Polymorphism in Generic Web Units

A corpus linguistic study

Alexander Mehler
Rüdiger Gleim
Department of Linguistics
Bielefeld University
{Alexander.Mehler,Ruediger.Gleim}@uni-bielefeld.de

Abstract

Corpus linguistics and related disciplines which focus on statistical analyses of textual units have substantial need for large corpora. More specifically, genre or register specific corpora are needed which allow studying variations in language use. Along with the incredible growth of the internet, the web became an important source of linguistic data. Of course, web corpora face the same problem of acquiring *genre* specific corpora. Amongst other things, web mining is a framework of methods for automatically assigning category labels to web units and thus may be seen as a solution to this corpus acquisition problem as far as genre categories are applied. The paper argues that this approach is faced with the problem of a many-to-many relation between expression units on the one hand and content or function units on the other hand. A quantitative study is performed which supports the argumentation that functions of web-based communication are very often concentrated on single web pages and thus interfere any effort of directly applying the classical apparatus of categorization on web page level. The paper outlines a two-level algorithm as an alternative approach to category assignment which is sensitive to genre specific structures and thus may be used to tackle the problem of acquiring genre specific corpora.

1 Introduction

Corpus linguistics and related disciplines which focus on statistical analyses of textual units have substantial need for large corpora. The more resources are available the better can noise be reduced and the better statistical results tend to be. Of course, size alone does not matter (Kilgarriff and Grefenstette, 2003). Depending on the field of research, corpora have to be balanced at least with respect to the membership of textual units in specific genres (Biber et al., 1998). Thus, the acquisition of suitable corpora is one of the key issues of corpus linguistic projects.

Along with the incredible growth of the internet, the web became an important source of linguistic data. Kilgarriff and Grefenstette (2003) give a comprehensive survey of the usage history and relevance of the web as a corpus. The web offers vast amounts of data, but also brought up new problems. One of the most important issues is the acquisition of a corpus of units which solely belong to a specific web genre (Rehm, 2004). Another aspect relates to the various data formats and malformed codings which complicate computational processing. However the web is and will be an important source for corpus linguistics.

An approach for solving the problem of web corpus acquisition comes from web mining (Kosala and Blockeel, 2000). Amongst others, web mining includes the categorization of *macro structures* (Amitay et al., 2003) such as web hierarchies, directories or corporate sites, but also of web pages (Fürnkranz, 2002) as well as the identification of page segments, i.e. of *web micro structures* (Mizuuchi and Tajima, 1999). The basic idea is to perform web structure mining as function learning. That is, web units *above*, *on* or *below* the level of single pages are mapped onto at most one predefined category label per unit (Chakrabarti et al., 1998). The majority of these approaches utilizes text categorization methods. But other than text categorization, they also use HTML markup, metatags and link structure beyond bag-of-word representations of the pages' wording as input of feature selection (Yang et al., 2002). Chakrabarti et al. (1998) and Fürnkranz (1999) extend this approach by including pages into feature selection which are interlinked with the focal pages to be categorized. Moreover, the aggregation of the representations of the wording, markup and linking of pages is demonstrated by Joachims et al. (2001).

In following section, we argue that this approach is confronted with the problem of a many-to-may relation between expression and content/function units which interferes the effort to directly apply the apparatus of categorization to web pages. In order to support this line of argumentation, section (4) outlines a corpus linguistic study of quantitative characteristics of hypertext graphs used to represent the structure of genre-specific websites. The underlying representation format is presented in section (3). Finally, section (5) proposes a two-level algorithm of hypertext categorization in order to tackle the problem of acquiring genre specific web corpora.

2 Aspects of Informational Uncertainty in Web-based communication

From a general point of view, the basic assumption behind the approach of categorization in web mining is that monolingual web units of similar content or function tend to have similar structures and wording. In reverse, this hypothesis implies that analyzing the wording of single web pages allows predicting their function or classifying them with respect to topic categories, which in the area of text and hypertext categorization are usually predetermined. In this framework, web pages are supposed to form reliable units of function and content structure, respectively. In case of *conference websites* as instances of a *web genre* (Yoshioka and Herman, 2000) *function* relates, for example, to *informing* about conference topics, *submitting* papers or *registering* to the conference, whereas *content* relates to the specific topics or scientific area the conference is about.

This approach is confronted with *polymorphism* and its reversal relation of *discontinuous manifestation* which establish a many-to-many relation of *function*, *content* and *expression* units (Mehler et al., 2004). As a result of this relation, the content/function structure of web sites is generally *not* directly accessible by (just segmenting and subsequently) categorizing *single web pages*, that is, by mapping them onto predetermined content or function categories. Polymorphism occurs if the same hypertextual unit manifests several categories by means of separate segments. This one-to-many relation of expression and content/function units is accompanied by a reversal relation according to which the same content/function unit is distributed over several expression units. This combines to a many-to-many relation between visible, manifesting and implicit, manifested structure.

Polymorphism is given, when, for example, the same web page of a conference website provides information about the *call for papers*, the *submission procedure* and *conference registration*, that is, when manifesting at least two functions. The reversal case occurs, when, for example, a call for papers is manifested by different pages each informing about another conference topic.

Both one-to-many relations, which — from a sign theoretical point of view — are directly opposed, result in two corresponding problem categories of corpus-based analysis of web documents:

- Polymorphism results in *multiple categorizations* without being reducible to ambiguity of category assignment since in this case several categories are actually manifested by the same expression unit. Thus, resolving polymorphism cannot be reduced to the task of disambiguating category assignment as applied in machine learning and related areas.

- On the other hand, discontinuous manifestation results in *flawed* or even *missing categorizations* since in this case the web pages under consideration manifest the focal content/function categories only in part, that is, by means of features which do not allow to infer the content/function category being instantiated.

If web pages manifest several categories, but some of them only in part (where these categories are distributed over the remaining pages), polymorphism and discontinuous manifestation occur simultaneously. For the time being, not very much is known about the distribution of this many-to-many relation and its impact on hypertext categorization, since corpus-based hypertext analysis rather concentrates on single pages abstracting from this relation. Insofar this many-to-many relation is prevalent, a proper corpus linguistic analysis of web-based units and their categorization is bound to a preliminary *structure analysis* which first resolves polymorphism and discontinuous manifestations. We hypothesize this structure analysis to be constrained as follows:

- The functional structure of many websites is determined by their membership in *web genres* (Yoshioka and Herman, 2000). Hypertext categories are — as far as they focus on the functions, web pages are intended to serve — specific to these genres, i.e. they vary with the focal genre of, for example, *conference websites* (Yoshioka and Herman, 2000), *personal home pages* (Rehm, 2004) or *online shops* etc.

- What is common to instances of different web genres is the existence of an implicit *Logical Document Structure* (LDS) — analogously to textual units

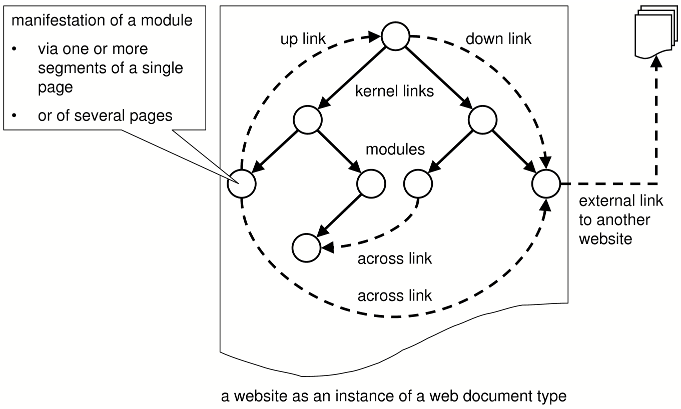
manifestation of a module
• via one or more segments of a single page
• or of several pages
up link
down link
kernel links
modules
across link
across link
external link to another website
a website as an instance of a web document type

Figure 1: Hyperlinks (denoted by arcs) of different types (Amitay et al., 2003; Eiron and McCurley, 2003) are connecting hypertext modules (denoted by circles), thereby forming a certain class of hypertext graphs with kernel hierarchical structure.

whose LDS (Power et al., 2003; Bateman et al., 2001) is described, for example, in terms of section, paragraph and sentence categories.

In case of instances of web genres we hypothesize their logical document structure to include four levels — we speak of types as abstract entities (e.g. the type of unit called *paragraph*) instantiated by concrete, observable events (e.g. a concrete paragraph unit); this terminology follows object oriented modeling where types or classes are distinguished from their (object) instances:

- The highest level is constituted by *web document networks* which consist of web documents serving different functions, manifested independently from each other by separate websites. The system of websites (e.g. personal academic home pages, project sites, library sites etc.) of a university contributing to the same corporate identity is an example of a web document network.

- *Web document types* classify the level of pragmatically closed acts of web-based communication, where each of these acts serves a complex function of, for example, *conference organization*, *personal presentation* or *online shopping.* Web document types correspond to *web genres* and are typically manifested by whole websites. They organize systems of dependent sub-functions each of which is seen to be manifested by a single instance of so called module types:

- *Module types* categorize functionally homogeneous units of web-based communication manifesting a single, but dependent function, e.g. *call for papers*, *program* or *conference venue* as sub-functions of the spanning function of *web-based conference organization.* Module types may be manifested by a single web page, its segments or even by segments of different web pages.

- On a lower level of structural resolution, *elementary building blocks* (e.g. logical *lists*, *tables*, *sections* and even paragraphs in the sense of the LDS of texts) occur as dependent parts of instances of module types.

	modules	documents	document networks
functionally homogeneous	+	−	−
functionally closed	−	+	−
thematically homogeneous	+	−	−
thematically closed	+/−	−	−

Table 1: Module types, document types and their networks distinguished by their functional and thematic homogeneity and independence (closeness), respectively.

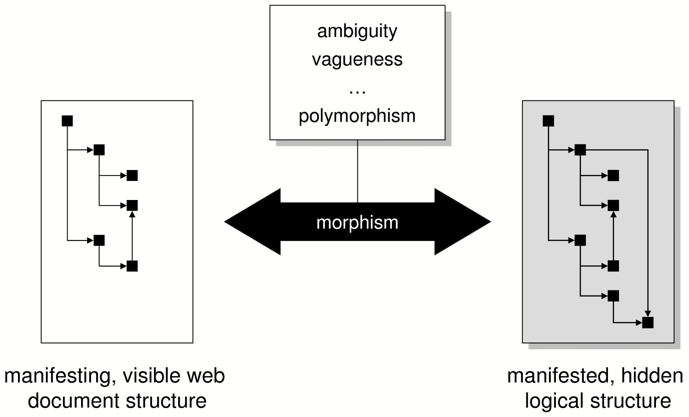
ambiguity
vagueness
...
polymorphism
morphism
manifesting, visible web document structure
manifested, hidden logical structure

Figure 2: Aspects of informational uncertainty.

Table (1) summarizes these terms utilizing the distinction of functional/thematic homogeneity and independence.

Web document types, module types and their building blocks are logical units manifested, for example, by websites, web pages, HTML tables and lists, respectively. Their logical status relates to their functional and thematic delimitation, which does not (necessarily) hold for the physical presentation units used to manifest them as web-based units. Thus, we view *web pages* as an analogue of (book) *pages* in the sense that they serve as physical, layout-based units for presenting functionally/thematically delimited units of web-based communication. That is, analogous to (book) pages, where thematically delimited paragraphs may occur on subsequent pages, functionally delimited units of web-based communication may be distributed over several web pages. Because of the nonlinear order and poly-sequentiality of hypertext units (Kuhlen, 1991; Storrer, 2002), this distribution may incorporate many more than only two pages and thus may be discontinuous.

A central hypothesis of this paper is that discontinuous manifestation relates to the instantiation of module types. In other words, the scope of distributed manifestation is delimited by the web document type to which the focal module type belongs. Analogously, polymorphism relates to the manifestation of several module types by the same presentation unit on web page level. These considerations have many implications regarding corpus-based analysis of web documents which can be summarized as follows:

- Web document types constitute the level on which polymorphism and discontinuous manifestations are most effective and therefore interfere any effort of directly applying the apparatus of classical text categorization to web-based units.

- Uncovering the LDS of websites contributes to breaking the many-to-many-relation of polymorphism and discontinuous manifestation. It aims to explicate which modules are manifested by which (segments of which) visible web pages of the same site and which links of which types – as distinguished in figure (1) – interlink these modules.

- As far as the units of this LDS are functionally delimited, it is indispensable to analyze websites as manifestations of web genres. In other words, in case of web-based units, structure analysis means *web genre analysis*.

- The central hypothesis of this paper is that hypertext categorization has to be reconstructed as a kind of *structure learning* focussing on prototypical, recurrent patterns of the LDS of websites *as instances of web genres* on the level of document types and their typing according to the functions their constitutive modules are intended to serve.

Polymorphism and discontinuous manifestation are just two aspects of informational uncertainty characterizing the relation of visible, manifesting web structure and hidden, manifested function or content structure. Figure (2) demonstrates this by listing, amongst others, vagueness and ambiguity as additional aspects of this informational uncertainty. In the following section, we concentrate on polymorphism.

3 Hypertext Representation

The structural analysis of hypertext resources demands for an adequate representational format which allows representing the different link structures hypertexts consist of. In case of web-based hypertexts, the raw material is usually based on `HTML` but may also consist of `PDF`-, `DOC`- or even plain text documents. In order to avoid the overhead of adjusting each analysis tool to these various input formats, a common representation is required which abstracts from the concrete realisation and focuses on accurately modeling the underlying hypertext structure. That is, what is needed is not just a technical representation *format*, but a data *model* which is based on corpus linguistic principles (e.g. interoperability, adequacy, readability and computational facility).

The mathematical model of directed graphs is widely used for hypertext representation (Botafogo et al., 1992). Graphs serve well for various applications and there exist numerous programming libraries to support the analysis of the structures they represent. They lack however the ability to model hyperlinks accurately. A URL, for example, allows addressing a certain anchor *within* a linked document. Thus, besides the fact that two pages are interlinked, the information about the corresponding anchors involved has to be preserved. This is not possible by means of digraphs since they only allow representing binary relations. *Hypergraphs* (Berge,

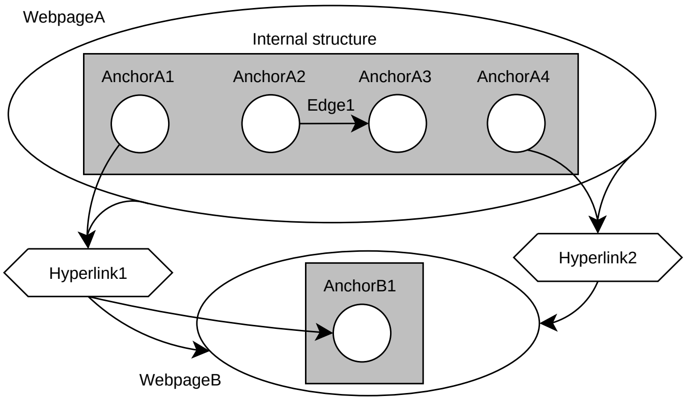

Figure 3: Three alternatives (Edge1, Hyperlink1, Hyperlink2) of `HTML` based linking.

1989) are a more general approach which allows representing the interconnection of arbitrary numbers of nodes by means of *hyperedges* which — in mathematical terms — correspond to heterogeneous relations of arbitrary arity. Another important aspect relates to accurately modeling substructures of hypertext graphs. An example is given by the interlinking of anchors of the *same* web page which can be considered as an embedded hypergraph. When dealing with web structures, these embeddings have to be represented, too.

The Graph eXchange Language (GXL) (Winter et al., 2002) is an XML based format for representing hypergraphs and thus can be utilized for hypertext representation (Mehler et al., 2004). It is designed as a common interchange format for applications that need to share complex graph structures. Elements of graphs can be attributed and typed. They also may contain an arbitrary number of substructures (e.g. a node can contain a graph) und thus allow the representation of graph embedding. The original resource (e.g. an `HTML`-document) can either be referenced by a `URL` or explicitly stored as an attribute. However the format is not designed to store large scale content but to explicate the underlying structuring. Figure (3) shows various ways of how two pages could be interlinked and how this information is modeled in the present framework (see table 2). Note that for each web page involved, anchors and page-internal hyperlinks are modeled as nodes and edges of an embedded graph.

For more information on this representation model see Mehler et al. (2004). In order to automatically extract and transform hypertext documents into the GXL we use the HyGraph system (Gleim, 2005) as a tool for preprocessing corpora of web documents. Starting from given resources (either extracted from within the system or by means of external tools) it automatically extracts the underlying hypertext structure, represents the information in GXL and adds various meta data. The resulting GXL documents are finally visualized and further explored.

```
<gxl xmlns:xlink="http://www.w3.org/1999/xlink">
 <graph hypergraph="true" edgemode="directed" id="Graph0">
  <node id="WebpageA">
   <graph edgemode="directed" hypergraph="false" id="GraphA">
    <node id="AnchorA1">...</node>
    <node id="AnchorA2">...</node>
    <node id="AnchorA3">...</node>
    <node id="AnchorA4">...</node>
    <edge id="Edge1" from="AnchorA2" to "AnchorA3"/>
   </graph>
  </node>
  <node id="WebpageB">
   <graph edgemode="directed" hypergraph="false" id="GraphB">
    <node id="AnchorB1">...</node>
   </graph>
  </node>
  <rel id="Hyperlink1">
   <relend direction="in" target="WebpageA" role="sourcepage"/>
   <relend direction="in" target="AnchorA1" role="sourceanchor"/>
   <relend direction="out" target="WebpageB" role="targetpage"/>
   <relend direction="out" target="AnchorB1" role="targetanchor"/>
  </rel>
  <rel id="Hyperlink2">...</rel>
 </graph>
</gxl>
```

Table 2: A GXL-based hypertext graph of a conference website.

4 Characteristics of Hypertext Graphs

In order to demonstrate the impact of polymorphism and discontinuous manifestation we analyze a corpus of conference websites in the area of computer and information science which are expected to serve recurrent, stable functions. The corpus is described in table (3). It collects conference websites entered via the conference calendars of ACM, IEEE and IFIP. All websites were automatically mapped onto their GXL representations. That is, each website was represented by a separate graph according to the graph model of section (2) and (3) which finally was input of computing quantitative graph characteristics.

The basic idea is that if polymorphism is effective in the focal area, website authors can chose among several alternatives to code the same functions. That is, we expect a broad spectrum of hypertext graphs being actually manifested by the websites of the corpus. This is (not only but also) in accordance with expecting skewed distributions of the graphs' characteristics, that is, distributions which have already been observed in the area of social network analysis (Newman, 2003). More specifically, we consider two questions:

- Does the distribution of the out degrees of the pages follow a power-law comparable to the distributions already observed in the Web? This question relates to the small world phenomenon (Watts and Strogatz, 1998).

Variable	Value
number of web sites	1,096
number of web pages	50,943
number of hyperlinks	303,278
maximum depth	23
maximum width	1,035
average size	46
average width	38
average height	3

Table 3: The sample corpus of 1,096 conference websites.

- Are the graphs uniformly composed or do they exploit the whole spectrum of possible graphs with kernel hierarchical structure?

In order to tackle the first question we fit a power law $p_0 + p_1 k^{-\alpha}$ to the distribution of empirically observed out degrees of the nodes (measured by the number of outgoing hyperlinks). This fitting is successful for $\alpha = 0.64782$ ($p_0 \approx 0$ and $p_1 \approx 3,000$). Constraining p_0, p_1 in order to get an α of 1.5 also leads to an acceptable fitting — a result which is in accordance with degree distributions observed in social and information networks (Newman, 2003). Thus, our data does at least not contradict those distributions which have already been investigated in the area of larger, more heterogeneous networks. In order to tackle the second question we fit three distributions: the distribution of the graphs' heights, widths and child imbalances (Botafogo et al., 1992). Power law fitting is now successful with the exponents 0.1, 0.42 and 0.47, respectively. Finally, we compute the correlations of these parameters and yield $\rho(\mathit{width}, \mathit{height}) = 0.98$, $\rho(\mathit{width}, \mathit{child\ imbalance}) = 0.76$ and $\rho(\mathit{height}, \mathit{child\ imbalance}) = 0.78$. These results support the following interpretation: On the one side, there is a small set of wide, high or imbalanced graphs which comprise short as well as long pathes starting from the root. On the other side, there is a large set of very small graphs with a couple or even only one node, i.e. academic homepages which code all information in a single page. At the same time, there is a smooth transition between both extremal cases as it is typical of the Zipfian order of social systems (Zipf, 1972).

These preliminary results confirm the hypothesis that actual practices of hypertext authoring as observable in the genre of conference websites result in skewed distributions of the characteristics of the corresponding hypertext graphs. Supposed that these websites serve more or less the same set of recurrent functions, this means that authors make use of many manifestation variants, but tend to concentrate the majority of functions on single and thus polymorphic web pages. This result is confirmed by an experiment in hypertext categorization which demonstrates deficits of the classical apparatus of text categorization when applied to web pages (Mehler et al., 2004).

5 A Two-Level Model of Structure-Sensitive Hypertext Categorization

A central implication of the impact of polymorphism and discontinuous manifestation is that, prior to hypertext categorization, the many-to-many relation of visible and hidden web structure has to be resolved at least with respect to the logical document structure (LDS). In other words: hypertext categorization is bound to a preliminary structural analysis. Insofar this analysis results in structured representations of web units, function learning as performed by text categorization is inappropriate to mining web genres. It unsystematically leads to multiple categorizations when directly applied to web units whose borders do not correlate with the functional or content-based categories under consideration. Rather, a sort of structure learning has to be performed, mapping these units onto representations of their LDS which only then are object to mining prototypical sub-structures of web genres. In this section, hypertext categorization is reconstructed along this line of argumentation. An algorithm is outlined which addresses structure learning from the point of view of prototypes of LDS. It is divided into two parts:

I. Logical Document Structure Learning Websites as supposed instances of web genres have first to be mapped onto representations of their LDS. That is polymorphism has to be resolved with respect to constituents of this structure level. This includes the following tasks:

- Visible segments of web pages have to be identified as manifestations of constituents of LDS.
- Visible hyperlinks have to be identified as manifestations of logical links, i.e. as kernel links or up, down, across or external links.
- Finally, functional equivalents of hyperlinks have to be identified as manifestations of logical links according to the same rules, i.e. of links without being manifested by hyperlinks.

Solving these tasks, each website is mapped onto a representation of its LDS based on the building blocks described in section (1). This means that websites whose visible surface structures differ dramatically may nevertheless be mapped onto similar LDS representations, and vice versa. So far, these intermediary representations lack any typing of their nodes, links and sub-structures in terms of functional categories of the focal web genre. This functional typing is addressed by the second part of the algorithm:

II. Functional Structure Learning The representations of LDS are input to an algorithm of *computing with graphs* which includes four steps:

1. *Input:* As input we use a corpus $C = \{G_i \,|\, i \in I\}$ of labeled *typed directed graphs* $G_i = (V, E, k(G_i), \tau)$ with kernel hierarchical structure modeled by

an ordered rooted tree $k(G_i) = (V, D, x, \mathcal{O})$ with root x and order relation $\mathcal{O} \subseteq D^2$, $D \subseteq E$. Typing of edges $e \in E$ is done by a function $\tau : E \to T$ where T is a set of type labels. In case of websites, vertices $v \in V$ are labeled as either accessible or unaccessible web pages or resources and edges are typed as kernel, across, up, down, internal, external or broken links. In case of logical hypertext document structure, vertices are logical modules whereas the set of labels of edge types remains the same.

2. *Graph similarity measuring:* The corpus C of graphs is input to a similarity measure $s\colon C^2 \to [0, 1]$ used to built a similarity matrix (Bock, 1974) $\vec{S} = (s_{kj})$ where s_{kj} is the similarity score of the pairing $G_i, G_j \in C$. s has to be sensitive to the graphs' kernel hierarchical structure as well as to the labels of their vertices and the types of their edges. We utilize the measure of Dehmer et al. (2004) which was successfully applied in the area of large graphs (Emmert-Streib et al., 2005).

3. *Graph clustering:* Next, the similarity matrix is input to clustering, that is, to unsupervised learning without presetting the number of classes or categories to be learned. More specifically, we utilize hierarchical agglomerative clustering (Bock, 1974) based on average linkage with subsequent partitioning. This partitioning refers to a lower bound (Rieger, 1989) $\theta = \bar{\eta} + \frac{1}{2}\sigma$, where $\bar{\eta}$ is the mean and σ the standard deviation of the absolute value of the differences of the similarity levels of consecutive agglomeration steps. This gives a threshold for selecting an agglomeration step for dendrogram partitioning whose similarity distance to the preceding step is greater than θ. We use the first step exceeding θ.

4. *Graph prototyping:* Next, for each cluster $X = \{G_{i_1}, \ldots, G_{i_n}\} \subseteq C$ of the output partitioning of step (3) a graph median $\hat{G}$ has to be computed according to the approach of Bunke et al. (2001):

$$\hat{G} = \arg\max_{G \in X} \frac{1}{n} \sum_{k}^{|X|} s(G, G_{i_k}) \tag{1}$$

   The basic idea of applying this formula is to use $\hat{G}$ as a prototype of the cluster X in the sense that it prototypically represents the structuring of all members of that set of graphs.

5. *Graph extraction:* The last step is to use the prototypes $\hat{G}$ as kernels of instance based learning. More specifically, the prototype graphs can be used as templates to extract sub-structures in new input graphs. The idea is to identify inside these graphs recurrent patterns and thus candidates of functional categories of the focal genre (e.g. *paper submission* or *conference venue* graphs in case of the genre of *conference websites*).

It is this last step which addresses the final categorization by using *structured categories in order to categorize sub-structures of the input graphs.* It replaces the mapping of visible segments of web units onto predefined categories by mapping sub-structures of the hidden LDS onto clusters of homogeneously structured instances of certain module types of the focal web genre. The basic idea for solving the problem of acquiring genre specific corpora is to use this two-level algorithm in order to derive structural profiles for different genres.

6 Conclusion

This paper argued that the structure of websites is an uncertain manifestation of their hidden logical document structure and thus does not allow immediately utilizing web mining for the task of acquiring web genre specific corpora. Rather, a prerequisite of hypertext categorization is the reconstruction of the logical document structure of web documents. A corpus analysis has been performed in order to indirectly show the impact of polymorphism and discontinuous manifestations. In order to find a solution to this problem, hypertext categorization has been reconstructed by means of an algorithm which reflects the difference of visible and hidden structure and utilizes the paradigm of structure instead of function learning. Future work aims at evaluating this algorithm.

References

Amitay, E., D. Carmel, A. Darlow, R. Lempel, and A. Soffer (2003) The connectivity sonar: detecting site functionality by structural patterns. In *Proc. of the 14th ACM conference on Hypertext and Hypermedia*, pp. 38–47.

Bateman, J. A., T. Kamps, J. Kleinz, and K. Reichenberger (2001) Towards constructive text, diagram, and layout generation for information presentation. *Computational Linguistics 27*(3), 409–449.

Berge, C. (1989) *Hypergraphs: Combinatorics of Finite Sets.* Amsterdam: North Holland.

Biber, D., S. Conrad, and R. Reppen (1998) *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Bock, H. H. (1974) *Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten.* Göttingen: Vandenhoeck & Ruprecht.

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

Bunke, H., S. Günter, and X. Jiang (2001) Towards bridging the gap between statistical and structural pattern recognition: Two new concepts in graph matching. In *Proceedings of the Second International Conference on Advances in Pattern Recognition*, Berlin, pp. 1-11. Springer.

Chakrabarti, S., B. Dom, and P. Indyk (1998) Enhanced hypertext categorization using hyperlinks. In L. Haas and A. Tiwary (Eds.), *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 307–318. ACM.

Dehmer, M., R. Gleim, and A. Mehler (2004) A new method of similarity measuring for a specific class of directed graphs. *Submitted to Tatra Mountain Journal*.

Eiron, N. and K. S. McCurley (2003) Untangling compound documents on the web. In *Proceedings of the 14th ACM conference on Hypertext and hypermedia, Nottingham, UK*, pp. 85–94.

Emmert-Streib, F., M. Dehmer, and J. Kilian (2005) Classification of large graphs by a local tree decomposition. In *Proceedings of DMIN'05, International Conference on Data Mining, Las Vegas, Juni 20-23*.

Fürnkranz, J. (1999) Exploiting structural information for text classification on the WWW. In *Proceedings of the Third International Symposium on Advances in Intelligent Data Analysis*, Berlin, pp. 487-498. Springer.

Fürnkranz, J. (2002) Hyperlink ensembles: a case study in hypertext classification. *Information Fusion 3*(4), 299–312.

Gleim, R. (2005) HyGraph: Ein Framework zur Extraktion, Repräsentation und Analyse webbasierter Hypertexte. In *Proc. of GLDV '05*, pp. 42–53.

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

Kilgarriff, A. and G. Grefenstette (2003) Introduction to the special issue on the web as corpus. *Computational Linguistics 29*(3), 333–347.

Kosala, R. and H. Blockeel (2000) Web mining research: A survey. *SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining 2*(1), 1–15.

Kuhlen, R. (1991) *Hypertext: ein nichtlineares Medium zwischen Buch und Wissensbank*. Berlin: Springer.

Mehler, A., M. Dehmer, and R. Gleim (2004) Towards logical hypertext structure — a graph-theoretic perspective. In T. Böhme and G. Heyer (Eds.), *Proceedings of the Fourth International Workshop on Innovative Internet Computing Systems (I2CS '04)*, Lecture Notes in Computer Science, Berlin. Springer.

Mizuuchi, Y. and K. Tajima (1999) Finding context paths for web pages. In *Proceedings of the 10th ACM Conference on Hypertext and Hypermedia*, pp. 13–22.

Newman, M. E. J. (2003) The structure and function of complex networks. *SIAM Review 45*, 167–256.

Power, R., D. Scott, and N. Bouayad-Agha (2003) Document structure. *Computational Linguistics 29*(2), 211–260.

Rehm, G. (2004) Ontologie-basierte Hypertextsorten-Klassifikation. In A. Mehler and H. Lobin (Eds.), *Automatische Textanalyse: Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte*, pp. 121–137. Wiesbaden: VSS.

Rieger, B. (1989) *Unscharfe Semantik*. Frankfurt a.M.: Peter Lang.

Storrer, A. (2002) Coherence in text and hypertext. *Document Design 3*(2), 156–168.

Watts, D. J. and S. H. Strogatz (1998) Collective dynamics of 'small-world' networks. *Nature 393*, 440–442.

Winter, A., B. Kullbach, and V. Riedinger (2002) An overview of the GXL graph exchange language. In S. Diehl (Ed.), *Software Visualization*, pp. 324–336. Berlin: Springer.

Yang, Y., S. Slattery, and R. Ghani (2002) A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems 18*(2-3), 219–241.

Yoshioka, T. and G. Herman (2000) Coordinating information using genres. Technical report, Massachusetts Institute of Technology — Sloan School of Management: Center for Coordination Science.

Zipf, G. K. (1972) *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology.* New York: Hafner Publishing Company.