Abstract

The World Wide Web is one of the most popular and quickly growing aspects of the Internet. Ways in which computer scientists attempt to estimate its size vary from making educated guesses, to performing extensive analyses on search engine databases. We present a new way of measuring the size of the World Wide Web using "Quadrat Counts", a technique used by biologists for population sampling. There has been an exponential growth in hypermedia and web modeling languages in the market. This growth has highlighted new problems and new areas of research. This paper categorizes and reviews the main hypermedia and web modeling languages showing their origin and their primary focus. It then concludes
with recommendations for further research in this field. When automatically extracting information from the world wide web, most established methods focus on spotting single HTML documents. However, the problem of spotting complete web sites is not handled adequately yet, in spite of its importance for various applications. Therefore, this paper discusses the classification of complete web sites. First, we point out the main differences to page classification by discussing a very intuitive approach and its weaknesses. This approach treats a web site as one large HTML-document and applies the well-known methods for page classification. Next, we show how accuracy can be improved by employing a preprocessing step which assigns an occurring web page to its most likely topic. The determined topics now represent the information the web site contains and can be used to classify it more accurately. We accomplish this by following two directions. First, we apply well established classification algorithms to a feature space of occurring topics. The second direction treats a site as a tree of occurring topics and uses a Markov tree model for further classification. To improve the efficiency of this approach, we additionally introduce a powerful pruning method reducing the number of considered web pages. Our experiments show the superiority of the Markov tree approach regarding classification accuracy. In particular, we demonstrate that the use of our pruning method not only reduces the processing time, but also improves the classification accuracy.

References


Index Terms

Computer Science

Advances In Computer Application
Keywords
Www Size Estimation Using Biological Techniques Modelling Methodologies Web Application
Web Site
Hypermedia
Hypertext