Probabilistic Web Content Analysis (Analyse Probabiliste du Contenu de Pages Web)

Résumé du projet de recherche (Langue 1)

An automatic identification of meaningful content sections on web pages, such as titles, paragraphs, advertisements, product images or user comments, facilitates a large number of applications, ranging from speech rendering for the visually impaired over contextual advertisement to structured web search. Ultimately, such an identification always necessitates both, a partitioning of the content and a classification of the resulting partitions into a number of application-dependent semantic categories. We hence propose to approach the analysis of web content in an interdependent classification framework, integrating semantic coherence, just as in segmentation, via interaction features which describe the semantic configuration of two or more semantically atomic content regions. One of the major obstacles to gaining meaningful access to web contents is their semantically inappropriate organisation and markup. As a consequence, it generally is impossible to characterise an interesting content region with certainty. In this thesis, we propose to treat the uncertainties arising in an analysis of web content in a coherent probabilistic framework, the Bayesian diagnostic paradigm, and attempt to illuminate the conditions under which some probability model might be justified, deriving its form of representation from assumptions about observable quantities such as region features and semantics, utilising the concepts of exchangeability, conditional independence and sufficiency. In particular, we examine different Markovian dependencies between the semantic content categories within individual web pages and discuss how to take into account the structure that exists between pages and sites. We equally present an informal feature analysis which elucidates the manifold information available in the content, structure and style of a web page. Such an analysis is a quintessential prerequisite to both formal probabilistic modelling and high predictive performance. Furthermore, we introduce a new, publicly available data set of 604 real-world news web pages from 177 distinct sites with accurate annotations based on 30 distinct semantic categories, termed the NEWS600 corpus. Finally, we conduct a series of experiments on the NEWS600 corpus to empirically compare a number of different approaches for web news content classification. It demonstrates that even relatively simple models in our framework achieve significantly better results than the current state of the art.