

# On Information Need and Categorizing Search

Sven Meyer zu Eißen

## 1 Scope

This dissertation is located in the area of information retrieval (IR), 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 thesis develops new concepts and algorithms in various aspects throughout the information retrieval process, with a special focus on automatic document categorization.

## 2 Contributions

Figure 1 outlines the core components of Asearch, our meta search engine that operationalizes the research presented in the thesis. In particular, Asearch categorizes Web search results according to topic and genre with the help of tailored document representations and new algorithms in the fields of clustering, topic identification, and genre classification as outlined in the following points.

(1) *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 approaches.

(2) *Cluster Validity.* The new cluster validity index  $\bar{p}$  for document clustering is proposed.  $\bar{p}$  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  $\bar{p}$  delivers reliable results in comparison to existing approaches in document categorization scenarios.

(3) *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 WCC algorithm to compute cluster labelings is proposed.

(4) *Genre Categorization.* In recent IR research, the term "categorization" is associated with an organization of documents according to topic. However, we show that the genre of a document is a very useful categorization criterion when

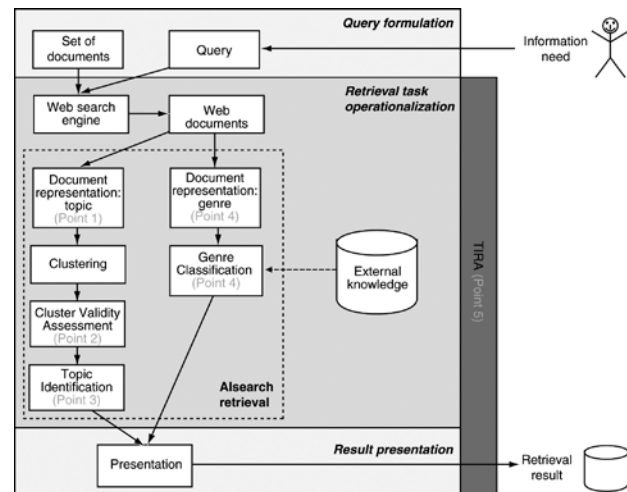


Figure 1: Overview of Asearch and the core contributions.

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.

(5) *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.

### Contact

Dr. Sven Meyer zu Eißen  
 Bauhaus-Universität Weimar  
 Fakultät Medien/Mediensysteme, 99421 Weimar  
 Tel.: +49 (0)3643-583720  
 Email: sven.meyer-zu-eissen@medien.uni-weimar.de  
 The dissertation can be downloaded at  
[http://ubdata.uni-paderborn.de/ediss/17/2007/meyer\\_zu](http://ubdata.uni-paderborn.de/ediss/17/2007/meyer_zu)



Sven Meyer zu Eißen studied computer science at the University of Paderborn and joined the working group "Knowledge-Based Systems" afterwards. Since 2005 he is member of the Web Technology and Information Systems group at the Bauhaus University Weimar. His research interests include Web technology, information retrieval, and machine learning.