A Multi-strategy Approach for Catalog Integration

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,takeda}@nii.ac.jp, hamasaki@grad.nii.ac.jp

Abstract. When we have a large amount of information, we usually use categories with a hierarchy, in which all information is assigned. This paper proposes a new method of integrating two catalogs with hierarchical categories. The proposed method uses not only the contents of information but also the structures of both hierarchical categories. We conducted experiments using two actual Internet directories, and the results show improved performance compared with the previous approach.

In this paper, we introduce a novel approach for catalog integration problem. The problem addressed in this paper is finding an appropriate category C_t in the target catalog T_C for each information instance I_{si} in the source catalog S_C. What we need to do is determine an appropriate category in T_C for an information instance. In order to solve the problem, we proposed the Similarity-based integration (SBI) [3]. SBI has a higher performance compared with the Naive Bayes (NB) approach, even with the extension proposed by [1]. In this paper, we propose a method which combines the SBI approach and the NB approach. In order to combine handling the meaning of information, we propose using NB after SBI.

A problem of SBI is that it is hard to learn a mapping 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 catalog. In order to avoid this type of rules, we propose to combine a contents-based classification method after we apply the SBI algorithm. Since NB 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 [4]. This method is applied after the rule induced by SBI decides the starting category for the Pachinko Machine NB.

In order to evaluate the proposed algorithm, we conducted experiments using real Internet directories collected from Yahoo! [5] and Google [2]. The data was collected during the period from December 2003 to January 2004. The locations in Yahoo! and Google are Photography. We conducted ten-fold cross validations for the links appeared in both directories. The shared links were divided into

C. Zhang, H.W. Guesgen, W.K. Yeap (Eds.): PRICAI 2004, LNAI 3157, pp. 944–945, 2004.

© Springer-Verlag Berlin Heidelberg 2004

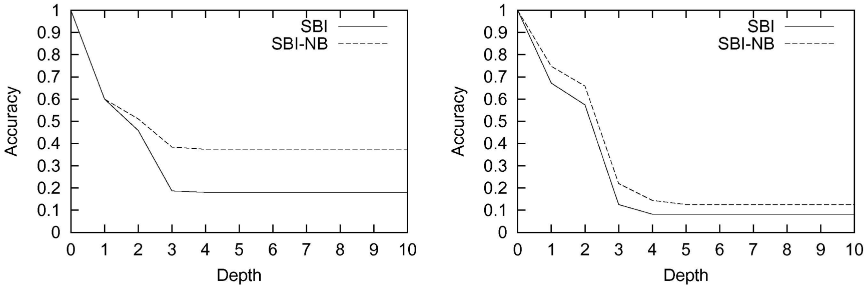

Fig. 1. Experimental Results

ten data sets; nine of which were used to construct rules, and the remaining set was used for testing. Ten experiments were conducted for each data set, and the average accuracy is shown in Figure 1. The accuracy is measured for each depth of the Internet directories. The vertical axes in Figure 1 show the accuracy and horizontal axes show the depth of the concept hierarchies. The left side of Figure 1 shows the results obtained using Google as the source catalog and Yahoo! as the target catalog, and the right side of Figure 1 shows the results obtained using Yahoo! as the source catalog and Google as the target catalog. For comparison, these graphs also include the results of SBI. SBI-NB denotes the results of the method proposed in this paper. The proposed algorithm performs much better in accuracy than the original SBI. One reason for this is that the NB works well. In other words, the contents-based classification is suited for this domain. According to [3], the NB method does not achieve the performance of SBI in the Photography domain. However, our proposed algorithm effectively combines the contents-based method with the category similarity-based method.

In this paper, a new technique was proposed for integrating multiple catalogs. The proposed method uses not only the similarity of the categorization of catalogs but also the contents of information instances. The performance of the proposed method was tested using actual Internet directories, and the results of these tests show that the performance of the proposed method is more accurate for the experiments.

References

1. R. Agrawal and R. Srikant. On integrating catalogs. In *Proc. of the 10th Int. World Wide Web Conf.*, 603–612, 2001.
2. Google. http://directory.google.com/, 2003.
3. R. Ichise, H. Takeda and S. Honiden. Integrating Multiple Internet Directories by Instance-based Learning. In *Proc. of the 18th Int. Joint Conf. on AI*, 22–28, 2003.
4. A. K. McCallum, 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.
5. Yahoo! http://www.yahoo.com/, 2003.