A Search Engine for Browsing the Wikipedia Thesaurus

Kotaro Nakayama, Takahiro Hara and Shojiro Nishio

Dept. of Multimedia Eng., Graduate School of Information Science and Technology
Osaka University, 1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
TEL: +81-6-6879-4513 FAX: +81-6-6879-4514
{nakayama.kotaro, hara, nishio}@ist.osaka-u.ac.jp

Abstract. Wikipedia has become a huge phenomenon on the WWW. As a corpus for knowledge extraction, it has various impressive characteristics such as a huge amount of articles, live updates, a dense link structure, brief link texts and URL identification for concepts. In our previous work, we proposed link structure mining algorithms to extract a huge scale and accurate association thesaurus from Wikipedia. The association thesaurus covers almost 1.3 million concepts and the significant accuracy is proved in detailed experiments. To prove its practicality, we implemented three features on the association thesaurus; a search engine for browsing Wikipedia Thesaurus, an XML Web service for the thesaurus and a Semantic Web support feature. We show these features in this demonstration.

1 Introduction

A thesaurus is a kind of dictionary that defines semantic relatedness among words. Although the effectiveness is widely proved in various research areas, automated thesaurus dictionary construction (esp. machine-understandable) is one of the difficult issues. Since it is difficult to maintain huge scale thesauri, they do not support new concepts in most cases. Therefore, a large number of studies have been made on automated thesaurus construction based on NLP. However, issues due to the complexity of natural language, for instance the ambiguous/synonym term problems still remain on NLP.

We noticed that Wikipedia, a collaborative wiki-based encyclopedia, is a promising corpus for thesaurus construction. According to statistics of Nature, Wikipedia is about as accurate in covering scientific topics as the Encyclopedia Britannica. It covers concepts of various fields such as Arts, Geography, History, Science, Sports and Games. It contains more than 2 million articles (Dec. 2007, English only) and it is becoming larger day by day. Because of the huge scale concept network with a wide-range topic coverage, it is natural that Wikipedia can be used as a knowledge extraction corpus. In fact, we already proved that it can be used for accurate association thesaurus construction[1, 2]. In this demonstration, we describe the overview of our thesaurus construction method and show three features which we developed for the thesaurus; a search engine for browsing Wikipedia Thesaurus, an XML Web service for the thesaurus and a Semantic Web (RDF) support feature.

2 pfibf

$pfibf$ (Path Frequency - Inversed Backward link Frequency), an association thesaurus construction method we proposed, is a link structure mining method which is optimized for Wikipedia mining. The relativity between any pair of articles (v_i, 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 concept graph and sharing hyperlinks to articles, the relativity is estimated to be higher than further ones.

The number of backward links on articles is also estimated as a factor of relativity because general/popular articles have a lot of backward links and these articles easily have high relativity to many articles. Therefore, we must consider the inversed backward link frequency ibf in addition to the two factors above. The relativity becomes stronger if there are many paths (sharing many intermediate articles) between them. In addition, the relativity becomes stronger according to the path length.

Therefore, if all paths from v_i to v_j are given as $T = \{t_1, t_2, ..., t_n\}$, the relativity pf (path frequency) between them can be expressed as follows:

$$pfibf(v_i, v_j) = pf(v_i, v_j) \cdot ibf(v_j), \tag{1}$$

$$pf(v_i, v_j) = \sum_{k=1}^{n} \frac{1}{d(|t_k|)}, \tag{2}$$

$$ibf(v_j) = \log \frac{N}{bf(v_j)}. \tag{3}$$

$d()$ denotes a function which increases the value according to the length of path t_k. A monotonically increasing function such as a logarithm function can be used for $d()$. N denotes the total number of articles and $bf(v_j)$ denotes the number of backward links of v_j. This means a page which shares hyperlinks with a specific page but not with other pages, has a high $pfibf$.

However, counting all paths between all pairs of articles in a huge graph is a computational resource consuming work. Therefore, we proposed an efficient data structure named "Dual binary tree" (DBT) and a multiplication algorithm for the DBT[2].

3 Architecture and Applications

We constructed a huge scale association thesaurus named Wikipedia Thesaurus by using $pfibf$ described above. We used the Wikipedia dump data created in Sept. 2006. It contains more than 1.3 million concepts and 243 million association relations among the concepts. The thesaurus is stored in a database (MySQL 5.0) and indexed to provide practical APIs for other applications such as information
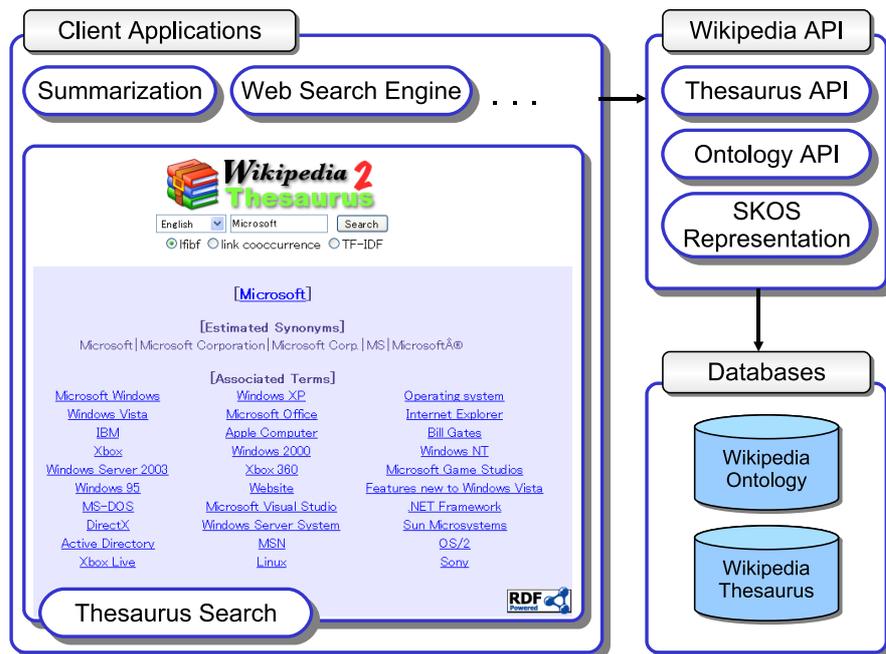
Fig. 1. Search engine for Wikipedia Thesaurus and architecture

retrieval and text summarization. The APIs are provided by XML Web services to achieve high interoperability among different environments. They allow developers to add association term extraction capability to their own applications.

We developed a thesaurus browsing system by using the APIs to prove the capability of this approach and the accuracy of our association thesaurus. We are also working on a huge scale Web ontology construction from Wikipedia. The ontology will be provided as an XML Web service. Figure 1 shows the architecture of our system and a screen shot of the thesaurus browsing system.

The concept search engine provides SKOS[3] (an application of RDF) representation function. SKOS is a developing specification and standard to support knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the Semantic Web. SKOS supports several basic relations between two concepts such as "broader," "narrower" and just "related." Further, since $pfibf$ extracts relatedness between two concepts, we extended SKOS in order to express the relatedness. We defined an extension for SKOS as the "The Wikipedia Thesaurus Vocabulary (WTV)." Currently, WTV supports some simple relations not in SKOS such as relatedness. The URLs for concepts in the Wikipedia Thesaurus correspond to the URLs of Wikipedia articles. For example, "http://wikipedia-lab.org/concept/Music" in the thesaurus corresponds to the article "http://en.wikipedia.org/wiki/Music" in Wikipedia. Figure 2 show an example of a SKOS representation of "Music."

Our thesaurus browsing system and XML Web services are available under the following URLs.

Fig. 2. Sample of SKOS (RDF) representation

- Wikipedia Thesaurus:
  http://wikipedia-lab.org:8080/WikipediaThesaurusV2
- Wikipedia API:
  http://wikipedia-lab.org/en/index.php/Wikipedia_API

We are going to extract much more complicated relations from Wikipedia by using NLP techniques. Relatedness is just a first step, but it will support much more relation types in the future.

References

1. Nakayama, K., Hara, T., Nishio, S.: A thesaurus construction method from large scale web dictionaries. In: Proc. of IEEE International Conference on Advanced Information Networking and Applications (AINA 2007). (2007) 932–939
2. Nakayama, K., Hara, T., Nishio, S.: Wikipedia mining for an association web thesaurus construction. In: Proc. of International Conference on Web Information Systems Engineering (WISE 2007). (2007)
3. World Wide Web Consortium: Simple knowledge organisation systems (skos). http://www.w3.org/2004/02/skos/ (2004)