Multi-agent Web Text Mining on the Grid for Enterprise Decision Support

Kin Keung Lai [1,2,3], Lean Yu [1,3], and Shouyang Wang [1,2]

[1] Institute of Systems Science, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100080, China
{yulean, sywang}@amss.ac.cn
[2] College of Business Administration, Hunan University, Changsha 410082, China
[3] Department of Management Sciences, City University of Hong Kong,
Tat Chee Avenue, Kowloon, Hong Kong
{msyulean, mskklai}@cityu.edu.hk

Abstract. In this study, a multi-agent web text mining system on the grid is developed to support enterprise decision-making. First, an individual intelligent learning agent that learns about underlying text documents is presented to discover the useful knowledge for enterprise decision. In order to scale the individual intelligent agent with the large number of text documents on the web, we then provide a multi-agent web text mining system in a parallel way based upon grid technology. Finally, we discuss how the multi-agent web text mining system on the grid can be used to implement text mining services.

1 Introduction

With the development of technology, the enterprises are suffering more pressures than ever. To obtain competitive edges, the utilization of intelligent mining techniques has received more and more attention. Currently, text mining, which can handle non-structured textual data, has becoming a new decision support tool for enterprise decision-makers. Since the most natural form of storing information is as text, text mining is believed to have a higher commercial potential than data mining [1]. A recent study, conducted by Delphi Group (http://www.thedelphigroup.com/), has indicated that 80% of a company's information is contained in textual documents. Thus, it is important to develop a web text mining system for enterprise decision-making.

In the web text mining, one crucial problem is how deal with the large numbers of available text documents over a tolerable limit. For this problem, a multi-agent intelligent learning system based on the grid technology [2] is proposed. The main motivation of this study is to develop a multi-agent web text mining system on the grid that offers valuable knowledge to support enterprise decision-making. The rest of this study is organized as follows. Section 2 presents a framework of the back propagation neural network (BPNN) based intelligent learning agent for text mining. To scale the computational load for the large-scale text mining task, a multi-agent web text mining system on the grid is proposed in Section 3. Section 4 concludes.

H.T. Shen et al. (Eds.): APWeb 2006, LNCS 3842, pp. 540 – 544, 2006.

2 The BPNN-Based Intelligent Agent for Web Text Mining

Web text mining, a new research field in knowledge discovery, refers to the process of using unstructured web-type textual document and examining it in an attempt to discover implicit patterns "hidden" within the web documents using interdisciplinary techniques from data mining, machine learning, and natural language processing [1]. One main goal of web text mining is to help people discover knowledge for decision support from large quantities of semi-structured or unstructured web text documents.

Actually, web text mining consists of a series of tasks and procedures, which involves many interdisciplinary fields mentioned above. Because the final goal of text mining is to support decision, the web text mining must adapt the dynamic change over the time as the web text documents increase rapidly. Thus, web text mining must have learning capability. In this study, the BPNN is used as a computational agent for web text mining. In the environment of our proposed approach, the BPNN agent is first trained with the many related web documents, and then the trained BPNN agent can project to new documents for decision when new web document arrives. Fig. 1 illustrates the main components of a BPNN agent and the control flows among them. Note that the control flows with thin arrow represent learning phase and the control flows with bold arrow represent discovering phase of web text mining system.

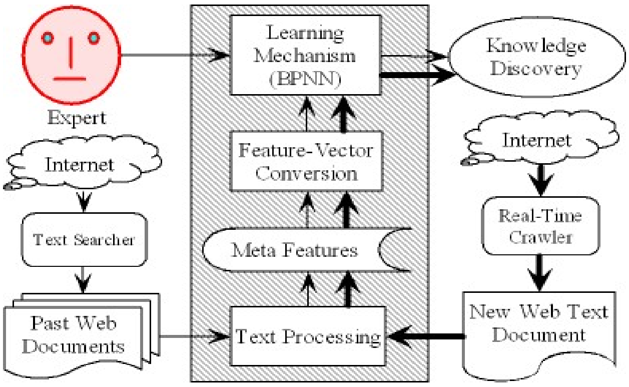

Fig. 1. The framework of a BPNN agent for web text mining

As can be seen from Fig. 1, the framework of the BPNN agent for web text mining consists of four main components: text search and collection, text processing, feature vector conversion, and learning mechanism, which is described in detail as follows.

(1) Text search and collection. Clearly, the first step in web text mining is to collect the text data. Its work is to find related text documents by retrieval tools.
(2) Text processing. When web text documents are collected, the collected text documents are mainly represented by semi- or non-structural information. Its aim is to extract typical features that represent the text contents from these collected texts.
(3) Feature vector conversion. Before using knowledge discovery algorithm, text feature data must be transformed into numerical data. Here we use binary form to

perform conversion. We simply check for the presence “1” or absence “0” of words by comparing with predefined indices to formulate a numerical table with binary form.

(4) BPNN agent learning mechanism. In this study, the BPNN is used as an intelligent agent to explore the hidden patterns. Actually, BPNN agent is a supervised learning mechanism in the form of the neural network associative memory as shown in Fig. 2 as the shaded rectangle. Thus the BPNN agent acts in two phases: a training phase (the top part) and a testing phase (the bottom part), as illustrated in Fig. 2.

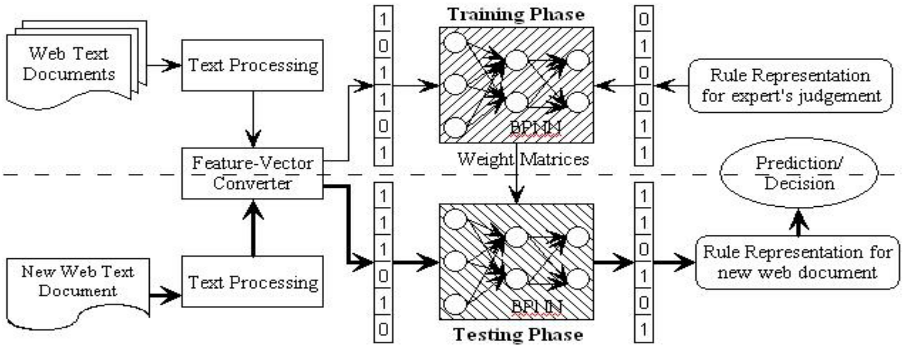

Fig. 2. The learning mechanism of BPNN agent for web text mining

During the training phase, the input and the output layer of the BPNN are set to represent a training pair (x, y) where x is produced by the feature vector converter and y is produced by the expert’s judgments. Commonly, $x \in R^n$ is an n-dimensional feature vector containing the independent variable or attributes, and $y \in \{0, 1\}$ is the dependent variable or truth. The goal is to construct a function or a model f,

$$y = f(x) = f_a(x) = f(x; a),\ a \in A, \tag{1}$$

where $f = f_a$ is defined by specifying parameters $a \in A$ from an explicitly parameterized family of models A. To search the optimal parameters a, the BPNN learning procedure is performed for all training pairs. The BPNN procedure repeatedly adjusts the link-weight matrices of BPNN in a way that minimizes the error for each training pair. When the average squared error is acceptably small, the BPNN stops and produces the link-weight matrices, which is stored as the knowledge for the use of testing phase.

During the testing phase, the input layer of the BPNN is activated by the feature vector produced by the feature-vector converter for a new web text document. This activation of the BPNN spreads from the input layer to the output layer using the link-weight matrices stored during the training phase. That is, the model $f = f_a$ determined by training phase is applied to previously unseen feature vectors x to produce the output of BPNN, a vector representation whose components are all between 0 and 1.

In principle, our approach proposed above offers the potential solution to the web text mining problem. But it is infeasible for a single BPNN agent to handle large-scale text documents. For this problem, a multi-agent web text mining system on the grid is proposed in the next section.

3 Multi-agent Web Text Mining on the Grid

3.1 The Structure of Multi-agent Web Text Mining System

Assume that we have a number of BPNN-based intelligent agents, each of which has its own available text documents described in the previous section. For handling the different web text documents with different categories, a two-layer multi-agent web text mining system is constructed, as illustrated in Fig. 3.

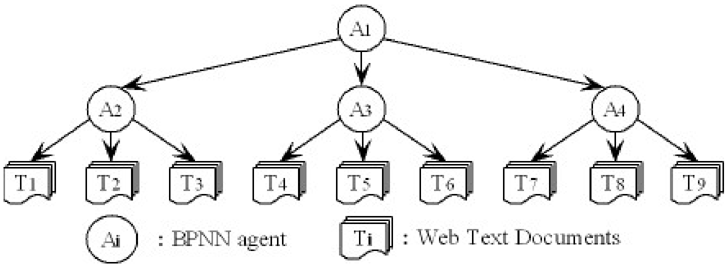

Fig. 3. The structure of multi-agent web text mining system

From Fig. 3, the two-layer multi-agent web text mining system includes one superordinate BPNN agent and several subordinate BPNN agents. Here the superordinate BPNN agent can treat the subordinate BPNN agents in the same way as the subordinate BPNN agents treat their available text documents. Usually, the multi-agent-based web text mining system can be performed by distributed system in a parallel way. However, the operation of multi-agent web text mining system may increase computational requirements, for example, adding some computers to deploy the BPNN agents. For this requirement, the grid technology is applied in this study.

3.2 The Multi-agent Web Text Mining System on the Grid

With the growing demands of computational requirements of the multi-agent web text mining system, grid infrastructures are foreseen to be one of the most critical yet challenging technologies to meet the practical demands for high performance and high efficiency text mining in a large variety of web text documents. In the past few years, many software environments for gaining access to very large distributed computing resources have been made available, such as Globus [3] and Condor [4]. Based upon the previous work, a multi-agent-based web text mining system on the grid is proposed. That is, our BPNN agent can be deployed in different grid and implemented collaboratively. Fig. 4 shows the architecture of the web text mining with three grids.

As a result, the multi-agent-based web text mining system on the grid can collaboratively discover some useful knowledge for enterprise decision support in an efficient way. For further explanation, a simulated study should be provided. Due to space limitation, the simulated experiment is omitted here.

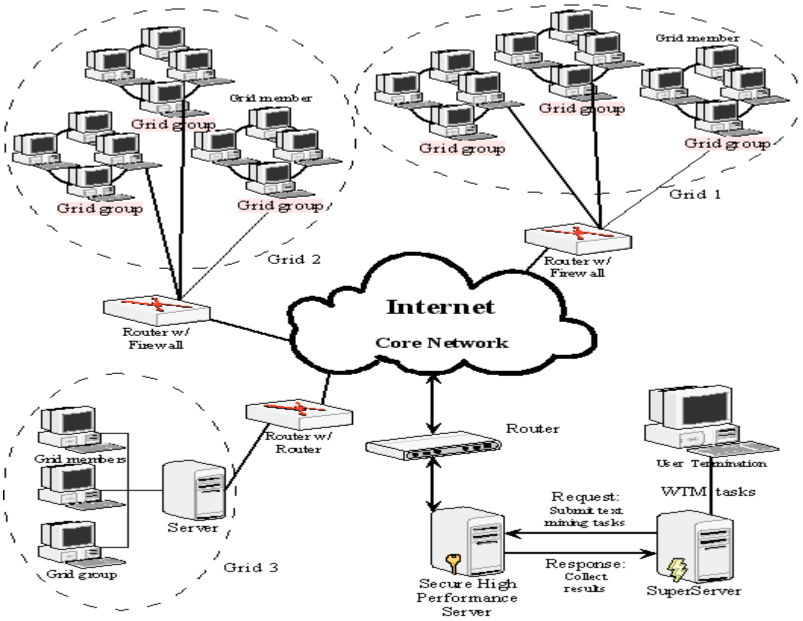

Fig. 4. The architecture of the multi-agent web text mining system on the grid

4 Conclusions

This study proposes a multi-agent web text mining system on the grid to support enterprise decision. In this study, we first propose a single intelligent agent to perform text mining. With the rapid increase of web information, a multi-agent web text mining system on the grid is then constructed for large-scale text mining application.

References

1. Yu, L., Wang, S.Y., Lai, K.K.: A Rough-Set-Refined Text Mining Approach for Crude Oil Market Tendency Forecasting. International Journal of Knowledge and Systems Sciences 2 (2005) 33-46
2. Foster. I., Kesselman, C., Tuecke, S.: The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International Journal of High Performance Computing Applications 15 (2001) 200-223
3. Litzkow, M., Livny, M.: Experience with the Condor Distributed Batch System. Proceedings of the IEEE Workshop on Experimental Distributed Systems (1990) 97-101
4. Foster, I., Kesselman, C., Tuecke, S.: Globus: A Metacomputing Infrastructure Toolkit. International Journal of Supercomputer Applications 11 (1997) 115-128