Site Abstraction for Rare Category Classification in Large-Scale Web Directory

Tie-Yan LIU[1], Hao WAN[2*], Tao QIN[2*], Zheng CHEN[1], Yong REN[2], Wei-Ying MA[1]

[1]Microsoft Research Asia
5F, Sigma Center, No. 49, Zhichun Road
Haidian District, Beijing, 100080, P. R. China
{t-tyliu, zhengc, wyma}@microsoft.com

[2]Dept. Electronic Engineering,
Tsinghua University
Beijing, 100084, P.R. China
{wanhao, qinshitao99}@mails.tsinghua.edu.cn
reny@tsinghua.edu.cn

ABSTRACT

Automatically classifying the Web directories is an effective way to manage Web information. However, our experiments showed that the state-of-the-art text classification technologies could not lead to acceptable performance in this task. Due to our analysis, the main problem is the lack of effective training data in rare categories of Web directories. To tackle this problem, we proposed a novel technology named Site Abstraction to synthesize new training examples from the website of the existing training document. The main idea is to propagate features through parent-child relationship in the sitemap tree. Experiments showed that our method significantly improved the classification performance.

Categories and Subject Descriptors

H.4.m [Information System Applications]: Miscellaneous; I.5.4 [Pattern Recognition]: Applications – Text processing.

General Terms

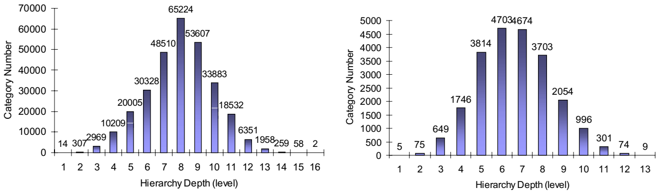
Performance, Design, Experimentation, Verification.

Keywords

Text Classification, Site abstraction, Web directory, Hierarchical Classification, Support Vector Machines (SVM).

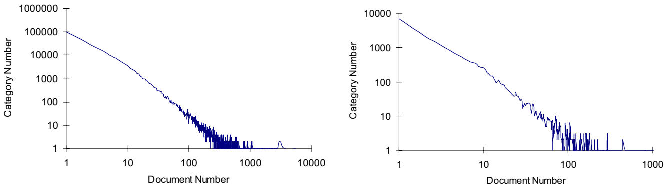
1. INTRODUCTION

With the explosive growth of the Web, it becomes more and more difficult to manage the Web information. In early stage, people manually categorized Web pages into Web directories such as Yahoo! Directory (http://dir.yahoo.com/) and Open Directory Project (ODP, http://dmoz.org/). However, manually labeling is time-consuming and labor-expensive, which makes it not scalable with respect to the high growing speed of the Web. Therefore, automated text categorization (TC) technologies were adopted in many previous works to categorize Web pages [1][2][6]. However, these works were more of demonstrations than solutions because the sampling strategies used in them (only top few levels or selective common categories) could not well reflect the characteristics of Web directories. To tackle this problem, in this paper, we propose to use a specific subset of Yahoo! Directory named MERG which has very similar statistics to the full set for the experimental study. MERG consists of five sub trees, namely "News and **M**edia", "**E**ntertainment", "**R**eference", "**G**overnment" and "Regional". It contains totally 22,803 categories and 54,542 documents which are organized into a 13-level hierarchy. Some comparisons between MERG and Yahoo! Directory are shown as below.



(a) Yahoo! Directory (b) MERG

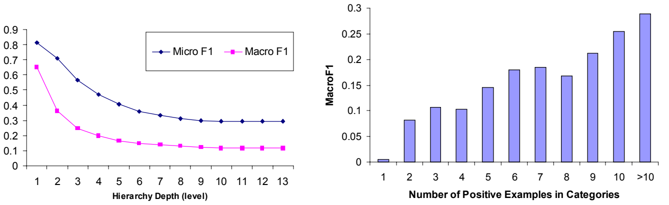
Figure 1. Category distributions over levels.



(a) Yahoo! Directory (b) MERG

Figure 2. Power law distributions of the category size.

Over the MERG data set, we ran hierarchical SVM classification with similar settings to [2] and [3]. The corresponding results were shown in Figure 3(a). From this figure, we can see that the classification performance was disappointing: Macro-F1 of 0.116 and Micro-F1 of 0.237 (see the performance for the lowest level, which corresponds to the classification of the full set of MERG) are not acceptable for real-world classification applications. To find out why hierarchical SVM performed so poor, we listed the classification performance with respect to category size (the number of positive examples in the training set) in Figure 3(b).



(a) (b)

Figure 3. Performance of hierarchical SVM classification over MERG.

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

*The works of Hao WAN and Tao QIN were performed at Microsoft Research Asia

From this figure we can see that the classification performance decreased with the decreasing category size. That is, the data sparseness problem in rare categories might be the major reason for the overall poor classification performance. This motivates us to investigate how to expand the training set in order to improve the classification performance. As a result, we propose a novel method named Site Abstraction which utilizes the structure of the website to synthesize new training examples.

2. SITE ABSTRACTION

Due to our observation, almost all the labeled documents in Web directories are entry pages (denoted by *e*) of websites (denoted by **S**). In conventional Web directory classification works [1][2][6], only the entry page *e* was used as training document. However, it is easy to understand that other pages in **S**-{*e*} are also relevant to the category label. Therefore, we propose to utilize the information embedded in **S**-{*e*} to synthesize some new training documents for training set expansion. As a demonstration of this idea, we develop a mechanism named Site Abstraction to propagate the features (terms) of those pages in the lower levels of the sitemap tree to their ancestors.

$$F^*(p_k)=\begin{cases} F(p_k) & \mathbf{CHILD}(p_k)=\Phi \\ F(p_k)+\alpha\dfrac{\sum_{p_{k+1}\in\mathbf{CHILD}(p_k)}F(p_{k+1})}{|\mathbf{CHILD}(p_k)|} \quad k>1 & \\ & \mathbf{CHILD}(p_k)\neq\Phi \\ \alpha\dfrac{\sum_{p_{k+1}\in\mathbf{CHILD}(p_k)}F(p_{k+1})}{|\mathbf{CHILD}(p_k)|} \quad k=1 & \end{cases} \tag{1}$$

where Φ represents the empty set, *F* is the original feature of page p_k, F^* is the refined feature after propagation, **CHILD**(p_k) denotes the set of child pages of p_k, |.| is the number of pages in a set.

The above propagation starts from the lowest level in the sitemap tree. When it eventually terminates at the first level, a new training document is synthesized. By including this new document into the training set, we can solve the data sparseness problem to some extent. Note that this new page is based on the pages in **S**-{*e*}, but independent of the entry page *e*. Therefore, it can serve as a good complement to the existing training document *e*.

As for the above propagation process, one may argue that there is possibly the risk of topic drift. We acknowledge this; however since the influences of the lower-level pages are restrained by the weighting factor *α*, this risk will not be so high.

3. EXPERIMENTS

In this section, we tested the effectiveness of our Site Abstraction technology. For simplicity, we selected those categories with only one positive training example in the original training set for experiments. We totally crawled 3 levels and no more than 100 pages from the website corresponding to the existing positive examples. After that, we used the URL information (folder depth) of each crawled page to construct the sitemap tree of a website according to [3]. Then we synthesized one more training documents for each category by Site Abstraction. For comparison, we also implemented some other works targeting training set expansion (i.e. LiveClassifier [4] and Document-self Expansion [7]), as well as the approach of using all downloaded pages directly as training examples without applying the propagation formula (1) (denoted by "Site Raw"). Our experimental results were shown in Figure 4.

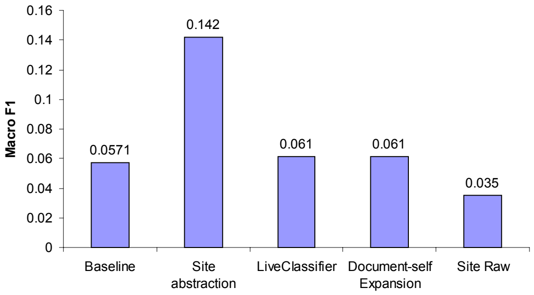


Figure 4. Classification results over the 1-doc categories.

From this figure, we can see that both LiveClassifier and Document-self Extension did not perform as well as expected. They did improve the classification accuracy, but only marginally. An interesting observation is that Site Raw actually deteriorated the performance, indicating the existence of concept drift. Comparatively, Site Abstraction did improve the classification accuracy significantly. The relative improvement was over 149%. This showed that Site Abstraction could inhibit the concept drift and construct informative training examples. If looking at the absolute value of the resulting classification performance of Site Abstraction, we found that although we only added one more training document, the performance over the categories with only one document had been almost as good as over those categories with five or six training documents (See Figure 3(b)). This verified from another aspect the effectiveness of Site Abstraction: one synthesized page is much more than one real page.

4. CONCLUSION

In this paper, we first conducted experiments to show that the state-of-the-art text classification methods could not well handle Web directory classification. Then we pointed out that the data sparseness problem in rare categories of Web directories is the major reason for the poor classification performance. To tackle this problem, we proposed a novel technology named Site Abstraction to synthesize new training examples based on the website of the existing training document. The main idea is to propagate features through parent-child relationship in the site structure. Experiments showed that our method significantly (relatively 149%) improved the classification performance.

5. REFERENCES

[1] Calvo R. A., Lee J. M. and Li X., Managing Content with Automatic Document Classification. *Journal of Digital Information*, Vol.5, No.282, 2004.

[2] Dumais S. and Chen H., Hierarchical classification of Web content. SIGIR 2000, 256-263.

[3] Feng G., Liu T. Y., Ma W. Y., *et al*, Level-based Link Analysis, *APWeb* 2005.

[4] Huang C. C., et al. Liveclassifier: creating hierarchical text classifiers through web corpora. *WWW* 2004, 184-192.

[5] Lewis, D. D., Yang, Y., Rose, T., Li, F. RCV1: A new benchmark collection for text classification research. *Journal of Machine Learning Research*. 5 (2004) 361-397

[6] McCallum, A., Rosenfeld, R., Mitchell, T. and Ng, A. Improving text classification by shrinkage in a hierarchy of classes. *ICML* 1998*,* 359-367.

[7] Tseng Y. H. and Juang D. W., Document-Self Expansion for Text Classification, *SIGIR* 2003, 399-400.