Visual Features in Genre Classification of HTML

Ryan Levering
SUNY at Binghamton
ryan.levering@binghamton.edu

Michal Cutler
SUNY at Binghamton
cutler@binghamton.edu

Lei Yu
SUNY at Binghamton
lyu@cs.binghamton.edu

ABTRACT
Automatic genre classification historically has focused on extracting textual features from documents. In this research, we investigate whether visual features of HTML documents can improve the classification of fine grained genres. Three different sets of features were compared on a genre classification task in the e-commerce domain - one with just textual features, one with HTML features added, and a third with additional visual features. Our experiments show that adding HTML and visual features provides much better classification than textual features alone.

Categories and Subject Descriptors
I.7.2 [**Document and Text Processing**]: Document Preparation – *hypertext/hypermedia, index generation*

General Terms
Experimentation

1. INTRODUCTION
Only recently have commercial search engines focused on goal-oriented interpretations of user's information need. Identifying the specific functions (or genres) of Web documents would enable search engines to provide users with results that attempt to match the user's query with the genre of the document the user needs.

The concept of "genre" has been around in literature for many years. Research from information science, like [2], attempted to define "genre" in its Web context by empirically evaluating the genre of random Web pages. For a larger study on user's perceptions of genres, their usefulness, and work on defining a more complete genre palette, see [7].

The field of genre classification historically has focused on extracting textual features from documents. However, in addition to text, Web pages also possess a URL address, images, other media, and extensive layout information. The goal of this research was to investigate whether visual layout features of HTML pages improve the classification of fine grained genres.

In this research, we use several methods to construct visual features from a rendered HTML document, using both the *position* and the *area* of more traditional features (links, text, emphasis tags, etc.), as well as those of other non-textual content. [6], a survey of the web, found that non-textual content accounts for a majority of many web pages' visual area. Therefore, being able to use information in these additional resources is important.

HT'07, September 10–12, 2007, Manchester, United Kingdom.

2. THE FRAMEWORK
In order to focus on the feature construction, we chose a task where the genres were relatively easy to determine and limited in generality. The theoretical goal was to train classifiers so that spiders that crawl known commercial store sites could recognize some genre of their pages.

We chose three genres for the task: store homepages, store product lists, and store product pages. These are very frequently occurring genres in online stores, with very common user perceptions as to what to expect on one of these pages. One person manually compiled a list of URLs for each genre after given an example of each type of genre. The websites were all popular stores that sold or at least advertised their products online. In addition, the person collected a larger set of URLs that did not fit into one of these genres. These pages serve as a negative class to better train and evaluate a general-purpose classifier.

3. FEATURE CONSTRUCTION
We attempted to include a generous number of commonly used non-visual features to serve as a baseline. For our textual feature set, we chose to use word frequencies, punctuation, and grammar/readability features consistent with [1]. For HTML level features, we included individual tag counts, tag depths, "table" tag depths, several JavaScript statistics, an in-page link analysis, and some form analysis. In addition, we included URL based features such as URL length and a tokenization of the URL.

The commonality in all of the visual features that follow is that they carry some context about the rendered location and size of an object from the Web page. For a rendering engine, we used JRex [5], based on Mozilla's rendering engine.

3.1 Image Counts and Statistics
It is becoming increasingly common to find pages without many words at all. The way image elements are laid out on these pages may be visual cues that viewers use to determine genre. Because the type of image often indicates the image usage (photographic JPEGs versus iconic GIFs), separate image counts were maintained for those types. We also included a standard deviation, minimum, maximum, and mean of all the image sizes.

3.2 Area Based Features
For this set of features, we wanted to measure the general layout of the page to use the information to assist classification. The idea was inspired by optical scanning segmentation [3].

By projecting different content type areas (text, image, form, and object) to a single dimension, we can get information about how the visual elements are spread along both the X and Y dimensions and thus how the content focus of the document changes. Thus if the genre was the type of page that often had a single image at the top followed by a heavy text passage, the top would have a strong image projection and weak text projection and the ratio would reverse as we came closer to the bottom.

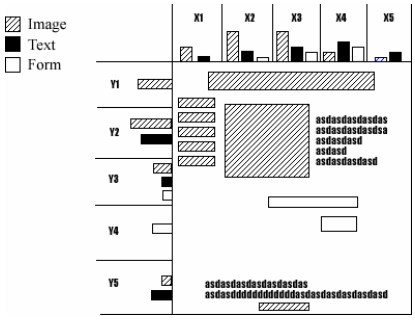

Figure 1: Projecting content types into buckets

Each page was broken into five proportional segments in each dimension as illustrated in Figure 1. The area of each content type within each segment was extracted as a feature to get an idea of the layout flow across the page.

3.3 Visually Central Features

Using the position of text in a page to increase its importance was first shown by [4]. This idea was applied by taking all the basic features and where it was possible and reasonable, extracting the location and area information for only those that occurred within the central area of the page. This includes link counts, form elements, image statistics, and common vocabulary terms. This has the main benefit of removing the noisy header/sidebar/footer information from the analysis of the page.

4. THE EXPERIMENTS

The main goal of our experiments was to look into how different combinations of feature sets performed in our classification scenario. At the same time, we wanted to make sure that our experiment performed well enough to be useful.

4.1 Feature Ranking

We first used information gain (a measurement of the change in entropy when a feature is used to divide data) to pre-select features for classification as well as using the results for evaluation. Features with high information gain most likely represent features with importance in determining a page genre.

Examining the ranked features, we found that visual features are important at some level to all of the classification tasks. While the features weren't usually the very top features, they definitely play a strong role in characterizing the data space. In addition, the Homepages ranking was especially interesting in that with the addition of the negative class, text features do not show much worth and are not in the top twenty-five features at all. This probably indicates negative documents hurting the discriminating power of words found in store homepages. Intuitively, it is hard to think of a word that should appear on store homepages that will not appear often in the rest of the site (like "buy/shop").

4.2 Realistic Classification Experiment

In this experiment, we are interested in testing the ability of our different feature sets to discriminate a particular genre of document in the presence of a large, noisy negative class. We build binary classifiers for each genre. A binary classifier alone would be useful for detecting a particular genre. In a system that needs to assign a single genre to each document, some form of voting scheme is necessary.

For the classification, we first applied the information gain feature ranking discussed above to choose the top 100 features. After feature selection, we used a linear SVM to perform unbalanced classification with 10-fold cross validation. The binary classifiers were then combined via probabilistic consensus to create a multi-class confusion matrix, which was used for overall accuracy.

The results show improvement as more complex features are added. The greatest gains as a whole occur when HTML features are added, bringing the overall accuracy from 73.5% to 87.6%, mainly due to the ability to distinguish the Homepages genre. For our particular experiment, there was much consistency in tokens used in URL strings, such as "product" or "category". Adding the visual features raises the accuracy to 90.1%, this time based on gains in the Product List classification. This is potentially due to the usefulness of average central JPEG size, which is a highly ranked feature for this genre

5. CONCLUSION

Genre classification can be useful in web search, as well as other IR tasks. On the Web, genres are particularly visually-oriented. This paper deals with the challenging problem of using visual features for improving genre classification. We have demonstrated that the visual features are beneficial for all our genre classification problems.

One important area of future work includes applying these features to different, broader genre palettes. In addition, we believe a more detailed examination of potential visual features and their creation will yield better genre-characterizing features.

6. REFERENCES

[1] Boese, E. 2005. Stereotyping the Web: Genre Classification of Web Documents. Masters thesis, Dept. of Computer Science, Colorado State University, Boulder, Colorado.

[2] Crowston, K. and Williams, M 1997. Reproduced and emergent genres of communication on the World-Wide Web. In *Proceedings of the 30th Hawaii International Conference on System Sciences*, 30. Washington, DC: IEEE Computer Society.

[3] Ha, J., Haralick, R.M., Phillips, I.T. Recursive X-Y cut using bounding boxes of connected components. In Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 2), p.952, August 14-15, 1995.

[4] Kovacevic, M., Diligenti, M., Gori, M. and Milutinovic, V. 2002. Recognition of Common Areas in a Web Page Using Visual Information: a possible application in a page classification. In *Proceedings of 2002 IEEE International Conference on Data Mining*, 250. Washington, DC: IEEE Computer Society.

[5] JRex 2007. http://jrex.mozdev.org/.

[6] Levering, R. and Cutler, M. 2006. The portrait of a common HTML web page. In *Proceedings of the 2006 ACM symposium on Document engineering*, 198-204. New York, NY:ACM Press.

[7] Rosso, M.A. 2005. Using Genre to Improve Web Search. Ph.D. diss., School of Information and Library Science, University of North Carolina, Chapel Hill, North Carolina.