DOM-Based Multiple Noise Pattern Elimination from Web Page using Neural Network

Thanda Htwe
University of Computer Studies, Yangon
thandahtwe.ucsy@gmail.com

Nan Saing Moon Kham
University of Computer Studies, Yangon
moonkham.ucsy@gmail.com

Abstract

Nowadays, with the rapid growth of the Web, a large volume of data and information is published in numerous Web pages. However, Web pages are often cluttered with distracting features around the body of an article that distract Web user from the actual content they are interested in. Moreover, web pages are heterogeneous; many pages are semi-structured and noisy. To extract information from these pages, several challenges must be overcome. An important problem for information extraction from the web is the identification and removal of noise. In this paper, we focus on the information extraction based noise elimination approach VNR (Various Noises Remover) to remove various noise data from Web documents. It contains DOM-based analysis to traverse the whole Web page and a back propagation Neural Network algorithm for noise detection and removal. Our system first builds DOM tree structure for an incoming Web page and then split it into sub trees to detect noise data. We also apply back propagation neural network algorithm to classify various noise patterns, data patterns and mixture patterns in current Web page. The classification result of neural network is used for eliminating various noise patterns. In order to evaluate our proposed system, we perform the experiment on several Web pages of different Web sites as data set.

1. Introduction

Internet has billions of web pages ...pying significant portion of it and the number ...still growing rapidly. The amount of ...mation that is currently available on the Internet is HTML Web pages. Computer users are spending more and more time on the Internet in today's world of online shopping and banking; meanwhile, web pages are getting more complex in design and content. However, there is a lot of redundant and irrelevant information on the Internet [1], such as contents of mirror sites or identical pages with different URLs. We call this kind of redundancy intra page redundancy.

Examples of intra page redundancy include company logos, navigation panels, advertisements, catalogs of services, and announcements of copyright and privacy policies. These contents are frequently texts or hyperlinks irrelevant to the meaning of the page. However, these irrelevant links increase the difficulty for Web site analyzers, search engines, and Web miners to perform their tasks. Obviously, the problem of intra-page redundancy affects two factors widely used to evaluate search engines: the precision of search and the size of index. The presentation of search results is also influenced by the problem since most search engines automatically capture first several sentences as the description of a page.

Automatic extraction of useful and relevant content from web pages has many applications, ranging from enabling end users to accessing the web more easily over constrained devices like Personal Digital Assistants (PDAs) and cellular phones to provide more accessibility to the Web for the visually disabled.

In this work, we focus on detecting and eliminating local noises in Web pages to improve the performance of Web mining, e.g., Web page clustering and classification. This paper develops an efficient approach to informative data extraction problem by reducing noise from Web pages. We apply back propagation neural network