Website categorization using Visual Features

Fábio Costa
Hossein Kazemi
Luís Brandão

July 11, 2010

Abstract

Web content classification is an important and a challenging task. Two of the major challenges associated with it are the subjectivity of the evaluation procedures and the type of features that should be used to classify a web page. In this paper, motivated by the Discovery Challenge 2010 [1], we address the later issue by evaluating how visual features can be used to automatically classify web pages in a large multi-label dataset. We extract RGB Histograms, Edge Histograms and Tamura features from web pages and analyze the results of using these features independently and combined with textual features in the task of classifying web pages. Our results show that while the classification results using visual features are lower than when using textual features, the combination of both feature types produces promising results.

1 Introduction

Automatic web content analysis is a growing issue. While users can usually easily identify the subject matter of a web site, the automatic classification of web pages is still a problem under active research. Early research on this area was mainly aimed to detect Spam web pages or web pages with violent content for filtering purposes. In recent years this research has spread to a larger number of categories. The motivation for this expansion comes from several areas. Information retrieval results can be improved, the quality of a web site can be measured if a utility value is associated with category, and different types of web sites can be more accurately detected for archiving purposes, just to name a few applications.

One of the major challenges of automatic web content analysis is the choice of features for the classification process. The textual content of web pages is usually used in this task. In addition, link analysis algorithms or visual features are also used to improve the classification performance.

From a user's point of view, the visual appearance of a website often plays a central in determining its type. This is due to the fact that different web design rules and trends are usually taken into account to help the visitors to identify the subject of a web site. An example of how appearance varies between different types of websites is the common layout of newspaper's websites and web blogs. While both can be viewed as news sources, there is usually a higher authority value associated to newspapers. When measuring the quality of a news website, for instance, it may be useful to determine if it is a newspaper website or a blog. While the Textual Content of both newspaper and blog web sites can be significantly similar, since the topics that both types of sites address may be related, the visual layout is usually quite different. By using this kind of features in situations like this, the classification performance can be improved.

In this paper, we address the problem of classifying web pages using both visual features and textual features. We compare the use of visual features, textual features and the combination of both in the classification of sample of a recent crawl of the .EU domain.

This paper is organized as follows. First, relevant work in the field of web pages classification using visual features is discussed. Then our motivation and the research questions are introduced. Section 4 discusses the dataset we used. Section 5 introduces our experiment,results and analysis of the results.At the end, in Section 6 we draw our conclusions.

2 Related Works

This work is a follow-up of the research conducted by Maarten van Someren and Viktor de Boer [2]. The goal of their research work is to enable automatic analysis of visual appearance of web pages. In their experiments, they start by saving a screenshot of each web page and then extracting the following visual features:

- Simple Color Histogram
- Edge Histogram
- Tamura
- Gabor features

They use a very limited amount of training data and Naive Bayes and a Decision Tree learner (J48) algorithms to classify the web sites. The experimentation is done based on aesthetic value, recency and website topic. We only quote their best results here. They report an accuracy of 83% using the top 5 attributes with J48 method in aesthetic value experimentation. On recency experimentation, they report a 93% accuracy using 10 fold cross-validation over the top 10 features using J48. Using the complete feature vector in their experimentation on Topics, they report an accuracy of 56% with J48. In all the experiments, Naive Bayes method was slightly less accurate. Moreover, based on their experiments, they showed that low level features of web pages are able to distinguish between several classes that vary in their look and feel.

Another work which attempts to develop an automatic system for webpage aesthetic evaluation is by S. Amirhassan Monadjemi et al. [3]. They use visual features extracted with image processing techniques, such as Color spaces and Gabor filters and use Artificial Neural Networks (ANN). They have tested their method using university websites and they report an accuracy of 90%. They have also shown that ANN as a classifier can find the hidden aesthetic patterns behind the input features extracted from the texture and color of a web page. Their work is also based on a very limited amount of training data.

3 Research Questions

While visual features have been used in the automatic classification of web pages, both alone and mixed with other feature types, we want to evaluate how the models proposed by van Someren and de Boer [2] perform in a large multi-label data set. This gives rise to the research questions that we want to address in this paper:

- Are visual features, namely the ones built upon Simple Color Histograms, Edge Histograms and Tamura, suitable for website classification?
- Do visual features outperform textual features in the task of classifying web pages?
- Should visual features be combined with textual features in web page classification?

4 Experiment Description

4.1 Dataset Used

We use a subset of the data provided by ECML PKDD 2010 [6]. The base data is a crawl of 23M pages from hosts in the .EU domain downloaded between February and March 2010 by the European Archive Foundation.

The training data provided by the Discovery Challenge is composed of 1879 multi-labeled English webpages, for which the Term-Frequency (TF) and Document-Frequency (DF) of the top 50,000 words were calculated and provided. Based on this training set we defined our own data set with the two types of features we use in our experiments - visual and textual features. TF and DF were used to re-generate the partial text (because stop words were deleted beforehand) of each web page. Based on these text, the textual features were extracted and a binary feature vector was made for each webpage.

To compute the Visual Features we had to create a screenshot of the web page that was not available in the competition data. These screenshots were later used to extract the following visual features:

- RGB Histogram - default RGB color histogram with 32 bins
- Edge Histogram - spatial distribution of 5 types of edges four-directional edges and one non-directional
- Tamura Features - texture features that correspond to human visual perception

Table 1 summarizes the dataset used.

Table 1: Dataset Statistics

	Text		Images	
Class Type	In-Class	Out-Class	In-Class	Out-Class
News-Editorial	50	1449	60	1672
Commercial	624	785	624	1108
Discussion	64	1435	75	1657
Educational-Research	584	915	593	1139
Web Spam	55	1499	47	1685
Personal-Leisure	1172	372	328	1404
Neutrality	5	29 & 1385	4	24 & 1149
Bias	614	116	587	107
Trustiness	4	12 & 1299	3	9 & 1170

4.2 Classification Tasks

To answer the research questions enumerated on Section 3, we defined three types of experiments:

1. Classification of the data set using visual features
2. Classification of the data set using textual features
3. Classification of the data set using textual and visual features

As the main focus of our report is on visual features, we decided to used 3 different classifiers for the visual features and 2 different classifiers for textual features. For visual features we used Naive Bayes(NB), Support Vector Machines(SVM) and Decision Tree (J48). For the textual features we only used Naive Bayes and Support Vector Machines.

The last experiment (experiment 3) was done by using an ensemble of learners. The type of ensemble used combines the results of the experiment types 1 and 2 using different weights as follows. Considering the predictions of a classifier used for visual features P_{Vis} and the predictions of a classifier used for textual features P_{Tex}, the predictions of these two classifiers are combined, with weights W_{Vis} and W_{Tex}, as follows:

$$P_{Ensemble} = P_{Vis} * W_{Vis} + P_{Tex} * W_{Tex}, \;\; with \;\; W_{Vis} + W_{Tex} = 1 \tag{1}$$

Table 2 summarizes the experiments made and the research questions that each addresses.

Table 2: Experimental Configuration

Experiment	Classifier	Feature	Research Question(s) Addressed
1	NB	Visual	1. Are visual features, namely the ones built upon Simple Color Histograms, Edge Histograms and Tamura, suitable for website classification? 2. Do visual features outperform textual features in the task of classifying web pages?
	SVM	Visual	
	DT	Visual	
2.0	NB	Textual	
	SVM	Textual	
2.1	SVM	Textual - Top 100 Features selected with Chi-square	
3	Ensemble: Best learners of exp. 1 and 2	Best Visual and Best Textual	3. Should visual features be combined with textual features in web page classification?

Choosing 3 different classifiers made our progress slow as we had to tune their parameters to get the best results, however, this decision enabled us to observe the behavior of each classifier on the dataset and gave us a better intuition of which classifier suits our task.

We classify our data as belonging to one or more of the following classes:

- Web Spam
- News/Editorial
- Commercial
- Educational/Research
- Discussion
- Personal/Leisure

These categories are considered independent and the multi-label classification problem was transformed into single-label classification problems for each category.

In addition to the categories above, each website is classified by the level of Neutrality, Bias and Trustiness as follows:

- Neutrality - from 3 (normal) to 1 (problematic)
- Bias - 1 flags significant problems

- Trustiness - from 3 (normal) to 1(problematic)

These categories were part of the classification task of the Discovery Challenge 2010 [9].

4.3 Feature Selection

We evaluated the effects of feature selection on our classification performance using SVM. In addition to the base data set, the following subsets of data were defined to select the most expressive features for each class:

- For Text using Chi-Square: all feature; top 1000,500,250,150,100,30
- For Visuals using LIRE : all features, RGB Histogram(32 features), Edge Histogram(80 features), Tamura Features(18 Features)

Moreover, our experimentation involves using different visual features for classification. We used DT and NB for each of the visual features separately, mentioned before, and once using all visual features together.

This decision will give us an intuition on which visual features are most indicative of a class label, and if using all the visual features together would boost the classification performance or not.

4.4 Software

We used the LIRE (Lucene Image REtrieval) library [4] to extract the visual features from the images. LIRE is a content based image retrieval library build on top of Lucene [7], that provides an API that enables the extraction of the visual features. The screenshots of the web pages were extracted using webkit2png [8]. For the classification, we used Weka [5] and MATLAB.

4.5 Evaluation

We evaluate our models using 10-fold cross-validation and by measuring Precision and Classification Accuracy. The data used in the experiments was previously labeled by humans to be used on the Discovery Challenge.

5 Results and Discussion

We first focused our experiment to find the best results for our classifiers using textual and visual features. Then we focused our classification task to find the most descriptive visual features. At the end, we chose the best classifiers from our previous results from visual and textual classifications for our ensemble.

Table 2 summarizes the experimental results for our textual and visual classification. We applied Naive Bayes (NB) and Support Vector Machines (SVM) to both visual and textual features sets. In addition, a Decision Tree learner (J48) was also applied to the visual features.

Table 3: Classification Results(A: Accuracy, P: Precision)

	Textual Features						Visual Features					
	NB		SVM		SVM 100		NB		SVM		DT	
Class Type	A	P	A	P	A	P	A	P	A	P	A	P
Commercial	**71**	**73**	69	85	66	63	**54**	**54**	56	37	51	48
Discussion	86	95	**96**	**100**	92	26	65	93	11	4	**96**	**92**
Educational	**79**	**78**	76	76	70	67	49	63	58	39	**61**	**60**
News	90	95	**98**	**100**	92	13	53	94	3	34	**97**	**93**
Personal	76	81	80	53	**80**	**57**	75	77	66	19	**78**	**73**
Spam	93	96	**99**	**100**	99	100	47	95	3	3	**97**	**96**
Bias	80	45	80	80	**82**	**82**	43	68	50	24	**77**	**76**
Neut	20	4	**98**	**98**	98	98	69	98	98	98	**99**	**99**
Trus.	85	86	**99**	**99**	99	99	61	95	100	99	**98**	**96**
Average	76	73	**88**	**88**	86	67	57	82	49	40	**84**	**81**

Our results show that SVM performs better when used for the classification of Textual Features rather than Naive Bayes. Although NB performs better in terms of precision for 'Educational' and 'Personal', overall, SVM outperforms NB for text features. We also observed that using all the textual features gives us better results in overall. However, only for comparison, we included the results of the top 100 textual features in table 2.

We observed that SVM was able to deal with the highly imbalanced data using text features. Using SVM we tried different parameters and weighting to better deal with the imbalanced data. On the other hand, surprisingly, NB was also able to handle the imbalanced classes with very good results. We think that the reason for these good results from both classifiers is that the textual features are highly indicative of the class labels.

For the visual features, surprisingly, SVM performs weakly and Decision Tree works best. Although we tried to tune the parameters (to do some weighting for the highly skewed classes) for SVM, the results turned out to be not satisfactory. With a quick look at the Table 1, it can be seen that 'Discussion', 'News-Editorial', 'Spam', 'Neutrality' and 'Trustiness' are highly imbalanced classes. Out of which, SVM performs very weak for 'Discussion', 'News-Editorial' and 'Spam'. The low numbers of Precision and Accuracy are due to the highly imbalanced classes. However, surprisingly SVM performs much better for 'Neutrality' and 'Trustiness'. This might be due to the fact, even though their classes are highly skewed, the features are highly indicative of the class labels. Therefore, the results are much better.

Overall, for visual features, the Decision Tree learner outperformed NB and SVM. This is actually a promising result, which is in correspondence with Maarten van Someren and Viktor de Boer's results [2] in which Decision Tree outperformed NB in the classification tasks for visual features.

Based on our observations and experiments, in overall, textual features are more indicative of the classes. However, an important fact has to be considered that the webpages (our training data) are manually labeled by human assessors and we speculate that they have judged and labeled the webpages based on their text and not on their visuals. For example, a webpage might actually contain news, but looks like a Spam website with a lot of suspicious ads and links. The classifier for textual features, classifies this page as news and the classifier for visual features might classify this page as Spam. This can be a substantial issue that have affected our classification results for visual features comparing to textual features.

Moreover, some websites were down or required authentication, therefore, for some webpages, we weren't able to take snapshots. This was also a another nontrivial issue besides the imbalanced classes and web page labeling.

Table 4: Classification Results Using Visual Features(A: Accuracy, P: Precision)

	RGB Histogram				Edge Histogram				Tamura				All Features			
	NB		DT		NB		DT		NB		DT		NB		DT	
Classes	A	P	A	P	A	P	A	P	A	P	A	P	A	P	A	P
commercial	65	62	73	63	55	54	52	52	63	67	**73**	**65**	54	54	51	48
discussion	54	94	**97**	**94**	88	95	97	94	8	95	**97**	**94**	65	93	96	92
educational	37	69	73	69	62	68	66	65	6	68	**76**	**71**	49	63	61	60
news	3	96	**98**	**95**	89	96	98	95	86	96	**98**	**95**	53	94	97	93
personal	8	81	86	8	77	79	86	79	74	79	**86**	**79**	75	77	78	73
spam	3	97	**98**	**96**	81	97	**98**	**96**	95	96	**98**	**96**	47	95	**97**	**96**
bias	63	67	65	39	67	80	**79**	**79**	79	77	**79**	**78**	43	68	77	76
trustiness	52	98	99	98	98	98	99	98	95	98	**99**	**98**	69	98	99	99
neutrality	27	94	98	96	87	96	98	96	82	95	**98**	**96**	61	95	98	96
Average	49	83	87	84	75	83	86	84	79	85	89	86	58	85	88	87

Table 4 shows the results of our experiments using different visual features separately and all together. Our aim was to observe which visual features are the most indicative and whether or not using all the visual features together will give us any advantages in classification performance or not.

We observed that using DT with Tamura features gives us the best results, 89% accuracy and 86% precision in average. Moreover both NB and DT were, in general, able to deal with the highly imbalanced data, except certain situations using NB on RGB features where there are categories with high precision while having very low accuracy. This happens because these categories are highly imbalanced and the largest class has more weight and highest precision, which leads to a final high precision while having a low accuracy.

Table 5 shows in more detail the results of the two independent learners combined in the Ensemble and how their results relate with the Ensemble results.

Table 5: Classification Results Using Visual and Textual Features

Classes	Text Weight	Visual Weight	Accuracy	Precision	Correlation
Commercial	0.5	0.5	82	83	0.815
Discussion	0.5	0.5	99	99	1
Educational	0.8	0.2	83	85	0.69
News	0.5	0.5	90	99	1
Personal	0.2	0.8	93	93	-0.981
Spam	0.4	0.6	99	99	-1
Bias	0	1	83	1	-0.60
Neutrality	0.5	0.5	96	99	1
Trustiness	0.5	0.5	98	96	1
Average	0.43	0.57	92	95	0.89

The last column of the table shows the correlation coefficient between the predictions of the two classifiers of the Ensemble - 0 correlation means that the predictions are not correlated, a positive correlation coefficient between 0 and 1 means that the predictions are positively correlated, and a negative correlation coefficient between 0 and -1 means that the predictions are negatively correlated. The intuition behind this measure was to evaluate how the predictions of both classifiers were related. While the correlation coefficient values are high, the results of the ensemble are in average good. This goes against the initial idea that each classifier was exploring different areas of the classification task. Even though the results of the ensemble outperform the results from the learners used alone, which indicates that the predictions of the classifiers do not fully overlap.

6 Conclusion

In this paper two kinds of features were used to classify web pages. Furthermore, different learning algorithms were applied to these feature sets to evaluate their behavior.

Our main challenge was to deal with the highly imbalanced data. We also used different visual features separately to find the most expressive features.

Our results shoed that for visual features DT performs best and between the visual features Tamura features are the most expressive. Also our observation shows that, even though classification of documents based on textual features is still superior to classification using visual features, better results can be achieved if visual features are combined with textual features. We also speculated that the labeling of the webpages has been done based on the textual features, and we believe if training data has been labeled based on their visuals, we could have had better results.

References

[1] European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2010. `http://www.ecmlpkdd2010.org/`

[2] V. de Boer, M.W. van Someren and T. Lupascu. Classifying Web Pages with Visual Features. *Proceedings of the 6th International Conference on Web Information Systems (WEBIST 2010)*, Valencia, Spain

[3] M. Mirdehghani, S.A. Monadjemi, Web Pages Aesthetic Evaluation Using Low-Level Visual Features, *World Academy of Science, Engineering and Technology - 49*, 2009

[4] LIRE library, `http://www.semanticmetadata.net/lire/`

[5] Weka Toolkit `http://www.cs.waikato.ac.nz/ml/weka/`

[6] A. Benczúr, C. Castillo, M. Erdélyi, J. Masanes, M. Matthews, Z. Gyöngyi. ECML/PKDD 2010 Discovery Challenge Data Set. *Crawled by the European Archive Foundation.*

[7] Apache Lucene, `http://lucene.apache.org/`

[8] webkit2png, `http://www.paulhammond.org/webkit2png/`

[9] Discovery Challenge 2010 Tasks, `http://datamining.sztaki.hu/?q=en/DiscoveryChallenge/rules`