Automated Alignment of Multiple Internet Directories

ICHISE Ryutaro
National Institute
of Informatics
2-1-2, Hitotsubashi,
Chiyoda-ku, Tokyo, Japan
ichise@nii.ac.jp

TAKEDA Hideaki
National Institute
of Informatics
2-1-2, Hitotsubashi,
Chiyoda-ku, Tokyo, Japan
takeda@nii.ac.jp

HONIDEN Shinichi
National Institute
of Informatics
2-1-2, Hitotsubashi,
Chiyoda-ku, Tokyo, Japan
honiden@nii.ac.jp

ABSTRACT

Directory services are tools for making useful information more accessible, but individual internet directories in directory services are of limited use in finding user-relevant web pages. In this paper, we propose a method for aligning URL information from one internet directory to another. This method can discover an appropriate position in a directory for a web page which is not shown in that directory but shown in another directory. By using this method, one can extend a favorite directory with information on other directory services. Discovery of alignment is based on categorization similarity in the concept hierarchies of the two internet directories. We adopted the "κ statistic" method to measure the similarity syntactically. We conducted an experiment using real-world internet directories. The results of this experiment show that the proposed method is promising, i.e., it can classify unknown web pages into appropriate categories within an internet directory.

Keywords

Machine learning, internet directory, Web, categorization

1. INTRODUCTION

With the rapid growth of the WWW, it is getting difficult for us to find information that we want. To find useful information, people often use search engines which show web pages including the specified keywords. But if they do not have knowledge well on thir searching domain, search engines are useless because they can not choose appropriate keywords. One answer to this kind of problem is to use directory service which have the pages evaluated and organized by humans before they are registered within the search engine archive. However, a single directory service is not enough for use since the search engine tends to have some speciality both in collecting pages and in categorizing them. To solve those problems, we propose a method in which multiple internet directories are used. There are many public internet directories which have been formulated by humans. Some of them are provided to cover a wide domain and some are for special domain. Such internet directories are hard to align, because of their variety of conceptual hierarchy and their wide distribution around the world. We propose a method for solving this problem. In this paper, we describe a machine-learning method for aligning multiple internet directories.

2. INTERNET DIRECTORY MODEL

Most internet directories are managed via a system of hierarchical categorization. Each internet directory contains only one conceptual hierarchy for the organization of its categories. Each category contains links to web pages (URLs). The diagram on the left side of Figure 1 represents a single internet directory.

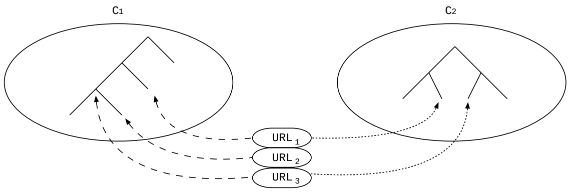

Figure 1: Internet Directory Model

In Figure 1, there are two different internet directories (C_1 and C_2) and three different URLs (URL_1, URL_2 and URL_3). Some of URLs are shared between two directories and some are not. It is important to keep in mind that these internet directories do not have the same conceptual hierarchy. The next step is to consider an appropriate way to align a URL from C_1 into C_2. In the example shown in Figure 1, C_2 does not contain URL_2. If URL_2 can be placed in the concept hierarchy of C_2, the user can then use C_2 to find useful information (at URL_2), because URL_2 has already been evaluated and categorized. In the next section, we propose a method of defining a rule for alignment from the concept hierarchy of C_1 to that of C_2, so that a web page which has already been categorized in C_1 can subsequently be categorized in C_2. The most important point of this approach is that the concept hierarchy of C_2 does not need to be adjusted to fit the concept hierarchy of C_1. Thus, a user can continue to use whichever internet directory they are accustomed to.

3. FINDING SIMILAR CATEGORY

In our method, alignment is discovered as alignment rules between two directories. We firstly find categories which are similar to each other ("similar categories"); the web page will then map one to another based on similarity between categories. To find similar categories, our algorithm starts by comparing the most general categories of the two internet directories. For each pair of categories, we can determine similarity based on URLs categorized in both categories. For each category, we can decide whether a particular URL belongs to that category. If the categorization methods of two categories are similar, then the system generates an alignment rule for them. It should be noted that, because internet directories are structured as trees, we can easily categorize URLs according to a nodal structure, such that URLs in lower (more specific) categories are included in higher (more general) categories.

To find similar categories, we used a statistical method for determining the degree of similarity between two categorization criteria. The "κ statistic" method [2] is an established method of evaluating similarity between two criteria. The relationship between two categorization criteria is examined from “top” to “bottom”. First, the most general categories in the two internet directories are compared using the “κ statistic”. If the comparison confirms that the two categories are similar, then the algorithm outputs an alignment rule for them. At the same time, the algorithm pairs one of these two similar categories with a “child” category of the other similar category. This new pair is then evaluated recursively using the “κ statistic” method. When a similar pair is not generated, the algorithm outputs the alignment rule between the two internet directories. We can then apply this rule to deciding whether a particular URL in C_1 fits the concept hierarchy in C_2.

4. EXPERIMENTAL EVALUATION

In order to evaluate our algorithm, we conducted experiments using the Yahoo! Japan [6] and LYCOS Japan [3] directories as internet directories. The Yahoo! directory contains approximately 41,000 categories and 224,000 URLs. In contrast, LYCOS contains approximately 5,700 categories and 48,000 URLs. Approximately 25,000 URLs can be found both in Yahoo! and in LYCOS. Generally speaking, Yahoo! contains more knowledge than LYCOS as an internet directory, but half of the URLs contained in LYCOS are not contained in Yahoo!. This fact implies that, to ensure access to useful Web pages, it is not sufficient that a single directory contains enormous number of URLs.

In this experiment, we used the categories Yahoo!:Arts /Humanities/Literature (and sub-categories) and LYCOS: Arts/Literature (and sub-categories) as our two internet directories. We conducted 10-fold cross validation for the shared URLs; i.e., the shared URLs were divided into ten data sets, and nine of these sets were used for training while the remaining set was used for testing. Ten experiments were conducted for each data set. The parameter of significance level for “κ statistic” was set at 5%.

The results of the experiments are shown in Figure 2. The vertical axis denotes average accuracy of the test data. “Use Exact Rules” denotes values of accuracy for a system which only uses the alignment rules for the category to which the URL belongs. “Use Parent Rules” denotes that, if the system does not generate an alignment rule for a category to which the URL belongs, it will use a rule generated for the parent category instead. “Criterion 1” means that a URL is categorized in the same category as the test data. “Criterion 2” means that a URL is categorized in the same category or its parent category as the test data. “Criterion 1” is very strict criterion because the target directory should have enough intermediate categories in comparison with the source directory, while “Criterion 2” is more general and more realistic because it does not matter which directories are rich in categorization. “Yahoo to Lycos” means that URLs are aligned from the concept hierarchy of Yahoo! to the concept hierarchy of LYCOS, and “Lycos to Yahoo” denotes vice versa.

5. DISCUSSION

More than 80% of the URLs we used in our experiments were categorized appropriately by our system. The data in Figure 2 imply that Yahoo-to-Lycos aligning was more accurate than the inverse operation. As implied by the total number of categories, the categorical hierarchy in Yahoo! is more complex than that of LYCOS. Thus, for Yahoo-to-Lycos aligning, the learned rules are likely to involve transfer of URLs from relatively complex categories to relatively general categories. Such a trend is likely to result in relatively accurate rules. For example, suppose concept hierarchy A contains category S under category X, and concept hierarchy B also has category X but does not have sub-categories under X. In such a case, it would be much easier to learn a rule for ”A:/X/S -> B:/X” than to learn a rule for ”B:/X -> A:/X/S”, because ”B:/X” contains URLs that belong in S and URLs that do not. The data shown in Figure 2 reflect this. In the above situation, nevertheless, our method works properly. When regarding parent categories as right answers (“Criterion 2’), both directions are almost the same results, i.e., ”B:/X -> A:/X” is learned instead of ”B:/X -> A:/X/S”.

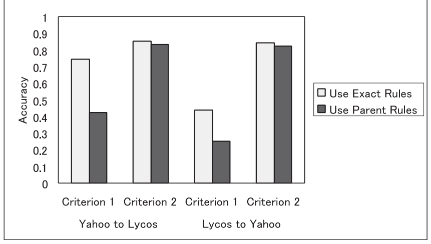

Figure 2: Result of Experiment

The bookmark-sharing systems of Siteseer [5] and Blink [1] are similar to our system. The main difference between our system and theirs is the use of hierarchy in categorization. Their systems only consider the number of URLs in a given category, but our method uses hierarchy structures. One of the merits of this approach is that, if there is no exact category into which a given URL fits, then the URL is aligned into the parent category. Bookmark-Agent [4] uses another approach to sharing bookmarks, based on keywords. Unlike a bookmark agent, our system only uses link information, not the contents of the page. Our system would therefore categorize ”Sherlock Holmes” and ”Conan Doyle” under the same concept, although they contain different words.

6. CONCLUSION

In this paper, we propose aligning internet directories as a new approach to integrate multiple directories and its method based on statistical method. To test our ideas, we conducted experiments using Yahoo! and LYCOS. Our experimental results show that the alignment rules learned by our system are reliable so that URLs in a directory can be located to the appropriate positions in the other directory. The advantage of using our method is that it can extend directories automatically with information in other directories. For example, a user can expand her/his favorite directory as she/he likes by introducing information in other directories.

7. REFERENCES

[1] Blink, 2000. http://www.blink.com/.

[2] J. L. Fleiss, *Statistical Methods for Rates and Proportions*, John Wiley & Sons, 1973.

[3] LYCOS JAPAN, 2000. http://www.lycos.co.jp/.

[4] M. Mori and S. Yamada, Bookmark-Agent: Information Sharing of URLs, Poster Proc. of the 8th Int. WWW Conf, 1999.

[5] J. Rucker and M. J. Polanco, Siteseer: Personalized Navigation for the Web, Comm. of the ACM, pp. 73-75, vol. 40, No. 3, 1997.

[6] Yahoo! JAPAN, 2000. http://www.yahoo.co.jp/.