Web Usage Mining with Web Logs

Xiangji Huang
York University, Canada

Aijun An,
York University, Canada

Yang Liu
York University, Canada

INTRODUCTION

With the rapid growth of the World Wide Web, the use of automated Web-mining techniques to discover useful and relevant information has become increasingly important. One challenging direction is Web usage mining, wherein one attempts to discover user navigation patterns of Web usage from Web access logs. Properly exploited, the information obtained from Web usage log can assist us to improve the design of a Web site, refine queries for effective Web search, and build personalized search engines.

However, Web log data are usually large in size and extremely detailed, because they are likely to record every aspect of a user request to a Web server. It is thus of great importance to process the raw Web log data in an appropriate way, and identify the target information intelligently. In this chapter, we first briefly review the concept of Web Usage Mining and discuss its difference from classic Knowledge Discovery techniques, and then focus on exploiting Web log sessions, defined as a group of requests made by a single user for a single navigation purpose, in Web usage mining. We also compare some of the state-of-the-art techniques in identifying log sessions from Web servers, and present some popular Web mining techniques, including *Association Rule Mining, Clustering, Classification, Collaborative Filtering, and Sequential Pattern Learning*, that can be exploited on the Web log data for different research and application purposes.

BACKGROUND

Web Usage Mining (WUM), defined as the discovery and analysis of useful information from the World Wide Web, has been an active area of research and commercialization in the recent years (Cooley, Srivastava, & Mobasher, 1997). In general, as shown in Fig1, the WUM process can be considered as a three-phase process, which consists of data preparation, pattern discovery, and pattern analysis (Srivastava, Cooley, Deshpande, & Tan, 2000).

This process implicitly covers the standard process of Knowledge Discovery in the Databases (KDD), and WUM therefore can be regarded as an application of KDD to the Web domain. Nevertheless, it is distinct from standard KDD methods by facing the unique challenge to dealing with the overwhelming resources on the Internet. To assist Web users in browsing the Internet more efficiently, it is widely accepted that the easiest way to find knowledge about user navigations is to explore the Web server logs. Generally, Web logs record all user requests to a Web server. A request is recorded in a log file entry, which contains different types of information, including the IP address of the computer making the request, the user access timestamp, the document or image requested, etc. The following is an example extracted from the *Livelink* Web server log (Huang, An, Cercone & Promhouse, 2002).[1]

Livelink is a database driven web-based knowledge management system developed by Open Text Corporation (http://www.opentext.com). It provides a web-based environment (such as an intranet or extranet) to facilitate collaborations between cross-functional employees within an organization. In this example, a user using the computer with the IP 24.148.27.239 has requested a query with object ID 12856199 on April 10th, at 7:22pm.

However, when statistical methods are applied to such log data, we tend to get results that are too refined or too specific than they should be. In addition, it is very likely that user browsing behaviour is highly uncertain. Different users may visit the same page for

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 1. Web usage mining process

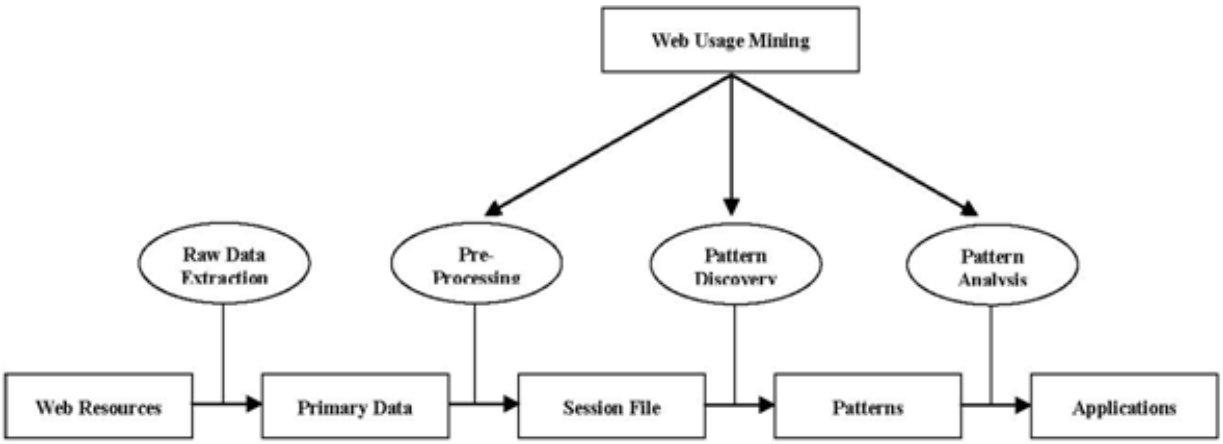

Figura 2. A livelink log entry

```
Wed Apr 10 19:22:52 2002
GATEWAY_INTERFACE = 'CGI/1.1'
HTTPS = 'on'
HTTPS_KEYSIZE = '128'
HTTPS_SECRETKEYSIZE = '1024'
HTTP_ACCEPT_LANGUAGE = 'en-us'
HTTP_CONNECTION = 'Keep-Alive'
HTTP_COOKIE = 'WebEdSessionID=05CAB314874CD61180FE00105A9A1626;
HTTP_HOST = 'intranet.opentext.com'
HTTP_REFERER  ='https://intranet.opentext.com/intranet/livelink.exe?func=doc.
     ViewDoc&nodeId=12856199'
HTTP_USER_AGENT = 'Mozilla/4.0 (compatible; MSIE 5.01; Windows NT 5.0)'
objAction = 'viewheader'
objId = '12856199'
QUERY_STRING = 'func=ll&objId=12856199&objAction=viewheader'
REMOTE_HOST = '24.148.27.239'
REQUEST_METHOD = 'GET'
SCRIPT_NAME = '/intranet/livelink.exe'
SERVER_NAME = 'intranet.opentext.com'
SERVER_PORT = '443'
SERVER_PROTOCOL = 'HTTP/1.1'
SERVER_SOFTWARE = 'Microsoft-IIS/5.0'
04/10/2002 19:22:52     Done with Request on socket 069DC4B0
04/10/2002 19:22:57     Processing Request on socket 09A87EF8
```

different purposes, spend various amount of time on the same page, or even access the same page from different sources. Hence, Web usage mining with original entry/ request-based logs may induce erroneous and worthless results. To solve this problem, Web usage session is introduced as a group of requests made by a single user for a single navigation purpose. A user may have a single session or multiple sessions during a period of time, and each session may include one or more requests accomplishing a single task.

MAIN FOCUS OF THE CHAPTER

Session identification method traditionally suffers from the problem of time threshold setting. Different users may have different browsing behaviours, and their

5 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/web-usage-mining-web-logs/11109

Related Content

Modeling Score Distributions

Anca Doloc-Mihu (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1330-1336).*
www.irma-international.org/chapter/modeling-score-distributions/10994

Text Mining for Business Intelligence

Konstantinos Markellos (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1947-1956).*
www.irma-international.org/chapter/text-mining-business-intelligence/11086

Efficient Graph Matching

Diego Reforgiato Recupero (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 736-743).*
www.irma-international.org/chapter/efficient-graph-matching/10902

XML-Enabled Association Analysis

Ling Feng (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 2117-2122).*
www.irma-international.org/chapter/xml-enabled-association-analysis/11112

Learning Temporal Information from Text

Feng Pan (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1146-1149).*
www.irma-international.org/chapter/learning-temporal-information-text/10966