An Efficient Approach for Analyzing User Behaviors in a Web-Based Training Environment

Show-Jane Yen, Fu Jen Catholic University, Taiwan, R.O.C.

ABSTRACT

Mining frequent traversal patterns is to discover the consecutive reference paths traversed by a sufficient number of users from Web logs in a Web environment where users can travel from one object to another through the corresponding hyperlinks. Previous approaches for mining frequent traversal patterns only consider the forward references, such that the information about backward references will be lost. In this paper, we propose an efficient algorithm to discover the non-simple frequent traversal patterns. The non-simple frequent traversal patterns include forward and backward references, and are used to suggest potentially interesting traversal path to the users, which are useful, especially, in a Web-based training environment. The experimental results show that the discovered patterns can present the complete browsing paths traversed by most of the users and our algorithm outperforms other algorithms in discovered information and execution times.

Keywords: web mining, web structure, traversal sequence, non-simple frequent traversal pattern

INTRODUCTION

Because the capacity of the storage is getting larger, a large amount of data can be stored in a database. Potential useful information may be embedded in large databases. Hence, how to discover the useful information in such databases is becoming a popular field in computer science. The purpose of data mining (Agrawal, 1994; Chen et al., 1995; Han et al., 2000; Tung et al., 1999; Yen, 2000) is to discover the useful information from the large databases such that the quality of decision-making can be improved.

Owing to the fact that available data on the Web is rapidly growing, discovery and analysis of useful information from the Web is becoming an important task, which is called Web Mining (Chen et al., 1999; Chen & Yun, 2000; Cooley et al., 1999; Cho et al., 2000; Yen, 1998). One of the
most important Web applications is Web-based training, from which users can learn anywhere.

However, most Web-based training systems statically provide homepages with course content and information services, but do not consider the learning directions and the interests of the users. Web users often waste a lot of time for searching and waiting for needed information because the designers of the Web-based training systems cannot understand what the Web users really want, i.e., the systems cannot provide correct and interesting information to the Web users efficiently. In this paper, we track the original user traversal paths in Web logs, and analyze the user access patterns to find the consecutive reference sequence traversed by a sufficient number of users.

The definitions about mining non-simple frequent traversal patterns are presented as follows: A page corresponds to a Web page or object on the Web. A traversal sequence is a set of pages ordered by increasing traversal-time, which is represented as \(<s_1, s_2, \ldots, s_n>\) and \(s_i\) is a page. For two traversal sequences \(<a_1, a_2, \ldots, a_n>\) and \(<b_1, b_2, \ldots, b_m>\), if there exists ordered integers \(i_1 < i_2 < \ldots < i_k\), \(1 \leq i_k \leq m\), such that \(a_1 \subseteq b_{i_1}, \ldots, a_n \subseteq b_{i_k}\), then \(<a_1, a_2, \ldots, a_n>\) is contained in \(<b_1, b_2, \ldots, b_m>\). A traversal sequence is maximal if this sequence is not contained in any other sequence.

A user sequence is a complete traversal sequence for a user, from entering to exiting the Web system. A user sequence database contains a set of user sequences traversed by all users in the system, which includes user identifiers, traversal pages and traversal-time. A user sequence \(c\) supports a traversal sequence \(s\) if \(s\) is contained in \(c\). The support for a traversal sequence \(s\) is the number of user sequences that supports \(s\). If the support for a traversal sequence \(s\) satisfies the user-specified minimum support threshold, then \(s\) is called frequent traversal sequence. The length of a traversal sequence \(s\) is the number of pages in the sequence. A traversal sequence of length \(k\) is called a \(k\)-traversal sequence, and a frequent traversal sequence of length \(k\) a frequent \(k\)-traversal sequence. If a frequent traversal sequence is maximal among all the other frequent traversal sequences, then the frequent traversal sequence is a frequent traversal pattern. If there are repeated pages in a traversal sequence, that is, the backward references are also retained in the sequence, then this sequence is called non-simple traversal sequence. If the traversal sequence only includes forward references, but not backward references, then this sequence is called simple traversal sequence. A non-simple traversal sequence can represent complete user behaviors because a user browsing path always includes backward and forward references. However, for simple traversal sequence, just a part of the user behaviors remain because the backward behaviors have been removed and lost.

In this paper, we propose an efficient algorithm to discover all the non-simple frequent traversal patterns from a user sequence database in the Web-based training environment. For convenience, the non-simple frequent traversal sequence is also called frequent traversal sequence in this paper. Before generating the frequent traversal sequences, we need to generate the candidate traversal sequences, and scan the database to count the support for each candidate traversal sequence to decide if it is a frequent traversal sequence. A candidate traversal sequence of length \(k\) is called a candidate \(k\)-traversal sequence.

The rest of this paper is organized as
The Sloan Consortium

www.igi-global.com/chapter/sloan-consortium/12323?camid=4v1a