Summary

of the dissertation "On Information Need and Categorizing Search" by S. Meyer zu Eißen

Information retrieval (IR) is a discipline that deals with the task of satisfying a person's information need with the help of a computer. IR systems let a user specify an information need, which is evaluated against large collections of digital documents. Techniques for satisfying information needs have meanwhile become ubiquitous, be it in the form of search engines at home or at work, on mobile devices or on workstations, or as retrieval components in file systems, document repositories, databases, or knowledge management tools. The reason for this pervasiveness is the growing information need, the diversity of retrieval tasks, and the desired degree of personalization. The thesis develops new concepts and new algorithms in various aspects throughout the information retrieval process, with a special focus on automatic categorization. In particular, the following points are addressed.

Document Representation. We propose the suffix tree document model along with new similarity measures for quantifying document similarity. In contrast to traditional methods, the suffix tree model is able to quantify both aspects in parallel, term matches as well as term order matches. Experiments show that our document representation improves document categorization performance compared to traditional vector space representations.

Cluster Validity. The new cluster validity index $\overline{\rho}$ for document clustering is proposed. $\overline{\rho}$ allows us to identify among a set of clusterings those which have been generated with adequate parameters, i.e. those that reflect the human idea of categorization. An experimental evaluation shows that $\overline{\rho}$ delivers reliable results in comparison to existing approaches in document categorization scenarios.

Topic Identification. When a categorization according to topic is determined using an unsupervised approach, it has to be presented to a user. In particular, the categories have to be labeled with characteristic terms for browsing. Desired properties for category labels are introduced and formalized, and the new WCC algorithm to compute cluster labelings is proposed.

Genre Categorization. In recent IR research, the term "categorization" is associated with an organization of documents according to *topic*. However, we will show that the *genre* of a document is a very useful categorization criterion when searching large document repositories. We propose a genre categorization scheme for Web documents, and we introduce a novel document representation that can be employed to classify documents according to genre. A feasibility study shows that that genre categorization is possible, even in a large and heterogeneous collection like the World Wide Web.

Software Engineering. A Model Driven Architecture (MDA) approach for composing and executing IR processes is proposed. In contrast to the commonly used library-based IR process design, our approach called TIRA clearly separates the specification of an IR process from its operationalization. TIRA allows to tailor IR processes with respect to personal preferences.

All concepts and algorithms have been operationalized, among others within AIsearch, a search engine that has been awarded at the EASA-2004 (European Academic Software Award) competition.