Construction of a Hierarchical Classifier Schema using a Combination of Text-Based and Image-Based Approaches

Cheng Lu and Mark S. Drew
School of Computing Science
Simon Fraser University
Vancouver, B.C., CANADA V5A 1S6
(604) 291-4682 Fax (604) 291-3045
{clu, mark }@cs.sfu.ca

ABSTRACT
Web document hierarchical classification approaches often rely on textual features alone even though web pages include multimedia data. We propose a new hierarchical integrated web classification approach that combines image-based and text-based approaches. Instead of using a flat classifier to combine text and image classification, we perform classification on a hierarchy differently on different levels of the tree, using text for branches and images only at leaves. The results of our experiments show that the use of the hierarchical structure improved web document classification performance significantly.

Topic Area:
Web IR

1. INTRODUCTION
Multimedia data on the web includes rich picture content accompanying text. Web documents are ideally suited to application of a combination schema that takes advantage of the images in web documents and integrates image information with text information to obtain a high-performance classifier.
We propose a new classification schema combining a text-based classification approach and an image-based approach to a structured hierarchy of topics. We divide the classification task into smaller classification problems corresponding to the branches in the classification hierarchy and apply a text-based approach and an image-based approach on different levels separately. Each of these subtasks is thus significantly simpler and their features are better discriminated than treating them as a unified task in a flattened classifier. It is then possible to make a determination based only on a small set of significant features corresponding to different levels. Thus the resulting models are more efficient and achieve a higher accuracy.

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA.

2. TEXT FEATURE TRAINING AND CLASSIFICATION
There are two major components in our text classification process: an HTML parser and a classifier.

2.1 HTML Parser
The HTML parser takes an input HTML file for a web page and outputs a feature list for the page parser, which includes the stems of the keywords and their occurrence within the page.

2.2 Classifier
The two main phases for the classifier are training and classifying. Training produces an evaluation model for classification. To construct the evaluation model, a feature selecting process is needed. We use Fisher's discriminant [1] to calculate scores of all the attributes from the HTML Parser and then carry out feature selection based on the scores.

We choose a Naïve Bayesian-based classifier [2] to perform the text-based classification task. This classifier uses Bayes' rule to carry out classification from the text features of the given web page. Via Bayes' rule, a posteriori probability can be calculated for each class given a web document.

3. IMAGE-BASED CLASSIFICATION APPROACH
We use an image classification approach that draws on visual information: OF*IIF scoring of images [3].

3.1 Illumination-Invariant Image Preprocessing
We adjust this technique to improve accuracy and efficiency of image classification by using a much more expressive image feature vector than the usual color histogram.
We normalize the length of each of the R,G,B color channels, taken as a long vector [4]. Such normalization has the upshot of effectively placing each image in the same illumination environment, and is based on well-established theory in color science. Then, we go to a 2D chromaticity space:

$$r = R/(R+G+B); \; g = G/(R+G+B)$$

This has the effect of normalizing away shading effects (in a Lambertian shading model), and together the two steps very effectively capture the salient features of an image.

3.2 Color Feature Histogram Clustering, Training and Classification

3.2.1 Clustering

Chromaticity histograms are derived from color-channel-normalized images, using 16 bins for each of *r* and *g*; however, since $r + g \leq 1$, only half the bins are populated, and we use a histogram with $16^2/2+16=144$ bins. Once visual features have been extracted for the blocks of all training images, feature vectors associated with the blocks are clustered using k-means clustering with k equal to the number of classes. The output is a set of k clusters that minimize the Euclidean distance.

3.2.2 Training and Classification

Cluster means (i.e., image features) are associated with class features via training. We use the OF*IIF approach [3] to perform the image-based classification task. In [3], each histogram in the target image has to investigate every feature by computing the Euclidean distance so as to decide to which feature the histogram belongs. Here, we propose a novel matching approach called *binary-tree matching* to avoid heavy computational load. Its key concept is to build a tree structure of feature histograms by pairing these histograms and then pairing the median vectors of previous pairs recursively until the last median is produced: this vector is the root of the tree. When the histogram for a new image is introduced, we start at the root and then go to the closer of the root's two child vectors. The process repeats until the histogram reaches one leaf (i.e., feature). It is easy to see that this new approach can decrease the computational load dramatically and improve efficiency of classification.

4. EXPERIMENTS

We first conduct a simple experiment where web documents from computer science department web sites are chosen to construct a hierarchical classifier that contains two levels of classes: research area classes and person-group classes. The structure of the classifier is given in Fig. 1(a). In this structured hierarchy, the basic guiding principle is that the topics on the 'area' level are very different for different classes; thus it is easy to retrieve significant text features. On the other hand, the text content of the pair of sub-classes on the person-group level may be quite similar, but the images on the person pages have many common features that are drastically different from the images on group pages.

We selected 25 web pages with information regarding Faculty, Graduate, Research Group, and Lab from each of three universities' computer science departments and labeled them according to different research areas and person or group (on the basis of human judgment). We then extracted 25 web pages from the computer science departments at another four universities to construct a test set. We compared our hierarchical classification schema to the flattened classification schema [3] summing scores from an image-based and text-based approach. Table 1 shows classification results using the above classification schemas, and shows a significant increase of accuracy.

We also conducted a challenging and more practical experiment on the document hierarchy of Yahoo. The hierarchy we studied includes two levels, shown in Fig.1 (b). The top level of Yahoo has 13 classes. We observe that even though there are many images accompanying the web documents, the image-based approach will not work well since the color information has such a wide dispersion that discriminating color features will not be able to characterize images sufficiently well.

For the second level, we chose the Recreation & Sports hierarchy (http://sports.yahoo.com) because its deep classes are well suited for studying the effectiveness of our schema. There are 10 classes on this level. For each of them, about 500 randomly selected documents were chosen from Yahoo Sports for training and 200 for testing. Results are shown in Table 2. The precision of our hierarchical classifier is very encouraging.

5. CONCLUSIONS

Our work introduces a novel classification schema that efficiently utilizes multimedia data for web data mining. The key to the success of this schema is the combination of two efficient techniques that are used to solve different sub-problems separately in a structured classification hierarchy. The normalized image chromaticity feature used, which more expressively captures image features, and the matching approach are both also new to web information retrieval and substantially improve classification performance.

Table 1. Accuracy percentages for classification schemas, Universities (100 Pages)

Test Data Set	Hierarchical	Flat
Research Area classes	87.0	78.0
Person-group classes	90.0	85.0

Table 2. Accuracy percentages for classification schemas, Yahoo (200 Pages)

Test Data Set	Hierarchical	Flat
Yahoo Area classes	78.0	70.0
Sports classes	87.0	78.0

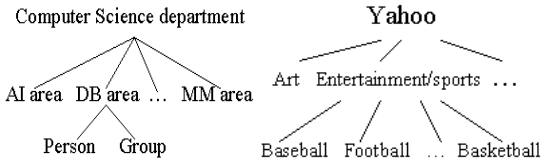

Figure 1. Hierarchical classifiers for (a) computer science department sites and (b) Yahoo site.

6. REFERENCES

[1] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan, Using taxonomy, discriminants, and signatures for navigating in text databases. In Proceedings of the 23rd VLDB Conference, Athens, 1997.

[2] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, 1997.

[3] S. Paek, C. L. Sable, V. Hatzivassiloglou, A. Jaimes, B. H. Schiffman, S.-F. Chang, and K. R. McKeown, Integration of visual and text based approaches for the content labeling and classification of Photographs, ACM SIGIR'99 Workshop on Multimedia Indexing and Retrieval, Berkeley, CA, Aug. 19, 1999.

[4] M.S. Drew, J. Wei, and Z.-N.Li, Illumination-Invariant Color Object Recognition via Compressed Chromaticity Histograms of Normalized Images, Int. Conf. on Comp. Vision, Bombay, pp. 533-540, 1998.