Classifying Web Sites

Lars Littig

Abstract

This thesis provides insights into and approaches for the classification of Web sites. Since the World Wide Web continues to grow at enormous speed, its heterogeneity and lack of structure increase as well. As a consequence, it is more and more difficult to identify the Web sites providing the information and services of interest. However, the automated classification of Web sites at different levels of granularity not only substantially increases the capability of today's search engines but is also important for the development of the future Web.

In the field of coarse-grained Web site classification, the contributions of this thesis are three-fold. Firstly, a rich set of effective features for the classification of Web sites is derived from their content and structure focusing on size, organization, composition of URLs, technical realization, and link structure. Secondly, these features are analyzed in a large measurement-based study to reveal the relation between structure and purpose of a Web site as well as the strengths and weaknesses of Web site classification solely based on structure or on content. Thirdly, this thesis proposes an approach for the classification of Web sites that combines structure and content in order to benefit from the advantages of both classification schemes. The effectiveness of this approach is evaluated on a dataset comprising more than 16,000 Web sites with about 20 million Web pages. It achieves a significant improvement of classification performance for the coarse-grained classification of Web sites into eight major classes of the Web.

In the field of fine-grained Web site classification, this thesis presents an effective approach for the classification of corporate Web sites into a taxonomy of 40 closely related business classes. The development of this approach involves a detailed analysis of the effects of data pre-processing. Therefore, the issues of structure- and content-based data pre-processing are stressed and a thesaurus-based denoising technique is employed. In addition to this, a methodology for the pre-classification of Web pages without the need for training data at page level is presented. Another contribution constitutes the proposal of a pruning strategy that is not restricted to the level of a Web page within the page tree but to its relevance for classification. Furthermore, additional features from both structure and content of Web sites are derived to deal with classes that exhibit only minor differences. As illustrated by a comprehensive performance study, which is based upon a dataset of about 12,000 corporate Web sites comprising more than 2.4 million Web pages, the narrow focus on the early stages of the classification process yields excellent classification accuracy.