Grouped ECOC Conditional Random Fields for Prediction of Web User Behavior

Yong Zhen Guo, Kotagiri Ramamohanarao, and Laurence A.F. Park

Department of Computer Science and Software Engineering
University of Melbourne, Australia
yzguo@csse.unimelb.edu.au

Abstract. Web page prefetching has shown to provide reduction in Web access latency, but is highly dependent on the accuracy of the Web page prediction method. Conditional Random Fields (CRFs) with Error Correcting Output Coding (ECOC) have shown to provide highly accurate and efficient Web page prediction on large-size websites. However, the limited class information provided to the binary-label sub-CRFs in ECOC-CRFs will also lead to inferior accuracy when compared to the single multi-label CRFs. Although increasing the minimum Hamming distance of the ECOC matrix can help to improve the accuracy of ECOC-CRFs, it is still not an ideal method. In this paper, we introduce the grouped ECOC-CRFs that allow us to obtain a prediction accuracy closer to that of single multi-label CRFs by grouping the binary ECOC vectors. We show in our experiments that by using the grouping method, we can maintain the efficiency of the ECOC-CRFs while providing significant increase in Web page prediction accuracy over ECOC-CRFs.

Keywords: Web Page Prediction, Conditional Random Fields, Error Correcting Output Coding, Grouping.

1 Introduction

Many Internet users have turned to wireless devices such as mobile phones or PDAs due to their mobility and convenience, even though the connection speed of these wireless devices are usually slow. Meanwhile, there are still many dial-up users with low bandwidth access to the Internet. On account of the limited bandwidth and low-speed connections, many wireless and dial-up Internet users need to spend long periods of time waiting for the requested Web pages to be transferred to them through Internet, which may lead to intolerable delays.

Web page *prefetching* is an effective way to reduce the access latency for Web users. However, if most of the prefetched Web pages are not visited by the users in their subsequent accesses, the limited network bandwidth and server resources will not be used efficiently, and hence may worsen the access latency problem. Therefore, the success of a prefetching method relies heavily on the Web page prediction accuracy.

Conditional Random Fields (CRFs) [1] are powerful probabilistic framework for classifying sequential data. Although the training complexity of a CRF is

T. Theeramunkong et al. (Eds.): PAKDD 2009, LNAI 5476, pp. 757–763, 2009.

very high, by using Error Correcting Output Coding (ECOC) [4] to decompose a multi-label CRF training into many binary-label sub-CRF trainings, the overall training complexity can be decreased. Therefore, ECOC-CRFs can be efficiently used on data sets containing many unique labels and are ideal for predicting an Internet user's future Web access patterns on large scale websites [5]. However, because the class information given to each sub-CRF is drastically decreased, the prediction accuracy of ECOC-CRFs is inferior to that of single multi-label CRFs. In this paper, we propose the grouped ECOC-CRFs, which can maintain the advantages of ECOC-CRFs while yielding a higher accuracy, thus can be used to improve the performance of ECOC-CRFs based Web page prediction further.

The rest of this paper is organized as follows: in Section 2, we briefly describe the ECOC-CRFs for Web page prediction, which is followed by the comparison of ECOC-CRFs and single CRFs in Section 3. We propose our novel grouped ECOC-CRFs to improve the prediction accuracy in Section 4 with experiments. Finally, we conclude this paper in Section 5.

2 ECOC-CRFs for Web Page Prediction

The complexity of a CRF training is proportional to the square of the number of labels (in Web page prediction scenario, unique Web pages), when used on large-scale websites containing thousands of unique Web pages, the CRF training will become highly resource intensive and even infeasible. If we can decompose a multi-label CRF training into a series of sub-CRF trainings, each with much fewer labels, the total training complexity will be reduced significantly. Error correcting output coding (ECOC) can be used to decompose a multi-label classification task into a set of binary-label classification sub-tasks [3], it consists of two steps: firstly, construct a code matrix for the original multi-label classification task and train a series of binary sub-classifiers for each column of the code matrix; secondly, combine the decoding results of each binary sub-classifier to obtain the possible result for the original classification task. Therefore, we can employ ECOC to decompose a multi-label CRF training into many binary-label sub-CRF trainings. Since all the sub-CRFs are binary, they can be trained very efficiently: an ECOC-CRF training can be completed in a fraction of the time of a single multi-label CRF training. In addition, all sub-CRFs can be trained in parallel to save the overall training time. It has been shown in [5] that we are able to efficiently predict Web pages that a user may access using ECOC-CRFs in large-size websites, in [5] it also explains how to use *Search Coding* to design an ECOC code matrix with good row separation.

Although by using ECOC-CRFs the total training complexity can be reduced, the class information given to each sub-CRF is drastically decreased as well. As a result, the prediction accuracy of ECOC-CRFs is lower than that of the single multi-label CRFs, which can be seen from the experimental results in next section. In this paper, we propose the grouped ECOC-CRFs to improve the accuracy of ECOC-CRFs based Web page prediction, while maintain the advantages of ECOC-CRFs.

3 Comparison of ECOC-CRFs and Single CRFs

In this section, we evaluate the Web page prediction performance of ECOC-CRFs and single CRFs on the publicly accessible *msnbc* dataset [6]. In this dataset, all the user visits are recorded in session format at the level of page categories. There are 17 different page categories which can also be treated as 17 distinct pageviews. We randomly selected $50,000$ distinct sessions with length more than 5 and less than 100 from the preprocessed dataset and divided them into two subsets: $45,000$ sessions for training and $5,000$ for testing.

For the implementation of CRF training, we use the CRF++ toolkit [8]. We used three CRF++ feature templates in our experiments. In the first template (referred to as CRF1), we define the current and previous one observation and their combination as the unigram features; for the second template (CRF2), we use the current and previous two observations and their combinations as the unigram features; we define the third template (CRF4) similarily. These templates share a same bigram feature, which will automatically generate the combination of the current label and previous label as the feature functions. The more abundant features in use, the more powerful CRF models can be obtained.

As a baseline, we trained 3 single multi-label CRFs on the *msnbc* dataset using the feature functions of CRF1 ,CRF2 and CRF4, and evaluated their page prediction accuracies by calculating the ratio of the number of correctly predicted labels to the total number of predicted labels. We also implemented the ECOC-CRFs by using *Search Coding* to design the code matrix. Figure 1 shows the accuracies of single multi-label CRFs and their corresponding ECOC-CRFs with ECOC code matrix parameters: code word length $n = 16$ and minimum Hamming distance between each code word $d = 8$, in which the prediction accuracies of the 1^{st} and 2^{nd}-order Markov Chain (referred to as 1^{st}-MC and 2^{nd}-MC respectively) are shown for comparison.

From Figure 1 we can see that all single CRFs produce much higher prediction accuracy than Markov Chains, indicating the superiority of CRFs to Markov Chains in predicting labels of sequences. We can also see that the accuracy of an ECOC-CRF is lower than its corresponding single multi-label CRF. This

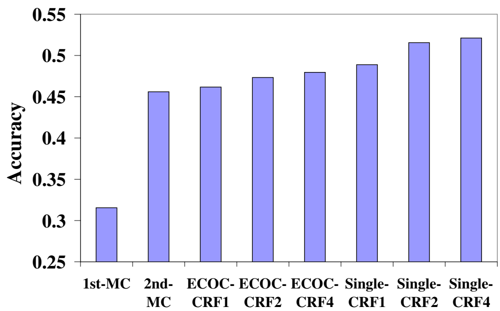

Fig. 1. Web page prediction accuracy of Markov Chains, single CRFs and ECOC-CRFs with code length $n = 16$ and minimum Hamming distance $d = 8$ on the *msnbc* dataset.

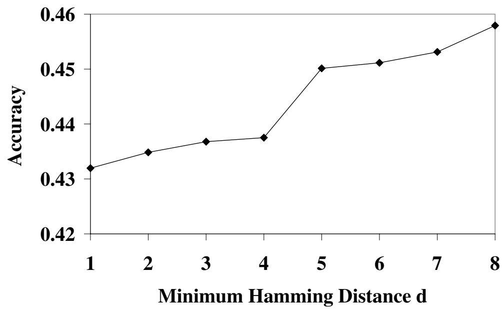

Fig. 2. Accuracy of ECOC-CRF1 with different ECOC code matrices on the *msnbc* dataset

reduction in accuracy is due to the limited class information provided to the binary sub-CRFs when compared to the single multi-label CRFs. However, we will show later that by using several approaches, the prediction accuracy of ECOC-CRFs can be improved further.

By choosing the ECOC code matrix with a larger minimum Hamming distance d, we can increase the number of sub-classification errors that can be corrected. Therefore, we can improve the accuracy of ECOC-CRFs by using an ECOC code matrix with larger minimum Hamming distance. We show this by an experiment on the *msnbc* dataset, in which we design a series of ECOC code matrices, whose minimum Hamming distances d are ranging from 1 to 8 and code lengths n vary from 5 to 16 accordingly(the minimum code length that satisfies the given minimum Hamming distances d and the number of code words m=17 by using Search Coding). Then we evaluated the accuracies of ECOC-CRF1 by using these 8 ECOC code matrices, the results in Figure 2 show that the prediction accuracy of ECOC-CRF1 increases as d increases.

However, as the minimum Hamming distances d increases, the length of code words n increases as well. Although long code words can lead to good row-separation, they increase the chance that two matrix columns are similar to each other as well, in which case the two corresponding sub-classifiers will learn similar concepts. Furthermore, the increase in the length of code words will also result in the increase in the number of sub-classifiers to be trained, which will add up the overall training time. Therefore, improve the accuracy of ECOC-CRFs by simply enlarging the minimum Hamming distances d is not an ideal method, we should find a tradeoff between the length of ECOC code words and accuracy. In the following section we will propose a grouping method which can improve the prediction accuracy of ECOC-CRFs without adding the length of ECOC code words.

4 ECOC-CRFs with Grouping

By decomposing a multi-label CRF classification problem into many independent binary classification problems, we achieved faster training times, but we also drastically reduced the class information given to each sub-CRF. If we can provide each sub-CRF with more abundant class information, the prediction accuracy can be improved. This leads to the consideration of grouping, which divides the columns of the ECOC code matrix into several groups and uses each group to train a sub-CRF. For example, an ECOC code matrix containing 12 columns can be divided into 4 groups containing 3 columns each, in which case the grouping size (the number of possible labels for each sub-CRF to learn) is $2^3 = 8$.

By grouping, the number of sub-CRFs needed will be cut down from n to $\lceil n/\log_2 G\rceil$, where n is the length of ECOC code words and G is the grouping size; meanwhile, the number of possible labels for each sub-CRF to learn will increase from 2 to G. Therefore, the training complexity can be reduced from $O(L^2)$ to $O(\lceil n/\log_2 G\rceil \times G^2)$, where L is the total number of labels. When the grouping size increases, each sub-CRF will obtain more refined information about the class labels, and hence, the accuracy should also increase. A grouping size of L (meaning all the columns are put into one group) implies that the ECOC-CRFs will reduce to single CRFs. An example grouping process is shown in Figure 3.

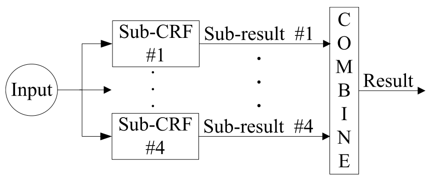

Fig. 3. Example grouping process when code length $n = 8$ and grouping size $G = 4$ (2 columns per group), therefore 4 sub-CRFs are needed

We performed experiments on the *msnbc* dataset to measure the effect of grouping on ECOC-CRF4 using two different experimental setups: (1) code length $n = 8$, minimum Hamming distance $d = 2$; and (2) $n = 16$, $d = 8$. These results can be found in Figure 4, from which we can see that as the grouping size of ECOC-CRFs grows, the overall trend of prediction accuracy increases. When the grouping size becomes big enough to include all L labels, the ECOC-CRFs can perform as well as the single multi-label CRFs. Therefore, by incorporating grouping, the accuracy of ECOC-CRFs can be improved further without increasing the code length.

In Figure 4 we also illustrate the change in training time of ECOC-CRFs (measured by adding the training time of all the sub-CRFs) versus the grouping size using the identical experimental setups. We can see that when the grouping size becomes larger, the total training time of ECOC-CRFs will increase and even exceed the training time of single multi-label CRFs. This is because the number of labels that a sub-CRF needs to learn increases as the grouping size increases, which in turn increases the training time dramatically, and since there are numerous sub-CRFs to be trained, the total training time can be longer than

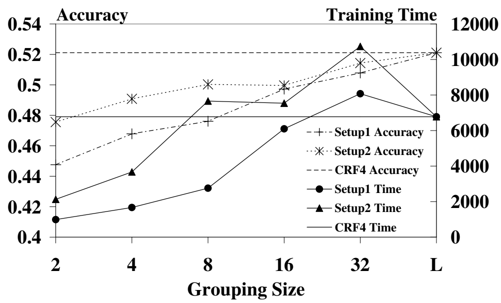

Fig. 4. Relationships between accuracy, training time (in seconds) and grouping size of Setup1 (n = 8, d = 2) and Setup2 (n = 16, d = 8) on *msnbc* dataset, where the accuracy and training time of single CRF4 are shown as a baseline. When grouping size is L, ECOC-CRFs reduce to single CRFs.

that of single CRFs. Therefore, it is important to select an appropriate grouping size to provide high accuracy and fast training.

We also conducted experiments on the CSSE dataset [7] which contains 3,829 unique Web pages to evaluate the scaling performance of grouped ECOC-CRFs on Web page prediction. After preprocessing, we randomly select 2,723 sessions as the training data and 544 sessions as the testing data. We conducted the experiments on this dataset by using different grouping sizes on two experimental setups with ECOC-CRF1: (1)code number $m = 3,829$, code length n=16, minimum Hamming distance d=1, and (2) $m = 3,829$, n=24, d=8. We recorded the prediction accuracy and the training time for different grouping size in Table 1. Due to the large number of labels (3,829 unique Web pages) and the limitation of our computational resources, the training of single multi-label CRFs and ECOC-CRFs with grouping size of L (in this case the grouped ECOC-CRFs will reduce to single CRFs) on this dataset are infeasible.

Table 1. Prediction accuracy and training time of Setup1 ($m = 3,829$, n=16, d=1) and Setup2 ($m = 3,829$, n=24, d=8) on CSSE dataset

Grouping Size	Setup1		Setup2	
(Label Numbers)	Accuracy	Time	Accuracy	Time
2	64.0%	51s	69.2%	75s
4	63.9%	63s	70.0%	84s
8	64.8%	108s	70.4%	146s
16	64.7%	267s	70.7%	343s
32	66.6%	736s	72.2%	942s
64	68.2%	1,995s	74.0%	3,815s
128	70.1%	8,151s	74.9%	12,042s
256	72.9%	35,302s	75.0%	49,629s
3,829(Single CRF)	infeasible		infeasible	

From Table 1 we can observe that on the CSSE dataset, with a bigger d, ECOC-CRFs can yield more accurate Web page prediction accuracy. We can also see that the grouping method can help to improve the accuracy of ECOC-CRFs significantly. When we compare the accuracy of grouping size of 256 and grouping size of 2 (no grouping), the improvements in accuracy are 8.9% for Setup1 and 5.8% for Setup2 respectively. Additionally, while the grouping size enlarges, the training time of both setups increase dramatically, and it is infeasible to train a single multi-lable CRF on this dataset. Therefore, we can use grouped ECOC-CRFs to obtain satisfactory Web page prediction performance in large scale websites.

5 Conclusion

By using ECOC-CRFs to decrease the training complexity we also reduce the class information provided to each sub-CRF, thus the prediction accuracy of ECOC-CRFs is lower than the corresponding single CRFs. In this paper, we proposed the grouped ECOC-CRFs. By dividing the columns of ECOC matrix into several groups, each sub-CRF can learn more class information which will lead to higher accuracy. Using grouped ECOC-CRFs we are able to obtain a prediction accuracy closer to that of single multi-label CRFs while maintain the advantages of ECOC-CRFs. Our experiments have shown that the grouped ECOC-CRF Web page prediction is highly accurate and scalable, and hence ready for use with large scale Web sites.

References

1. Lafferty, J., McCallum, A., Pereir, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, In: Proceedings of the 18th International Conference on Machine Learning, pp. 282–289 (2001)
2. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations 1, 12–23 (2000)
3. Berger, A.: Error-correcting output coding for text classification. In: Proceeding of IJCAI: Workshop on machine learning for information filtering (1999)
4. Dietterich, T., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research (1995)
5. Guo, Y.Z., Ramamohanarao, K., Park, L.: Error Correcting Output Coding-Based Conditional Random Fields for Web Page Prediction. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 743–746 (2008)
6. UCI KDD Archive, `http://kdd.ics.uci.edu/databases/msnbc/msnbc.html`
7. The Department of Computer Science and Software Engineering, University of Melbourne, `http://csse.unimelb.edu.au`
8. CRF++: Yet another CRF toolkit, `http://crfpp.sourceforge.net`