Chapter XIII
A Multi-Agent Neural Network System for Web Text Mining

Lean Yu
Chinese Academy of Sciences, China & City University of Hong Kong, China

Shouyang Wang
Chinese Academy of Sciences, China

Kin Keung Lai
City University of Hong Kong, China

ABSTRACT

With the rapid increase of the huge amount of online information, there is a strong demand for Web text mining which helps people discover some useful knowledge from Web documents. For this purpose, this chapter first proposes a back-propagation neural network (BPNN)-based Web text mining system for decision support. In the BPNN-based Web text mining system, four main processes, Web document search, Web text processing, text feature conversion, and BPNN-based knowledge discovery, are involved. Particularly, BPNN is used as an intelligent learning agent that learns about underlying Web documents. In order to scale the individual intelligent agent with the large number of Web documents, we then provide a multi-agent-based neural network system for Web text mining in a parallel way. For illustration purpose, a simulated experiment is performed. Experiment results reveal that the proposed multi-agent neural network system is an effective solution to large scale Web text mining.

INTRODUCTION

Web text mining is the process of using unstructured Web-type text documents and examining it in an attempt to find implicit patterns hidden in the Web text documents. With the amount of online Web
text information growing rapidly, the need for a powerful Web text mining method that can analyze Web text information and infer useful patterns for prediction and decision purposes has increased. To be able to cope with the abundance of available Web text information, Web users need assistance of some software tools and software agents (often called “softbots”) for exploring, sorting, and filtering the available Web text information (Etzioni, 1996; Kozierok & Maes, 1993).

Much effort for Web text mining has been made and some important progresses are obtained. For example, Joachims (1996) utilized probabilistic TFIDF method and naïve Bayes method to perform text categorization task, one subtasks of the text mining. Feldman and Dagan (1998) proposed a keyword-frequency approach to explore unstructured text collections. Tan (1999) presented a two-phase text mining framework for knowledge discovery in text. Lee and Yang (1999) used a self-organizing map (SOM) method to perform Web text mining task. Chen and Nie (2000) proposed a parallel Web text mining approach for cross-language information retrieval. Choi and Yoo (2001) utilized a neural network approach to text database discovery on the Web. Recently, Yu, Wang, and Lai (2005) utilized rough set theory to refine text mining for prediction purpose. Generally speaking, existing research concentrated on the development of agents that are high level interfaces to the Web (Etzioni & Weld, 1994; Furnkranz, Holzbaur, & Temel, 2002), programs for filtering and sorting e-mail messages (Maes, 1994; Payne & Edwards, 1997), or usenet netnews categorization (Joachims, 1996; Lang, 1995; Lashkari, Metral, & Maes, 1994; Mock, 1996; Sheth & Maes, 1993). More examples about Web text mining can be found in two recent surveys (Chakrabarti, 2000; Kosala & Blockeel, 2000). In the meantime, a number of Web text mining systems, such as IBM Intelligent Miner (http://www-306.ibm.com/software/data/iminer/) and SAS Text Miner (http://www.sas.com/), have already been developed.

Although some progress in Web text mining has been made, there are still several important issues to be addressed. First of all, most text mining tasks focus on the text categorization/classification, text clustering, concept extraction, and document summarization (Yu, Wang, & Lai, 2006). But the text content and entity relation modeling (i.e., the causality relationship between entities) is less explored in Web text documents; the essential goal of text mining is often neglected. As we know, the basic motivation of text mining is to find and explore some useful knowledge and hidden relationships from some unstructured text data to support decision-making, similar to data mining in structured data. In the existing literature, these text mining tasks little involved in hidden entity relationship modeling. For example, the main function of text categorization is to classify different documents into different prelabeled classes (e.g., Joachims, 1996), but how the different documents with different categories support decision-making is not clear. Differing in the previous Web text mining tasks, this chapter attempts to explore some implied knowledge hidden in Web text documents to support business prediction and decision-making (Yu et al., 2006).

Second, most text mining models usually utilize some well-known tools such as the vector space model (VSM) (e.g., TFIDF algorithm (Salton, 1971, 1991; Salton & Yang, 1973; Sparck Jones, 1972)) and some traditional statistical models, for example, naïve Bayes algorithm (Feldman & Dagan, 1998; Joachims, 1996; Katz, 1995)). Nevertheless, a distinct shortcoming of these models is that their extrapolation and generalization capabilities are often weak (Yu et al., 2006). To remedy this drawback, this chapter adopts an intelligent learning agent instead of traditional statistical models to perform the Web text mining tasks. Because the learning agent has good learning and flexible mapping capability between inputs and outputs, the generalization capability may be much stronger than the traditional models (Yu et al., 2006).