Chapter XIV
Improving Web Clickstream Analysis:
Markov Chains Models and Genmax Algorithms

Paolo Baldini
University of Pavia, Italy

Paolo Giudici
University of Pavia, Italy

ABSTRACT

Every time a user links up to a Web site, the server keeps track of all the transactions accomplished in a log file. What is captured is the “click flow” (clickstream) of the mouse and the keys used by the user during the navigation inside the site. Usually every click of the mouse corresponds to the viewing of a Web page. The objective of this chapter is to show how Web clickstream data can be used to understand the most likely paths of navigation in a Web site, with the aim of predicting, possibly online, which pages will be seen, having seen a specific path of other pages before. Such analysis can be very useful to understand, for instance, what is the probability of seeing a page of interest (such as the buying page in an e-commerce site) coming from another page. Or, what is the probability of entering (or exiting) the Web site from any particular page. From a methodological viewpoint, we present two main research contributions. On one hand, we show how to improve the efficiency of the Apriori algorithm; on the other hand, we show how Markov chain models can be usefully developed and implemented for Web usage mining. In both cases, we compare the results obtained with classical association rules algorithms and models.

INTRODUCTION

In the last few years, the number of people that have used the Internet has enormously increased. Companies promote and sell their products on the Web, institutions provide information about their services, and single individuals exploit personal Web pages to be introduced to the whole Internet community.
We will show how the information, concerning the order in which the pages of a Web site are visited, can be profitably used to predict the visit behaviour at the site itself.

Every time a user links up to a Web site, the server keeps track of all the actions accomplished in a log file. What is captured is the “click flow” (clickstream) of the mouse and the keys used by the user during the navigation inside the site. Usually every click of the mouse corresponds to the viewing of a Web page. Therefore, we can define the click-stream as the sequence of the Web pages requested. The succession of the pages shown by a single user during navigation inside the Web identifies a user session. Typically, the analysis only concentrates on the part of each user session concerning the access at a specific site. The set of the pages seen, inside a user session, coming from a determinate site, is known with the term server session.

All this information can be profitably used to efficiently design a Web site. A Web page is well designed if it is able to attract users and address them easily to other pages within the site. A very important area in Web mining is the application of data mining techniques to discover usage patterns from Web data in order to optimally design a Web site, and to better satisfy needs of different visitors. This problem is known as Web usage mining, in contrast to Web context mining (analysis of the content of Web sites) and Web structure mining (analysis of Internet links): for more details on this see, for instance, Baldi et al (Baldi, Frasconi, & Smyth, 2003) or Chakrabarti (2003).

The objective of our analysis is to use Web clickstream data to understand the most likely paths of navigation in a Web site, with the aim of predicting, possibly online, which pages will be seen, having seen a specific path of other pages before. Such analysis can be very useful to understand, for instance, what is the probability of seeing a page of interest (such as the buying page in an e-commerce site) coming from another page. Or, what is the probability of entering (or exiting) the Web site from any particular page.

The most frequent type of statistical analysis of Web clickstream data is the search of the most interesting association and sequence rules (see, for an introduction, Han & Kamber, 2000 or Hand, Heikki, & Smyth, 2001); this search is accomplished by means of the well known Apriori algorithm (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1995). Our research proposal is twofold: it improves both the statistical analysis, by considering different Markov chain models; and the computational search algorithm, considering a Genmax type proposal.

According to what typically is done in the data-mining literature and practice, we shall compare our proposal with standard approaches by means of a real case study. In the description of the analysis, we shall follow the steps of the data mining process as described, for instance, in Berry and Linoff (1997), Giudici (2003) or Hastie et al. (Hastie, Tibshirani, & Friedman, 2001).

The database from which we start to illustrate our methodology is the result of the elaboration

<table>
<thead>
<tr>
<th>(c_value)</th>
<th>(C_time)</th>
<th>(c_caller)</th>
</tr>
</thead>
<tbody>
<tr>
<td>70ee683a6df...</td>
<td>14OCT97:11:09:01</td>
<td>Home</td>
</tr>
<tr>
<td>70ee683a6df...</td>
<td>14OCT97:11:09:08</td>
<td>Catalog</td>
</tr>
<tr>
<td>70ee683a6df...</td>
<td>14OCT97:11:09:14</td>
<td>Program</td>
</tr>
<tr>
<td>70ee683a6df...</td>
<td>14OCT97:11:09:23</td>
<td>Product</td>
</tr>
<tr>
<td>70ee683a6df...</td>
<td>14OCT97:11:09:24</td>
<td>Program</td>
</tr>
</tbody>
</table>
Related Content

Improvement of Organizational Knowledge Transfer through Integration of Functional Silos in Smart Network: Case Study of Public Enterprises
Ivan Todorovi, Mladen udanov and Stefan Komazec (2015). Knowledge Management for Competitive Advantage During Economic Crisis (pp. 299-309).
www.igi-global.com/chapter/improvement-of-organizational-knowledge-transfer-through-integration-of-functional-silos-in-smart-network/117855?camid=4v1a

A Case Study: Closing the Assessment Loop with Program and Institutional Data
Robert Elliott (2012). Teaching Cases Collection (pp. 326-338).
www.igi-global.com/chapter/case-study-closing-assessment-loop/60858?camid=4v1a

Mining Clinical Trial Data
Jose Ma. J. Alvir and Javier Cabrera (2007). Knowledge Discovery and Data Mining: Challenges and Realities (pp. 31-47).
www.igi-global.com/chapter/mining-clinical-trial-data/24900?camid=4v1a

The Birth of a Survey: Developing an Assessment of Preservice Teachers’ Diversity Awareness
Mary Cain Fehr, Mary Frances Agnello, Steven M. Crooks, Fethi Inan and Raymond Flores (2012). Teaching Cases Collection (pp. 288-304).
www.igi-global.com/chapter/birth-survey-developing-assessment-preservice/60855?camid=4v1a