International Journal of Scientific Research in Computer Science and Engineering
Vol.5, Issue.3, pp.119-123, June (2017)

Review Paper

E-ISSN: 2320-7639

# Prediction of User Interest and Behaviour using Markov Model

**Neetu Anand[1*], Tapas Kumar[2]**

[1*]Department of Computer Sciences, Maharaja Surajmal Institute, GGSIP University, Delhi
[2]School of Computer Science and Engineering, Lingayas University, Faridabad, India

**Corresponding Author: neetuanand@msi-ggsip.org, Mob: 9811396950*

**Available online at: www.isroset.org**

Received 20th May 2017, Revised 30th May 2017, Accepted 24th Jun 2017, Online 30th Jun 2017

***Abstract-*** Data mining techniques are foreseeable to be a more expedient tool for analysing user behaviour. A main research area in Web mining focused on learning Web users and their interactions within Websites is Web usage mining. In fact, this research has spread outside of computer science to the management sciences and social sciences due to its importance to business and society as a whole.Children and youngsters have embraced the Internet in conducting their daily activities, and therefore, they use the Internet in ways that differ from elders. While elders tend to use the Internet to check for news, sports, weather, or research products and services, children and young adults are more likely to use the Internet to complete school assignments or play games. And while very high proportions of all age groups – adults and children alike – use e-mail; older children and young adults are doing so at much higher levels. Students are living their lives immersed in technology, 'surrounded by and using computers, videogames, digital music players, video cams, cell phones, and all the other playthings and tools of the digital age'. So it becomes very important to analyze the pattern of access of internet usage of child to identify the next page in the access series which in turn help in behavioural analysis of children.

***Keywords*** **-**Web mining, System monitoring, Behaviour pattern, Markov Model, Pre-processing

## I. INTRODUCTION

The iterative and interactive method of uncovering rational, innovative, valuable, and comprehensible knowledge (in form of patterns, models, rules etc.) from massive databases is called data mining. Over the last decade, it has been seen that emergence of Data Mining (DM) techniques help us in analyzing of structured data.The two goals of data mining are: Prediction (what?) and description (why?).In *predictive data mining* the goal is to build an executable model from data which can be used for classification, prediction or estimation, and in *descriptive data mining the* goal is to discover interesting patterns and relationships in data.Data mining is considered as best for Web data management, including Web documents, Web linkage structures, Web user transactions and Web Semantics. Mining knowledge of various types of Web data helps in realizing and understanding the relationships among various Web objects, and that will be utilized to benefit the improvement of Web data management [1].

Web usage mining is the automatic inventiveness of user access patterns from web servers and tries to discover valuable information from users' transactional data [2]. Using log data behavioural access of user can be determined by using a Markov models. Markov model [3] are widely used to model sequential processes, and have achieved many practical successes in areas such as web log mining, computational biology, speech recognition, natural language processing, robotics, and fault diagnosis.

The task of detecting usage profiles is a systematic activity, it required immense exploitation of all the available clickstreams on computer, then gathering, transforming and analyzing them with the mining techniques. To find out student profiles and make out their behaviour when using a eLearning site, we need to monitor their daily usage and thus collecting all the information about what kind of links they use to follow, services or documents they use to access, their frequency of usage, or simply what are their usual entry points. All of this can be done by observing the web log files. In this paper we will present the way how Markov Model is used in the establishment of usage profiles. Section II, describes Motivation and related work. In Section III we explain Web usage mining, its features, characteristics of web information and Challenges in detail. Section IV deals with Behavioural Modelling using Markov Model. In Section V, we present Conclusion and future scope and finally references are displayed.

## II. MOTIVATION AND RELATED WORK

Billions of users retrieve large amount of data over the internet around the world. These data are recorded in log

 

files, which is very important because user repetitively access the same kind of web pages. These series can be considered as a web access pattern which is helpful to find out the user behaviour. Web page prediction can be categorized into two type's path-based and point-based prediction.

*Path-based prediction* is based on user's previous and historic route data, while point-based prediction is based on currently noticed actions.

*Point-based models precision* is low due to the relatively small amount of information that could be extracted from each session to build the prediction model.

Researchers have used various prediction models including k-nearest neighbour (kNN), ANNs [4], fuzzy inference [5] SVMs, Bayesian model, Markov model and others.

Web usage analysis[6] provides us the means to cover all the essential phases to discover a Web profile based on the page views request sequences records (clickstream data) we have available. Markov models are commonly used method for modelling stochastic sequences with an underlying finite-state structure and were shown to be well-suited for modelling and predicting a user's browsing behaviour on a Web site [7]. Markov chains are especially useful to build prediction models [8], allowing for the establishment of future user behaviour while users are interacting with the sites. This is done with the analysis of previously user's behaviour with similar interests. Siriporn Chimphlee et al. [9] proposed a method for constructing the first-order and second-order Markov models for Web site access prediction based on past visitor behaviour and then compare it with association rules technique.

## III. WEB USAGE MINING

Web usage mining (WUM) can be defined as the application of machine learning techniques on web data for the automatic extraction of behavioural patterns from web users. In this sense, web usage patterns can be used for analysing web user preferences. Traditional data mining methods need to be pre-processed and adapted before being employed on web data. Several efforts have been made to improve the quality of the resulting data. Once a repository of web user behaviour (Web Warehouse) is available [10], specific machine learning algorithms can be applied in order to extract patterns regarding the usage of the web site. As a result of this process, several applications can be implemented on adaptive web sites, such as recommender systems and revenue management marketing, among others. Thus the prediction of users requested pages is an important issue for all web usage applications.

### *A. Feature of WUM*

Web mining has many advantages which make this technology to be used in corporations including the government agencies. This technology has enabled ecommerce to do personalized marketing, which eventually results in higher trade volumes. The government agencies are using this to classify threats and fight against terrorism. It also helps in predicting and identifying illegal activities. The companies can establish better CRM, they can understand the needs of the customer better and they can react to customer needs faster. Web mining technology when used on data of personal nature might cause concerns. The most criticized ethical issue involving web mining is the invasion of privacy [11].

### *B. Characteristics of Web Information*

- Boundless Data: Due to exponential progress of Internet the extensive amount of data volume is generated every day.
- Scattered Data: Web is integrated set of heterogeneous node, so web data is distributed across a wide range of computers or servers, which are located at different places around the world.
- Amorphous Data: There is no uniform format or schema for storing web pages. They are stored according to user defined format and later transferred to the conventional database.
- Vigorous Data: Web data is dynamic and its structure is updated frequently.

### *C. Challenges in Web Usage Mining*

The following challenges were encountered by Information user while interacting with web:

- Uncovering Relatable Data: - People either use the service or surf the web when they want to find specific information on the web. These search tools have several anomalies like short exactitude which is due to irrelevance of many of the search results. The consequences of this are difficulty in finding the relevant information. Additionally there is a problem of low recall due to incapability to index all the information available on the web.
- Crafting new knowledge out of the reports accessible from the web: - This problem is basically sub problem of the above problem. Above problem is query triggered process (retrieval oriented) but this problem is data triggered process that presumes that already has collection of web data and extract potentially useful knowledge out of it.
- Personalization of information: - Interaction of the people with the web differs in contents and presentation they desire.

 

- Learning about distinctive users: - This problem is about what the consumer do and desire. This problem is divided into other sub problem such as customizing the information to the intended consumers or even to personalize it to single user, problem related to web site design, management and marketing.

## IV. BEHAVIORIAL PROFILING

Activity recognition to monitor the behavior of the child is composed of four steps:

- *Internet usage detection*
- *Behavior tracking,*
- *Activity recognition*
- *High-level activity evaluation*.

The concept of behavioural Profiling (also known as "targeting") consists of collecting and analysing several events in log files. Log files records information can be viewed as in the form of client IP address, URL requested etc., in different formats such as Common Log format, Extended Common Log format which is issued by Apache and IIS.

Behavioural profiling involves collecting data (recording, storing and tracking) and searching it for identifying patterns (with the help of data mining algorithms). The data collection phase is often referred to as Behavioural Tracking [12].There are different techniques namely SVM, Markov model, Association rule mining, Markov model with clustering etc. has been used for web page prediction. The next page can be identified by tracing the users visiting behaviour and then extract their interest using patterns.

### *A. Markov Model*

Markov model have realised many real-world accomplishments in areas such as web log mining, computational biology, speech recognition, natural language processing, robotics, and fault diagnosis.

The concept of Markov model is to predict the next action depending on the result of previous actions. In Web prediction, the next action corresponds to predicting the next page to be visited. The previous actions correspond to the previous pages that have already been visited.

Markov models are characterized by three factors (*A, S,T* ), where *A* is the set of all possible *actions* that can be performed by the user; *S* is the set of all possible states for which the Markov model is built; and *T* is a *SX A Transition Probability Matrix* (TPM), where each entry $t_{ij}$ corresponds to the probability of performing the action *j* when the process is in state *i*.

The simplest first-order Markov model predicts the next action by only looking at the last action performed by the user. Second-order Markov model computes the predictions by looking at the last two actions performed by the user. This technique is generalized to the Kth-order Markov model, which computes the predictions by looking at the last K actions. In Web prediction, the Kth-order Markov model is the probability that a user will visit the kth page provided that she has visited the ordered $k - 1$ pages.

Let us consider the example of user navigation session concluded from a log file is modelled as a stream [13], which represents a sequence of requests made by the user within a defined time interval, and is given in tabular form as follows:

Table 1: User navigation Session and Frequency

| Sessions | Frequency |
|---|---|
| F,1,2,3,5,L | 3 |
| F,2,4,6,L | 5 |
| F,1,3,5,L | 6 |
| F,2,3,L | 4 |
| F,1,2,3,L | 2 |
| F,2,3,L | 4 |

In the Table 1, the first column represents a collection of navigation sessions with starting page F and a Last page L. The term 'frequency ' denotes the number of times the corresponding sequence of pages was traversed or visited in the session. These details are then plotted as a weighted directed graph (G) known as Markov model. The Markov model comprises of set of states for all web pages in the sites and a link or edges between two web pages represents page sequence. Each state or web page is defined by the identity number called page number. Each link or edges are denoted by a non-negative number represent number of visits of the pair of pages or page sequences. Each state has the numeric value for web page identity and the ratio value defines the page probability. Each link represents the transition probability that is number of times the link is followed after the anchor page is visited. This ratio is obtained by dividing number of times the page was viewed by the total number of page views. The transition probability is then represented in transition matrix which records the transition probabilities which are estimated by the proportion of times the corresponding link was traversed from the anchor [14,15].

*Markov chains* have been used to model user sequential navigational behaviour on the web site; because the main functionality of Markov chain is that the present state always depends upon the previous state. It is especially useful to build prediction models which allows for the establishment of future user behaviour while users are interacting with the sites. This is done with the analysis of previously user's behaviour with similar interests. More formally, a Markov chain is characterized by a set of states {s1, s2, …sn} and a

 

matrix of probabilities [$Pr_{ij}$] nxm, where $Pr_{ij}$ represents the probability to move from state i to state j[16]. Markov chains have many attractive properties. They can be easily estimated statistically. Since the Markov chain model is also generative, navigation tours can be automatically derived. The shortcoming of Markov chain is very large size of input is required i.e.it is feasible only in relatively small state spaces. Transition matrix created from transition graph is usually of very big size due to more number of web pages.

Hidden Markov models (HMMs) is a variation to Markov process is comprises of two variables: the (hidden) state and the observation. In addition to the transition probabilities, HMMs specify the probability of making each observation in each state. Because the number of parameters of a first-order Markov model is quadratic in the number of states (and higher for higher-order models), learning

## V. CONCLUSION AND FUTUREWORK

The main goal of this study was to understand how the students used the internet resources by a brief explanation about how Markov chains can be used to discovery web usage profiles. Markov models have been widely used to represent and analyze user Web navigation data. The ability of the model to make predictions is important in order to foresee the next link choice of a user after following a given trail. The conclusion based on the literature survey is that various research works had been done to predict user browsing behaviour. A model of past user navigation behaviour can be used to identify frequent usage patterns. In addition, being able to predict the near future navigation intentions of an individual user will enable the provision of pages adapted to the recent behaviour of this user. For future works, we will apply this model to predict the behaviour of children using internet usage log file. With this we can help the parents to restrict the internet usage of their child.

## ACKNOWLEDGEMENT

I am thankful to my PhD thesis supervisor, Professor Tapas Kumar for guiding me in preparing this paper.

## References

[1]. G. Chang, "*Mining the World Wide Web: An Information Search Approach*" In ACM SIGMOD, Vol.31, Issue.2, pp. 69-70, 2002.

[2]. J. Srivastava, "*Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*", In SIGKDD Explorations, Vol.1, No. 2, pp.12-23, 2000.

[3]. M. Hildebrandt," *Profiling: from data to knowledge*", DuD: Datenschutz und Datensicherheit, Vol.30, Issue.9, pp.548-552, 2006.

[4]. O. Asraoui and R. Krishnapuram, "*One step evolutionary mining of context sensitive associations and Web navigation patterns"*, In Proceedings of .SIAM International Conference on Data Mining, Arlington, VA, pp.531–547, Apr.2002.

[5]. O. Nasraoui, C. Petenes, "*Combining Web usage mining and fuzzy inference for Website personalization*", In Proceedings of Web KDD, USA, pp.37-46, 2003.

[6]. B. Bakariya, G.S. Thakur, "*Effectuation of Web Log Preprocessing and Page Access Frequency using Web Usage Mining*", International Journal of Computer Sciences and Engineering, Vol.1, Issue.1, pp.1-5, 2013.

[7]. M. Deshpande, G. Karypis, "*Selective Markov models for predicting web page accesses*", In ACM Transaction on Internet Technology, Vol.4, Issue.2, pp.163-184, 2004.

[8]. Neetu Anand and Mayank Singh, "*A New Approach to Monitor Children's Computer Usage Pattern*", In International Journal of Computer Science and Information Security, Vol. 11, No. 1, pp. 11-14, 2013.

[9]. Gery Mathias , Haddad Hatem, "*Evaluation of Web Usage Mining Approaches for Users Next Request Prediction",* In Proceedings of the 5th ACM international workshop on web information and data management, USA, pp.74-81, 2003.

[10]. Siddu P. Algur, Prashant Bhat, "*Abnormal Web Video Prediction Using RT and J48 Classification Techniques*", International Journal of Computer Sciences and Engineering, Vol.4, Issue.6, pp.101-107, 2016.

[11]. Poonamkaushal, *"Prediction of User's Next Web Page Request By Hybrid Technique"*, International Journal of Emerging Technology and Advanced Engineering, Vol.2, Issue.3, pp.338-342, 2012.

[12]. Yogesh Bhalerao, P. P. Rokade, "A Survey on User Navigation Pattern Prediction from Web Log Data", International Journal of Computer Sciences and Engineering, Vol.3, Issue.5, pp.133-137, 2015.

[13]. Jose Borges, Mark Levene,*"Evaluating Variable-Length Markov Chain Models for Analysis of User Web Navigation Sessions"*, IEEE Transactions on Knowledge and Data Engineering, Vol.19, No.4,pp. 441-452, 2007.

[14]. Marie Fernandes, "*Data Mining: A Comparative Study of its Various Techniques and its Process*", International Journal of Scientific Research in Computer Science and Engineering,Vol.5, Issue.1, pp.19-23, 2017.

[15]. M. Eirinaki, M. Vazirgiannis, "*Web Path Recommendations based on Page Ranking and Markov Models*", Proceedings on 7th ACM International Workshop Web Information and Data Management (WIDM '05), US, pp. 2-9, 2005.

[16]. Alice Marques and Orlando Belo, "*Discovering Student web Usage Profiles Using Markov Chains"*, In Electronic Journal of e-Learning, Vol.9 Issue.1, pp. 63-74, 2011.

## Authors Profile

*Ms.Neetu Anand* pursed Master of Information technology from GJU, Hissar, Haryana. She is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Sciences, Maharaja Surajmal Institute,GGSIP University ,Delhi. She is a member of CSI .She has published more than 15 research papers in reputed International journals and conferences including IEEE,Elsevier and Springer.Her main research work focuses on Web mining, Cloud Security and Privacy, Big Data Analytics and Data Mining. She has 17 years of teaching experience .

 

*Prof. (Dr.) Tapas kumar* pursed B.Tech in CSE from Amravati University,Maharashtra ; Master of Computer Science from Guru Jambeshwar University, Hissar, Haryana and PhD (Engineering) on *"An Application of Cellular Automata Paradigm in Image Processing",* BITS, Mesra, Ranchi. He is currently working as Associate Dean & Head - School of Computer Science& Engineering in Lingayas University,Faridabad. He is a member of ISTE,IAENG and CSI. He has published more than 30 research papers in reputed International Journals including scoupus Indexed Journal and Conferences including IEEE, Elsevier, Springer. His main Areas of Interest includes Cellular Automata, Image Processing, Data Sciences and Cloud Computing.He is currently supervising 8 PhD scholars and 2 PhD theses has been submitted. He supervised 14 M.Tech Thesis and guided more than 100 students of B.E in their research based & application based projects. He has 19 years of teaching experience.

 