

# Unifying Content and Structure Learning: A Model of Semi-Supervised Hypertext Zoning

Alexander Mehler & Armin Wegner

Fakultät für Linguistik und Literaturwissenschaft

Universität Bielefeld, Universitätsstraße 25, D-33615 Bielefeld

{Alexander.Mehler, Armin.Wegner}@uni-bielefeld.de

The majority of approaches to content and structure mining rely on the vector space model or on some of its variants as, for example, latent semantic analysis or self-organizing feature maps. Vector-based representation models are used in topic detection & tracking by either preset or explored topic categories as well as in text classification by functional (hyper-)text types or (web)genres. That is, input documents are mapped onto feature vectors which in most cases are insensitive to their structure. However, these models are efficient in terms of their time complexity while their space complexity goes much beyond that of small-world graphs. On the other hand, approaches to structure learning suffer from the time complexity of measuring similarities of graphs. Moreover, graph similarity and related approaches tend to lack similarity models of document *content*, but focus on document *structure*. Thus, there is a trade-off between time and space complexity of document representation on the one hand and the integration of content & structure modeling on the other hand. This holds all the more for *hypertext* documents which go beyond tree-like structures of textual units as their hyperlinks induce graphs up to the complexity of large networks.

This paper presents a semi-supervised model of content and structure learning from web documents. It presents a unified model for exploring similarities of web documents in terms of their content and structure and, thus, tackles the latter trade-off. This is of interest to all those approaches which utilize similarity measures as, for example, exemplar-based methods. Moreover, the paper focuses on a central deficit of supervised approaches to web mining. This relates to the question about the thematic and structural identity of web documents as instances of certain hypertext types. That is, other than the majority of approaches to web mining we do not generally rely on preset training corpora of well-defined and delimited web documents. Rather, we make this delimitation an object of learning itself. As a consequence, we present an algorithm of *hypertext zoning* which explores the limits of instances of a given hypertext type. Because of the semantic implicitness of web document structures this is an indispensable task in order to build large corpora of web documents for the different fields of web mining facing the thematic and structural dynamics of web-based communication.