Online Analytical Mining of Path Traversal Patterns for Web Measurement

Joseph Fong, City University of Hong Kong, Hong Kong
H.K. Wong, City University of Hong Kong, Hong Kong

ABSTRACT

The WWW and its associated distributed information services provide rich world-wide online information services, where objects are linked together to facilitate interactive access. Users seeking information of Internet traverse from one object via links to another. It is important to analyze user access patterns which will help improve Web page design by providing an efficient access between highly correlated objects, and also assist in better marketing decisions by placing advertisements in frequently visited document. We need to study the user surfing behavior by examining the Web access log, browsing frequency of Web pages and computing the average duration time of visit. This paper offers an architecture to store the derived web user access paths in a data warehouse, and facilitates its view maintainability by use of a metadata. The system will update the user access paths pattern with the data warehouse by the data operation functions in the metadata. Whenever a new user access path occurs, the view maintainability is triggered by a constraint class in the metadata. The data warehouse can be analyzed on the frequent pattern tree of user access paths on the Website within a period and duration. The result is an online analytical mining path traversal pattern. Our experimental and performance studies have demonstrated the effectiveness and efficiency of our system with the following contributions: an architecture of online analytical mining (OLAM) using frame model metadata, a methodology (stepwise procedure) of implementing OLAM and the resultant cluster of web pages frequently visited by users for marketing use.

Keywords: OLAM, web measurement, path traversal patterns, view maintenance

INTRODUCTION

As the popularity of WWW explodes, a massive amount of data is generated by Web servers in the form of Web access logs. This is a rich source of information for understanding Web user surfing behavior. Web usage mining is one type of Web mining activity that involves the automatic discovery of user access patterns on Web server(s). On the other hand, it is an application of data mining algorithms to Web access logs to find the trends and regularities in Web users’ traversal patterns.

Analysis of these access data can provide useful information for server performance enhancements, restructuring a Website, and direct marketing in electronic
commerce. As a result, Web usage mining has been widely used in improving Website design, business and marketing decision support, user profiling, and Web server system performance.

Among discovering various kinds of knowledge in large databases, mining association rule has attracted great attention in database research communities in recent years (Agrawal, Imielinski, & Swami, 1993; Agrawal & Srikant, 1994; Miller, & Yang, 1997; Srikant & Agrawal, 1995; Savasere, Omiecinski & Navathe, 1995). Association rule mining is a form of data mining to discover interesting relationships among attributes in those data. The discovered rules may help marketing decision support, and business management. Association rules have two important measurements: Support and Confidence. Support is an argument that decides whether the candidate is frequent or not. Confidence is an argument that describes the believable degree of association rules.

The Frequent Pattern Growth (FP-growth) algorithm is one of the association rule algorithms to find frequent itemsets, but unlike Apriori, it avoids the expense of candidate generation by generating only candidate itemsets. Because FP-growth does not need to examine both candidate and non-candidate sets and requires only two scans of the database, it is a fast algorithm for mining association patterns. We will study this algorithm in-depth in our proposed algorithm that is called Sequential FP-growth.

In this paper, we propose and develop an interesting method that is called online analytical mining of path traversal patterns, which integrates the recently developed data warehouse technology with efficient association mining methods. The system stores the derived Web user access paths in a data warehouse, and facilitates its view maintainability by use of a frame metadata. The system will update the user access paths pattern with the data warehouse by the data operation functions in the frame metadata. Whenever a new user access path occurs, the view maintainability is triggered by a constraint class in the frame metadata. The data warehouse can be analyzed on the frequent pattern tree of user access paths on the Website within a duration. The developed method achieves incremental, extensible, and multi-dimensional association rule mining with high performance.

RELATED WORK

Association Rules Discovery

The concept of association rules was first introduced in Agrawal, Imielinski, & Swami (1993). Since then, the problem of data mining for association rule has been studied extensively (Agrawal, & Srikant, 1994; Cheung, Han, Ng, & Wong, 1996; Han, Karypis, & Kumar, 1997; Park, Chen, & Yu, 1995; Savasere, Omiecinski, & Navathe, 1995; Savasere, Omiecinski, & Navathe, 1995; Savasere, Omiecinski, & Navathe, 1995). These studies covered a broad range of topics with variations studied, and aimed for further improvements of the performance of the algorithm. For example, fast algorithms based on the Apriori Algorithm (Agrawal, & Srikant, 1994), incremental updating and parallel algorithms (Cheung, Han, Ng, & Wong, 1996; Han, Karypis, & Kumar, 1997; Park, Chen, & Yu, 1995), mining of generalized, multi-level rules, and multi-dimensional rules (Zhao, Deshpande, & Naughton, 1997).

Sequential Patterns Mining

The problem of discovering sequential patterns mining is to find inter-transac-
Related Content

Data Warehouse Design to Support Customer Relationship Management Analysis
www.igi-global.com/article/data-warehouse-design-support-customer/3353?camid=4v1a

Principles on Symbolic Data Analysis
www.igi-global.com/chapter/principles-symbolic-data-analysis/20690?camid=4v1a
User Resistance to Software Migration: The Case on Linux
[www.igi-global.com/article/user-resistance-to-software-migration/109932?camid=4v1a](www.igi-global.com/article/user-resistance-to-software-migration/109932?camid=4v1a)

Modeling Design Patterns for Semi-Automatic Reuse in System Design
[www.igi-global.com/chapter/modeling-design-patterns-semi-automatic/63661?camid=4v1a](www.igi-global.com/chapter/modeling-design-patterns-semi-automatic/63661?camid=4v1a)