Towards Logical Hypertext Structure
A Graph-Theoretic Perspective

Alexander Mehler\(^1\), Matthias Dehmer\(^2\), and Rüdiger Gleim\(^2\)

\(^1\) Universität Bielefeld, D-33501 Bielefeld, Germany
\texttt{Alexander.Mehler@uni-bielefeld.de}
\(^2\) Technische Universität Darmstadt, D-64289 Darmstadt, Germany
\texttt{\{dehmer,gleim\}@informatik.tu-darmstadt.de}

Abstract. Facing the retrieval problem according to the overwhelming set of documents online the adaptation of text categorization to web units has recently been pushed. The aim is to utilize categories of web sites and pages as an additional retrieval criterion. In this context, the bag-of-words model has been utilized just as HTML tags and link structures. In spite of promising results this adaptation stays in the framework of IR specific models since it neglects the content-based structuring inherent to hypertext units. This paper approaches hypertext modelling from the perspective of graph-theory. It presents an XML-based format for representing websites as hypergraphs. These hypergraphs are used to shed light on the relation of hypertext structure types and their web-based instances. We place emphasis on two characteristics of this relation: In terms of realizational ambiguity we speak of functional equivalents to the manifestation of the same structure type. In terms of polymorphism we speak of a single web unit which manifests different structure types. It is shown that polymorphism is a prevalent characteristic of web-based units. This is done by means of a categorization experiment which analyses a corpus of hypergraphs representing the structure and content of pages of conference websites. On this background we plead for a revision of text representation models by means of hypergraphs which are sensitive to the manifold structuring of web documents.

1 Introduction

Text representation is a preliminary task of any approach to automatic text analysis. Among other things, this relates to the summarization, categorization, and mining of textual units. Analogously, hypertext representation is fundamental to automatic hypertext analysis \cite{31}. This comprises, for example, the identification of patterns (e.g. compound documents \cite{10}, or small worlds in WWW graphs \cite{11}), the categorization of links \cite{3,27} or the retrieval of information from large hypertext bases \cite{2}. In this context, the bag of words model of IR has been utilized as a starting point for hypertext representation just as HTML metadata, tags and link structures. That is, hypertexts are represented as vectors of features reflecting wording or markup as found in the hypertexts to be analyzed.
In spite of promising results this approach stays in the framework of text representation as elaborated in IR. Particularly with regard to categorization the predominance of traditional representation models is evident: Categorization is designed as an assignment of predefined category labels to feature vectors without the preceding exploration of hypertext structures (e.g. compound nodes or paths).

In this paper we plead for an integrative view of graph-theoretical analysis and categorization. Our starting point is a system of hypertext structure types and their nondeterministic manifestation by web-based units. We propose a four-layer model of hypertext structure types and focus on the many-to-many relation to its instances as units of Internet-based communication. We place emphasis on two characteristics of this relation: In terms of realizational ambiguity we speak of functional equivalents to the manifestation of the same structure type. Conversely, in terms of polymorphism the phenomenon is addressed that the same hypertext unit may manifest different structure types. Polymorphism occurs when, for example, the same page provides information about different topics (e.g. a page as part of an academic’s homepage lists courses beneath her biographical information) or serves different functions (e.g. a page offers the registration form of a research group beneath its brief description).

Our central hypothesis is that (comparable to natural language texts) realizational ambiguity and polymorphism are prevalent characteristics of web-based units. This has fundamental implications for hypertext categorization which normally presupposes to result in a non-overlapping separation of the object space, i.e. into an assignment of at most one category per object. If polymorphism is prevalent in this area, it does not make sense to view hypertext categorization as a process of disambiguating category assignments. As a consequence, two implications have to be balanced: Either the category system has to be revised, or – and this is our central thesis – the object space has to undergo a structural analysis as the result of which categorization and segmentation of the focal objects occurs. Since we view polymorphism to be a characteristic of web-based hypertexts, we expect multiple, interdependent categorizations to occur regularly. In other words: Proper hypertext categorization is bound to a preliminary structure analysis in which the regular realizational ambiguity and polymorphism of hypertext units is resolved. In order to support this line of argumentation we present a categorization of web pages of an area which is supposed to follow more stable authoring patterns and thus to be a profitable field of categorization: conference websites. This analysis operates on an XML-based representation format of hypertexts whose presentation is the second central focus of this paper. It is based on the idea to represent web-based units, their content and links as attributed typed directed nested hypergraphs [5].

The paper is organized as follows: After an outline of related work, our conceptual framework is presented in section (3): a four-layer model of hypertext structure types. This framework is used as the background of an XML-based format for representing web-based hypertexts as hypergraphs. The basic idea is to combine data-oriented representations of link structure, wording and markup