11th International Protégé Conference 2009
Poster Description

Extracting an ontology of persons from Wikipedia

Dong Nguyen, Arnold Overwijk

dong.p.ng@gmail.com, arnold.overwijk@gmail.com
Utrecht University, The Netherlands

Keywords
Automatic Ontology Construction, Wikipedia, Social Network

Abstract
The focus of this research is the automatic extraction of an ontology of persons in Information Technology. Our approach involves the extraction of a categorization hierarchy of Wikipedia, the extraction of information about persons and the extraction of relations between persons. We have investigated the suitability of Wikipedia to extract social relations. Our research indicates that the infoboxes are reliable sources to extract attributes and relations, however extracting relations from the article text itself is much more difficult.

The Web is an enormous resource to gather information about people and the relations between them. Since the rise of the Web a lot of research has been carried out to mine social networks from it. Previous research has mainly focused on the use of normal web pages to extract such a network (e.g. [1], [2]).

We investigate the use of Wikipedia as a resource to extract an ontology of people. In particular, we will create an ontology of people in Information Technology. We have mainly focused on the extraction of social relations. Besides acquiring relations among people, we will also extract other information, for example the field they are working in. Such an ontology could give a quick overview of a particular field (e.g. who has a lot of relations) or might be a good start ontology which could be extended with other information.

Wikipedia has several advantages compared to other resources. Information about people and their relations is often explicitly stated in predefined templates, which are present on the pages as an ‘infobox’. Moreover, we can avoid the problem of word sense disambiguation for the names of people, since people in Wikipedia are uniquely identified by the URI of their corresponding page.

Our approach can be applied to construct an ontology of all people present in Wikipedia. It consists of the following steps: First, we extract the categorization structure of Wikipedia of people (in our case, we restrict our approach by only extracting a categorization hierarchy of persons in IT). Then we use this categorization structure to determine if a particular page in Wikipedia actually represents a person (i.e. it has to belong to the previous extracted categorization structure). Once we have acquired all pages that represent a person, we extract attributes of these persons (e.g. name and institution) and their relations with others. Relations extracted from an infobox can be labeled directly. For internal links, we first use the Stanford Parser to obtain a parse tree of the sentence in which it occurs. We use this tree to extract the relation when the particular link is part of a prepositional phrase. The final ontology is represented in OWL.

In total we extracted 151 categories. We retrieved 2843 persons, 787 out of them had an infobox on their page. The extracted ontology can be found below.

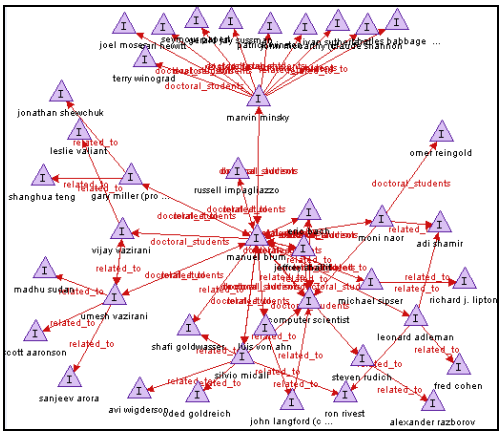

Fig. 1: Sub ontology of persons in IT

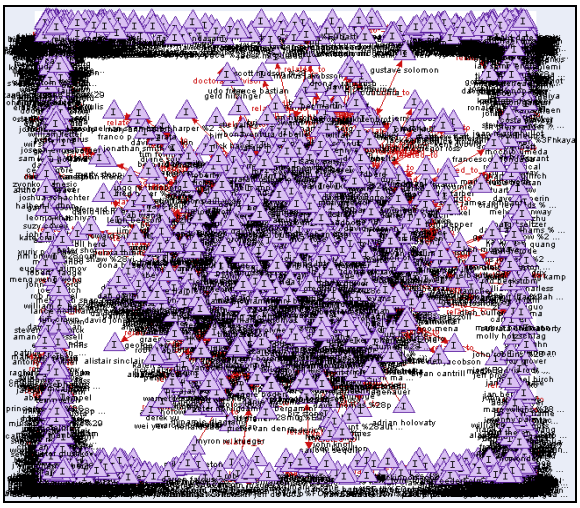

Fig. 2: Extracted ontology of people in IT

We used 300 articles to adjust our approach. We then took a set of 700 articles from which we manually evaluated 48 labeled relations. Extracting information and relations from infoboxes proved to be very reliable. However, extracting relations using the other links is not trivial. The main problem is the correct labeling of the relations. Sentences in Wikipedia are sometimes not grammatically correct and in many cases very complex. Therefore parsing such a sentence often does not result in a correct parse tree.

The relations extracted between people are looser than the conditions normally used for two persons to be related. Persons in our ontology can be related under loose circumstances, for example person X was influenced by person Y or X said something about Y. Some relations between persons were hardly relevant, and might be a too loose relation to be taken up in an ontology.

We believe that Wikipedia alone is not a sufficient resource to obtain an ontology of persons. Most persons are not represented in Wikipedia by an article, thus the created ontology gives a very limited view. However, we do believe that the created ontology might be a very suitable starting ontology to enrich with information from other resources (e.g. normal websites), and that especially information and relations extracted from the infoboxes are very reliable. These relations also indicate strong social relationships (*e.g. spouse*, *doctoral student, etc.*), while the links in the articles itself are sometimes hardly related to the person the article is about.

In our evaluation, we observed that some pages were wrongly categorized as persons. This could be filtered by applying heuristics such as only including pages which contain a birth date. Furthermore, a more interesting ontology can easily be constructed by combining the obtained categorization hierarchy with the extracted person instances.

[1] Yutaka Matsuo, Hironori Tomobe, Koiti Hasida, Mitsuru Ishizuka. Mining Social Network of Conference Participants from theWeb. In Proceedings of the International Conference on Web Intelligence, pp.190–194, 2003.

[2] Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. Social Networks, 25(3):211-230, July 2003.