Wikitology: Using Wikipedia as an Ontology

Zareen Saba Syed, Tim Finin, Anupam Joshi
University of Maryland, Baltimore County
1000 Hilltop Circle, Baltimore, MD 21250, USA
410-455-3971, {zarsyed1, finin, joshi}@umbc.edu

Identifying topics and concepts associated with a set of documents is a task common to many applications. It can help in the annotation and categorization of documents and be used to model a person's current interests for improving search results, business intelligence or selecting appropriate advertisements. We have investigated using Wikipedia's articles and associated pages as a topic ontology for this purpose. The benefits of the approach are that the ontology terms are developed through a social process, maintained and kept current by the Wikipedia community, represent a consensus view, and have meaning that can be understood by reading the associated pages.

Introduction

There are two popular techniques for describing what a document is about: using statistical techniques to describe the words and phrases it contains and assigning terms to the document that represent semantic concepts traditionally drawn from a standard hierarchy or ontology such as the Dewey Decimal System (Dewey 1990) or ACM Computing Classification System (Coulter *et al*. 1998). More recently, many Web 2.0 systems have allowed users to tag documents and Web resources with terms without requiring them to come from a fixed vocabulary, a process by which a community ontology can emerge.

An advantage of using the "ontology" approach, whether based on a designed or emergent ontology, is that the terms can be explicitly linked or mapped to semantic concepts in other ontologies and are thus available for reasoning in more sophisticated language understanding systems (Nirenburg *et al*. 2004) or specialized knowledge-based systems, or in Semantic Web applications. Using the traditional approach of a controlled, designed ontology has many disadvantages beginning with the difficult task of designing, implementing and also maintaining the ontology, especially in domains where the underlying concepts are evolving. As a final problem, assigning ontology terms to a document requires a person to be familiar with all of the possible choices, understand the consensus meaning of each, and select the best set of terms. The use of an implicit ontology emerging from the tags of a community solves some of these problems, but also has significant disadvantages. Some of these are inherent and others are being addressed in the research community and may ultimately admit good solutions. These problems are worth addressing because the result will be an ontology that represents a consensus view of a community and is constructed and maintained by the community without cost to any organization. It remains unclear how the terms in such an ontology should be organized structurally, understood informally by end users, or mapped to a more formal ontology such as Cyc (Lenat 1995) or popular Semantic Web ontologies like FOAF (Ding *et al*. 2005).

We are developing a system that is a blend of the two approaches based on the idea of using Wikipedia as an ontology in which each of the approximately 2.6M articles and 180K categories in the English Wikipedia represents a concept. This offers many advantages: Wikipedia is broad and fairly comprehensive, of generally high quality, constructed and maintained by tens of thousands of users, evolves and adapts rapidly as events and knowledge change, free and "open sourced", and has pages whose meaning can be easily comprehended by people. Finally, Wikipedia's pages are already linked to many existing formal ontologies though efforts like DBpedia (Auer *et al*. 2007) and Semantic MediaWiki (Krotzsch *et al*. 2006.) and in commercial systems like Freebase and Powerset.

Methodology

We use Wikipedia's articles and the category and article link graphs[1] to predict concepts common to a set of documents. Several algorithms were implemented and evaluated to aggregate and refine results, including the use of spreading activation (Crestani 1997) to select the most appropriate terms. While the Wikipedia category graph can be used to predict generalized concepts, the article links graph helps by predicting more specific concepts and concepts not in the category hierarchy. Our experiments show that it is possible to suggest new category concepts identified as a union of pages from the page link graph. Such predicted concepts can be used to define new categories or sub-categories within Wikipedia.

We use three different methods for our experiments. In the first method we use a set of related documents as search query to an information retrieval system populated with Wikipedia articles. After getting top *N* matching Wikipedia articles (based on cosine similarity) for each document in the set, we extract their Wikipedia categories and score them based on number of occurrences and the similarity score between test documents and retrieved Wikipedia articles. In the second method we also use the Wikipedia category links network for prediction of related concepts. We take the top Wikipedia categories predicted as a result of method one and use them as the initial set of activated nodes in the category links graph. After *K* pulses

[1] The version we used had about 375K category links and 90M links between articles. We removed some administrative pages and their links.

of spreading activation, the category nodes are ranked based on their activation score. In the third method we take the top Wikipedia matching documents for as initial set of activated nodes in the article links graph. To further refine the links in the article links graph we filter out all links between documents whose cosine similarity scores are below a threshold (e.g., 0.4) that represents semantic relatedness.

<table>
<tr><th colspan="3">Document Titles of Test Set</th></tr>
<tr><td colspan="3">Crop_rotation, Permaculture, Beneficial_insects, Neem, Lady_Bird, Principles_of_Organic_Agriculture, Rhizobia, Biointensive, Intercropping, Green_manure</td></tr>
<tr><th>Method 1</th><th>Method 2 (2 pulses)</th><th>Method 3 (2 pulses)</th></tr>
<tr><td>Agriculture
Sustainable_technologies
Crops
Agronomy
Permaculture</td><td>Skills
Applied_sciences
Land_management
Food_industry
Agriculture</td><td>Organic_farming
Sustainable_agriculture
Organic_gardening
Agriculture
Companion_planting</td></tr>
</table>

Table 1. Document titles of test set and results of concept prediction using different methods

Experiments and Results

We conducted several experiments to evaluate how well the Wikipedia categories represent concepts in individual documents and whether Wikipedia articles can help in predicting concepts not present as Wikipedia categories. Our methods were applied to a test set consisting of articles downloaded from the Web belonging to a particular topic. The results of one of our experiments where all the test documents were related to the topic "Organic Farming" are shown in Table 1. Methods one and two predict "Agriculture" amongst the top five categories which is a broader category of "Organic Farming" whereas, using the Wikipedia articles graph in method three predicts a more specific concept, i.e., "Organic Farming" which is not present as a category.

We applied a more formal evaluation of our system by creating a test set of 100 random Wikipedia articles, which were then removed from the IR index and associated data structures. We used our system to find related articles and categories for each of them, comparing the results to the actual Wikipedia categories and article links, which we took as the "ground truth". We then computed measures for precision, average precisions, recall and F-measure. We observed that the greater the average similarity between the test documents and the retrieved Wikipedia articles the better the prediction. Method two (with two spreading activation pulses) outperformed method one. At 0.8 average similarity threshold the F-measure was 100% for both methods, whereas for 0.5 it was 77% and 61% for methods two and one, respectively. For method three (using page links graph), the F-measure at 0.8 and 0.5 average similarity threshold was 80% and 67% respectively.

Conclusions and Future Work

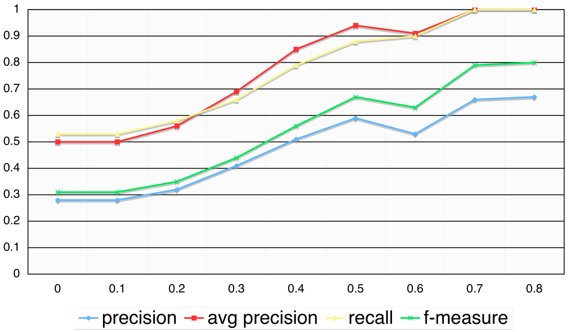

Figure 1. For a concept prediction test, values for precision, average precision, recall and f-measure increased with a threshold on similarity of pages.

We described the use of Wikipedia and spreading activation to find generalized or common concepts related to a set of documents using the Wikipedia article text and hyperlinks. Our experiments show that it is possible to predict concepts common to a set of documents using the Wikipedia categories, article text and links. We are currently investigating the application of machine learning techniques to classify links between Wikipedia articles, providing independent evidence to predict an article's semantic "type" (e.g., person, event, location) and to control the flow of spreading activation semantically. We are also applying Wikitology to model a person's current context in a collaborative environment and to improve the performance of an information retrieval system.

References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. 2007. DBpedia: A Nucleus for a Web of Open Data. Proc. 6th Int'l Semantic Web Conf., Springer, Nov. 2007.

Crestani, F. 1997. Application of Spreading Activation Techniques in Information Retrieval. Artificial Intelligence Review, 1997, vol 11; No. 6, 453-482.

Coulter, N. et al. 1998. Computing Classification System 1998: Current Status and Future Maintenance. Computing Reviews, 1998, ACM Press, New York, NY, USA.

Dewey, M. 1990. Abridged Dewey Decimal Classification and Relative Index, Forest Press.

Ding, L., Zhou, L., Finin, T., and Joshi, A. 2005. How the Semantic Web is Being Used: An Analysis of FOAF Documents. 38th Hawaii Int Conf. on System Sciences.

Krotzsch, M., Vrandecic, D. and Volkel, M. 2006. Semantic MediaWiki. 5th Int Semantic Web Conf., pp935-942, Springer, Nov. 2006.

Lenat, D. B. 1995. CYC: a large-scale investment in knowledge infrastructure. Communications of the ACM, v38, n11, pp. 33-38, 1995, ACM Press, New York, NY.

Nirenburg,S., Beale, S., and McShane, M. 2004. Evaluating the Performance of the OntoSem Semantic Analyzer. ACL Workshop on Text Meaning Representation.

Syed, Z, Finin, T and Joshi, A., Wikipedia as an Ontology for Describing Documents, Tecnical Report, UMBC, Dec. 2007.

Syed Z., Finin T., and Joshi A., Wikipedia as an Ontology for Describing Documents, Proc. 2nd Int. Conf. on Weblogs and Social Media, AAAI Press, March 2008.