Modeling and Analysis the Web Structure Using Stochastic Timed Petri Nets

Po-Zung Chen, Chu-Hao Sun
Department of Computer Science and Information Engineering, Tamkang University, Taipei County, Taiwan
Email: pozung@mail.tku.edu.tw, 894190320@s94.tku.edu.tw

Shih-Yang Yang
Department of Information Management, Kang-Ning Junior College of Medical Care and Management , Taipei, Taiwan
Email: shihyang@knjc.edu.tw

Abstract—**Precise analysis of the Web structure can facilitate data pre-processing and enhance the accuracy of the mining results in the procedure of Web usage mining. STPN（Stochastic Timed Petri Nets） is a high-level graphical model widely used in modeling system activities with concurrency. STPN can save the analyzed results in an incidence matrix for future follow-up analyses, and some already-verified properties held by STPN, such as reachability, can also be used to solve some unsettled problems in the model. In the present study, we put forth the use of STPN as the Web structure model. We adopt Place in the STPN model to represent webpage on the websites and use Transition to represent hyperlink. Through the model, we can conduct Web structure analysis. We simultaneously employ the Web structure analysis information in the incidence matrix and the reachability properties, obtained from the STPN model, to help proceed with pageview identification and path completion at the data preprocessing phase.**

Index Terms—**Web usage mining, data preprocessing, Stochastic Timed Petri Nets, reachability behavior, pageview identification, path completion.**

I. Introduction

As the Internet is increasingly prevalent worldwide and bringing out variations in business transactions, website management and design capability have been two of the potent constituents in the area of information science. To achieve better website management and design capability, many website management professionals started to examine site-users' webpage browsing frequency, sequence and even duration through the Web usage profiles. Hence, Web usage mining has become a hot research topic.

Web mining is the application of data mining techniques to discover patterns from the Web. According to analysis targets Web mining is divided into three types, namely Web content mining, Web structure mining and Web usage mining [1][2][3].

The main task of Web usage mining is to retrieve the information meaningful to the system management personnel from the Web server's accumulated usage profiles left by all the browser users. As the profiles are only the sequentially recorded contents of the services provided by the Web server, the profiles not only could contain multiple browsing profiles of different browser users but also could take in some extra or erroneous profiles. The website management personnel must proceed with preprocessing to these usage profiles if they are to correctly analyze said users' webpage-contents usage sequence. Hence, a data preprocessing is needed to enhance information processing before we can analyze the usage profiles. The first step of data preprocessing often is to delete the erroneous or useless data or columns in the usage profiles via data cleaning. After finishing data cleaning, we next need to extract different users' usage profiles with user identification, using the user's IP column. Each user's website usage profile might include his multiple website usage records within a period of time; hence, we need to divide said user's usage profile into his multiple browsing session log files. After completing said session identification, we still need to face problems related to path complete [4][6][7] and pageview identification [5] during data preprocessing.

In [8] we propose to use STPN to model a website structure, we can further apply the incidence matrix and related properties obtained from the STPN model to help proceed with pageview identification and path complete. The data preprocessing process includes into two parts. (Figure 1).

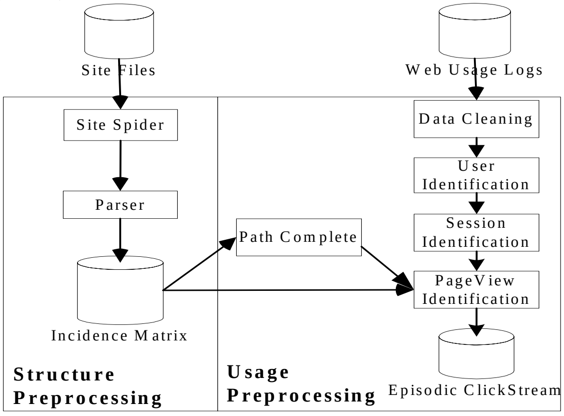

Figure 1: The data preprocessing process