دهمین کنفرانس بین المللی مهندسی صنایع
Tenth International Industrial Engineering Conference
۷ و ۸ بهمن ماه ۱۳۹۲
27-28 January, 2014

SOM as a means to extract visitors preference in a webpage clustering task

A. Keramati, R. Jafari-Marandi
School of Industrial and System engineering, College of engineering, University of Tehran, Tehran, Iran

***Abstract*— over the years by expansion of WWW we have reached to a point that achieving this goal, having more visitors, does not seem as simple as it used to be. As to address these complexity and to survive in the bloody battle of attracting and holding on to visitors, a significant amount of websites owners' and researchers' attentions has been drawn to data mining techniques. Among others, SOM used to tackle webpage clustering task is one of the areas under spotlight. Thus, this paper has used the data extracted from an Iranian website's database to examine the capability of SOM in order to extract visitor preference. Additionally, this paper offers the process one should go through so as to extract appropriate data from a dataset, and which set of data are most suitable for this paper main purpose.**

Keywords- Self Organazing Map (SOM), Webpage clustering, Visitors' preference

I. INTRODUCTION

Nowadays websites have started to merit their visitors as potential customers or profits. However, this is mostly the case for company or enterprise websites whose principal earning ways lays upon their customer spending [1]. Even before these websites, there are business and websites in the World Wide Web (WWW) whose only reason of existence is because of their visitors. To put it another way, for world-wide-known websites such as Facebook or even Google the only reason behind their existence and working is their users (visitors). Being visited by a significant number of visitors, these websites grasp every opportunities to make profits. Therefore, today the front line of WWW's battles can be summarized into three words which is having more visitors.

Over the years by expansion of WWW we have reached to a point that achieving this goal, having more visitors, does not seem as simple as it used to be. As to address these complexity and to survive in the bloody battle of attracting and holding on to visitors, a significant amount of websites owners' and researchers' attentions has been drawn to data mining techniques. Adaption of different data mining techniques for different website's task and problems has been the subjects of many researches in the past few decades. In general, web mining, the term is used for applying data mining in web, has three distinct areas [2]: Web content mining, web structure mining, web usage mining. Web content mining is related to the web's content such as text, image, audio, video, metadata, and hypertexts and the effort is to extract useful concepts and rules and summarize the content on the web. Whereas, web structure mining is related to underlying link structures of the Web and its aim is to categorize Web pages, measure similarities and reveal relationships between different Web sites. Last but not least, web usage mining is related to Web users' interactions with the web and its users aims is to extract patterns and trends in Web users' behaviors.

One of the very first step for applying data mining techniques to real world business or commercial problem is the recognition of business problem [3]. However, this fact necessitate the recognition of the business itself which is only possible, for the most cases, for an expert inside the business and not a newly-recruited Data miner. Now for a recognition of a website problem, coming to know a website is of the website owner's or its data miner's interest. Fortunately, if we take a wise step, Data mining has load of techniques to offer, not to address site problem or task, but to gain information and knowledge about and to better our understanding of a website, and consequently to classify and categorized these gained information to use for future data mining or other purposes. However, summarizing, profiling and visualization of a website is not a small task and can be broken into three categories: website structure [4], website content [5], and website visitors [6]. Incidentally, this paper main focus is webpage clustering which is a subtask of website content profiling. Ease of Use.

The rest of this paper is organized as follow: the section 2 is the summary of studied literature. Section 3 represents a compendium of SOM algorithm and its implications. The section 4 is this paper research framework which outlines the techniques used and steps taken in this paper. And, finally section 5 and 6 are respectively are numerical results and conclusion.