

Functionality-Based Web Image Categorization

Jianying Hu
Avaya Labs Research
233 Mount Airy Road
Basking Ridge, NJ 07920
jianhu@avaya.com

Amit Bagga
Avaya Labs Research
233 Mount Airy Road
Basking Ridge, NJ 07920
bagga@avaya.com

ABSTRACT

The World Wide Web provides an increasingly powerful and popular publication mechanism. Web documents often contain a large number of images serving various different purposes. Identifying the functional categories of these images has important applications including information extraction, web mining, web page summarization and mobile access. In this paper we outline a novel algorithm for automatic identification of two of the most important image categories, namely story and preview images.

Keywords

Web document analysis, image classification, text categorization

1. INTRODUCTION

Many web documents contain a large number of images, and these images tend to be highly heterogeneous in terms of their functionalities. The automatic identification of the functional categories of images contained in web documents is desirable in many web based information processing tasks, including information extraction, web mining, web page summarization and mobile ccess.

In an earlier study [3] we carried out a survey on the functional categories of Web images using data collected from 25 news web sites containing 899 images. Seven major functional categories were found: Story (S), Preview (P), Host (A), Commercial (C), Icons/Logos (I), Headings (H) and Formatting (F).

The task of automatic classification of images into these categories is challenging and requires a combination of sophisticated image understanding and text analysis techniques [3]. Instead of tackling all categories at once, we took a more practical approach, which is to start with the identification of a subset of these categories whose identification has useful applications by itself. We developed an algorithm designed to specifically identify web images belonging to the S (Story) and P (Preview) categories.

A quick study of the image categories reveals that the seven defined categories can be grouped into two *super classes*. The first super class, denoted SPA, includes categories S (Story), P (Preview) and A (Host), and the second one, denoted CIHF, contains the rest of the categories: C (Commercial), I (Icons and Logos), H (Heading) and F (Formmatings). The first super class is more likely to contain photographic images of “regular” aspect ratios, and they are often associated with some story. On the other hand, images in the second super class are more likely to be graphic, have “irregular” aspect ratios (e.g. extremely long or wide), and are often not associated with a story.

Copyright is held by the author/owner(s).
WWW2003, May 20–24, 2003, Budapest, Hungary.
ACM xxx.

Based on this observation, we designed our classification procedure as follows. First, an image is removed if its height is less than 20 pixels or the ratio between the larger dimension and the smaller dimension is greater than 2.5. For the remaining images, the main classifier to separate the SPA and CIHF classes is built using both image features and features extracted from the associated text. Then a secondary classifier using only text features is used to further separate out Host images from the SPA class. The remaining images are considered Story and Preview images.

2. IMAGE FEATURES

The image characteristic that stands out most at the first glance of a web page is whether an image is photographic or graphic. As mentioned above, this feature has strong correlation with the functional categories. We designed a new algorithm incorporating features from both the frequency domain and the color domain.

Frequency domain features were computed using the DCT coefficients. The 8×8 DCT results in 64 coefficients. A subset of these are selected using a discriminative analysis carried out on data extracted from a set of training images. While such DCT features were used directly with success in past efforts to classify an image block as text or non-text, our experiments showed that such a strategy does not work well for photographic and graphic image classification because both categories contain a large range of different image blocks.

To accommodate the large variation within each class, we apply unsupervised clustering on the training image blocks using the M selected DCT coefficients. To be specific, the K-means clustering algorithm [4] is used to group the training image blocks into a pre-determined number of K clusters. Each training image block is then labeled by its cluster index. Finally a normalized cluster histogram is computed for each image, yielding a K dimensional feature. Parameters M and K are chosen empirically and we settled on $M = 18$ and $K = 15$ in our experiments. For classification, each image block is assigned to the cluster with nearest cluster center and the same K dimensional cluster histogram is computed and used as the feature representing the whole image.

Swain *et al.* proposed 8 color related features to distinguish graphic and photographic images [2]. We selected 2 of the color features that are completely independent from the frequency domain features and thus add most discriminative power. The two features are the “band difference” feature, which is a rough measure of the degree of saturation in the image, and the “most common colors” feature, which is a rough measure of the degree of color concentration.

A two stage approach was used to combine the DCT and color features. First a frequency domain classifier is trained using the 18 DCT features. The same classifier is then applied to both train-

ing and testing images, giving a classification score for each image. This single score is then used as the frequency domain feature, which is concatenated with the 2 color features to form a 3 dimensional image feature. The photographic/graphic image classifier is then trained using these 3 features.

This approach is tested on 462 images collected from the web sites listed in [3], which includes 232 photographic images and 230 graphic images. The data set is divided into five roughly equal parts, each containing roughly equal numbers of graphic and photographic images. A Support Vector Machine (SVM) classifier was then trained on each four of the five parts and tested on the remaining part. The five-fold validation method was employed to arrive at the overall accuracy of the classification algorithm. For SVM classifier training and testing, we used the SVM^{light} system implemented by Thorsten Joachims [5] and tested both the linear kernel and the Radial Basis Function (RBF) kernels. The RBF kernels performed better than the linear kernel and achieved an accuracy of 92.5 %.

3. TEXT FEATURES

Images on the web are almost always accompanied by text and such text often contains useful information about the nature and content of the images. Much research has been carried out in the past on using the associated text for image searching and indexing on the web [2]. For that particular task, it was found that the most relevant text fields are: image file names, image captions and the `<alt>` tag in HTML. For functional image categorization, where the goal is not to search for a particular image, but rather to classify any given image into one of several broad functional categories, the text fields mentioned above are too specific. Instead, the surrounding text of an image (text found in the immediate neighborhood of the image) plays a much more important role in identifying its functionality.

For each image, we extracted text nodes in the neighborhood of the image node in the DOM tree, within a maximum of 2 levels. A maximum of 20 words each are extracted for “before text” (from text nodes to the left of the image node) and “after text” (from text nodes to the right of the image node). Structural features such as node boundary and whether each node is a hyperlink are preserved during extraction.

The following feature values are computed over the set of text nodes:

Hyperlink Count This is simply a count of the number of nodes that are hyperlinks. Images in class H (Heading) are likely to have larger values for this feature.

Number Count is a count of the number of all numeric words in the nodes. Images in class C are likely to have larger values for this feature.

Caps Count is a count of the number of capitalized words (excluding the first word) present in the text nodes. Images in the SPA superclass are likely to have higher values for this feature as their contexts usually contain proper names.

Non-dictionary Word Count This feature computes the number of words in the text nodes that do not belong to a dictionary. It is complementary to caps count feature. The dictionary used is WordNet [6], an on-line lexical database.

Maximum Words Count This feature computes the maximum number of words in any of the text nodes. Since SPA superclass images are likely to be accompanied by descriptions, the value of this feature for the superclass will likely be high.

4. COMBINING FEATURES

The image and text features were combined and a Support Vector Machine classifier using SVM^{light} [5] was trained and tested on the training and testing sets described in Section 5. The best results were obtained using the linear kernel of SVM^{light} .

Once the SPA superclass is identified, we used the following rules to separate out the host images (A):

1. For each text node corresponding to an image, identify the proper names (using BBN’s *IdentiFinder*[1]) in the node (if any) and then compute the percentage of the words in the node that belong to the person proper names (for example, in “Larry King Live,” 66.67% of the words belong to the name “Larry King”). Determine the maximum value for all text nodes corresponding to an image. If this value is greater than 50%, then proceed to the second rule. Otherwise, the image is not a host image.
2. For the text node which contains the maximum percentage value, check if the node is also a hyperlink. If so, then identify the current image as a host; otherwise, it is not.

5. EXPERIMENTS

The data collected from our previous study [3] consists of 899 images. To increase the training and testing sets, we collected a second set of front pages from the same sites but a different date. The new set consists of 960 images. The resulting set of 1859 images was subjected to the simple size screening test described in Section 1. After the screening, the resulting set consisted of 462 images.

The set of 462 images was then divided into 5 roughly equal parts containing an approximately equal number of graphic and photo images. Four of these parts are used for training while one is used for testing. The five-fold validation method was employed. In other words, the experiments were run five times where, in each run, one of the five parts was designated as a test set with the remaining four acting as training sets.

The precision and recall numbers achieved by the SVM classifier for the SPA super class are 90.5% and 95.4% respectively. After the rule based host image identification and removal, the final precision and recall numbers for the story and preview images are 82.6% precision and 95.3% recall.

6. REFERENCES

- [1] D. Bikel, R. Schwartz, and R. Weischedel. An Algorithm that Learns What’s in a Name. *Machine Learning*, 34:1–3, 1999.
- [2] C. Frankel, M. Swain, and V. Athitsos. Webseer: an image search engine for the world wide web. *University of Chicago Technical Report TR96-14*, 1996.
- [3] J. Hu and A. Bagga. Categorizing images in web documents. In *Proc. SPIE Conference on Document Recognition and Retrieval X*, Santa Clara, US, January 2003.
- [4] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.
- [5] T. Joachims. Making large-scale svm learning practical. In B. Scholkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.
- [6] G. A. Miller. Five Papers on WordNet. Technical Report 43, Cognitive Science Laboratory, Princeton University, July 1993.