1. Enhanced information retrieval by using HTML tags

Werner, Lars (1); Böttcher, Stefan (2); Beckmann, Ralph (3)
Source: *Proceedings of the 2005 International Conference on Data Mining, DMIN'05*, p 24-29, 2005, *Proceedings of the 2005 International Conference on Data Mining, DMIN'05*; **ISBN-13:** 9781932415797; **Conference:** 2005 International Conference on Data Mining, DMIN'05, June 20, 2005 - June 23, 2005; **Publisher:** CSREA Press
Author affiliation: (1) University of Paderborn, C-LAB, Paderborn, Germany (2) University of Paderborn, Computer Science Paderborn, Germany (3) SETIS Informatik and Consulting GmbH, Darmstadt, Germany
Abstract: Whenever digital libraries or knowledge management systems are to be automatically filled with web pages from the internet, document classification of the web pages is one of the major challenges. We present an approach which uses HTML tags in order to improve the quality of the hypertext document classification. Our approach uses weighting of HTML tags for separating relevant information in hypertext documents from the noise. We have evaluated our approach on the basis of a document classification algorithm. The results show that our weighting approach yields a classification which is approximately 35% better than a classification without the use of the HTML tagging information. (11 refs)
Main heading: Classification (of information)
Controlled terms: Data mining - Digital libraries - HTML - Hypertext systems - Information use - Knowledge based systems - Knowledge management - Websites
Uncontrolled terms: Document Classification - HTML tags - Hypertext documents - Knowledge management system - Text classification - Typographical information - Weighting approaches
Classification Code: 723 Computer Software, Data Handling and Applications - 903 Information Science
Database: Compendex

Data Provider: Engineering Village