Chapter VIII
Classification of Web Pages Using Machine Learning Techniques

K. Selvakuberan
Innovation Labs (Web 2.0), TATA Consultancy Services, India

M. Indra Devi
Thiagarajar College of Engineering, Madurai, India

R. Rajaram
Thiagarajar College of Engineering, Madurai, India

ABSTRACT

The explosive growth of the Web makes it a very useful information resource to all types of users. Today, everyone accesses the Internet for various purposes and retrieving the required information within the stipulated time is the major demand from users. Also, the Internet provides millions of Web pages for each and every search term. Getting interesting and required results from the Web becomes very difficult and turning the classification of Web pages into relevant categories is the current research topic. Web page classification is the current research problem that focuses on classifying the documents into different categories, which are used by search engines for producing the result. In this chapter we focus on different machine learning techniques and how Web pages can be classified using these machine learning techniques. The automatic classification of Web pages using machine learning techniques is the most efficient way used by search engines to provide accurate results to the users. Machine learning classifiers may also be trained to preserve the personal details from unauthenticated users and for privacy preserving data mining.

INTRODUCTION

Over the past decade we have witnessed an explosive growth on the Internet, with millions of web pages on every topic easily accessible through the Web. The Internet is a powerful medium for communication between computers and for accessing online documents all over the world but it is not a tool for locating or organizing the mass of information. Tools like search engines assist users in locating information on the Internet. They perform excellently in locating but provide limited ability in organizing the web pages. Internet users are now confronted with thousands of web pages returned by a search engine using simple keyword search. Searching through those web pages is in itself becoming impossible for users. Thus it has been of more interest in tools that can help make a relevant and quick selection of information that we are seeking. There is also estimation that 15 to 30 billion pages are accessible on 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. Accomplishing this in a meaningful way will require consistent use of metadata and other descriptive data structures such as semantic linking. We find that HTML meta tags are a good source of text features, but are not in wide use despite their role in search engine rankings. But most of the pages will not contain meta data. (John Pierre, M., 2001, Tom Mitchell, M., 1999, Sebastiani, F.,2002)

Automatic Web-page classification by using hypertext is a major approach to categorizing large quantities of Web pages. Two major kinds of approaches have been studied for Web-page classification: content-based and context-based approaches. Typical content-based classification methods utilize words or phrases of a target document to train the classifier. Context based approaches take into account the structure of the HTML pages to train the classifier. This is because sometimes a Web page contains no obvious clues textually for its category. For example, some pages contain only images and little text information. By exploiting the hyper textual information, context-based approaches additionally exploit the relationships between the Web pages to build a classifier. Web page classifiers trained both using content and contextual features classify the web pages more accurately even though following any one approach produces the desirable result. The main objective of this chapter is to focus on the web page classification issues and the machine learning techniques practiced by researchers to solve the web page classification problem.

LITERATURE SURVEY

Susan Dumais and Hao Chen (2000) explore the use of hierarchical structure for classifying web Pages using Support Vector Machine Classifiers. The hierarchical structure is initially used to train different second-level classifiers. In the hierarchical case, a model is learned to distinguish a second-level category from other categories within the same top level.

In the past, classification of the news has been done manually. Chee-Hong Chan, Aixin Sun, & Ee-Peng Lim (2001) experiment an automated approach to classify news based on SVM classifier which results in good classification accuracy. In personalized classification users define their own categories using specific keywords. By constructing search queries using these keywords, categorizer obtains positive and negative examples and performs classification. Online news represents a type of web information that is frequently referenced. The categorizer adopts SVM to classify the web pages into pre-defined categories (general categories) or user-defined categories (special categories). With personalized categories users are allowed to search their related article with the minimum effort.

Gautam Pant and Padmini Srinivasan (2006) proposed a method for classifying web pages us-