A TWO-LEVEL APPROACH TO WEB GENRE CLASSIFICATION

Ulli Waltinger, Alexander Mehler, Armin Wegner
Text Technology, Bielefeld University, Universitätsstr. 25, 33615 Bielefeld, Germany
{*ulli_marc.waltinger, alexander.mehler, armin.wegner*}*@uni-bielefeld.de*

Keywords: hypertext types, web genre classification, web structure mining, two-level classifier

Abstract: This paper presents an approach of two-level categorization of web pages. In contrast to related approaches the model additionally explores and categorizes functionally and thematically demarcated segments of the hypertext types to be categorized. By classifying these segments conclusions can be drawn about the type of the corresponding compound web document.

1 INTRODUCTION

Hypertext categorization is mainly performed as function learning irrespective of relational structure learning, that is, of genre-related structures internal to single sites and pages. Consequently, approaches to hypertext categorization mostly explore simple text features (Karlgren and Cutting, 1994) (Kessler et al., 1997) or additionally include structural features (Lee and Myaeng, 2002; Lee and Myaeng, 2004; Lim et al., 2005; Santini et al., 2006). Machine learning techniques which incorporate thematic and structural features of web pages have also been applied successfully (Joachims et al., 2001; Eissen and Stein, 2004; Lindemann and Littig, 2006). A basic premise of these approaches is that web genres are manifested by thematic and structural features either on the level of a single page or of a site *as a whole*. In contrast to this, (Mehler et al., 2005; Mehler, 2007) refer to polymorphism as an aspect of informational uncertainty which says that hypertext units are compound manifestations of web genres so that their categorization goes hand in hand with their genre-related segmentation. In this sense, the segmentation of recurrent structural units of a given hypertext unit precedes its genre-related categorization. (Mehler et al., 2007) analyze various notions of informational uncertainty in support of this two-level approach of hypertext categorization. The present paper follows this line of research. That is, we, firstly, focus on hypertext segment types as a reference point of hypertext segmentation in order to, secondly, use the learnt segmentation of a hypertext unit as a reference point of its categorization.

2 METHODOLOGY

In order to implement the two-level model of hypertext categorization we proceed as follows. We expect that differences of hypertext types correlate with differences of their segment types. This leads to the assumption that when having detected the genre-related segments of a hypertext unit we can draw conclusions about its genre. However, because of polyfunctionality this is not a trivial task: On the one hand, a personal academic home page can be detected by identifying a segment of type *publications* which solely consists of references all of which contain the same author name. On the other hand, a *contact* segment is common to instances of different web genres, e.g. conference websites and project websites. Moreover, segments vary in terms of their location within websites of the same genre. Based on this observation a web page of a website as an instance of a web genre is called *polymorphic* if it contains at least two segments of different types (Mehler et al., 2007). Such pages are problematic input to hypertext categorization as their polymorphic segments are responsive to different types. If in contrast to this a

website includes only segments of the same type it is called *monomorphic*. Such pages allow to apply the classical apparatus of hypertext categorization as their monomorphism guarantees separability compared to other monomorphic pages. As a matter of fact, polymorphism is the predominant case and therefore interferes with applying the classical apparatus (Mehler et al., 2007). In order to successfully distinguish (i) polymorphic from (ii) monomorphic pages and in order to successfully segment the former while directly categorizing the latter we perform two consecutive tasks: Firstly, we describe how to automatically extract segments within instances of different hypertext types – this is done in Section 2.1. Secondly, we classify all extracted segments according to types of hypertext segments as recurrent building blocks of the hypertext types under consideration. This is described in Section 2.2.

2.1 HYPERTEXT TYPE SEGMENTATION

Our general notion of segment types of a website is the actual visual depiction of a hypertext. When focusing on the structure – abstracting from layout – the logical document structure is the central reference point to be considered. When focusing on layout, the stylesheet language specifying the presentation of a hypertext unit has to be interpreted. We infer that visually separable sections correlate with sections on the content level. According to this approach, segment borders are seen to be marked by staginess text phrases (as, e.g., headings expressed through the font-size), by image intersections or – as it is done most often – by visual spaces with no textual or graphical presentation. This step of segmentation is called *Segment Cutting*. The code of a web document includes style- and structure-related elements so that both can be considered. The reason is that identifying headers only is insufficient since the visual depiction can be coded by tags or by CSS code (e.g. class- or font-size values). Therefore, we need to process the logical document structure of a web document in conjunction with all its document internal and external layout information. In the next step, we search for the most prominent segment separation features occurring in the input document. These predefined features (e.g. `div`, `h1`, `h2`, `a`, font-size that exceeds a certain threshold) are explored as indicators of content section boundaries. As a result of this *Segment Cutting* step we often get segments which are too small. This relates, for example, to navigation items and headings which are segmented as single sections. In order to overcome this problem we perform a second step called *Segment Re-Connecting* which amalgamates small segments with their subsequent segments. A segment is considered as being too small if its size is below a specified threshold. As a result we gain a set of segments for each document which are next used as an input to segment categorization.

2.2 HYPERTEXT SEGMENT TYPE CLASSIFIER

Our approach of two-level web genre categorization is to attribute the genre of a website or page by classifying its segments. As we are able to partition a hypertext into its content-related segments (see Section 2.1), the process of hypertext segment type categorization can be performed next. Categorization is done by means of Support Vector Machine (SVM) which is a popular technique for data categorization. More specifically, we utilize SVM-Light (Joachims, 1997). Since an SVM produces a model which predicts the class label based upon the given instance features, feature selection is an important part in training a segment type classifier. We explore three classes of segment features: *Frequency of HTML-tags*, *frequency of tokens* and *frequency of segment structure-related numerical features*. In the process of feature extraction the frequency of HTML tags is defined by all occurred tags within the segment except script code and comments which are removed. We included only stemmed nouns, verbs, adjectives, adverbs, numerals, punctation marks and named entities. Entities are split into subcategories as for example email, proper names, location and country entities. Thirdly, numerical characteristics are extracted by computing the standard deviation of section, paragraph and sentence lengths of segments. We argue that e.g. sentence lengths of a contact section differs from that of a project information section.

3 EXPERIMENT

For evaluating our approach we conducted a categorization experiment by focusing on three hypertext types: *conference websites*, *personal academic websites* and *academic project websites*. Previous studies focused on distinguishing thematically clearly separated web genres such as web shops and web logs, listings and search pages. In contrast to this, we deal with hypertext types which are closely related based on their common thematic background. The reason is to deal with a more realistic scenario of web genre categorization. Although there is much effort

on building reference corpora of web genre categorization (Rehm et al., 2008), these data are still out of reach so that we needed to compile our own training corpora. Training corpus building has been done by three volunteers downloading 50 German web pages per hypertext type. Because of our two-level approach each of these 150 pages had been manually segmented in terms of their genre-related sections (e.g. *contact*, *research*, *call for papers*). That is, for each of the segment types distinguished in this study monomorphic segments have been identified as typical examples in order to learn classifiers which can detect these types of segments even in polymorphic web pages. As a result 1,250 segments have been manually typed and used for training our SVM-based segment classifiers. All annotated segments with their associated segment labels were used for feature selection. We also performed a feature selection procedure based on the GSS coefficient (Galavotti et al., 2000). However, feature selection did neither improve nor deteriorate our categorization. We trained for each hypertext type one SVM, the corresponding segment types against each other. For evaluation purposes, we used the Leave-One-Out F-Measure calculated by the SVM-Light implementation. In order to determine the hypertext type of a compound web document we developed a weighted finite-state transducer. Doing this, we argue that each web genre is represented by a corresponding document grammar which can be represented by a weighted directed graph. The grammar is determined by its transition probability accumulated through its segments. For the experiment on an overall categorization scenario, we randomly chose 60 websites from the annotated corpus – 20 for each hypertext type – using both *polymorphic* and *monomorphic* websites.

4 RESULTS

Tables 1–3 show the results of our first level categorization with a focus on hypertext segment types. Table 4 shows the results of the second-level categorization with a focus on web genres or hypertext types. The first level categorization shows that we clearly outperform the corresponding baseline scenario. However, our gain in F-measure values is round about .65 in the case of all three hypertext types – far away from more desirable values above .9. As a consequence of this the second level categorization results in an F-measure value also round about .625 – in conjunction with remarkably balanced recall and precision values which also outperform the corresponding baseline however to a minor degree.

Classes (11)	Recall	Precision	F-Measure
about	.578	.703	.634
accommodation	.680	.700	.690
call	.350	.389	.368
committees	.609	.609	.609
contact	.581	.720	.643
disclaimer	.706	.667	.686
organizer	.455	.417	.435
program	.692	.838	.758
registration	.729	.771	.749
sightseeing	.708	.739	.723
sponsors	.542	.650	.591
Average	.603	.655	.626
Baseline			.200

Table 1: Evaluation Results: Conference Sites

Classes (6)	Recall	Precision	F-Measure
contact	.947	.857	.899
links	.583	.636	.608
personal	.661	.709	.684
publications	.795	.720	.756
research	.485	.800	.604
teaching	.581	.643	.610
Average	.675	.728	.694
Baseline			.280

Table 2: Evaluation Results: Personal Sites

Let us look on related experiments in order to interpret these results and to shed light on the range of F-measure values gained by our experiment. (Santini, 2006) and (Eissen and Stein, 2004), for example, report on an accuracy of .67 (Nave Bayes Classifier) and an F-measure value of .89 (SVM Classifier), respectively. Although these related experiments in web genre categorization cannot be directly compared to ours, their results seem to be better. However, if we have a closer look on the experiments being conducted we get insight into experimental differences which put this statement into perspective: (Santini, 2006) and (Eissen and Stein, 2004) both deal with web genres which because of their thematic and functional divergence are obviously more separable than the ones considered here. Web genres as, for example, *Weblogs*, *FAQ*s, *Search Engine Pages* or *Listings* etc. are thematically and functionally more divergent than project pages and conference websites which both stem from the area of academics. Thus, we expect that there is a larger divergence of lexical and other features between the genres considered in their experiments compared to the genres explored in our experiment.

Classes (9)	Recall	Precision	F-Measure
contact	.823	.869	.849
events	.525	.636	.575
framework	.447	.568	.500
links	.471	.421	.444
news	.539	.560	.549
objectives	.603	.734	.662
project	.799	.789	.794
publications	.761	.761	.761
staff	.500	.807	.617
Average	.608	.683	.639
Baseline			.240

Table 3: Evaluation Results: Project Sites

Classes (9)	Recall	Precision	F-Measure
conference	.640	.640	.640
personal	.618	.627	.622
project	.620	.608	.614
Average	.626	.625	.625
Baseline			.428

Table 4: Evaluation Results: Hypertext Type Classification

5 CONCLUSIONS

This paper presented a model of two-level categorization of web genres. In contrast to related approaches the model presented here additionally explores and categorizes functionally and thematically demarcated segments of the hypertext types to be categorized. By classifying these segments conclusions can be drawn about the type of the corresponding compound web document. Our research provides results in support of solving this task and, thus, goes beyond the narrow focus of classical approaches to functional hypertext categorization.

6 ACKNOWLEDGEMENTS

Financial support of the German Research Foundation (DFG) through the Research Group 437, Project A4 and *Topic-Oriented Peer-to-Peer Agents in Digital Libraries* (LIS) at Bielefeld University is gratefully acknowledged.

REFERENCES

Eissen, S. M. Z. and Stein, B. (2004). Genre classification of web pages: User study and feasibility analysis. In *In: Biundo S., Fruhwirth T., Palm G. (eds.): Advances in Artificial Intelligence*, pages 256–269. Springer.

Galavotti, L., Sebastiani, F., and Simi, M. (2000). Experiments on the use of feature selection and negative evidence in automated text categorization. In *ECDL '00: Proc. of the 4th European Conf. on Res. and Adv. Tech. for DL*, pages 59–68, London, UK. Springer-Verlag.

Joachims, T. (1997). Text categorization with support vector machines: Learning with many relevant features. Technical report.

Joachims, T., Cristianini, N., and Shawe-Taylor, J. (2001). Composite kernels for hypertext categorisation. In *Proc. of the 11th Int. Conf. on Machine Learning*, pages 250–257. Morgan Kaufmann.

Karlgren, J. and Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *Proc. of the 15th Conf. on CL*, pages 1071–1075. ACL.

Kessler, B., Nunberg, G., and Schiitze, H. (1997). Automatic detection of text genre. pages 32–38.

Lee, Y.-B. and Myaeng, S. H. (2002). Text genre classification with genre-revealing and subject-revealing features. In *SIGIR '02: Proc. of the 25th Int. ACM SIGIR*, pages 145–150, New York, NY, USA. ACM.

Lee, Y.-B. and Myaeng, S. H. (2004). Automatic identification of text genres and their roles in subject-based categorization. In *HICSS '04: Proc. of the 37th HICSS'04*, page 40100.2, Washington, DC, USA. IEEE Computer Society.

Lim, C., Lee, K., and Kim, G. (2005). Automatic genre detection of web documents. In *Su K., Tsujii J., Lee J., Kwong O. Y., NLP*, Berlin. Springer.

Lindemann, C. and Littig, L. (2006). Coarse-grained classification of web sites by their structural properties. In *Proc. of the 8th ACM - WIDM'06*, pages 35–42, New York, NY, USA. ACM Press.

Mehler, A. (2007). Structure formation in the web. toward a graph-theoretical model of hypertext types. In Witt, A. and Metzing, D., editors, *Linguistic Modelling of Information and Markup Languages*. Springer, Dordrecht.

Mehler, A., Gleim, R., and Dehmer, M. (2005). Towards structure-sensitive hypertext categorization. In *Proc. of the 29th Annual Conf. of the GCS, Universität Magdeburg, March 9-11, 2005*, Berlin/New York. Springer.

Mehler, A., Gleim, R., and Wegner, A. (2007). Structural uncertainty of hypertext types. An empirical study. In *Proc. of Towards Genre-Enabled Search Engines: The Impact of NLP, September, 30, 2007, Borovets, Bulgaria*, pages 13–19.

Rehm, G., Santini, M., and Alexander Mehler, e. (2008). Towards a reference corpus of web genres for the evaluation of genre identification systems. In *Pro. of the 6th LREC 2008, Marrakech (Morocco)*.

Santini, M. (2006). Identifying genres of web pages. In *Proc. of TALN 2006*.

Santini, M., Power, R., and Evans, R. (2006). Implementing a characterization of genre for automatic genre identification of web pages. In *Proc. of the COLING/ACL*, pages 699–706, Morristown, NJ, USA. ACL.