A Hybrid Algorithm for Alignment of Concept Hierarchies

Ryutaro Ichise[1,2], Masahiro Hamasaki[2], and
Hideaki Takeda[1,2]

[1] National Institute of Informatics,
Tokyo 101-8430, Japan
[2] The Graduate University for Advanced Studies,
Tokyo 101-8430, Japan
{ichise@,hamasaki@grad.,takeda@}nii.ac.jp

Abstract. Hierarchical categorization is a powerful and convenient method so that it is commonly used in various areas, such as ontologies. Although each hierarchy is useful, there are problems to manage multiple hierarchies. In this paper, we propose an alignment method between concept hierarchies by using the similarity of the categorization and the contents of the instance. By using this method, instances that exist in one hierarchy system but does not in the other can be located in a suitable position in the other. The experimental results show improved performance compared with the previous approaches.

Hierarchical categorization is a powerful and convenient method so that it is commonly used in various areas. Although each hierarchy is useful, there are problems to manage multiple hierarchies. Similarity-based integration (SBI) [2] is an effective method for solving this problem. By using this method, instances that exist in one hierarchy system but does not in the other can be located in a suitable position in the other. The main idea of SBI is to utilize only the similarity of categorizations across concept hierarchies. Namely, SBI does not analyze the contents of information assigned to the concept hierarchies. In this paper, we propose an extension of SBI which uses the contents in information instances.

In order to state our problem, we describe a model of the nature of concept hierarchies. Many information management systems for use with conceptual information like ontologies are managed via a system of hierarchical categorization. Such information management system is comprised of 2 elements, i.e, categories and information instances. The problem addressed in this paper is finding an appropriate category in the target concept hierarchy for each information instance in the source concept hierarchy. The important point of this approach is that the source concept hierarchy does not need to be adjusted to fit the target concept hierarchy. Thus, a user can apply our method while continuing to use whichever concept hierarchy they are accustomed to.

E. Motta et al. (Eds.): EKAW 2004, LNAI 3257, pp. 474–476, 2004.

A problem of SBI is that it is hard to learn an alignment rule when the destination category is in a lower category in the target concept hierarchy. In other words, the learned rules are likely to assign relatively general categories in the target concept hierarchy. In order to avoid this type of rules, we propose to combine a contents-based classification method after we apply the SBI algorithm. Since Naive Bayes (NB) [4] is very popular and easy to use, we adopt NB as the contents-based classification method. In order to apply the NB algorithm for hierarchical classification, we utilize the simple method of the *Pachinko Machine* NB. The Pachinko Machine classifies instances at internal nodes of the tree, and greedily selects sub-branches until it reaches a leaf [3]. This method is applied after the rule induced by SBI decides the starting category for the Pachinko Machine NB.

In order to evaluate this algorithm, we conducted experiments using the Yahoo! [5] and Google [1] directories as concept hierarchies, and the links (URLs) in each directory as information instances. We conducted ten-fold cross validations for the shared instances. The accuracy is measured for each depth of the Internet directories and is shown in Figure 1. The vertical axes show the accuracy and horizontal axes show the depth of the concept hierarchies. The experimental domains of the graphs are integration into Yahoo!: Automotive, Google: Autos, Yahoo!: Photography, and Google: Photography, from left to right. We compared the proposed system called SBI-NB with the SBI system and the NB system. From the comparison of the results of both NB and SBI with SBI-NB, we can expect high accuracy when NB produces a good result. On the other hand, when the NB method has poor accuracy, our method has at least the same performance of SBI and does not have any side effect from NB. From this, we can conclude that our approach is a good method for integrating the approach of contents-based classification method and the approach of the similarity-based integration method.

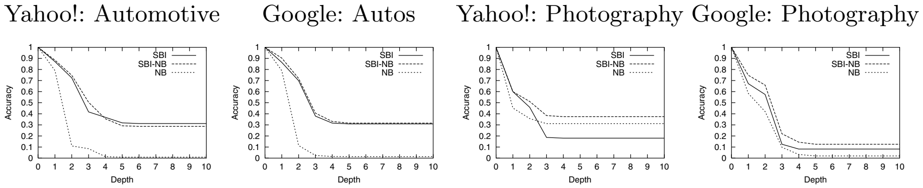
Yahoo!: Automotive Google: Autos Yahoo!: Photography Google: Photography

Fig. 1. Experimental Results

In this paper, we propose a new method for aligning concept hierarchies as a new approach to utilizing information in multiple concept hierarchies. Our experimental results show improved performance compared with the previous approaches.

References

1. Google., http://directory.google.com/, 2003.
2. Ichise R., Takeda H. and Honiden S., Integrating Multiple Internet Directories by Instance-based Learning. In *Proc. of the 18th Int. Joint Conf. on AI*, 22–28, 2003.
3. McCallum A., et al., Improving text classification by shrinkage in a hierarchy of classes. In *Proc. of the 15th Int. Conf. on Machine Learning*, 359–367, 1998.
4. Mitchell. T., *Machine Learning.* McGraw-Hill, 1997.
5. Yahoo!, http://www.yahoo.com/, 2003.