Abstract

Web page classification is achieved using text classification techniques. Web page classification is different from traditional text classification due to additional information, provided by web page structure which provides much information on content importance. HTML tags provide visual web page representation and can be considered a parameter to highlight content importance. Textual keywords are base on which Information retrieval systems rely to index and retrieve documents. Keyword-based retrieval returns inaccurate/incomplete results when differing keywords describe the same document and queries concept. Concept-based retrieval tried to tackle this by using manual thesauri with term co-occurrence data, or by extracting latent word relationships and concepts from a corpus. Semantic search motivates Semantic Web from inception for classification and retrieval processes. In this paper, a model for the exploitation of semantic-based feature selection is proposed to improve search and retrieval of web pages over large document repositories. The features are classified using Support Vector Machine (SVM) using different kernels. The experimental results show improved precision and recall with the proposed method with respect to keyword-based search.
technical report, 14.

Index Terms

Computer Science   Information Sciences

Keywords
Web Mining   Feature extraction   Inverse document frequency   HTML Tag   Support Vector Machines