Structural Uncertainty of Hypertext Types. An Empirical Study

Alexander Mehler, Rüdiger Gleim, Armin Wegner
Bielefeld University
Universitätsstraße 25
D-33615 Bielefeld
Alexander.Mehler,Ruediger.Gleim,Armin.Wegner@uni-bielefeld.de

Abstract

This paper presents a comparative study of three webgenres. It analyzes the distribution of their instances regarding hyperlink-based structures. The starting point is the notion of polymorphism as an aspect of informational uncertainty. The main result of the study is that hypertext graphs are multidimensionally distributed in a Zipfian manner which demands adapting algorithms of web structure mining to different structural classes.

Keywords

web structure mining, webgenre, hypertext type, quantitative structure analysis, search engine

1 Introduction

As far as web search engines are seen to be more than information retrieval on web pages, retrieved pages might be classified by the document type they manifest or these types might be made a retrieval criterion *ex ante* [13]. The idea behind this approach is that documents are not only distinguished in terms of their topics, but also of the various, though recurrent functions they serve [1]. In this sense, documents may deal with the same topic while serving different functions and, vice versa, they may be functionally equivalent while dealing with different topics. Thus, *function* and *content* are different, though not orthogonal reference points of document classification. There had been much research to introduce the notion of a *webgenre* in order to reflect this distinction [8, 13, 16, 17] – for an overview see [14, 15]. That is, webgenres or hypertext types – as they are interchangeably called in this paper – are functional units whose instances are at least alike or even equivalent in terms of the functions they serve. The premise underlying this research is that *hypertext categorization* can be realized by analogy to *text categorization* in order to reliably identify the range of existing webgenres and to classify their instances. Note that the majority of approaches plead for a straightforward application of text categorization methods to hypertext units [8].

This paper is in the line of research on utilizing hypertext categorization as a means to build webgenre-sensitive search engines. But other than the standard approach it focuses on *websites* as relevant search units instead of viewing *web pages* as the focal units of retrieval. That is, we conceive a search engine which retrieves *websites* possibly consisting of several pages in order to meet a certain content-related search query. More specifically, we think of a search engine which starting from instances of a certain hypertext type demarcates all its relevant constituents meeting the query – above or below the level of single web pages. As will be clarified in Section 2 we call such a search engine *sensitive to the logical document structure of hypertext types*. In order to make this program work, we need to automatically classify hypertext types and to demarcate their constitutive pages. *That is, we need to identify the borders of hypertext units which are underspecified due to the limits of HTML and related standards.* In this paper, we demonstrate why and also to which extent this is a hard task. This is mainly done by means of a comparative evaluation of structure formation on the level of websites of three webgenres: the genre of conference websites [8], of personal academic homepages [13] and of so called city websites which serve as official portals of communities or cities.

The paper is organized as follows: Section 2 distinguishes several types of informational uncertainty where polymorphism and temporal variability turn out to be specific to hypertext categorization. Section 3 analyzes the genre-specific distribution of hypertext graphs as models of websites. Finally, Section 4 concludes and prospects future work.

2 Structural Uncertainty

In [8] we introduced the notion of polymorphism as a characteristic of the functional organization of websites as instances of hypertext types. It occurs if the same expression unit (e.g. a web page) manifests several functions of a webgenre by distinct segments. We have shown that polymorphism makes hypertext categorization a hard task [8]. In this section, we compare polymorphism with *ambiguity*, *polyfunctionality*, *vagueness* and *temporal variability* as alternative aspects of the structural uncertainty of hypertext types. This is done in order to clarify the machine learning task induced by polymorphism.

We start from a form-meaning model in which segments of websites (e.g. web pages, their segments and hyperlinks) are distinguished as manifestation units from functions *served* and meanings *encoded* by these units:

- Firstly, we suppose a segmentation $S = \{s_i \,|\, i \in \mathfrak{I}\}$ of a website $x \in S$ which segments x into manifestation units $s_i \in S$.

- Secondly, we suppose a set of (e.g. functional or semantic) categories $C = \{c_1, \ldots, c_m\}$ together with a relation $F \subseteq S \times C$ where $(s, c) \in F$ iff segment s serves the function, has the meaning or is of type c whether as a whole or because of one of its segments.

Without loss of generality we say that segment s *serves* the category c iff $(s, c) \in F$. Let now $s \in S$ be a segment of x and $\sigma_{S=s}(F) = \{(s, c) \in F \,|\, c \in C\}$ such that $|\sigma_{S=s}(F)| > 1$. Then we distinguish five types of structural uncertainty in terms of $\sigma_{S=s}(F)$:

- **Ambiguity** occurs if the available information provides evidence for mapping s *as a whole* (e.g. a web page as a constituent of the webgenre *conference website*) onto several categories $c \in C$ (e.g. *call for workshops*, *call for papers* or *call for posters*) where, actually, s serves a single category. *As a whole* means that s has no constituent $s' \in S$ which instead of s serves any of these categories. We call the elements of the projection $\pi(\sigma_{S=s}(F))[C] = \{c \,|\, (s, c) \in \sigma_{S=s}(F)\}$ of $\sigma_{S=s}(F)$ onto C *competing interpretations* of s. In accordance with [5], the selection $\sigma_{S=s}(F)$ is said to be *non-specific* in proportion to the number $|\sigma_{S=s}(F)|$ of these interpretations, *dissonant* in the case that $\pi(\sigma_{S=s}(F))[C]$ is partitioned into two disjunct sets which map opposite, but nevertheless non-specific interpretations, and, thirdly, *confuse* to the degree that the latter partition contains more than two sets. As ambiguity presupposes that s actually serves a single, though underspecified category, this case of informational uncertainty matches the classical application scenario of disambiguating competing categorizations. *Therefore, it is not specific to hypertext categorization, but solvable by extending the informational input to categorization.*

- **Polyfunctionality** occurs if the available information provides evidence for mapping s *as a whole* onto several, but noncompeting interpretations $c \in C$. This case supposes that s simultaneously serves several categories so that $\pi(\sigma_{S=s}(F))[C]$ might be called the set of *concomitant* interpretations of s. Obviously, this scenario matches the multilabel categorization paradigm and, *therefore, is likewise not specific to hypertext categorization, but manageable by extending the categorization function to a relation.*

- **Polymorphism** occurs if the available information provides evidence for mapping s onto several categories $c \in C$ where each of them is served by a separate segment of s, *but not by s as a whole.* A special case of this occurs if s serves the same category several times by separate segments. At first glance, polymorphism might be considered as a sort of polyfunctionality and, thus, to be manageable in terms of a multilabel categorization. Actually, this is inappropriate if we look for search engines that are sensitive to the segmental structure of websites and their pages. As an example think of a wiki-related search engine where users do not only want to retrieve relevant articles, but need to get the exact section answering their query. *It is this segmentation-sensitive search scenario which makes polymorphism a relevant problem in hypertext categorization.* Obviously, mapping s onto all categories served by its segments does not suffice as we first need to decompose s in order to know which segment serves which category (cf. [9]). Note that polymorphism is a trivial consequence of superization, that is, constituents (e.g. phrases) of complex signs (e.g. sentences) normally diverge in terms of the categories (e.g. phrase types) they serve. But as far as the segmental structure of hypertext units is partly manifested by hyperlinks (as, e.g., in the case of portals in Wikipedia), *polymorphism is specific to hypertext categorization where segmentation has to go along with processing these links (see below).*

- **Vagueness** occurs if the available information provides evidence for mapping s onto a category to a degree smaller than one. In this case, F is generalized to a fuzzy relation over $S \times C$ where $\mu_F(s, c)$ denotes the membership degree. Although categorization is almost always vague in this sense, vagueness is normally abstracted by a sort of signum function where above a certain threshold a segment is said to serve the focal category and otherwise not. *Obviously, this type of uncertainty is not specific to hypertext categorization.*

- **Temporal variability** relates to the life cycle of hypertext types whose instances normally retain their access point (i.e. URL) while they age.[1] That is, if a corpus of, e.g., conference websites is built at a certain point in time it may contain instances of completely different time stages (e.g. *call for papers*, *submission closed*, *conference over*) of the life cycle of that webgenre. Note that the membership to such a stage is normally implicit (where the archive function of wiki-based systems is just an exception). This kind of temporal variability or structural fluency is untypical to text categorization which deals with "completed" texts whose change history is disregarded. *Thus, this sort of uncertainty is specific to hypertext categorization.*

Note that aspects of informational uncertainty which are said to be *non*specific to hypertext categorization are nevertheless relevant to it. Note further, that these aspects may co-occur. Polyfunctionality co-occurs, for example, with vagueness, if s serves several functions as a whole, but each of it to a separate degree. Likewise, ambiguity and polymorphism co-occur if separate constituents of s serve different functions where at least one of them is ambiguously attributed to different categories.

[1] Of course, things get more complicated if this access point changes, too.

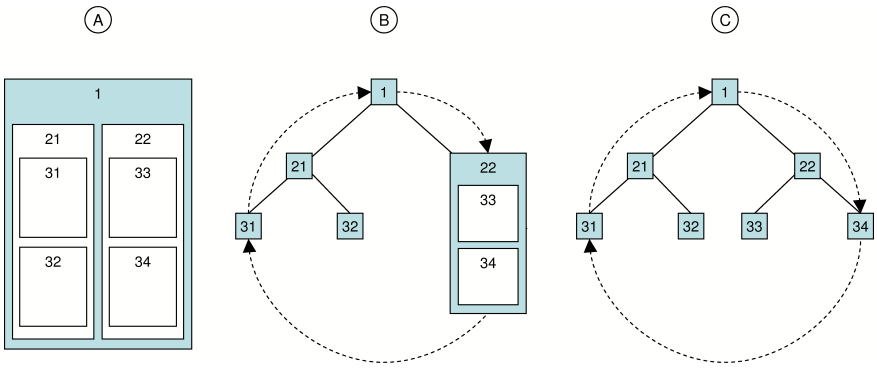

Fig. 1: *Alternative instantiations of a webgenre indicating a spectrum of structural equivalents in web-based communication—filled squares denote web pages, white squares represent segments and arcs denote hyperlinks.*

Because of its relevancy to hypertext categorization we now focus on polymorphism which becomes a serious problem, if webgenres are categorized in terms of layout units (e.g., web pages) as done by the majority of approaches to hypertext categorization (cf. [7] for an overview). This is demonstrated in Figure 1 where the same instance of a webgenre is transformed into structurally divergent, but supposably functionally equivalent documents: Variant A, for example, consists of a single page which in variant C is broken down into different ones. More specifically, C uses hyperlinks to manifest a subordination hierarchy which A manifests as an inclusion hierarchy. Variant B—which only partly breaks down A's inclusion hierarchy—is settled in-between these extreme cases.[2] Figure 1 demonstrates a spectrum of structural variants whose distribution is approached in Section 3. Its empirical relevancy has several implications to hypertext categorization:

- Firstly, *there is no unique vertical order of layout units into graph levels if these levels are specified in terms of hyperlinks:* Supposed that the variants A, B and C are functionally equivalent, it is hardly possible to speak of their leveled organization which is immediately alignable.[3] Node 34, for example, belongs to the putative level 1 in variant A, while it belongs to the putative levels 2 and 3 in variant B and C, respectively. That is, layout units as web pages lack a reference point which guarantees comparability along their arrangement into levels.
- Secondly, *there is no unique horizontal order of layout units into sibling vertices:* Node 34 in B cannot be uniquely identified as the neighbor of 33 as long as there is evidence that the functionally equivalent variants A and C map the same nodes onto different levels.
- Thirdly, *there is the fallacy of missing structure:* Focusing on web pages as the input units to hypertext categorization bears the risk to mistakenly leave out structural information distributed over several pages. This is most drastically shown by variant C in contrast to variant A.

Obviously, the true reference point of segmenting polymorphic units in hypertext categorization is the website as a whole. That is, properly categorizing webgenres presupposes to delimit their instances—internally and externally. Once more, it is the mixing of subordination and inclusion hierarchies within various instances of the same webgenre which complicates this delimitation task as exemplified in Figure 2: Supposed that all pages in variant C are monomorphic, node 1 in A and node 22 in B are polymorphic. The problem induced by this polymorphism is not that variants A, B and C are incomparable—as a matter of fact, they are supposed to be equivalent in terms of the categories they serve. Rather, their comparison cannot rely on web pages. That is, providing a proper *tertium comparationis* presupposes to break down polymorphism starting from websites as a whole so that putative structural differences based on polymorphism are removed. In [7], we call this tertium comparationis the *logical hypertext document structure* of websites.

Based on these specifications, we can now dispute two predominant premises of hypertext categorization:

- Firstly, we make a distinction between the visible layout and the hidden logical document structure of instances of hypertext types.
- Secondly, we do not view web pages as the primary instances of webgenres. Rather, we focus on websites as the relevant manifestation units of hypertext types which in some cases might be manifested by single pages. Thus, we necessarily view hyperlink structures as an indispensable resource of delimiting hypertext types.[4]

So far, we have diagnosed that temporal variability and, especially, polymorphism, are specific to hypertext categorization which demand to revise or at least to extend its underlying apparatus. The next section sheds light on the distribution of polymorphism in three webgenres.

[2] Note that ideally variant C breaks down the inclusion hierarchy of A in a way that none of its web pages is anymore polymorphic.

[3] [6] reports on an analysis in which the maximum website level is 179.

[4] This is certainly self-evident. However, as most experiments in hypertext categorization focus on pages only, they disregard hyperlink structure.

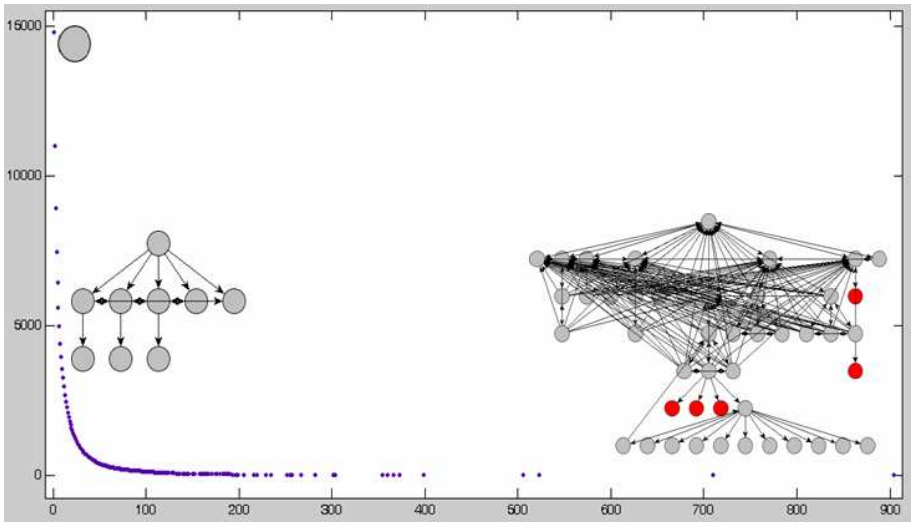

Fig. 2: *Schematic model of a multidimensional Zipfian distribution of hypertext graphs as instances of hypertext types.*

3 An Empirical Study of the Distribution of Hypertext Structures

What can we say about the distribution of the instances of hypertext types in terms of their structure? In order to tackle this question we concentrate on the impact of polymorphism as motivated in Section 2 (leaving out the temporal variability for future work). As explained above, we expect to find a wide range of structurally divergent instances of the same type which complicates their classification. This can be explained as follows: In Figure 1 we have distinguished three prototypical examples which span the spectrum of polymorphic websites as extreme cases: starting from polymorphic websites (Case A) going via intermediary cases (Case B) to non-polymorphic ones (Case C). Now we ask about the frequency distribution of these and related cases. In order to motivate this, look at Figure 2. It shows a putative distribution of the structural patterns in question: A large set of highly polymorphic websites which all consist of a single page switches over very rapidly into a set of more structured websites till we finally reach the very small set of highly structured ones (characterized by a large number of pages spanned by a deeply as well as broadly structured kernel tree in conjunction with many cross-referencing links).

Actually, Figure 2 manifest a *hypothesis* about the distribution of the instances of a given hypertext type in terms of their structure. It prognoses that this distribution resembles a multidimensional Zipfian distribution which might be fitted by a multidimensional power law by simultaneously reflecting several structural characteristics (e.g. size, height or width of the corresponding graphs and trees, respectively). As the fitting of such a multidimensional distribution is difficult, we follow an alternative approach: First, we independently observe and fit the distributions of some quantitative characteristics of hypertext graphs, test whether we can successfully fit corresponding power laws in each of these cases and finally show that the characteristics in question correlate positively so that the hypothesis of a multidimensional Zipfian distribution of hypertext graphs of the same webgenre is supported. An interesting byproduct of this approach is that we might also find evidence for the non-existence of a *typical instance* of the respective hypertext type supposed that the exponents of the successfully fitted power laws fall below a certain threshold [11]. Obviously, it is worth deciding this question.

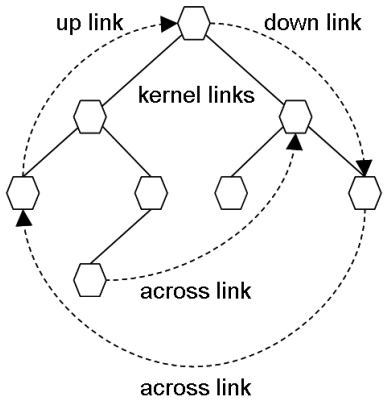

Fig. 3: *A schematic model of a generalized tree.*

In order to do this and thereby measuring the extent to which polymorphism is a prevalent characteristic of hypertext types, we now analyze the distributions of several structural characteristics of so called *hypertext graphs*. Hypertext graphs have been introduced as a graph model of hypertext types on the level of websites [7]. Generally speaking, hypertext graphs are generalized trees [4] which consist of a kernel tree in conjunction with graph-inducing across, up and down links – see Figure 3 for a schematic model of generalized trees.[5] The first step to measure polymorphism is to automatically map each input website on its corresponding hypertext graph model – as described in [8]

[5] See [10] for an extension of generalized trees as a text representation model.

Table 1: *The corpora and some of their quantitative characteristics (showing the respective min and max value using web pages as the counting unit).*

hypertext type	#websites	language	size	height	width
city websites	180	German	(1;2094)	(0;28)	(1;1866)
conference websites	1,460	English	(1;1130)	(0;23)	(1;1079)
personal academic homepages	1,350	English	(1;1258)	(0;10)	(1;1100)

– where vertices denote web pages and directed edges represents hyperlinks connecting these pages. As a result, we get three corpora of hypertext graphs as summarized in Table 1, that is, of the genre of conference websites, of personal academic homepages and of city websites. The reason to analyze the first two webgenres is that they have already been analyzed in the literature on hypertext categorization (cf., e.g., [13]). The third type is added to extend the basis of comparison.

In order to analyze the structural diversity of the instances of these types, we concentrate on the following quantitative characteristics of hypertext graphs:[6]

- **Out degrees:** Our first characteristic is the *out degree* on the level of web pages. The out degree of a web page is the number of edges starting from that page. Obviously, hyperlinks induce directed edges including across, up, down and also external links (leading out of the focal website). As we take all these types of links into account we operate on generalized trees. Note that we consider multiple and even parallel edges linking the same vertices (i.e. hyperlinks that are anchored at different positions of a page). Starting from the distribution of the out degrees of all page nodes of all hypertext graphs of a given webgenre corpus we build the *rank out degree distribution* of that corpus in which the page with the highest number of outgoing links has the highest rank followed by the page with the second highest out degree etc. Note that this is the sole distribution considered here which refers to web pages as the reference point of distribution building – in all other cases, websites are referred to as sampling units. One reason to include page-related out degrees as a quantitative characteristic of genre-specific websites is to show that linking has a *Zipfian* preference order by analogy to the Web as a whole [12].

- **Size, height and width:** The second characteristic is the *size* of websites in terms of the number of their constitutive pages. We build a rank size distribution per corpus which is input to curve fitting in order to clarify the Zipfian nature of the distribution of this feature. The same is done with the *width* (i.e. the number of leafs) and the *height* (i.e. the maximum geodesic distance from the root of the kernel tree of a hypertext graph to its leafs). The size, the width and the height are computed for the roots of the kernel trees of the corresponding hypertext graphs so that these characteristics are actually computed per website. Thus, the rank size, width and height distributions are – as all other distributions introduced subsequently – website-related.

- **Depth imbalance:** As a measure of the imbalance of the kernel trees of hypertext graphs we compute their A*bsolute* D*epth* I*mbalance* (ADI) according to [2]. Starting from an input vertex v, this measure basically computes the standard deviation of the adjusted heights of v's child nodes. As before, we compute the ADI for the root vertex r of the kernel tree T of each hypertext graph of our webgenre corpora, where the higher $\mathrm{ADI}(r) \in \mathbb{R}_+$ the higher the variance among the heights of r's child nodes, the more imbalanced T.

- **Child imbalance:** By analogy to the ADI we also compute the A*bsolute* C*hild* I*mbalance* (ACI) [2]. Whereas the ADI evaluates imbalance in terms of the heights of child nodes, the ACI focuses on the sizes of the trees dominated by these nodes. As before, size is measured as the number of vertices of the respective tree. Obviously, the ADI also reflects the width of a tree and, thus, provides complementary information to the ACI.

- **Compactness and stratum:** Finally, the stratum and the compactness measure – as introduced by [2] – operate on graphs, that is, in the present case on generalized trees. The Stra*tum* (Stra) is a metric which measures, so to speak, the deviation of a given hypertext graph from a purely linearly organized graph with the same number of vertices where a stratum of 1 indicates a maximally hierarchically organized hypertext. The C*ompactness* (C) analogously varies from 0 (i.e. graphs that are completely disconnected) to 1 (i.e. hypertexts that correspond to completely connected graphs). Other than stratum which explores hierarchical structures, the compactness focuses on the role of cross-referencing links, that is, of up, down and across links whose usage raises the degree of connectedness. Other than the ACI and the ADI which measure the imbalance of trees, stratum and compactness operate on graphs and thus provide additional information as they also explore graph-inducing links.

The results of fitting the rank distributions of these characteristics for the three webgenres considered here are reported in Table 2. We observe that except for the stratum and the compactness distribution (and a sole exception) all other six characteristics are distributed

[6] See [3,] for the role of quantitative indices in hypertext modeling.

Table 2: *Results of curve fitting based on either the power law (pl) model $c \cdot x^{-\alpha}$ or on the exponential (exp) model $c \cdot e^{-\alpha \cdot x}$. A is the corpus of city websites, B the corpus of conference websites and C the corpus of personal academic homepages. The quantitative characteristics operate on either the kernel Tree (T) or on the Generalized Tree (GT) as the focal graph model (GM). All values are round to 2 decimal places. $\bar{R}^2$ is the adjusted coefficient of determination. The column focus distinguishes fittings based on the rank and on the complementary cumulative (cc) distribution.*

				A		B		C	
distribution	focus	GM	fitting	α	$\bar{R}^2$	α	$\bar{R}^2$	α	$\bar{R}^2$
size	rank	T	pl	0.73	.97	1.2	.93	1.29	.99
height	cc	T	pl	0.18 (exp)	.95	3.57	.97	1.58	.99
width	rank	T	pl	0.76	.97	1.2	.94	0.77	.91
out degree	cc	GT	pl	1.1	.99	2.28	.99	2.03	.99
child imbal.	rank	T	pl	0.84	.96	0.82	.93	0.81	.90
depth imbal.	rank	T	pl	0.53	.94	0.42	.96	0.45	.94
stratum	rank	GT	exp	0.81 (pl)	.99	0.003	.99	0.01	.99
compactness	rank	GT	exp	0.014	.98	0.002	.96	0.004	.95

according to some power law indicating a Zipfian behavior in terms of very skewed distributions. In the case of the rank size distribution this means, for example, that there is only a very small set of websites consisting of very many pages while there is a large set of websites each consisting of a single page. This holds for all three webgenres. The same can be said for the out degree, height, width as well as for the child and depth imbalance distributions of the sites and pages analyzed (the only exception is the height of city websites). A note on fitting: We do not consider grouping which is a usual method to deal with 'fraying tails'. In some of the cases, we start instead from the complementary cumulative (or Pareto-like) distribution derived from the corresponding rank distribution to get the value of the exponent α which maximizes the determination coefficient among all candidate functions when fitting the power law model $cx^{-\alpha}$.

The next step is to evaluate wether the ranks of the websites of a webgenre computed according to the rank distributions induced by the eight characteristics correlate or not. If so we would know, for example, that a large website (in terms of its size) tends to be deeply and broadly structured in terms of its height and width. Table 3 shows that in the case of conference websites this is true for most of the characteristics. More interestingly, this is always true for personal academic homepages: Whenever we have such a homepage in our corpus with a large amount of absolute depth imbalance it also tends to have a high value of height, width and size. Alternatively, a *compact* personal academic homepage tends to have a hierarchical structure (i.e. a high stratum value) while a low value of compactness goes together with a lack of hierarchical structure. These and related positive correlations are reported in Table 3 which is in support of a multidimensional Zipfian distribution in the case of personal academic homepages and of conference websites (though to a minor degree because of some insignificant and negative correlation, respectively). In contrast to this, the smaller corpus of city websites does *not* confirm this picture – whether this is due to the small corpus size or to the fact that this is a feature of that webgenre cannot be answered here. Anyhow, the results of Table 3 support the view that there are different structural classes within the same hypertext type whose frequencies tend to be distributed according to a power law. Interestingly, this means that there is a large class of websites which hide, so to speak, their structure within a couple or even a single page – this observation confirms approaches which focus on page internal structuring as done by Rehm [13, 14]. But the results also show that *the other half* of websites tends to be well structured by means of hyperlinks so that we shall not omit these structural classes. This diagnosis is confirmed by some of the absolute values of the exponents of the power laws being fitted which because of being smaller than 2 (or 1 in the case of rank distributions) indicate the non-existence of a finite expectation value (cf. [11] for the details of this argumentation). Thus, the corpora analyzed so far do *not* allow to speak of a *typical hyperlink-based website structure*. Of course, this negative result includes the class of "unstructured" websites each of which consists of a single page. Rather, in order to cover a given webgenre we need to take all structure classes into account. A last remark on fitting: So far we did not control the effect of temporal variability which may displace the distributions unexpectedly. Future work will need to control this impact in order to give a more reliable picture of the genre-sensitive distribution of hypertext graphs.

Summarizing our findings w.r.t genre-sensitive search engines, we come up with the following diagnosis: As far as we need search engines that can retrieve query-related segments *above* and *below* the level of web pages we need to deal with polymorphism of websites. Because of the Zipfian distribution of hypertext graphs this seems to be a hard task as we cannot train our algorithms to a sort of typical website structure. Rather, we have to adapt our algorithms to several different classes thereof.

4 Conclusions

In this paper we have presented a quantitative analysis of the instances of three hypertext types. A basic

Table 3: *Rank correlations ρ of seven quantitative characteristics of city websites (A), conference websites (B) and of personal academic homepages (C). The characteristics are Compactness (C), Stratum (Stra), Absolute Child Imbalance (ACI), Absolute Depth Imbalance (ADI), Height (H), Size (S) and Width (W). All values are rounded to two decimal places. Rank correlations which are judged to be insignificant due to a t-Test are underlined ($\alpha = 0.05$; $H_0 : \rho_S = 0$).*

A	C	Stra	ACI	ADI	H	S	W
C	1	0.32	0	0.07	-0.56	0.11	0.15
Stra		1	0.04	0.16	-0.5	0.16	0.2
ACI			1	0.77	-0.36	0.72	0.71
ADI				1	0.13	0.59	0.62
H					1	0.03	0.18
S						1	0.99
W							1

B	C	Stra	ACI	ADI	H	S	W
C	1	0.27	0.48	0.18	-0.35	0.15	0.12
Stra		1	0.4	0.53	0.23	0.53	0.51
ACI			1	0.58	0.03	0.51	0.47
ADI				1	0.87	0.98	0.98
H					1	0.91	0.92
S						1	1
W							1

C	C	Stra	ACI	ADI	H	S	W
C	1	0.86	0.97	0.93	0.87	0.94	0.94
Stra		1	0.94	0.98	0.99	0.97	0.97
ACI			1	0.98	0.94	0.99	0.99
ADI				1	0.99	1	1
H					1	0.98	0.98
S						1	1
W							1

result of our study is that there exist highly skewed distributions of the instances of the genres involved. This indicates the need to explore the logical document structure of websites or, at least, to distinguish structural classes in order to overcome negative implications of this skewness. Future work will deal with both of these approaches: (i) adapting methods of web structure mining to the specifics of these structural classes and (ii) exploring the logical document structure of different webgenres. By means of this we expect to contribute to making search engines webgenre-sensitive. The reason is that this research focuses on identifying segments of pages as well as of websites and, thus, on overcoming the limits of page-centered search engines.

Acknowledgements

Financial support of the German Research Foundation (DFG) through the Research Group 437 *Text Technological Modelling of Information* and through the Sonderforschungsbereich 673 *Alignment in Communication* at the University of Bielefeld is gratefully acknowledged.

References

[1] D. Biber. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge, 1995.

[2] R. A. Botafogo, E. Rivlin, and B. Shneiderman. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, 10(2):142–180, 1992.

[3] M. Dehmer. Data Mining-Konzepte und graphentheoretische Methoden zur Analyse hypertextueller Daten. *LDV-Forum*, 20(1):113–141, 2005.

[4] M. Dehmer, A. Mehler, and F. Emmert-Streib. Generalized trees. In *Proceedings of the 2007 International Conference on Machine Learning: Models, Technologies & Applications (MLMTA'07), June 25-28, 2007, Las Vegas*, 2007.

[5] G. J. Klir and T. A. Folger. *Fuzzy Sets, Uncertainty, and Information*. Prentice Hall, Englewood, 1988.

[6] W. Koehler. Classifying web sites and web pages: the use of metrics and URL characteristics as markers. *Journal of Librarianship and Information Science*, 31(1):297–307, 1999.

[7] A. Mehler. Structure formation in the web. A graph-theoretical model of hypertext types. In A. Witt and D. Metzing, editors, *Linguistic Modeling of Information and Markup Languages. Contributions to Language Technology*, Text, Speech and Language Technology. Springer, Dordrecht, 2007.

[8] A. Mehler and R. Gleim. The net for the graphs – towards webgenre representation for corpus linguistic studies. In M. Baroni and S. Bernardini, editors, *WaCky! Working Papers on the Web as Corpus*, pages 191–224. Gedit, Bologna, 2006.

[9] A. Mehler, R. Gleim, and M. Dehmer. Towards structure-sensitive hypertext categorization. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, and W. Gaul, editors, *Proceedings of the 29th Annual Conference of the German Classification Society, March 9-11, 2005, Universität Magdeburg*, pages 406–413, Berlin/New York, 2006. Springer.

[10] A. Mehler, U. Waltinger, and A. Wegner. A formal text representation model based on lexical chaining. In *Learning from Non-Vectorial Data – Workshop at the KI 2007, September 10, 2007, University of Osnabrück*. 2007.

[11] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46:323–351, 2005.

[12] R. Pastor-Satorras and A. Vespignani. *Evolution and Structure of the Internet*. Cambridge University Press, Cambridge, 2004.

[13] G. Rehm. Towards automatic web genre identification – a corpus-based approach in the domain of academia by example of the academic's personal homepage. In *Proc. of the Hawaii Internat. Conf. on System Sciences*, 2002.

[14] G. Rehm. *Hypertextsorten: Definition, Struktur, Klassifikation*. Books on demand, Norderstedt, 2007.

[15] M. Santini. *Automatic Identification of Genre in Web Pages*. Phd thesis, University of Brighton, Brighton , United Kingdom, 2007.

[16] M. Santini. Characterizing genres of web pages: Genre hybridism and individualization. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, 2007.

[17] B. Stein and S. M. zu Eißen. Automatische Kategorisierung für Web-basierte Suche – Einführung, Techniken und Projekte. *Künstliche Intelligenz*, 4:11–17, 2004.