Chapter XII
Web Mining by Automatically Organizing Web Pages into Categories

Ben Choi
Louisiana Tech University, USA

Zhongmei Yao
Louisiana Tech University, USA

ABSTRACT

Web mining aims for searching, organizing, and extracting information on the Web and search engines focus on searching. The next stage of Web mining is the organization of Web contents, which will then facilitate the extraction of useful information from the Web. This chapter will focus on organizing Web contents. Since a majority of Web contents are stored in the form of Web pages, this chapter will focus on techniques for automatically organizing Web pages into categories. Various artificial intelligence techniques have been used; however the most successful ones are classification and clustering. This chapter will focus on clustering. Clustering is well suited for Web mining by automatically organizing Web pages into categories each of which contain Web pages having similar contents. However, one problem in clustering is the lack of general methods to automatically determine the number of categories or clusters. For the Web domain, until now there is no such a method suitable for Web page clustering. To address this problem, this chapter describes a method to discover a constant factor that characterizes the Web domain and proposes a new method for automatically determining the number of clusters in Web page datasets. This chapter also proposes a new bi-directional hierarchical clustering algorithm, which arranges individual Web pages into clusters and then arranges the clusters into larger clusters and so on until the average inter-cluster similarity approaches the constant factor. Having the constant factor together with the algorithm, this chapter provides a new clustering system suitable for mining the Web.
INTRODUCTION

Web mining aims for finding useful information on the Web (Scime & Sugumaran, 2007; Linoff & Berry, 2001; Mena, 1999). The first stage of Web mining is searching. search engines, such as Google, focus on searching (Berry & Browne, 1999). Search engines first try to find as many Web pages as possible on the Internet. This is done by Web crawlers, which go from Web pages to Web pages to retrieve as many addresses (URLs) of Web pages as possible. Since current search engines use keyword search, keywords on each Web page found by the Web crawler are stored on databases for fast retrieval (Baberwal & Choi, 2004).

The next stage of Web mining is the organization of Web contents, which is the objective of this chapter. Since majority of Web contents are stored in the form of Web pages, current search engines and most current researches focus on organizing Web pages (Choi, 2001). Search engines, such as Google, focus on ordering Web pages based on the relevance of the Web pages in relating to the search keywords. Some search engines, such as Yahoo, also try to organize Web pages into categories. Yahoo tries to classify Web pages manually by having people read the contents of the Web pages and assign them to categories. Since the number of Web pages on the Internet has grown to the order of several billions, the manual method of classifying Web pages has been proved to be impractical. Thus, most current researches in Web mining focus on automatically organizing Web pages into categories (Choi & Yao, 2005; Yao & Choi 2007).

Various Artificial Intelligence techniques have been used to facilitate the process of automatically organizing Web pages into categories. Two of the most successful techniques are automatic classification and clustering. Web page classification assigns Web pages to pre-defined categories (Choi & Yao, 2005). Since defining a category is not an easy task, machine learning methods have been used to automatically create the definition from a set of sample Web pages (Choi & Peng, 2004). Web page clustering does not require pre-defined categories. It is a self-organization method based solely on measuring whether a Web page is similar to others. It groups Web pages having similar contents into clusters. This chapter will focus on automatic clustering of Web pages.

The organization of Web contents will then facilitate the final stage of Web mining, which is the extraction of useful information from the Web. Nowadays the extraction of useful information from the Web is usually done by search engine users, who have to scan Web pages after Web pages in hope of finding the useful information and often give up without getting the needed information. The results of organizing Web pages into categories or clusters will allow the users to focus on the groups of Web pages that are relevant to their needs.

The future of Web mining is moving toward Semantic Web, which aims for automatically extracting useful information from the Web (Antoniou & van Harmelen, 2004). For a computer to automatically extract useful information from the Web, the computer first needs to understand the contents of Web pages. This is done with the help of natural language understanding and with the help of assigning meaningful tags to strings of characters. For instance, a string of digits may be assigned as phone number or a string of digits and letters may be assigned as address. Understanding of Web contents will also help organizing Web pages into categories and on the other hand the organization of Web contents can facilitate the understanding (Choi & Guo, 2003; Peng & Choi, 2005).

In this chapter, we are interested in cluster analysis that can be used to organize Web pages into clusters based on their contents (Choi & Yao, 2005; Yao & Choi, 2007). Clustering is an unsupervised discovery process for partitioning a set of data into clusters such that data in the same cluster is more