# Classifying Web Sites

Christoph Lindemann and Lars Littig
University of Leipzig
Department of Computer Science
Johannisgasse 26
04103 Leipzig, Germany
http://rvs.informatik.uni-leipzig.de

## ABSTRACT

In this paper, we present a novel method for the classification of Web sites. This method exploits both structure and content of Web sites in order to discern their functionality. It allows for distinguishing between eight of the most relevant functional classes of Web sites. We show that a pre-classification of Web sites utilizing structural properties considerably improves a subsequent textual classification with standard techniques. We evaluate this approach on a dataset comprising more than 16,000 Web sites with about 20 million crawled and 100 million known Web pages. Our approach achieves an accuracy of 92% for the coarse-grained classification of these Web sites.

**Categories and Subject Descriptors:** H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *information filtering, selection process*; H.2.8 [Database Management]: Database Applications – *data mining*.

**General Terms:** Algorithms, Experimentation, Measurement.

**Keywords:** web site classification, web mining, web measurement

## 1. INTRODUCTION

Due to the tremendous size of the World Wide Web, it is more and more difficult to identify Web sites providing the information and services of interest. Therefore, the ability to discern the functionality of a Web site clearly improves the capability of search engines to present high quality search results. E.g. Yahoo! Mindset [4], which is still in research status, distinguishes between search results from Web sites of two functional classes. This allows for an adaptive ranking by favoring those results that correspond to the intent behind the search. Further opportunities that arise from coarse-grained classification of Web sites include personalized ranking and site tagging.

Research on the classification of Web sites aims at discovering useful knowledge for establishing structure in the Web. It can be divided into two sub areas that differ in the purpose and granularity of classification. Coarse-grained classification seeks to discern the functionality of a Web site in order to improve the quality and ranking of search results, e.g. [1], [3]. Fine-grained classification deals with the automated categorization of Web sites in order to build Web directories, e.g. [2].

Following this motivation, we present an approach for the coarse-grained classification of Web sites. This approach exploits structure and content of Web sites in order to classify them into one of the eight most relevant functional classes, namely *Academic*, *Blog*, *Community*, *Corporate*, *Information*, *Nonprofit*, *Personal*, and *Shop*. As major contribution, we derive a rich set of features from the structure of a Web site and show that utilizing them for a pre-classification of Web sites into sets of aggregated classes enhances a subsequent classification by content.


*WWW 2007*, May 8-12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

## 2. FEATURES FOR CLASSIFICATION

Deriving features for the classification of Web sites from a Web crawl is a process of knowledge discovery from data. Due to the heterogeneity of the Web and its lack of structure, it is crucial to identify properties of a Web site that best reflect its functionality. Therefore, we deal with the identification of properties that describe the structure of a Web site with respect to several aspects. In particular, we focus on size, organization, URL composition, technical realization, and link structure.

Overall, we identify 30 properties that reflect the functionality of a Web site so that they can be used as features for classification. We compute the information gain of each single feature, which describes its ability to distinguish between Web sites of the considered functional classes. Table 1 lists the feature with the highest information gain (IG) for the five different types of features. We exploit the link structure by considering the average external site outdegree, which describes the average number of links to different external Web sites. Furthermore, we analyze the size of a Web site in terms of page count, i.e. number of known pages. Regarding the organization of a Web site, the fraction of PDF and PS documents is the feature with the highest information gain. The fraction of pages per Web site that contain javascript is a feature that focuses on the technical realization of Web sites. Finally, features describing the general composition of the URLs of a Web site include the average number of digits within the URL path. In addition to this, Table 1 shows that some features can be derived from known pages. Since it is much more expensive to download a remote Web page than to perform in-memory classification operations [2], the ability to derive structural properties from non-crawled pages is a big advantage.

We further delve into the relation between structure and functionality by visualizing the differences between Web sites of different functional classes with box-and-whisker plots as exemplarily shown in Figure 1 and 2. We observe from Figure 1 that the largest Web sites belong to the classes Academic and Information as 50% of these Web sites have more than 10,000 pages. Figure 2 displays the maximum level of the page tree, which is a feature of type organization. Here, the classes Academic and Information stand out again since they possess the

**Table 1. Features derived from structure of Web sites**

| Feature | Type | Derived | IG |
|---|---|---|---|
| average external site outdegree | link struct. | crawled | 0.54 |
| number of known pages | size | known | 0.36 |
| fraction of PDF/PS documents | organization | known | 0.34 |
| fraction of pages with javascript | technical | crawled | 0.31 |
| average number of digits in path | URL | known | 0.29 |

deepest page tree. Furthermore, we observe a relation between the classes Academic and Information, Blog, Community, and Shop, and between the classes Corporate, Nonprofit, and Personal from both figures.

These relations fuel the necessity to explore further features for classification. Therefore, we aim at deriving features from the content of Web sites by creating a domain-specific dictionary (DSD) for each functional class. A DSD contains the most representative terms of a domain-specific corpus compared to a general corpus. The domain-specific terms within the DSD stand out as prominent features.

## 3. CLASSIFICATION OF WEB SITES

We evaluate the effectiveness of all considered approaches for classification on a dataset comprising 16,256 Web sites selected from publicly available Web directories for each of the eight functional classes. Crawling this dataset yields 20,731,273 crawled and 100,321,069 known pages.

In order to compare the accuracy of classification by structure and by content, we first employ a naive Bayesian classifier. Prior to classification, we perform data preprocessing which involves unsupervised data discretization, data transformation, and feature subset selection. The results of classification are evaluated by employing 10-fold cross validation. We observe that the classification by structure achieves a micro-averaged F1 score of 70%. Thus, structural properties are appropriate as features for classification but the achievable classification accuracy is limited due to the relation between some functional classes as mentioned before. This observation is further approved by the confusion matrix which is omitted due to space limitations. As a consequence, we aggregate the related classes into three sets. The classification of Web sites into these sets of functional classes results in a high F1 score of 94%.

We apply a standard technique for matching unknown text against our DSDs to classify Web sites by their content. Since the created DSDs are very descriptive for the functional classes, we achieve a micro-averaged F1 score of 84%. However, we observe that there are difficulties in correctly classifying Web sites of the functional classes Blog, Community, Information, and Personal as highlighted by a F1 score of at most 72%.

Considering the strengths and weaknesses of both approaches we conclude that classification by structure suffers from Web sites that possess a similar structure although being created for different purposes. Classification by content is affected by Web sites of some functional classes which cover a very broad spectrum of topics. Our novel approach overcomes these disadvantages by at first concentrating on the three sets of aggregated classes. Subsequently, the results of this pre-classification are used to enhance the classification by content.

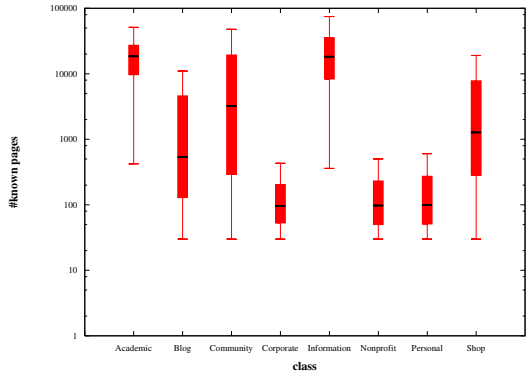


**Figure 1. Number of known pages**

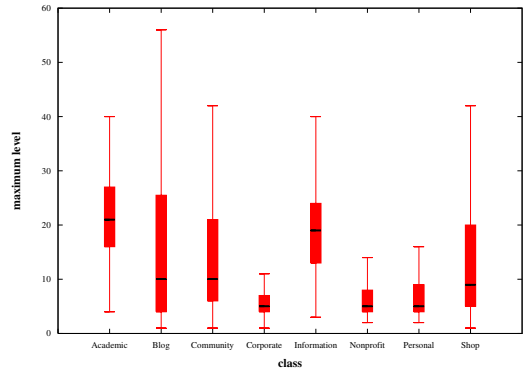


**Figure 2. Maximum level of page tree**

In particular, if the classification by content does not result in a clear decision for one class, i.e. the highest probability is below 0.75, or if one of the problem classes is predicted, the final decision is based upon the combined votes of both classifiers. This approach yields a micro-averaged F1 score of 92%. Thus, the achievements of classification solely based on structure or content are clearly surpassed.

## 4. CONCLUSIONS

We presented a novel approach for the coarse-grained classification of Web sites. This approach utilized 30 structural properties for a pre-classification of Web sites into three aggregated classes that considerably improved a subsequent classification by content using standard techniques.

## 5. REFERENCES

[1] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer, The Connectivity Sonar: Detecting Site Functionality by Structural Patterns, *Proc. 14th Conf. on Hypertext and Hypermedia*, Nottingham, United Kingdom, 2003.

[2] M. Ester, H.-P. Kriegel, and M. Schubert, Web Site Mining: A New Way to Spot Competitors, Customers and Suppliers in the World Wide Web, *Proc. 8th Int. Conf. on Knowledge Discovery and Data Mining*, Edmonton, Canada, 2002.

[3] C. Lindemann and L. Littig, Coarse-grained Classification of Web Sites by Their Structural Properties, *Proc. 8th Int. Workshop on Web Information and Data Management*, Arlington, VA, 2006

[4] Yahoo! Mindset, http://mindset.research.yahoo.com