Chapter 5.15
Feature Selection for Web Page Classification

K. Selvakuberan
Tata Consultancy Services, India

M. Indra Devi
Thiagarajar College of Engineering, India

R. Rajaram
Thiagarajar College of Engineering, India

ABSTRACT

The World Wide Web serves as a huge, widely distributed, global information service center for news, advertisements, customer information, financial management, education, government, e-commerce and many others. The Web contains a rich and dynamic collection of hyperlink information. The Web page access and usage information provide rich sources for data mining. Web pages are classified based on the content and/or contextual information embedded in them. As the Web pages contain many irrelevant, infrequent, and stop words that reduce the performance of the classifier, selecting relevant representative features from the Web page is the essential preprocessing step. This provides secured accessing of the required information. The Web access and usage information can be mined to predict the authentication of the user accessing the Web page. This information may be used to personalize the information needed for the users and to preserve the privacy of the users by hiding the personal details. The issue lies in selecting the features which represent the Web pages and processing the details of the user needed the details. In this article we focus on the feature selection, issues in feature selections, and the most important feature selection techniques described and used by researchers.

INTRODUCTION

There are an estimated 15 to 30 billion pages available in the World Wide Web with millions of pages being added daily. Describing and organizing this vast amount of content is essential for realizing the web's full potential as an information resource. Automatic classification of web pages is needed for the following reasons. (a) Large amount of information available in the internet makes it difficult for

DOI: 10.4018/978-1-60566-196-4.ch012

the human experts to classify them manually (b) The amount of Expertise needed is high (c) Web pages are dynamic and volatile in nature (e) More time and effort are required for classification. (f) Same type of classification scheme may not be applied to all pages (g) More experts needed for classification. Web page classification techniques use concepts from many fields like Information filtering and retrieval, Artificial Intelligence, Text mining, Machine learning techniques and so on. Information filtering and retrieval techniques usually build either a thesauri or indices by analyzing a corpus of already classified texts with specific algorithms. When new text is to be classified, thesaurus and index are used to find the similarity with already existing classification scheme to be associated with this new text.

Until the late 1980s, the most effective approach to web page classification seemed to be that of manually by building classification systems by means of knowledge-engineering techniques, i.e. manually defining a set of logical rules that encode expert knowledge on how to classify web page documents under the given set of categories. In the 1990s this perspective has been overturn, and the machine learning paradigm to automated web page classification has emerged and definitely superseded the knowledge-engineering approach. Within the machine learning paradigm, a general inductive process automatically builds an automatic text classifier by "learning", from a set of previously classified web documents, the characteristics of the categories of interest. The advantages of this approach are accuracy comparable to human performance and a considerable savings in terms of expert manpower, since no intervention from either knowledge engineers or domain experts is needed. Currently web page categorization may be seen as the meeting point of machine learning and information retrieval. As Machine Learning aims to address larger, more complex tasks, the problem of focusing on the most relevant information in a potentially overwhelming quantity of data has become increasingly important. For instance, data mining of corporate or scientific records often involves dealing with both many features and many examples, and the internet and World Wide Web have put a huge volume of low-quality information at the easy access of a learning system. Similar issues arise in the personalization of filtering systems for information retrieval, electronic mail, net news.

The main objective of this chapter is to focus on the feature selection techniques, need for feature selection, their issues in web page classification, feature selection for privacy preserving data mining and the future trends in feature selection.

LITERATURE SURVEY

Rudy Setiono and Huan Liu (1997) proposed that Discretization can turn numeric attributes into discrete ones. χ^2 is a simple algorithm. Principal Component Analysis-compose a small number of new features. It is improved from simple methods such as equi-width and equal frequency intervals. For each and every attributes calculate the χ^2 value for each and every interval. Combine the lowest interval values while approximation.

Shounak Roychowdhury (2001) proposed a technique called granular computing for processing and expressing chunks of information called granules. It reduces hypothesis search space, to reduce storage. Fuzzy set based feature elimination techniques in which subset generation and subset evaluation are employed. For optimal feature selection brute force technique is employed.

Catherine Blake and Wander Pratt (2001) suggested the relationship between the features used to represent the text and the quality model. A comparison of association rules based on three different concepts: words, manually assigned keywords, automatically assigned concepts are made. Bidirectional association rules on concepts or keywords are useful than the words used. Each individual feature should be informative. The quality of features should be meaningful. The