Automatic related term extraction from Web

Masato Hoshi (210083)

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

February 13, 2004

Keywords: Web, Search Engine, Term Extraction, Glossary, Portal Site.

By using a portal site, the Internet can be used effectively. However, the portal site which I need does not necessarily exist. Therefore, it is convenient if a portal site can be made automatically. Portal site has contents, such as a glossary, a link collection, and other etc. I aim at extracting a realted term automatically for the glossary.

Term extractoon consists of the following three steps.

1. Collecting rarated documents
2. Selecting potentioal terms
3. Calculating the degree of relation

First step, receive the theme of the portal site and collect the documents relavant to the theme from the Web. In more detail, collect 500 web pages using search engine goo.

Second step, extract noun phrase and score them using statistical infomation. At this step, I experiment in three way of techniques.

Third step, calculate the degree of ralation potential term and theme. If potential term appears together with the theme, the degree of relation is high.

In this research, three kinds of experiments by the difference in second step were conducted.Three kinds of the approach is as follows.

1. Approach based on frequency of appearance
2. Approach based on statistics of compound nouns
3. Approach by the combination of a two method

First approach is count the frequency of term which consists of n-gram, and divides it by avelage of same n-gram. It means how does the term appear frequently in documents.

Second approach is giving a score to the simple noun which constitutes a noun phrase, and has calculated evaluation of the whole noun phrase. This approach is based on the idea of being as important as the single noun contained in many compound nouns.

Third approach is the combination of two approaches. This approach combines the output of two approaches simply.

By these three kinds of approaches, experimented in 20 kinds of themes. From the experiment result, the difference in the output by the difference in approach was considered. When the number of correct answers was compared, first approach is most excellent and second approach was most inferior. However, each approach has the different feature. First approach tends to output person name. Second approach tends to output the term containing the theme of a portal site. But, for the term containing the theme of a portal site, an appropriate result is not necessarily obtained on account of processing of the degree calculation stage of relation. Moreover, it is influenced of the problem of a term that third approach also contains the theme of a portal site.

From these results,if the problem of the term containing the theme of a portal site is solvable , it will be predicted that this system could take out various terms.