Acquisition of Term Explanations for Automatic Construction of a Portal Site

Shunsuke Sugai (210046)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 13, 2003

Keywords: portal site, acquitision of term explanation, web surfing support, information extraction.

A portal site is a website which was made for the purpose of visiting first when investigating the matter relevant to the theme with a user. The site collects various information about the theme. The main structure in a portal site is a search engine, a link collection, a glossary and so on. Now, many of such sites are made by human, however, automatic creation of a portal site isexpected.

This research aims of acquiring a terminological explanatory note automatically for the purpose of creation of a glossary in a portal site. Explanation of a term is dynamically acquired from a Web document. Because web documents contain the technical terms and coined words and new words. And web documents are renewed frequently. So this feature is suitable for our purpose. When someone wants to know the meaning of a term or make a glossary, he or she may obtain term explanation by using search engines. But it is difficult to find the page containing explanation of terms. On the other hands, this research will acquire the term explanation automatically which a user wants.

There is the case that words have several senses. The term explanation for the theme of the portal site used as a target is sorted out automatically, and ones with highest score are provided to users. Concretely speaking, the system finds the specific nouns related with the portal site theme, and

removes unrelated term explanation. Moreover, theme always changes. So, we should obtain the theme field corpus synamically. And the characteristics of this paper is in the point of asking for the degree of relation of term explanation.

This system acquires term explanation after recievingthe theme of a portal site and a technical term. It consists of the following three processes.

1. This system acquires the candidate pages considered that there is explanation of the term.

   The term X carried to a glossary is given. The system uses two query, "X-towa", and "X-wa". And it inputs this query into search engine "goo", and acquires Web pages. A character code fo them is converted to EUC.

2. Removal of the unnecessary tags from the candidate pages and formatting the pages.

   We prepare the list of a character ornament tag, style tags, etc. and according to this list, remove unnecessary tags from the candidate pages. Next, a new-line is inserted before and after a tag when tags continue, blank is inserted between them and a terminal is added to the last of the page.

3. Extraction of the term explanation using HTML tags.

   Using HTML tags, extract the term explanation from the candidate pages. There are two patterns for extraction of the term explanation. (1) The term is located in title tag, and the next sentence is the term explanation. (2) Term or query is in the sentence. We consider both cases. And by carrying out a morphological analysis, when a subsequent sentence starts in a demonstrative or a conjunction, the sentence is also extracted as term explanation. Term explanation is classified into some of phrase templates at this time.

4. Deletion of the unnecessary.

The unsuitable term explanation are removed when it matches the specific phrase template, it is too long, and it is a head of a compound nown.

5. Scoring of term explanation.

   The degree of relation of a theme and term explanation is judged by TF-IDF and RDF. For calculating a score of term explanation, the following formula is used.

$$Score(E) = \frac{1}{|E|} \sum_{n \in E} score(n) \tag{1}$$

   In order to obtain score(n), the system searches the Web pages relevant to a theme. We propose the two kinds of definitions of score(n), one is "RDF" based on document frequency, and the other is "TF-IDF" based on the term frequency and document frequency of a word. Anyway, a high score is given to a noun related with a theme.

Experiments are conducted in order to evaluate this system. In these experiments, we try two patterns of term explanation. This is the words with a meaning, and the words with many different meanings. Moreover, both score of TF-IDF and RDF were used to rank term explanations. We evaluate ten higher candidates, and judge whether it is suitable or not. Experimental results show at least, one appropriate explanation is found in the top 10 candidates.RDF is superior to TF-IDF as scoring measure. However, term explanation acquisition of polysemous words was difficult. This was because much explanation corresponding to many meanings were extracted, and the technique of measuring the degree of relation did not work well. In order to overcome this proglam, a method to use the phrase template obtained in procedure 3 for scoring should be considered.