# Towards Language–Independent Web Genre Detection

Philipp Scholl, Renato Domínguez García, Doreen Böhnstedt,
Christoph Rensing, Ralf Steinmetz
Multimedia Communications Lab, TU Darmstadt
Merckstraße 25
64283 Darmstadt, Germany
{scholl, renato, boehnstedt, rensing, ralf.steinmetz}@kom.tu-darmstadt.de

## ABSTRACT

The term *web genre* denotes the type of a given web resource, in contrast to the topic of its content. In this research, we focus on recognizing the web genres *blog*, *wiki* and *forum*. We present a set of features that exploit the hierarchical structure of the web page's HTML mark-up and thus, in contrast to related approaches, do not depend on a linguistic analysis of the page's content. Our results show that it is possible to achieve a very good accuracy for a fully language independent detection of structured web genres.

## Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods*

## General Terms

Algorithms

## 1. INTRODUCTION AND MOTIVATION

Nowadays the World Wide Web's content is increasing in a scale never seen before. Especially the upcoming of social web applications (e.g. blogs and wikis) that make it easy for people to create User Generated Content has substantially contributed to the huge amount of available information in the Web. Their increasing popularity has given rise to search and analysis engines focusing on the social web, e.g. Google's blog search. A key requirement of such systems is to identify the *genre* of the respective web pages as they crawl the Web. The genre of a web page – as opposed to the topic of its content – comprises its functionality, purpose and conventions of usage, content creation, authorship and reception, thus defines a typical appearance and way of providing content with similar web pages. For example, web pages enable displaying a product (e.g. the web genre *e–shop*), representing a person in a social or organizational context (*personal* or *academic homepage*) or allowing to follow and take part in discussions in a *forum*.

In a user study, Meyer zu Eissen et al. [2] show the importance of web genres for the expectations of users. They, as well as Santini [4], employ different machine learning algorithms in order to automatically detect web pages' genres between different genre classes like *discussion*, *personal homepages*, etc. . Both apply different feature types like linguistic features (e.g. part-of-speech tagging and document terms), structural features (e.g. HTML tag frequencies, use of facets used to enable functionalities like form input elements) and simple text statistics (e.g. frequencies of punctuation).

However, a fact often neglected by related work is that the absolute dominance of the English language on the web is decreasing. Thus, it is important to develop a way of recognizing web genres independently of the language used on the respective web page. As many genres exhibit a certain structural and visual layout, this property enables to ignore linguistic features altogether.


*WWW 2009,* April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

## 2. OUR GENRE CLASSES AND FEATURES

In this research we restrict ourselves to the challenge of recognizing the web genres *wiki*, *forum* and *blog*, as they are content-creation backbones of online communities, widely adopted and used, and support different paradigms of content creation and collaboration in different languages. Therefore, use of language in our scenario is not a discerning feature and should be neglected in favour of a truly language agnostic approach.

Instead, we apply conventional features that are language-independent and are commonly used by related approaches: HTML tag frequencies, *layout*, *functional* and *typographical facets* as described in [4], content word count, punctuation frequencies, URL properties, text / mark-up ratios and CSS rule counts.

Further, we propose some novel features that base on the logical structure of a web genre. Specific genres often exhibit typical content structures (e.g. a blog basically consists of blog posts and any number of comments to each of this blog posts). As can be easily perceived by a human, this structure is mirrored in the visual layout of the content blocks – and in the mark-up structure rendered by the application's underlying template engine. For example, all comments on a blog post page share a common structure, only the user generated content (i.e. the text entered by a user) contained in this structure varies (e.g. multiple paragraphs, additional links or images etc. ). Thus, some similar mark-up is repeated. We call this repeated mark-up structure a *structural pattern*.

Based on these structural patterns, we compute features that represent properties of a web resource that relate to the number, size, hierarchy, structure, in-page location and content ratio of identified patterns. The pattern extraction method is based on [3], a computationally affordable approach to measure (sub)-tree similarity. It recursively walks the HTML Document Object Model, abstracting each node

and its sub-tree into a representation that only takes into account the sub-tree structure by ignoring textual content. Elements that are less likely to be a structural part of a pattern are ignored (like inline elements) or contracted (like paragraphs). The structure representations that share a common parent node are compared with each other, and if their similarity exceeds a certain threshold, they are considered to be pattern candidates. Finally, all pattern candidates that do not fulfil requirements like e.g. a certain level of complexity are discarded. The remaining patterns are taken as a basis for computation of the pattern features mentioned above. In [1], the pattern extraction algorithm and the features are explained in more detail.

# 3. EVALUATION AND RESULTS

After a preliminary analysis of our focused web genres, we saw the need to split the *blog* and the *forum* genre corpora into sub-genres in order to reflect the structural diversity within the different web page types in the web genres themselves. For example, the respective start pages that serve to give an overview of all contained blog posts or forum threads differ structurally from the pages that present the content (in this case the blog post pages and the forum thread pages).

There is no corpus with our genre and multi-language requirements available, so we compiled a corpus containing example instances for machine learning by classified examples and to validate the selected features. From this corpus we randomly selected 200 sample instances per (sub-)genre, getting a corpus containing 1000 multi-lingual instances (of which 65% are English, 7% German, 7% French and the rest in about twenty different European and Asian languages) in five different genres or sub-genres. Further, we took great care to include different applications per genre, e.g. for wikis we sampled pages from wiki engines like *MediaWiki*, *MoinMoin* and many others.

For our evaluation, we applied Support Vector Machines with Sequential Minimal Optimization (SMO) for classification. All classification results were subjected to ten-fold cross validation.

**Table 1: Confusion Matrix for classification using all features with SMO**

| a | b | c | d | e | ← classified as |
|---|---|---|---|---|---|
| 181 | 10 | 4 | 4 | 1 | a = Blog_Page |
| 21 | 175 | 1 | 3 | 1 | b = Blog_Post |
| 10 | 1 | 184 | 3 | 1 | c = Wiki_Page |
| 9 | 0 | 2 | 184 | 5 | d = Forum_Start |
| 4 | 4 | 1 | 7 | 184 | e = Forum_Thread |
| 0.80 | 0.92 | 0.96 | 0.92 | 0.96 | Precision |
| 0.91 | 0.87 | 0.93 | 0.92 | 0.92 | Recall |
| 0.85 | 0.89 | 0.94 | 0.92 | 0.94 | F–Measure |
| | | | | 90.8% | Accuracy |

Using all features, we achieved 90.8% accuracy (i.e. correctly classified instances) in our results. From the result's confusion matrix (see Table 1), one can see that a major source of incorrect classification is the distinction between blog start pages (here labeled as class `Blog_Page`) and blog post pages (`Blog_Post`). As these are affiliated with the same superordinate genre *blog* (same with `Forum_Start` and `Forum_Thread`), we may integrate these results if we are not interested in detecting the exact sub-genre, getting an overall accuracy of 95.1%.

Ranking all features by *information gain* shows that among the 20 most important features are HTML tag frequencies, syntactic URL analysis, link analysis, HTML facets and two of the pattern ratio features. This means that the ratio of how much content of a web resource is contained in recurring patterns is significant to the web genre of this resource.

# 4. CONCLUSIONS

In conclusion, it is possible to achieve a good accuracy for detection of a web genre like *blog*, *wiki* and *forum* and their respecting sub-genres taking into account traditional features and pattern features. The latter base on the reoccurring HTML mark-up structures and do not demand knowledge of the language of the HTML's content. Thus our approach is fully language independent. Further, it works with different systems and applications.

We obtained reasonable results with only a small set of 144 features. Other approaches – particularly those making use of linguistic analysis – often have several thousand features. Thus the limited number of features in our approach reduces the computational complexity of the actual classification task.

Further research will include the – in other web genre detection research often neglected – issue of classifying outliers as well, i.e. detecting if a web page belongs to one of our genres or not. This is vital for a real–world application of this genre detection approach, as we rarely know in advance that our analyzed web resources will only consist of the web genres mentioned here. Preliminary results show that the accuracy with an included outlier-class performs still well with 86% accuracy. Finally, regarding the web genres presented here, we will apply the results gained in our evaluations to the field of Community Mining.

# 5. ACKNOWLEDGEMENTS

The content presented in this paper is a result of the project "Web 2.0 Resources and Artifacts", which was funded by SAP Research. The authors take responsibility for the content.

# 6. REFERENCES

[1] R. Domínguez García, P. Scholl, D. Böhnstedt, C. Rensing, and R. Steinmetz. Automatic web genre classification using structural features. Technical Report KOM–TR–2008–06, Multimedia Kommunikation – TU Darmstadt, Germany, July 2008.

[2] S. Meyer zu Eissen and B. Stein. Genre classification of web pages — user study and feasibility analysis. In *KI 2004: Advances in Artificial Intelligence*, volume 3238 of *LNCS*, pages 256–269. Springer Berlin / Heidelberg, 2004.

[3] D. Rafiei, D. L. Moise, and D. Sun. Finding syntactic similarities between xml documents. In *Proceedings of the Conference on Database and Expert Systems Applications (DEXA'06)*, volume 0, pages 512–516, Los Alamitos, CA, USA, 2006. IEEE Computer Society.

[4] M. Santini. *Automatic Identification of Genre in Web Pages*. PhD thesis, University of Brighton, January 2007.