FTT Algorithm of Web Pageviews for Personalized Recommendation

Shen Yunfei¹, Qin Zheng¹, Yuan Kun², and Luo Xiaowei³

¹ Software School of Tsinghua University, Beijing, 100084, China
  utcloud@163.com, qingzh@mail.tsinghua.edu.cn
² Research center of University of Science and Technology of China,
  Anhui Hefei, 230062, China
  cloudskysea@163.com
³ Department of Construction Management, Tsinghua University,
  Beijing 100084, China
  xiaowei99@mails.tsinghua.edu.cn

Abstract. As the need for personalized services sharply increases caused by the
booming of Internet, Web-based data-mining is becoming a valuable sources of
thoughts and theory to satisfy the personalized system function. The characters
of personalized data-mining is reviewed and discussed in the beginning, and
then an innovative algorithm (FP-Tree time—validity algorithm ) of Web
pageviews, based on personalization, is raised. More authentic information can
be efficiently got by adding time-validity coefficient to FTT-Tree storage
structure to implement increment mining.

Keywords: Data mining, Web mining, Personalization, Association rule, Time-
validity.

1 Introduction

While Internet rapidly expands, lots of Websites turns to offer exact information to
specific users conveniently and speedily through re-organizing their information
service pattern. Correlation rules, one of the data-mining measures, acts important
roles during the process of searching users’ interest hidden behind data and
correlations between them to forecast the trend of the development.

2 Literature Reviews

R. Agrawal and others gave the definition¹ in 1993, making it possible to search
association rules in mass transaction data. On the basis of association rules, they
introduced a more efficient Apriori Algorithm in 1994². But Apriori Algorithm needs
to scan the database many times and may bring lots of candidate sets. To solve the
problem, Han and others³ introduce Fp-Tree structured algorithm in 2000, which
improves the data-mining efficiency obviously.

Universities and academic institutes all over the world proposed many helpful
ideas in research on web log. PageGather⁴ selects unlinked but attractive web pages

© Springer-Verlag Berlin Heidelberg 2006
for clustering by analyzing users' behaviors; AVANTI\(^5\) ask users to raise their interests first, then based on the known users' information, he forecasts not only the next web page which specific user will visit, but the purpose user surf the websites.

Since traditional association rules seldom consider data time-validity while valuable information contained in data record is of great time-validity, some data's value may change as time passes. An improved data-mining algorithm (FP-Tree time-validity algorithm) based on association rules is introduced in the following paragraph. FP-Tree time-validity algorithm (FTT) uses FP-tree structure to storage frequent item sets, and with time-value coefficient added, produce personalized recommendation system with desirable time-validity.

3 Formularized Description of Web Log Data

3.1 Association

Web log data can be expressed by quadruple:

\[ Z=<U,S,T,D> \]

In the quadruple, \( U \) refers to Visitor (usually represented by IP or username); \( S \) refers to previous sets of visit; \( T \) refers to time of visit; \( D \) refers to intention sets of visit. Web log session is generated by visitors and then saved with precision of seconds. Web log data is always the only one that can determine what the users' sessions contain. Via analyzing the preceding quadruples, this paper discloses the trend of users' interest changes during a specific period.

Example 1: Weblog (shown in table 1)

<table>
<thead>
<tr>
<th>U</th>
<th>S(url)</th>
<th>T</th>
<th>D(url)</th>
</tr>
</thead>
<tbody>
<tr>
<td>143.15.16.20</td>
<td>/admissions/</td>
<td>2002-03-11 15:8:57</td>
<td>/admissions/_vti_cnf/general.asp</td>
</tr>
</tbody>
</table>

**Table 1.** Weblog Example

**Definition 1.** Session predication \( \varphi \) ---a property expression of session, which can be formulated when expressing web log’s quadruple: \( Z=<U,S,T,D> \).

Preceding process was a sample of data-mining for a single user session. Using the method\(^6\), we can get all association rules from session files consisted of all users’ sessions. All association rules applied in the personalized recommendation system have following definitions\(^[7][8]\):

**Definition 2.** If \((d_1,d_2,d_3,d_4,d_5)\in S, dj\in D, (d_1,d_2,d_3,d_4,d_5) \rightarrow dj\) dj is a forward rule with length of 1 and \((d_1,d_2,d_3,d_4,d_5)\) is a backward rule with length of 5.

3.2 Episode Rule

Web log data is composed of ordered time sequence and characterized with obvious time marker: timestamps.