Incremental Mining of the Schema of Semistructured Data

ZHOU Aoying (周傲英), JIN Wen (金 文), ZHOU Shuigeng (周永庚),
QIAN Weining (钱卫宁) and TIAN Zenping (田增平)

Department of Computer Science, Fudan University, Shanghai 200433, P.R. China
E-mail: {ayzhou,970220,970218,zptian}@fudan.edu.cn
Received December 14, 1998; revised October 22, 1999.

Abstract Semistructured data are specified in lack of any fixed and rigid schema, even though typically some implicit structure appears in the data. The huge amounts of on-line applications make it important and imperative to mine the schema of semistructured data, both for the users (e.g., to gather useful information and facilitate querying) and for the systems (e.g., to optimize access). The critical problem is to discover the hidden structure in the semistructured data. Current methods in extracting Web data structure are either in a general way independent of application background, or bound in some concrete environment such as HTML, XML etc. But both face the burden of expensive cost and difficulty in keeping along with the frequent and complicated variances of Web data. In this paper, the problem of incremental mining of schema for semistructured data after the update of the raw data is discussed. An algorithm for incrementally mining the schema of semistructured data is provided, and some experimental results are also given, which show that incremental mining for semistructured data is more efficient than non-incremental mining.

Keywords data mining, incremental mining, semistructured data, schema, algorithm

1 Introduction

With more and more applications working successfully, Data Mining in very large databases has recently been focused on by many database research communities because of its promising future in many areas, such as OLAP in insurance safety, decision support, market strategy and financial forecast [1,2]. In the early years, much research energy was targeted at the mining of transaction databases, relational databases, spatial databases etc., which generally hold structured data [3-7]. However, as the maturity of Internet develops fast nowadays, the volume of data available on-line grows much rapidly too. Generally, data on network or in the digital library have no absolute schema fixed in advance, and the structure of data may be fairly irregular or incomplete. This kind of data is termed as semistructured data. Semistructured data are self-describing data that have some structure but which is neither regular nor known a-priori to the system. In order to facilitate querying and optimizing access to the Web, mining certain common structure or schema from the semistructured data is particularly important.

The large amount of interesting data on the Web encourages people to try to solve the problem above. Some effective work such as [8, 9] has made to extract the schema of semistructured data in a general way which is independent of the application environment, while [10-12] reached the same goal through different paths. S. Nestorov's work in [8] refers to approximately classifying the semistructured objects into a hierarchical collection of types.
to extract the implicit structure in large sets of semistructured data. K. Wang in [9] at the first time directly raises the problem of discovering schema for semistructured data using association rule, and a framework and the corresponding algorithm for it are proposed. But their efforts in [8] and [9] both are of the drawback of redundant computation work whose cost is expensive when the variances in Web data are so often. The work of Mendelzon's group in Toronto University tried to start their step from some applications field such as HTML etc. and with the help of cost locality, range expression bounds, finally the web data structure could be obtained through the calculus based on virtual graph in [10-12]. Their idea is easy in implementation but is limited to the application field they applied first. As the data on the Internet/Web increasing continuously with the time moves on, mining data on Internet/Web cannot be done once and for all. The mined results need to be recast as the raw data grow. In the course of this period, a useful way to update the mined results is to take advantage of the already discovered results rather than do the whole work from scratch. So naturally, an incremental mining philosophy must be adopted. However, to our best knowledge, we have found no literature on incrementa! mining of schema for semistructured data up to now.

This paper solves the problem of incremental mining of schema for semistructured data, and an algorithm for this purpose is also presented. The other parts of this paper are organized as follows. We first introduce and describe some basic concepts in Section 2, including the object exchange model (OEM) used for representing semistructured data, schema mining for semistructured data, and incremental mining problem. In Section 3, the algorithm for incrementally mining semistructured data is given in detail. Section 4 provides some experimental results that show the advantage of incremental mining method. Finally, we summarize the paper in Section 5.

2 Defining the Problem

In this section, we assume that one description model, say OEM, is applied and the corresponding concepts or pre-definitions of schema and incremental mining on semistructured data are also introduced.

2.1 Object Exchange Model (OEM)

We represent semistructured data by object exchange model (OEM), which has been proposed for modeling semistructured data in [13]. In OEM, semistructured data are represented as a rooted, labeled, and directed graph with the objects as vertices and labels

Fig.1. Example of semistructured database TD in OEM.