Web Page Classification Using Image Analysis Features

Viktor de Boer¹, Maarten W. van Someren¹, and Tiberiu Lupascu²

¹ Informatics Institute, Universiteit van Amsterdam
Science Park 107, 1098 XG Amsterdam, The Netherlands
² EURO IT&C B.V., Haarlem, The Netherlands
{v.deboer,m.w.vansomeren}@uva.nl, tiberiu.lupascu@euroitc.com

Abstract. Classification of web pages is usually done by extracting the textual content of the page and/or by extracting structural features from the HTML. In this work, we present a different approach, where we use the visual appearance of web pages for their classification. We extract generic, low-level visual features directly from the page as it is rendered by a web browser. The visual features used in this document are simple color and edge histograms, Gabor and texture features. These were extracted using an off-the-shelf visual feature extraction method. In three experiments, we classify web pages based on their aesthetic value, their recency and the type of website. Results show that these simple, global visual features already produce good classification results. We also introduce an online tool that uses the trained classifiers to assess new web pages.

Keywords: Web design, Computer vision, Image analysis, Machine learning.

1 Introduction

The ”Look and Feel” or is an important property of a website. Most research and development that is aimed at analysis of websites focuses on the content, in particular on the words and their meanings. In addition to the content the form of a webpage is used to convey or even to induce emotional aspects. Individuals and organizations attach much importance to the image that is created by their presence on the web. A bank should appear reliable, an artist creative, an IT company technically advanced and user friendly. Designers use their creativity to find a form that conveys and evokes emotion, trust, authority or a range of impressions like creativity, innovation, political or environmental awareness, religious background, etc. etc. In this paper we describe preliminary experiments with several dimensions of Look and Feel.

Look and feel can take many forms, as can easily be seen by reviewing for example home pages of persons and organizations. In the design of a website the visual appearance or look and feel is constructed by colors and color combinations, type fonts, images and videos, dimensions of page layout such as contrasts. Look and Feel is produced by designers in an intuitive way, using design tools that enable manipulation of visual elements.

Our goal is to enable automatic analysis of this visual appearance of web pages. This goal is part of a wider effort to achieve automated analysis of websites. In earlier
studies methods and tools were developed that analyze websites by their content, in particular their vocabulary and structure. The practical goal of this is to develop a tool that supports the design of web-based information systems by constructing a first draft of the information architecture or by critiquing a first draft. This is done by modeling a given collection of sites and comparing the model with the draft. The first version of the tool only considered the content, the way in which this is organized over pages and the hyperlinks between pages [3].

Our approach is based on using the page as it appears to the user. Analysis of documents on the web is normally based on data that are extracted from the HTML. This is the approach that is typically followed for analyzing the content of web pages. The HTML is removed and the natural language words are used as properties of the page and used for classification or extracting information [5,9, e.g.]. For analysis of Look and Feel this approach seems hardly feasible because Look and Feel elements are difficult or impossible to identify in the HTML code. Some systems allow selection of Look and Feel elements in the form of possibilities for color schemes or the shape and layout of menus, buttons, etc. but many designers construct their own layout, colors and style for objects. For this reason we decided to use low level features of a page, taken as an image. This makes it independent of how the page is produced and analyzes it directly in terms of how the user sees it.

In [1], the authors describe a study into the perceived quality of web sites. The results show that the number of images on a web site is one of the five features that has the highest correlation with high quality web sites as perceived by users. Research into the perceived quality of web pages has shown that the visual appearance of web pages is also important for the perceived credibility [7].

In [11] the author describes the AQUAINT system, a quality based search engine. The system uses 113 features to describe web pages, which are extracted at runtime. Among these features are also color features: notably the number of colors, the number of unique colors, the RGB values of most frequent color, the text color and the background color. Other visual features include the number of graphics on a page, the number of links to graphics, the relation between the number of graphics and the file size. These features are at least partially derived from the underlying HTML. In our approach, we extract the visual features from the pages, as rendered by a web browser. These visual features were combined with other features (e.g. textual content) to train a classifier that distinguishes between high and low quality web pages. The relation between the number of graphics and the file size was among the most important distinguishing features.

Below we summarize our method for training classifiers, the evaluation procedure and the results of experiments with visual attributes of web pages.

2 Visual Features

We use the Firefox web browser to render an image for a web page. Of each page, we save a screen shot using the Fireshot plugin[^1] for the Firefox web browser. These screen shots are stored as .PNG files.

[^1]: http://screenshot-program.com/fireshot/