A Probabilistic Model for Web Usage Mining

NICOLETA DAVID*, LUCIAN PATRASCU**, ADELA SASU***, DANIELA DAMIAN*
Department of Mathematics, Informatics and Socio-Human Sciences*, Department of Economics**, Department of Mathematical Analysis and Probabilities***
“George Baritiu” University*, “Transilvania” University***
Lunii Street, No. 6*, B-dul Eroilor No. 29***, Brasov
ROMANIA
nicoletadavid@yahoo.com, luccianpatrascu@yahoo.com, adelasasu@yahoo.com, dana.damian@yahoo.com

Abstract: - In this paper we propose the DTTS method for determining the typical trails in a web site. We use a probabilistic model for a web site and we use the entropy of a Markov chain in order to compute the length of the typical trails.

Key-Words: - Entropy, Markov Chain, Web Application, Web Site, Web Usage Mining, Typical Trail

1 Introduction
Web usage mining is defined as the application of data mining techniques to discover user web navigation patterns in web log data. Log files provide a list of the page requests made to a given web server in which a request is characterized by the IP address of the machine placing the request, the date and time of the request and the URL of the page requested [15].

From this information it is possible to reconstruct the user navigation sessions within a web application, where a session consists of a sequence of web pages viewed by a user in a given period of time.

Several authors have proposed the use of Markov models to model user requests on the web. Since many papers use the simple Markov models, another proposes the n-gram model approach [3, 13, 15, 19].

In [16] the authors present a statistical analysis of models of different orders, respectively 1-gram, 2-gram, etc. The experimental results have shown that the shorter navigation paths are more relevant for the estimation of the conditional probabilities.

In [12] is presented a theoretical approach of the users’ navigation on the web in terms of entropy and an estimation of the navigation paths in order to obtain that entropy.

The paper is organized in the following way: in section 2 we present an overview of the web applications technologies, in section 3 we present a probabilistic model for web usage mining. In section 4 we give an experimental evaluation of our typical trails length estimation using the Markov model presented in section 3. The experiments were conducted using data collected from the log files of an e-commerce web site during one week. In section 5 we present the method DTTS for determining the typical trails in a web site. Finally, in section 6 we present our conclusion and future work.

2 An Overview of Web Application Technologies
Web Applications are software which can be accessed via the Internet or an Intranet using a web browser [6], [9].

Web applications are not limited to one type of application. They can range from simple static web pages to sophisticated applications.

Generally, the tasks executed by the web applications are the following:
1. Web applications offer an interface for requests, respectively an interface used by the users for data inserting. Examples of requests are the following: searching in a database, ordering a book, requesting of a file, etc.
2. Transmission of the requests defined by users, meaning that the data inserted by users are sent to the Web server;
3. Processing execution of the data on the server, meaning that the Web server is processing the user’s data;
4. Transmission of the requests’ results – processed data are sent to the client;
5. Data processing – data returned by the Web server are displayed in browser as HTML code.

The technologies and tools used in Web Applications development are split in the following categories:
• Display Layer (GUI-Graphical User Interface)

Proceedings of the 8th WSEAS International Conference on Telecommunications and Informatics

ISSN: 1790-5117 129 ISBN: 978-960-474-084-0
• Communication Layer
• Middle Layer
• Data Layer

In figure 1 these conceptual categories are presented.

Display Layer is the application interface that executes tasks of translating the user actions into executable code. The browser displays the information in HTML code and also sends the user information inserted in HTML forms.

JavaScript language is used for animation, validation of client forms and events handling. VBScript is a programming language similar with JavaScript and is used for including the controls in a web application.

A cookie is a file transferred by a web site to the users’ computer. Cookies are used for authentication, for example in the case of e-commerce web sites they are used from simple navigation to product transaction.

Fig. 1 Technologies used in web applications

The transmission of information is made by the Communication Layer. The information must be accurate, so the software and hardware must follow some rules named protocols. There are three types of communication protocols used for web applications:

1. computer-computer communication, executed by the IP protocol;
2. text encryption, executed with the encryption keys;
3. software-software communication.

Generally, the TCP protocol gets the information from the hardware level to the software level; the software packs contain communication protocols, like HTTP or SSL for encrypted communication.

The Middle Layer accepts and processes the requests by using the resources offered by the server or the server’s network. The tools used for accessing the system resources are the following:

1. CGI (Common Gateway Interface) is the most important instrument used for accessing the system resources. CGI is a service offered by all the servers that permits to create the executable scripts, to transmit the GET and POST data to the CGI script and to filter the answers generated by CGI.
2. The SSI (Server Side Includes) concept means including some special tags into HTML document that are interpreted by the web server in the same time when the HTML document is sent to the browser.
3. ASP (Active Server Pages) are among the first web applications developing tools that integrated the execution of the application directly in the web server. The scope of this integration was to obtain performance as against using the CGI scripts. ASP use VBScript or JavaScript.
4. COM/ActiveX (Component Object Model) is an object-oriented model developed by Microsoft that defines the way the objects interact each other in a web application.
5. DCOM (Distributed Component Object Model) is an extension of a COM model that helps the objects communication and manipulation in network and on Internet.

Data Layer displays the data sent by the web server in HTML format in browser.

XML (Extensible Markup Language) is a set of specifications used for creating a personalized markup language. It is an extensible language because permits the user to define their own elements.

CSS (Cascading Style Sheets) is a standard for formatting the HTML document elements, like color, text font, text size, document layout. CSS is used for the definition of the styles in an HTML document. The extension of the CSS standard is XSL (eXtensible Style Sheet Language) that contains a language for transformation of the XML documents.

3 A Probabilistic Model for Web Mining

In this section we present a probabilistic web usage mining model for web applications. We propose a method for determining the typical trails length in a
Let us consider a Markov chain \( M=(X, P, \mu) \) defined by:

a) A set of states \( X = \{X_i\}_{i=1,n} \);

b) A transition matrix \( P = \{p_{ij}\}_{i,j=1,n} \);

c) A vector of initial probabilities \( \mu = \{\mu_i\}_{i=1,n} \).

In [2] the authors propose to model a web application in terms of a Markov chain in the following way: the set of states \( X = \{X_i\}_{i=1,n} \) represents the pages of the web application, the probability of the transition from the page \( X_i \) to page \( X_j \) is estimated by the ratio of the number of times the corresponding sequence of pages was traversed and the number of times the anchor page was visited, and the initial probability of a state is estimated as the proportion of times the corresponding page was requested by the user.

The Markov property in this model is that the choice of the following page depends only of the current page visited by the user.

In figure 2 is presented this Markov model. The pages are organized on levels like the following: on 1st level there is the homepage, on 2nd level there are the pages that have in-links from homepage, on the level \( j \) there are the pages that have in-links from the pages from the level \( j-1 \), etc.

**Fig. 2 A Markov model for web applications**

**Definition 1:** A trail in \( M \) is a finite sequence \( T = \{X_i, X_{i+1}, ..., X_j\} \) of states in \( M \) such that \( \{X_i, X_{i+1}\} \in \{1, ..., t-1\} \) are consecutive pages in the web site.

The probability \( p(T) \) of a trail \( T \) of length \( t \) is given by:

\[
p(T) = P_{k_1} P_{k_2} P_{k_3} ... P_{k_t},
\]

where \( k_1, k_2, ..., k_t \) is a permutation of \( 1, 2, ..., t \) [2].

**Definition 2:** A typical trail in a web site is a trail \( T \) whose estimated probability is above a specified cut-point \( \theta \in (0,1) \).

In [2] the cut-point is composed of two distinct thresholds, with \( \lambda = \theta \delta \), where \( \theta \in (0,1) \) is the support threshold and \( \delta \in (0,1) \) the confidence threshold.

We calculate the two parameters in the following way: \( \theta = \frac{1}{n} \), where \( n \) is the number of states in this model, respectively \( \delta = \frac{1}{m} \), where \( m \) is the average number of out-links on a page.

In [7] we have proposed an estimation of the length of the typical trails in a web site like the following:

\[
t_{\lambda} \leq -\frac{\ln \lambda}{H(M) \ln 2}, \lambda \in (0,1)
\]

where \( H(M) \) is the entropy of the Markov chain [4], [17].

### 4 Experimental Evaluation

In this section we present the experimental evaluation of the navigation trails collected from the log files of an e-commerce web site having 10 states.

We have collected the data during one week, between 10.02.2007-17.02.2007 and we have obtained the length of the typical trails \( t_{\lambda} \leq 5 \) for \( \lambda = 0.03 \).

In figure 3 we present the frequencies of the lengths of typical trails for the collected data in which we pruned out the pages rarely visited and the transitions with low probability.

**Fig. 3 The lengths of typical trails for the data set 10.02.2007-17.02.2007**

One can observe that for trails lengths less or
equal with 5 we have obtained the biggest frequencies. The conclusion is that the estimation of the length $\tau$ is correct.

Another remark is that the users prefer shorter trails than the longer ones during the navigation on a web site.

Our estimation of the length of the typical trails in a web site can improve the web mining algorithms that search behavior patterns in log data.

5 DTTS Method

In this section we will use the estimation of the typical trails length in the formula (2) from Section 2 in order to propose a method for determining the typical trails in a web site. The DTTS (Determining the Typical Trails in a web Site) method can be used for the optimization of the web sites’ structure.

Several authors have proposed methods for improving the structure of web sites. In [20] the authors propose an algorithm for discovering the locations (pages) where the users expect to find their target pages. An algorithm for selecting expected locations and to optimize the benefit to the web site or the visitor is presented.

Our DTTS method uses the entropy of Markov chain entropy and discovers the typical trails in a site, that are the navigation trails having the length above a specified cut-point $\lambda \in (0,1)$.

In figure 4 we present the calculation of the parameter $\lambda \in (0,1)$ according with the number of pages $n$ and the average number of out-links on a page $m$.

5.1 DTTS Method

Input:
- $n$: number of the pages in web site
- $\mu$: the average number of out-links on a page

Compute $\lambda$:
\[
\lambda = \frac{\mu}{n}
\]

output $\lambda$

The DTTS method has the following steps:

6 Conclusion

Many authors have been studying the problem of mining web usage patterns from log data. Patterns inferred from past user navigation behavior in a web site are useful to provide insight on how to improve the web site structure and design.

The Markov models have been used in modeling the user behavior on the web. From simple Markov models to models of $n^{th}$ order, known as $n$-gram, there is a lot of analysis about mining web usage.

In this paper we have presented a Markov chain model for an e-commerce web site. Our central approach was the typical navigation trails on a web site. We have used the entropy of a Markov chain concept in order to estimate the length of the typical navigation trails. We have proposed the DTTS method for determining the typical trails using the estimation from formula (2).

Our experimental evaluation has been conducted on a web site having 10 states and has shown that the estimation is correct.

As future work we propose to use the method DTTS in order to improve the web sites structure.

References:


