Probability Measure of Navigation pattern predition using Poisson Distribution Analysis

Dr.V.Valli Mayil
Director/MCA
Vivekanandha Institute of Information and Management Studies
Tiruchengode

Ms. R. Rooba, Asst. Professor
Dept. of Computer Technology
Kongu Arts and Science College
Erode – 638 107

Ms. C. Parimala, Asst. Professor
Dept. of Computer Applications
T.John College
Bangalore – 560 086

Abstract

The World Wide Web has become one of the most important media to store, share and distribute information. The rapid expansion of the web has provided a great opportunity to study user and system behavior by exploring web access logs. Web Usage Mining is the application of data mining techniques to large web data repositories in order to extract usage patterns. Every web server keeps a log of all transactions between the server and the clients. The log data which are collected by web servers contains information about every click of user to the web documents of the site. The useful log information needs to be analyzed and interpreted in order to obtain knowledge about actual user preferences in accessing web pages. In recent years several methods have been proposed for mining web log data. This paper addresses the statistical method of Poisson distribution analysis to find out the higher probability session sequences which is then used to test the web application performance.

The analysis of large volumes of click stream data demands the employment of data mining methods. Conducting data mining on logs of web servers involves the determination of frequently occurring access sequences. A statistical poisson distribution shows the frequency probability of specific events when the average probability of a single occurrence is known. The Poisson distribution is a discrete function wich is used in this paper to find out the probability frequency of particular page is visited by the user.

1. Introduction

Quantitative assessment of navigational behavior is a fundamental task to understand the phenomenon of web navigations. Quantitative measures of user behavior will provide a better characterization of user navigation and this will, in turn, suggest better ways of designing the structure of web sites. The information of web access paterns can be generated from log files via a cleaning process, from which a set of navigation sessions or trails are identified. Quantitative operations can be performed on session information which predicts important characterization of navigation behavior. the complete web site usage statistics can be availed by analysing web site visitor profile and access behavior. Several tools have been proposed to view reports on accessed site's resources, visitors' activity and

navigation, sites that refer web traffic to you, search queries, search engine spiders, user browsers and operating systems, web server errors and much more.

The authors in [1,2] adopt a matrix-theoretic approach in modeling web log data and propose a set of algebraic operators, collectively called navigation operations, which can be employed to manipulate navigation matrices. The information of web usage can be generated from log files via a cleaning process, from which a set of navigation sessions that represent the trails are formed during the navigation process. The trails are modeled as a weighted directed graph, called a transition graph, and then a corresponding navigation matrix is computed with respect to the underlying web topology. The author in [1] defines a minimal set of binary operations, includes sum, union, intersection and difference operations on the matrices. These operations enable the user to analyze navigation from the contents of two given navigation matrices.

In ref [3] the author applies descriptive statistical, network and graph analysis methods on user behavior data to derive user profiles. For graph analysis, the log file is first converted to an adjacency matrix that represents the usage pattern of a web site for a certain user. The matrix shows which pages were requested by the user and how the user got to the different pages. The graph structure is created with node and arcs, where node and arc represents web pages and links respectively. Various network and graph analysis methods are applied on the structure to derive quantitative measures of navigation patterns.

The authors E. Pitkow & P. Pirolli [5,6] used the longest repeated sequence algorithm to predict user surfing behavior and Wang & Zaïane [7] employed a sequence alignment algorithm to cluster user web navigation sessions. Session sequences are represented in a Markov model and various probability measures of navigation pattern is analysed under Markov model

Many authors proposed various statistical, network and probability analysis methods. The part of the thesis work discussed in this chapter uses a statistical and poisson distribution analysis methods to derive the probability measures from the log file. The integration of statistical and probability analysis is an important part of navigation access profile derivation

2. Simple Navigation Metrics

Statistical techniques are applied on preprocessed log file to obtain descriptive session information. For each user session, number of web pages visited in the particular session and amount of bytes transferred in a web page are calculated. The simple navigation metrics includes dwelling time of each web page in a session. The analysis work begins with statistical methods and calculates the frequency of the individual pages and the time spent on each page. The time factor is the most meaningful factor in the analysis and a positive correlation of time spent on a web page and user interest has already been identified in [62]. The work measures the dwell time between each page in a session and total time spent on each session.

3. Probability Evaluation of Log files using Poisson Distribution

A Poisson Process is a stochastic process which consists of a collection of (random) points in time. An example of a Poisson process is the points of time where customers arrive in a shop. The concept of a Poisson process can be generalised to processes with points in arbitrary sets (instead of points in time).

Poisson distribution is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event. It gives theoretical probabilities and theoretical frequencies of a discrete variable. This distribution can be applied when the happening of the event must be of two alternatives such as success or failure. It is applicable when the number of trails ‘n’ is very large. Examples of events that may be modeled as a Poisson distribution include: The number of phone calls at a call centre per minute, the number of times a web server is accessed per minute and the number of mutations in a given stretch of DNA after a certain amount of radiation.

The probability distribution of a random variable X representing the number of successes occurring in a

given time interval or a specified region of space is given in the formula (4.1):

$$P(X) = (m^x e^{-m})/x! \text{ , where } x=0, 1, 2, 3\ldots \quad (4.1)$$

Where 'e' is the base of the natural logarithm (e = 2.71828...). The variable 'x' is the actual number of successes that result from the experiment. The variable 'm' is average number of successes in the given time interval or region of space. The poisson distribution is applied for web log data, since it contains large volume of web page hits. The method is used to find the probability measure of each page visited against number of times in the web log.

4.Experimental Results

The goal of the work in this section is to find the probability of occurrences of every web pages using poisson probability technique. The method finds the probability of number of page occurring in a fixed time period.

The experiment is conducted on four days web log transactions of Kongu Arts and Science college web server from 14.10.07 to 17.10.07 are collected and preprocessed with the data cleaning code. The cleaned log records are converted in the session sequence format which contains <session no, page sequences, number of web pages in the session>.

The following table 1shows the session details of the web log transactions. In order to find the poisson probability, it requires in finding the frequency of 1,2,…n time occurrences of every web pages in each session in the given time period. Using Poisson distribution the expected frequency of 4 times occurring of each web page is calculated. The poisson probability result is shown in table 1.

Table 1 Session sequences

Session No.	Sequence of pages	Freq. web Pages
1	Magazine.html	1
2	Main.html, course.html, biodept/bscbio.html, biodept/bscbio.html	4
3	Main.html, college.html	2
4	Magazine.html, main.html	2
5	Aicte.html	1
6	Main.html	1
7	College.html	1
8	Phdhistory.html	1
9	Mphileconomics.html	1
10	Biodept.html	1
11	College.html	1
12	Cspgdept, cspgdept/cspg.html, cspgdept/cspg.html	3

The result in table 2 shows the probability of occurrence of all web pages of the web site www.kasc.ac.in. The entries in the table 2 show that 1-time hit occurrence of magazine.html page is available at two different sessions. The frequency of 1-time, 2-time and 3-time occurrences of every page in the site is reported in the table 4.2. This poisson metrics gives the probability of 4-time occurrence of each page value and is reported in the table 2. The graph in figure 1 shows the probability of occurrence of web pages in the web site www.kasc.ac.in. The graph result shows the probability metrics that can predict the navigation behavior of web users.

Table 2 Poisson Probability Evaluation

	Frequency of occurrences in the session			
	1-time occurrences	**2- time occurences**	**3- time occurences**	**Poisson value for x=4**
Magazine.html	2	0	0	0.9998278
Main.html	2	1	0	0.9963401
Courses.html	1	0	0	0.9999933
Biodept/bscbio	1	0	0	0.9999933
Biodept/bio	0	1	0	0.9998278
College.html	1	1	0	0.9989353
Aicte.html	1	0	0	0.9999933
Phdhistory…	1	0	0	0.9999933
mphilecon	1	0	0	0.9999933
Cspgdept/…	1	0	0	0.9999933

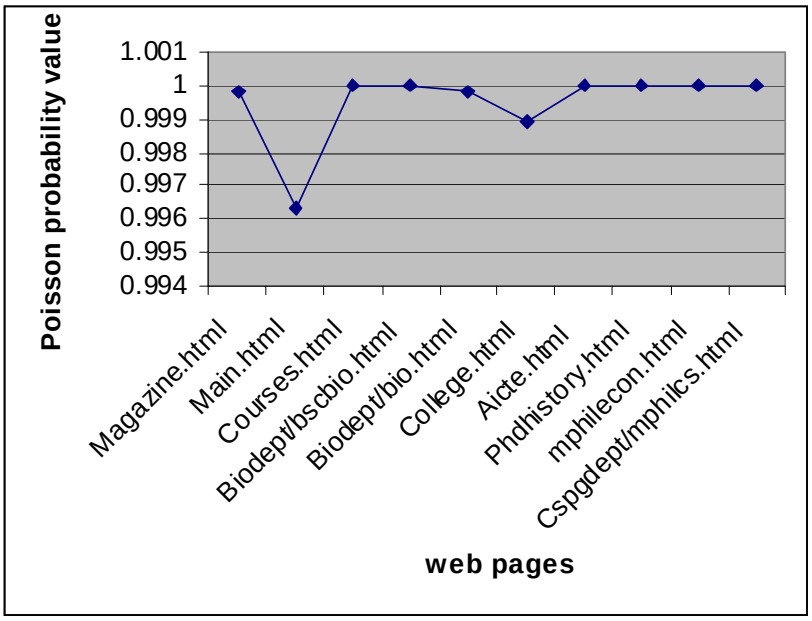

Figure 1 Poisson Probability Analysis of Web Log data

5.Conclusion

Appropriate metrics can provide useful characterizations of user web navigation behavior and can diagnose a variety of problems. The ability to predict the chances of occurrences with precision would be extremely useful in practice. The work proposes a probability analysis of web log file using Poisson distribution. The four days web log transactions from 14.10.07 to 17.10.07 of Kongu Arts and Science College web server have been collected for the Poisson probability analysis. The approach finds the probability and frequency of viewing every page in the website. The figure 4.1 shows that the web pages "magazine.html", course.html", "biodept/bio.html", "aicte.html","phdhostory.html", "mphilecon.html", "cspgdept/mphilcs.html" have more probability value. Hence the probability of occurrences of these pages in the future is higher than the other pages in the web site.

References

[1] Wilfred Ng, "Capturing the Semantics of Web Log data by navigation matrices", Proceedings of the IFIP TC2/WG2.6 Ninth Working Conference on Database Semantics, Pages: 155 – 169, 2001.

[2] N. Zin and M. Levene, "Constructing web-views from automated navigation sessions", In Proceedings of the ACM Digital Library Workshop on Organizing Web Space, pp. 54-58, 1999.

[3] Guandong Xu, Xiaofang Zhou and Yanchun Zhang, "A latent usage approach for clustering web transaction and building user profile, Advanced Data mining and Applications, volume 3584, pp: 31-42, 2005.

[4] Jian Pei, Jiawei Han, Behzad Mortazavi-asl and Hua Zhu, "Mining access patterns efficiently from web logs", In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 396–407, 2000.

[5] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan and A. S. Tomkins, "The web as a graph: measurements, models and methods", Lecture Notes in Computer Science, Vol. 1627, pp: 1-18, 1999

[6] Qiang Yang, Hui Wang and Wei Zhang, "Web-log mining for quantitative temporal event prediction", IEEE Computational Intelligence Bulletin, Vol.1, No.1, December 2002.

[7] W. Wang and O. R. Zaïane, "Clustering Web Sessions by Sequence Alignment", Proceedings of DEXA Workshops, pp: 394-398. 2002.