Coarse-grained Classification of Web Sites by Their Structural Properties

Christoph Lindemann and Lars Littig

University of Leipzig
Department of Computer Science
Augustusplatz 10-11
04109 Leipzig, Germany

http://rvs.informatik.uni-leipzig.de

ABSTRACT

In this paper, we identify and analyze structural properties which reflect the functionality of a Web site. These structural properties consider the size, the organization, the composition of URLs, and the link structure of Web sites. Opposed to previous work, we perform a comprehensive measurement study to delve into the relation between the structure and the functionality of Web sites. Our study focuses on five of the most relevant functional classes, namely *Academic*, *Blog*, *Corporate*, *Personal*, and *Shop*. It is based upon more than 1,400 Web sites composed of 7 million crawled and 47 million known Web pages. We present a detailed statistical analysis which provides insight into how structural properties can be used to distinguish between Web sites from different functional classes. Building on these results, we introduce a content-independent approach for the automated coarse-grained classification of Web sites. A naïve Bayesian classifier with advanced density estimation yields a precision of 82% and recall of 80% for the classification of Web sites into the considered classes.

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 mining, web site classification, web structure mining, web measurement, naïve bayesian classification, search engines

© ACM, 2006. This is the author's version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution. The definitive version was published in *Proceedings of WIDM 2006,* http://doi.acm.org/10.1145/nnnnnn.nnnnnn

1. INTRODUCTION

The World Wide Web comprises an ever-growing number of Web sites providing different information and serving different needs. Therefore, it becomes more and more difficult to judge and classify Web sites. However, while Web sites differ in design and content, many Web sites are created for the same purpose. As a consequence, they are related by their functionality like the Web sites of two universities or two competing corporations.

The ability to classify Web sites with different functionality would be extremely valuable for improving the capabilities of search engines. In fact, the coarse-grained classification of Web sites could improve the quality of search results. This holds especially in case of an uncertain information need behind the query of an user. A coarse-grained classification would allow for marking search results in accordance to their corresponding Web sites' class. This would enable the user to easily choose results from the right context. First demo applications like Yahoo! Mindset [17] build on this idea. In addition to this, results from different classes could be included within the top 10 hits to present widespread results. Personalized ranking could be performed by favoring results from a certain class of interest with regard to the search history of the user. Furthermore, the freshness of the index of a search engine could be increased by adjusting the refreshing strategy in accordance to the change ratios of Web sites from different classes [5].

A very broad classification of Web sites as commercial, organizational, or educational can be performed by considering their top level domain, e.g. .com, .org, or .edu. As a consequence, this kind of classification is in part trivial for some Web sites. These Web sites are often physically located in the US. But as online shops and companies share the top level domain .com, the classification becomes inaccurate. Furthermore, Web sites of various genres reside in the same top level domain in countries other than the US, e.g. .ch, .de, or .fr.

The aim of this research is a coarse-grained classification of Web sites into classes describing their functionality. In other words, we want to identify what a Web site is, and not what it is about. Since our approach is solely based on the structural properties of Web sites, we identify properties reflecting their functionality. These structural properties consider the size, the organization, the composition of URLs, and the link structure of Web sites. Our study reveals to examine, amongst others, the number of pages, the fraction of HTML documents, and the average internal outdegree. Opposed to previous work, we perform a comprehensive measurement study to delve into the relation between the structure and the functionality of Web sites. Our study focuses on five of the most relevant functional classes, namely *Academic*, *Blog*, *Corporate*, *Personal*, and *Shop*, as discussed in Section 3. The resulting benchmark is based upon 1,461 Web sites composed of 7,020,191 crawled and 47,225,117 known pages, whereas known pages include the crawled pages and further discovered but not crawled pages. We present a detailed statistical analysis which identifies distinguishing measures for the coarse-grained classification of Web sites. Furthermore, we perform an advanced density estimation based on the distributions of these measures fitted from the measured data. Subsequently, we show that a naïve Bayesian classifier with advanced density estimation yields a precision of 82% and recall of 80% for the coarse-grained classification of Web sites into the five considered functional classes.

The remainder of this paper is organized as follows. Section 2 summarizes related work on Web site classification and Web measurement. Section 3 introduces our Web measurement methodology, which defines how to select and gather the data necessary for the analysis of structural properties. This analysis is described in Section 4. In Section 5, we build on the presented Web measurement results by providing an advanced approach for the automated classification of Web sites by their structural properties. Finally, concluding remarks are given.

2. RELATED WORK

In recent years, several approaches for the classification of Web sites have been introduced. These approaches have to be distinguished by the purpose of classification, i.e. approaches aiming on discerning the functionality of a Web site by analyzing their structure, e.g. [1], compared to approaches focusing on the categorization by topical content, e.g. [10], [12], [16].

Amitay et al. [1] examined the structural patterns of 202 Web sites in order to detect their functionality. They achieved a precision of 54.5% and 59%, respectively, for the classification of these Web sites into functional classes by applying two classification algorithms. Thus, their results provide evidence for the fact that the structure of a Web site reflects its functionality. Like [1], our approach for the coarse-grained classification of Web sites is based upon structural properties and driven by the same motivation. However, while [1] focused mainly on the link structure of Web sites, we aim at gaining a deeper insight into the relation between structure and functionality by analyzing different types of structural properties in a comprehensive measurement study. Furthermore, we introduce an advanced approach for classification resulting in a higher classification accuracy.

Pierre [16] discussed several issues related to the content-dependent, fine-grained classification of Web sites. He introduced a superpage-based approach for the classification of Web sites into industry categories utilizing metatags. Kriegel et al. [10] further exploited the semantic structure and local context information by representing a Web site as a tree of pages with topics. They employed the k-order Markov tree classifier and evaluated their approach based upon a testbed of 82,842 Web pages representing 207 corporate Web sites from two specific categories. Tian, Huang, and Gao [12] proposed an approach for Web site classification based on content, structure, and context information of Web sites. Their approach represented the Web site structure as a two-layered tree, i.e. a DOM tree for each page and a page tree for linking all pages. They presented a two-phase classification algorithm using the Hidden Markov Tree as the statistical model of both trees. Opposed to our work, all of these approaches [10], [12], [16] depend on textual content, focus on different small fractions of the Web, and are in part evaluated upon small benchmarks of Web sites. However, the outlined development towards more sophisticated classification approaches highlights the benefit of structural properties for the classification of Web sites even in the field of fine-grained classification.

Web measurement focusing on the link structure and the evolution of the Web has also been an active area of research in recent years. The papers [3], [6] reported large-scale measurement studies of the Web's link structure. Broder et al. [3] provided evidence that the macroscopic structure of the Web comprises three main components: IN, SCC, and OUT, i.e. looks like a bow tie. Building on these results, Dill et al. [6] discovered that both micro- and macroscopic graph structures possess the bow tie structure, i.e. cohesive sub-regions of the Web, introduced as so called thematically unified clusters, display the same characteristics as the Web at large. These unified clusters are for example a random collection of Web sites. We adopt the idea of considering unified clusters. Though, instead of building random collections, we consider classes of Web sites related by their functionality. [3], [6], aimed at gaining insight into the graph structure of the Web. Opposed to this, the aim of our study lies in gaining insight into the structure of Web sites and its relation to the functionality as basis for their coarse-grained classification.

Several other Web measurement studies, e.g. [5], [11], investigate the change frequency of individual Web pages across different top level domains. Cho and Garcia-Molina [5] studied the evolution of more than 500,000 Web pages drawn from 270 US Web sites. They reported that pages from university Web sites changed less frequently than pages on Web sites with .com top level domain. In a complementary study [11], Fetterly et al. studied the evolution of more than 150 million Web pages across European, US, and Asian top level domains. They observed a strong relation between top level domain and frequency of change. In summary, the evolution studies state that pages drawn from Web sites belonging to different domains change at different rates [5], [11]. We argue that Web sites with different functionality, which are in part comparable to Web sites from different top level domains, e.g. .com in contrast to Web sites from .edu, also differ in their structural properties. Therefore, they can be classified without inspecting their content and crawling them repeatedly.

3. MEASUREMENT METHODOLOGY

3.1 Selection of Web Sites

Since our studies focus on the functionality of a Web site and not on its topical content, we concentrate on the following five functional classes:

- *Academic*: Web sites of universities and research institutions.
- *Blog*: The group of Web logs as a popular representative of community Web sites with many individual content creators.
- *Corporate*: The Web presence of enterprises.
- *Personal*: The homepages of individuals or small groups.
- *Shop*: Online shops and auction portals offering products usually for sale.

Similar classes are considered whenever importance is attached to distinguishing Web sites because of their functionality [1], [13]. Considering the tremendous size of the Web and the various genres and modes of Web sites that occur on it, this set of functional classes cannot be complete. We intentionally omit the class comprising spam Web sites in order to have a clean index. Furthermore, we do not focus on search engines and Web directories as these types of Web sites are normally not searched for but used to find Web sites of the considered classes. Remember that our classification approach is motivated by achieving better search results. The consideration of further interesting classes like non-profit organizations or information portals is left for future work. However, we believe that the considered functional classes belong to the most relevant classes in accordance to the motivation and due to the fact that a vast amount of Web sites can be assigned to these classes.

In order to analyze the structural properties as general basis of classification, we select several Web sites from each considered functional class. This is achieved by randomly choosing the corresponding URL from Web directories like the Open Directory Project [7] within an appropriate category. Our measurements are strictly focused on the German part of the Web, i.e. we only obtain URLs from the top level domain .de. This is due to several reasons: Firstly, we want to avoid noisy data due to mixing Web sites from different countries whose structure and organization might be influenced by national distinctions. Secondly, we manually verified that the selected Web sites really belong to their assigned classes in order to have a solid benchmark. This aim can be achieved more efficiently with our knowledge of the German part of the Web. Thirdly, the top level domain .de is a very good example of the Web of an industrialized nation where Web sites of various functionalities reside in the same top level domain. In addition to this, we select only URLs pointing at the entry page of a Web site (e.g. http://www.uni-leipzig.de), i.e. URLs without a subdomain or a path leading to a page in a subdirectory. This approach guarantees that the data for each Web site is collected beginning with the entry page of a Web site.

3.2 Collecting the Data

For collecting data from the selected Web sites, we employ a search engine software system developed by our group. Our search engine runs on a Linux dual-processor PC server with 3.0 GHz Intel Pentium IV Xeon processors and 6 GB RAM.

The crawl is seeded from the sets of URLs of the Web sites belonging to the given functional classes. These sets are disjoint so that every Web site is assigned to exactly one class. We define a Web site as the set of Web pages which belong to the same domain, e.g. uni-leipzig.de. Thus, according to our definition the pages located in a subdomain of a Web site, e.g. informatik.uni-leipzig.de, are also considered as belonging to this Web site.

Crawling the entry page of each Web site first, the content of the page is parsed to extract links to other pages. Our crawler scans every single Web site in a breadth-first-search manner following only internal links, i.e. links pointing at a page within the same domain. External links are counted for later analysis, but are discarded afterwards. By crawling the Web sites in this way, we assure on the one hand that only pages from the pre-selected Web sites are downloaded and considered for our measurement study. On the other hand, we are able to determine the level of a page, i.e. the minimum number of clicks it is away from the entry page. Furthermore, we detect the language of the page's content by applying several heuristics like checking the http-header, the metatags, and counting language-specific stopwords.

Table 1. Measurement statistics of functional classes

Class	#Web Sites	#Crawled Pages	#Known Pages
Academic	158	2,233,615	11,860,670
Blog	222	751,717	2,071,394
Corporate	449	571,492	2,188,385
Personal	274	273,839	576,266
Shop	358	3,189,528	30,528,402
Total	1,461	7,020,191	47,225,117

To reduce the traffic placed on the servers hosting the selected Web sites, we crawled at most 20,000 pages per Web site or at most 2 GB of data. This boundary allows collecting data in a sufficient way as most Web sites comprise less than 20,000 pages. Our results show that even Web sites consisting of more than 20,000 pages can be well classified based upon the analysis of the data from the crawled pages. This is because we are able to use additional data from further discovered but not crawled pages of these Web sites as described in the following subsection. In addition to this restriction, our crawler obeys the robots exclusion protocol and the netiquette by keeping a timeout of at least two seconds between two successive requests to the same server. The crawl is completed when no further pages belonging to the pre-selected Web sites can be retrieved obeying the restrictions described before.

Collecting data in order to analyze the structural properties of a Web site in this manner is independent of a page's textual content, the importance of the content to a human user, the freshness of the content, and its change ratio as none of these measures is considered for the classification. We examine all Web sites from which at least 100 pages could be crawled correctly. On the one hand, this minimizes measurement errors due to Web sites which could not be crawled properly, e.g. because of flash intros or redirections. On the other hand, we assure that our analysis is based on statistical significant sample sets of pages within each Web site. The entire Web sites are characterized and classified based upon the analysis of these fetched portions.

Table 1 shows the resulting number of Web sites per functional class. Furthermore, it summarizes how many pages have been overall crawled per class and how many pages are known. The number of known pages includes the number of crawled pages plus further pages belonging to Web sites within the class. These Web pages have been discovered but have not been downloaded. All in all we analyze the structural properties of 1,461 Web sites from five distinct functional classes. Each class comprises at least 158 Web sites. The analysis is based upon 7,020,191 crawled and 47,225,117 known Web pages.

Since collecting the data in order to analyze structural properties is an essential step for our classification approach, it is important to consider its computational cost. The download of a remote Web page is much more expensive than in-memory classification operations [10], [12]. As described in the following subsection, most of the identified measures can be derived from the known pages of a Web site. Thus, they can be derived not only from crawled pages but from pages which have not been downloaded. Our benchmark contains 40,204,926 such non downloaded Web pages. The ability to analyze structural properties of these pages is a major advantage of our approach as it saves computational cost and enhances the accuracy at the same time.

3.3 Structural Properties of Web Sites

In order to identify and analyze structural properties which reflect the functionality of a Web site, we focus on structural properties that consider the size, the organization, the composition of URLs, and the link structure of Web sites.

We analyze the size of a Web site by determining the number of known pages per Web site (page count) and the average document size, which describes size in terms of amount of available data.

The organization of a Web site is spotted by counting the number of distinct subdomains per Web site, analyzing the fraction of document types, checking the number of different languages, and by detecting the average and maximum level of its pages. We compute the number of subdomains by adding up the number of different host parts within the URLs of one Web site. The document type is determined by inspecting the file extension. We checked the URLs for many of the most common file extensions including .html, .xml, .txt, .pdf, .ps, .php, .asp, .jsp, .pl. Further detected file extensions are added up together, defined as other. A language is counted for the observed Web site if at least ten pages present their content in this language. We are able to detect the languages German, English, French, Spanish, Italian, and Dutch. The average and the maximum level of a Web site are determined by checking the level of every page, i.e. by counting the minimum number of links that have to be followed beginning from the entry page, which has level 0, in order to reach this page.

Further properties describing the general composition of the URLs of a Web site can be directly derived from the URLs. We determine the length of the site name, i.e. the string length of the top domain without subdomains, the average length of all URLs, and the average length of the URL path. Furthermore, we count the number of slashes and digits within the path. Although these properties might not be promising for reflecting the functionality at first sight, we find that Web sites of distinct classes differ in these measures as presented in Section 4.

Obviously, the link structure provides another source to gain further insight into the structural properties of a Web site. We consider the link structure by calculating the average and maximum overall, internal, and external outdegree as well as the external site outdegree. The outdegree of a page is defined as the number of links within a page pointing at other pages belonging to the same Web site, i.e. internal links, or at pages on other Web sites, i.e. external links. The sum of the internal and the external outdegree is the overall outdegree. Duplicated links within one page are counted only once. The difference between the external outdegree and the external site outdegree is that the former counts the number of links to distinct external pages whereas the latter counts the number of links to distinct Web sites.

Tables 2 and 3 summarize all measures for the different types of properties and state whether a measure can be derived from all known or just from the crawled pages of a Web site. Since most of the measures can be derived from all known pages, the available amount of data for the analysis grows rapidly as stated before.

Table 2. Measures derived from known pages

Measure	Type
Number of known pages	Size
Avg. number of slashes in URL path	URL
Avg. number of digits in URL path	URL
Avg. URL length	URL
Avg. length of URL path	URL
Length of sitename	URL
Level (avg., max)	Organization
Number of subdomains	Organization
Fraction of document types (HTML, PDF, PS, PHP, TXT, ASP, JSP, XML, PERL, Other)	Organization

Table 3. Measures derived from crawled pages

Measure	Type
Outdegree (max., avg.)	Link structure
Internal outdegree (max., avg.)	Link structure
External outdegree (max., avg.)	Link structure
External site outdegree (max., avg.)	Link structure
Number of different languages	Organization
Avg. document size	Size

Although our analysis is focused on the German part of the Web, we believe that our methodology is applicable for other parts of the Web, too. This holds especially for Web sites within the top level domain of other industrialized countries. Since the identified structural properties are independent of a page’s content, they can be determined easily for the Web sites representing the relevant functional classes. These Web sites can be obtained from trusted Web directories for the considered top level domain. As a consequence, applying our measurement methodology would gain insight into the relation between structure and functionality of Web sites within top level domains other than .de as basis for their coarse-grained classification.

4. MEASUREMENT RESULTS

In this section, we provide insight into how the structure reflects the functionality of a Web site. Therefore, we analyze the structural properties by plotting the cumulative distribution of each measured parameter. The goal of this analysis is to identify distinguishing measures for the coarse-grained classification of Web sites from different functional classes. Due to space limitations we do not present the distributions of all considered properties, but concentrate on those for which the differences between the classes are most impressing. However, for the classification process outlined in Section 5 all measured properties stated in Section 3 are taken into consideration.

As a first example of a distinguishing measure derived from the structural properties, we analyze the size of a Web site in terms of number of Web pages. Obviously, Web sites from the functional class Academic are in general much larger than Web sites from the class Personal. This intuition is underlined by Figure 1 which shows that more than 80% of the Web sites from class Academic have at least 6,000 pages. For Web logs this fraction is only about 17% while Web sites from class Shop are also quite large. 80% of the Web sites from this class consist of more than 1000 pages. The smallest Web sites in terms of number of known pages belong to class Personal, which is an intuitive result. Only 10% of the Web sites from the class Corporate have more than 2000 known pages. This is due to the fact, that our benchmark for this class comprises many small and medium-sized enterprises. We conclude that the number of known pages of a Web site is an important indicator of the functional class it belongs to.

As an impressing example of the measures representing the fraction of document types used within a Web site, Figure 2 provides the distribution of the fraction of HTML documents for each functional class. We observe in this figure that more than 70% of the Web sites from class Personal consist almost entirely, i.e. with more than 95%, of HTML documents. Web sites from class Corporate have on average less HTML documents. Here, a larger fraction of about 40% of the Web sites has almost no HTML documents. Another 20% of the Web sites from this class consist with more than 80% of HTML documents. The remaining 40% of these Web sites are partly composed of HTML and non-HTML documents. The fraction of HTML documents on Web sites of class Academic is rather uniformly distributed between 0% and 100%. Web sites from the classes Shop and Blog are composed to the least extend of HTML documents. Only 15% and 7% of the sites, respectively, consist almost entirely of HTML documents, and about 50% and 70%, respectively, have almost no HTML document on the entire Web site.

Figure 3 shows that quite simple structural properties regarding the composition of URLs nevertheless also reflect the functionality. We observe that especially Web sites of class Shop differ from the other Web sites as the pages of 80% of these Web sites have on average more than five digits per URL. A frequent use of session IDs to keep state, e.g. during the shopping session of a customer, is a possible reason for this observation. Web sites from class Personal have on average the smallest number of digits within their URLs, i.e. about 80% have five or less digits. This is an intuitive result as the creation of personal homepages rather seldom utilizes more sophisticated techniques.

Finally, significant differences between the structural properties of the considered functional classes can be observed in the average internal outdegree depicted in Figure 4. Web sites from the class Shop have on average the strongest internal navigational structure, i.e. provide many links to other pages of the same site. For example about 50% of the Web sites from this class contain on average more than 40 internal links per page. This might be due to the fact, that shops often contain comprehensive product listings with links to Web pages describing the listed products in more detail. In contrast, Web sites from the class Personal have on average only few internal links per page, i.e. only about 20% of these sites have on average more than 10 internal links per page. Between these extremes 40% of the Web sites from class Academic have more than 20 internal links per page. Thus, the average internal outdegree provides significant discriminative power to distinguish between the different functional classes.

In summary, the figures show that all considered types of structural properties, i.e. properties regarding the size, the organization, the composition of URLs, and the link structure, reflect the functionality of a Web site. Therefore, the differences in the distributions of structural properties of Web sites with different functionality allow for deriving measures with discriminative power, denoted as *discriminators* in the rest of this paper. We exploit these discriminators for the coarse-grained classification of Web sites.

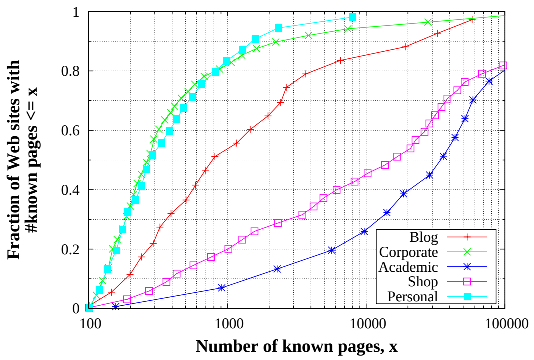

Figure 1. Number of known pages of functional classes

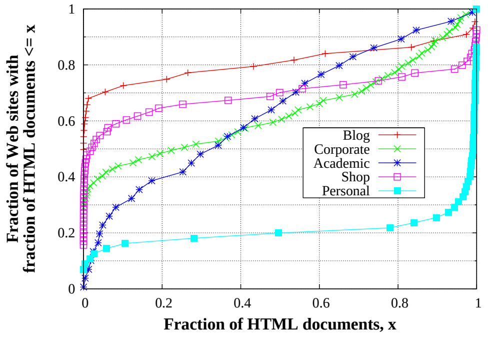

Figure 2. Fraction of HTML documents of functional classes

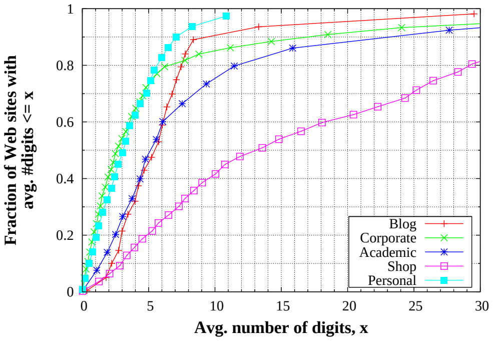

Figure 3. Avg. number of digits of functional classes

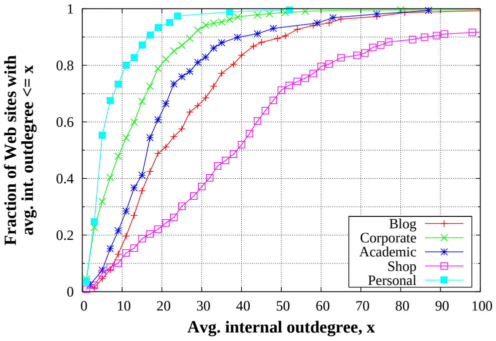

Figure 4. Avg. internal outdegree of functional classes

5. COARSE-GRAINED CLASSIFICATION

5.1 Naïve Bayesian Classifier with Advanced Density Estimation

The naïve Bayesian classifier [9] is known to be a simple but effective technique for classification tasks in several application domains like spam-filtering and pattern recognition. As the name “naïve” suggests this method makes the important simplifying assumption that the discriminators are conditionally independent given the class. Although this assumption does not hold in many applications, the naïve Bayesian classifier nevertheless provides excellent classification performance [8]. As a consequence, our approach builds on a naïve Bayesian classifier. It utilizes an advanced density estimation procedure by fitting the measured data to a comprehensive set of distribution functions to improve the classification accuracy as outlined in this section.

In our application, the naïve Bayesian classifier computes the probability of a Web site belonging to one of the five considered functional classes, denoted as C_i with i=1,…,5, given the set of discriminators. This probability is denoted as $P(C_i| \vec{x})$ where $\vec{x}$ = $<x_1,\ldots,x_d>$ is a vector composed of the particular values observed for the discriminators of the Web site to classify. d is the number of used discriminators. The computation is based on several components. Firstly, the likelihood of the discriminators given the considered class, denoted as $P(\vec{x}|C_i)$. Secondly, the prior probability reflecting the fraction of existing Web sites for each class, denoted as $P(C_i)$. Thirdly, a normalizing constant in the denominator, denoted as $P(\vec{x})$, which is invariant across classes. Putting it all together, according to Bayes’ theorem the probability of a Web site belonging to a specific class given the set of discriminators can be computed by

$$P(C_i \mid \vec{x}) = \frac{P(C_i)P(\vec{x} \mid C_i)}{P(\vec{x})} = \frac{\frac{1}{5}\prod_{j=1}^{d} P(x_j \mid C_i)}{\sum_{i=1}^{5}\left[\frac{1}{5}\prod_{j=1}^{d} P(x_j \mid C_i)\right]} \qquad (1)$$

Since the fractions of Web sites from the considered classes are a priori unknown for the World Wide Web, we assign for each class the same prior probability, i.e. $P(C_i)$=1/5 for i=1,…,5. The probabilities $P(x_j \mid C_i),\ j = 1,...,d,\ i = 1,...,5$ can be deduced by evaluating the appropriate probability density function $f_{j,i}$ modeling the distribution of discriminator j for class i.

The computation of the probability of each class C_i given the set of discriminators $\vec{x}$ by a naïve Bayesian classifier heavily depends on the set of discriminators used for classification [15]. On the one hand, it is important to use as many discriminators as possible to distinguish between classes. On the other hand, the classification performance suffers from discriminators which are redundant due to correlations. Thus, we employ the discriminator selection approach outlined in [15], which chooses an appropriate set of discriminators according to the best classification accuracy.

A second simplifying assumption often used in naïve Bayesian classifiers is the normality assumption of the discriminators. Since this assumption does not hold in general, we aim to improve the classification accuracy with advanced density estimation. That is, while retaining the independence assumption of the discriminators our approach accounts for the fact that discriminators may follow probability distributions other than normal. Thus, we fit the measured data to the probability density functions (pdf) exponential, normal, lognormal, Weibull, Pareto, and additionally to the step function given in equation (2).

$$f(x) = \begin{cases} \frac{p_1}{m_1} & \text{if} \quad x \le m_1 \\ \frac{p_{k+1}}{m_{k+1} - m_k} & \text{if} \quad m_k < x \le m_{k+1} \quad \text{for} \quad k = 1,...n-1 \\ 0 & \text{if} \quad x > m_n \end{cases} \qquad (2)$$

with $p_k \ge 0,\ m_k \ge 0$ for $k = 1,...,n$
and $m_k < m_{k+1}$ for $k = 1,...n-1$

This equation defines a probability density function used for mathematically capturing discretization techniques for Bayesian classifiers. The density function defines the probability p_k that a measured data value x falls into the interval number k, defined by the interval boundaries m_{k-1} and m_k. The parameters for this distribution are the interval boundaries m_k and the relative frequencies p_k for interval number k, k=1,…,n. We employ the weighted proportional k-interval discretization approach by Yang and Webb [18] for determining m_k. The parameters p_k are derived from the measured data by calculating the relative frequencies of measured data values for each interval k given by m_k. Furthermore, the parameters of the well-known continuous distributions exponential, normal, lognormal, Weibull, and Pareto are determined by fitting the cumulative distribution functions (CDF) of the distributions to the measured data by least-squares regression utilizing the Levenberg-Marquardt algorithm [2]. We employ the CDF instead of the pdf for fitting because it is not biased by discretization.

By this choice of possible distribution functions we account for the observation made in Section 4 that the measured distributions of the discriminators may differ extremely both among different discriminators and even for the same discriminator among different classes. For example recall from Figure 2, that the fraction of HTML documents for class Personal roughly takes only the two values 0 and 1, i.e. describing a discrete distribution. Opposed to this, the values of the same discriminator for class Academic are more evenly spread between 0 and 1, i.e. describing a continuous distribution. With our approach we combine the advantage of the independence from discretization, given by the continuous distributions, with the advantage of the high flexibility of the step function.

The choice of the best suited distribution for each discriminator and class is based upon the root-mean-square of residuals [14] denoted as Δ and defined by equation (3).

$$\Delta = \sqrt{\frac{\sum_{l=1}^{s}\left(CDF_m(x_l) - F(x_l)\right)^2}{s}} \qquad (3)$$

s denotes the number of measured values, CDF_m denotes the CDF of the measured values, and F(x) denotes the CDF of the fitted probability distribution. Thus, after fitting each of the six distributions to the measured data for each discriminator and class, the error terms are calculated and compared. The distribution with the smallest Δ fits best to the measured data and is therefore chosen for the classification task.

Table 4. Results of fitting distributions to measured data

Discriminator	Class	Distribution	Parameter 1	Parameter 2	Error term Δ
Number of known pages	Academic	lognormal	$\mu = 10.2024$	$\sigma = 1.7678$	0.03270
	Blog	lognormal	$\mu = 6.74803$	$\sigma = 1.9701$	0.02440
	Corporate	lognormal	$\mu = 5.18206$	$\sigma = 1.56141$	0.03139
	Personal	lognormal	$\mu = 5.31639$	$\sigma = 1.57409$	0.01135
	Shop	lognormal	$\mu = 9.35839$	$\sigma = 2.68941$	0.00263
Avg. internal outdegree	Academic	lognormal	$\mu = 2.74898$	$\sigma = 0.72313$	0.02123
	Blog	Weibull	$\alpha = 1.47891$	$\lambda = 0.00717598$	0.01859
	Corporate	Weibull	$\alpha = 1.08249$	$\lambda = 0.0621629$	0.01334
	Personal	lognormal	$\mu = 1.5701$	$\sigma = 0.901351$	0.02181
	Shop	Weibull	$\alpha = 1.5335$	$\lambda = 0.0027379$	0.03157

Table 4 exemplifies the distributions, corresponding parameters and Δ values for the two discriminators number of known pages and average internal outdegree. We observe that the number of known pages follows a lognormal distribution for all considered classes. Furthermore, the small values of the error terms Δ indicate a close fit of the distribution functions to the measured data. Recall that we do not consider Web sites from which less than 100 pages could be crawled. Thus, for fitting the distribution functions to the measured data we subtract 100 from the measured values of the number of known pages in order to let the distribution start at zero instead of 100. We further observe, that there are significant differences in the parameters of the distribution between the individual classes, reflecting the power of the discriminator. For the average internal outdegree, Table 4 points out that this discriminator follows a lognormal distribution for the classes Academic and Personal instead of a Weibull distribution for the other classes. Thus, we conclude that the distributions of a particular discriminator for different classes may not only differ in terms of distribution parameters but furthermore in terms of the distribution function.

Most measured distributions of the discriminators not stated in Table 4 can be well modeled by the lognormal or Weibull distribution, respectively. The discriminators constituting fractions of documents of a specific type are generally modeled by the discrete distribution with smallest root-mean-square of residuals. This is due to the fact that for many of these discriminators the measured data only take on a small number of distinct values. As a consequence, this behavior can only be appropriately modeled by the discrete distribution. Further discriminators which are best modeled by the discrete distribution include the number of languages and the number of subdomains, clearly because these discriminators can only take discrete values.

5.2 Classification Results

As [4], [10] we evaluate the accuracy of our classification approach by employing 10-fold cross validation. This method divides the overall collected data set of the Web sites of all considered classes randomly into ten sets of equal size. In each of the ten turns, another set is utilized for the automated classification and the remaining nine sets are used as training data. Subsequently, we determine for each turn the achieved precision and recall for the considered five major classes of Web sites individually and then average them over all turns of the cross validation. In the actual classification process a Web site is assigned to the class with the highest probability, so that all Web sites are classified. In addition to this, we examine different confidence levels by assigning a Web site to a particular class only if the probability for that class is above a certain threshold, i.e. 0.8 and 0.9, respectively. Otherwise the Web site is marked as undefined.

We present the results of the classification process for the naïve Bayesian classifier with advanced density estimation in Table 5. The precision of the classification for one class is defined as the fraction of Web sites classified as members of that class which also actually belong to the class. The recall is defined as the fraction of Web sites which belong to a class and are also classified as members of that class. Micro-averaging and macro-averaging aggregate the precision and recall for each class into an overall measure [4]. Micro-averaging makes the overall precision and recall depend mostly on the precision for classes with a large number of sites in the sample set (i.e., Corporate and Shop). Opposed to this, the macro-averaged measures pay equal importance to each class. Furthermore, the F1 score describes the overall performance of a classifier with respect to recall and precision at the same time [4]. We calculate the F1 score for the macro-averaged values. Table 5 shows that our approach yields a precision of 82% and recall of 80% for the classification of all Web sites resulting in a F1 score of 81%. The micro-averaged precision and recall are 80%. Considering the different confidence levels, the precision of our classifier can be increased to 87% by applying a threshold of 0.9 for the classification probability. However, this approach is at the cost of a smaller recall of 70% as 278 of the 1,461 Web sites remain unclassified. In addition to this, Table 5 shows that Web sites of the classes Academic and Corporate can be best classified with a precision of 98% and 85%, respectively.

Table 5. Classification accuracy at different confidence levels

Conf. / Class	No Thresh.		Thresh.=0.8		Thresh.=0.9	
	Pre.	**Rec.**	**Pre.**	**Rec.**	**Pre.**	**Rec.**
Academic	0.98	0.85	0.99	0.84	0.99	0.82
Blog	0.79	0.78	0.84	0.72	0.86	0.70
Corporate	0.85	0.83	0.86	0.75	0.86	0.68
Personal	0.74	0.78	0.81	0.69	0.85	0.64
Shop	0.74	0.78	0.76	0.72	0.76	0.67
# Undef.	0		176		278	
Micro-avg.	0.80	0.80	0.84	0.73	0.85	0.69
Macro-avg.	0.82	0.80	0.85	0.74	0.87	0.70
F1 score	0.81		0.79		0.78	

Table 6. Confusion matrix of classifier without threshold

Class	C1	C2	C3	C4	C5
C1: Academic	**134** **84.8%**	1 0.6%	14 8.9%	1 0.6%	8 5.1%
C2: Blog	0 0.0%	**174** **78.4%**	2 0.9%	18 8.1%	28 12.6%
C3: Corporate	1 0.2%	2 0.4%	**372** **82.9%**	25 5.6%	49 10.9%
C4: Personal	0 0.0%	28 10.2%	20 7.3%	**215** **78.5%**	11 4.0%
C5: Shop	2 0.6%	15 4.2%	32 8.9%	30 8.4%	**279** **77.9%**

To shed light on the distinction of the considered classes, Table 6 presents the confusion matrix for the classification approach without a threshold. The single rows of Table 6 show for each functional class how many Web sites have been assigned to the different classes and the corresponding percentage. The diagonal highlights the number of correct classified Web sites, so the percentage equals the recall for this class. Examining the confusion matrix, we see, for example, that the classifier sometimes confuses Web sites from the classes Blog and Personal. 8.1% of the Web sites from class Blog are assigned to class Personal and vice versa 10.2% of class Personal are classified as Blog. This is due to the fact that personal homepages sometimes contain small private Web logs resulting in a similar structure. Furthermore, 10.9% of the Web sites from class Corporate are misclassified as Shop. We observe that corporate Web sites sometimes include small shops or provide extensive listings of their products without selling them. A hybrid approach considering structural properties and content at the same time might overcome these difficulties resulting in an even higher classification accuracy.

6. CONCLUSION

We identified and analyzed structural properties of Web sites considering their size, their organization, the composition of URLs, and their link structure. We presented a comprehensive measurement-based study to examine the relation between structure and functionality of Web sites. This study was focused on five of the most relevant functional classes, namely *Academic*, *Blog*, *Corporate*, *Personal*, and *Shop*. The analysis of our study revealed substantial differences between the Web sites from these classes with respect to all types of structural properties. In fact, Web sites from the functional class Academic differ from the other classes in their number of known pages. Shops can be easily distinguished inspecting their average internal outdegree while we observe a different fraction of HTML documents for Web sites from class Personal.

We showed that the distributions of structural properties can be well captured by exponential, normal, lognormal, Weibull, Pareto probability distributions, and a step function. These results allow for the effective coarse-grained classification of Web sites by their structural properties using a naïve Bayesian classifier with advanced density estimation. Depending on the confidence level, our approach yields a macro-averaged precision of up to 87% for the classification of Web sites from different functional classes.

As our results are very promising, future work will focus on the identification of further structural properties of Web sites with discriminative power. We will extend our analysis to consider additional functional classes as mentioned before. Furthermore, applying our methodology to Web sites from top level domains other than .de will gain insight into their structural properties in comparison to the German part of the Web.

7. 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] D. Bates and D. Watts, *Nonlinear Regression and Its Applications*, Wiley, 1988.

[3] A. Broder, R. Kumar, F. Maghoul, P. Rhaghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, Graph Structure in the Web, *Proc. 9th Int. WWW Conf.*, Amsterdam, The Netherlands, 2000.

[4] S. Chakrabarti, *Mining the Web*, Morgan Kaufmann, 2003.

[5] J. Cho and H. Garcia-Molina, The Evolution of the Web and its Implications for an Incremental Crawler, *Proc. 26th VLDB Conf.*, Cairo, Egypt, 2000.

[6] S. Dill, R. Kumar, K. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins, Self-Similarity in the Web, *ACM Trans. on Internet Technology*, 2, 205-223, 2002.

[7] DMOZ: open directory project, www.dmoz.org

[8] P. Domingos and M. Pazzani, On the Optimality of the Bayesian Classifier under Zero-One Loss, *Machine Learning*, 29, 103-130, 1997.

[9] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley and Sons, 2nd Edition, 2001.

[10] M. Ester, H. 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.

[11] D. Fetterly, M. Manasse, M. Najork, and J. Wiener, A Large-scale Study of the Evolution of Web Pages, *Proc. 12th Int. WWW Conf.*, Budapest, Hungary, 2003.

[12] W. Gao, T.-J. Huang, and Y-H. Tian, Two-phase Web Site Classification Based on Hidden Markov Tree Models, *Web Intelligence and Agent Systems*, 2004.

[13] D. Gibson, K. Punera, and A. Tomkins, The Volume and Evolution of Web Page Templates, *Proc. 14th Int. WWW Conf.*, Chiba, Japan, 2005.

[14] J. Kenney and E. Keeping, Root Mean Square, *Mathematics of Statistics*, Van Nostrand, 3rd Edition, 59-60, 1962.

[15] R. Kohavi and G. John, Wrappers for Feature Subset Selection, *Artificial Intelligence*, 97, 273-324, 1997.

[16] J. M. Pierre, On the Automated Classification of Web Sites, Linköping Electronic Articles in Computer and Information Science, Sweden 6, 2001.

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

[18] Y. Yang and G. Webb, Weighted Proportional k-Interval Discretization for Naive-Bayes Classifiers, *Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, Seoul, Korea, 2003.