Uses a Statistical and Poisson Distribution Analysis Method to Derive the Probability Measures from the Log File

J.MALARVIZHI, M.C.A., M.Phil., Scholar, DCS, Bharathiar University, Coimbatore  
E-Mail: malar76@gmail.com.  

Dr. S. THABASU KANNAN, Principal, M.Tech, Ph.D, MBA, FIE., PCET, Sivagangai, Tamilnadu, India  
Mail: thabasukannan@gmail.com

ABSTRACT: Due to increasing the number of Engineering Colleges in Tamilnadu, the level of competition for admission is also increased. By implementing some dynamic strategies only the academic institutions can meet their own competition. One survey clearly states that more than 75% of the Engineering Colleges their strength is less than 30% of their actual intake. Hence the surveillance is the problem for the institutions. One more survey shows that every year 10% of the engineering colleges’ windup their affiliation and approval due to lack of admissions, and 5% of the engineering colleges have decided to sell due to lack of strength. With the strong effort and dynamic strategy framed by the institution, they can even manage in the society. The candidate finds admission in an institution only when their own preference matches exactly, otherwise the candidate continues to go by the next alternate in the list of preference. This paper clearly stresses some factors influenced to identify the pattern for getting the potential of the students to meet at least the breakeven point. 
In addition to the above, the World Wide Webin internet plays an important role to store, share and distribute information about the academic institutions. A social survey states that more than 65% of the admissions gained by their effective web pages. The exponential growth of the World Wide Web has provided an excessive prospect to study the potential student and their behavior by using web access logs. If the institution’s web domain clearly contains the information required for the potential students, surely they can attract the above by which they can get more number of admissions even beyond our jurisdiction. Some of the attractions from the potential students while accessing the web site for getting the admission are:

- get the required information by clicking minimum number of hits from the web pages,
- no web traffic occurred while accessing and navigating the college web site
- Search queries will be rectified within a short period of response time by implementing the practice of search engine optimization, search engine spiders.
- Always use fastest and latest browsers and operating systems in their WWW.
- not to display much more web server errors while navigating the college web site.

For attracting the counseling category and other state students, this web page plays an important role. Web usage Mining is the application of data mining techniques to very large data repositories to extract practice patterns. In general every web server keeps a record of all transactions needed for the potential students and act as a bridge between the potential students and institutions. The record contains full details about every user click to the web documents of the web site. The useful record details needs to be scrutinized and inferred to gather knowledge about actual potential student and their parent preferences in accessing web pages. In recent years several methods have been proposed for mining web log data. This paper its main intention is to use the statistical method of Poisson distribution analysis to find out the higher probability session sequences and also compare the efficiency of our developed algorithm with Poisson value. 

The study of large volumes of click stream data demands the employment of data mining methods. Conducting data mining on records of web servers contains the determination of frequently occurring access sequences. A statistical method of poison distribution clearly shows the probability of frequency of specific events when the average probability of a single occurrence is known. Here the probability of poison value is compared with the efficiency of our developed algorithm. For more number of transactions, our developed algorithm its performance is better than poison value. Because our algorithm extracts the confidence level as dependent rather independent. The Poisson distribution is used in this paper to find out the probability frequency of particular page is visited by the user as independent, but the result of the developed algorithm is dependent.

Keywords: Sequence Alignment Algorithm, Appropriate Model, Statistical Methods, Data Mining Methods, Log Transactions.

I. INTRODUCTION
The assessment of navigational behavior in quantitative manner is a fundamental task to understand the occurrence of web navigations. The quantitative measures of potential student behavior will provide a better characterization of user navigation and this will, in turn, suggest better ways of
designing the structure of the college web sites. The outcome of patterns used for web access can be generated from record files through which a set of navigation sessions or trails are identified. The operations can be performed on session information which predicts important characterization of navigation behavior. The complete web site usage details can be availed by scrutinizing web site potential student profile and their own access behavior. Several tools have been proposed to view reports on accessed site’s resources, potential student activity and navigation, sites that refer web traffic, search queries, search engine spiders, user browsers and operating systems, web server errors and much more.

Some authors 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.

Regarding 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 descriptive statistical, network and graph analysis methods on user behavior data to derive user profiles. For graph analysis, the record 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.

And also the longest repeated sequence algorithm used to predict user surfing behavior and 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 analyzed under Markov model.

Many authors proposed various statistical, network and probability analysis methods. This paper uses a statistical and Poisson distribution analysis method 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.

II. SIMPLE NAVIGATION METRICS

Statistical techniques are applied on preprocessed record file to obtain descriptive session information. For each user session, number of transactions required in the particular session to convert a particular enquiry in to admission and the number of times called for the above conversion. The simple navigation metrics includes dwelling time or lodging time of each transaction in a particular period of time. The analysis work begins with statistical methods and calculates the frequency of the individual transactions and the time spent on each transaction. The time factor is the most meaningful factor in the analysis and a positive correlation of time spent on a transaction and student interest has been identified. The work measures the dwell time between each transaction in a particular period of time and total time spent on each session.

III. PROBABILITY EVALUATION OF RECORD 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 the potential of students arrive in a College. The concept of a Poisson process can be generalized 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 alternatives such as success or failure. It is applicable when the number of trails ‘n’ is very large. Examples of events as a Poisson distribution include: The number of phone calls made to convert an enquiry in to admission in any academic institution, the number of times the institution web site is accessed for admission purpose per minute and the number of times a student can be called for admission after a certain amount of time.

The Poisson distribution may be useful to model events such as

- The number of students admitted in a particular academic year
- The number of occurrences needed to convert the enquiry in to admission.
- The number of students admitting beyond 35 km radius of the college.
- The number calls needed to convert the enquiry into admission
- The dynamic strategy used to convert
- Generalization of the procedure of admission by identifying the pattern.

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 following formula:

\[ P(X) = \frac{(m^x e^{-m})}{x!} \]  

[Equation 1]

Where
- \( e \) - base of the natural logarithm (\( e = 2.71828 \)).
- \( x \) - actual no of successes that result from the experiment (takes values 0,1,2…).
- \( m \) - average no of successes per interval.

\[ P(X = \text{got admission in 10 steps}) = \frac{m^x e^{-m}}{x!} = \frac{1^x e^{-1}}{x!} \]

\[ P(X=0) = \frac{1^0 e^{-1}}{0!} = e^{-1} = 0.368; P(X=1) = \frac{1^1 e^{-1}}{1!} = 0.368 \]

<table>
<thead>
<tr>
<th>X</th>
<th>P(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.368</td>
</tr>
<tr>
<td>1</td>
<td>0.368</td>
</tr>
<tr>
<td>2</td>
<td>0.184</td>
</tr>
<tr>
<td>3</td>
<td>0.061</td>
</tr>
<tr>
<td>4</td>
<td>0.015</td>
</tr>
<tr>
<td>5</td>
<td>0.003</td>
</tr>
<tr>
<td>6</td>
<td>0.0005</td>
</tr>
</tbody>
</table>

The average number of transactions needed to convert the admission for this year is approximately 2.5 and the Poisson model is appropriate. Because the average event rate is 2.5 transactions per admission \( m = 2.5 \).

\[ P(X=0; \text{student in a admission}) = \frac{2.5^0 e^{-2.5}}{0!} = e^{-2.5} = 0.082 \]

\[ P(X=1; \text{student in a admission}) = \frac{2.5^1 e^{-2.5}}{1!} = 0.205 \]

**Table : Probability for 0 to 7 students in an admission**

<table>
<thead>
<tr>
<th>X</th>
<th>P(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.082</td>
</tr>
<tr>
<td>1</td>
<td>0.205</td>
</tr>
<tr>
<td>2</td>
<td>0.257</td>
</tr>
<tr>
<td>3</td>
<td>0.213</td>
</tr>
<tr>
<td>4</td>
<td>0.113</td>
</tr>
</tbody>
</table>

**The Poisson distribution is applied for web record 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.**

**When is the Poisson distribution an appropriate model?**
The Poisson distribution is an appropriate model if the following assumptions are true.

- \( K \) is the number of times an event occurs in an interval and \( K \) can take values 0, 1, 2,…
- The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.
- The rate at which events occur is constant. The rate cannot be higher in some intervals and lower in other intervals.
- Two events cannot occur at exactly the same instant.
- The probability of an event in an interval is proportional to the length of the interval.

If these conditions are true, then \( K \) is a Poisson random variable, and the distribution of \( K \) is a Poisson distribution.

**IV. Experimental Results**
The goal of the work is to find the probability of occurrences of every transactions using poison probability technique. The method finds the probability of number of transaction occurring in a fixed time period. Its performance is compared with our developed algorithm. The experiment is conducted on forty days repository log transactions of Pannai College of Engineering and Technology, Sivagangai web server from 14-05-2015 to 23-06-2015 are collected and preprocessed with the data cleaning code. The cleaned repository records are converted in the session sequence format which contains <session no, no of sequences, number of transactions made in the session>.

The Table-1 shows the session details of the repository log transactions and the probability of occurrence of all transactions of the above college. To find the Poisson probability, it requires in finding the frequency of 1,2,…n time occurrences of every transactions in each session in the given time period. Using Poisson distribution the expected frequency of 4 times occurring of each transaction is calculated. The Poisson probability result is shown in Table-1. The frequency of 1-time, 2-time and 3-time occurrences of every transaction in the college is reported in the table. The result in the following table shows that 1-time hit occurrence
of placement transaction is available at four different ses-

<table>
<thead>
<tr>
<th>S. No</th>
<th>Transactions</th>
<th>No. of occurrences</th>
<th>1-time occurrence</th>
<th>2-time occurrence</th>
<th>3-time occurrence</th>
<th>Poison value for x=4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Moderate fees</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0.9998 278</td>
</tr>
<tr>
<td>2</td>
<td>State of the art lab</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0.9963 401</td>
</tr>
<tr>
<td>3</td>
<td>Library</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0.9999 933</td>
</tr>
<tr>
<td>4</td>
<td>Experienced faculty</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0.9999 933</td>
</tr>
<tr>
<td>5</td>
<td>Good result</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0.9998 278</td>
</tr>
<tr>
<td>6</td>
<td>Placement</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0.9889 353</td>
</tr>
<tr>
<td>7</td>
<td>Hostel facility</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0.9999 933</td>
</tr>
<tr>
<td>8</td>
<td>Sports activity</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0.9999 933</td>
</tr>
<tr>
<td>9</td>
<td>Basic amenities</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0.9999 933</td>
</tr>
<tr>
<td>10</td>
<td>Transport facility</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0.9999 933</td>
</tr>
</tbody>
</table>

For more number of transactions the newly developed algorithm its performance is better, for some transaction the poison value is better and for some transactions both the values are deviated in smaller manner. The confidence level is greater than 100% in our developed algorithm for the transactions of placement, state of the art lab and good result. Because the algorithm its result is dependent i.e the result can be calculated by not only the primary transactions but also adding the result of primary and secondary transactions.

<table>
<thead>
<tr>
<th>S. No</th>
<th>Transaction</th>
<th>Confidence</th>
<th>S. No</th>
<th>Transaction</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Moderate fees</td>
<td>100</td>
<td>6</td>
<td>Placement</td>
<td>127.5</td>
</tr>
<tr>
<td>2</td>
<td>State of the art lab</td>
<td>112.75</td>
<td>7</td>
<td>Hostel facility</td>
<td>86.6</td>
</tr>
<tr>
<td>3</td>
<td>High volume Library</td>
<td>88.5</td>
<td>8</td>
<td>Sports activity</td>
<td>84</td>
</tr>
<tr>
<td>4</td>
<td>Experienced faculty</td>
<td>99.5</td>
<td>9</td>
<td>Basic amenities</td>
<td>98.5</td>
</tr>
<tr>
<td>5</td>
<td>Good result</td>
<td>112.5</td>
<td>1</td>
<td>Transport facility</td>
<td>93.75</td>
</tr>
</tbody>
</table>

But the poison value calculation does not exceed 100%, because it is extracted by the primary transaction only since it is independent.

Table 1 Poisson Probability for admission transaction

Table 2 Confidence value derived from our algorithm

Figure 1: Admission related value derived from Poisson Probability and our new algorithm
V. CONCLUSION

Appropriate metrics can provide useful characterizations of potential student and their parent 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 transaction file using Poisson distribution. The forty days log transactions from 14-05-2015 to 23-06-2015 of Pannai College of Engineering and Technology repository have been collected for the Poisson probability analysis. The approach finds the probability and frequency of viewing every transaction in the College. The Figure-1 shows that the transactions like Placement, Good result, experienced faculty, State of the art lab and Library have more probability value. Hence the probability of occurrences of these transactions in the future is higher than the other transactions in the Engineering Colleges.

REFERENCES


[4] Jian Pei, Jiawei Han, BehzadMortazavi-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.


[14] Haisun Wang, Wei Wang, Jiong Yang, and Philip S. Yu. Clustering by pattern similarity in large data sets. In ACM SIGM