

Wikipedia Link Structure and Text Mining for Semantic Relation Extraction Towards a Huge Scale Global Web Ontology

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, a collaborative Wiki-based encyclopedia, has become a huge phenomenon among Internet users. It covers huge number of concepts of various fields such as Arts, Geography, History, Science, Sports and Games. Since it is becoming a database storing all human knowledge, Wikipedia mining is a promising approach that bridges the Semantic Web and the Social Web (a. k. a. Web 2.0). In fact, in the previous researches on Wikipedia mining, it is strongly proved that Wikipedia has a remarkable capability as a corpus for knowledge extraction, especially for relatedness measurement among concepts. However, semantic relatedness is just a numerical strength of a relation but does not have an explicit relation type. To extract inferable semantic relations with explicit relation types, we need to analyze not only the link structure but also texts in Wikipedia. In this paper, we propose a consistent approach of semantic relation extraction from Wikipedia. The method consists of three sub-processes highly optimized for Wikipedia mining; 1) fast pre-processing, 2) POS (Part Of Speech) tag tree analysis, and 3) mainstay extraction. Furthermore, our detailed evaluation proved that link structure mining improves both the accuracy and the scalability of semantic relations extraction.

1 Introduction

Wikipedia, a collaborative Wiki-based encyclopedia, has become a huge phenomenon among Internet users. According to statistics of Nature, Wikipedia is about as accurate in covering scientific topics as the Encyclopedia Britannica[1]. It covers concepts of various fields such as Arts, Geography, History, Science, Sports, Games. It contains more than 2 million articles (Oct. 2007, English Wikipedia) and it is becoming larger day by day while the largest paper-based encyclopedia Britannica contains only 65,000 articles.

As a corpus for knowledge extraction, Wikipedia's impressive characteristics are not limited to the scale, but also include the dense link structure, sense disambiguation based on URL, brief link texts and well structured sentences. The fact that these characteristics are valuable to extract accurate knowledge from Wikipedia is strongly confirmed by a number of previous researches on

Wikipedia Mining[2–5]. These researches are mainly about semantic relatedness measurements among concepts. Besides, we proposed a scalable link structure mining method to extract a huge scale association thesaurus in a previous research [4]. In that research, we developed a huge scale association thesaurus dictionary extracting a list of related terms from any given term. Further, in a number of detailed experiments, we proved that the accuracy of our association thesaurus achieved notable results. However, association thesaurus construction is just the beginning of the next ambitious research *Wikipedia Ontology*, a huge scale Web ontology automatically constructed from Wikipedia.

Semantic Wikipedia [6] is an impressive solution for developing a huge scale ontology on Wikipedia. Semantic Wikipedia is an extension of Wikipedia which allows editors to add semantic relations manually. Another interesting approach is to use Wikipedia’s category tree as an ontology [7–9]. Wikipedia’s categories are promising resources for ontology construction, but categories can not be used as an ontology since the structure of Wikipedia category is just a taxonomy and do not provide explicit relation types among concepts.

In contrast to these approaches, we propose a full-automated consistent approach for semantic relation extraction by mining Wikipedia article texts. Since a Wikipedia article is a set of definitive sentences, the article text is yet another valuable resource for ontology construction. The method consists of three sub-processes highly optimized for Wikipedia mining; 1) fast preprocessing, 2) POS (Part Of Speech) tag tree analysis, and 3) mainstay extraction. Furthermore, we show the potential of important sentence analysis for improving both accuracy and scalability of semantic relations extraction.

The rest of this paper is organized as follows. In section 2, we explain a number of researches on Wikipedia Mining for knowledge extraction in order to make our stance clear. In section 3, we describe our proposed integration method based on NLP and link structure mining. We describe the results of our experiments in section 4. Finally, we draw a conclusion in section 5.

2 Related Works

2.1 Wikipedia Mining

As we mentioned before, Wikipedia is an invaluable Web corpus for knowledge extraction. Researches on semantic relatedness measurement are already well conducted[2–5]. WikiRelate [5] 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 graph. 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 ($2 \text{ million} \times 2 \text{ million}$). Furthermore, using the inversed path length as semantic relatedness is a rough method because categories do not represent semantic relations in many cases. For instance, the concept “Rook (chess)” is placed in the category “Persian loanwords” together with “Pagoda,” but the relation is not semantic, it is just a navigational relation. Therefore, in our previous research, we

proposed *pfibf* (Path Frequency - Inversed Backward Link Frequency), a scalable association thesaurus construction method to measure relatedness among concepts in Wikipedia.

2.2 Wikipedia and Web Ontology

Semantic Wikipedia[6] is an impressive predecessor of this research area. It allows editors to put additional tags to define explicit relations between concepts. For example, assume that there is a sentence written in Wiki format like this;

```
'London' is the capital city of [[England]]
```

“[[...]]” is a hyperlink tag to another article (concept) and will be translated into a hyperlink when it is shown to readers, so the readers can understand that “London” is the capital of “England.” However, obviously, machines can not understand the relation type if no NLP techniques are used because the relation is written in natural language. To solve this problem, Semantic Wikipedia allows users to add special annotations like this;

```
'London' is the capital city of [[capitalof::England]]
```

Semantic Wikipedia is a promising approach for a huge scale Web ontology construction but we wish an automated approach without any additional human-effort since a Wikipedia article already includes rich semantic relations.

3 Proposed method

To achieve full-automated Web ontology construction from Wikipedia, we propose a consistent approach for semantic relation extraction by mining Wikipedia article text. Basically, the proposed method extracts semantic relations by parsing texts and analyzing the structure tree generated by a POS parser. However, parsing all sentences in an article is not efficient since an article contains both valuable sentences and non-valuable sentences by mixture. Our assumption is that it is possible to improve accuracy and scalability by analyzing only important sentences for the topic.

In this section, we describe our proposed method for semantic relation extraction from Wikipedia. The whole flow of the proposed method is performed in the following three phases;

1. Preprocessing
(Trimming, chunking and partial tagging)
2. Parsing and POS structure tree analysis
3. Mainstay extraction.

These phases are described in detail in the following subsections.

3.1 Preprocessing

Before we parse sentences, we need to trim, chunk and segment the sentences in order to make them processable for the parser. For this aim, we usually use statical NLP tools, however these tools cannot process the Wikipedia articles correctly since the articles are written in a special syntax composed of HTML tags and special Wiki command tags such as triple quotations, brackets for hyperlinks and tables. That is why we developed our own Preprocessor for this aim. Preprocessing is accomplished in three sub steps; 1) Trimming, 2) Chunking and 3) Partial tagging.

First, the preprocessor trims a Wikipedia article to remove unnecessary information such as HTML tags and special Wiki commands. We also remove table tags because table contents are usually not sentences. However, we do not remove link tags (“[[...]]”) because links in Wikipedia are explicit relations to other pages and we use this link information in the following steps.

Second, the preprocessor separates the article into sentences. Basically, an article is separated into sentences by periods (“.”). However, abbreviations etc. also use “.”, so the preprocessor does not separate a sentence if the following character is a small letter. This simple strategy works very well in almost all cases (Over 99%) for Wikipedia articles. Furthermore, since it is based on neither semantic nor statistic methods, the process is much faster than ordinary chunkers. After separating an article into sentences, each sentence is separated into semantic chunks (phrases). Basically, terms are separated by white space (“ ”), but terms are bounded if these terms are placed in quotations or link tags.

Finally, phrases in quotations and link tags are tagged as nouns to help the following parsing phase. Bounding and partial tagging are helpful information for the parsing process because one of the most difficult technical issues in parsing natural language is chunking and bounding. Especially for domain specific terms or new terms, parsers often cannot parse the sentence structure correctly.

3.2 Parsing and Structure Tree Analysis

After the preprocessing, partially tagged and chunked sentences are given. In this phase, we parse each sentence to get a structure tree and analyze that structure tree to extract relations between concepts. To parse sentences, we adopted an unlexicalized PCFG (Probabilistic Context-Free Grammars) parsing method based on the factored product model. We used the Stanford NLP parser[10] for this purpose. It can parse a sentence accurately if the sentence is trimmed, chunked and tagged correctly, even if the sentence contains hyperlink tags (“[[...]]”). Figure 1 shows the detailed illustration of this phase.

“/NN” is a special POS tag for nouns, which is added in the partial tagging process. A list of main POS (Part Of Speech) tags used in this research is shown in Table 1.

The parser gets a partially tagged sentence and constructs a structure tree for the given sentence. For instance, assume that there is a semi-tagged sentence like this: “[[Madrid]]/NN is the [[capital]]/NN and largest city of [[Spain]]/NN .

