ABSTRACT

The business needs, the availability of huge volumes of data and the continuous evolution in Web services functions derive the need of application of data mining in the Web service domain. This article recommends several data mining applications that can leverage problems concerned with the discovery and monitoring of Web services. This article then presents a case study on applying the clustering data mining technique to the Web service usage data to improve the Web service discovery process. This article also discusses the challenges that arise when applying data mining to Web services usage data and abstract information.

Keywords: data mining; knowledge discovery; service discovery; Web services

INTRODUCTION

Simplicity and ubiquity are factors that brought upon the success of the Web. Once solely a repository that provides access to information for human interpretation and use, the Web has evolved to become a repository of software components (Alesco & Smith, 2004; Clark 2002; Newcomer, 2002). These service-oriented components, or Web services, are emerging as the key enabling technology for today’s e-commerce and business-to-business applications, and are transforming the Web into a distributed computation and application framework. The implication of this transformation is that the volume of data generated on the Web is increasing exponentially. Whilst human users have generated huge volumes of data from using the Web and warranted a myriad of research in the data mining community (Kosala & Blockeel, 2000), the volume of data generated by machines using and providing Web services would certainly make the former dwarf in size. Since Web servers store the access request information in Web server access logs, all interactions between Web services are recorded, similar to the human-server interactions that are prevalent on the Web now. Web server access logs recording site accesses have been a rich source for discovering the browsing behavior of site visitors in Web mining. In a similar manner, logs that record the Web services accesses can be mined for Web services usage patterns by organizations. With the huge volume of accesses, the amounts of logs that can be collected make a feasible source for data mining. Data mining (DM) methods search for distinct patterns and trends that exist in data but are ‘hidden’ among the vast amount of data (Han & Kamber, 2001).
Since Web service is a relatively new technology, there are a lot of areas where improvements can be made to complement the current state of the art. Not only these improvements enhance the technical workings of Web services, they can also generate new business opportunities. As will be shown in this article, some of these improvements can be made possible through data mining. For example, as the number of Web services increases, it becomes increasingly important to provide a scalable infrastructure of registries that allows both developers and end-users to perform service discovery.

Given the competitiveness of the current business market place, the need to exploit sources of data for business intelligence is greater than ever before. The ability for a business to obtain previously unknown knowledge about itself and its customers is a valuable asset. Business needs, the availability of huge volumes of data, and in particular, the possible Web services improvement areas form the premises for the application of data mining using the Web services data. The goal of this article is thus to present how Web services data can be utilized in data mining and the kinds of benefits that can be gained from such operations.

This article discusses the Web services in the context of data mining. It first discusses the basics of data mining, and then recommends a set of data mining applications that can leverage problems concerned with the discovery and monitoring of Web services. A case study is presented that demonstrates the improvement in the Web discovery with a proposed clustering data mining method on using the Web query logs. Some of the challenges that arise while mining Web services data are also presented. In the end, this article looks at the current technologies in use for Web service discovery and log mining.

**DATA MINING AND ITS OPEARTIONS**

Data mining is the search for distinct patterns and trends that exist in datasets but are ‘hidden’ among the vast amount of data. A data mining task includes problem identification, data pre-processing, data modeling and pattern evaluation. With the high costs associated with data mining operations, it is essential for businesses to know whether their investments are worthwhile. This requires the analyst to have an understanding of the application domain, the relevant prior knowledge, the data that is available for mining, and the goals of the end user. Once the objective of data mining task is identified, the next step is to prepare the data for mining. This pre-processing step involves basic operations such as the removal of noise or outliers, handling inconsistencies and missing values to ensure the quality of the data set, and transforming it into data structures specific to the data mining task. The preprocessing step is particularly relevant to this study as its assists to determine how the Web services data can be mined. Data modeling is next used to infer rules from the pre-processed data or build model that fits best on this data set. The data modeling step refers to the application of a DM technique or a combination of techniques for identifying patterns from the derived data set. The kinds of patterns that can be mined depend on the data mining operation as listed in Table 1.

Predictive modeling solves a problem by looking at the past experiences with known answers, and then projecting to new cases based on essential characteristics about the data. This process enables the prediction of the unknown value of a variable given known values of other variables. The classification predictive modeling is used to predict discrete nominal values, whereas value prediction, or regression, is used to predict continuous values. Clustering or segmentation solves a problem by identifying objects with similar characteristics in a data set, and thus forming the groups of similar objects. The link analysis operation establishes
The Library Big Data Research: Status and Directions
www.igi-global.com/chapter/the-library-big-data-research/217944?camid=4v1a