Extracting Structured Knowledge for Semantic Web by Mining Wikipedia

Kotaro Nakayama
The Center for Knowledge Structuring
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
nakayama@cks.u-tokyo.ac.jp

ABSTRACT

Since Wikipedia has become a huge scale database storing wide-range of human knowledge, it is a promising corpus for knowledge extraction. A considerable number of researches on Wikipedia mining have been conducted and the fact that Wikipedia is an invaluable corpus has been confirmed. Wikipedia's impressive characteristics are not limited to the scale, but also include the dense link structure, URI for word sense disambiguation, well structured Infoboxes, and the category tree. One of the popular approaches in Wikipedia Mining is to use Wikipedia's category tree as an ontology and a number of researchers proved that Wikipedia's categories are promising resources for ontology construction by showing significant results. In this work, we try to prove the capability of Wikipedia as a corpus for knowledge extraction and how it works in the Semantic Web environment. We show two achievements; Wikipedia Thesaurus, a huge scale association thesaurus by mining the Wikipedia's link structure, and Wikipedia Ontology, a Web ontology extracted by mining Wikipedia articles.

1. WIKIPEDIA THESAURUS

WikiRelate [3] is one of the pioneers in this research area. The algorithm finds the shortest path between categories which the concepts belong to in a category tree. As a measurement method for two given concepts, it works well. However, it is impossible to extract all related terms for all concepts because we have to search all combinations of category pairs of all concept pairs (2 million × 2 million). Therefore, in our previous research, we proposed $pfibf$ (Path Frequency - Inversed Backward Link Frequency)[1], a scalable association thesaurus construction method to measure relatedness among concepts in Wikipedia. The basic strategy of $pfibf$ is quite simple. The relativity between two articles v_i and v_j is assumed to be strongly affected by the following two factors:

- the number of paths from article v_i to v_j,
- the length of each path from article v_i to v_j.

The relativity is strong if there are many paths (sharing of many intermediate articles) between two articles. In addition, the relativity is affected by the path length. In other words, if the articles are placed closely together in the graph of the Web site, the relativity is estimated to be higher than that of farther ones. Therefore, by using all paths from v_i to v_j given as $T = \{t_1, t_2, ..., t_n\}$, the relativity pf (Path Frequency) between them is defined as follows:

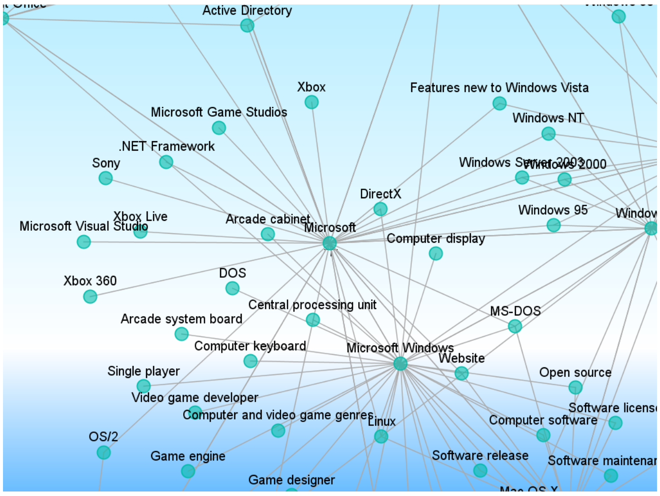

Figure 1: Wikipedia Thesaurus Visualization

$$
\begin{aligned}
pf(v_i, v_j) &= \sum_{k=1}^{n} \frac{1}{d(|t_k|)}, && (1) \\
pfibf(v_i, v_j) &= pf(v_i, v_j) \cdot \log \frac{N}{bf(v_j)}. && (2)
\end{aligned}
$$

$d()$ denotes a function which increases the value according to the length of path t_k. N denotes the total number of articles and $bf(v_j)$ denotes the number of backward links of the page v_j. Wikipedia Thesaurus [1] [2] is an association thesaurus search engine that uses $pfibf$ in its behind. It provides over 243 million relations for 3.8 million concepts in Wikipedia. We implemented a search engine for the thesaurus with capabilities of RDF export and association viusalization (Figure 1).

[1] The method name was $lfibf$ in the past and was changed to $pfibf$

[2] http://wikipedia-lab.org:8080/WikipediaThesaurusV2

2. WIKIPEDIA ONTOLOGY

In order to extract semantic relations from Wikipedia, we propose a method that analyzes both the Wikipedia article texts and link structure. Basically, the proposed method extracts semantic relations by parsing texts and analyzing the structure tree generated by a parser. However, parsing all sentences in an article is not efficient since an article contains both valuable sentences and non-valuable sentences. We assume that it is possible to improve accuracy and scalability by analyzing only important sentences on the page. Furthermore, we use synonyms to enhance co-reference resolution. In a Wikipedia article, usually a number of abbreviations, pronouns and different expressions are used to point to an entity, thus co-reference resolution is one of the technical issues in order to make the parsing process accurate.

The method consists of three main phases; parsing, link (structure) analysis, and integration. First, for a given Wikipedia article, the method extracts a list of related terms for an article using $pfibf$ [1]. At the same time, it provides synonyms by analyzing the link texts of backward links of the article. Second, the method analyzes the article text to extract explicit semantic relations among concepts by parsing the sentences. Finally, in the integration phase, three steps for triple extraction are conducted; 1) analyzing the structure tree generated by the parser, 2) filtering important semantic information using parsing strategies, and 3) resolving co-references by using synonyms. The main steps of the proposed method are described as follows.

2.1 Co-reference Resolution

In terms of Wikipedia mining, co-reference resolution is a task to determine whether the subject of a sentence is same as the main topic of the article. In a Wikipedia article, usually a number of abbreviations, pronouns and different expressions are used to point an entity, thus co-reference resolution is one of the technical issues in order to make the parsing process accurate. We employed three strategies for co-reference resolution; Article title ($C1$), Frequent pronouns ($C2$) and Synonyms ($C3$).

$C1$ is an approach to detect co-references if the terms used in s_a are all contained in the title of A_t. $C2$ uses pronouns for the judgment. It judges s_a as a co-reference to t if s_a is the most frequently used pronoun in A_t. $C1$ and $C2$ were proposed in previous research [2], but $C3$ is a novel approach proposed by us. The main idea of the approach is to detect co-references if the s_a is a synonym of t. In addition, we investigated the effectiveness of combining these three approaches in detail.

2.2 Parsing Strategies

We provide two strategies for sentence parsing in order to improve the performance; LSP and ISP.

LSP (Lead Sentence Parsing) is a strategy that parses only the lead sentences (first n sentences). After a simple inspection, we realized that a considerable number of Wikipedia articles begin with definitive sentences containing relations (hyperlinks) to other articles (concepts). Especially, the first sentence often defines "is-a" relation to other article. The statistics on lead sentence unveiled that a large number of pages in Wikipedia has a high potential for extracting "is-a" relations to other concepts thus the first sentence analysis seems a promising approach.

ISP (Important Sentence Parsing) detects important sentences in a page if the sentence contains important words/phrases for the page. Our assumption is that the sentences containing important words/phrases are likely to define valuable relations to the main subject of the page, thus we can make the co-reference resolution accurate even if the subject of the sentence is a pronoun or another expression for the main subject. We use $pfibf$ to detect important sentences. By using $pfibf$, a set of important links for each article (concept) in Wikipedia can be extracted. ISP detects important sentences in a page from sentences containing important words/phrases for the page. It crawls all sentences in the article to extract sentences containing links to the associated concepts. The extracted sentences are then parsed as the important sentences in the article. For each links in a sentence, the parser calculates $pfibf$ and the max value denotes the importance of the sentence. The importance can be used for filtering unimportant sentences by specifying thresholds.

For example, when analyzing the article about "Google," associated concepts such as "Search engine", "PageRank" and "Google search" are extracted from the association thesaurus. Therefore, ISP crawls all sentences in the article to extract sentences containing links to the associated concepts.

Acknowledgment: This research was supported in part of the Microsoft Research IJARC Core Project and Grant-in-Aid for Scientific Research (C) 20500093. We appreciate helpful comments and advices from Prof. Yutaka Matsuo at the University of Tokyo as well as from Prof. Takahiro Hara and Prof. Shojiro Nishio at Osaka University.

3. REFERENCES

[1] K. Nakayama, T. Hara, and S. Nishio. Wikipedia mining for an association web thesaurus construction. In *Proc. of IEEE International Conference on Web Information Systems Engineering (WISE 2007)*, pages 322–334, 2007.

[2] D. P. T. Nguyen, Y. Matsuo, and M. Ishizuka. Relation extraction from wikipedia using subtree mining. In *Proc. of National Conference on Artificial Intelligence (AAAI-07)*, pages 1414–1420, 2007.

[3] M. Strube and S. Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proc. of National Conference on Artificial Intelligence (AAAI-06)*, pages 1419–1424, July 2006.