On Hierarchical Web Catalog Integration with Conceptual Relationships in Thesaurus

Ing-Xiang Chen  
Yuan Ze University  
135 Yuan-Tung Rd.  
Chungli, Taiwan, 320

Jui-Chi Ho  
Yuan Ze University  
135 Yuan-Tung Rd.  
Chungli, Taiwan, 320

Cheng-Zen Yang  
Yuan Ze University  
135 Yuan-Tung Rd.  
Chungli, Taiwan, 320

{sean, ricky, czyang}@syslab.cse.yzu.edu.tw

ABSTRACT

Web catalog integration is an interesting problem in current digital content management. Past studies have shown that using a flattened structure with auxiliary information extracted from the source catalog can improve the integration results. However, the nature of a flattened structure ignores the hierarchical relationships, and thus the performance improvement of catalog integration may be reduced. In this paper, we propose an enhanced hierarchical catalog integration (EHCI) approach with conceptual thesauri extracted from the source catalog. The results show that our enhanced hierarchical integration approach effectively boosts the accuracy of hierarchical catalog integration.

Categories and Subject Descriptors

H.3.2 [Information Storage]: Record classification; H.3.5 [On-line Information Services]: Web-based services

General Terms

Algorithms, Experimentation

Keywords

hierarchical catalog integration, conceptual relationships, thesaurus

1. INTRODUCTION

With the explosive growth of various kinds of Web information, an integrated Web catalog is becoming important service for on-line vendors and Internet users [1, 2, 4]. An early study has shown that only about 20% of the categorized sites retrieved from both Yahoo! and Google catalogs are the same [2], which means that users may need to spend much effort browsing different Web catalogs to obtain the required materials.

As noted in [1], catalog integration is not just a classification task because when some implicit source information is exploited, the integration accuracy can be highly improved. A foremost approach has first been proposed to enhance the Naive Bayes classifier with implicit source information [1]. Other studies (e.g. [2, 4]) employ SVMs to enhance the accuracy performance. Since these studies only consider a flattened integration structure, they completely ignore the hierarchical relationships between the categories and subcategories in the destination catalog, and thus the integration accuracy is restrained.

Past studies have shown that exploiting a hierarchical structure in classification may bring better advantages than using a flattened structure [3]. However, it has not been testified in Web catalog integration. In this paper, we propose an enhanced hierarchical catalog integration (EHCI) approach with conceptual relationships extracted from the source catalog thesaurus to improve the integration performance. We applied SVM classifiers to the EHCI approach in our experiments and compared its performance with that of a simple hierarchical catalog integration approach (SHCI), which is designed referring to [3] and [5]. The experimental results have shown that the EHCI approach effectively boosts the accuracy of hierarchical catalog integration over all categories.

2. HIERARCHICAL INTEGRATION

In the integration process, we assume that there are two hierarchical catalogs participating in the integration process. Figure 1 illustrates the process of automated hierarchical catalog integration. One is the source catalog \( S \) with a set of \( m \) categories \( S_1, S_2, \ldots, S_m \), their subcategories, and so on down to the lowest-layered subcategories. The other is the destination catalog \( D \) with a set of \( n \) categories \( D_1, D_2, \ldots, D_n \), their subcategories, and so on down to the lowest-layered subcategories. The integration process is performed by merging each document \( d \) in \( S \) into a correspondent destination category in \( D \). That is, for each layered directory in the hierarchy, training documents in each directory are trained as directory classifiers and local classifiers to help each document \( d \) integrate into a corresponding directory. Only the documents integrated into the corresponding layered categories and subcategories are viewed as correctly integrated.

3. THE INTEGRATION APPROACH

In hierarchical catalog integration, a one-against-rest strategy is used at each decision point in the hierarchy to integrate documents into the matching directory, and the destination classifiers are trained with the hierarchy labels for further enhancement. To improve the integration accuracy, a weight formula is designed to extract the semantic concepts existing in the source catalog. In Equation 1, the weight of each thesaurus is exponentially decreased and accumulated according to the increased layers to represent the semantic
The threshold more distinctive to classify the documents into the correct label information, the classifiers can thus be trained to be transformed and added into the test documents.

W. With such a thesaurus weighting scheme, the conceptual relationships of the layered source categories can be accommodated to enhance the weights of the source thesauri to enhance the destination classifiers, and so is the value of \( \lambda \) set in the native destination category.

Likewise, to build enhanced classifiers in destination categories, the EHCI scheme is used in the destination catalog. With the enhancement of the features and native category label information, the classifiers can thus be trained to be more distinctive to classify the documents into the correct categories. The threshold \( \lambda \) is heuristically set as 0.05 to accommodate the weights of the source thesauri to enhance the destination classifiers, and so is the value of \( \lambda \) set in the native destination category.

\[
FeatureWeight = \lambda \times \frac{L_x}{\sum_{i=0}^{n} L_i} + (1 - \lambda) \times f_x
\]

4. EXPERIMENTAL RESULTS

The EHCI approach described in the previous section was applied to a collection of 17888 pages retrieved from the Yahoo! catalog and a collection of 17890 pages retrieved from the Google catalog. In Table 1, a set of 1472 classes in the Yahoo! catalog \( (C_Y) \) and a set of 3343 classes in the Google catalog \( (C_G) \) were organized according to the original hierarchy in a depth of six layers. The test documents are selected by intersecting documents of Yahoo! with those of Google, namely \( Y \cap G \), in which the number of \( t_Y \) and \( t_G \) is different in the sense that some test documents may appear in more than one class simultaneously. The training documents of the Yahoo! catalog \( (T_{Y-C}) \) and the Google catalog \( (T_{G-C}) \) are gathered by subtracting the intersected documents, namely \( Y \cap G \). In this experiment, the documents are both integrated from Yahoo! into Google and from Google to Yahoo!. Figure 2 illustrates that the overall performance of EHCI outperforms the original HCI approach in both Yahoo!-to-Google and Google-to-Yahoo! catalog integration. The result further shows that the conceptual relationships are consistently effective in catalog integration.

5. CONCLUDING REMARKS

This paper reports our studies on the effects of a hierarchical scheme to enhance the integration accuracy. By exploiting the hierarchical relationships between categories and subcategories, the improvement of integration accuracy is very promising. It shows that a hierarchical integration scheme is effective for Web catalog integration, and our EHCI approach consistently achieves improvements on real-world catalogs with SVM classifiers.

6. ACKNOWLEDGMENTS

This work was supported in part by National Science Council of R.O.C. under grant NSC-94-2213-E-155-050.

7. REFERENCES

[1] R. Agrawal and R. Srikant. On integrating catalogs. Proc. WWW10, pages 603–612, May 2001.
[2] I.-X. Chen, J.-C. Ho, and C.-Z. Yang. An iterative approach for web catalog integration with support vector machines. Proc. AIRS’05, pages 703–708, Oct. 2005.
[3] S. Dumais and H. Chen. Hierarchical classification of web content. Proc. SIGIR’00, pages 256–263, Jul. 2000.
[4] S. Sarawagi, S. Chakrabarti, and S. Godbole. Cross-training: Learning probabilistic mappings between topics. Proc. SIGKDD’03, pages 177–186, Aug. 2003.
[5] A. Sun, E.-P. Lim, and W.-K. Ng. Performance measurement framework for hierarchical text classification. JASIST, 54(11):1014–1028, Jun 2003.