World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:06, 2015

An Automatic Bayesian Classification System for File Format Selection

Authors : Roman Graf, Sergiu Gordea, Heather M. Ryan

Abstract : This paper presents an approach for the classification of an unstructured format description for identification of file formats. The main contribution of this work is the employment of data mining techniques to support file format selection with just the unstructured text description that comprises the most important format features for a particular organisation. Subsequently, the file format indentification method employs file format classifier and associated configurations to support digital preservation experts with an estimation of required file format. Our goal is to make use of a format specification knowledge base aggregated from a different Web sources in order to select file format for a particular institution. Using the naive Bayes method, the decision support system recommends to an expert, the file format for his institution. The proposed methods facilitate the selection of file format and the quality of a digital preservation process. The presented approach is meant to facilitate decision making for the preservation of digital content in libraries and archives using domain expert knowledge and specifications of file formats. To facilitate decision-making, the aggregated information about the file formats is presented as a file format vocabulary that comprises most common terms that are characteristic for all researched formats. The goal is to suggest a particular file format based on this vocabulary for analysis by an expert. The sample file format calculation and the calculation results including probabilities are presented in the evaluation section.

Keywords : data mining, digital libraries, digital preservation, file format
Conference Title : ICSDE 2015 : International Conference on Software and Data Engineering
Conference Location : Copenhagen, Denmark
Conference Dates : June 11-12, 2015