rank	Expt. 1(a)	Expt. 1(b)	Expt. 2(a)	Expt. 2(b)
3	1	1	1	1	1
10	2	8	5	8	5
6	3	2	2	2	2
2	4	6	8	5	6
5	5	5	6	6	7
1	6.5	3	3	3	3
9	6.5	7	7	7	8
8	8	4	4	4	4
7	9	10	10	10	10
4	10	9	9	9	9

Table 3: Comparison of rankings from system experiments along with average human rank

Rankings	Pearson Correlation Coefficient
Humans:Rank1a	0.565
Humans:Rank1b	0.650
Humans:Rank2a	0.577
Humans:Rank2b	**0.693**
Humans:No WordNet	0.492

Table 4: Comparison of ranks with those given by users

5 CONCLUSION

Ontologies can be evaluated and ranked is many different ways, based on variant characteristics. Here we experimented with evaluating ontologies based on their coverage of the domain of interest. As we are simply looking for a set of particular concepts there is no attempt to look at how well they are connected or how well each is defined in an ontology. Combining our method with some of the ranking metrics from [1] would possibly be beneficial. The set of potential concepts which are extracted from the corpus were very acceptable in our experiment. However, retrieving a suitable corpus can not be guaranteed for all ontology search queries, especially if the search terms are too specific.

The results from a Wikipedia-only corpus where better in our experiments, but it did not differ dramatically from the unrestricted one. This renders limiting the corpus to Wikipedia questionable, especially that the ontology topics that users might be after may not be covered well enough in Wikipedia.

REFERENCES

[1] Alani, H. and Brewster, C., Ontology Ranking based on the Analysis of Concept Structures. In Proceedings of the 3rd International Conference on Knowledge Capture (K-Cap), Banff, Alberta, Canada. 2005.

[2] Brewster, C., Alani, H., Dasmahapatra, S. and Wilks, Y.. Data driven ontology evaluation. In International Conference on Language Resources and Evaluation, Lisbon, Portugal, 2004.

[3] Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Reddivari, P., Doshi, V. C. and Sachs, J. Swoogle: A semantic web search and metadata engine. In Proceedings of the 13th ACM Conference on Information and Knowledge Management, Nov. 2004.

[4] Page, L., Brin, S., Motwani, R. and Winograd, R. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Uni., 1999.

[5] Salton, G. and Buckley, C. 1988 Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5): 513–523