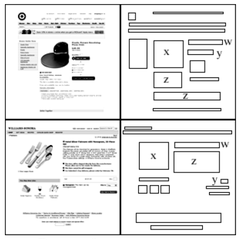
Using visual features in genre classification of HTML

Ryan Levering, Michal Cutler and Lei Yu
Department of Computer Science, Binghamton University, Binghamton, NY

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. For information on genre on the web, see [1][2][7].

The goal of this research was to investigate whether visual layout features of HTML pages improve the classification of fine grained genres. We do not focus on whether they will always be useful to genre classification, but rather on whether they can be and how these visual patterns can be translated into features for classification.

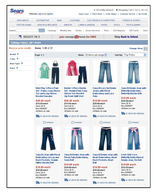


Pages of a genre appear to show consistency in layout

In this research, we use several methods to construct layout 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. Being able to use information in these increasingly important [6] resources is important.

Framework For Evaluation

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.



?

Product Description
Product List
Store Homepage
None of the Above

Feature Construction

We used three feature sets for the classification task in order to compare their relative worth. For an HTML browser, we used JRex [5], based on Mozilla's rendering engine.

Feature Sets

- **Pure Text – Only based on the textual contents of the page**
- **+HTML – Added URL and DOM-based statistics**
- **+Visual – Added features discussed below**

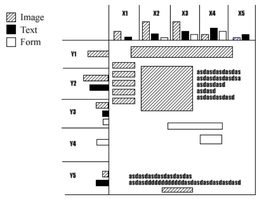
Image Counts and Statistics

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

Large JPEG can indicate product description

Area Based Features

For this set of features, we wanted to measure the general layout of the page. The idea was inspired by optical scanning segmentation methods [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.



Projecting content types to dimensional buckets

Each page was broken into five columns X1 to X5 and five rows Y1 to Y5, as seen to the left. The area of each content type within each column/row was extracted as a feature to get an idea of the layout flow across the page.

Visually Central Features

By focusing on the center, we lose header noise

Using the position of text in a page to increase its importance was shown in [4]. This idea was applied by extracting the location and area information for only features 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.

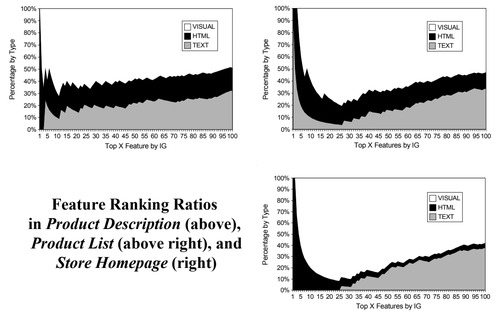
Experimental Evaluation

The main goal of our experiments was to look into how different combinations of feature sets performed in our classification scenario.

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 text features do not show much worth and are not in the top twenty-five features at all. 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 (e.g. "buy/shop").

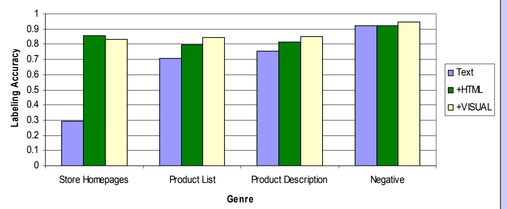


Feature Ranking Ratios in *Product Description* (above), *Product List* (above right), and *Store Homepage* (right)

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.

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.



Genre classification accuracy with varying feature sets

Conclusions

Genre classification can be useful in web search, as well as other IR tasks. On the Web, some 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.

The more important question of how important these features are to genre classification in general depend on how genre classification is used on the Web. We hypothesize that for broader genre classification, like determining whether a page is commercially-driven or not, these features will not be as useful beyond filtering noise. However, we can imagine many search tasks with transactional type goals that can be met with pages that have some visual archetype that translate into a genre.

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.

Literature cited

[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.

Acknowledgments

I would like to thank my wife, Terrill Levering for contributing large amounts of time in HTML sample collection. In addition, the Computer Science department at Binghamton University provides the assistantships that let me work unfunded.

For further information

Please contact *ryan.levering@binghamton.edu*. More information on this and related projects can be obtained at http://webseer.sourceforge.net. A link to a digital version of this poster can be found at http://webseer.sourceforge.net/ht07-poster.pdf and the paper version can be found at http://webseer.sourceforge.net/ht07-paper.pdf.

BINGHAMTON UNIVERSITY
State University of New York