Analysis of Clickstream Data using Markov Chains

ISBN: 978-1-943295-14-2

Swapna Datta Khan
Army Institute of Management
(captsdk@gmail.com)

Markov Chains help predict Consumer Behaviour by analyzing the switching process of customers from one brand to another Contemporary predictive analytics enabled by Markov Chains uses Clickstream Data to realize customer preferences in online retailing and online services by predicting the next click or the destined click given a pattern of clicks a user generates. This generic paper studies literature to throw light on the analysis of Clickstream data to study online Consumer Behaviour.

Keywords: Clickstream Data, Markov Chain, Markov Model

1. Introduction

Markov Chains help predict Consumer Behavior with respect to Brand Loyalty by analyzing the switching process of customers from one brand to another. Markovian Decision Process also enables the study of the maintenance condition of deteriorating equipment by analyzing the status of equipment and commenting quantitatively on productivity. Markov Chains facilitate the estimation of the portion of accounts receivable, which will eventually be uncollectible. As regards the Internet, while searching results on Google, the probability to be directed to a certain page is given by the stationary distribution on the following Markov chain on all (known) web pages visited. Contemporary predictive analytics, enabled by Markov Chains, uses data generated from an observed Clickstream to throw light on customer preferences in online retailing and online services by predicting the next click or the "destined" click given a pattern of clicks a user has followed till now. The objectives of this paper is to gain an insight into the contribution of Markov Chains towards the analysis of Clickstream data.

2. Objective and Methodology

Objective of study: To study the contribution of Markov Chains towards the analysis of Clickstream data

Methodology and Scope of Research: The research is a content analysis, post review of associated literature that analyzes relevant Clickstream data using Markov Chains to draw results. The possible findings will enable related analysis of Clickstream data with Markov Chains to understand customer needs, segment the customer base and identify communities of online customers with similar interests. The findings of this research are generic in nature and dos not dwell on any case.

3. Review of Relevant Literature

A Markov chain is a set of transitions, which are determined by some probability distribution, that satisfy the Markov property. The Markov property refers to the memory less property of a stochastic process: the conditional probability distribution of future states of the process depends only upon the present state, not on the sequence of events that preceded it. The different instances of Markov processes for different levels of state space generality and for discrete time vs continuous time are shown in Figure 1 below.

	Countable state space	Continuous or general state space
Discrete-time	(discrete-time) Markov chain on a countable or finite state space	Harris chain (Markov chain on a general state space)
Continuous-time	Continuous-time Markov process or Markov jump process	Any continuous stochastic process with the Markov property, e.g., the Wiener process

Figure 1 *Source: (https://en.wikipedia.org/wiki/Markov_chain)*

The memory less property of a discreet time Markov Chain could be defined as P ($X_n=x_n$/ $X_{n-1}=x_{n-1}$. $X_{n-2}=x_{n-2}$… $X_1=x_1$) = P($X_n = x_n$/ $X_{n-1}=x_{n-1}$); where X_i is a random variable and both conditional probabilities are well defined.

If $\alpha \approx 0.85$, N is the number of known web pages and the page i has k_i links, then it has the following transition probability M_i (to all pages that are linked to)

$$\frac{\alpha}{k_i} + \frac{1-\alpha}{N}$$

The following Figure 2 represents the Page Rank Algorithm with a transition probability of M

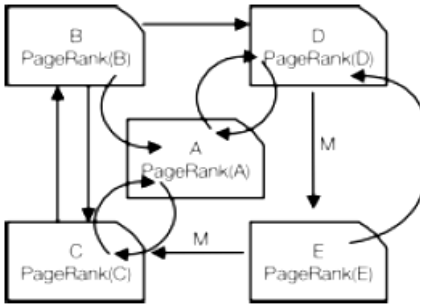


Figure 2 *Source: (https://en.wikipedia.org/wiki/Markov_chain)*

(Soni), (https://en.wikipedia.org/wiki/Markov_chain)
Prior to the Application of Markov Chains, we need to check the following

- Irreducibility or the approachability of one state from another
- Periodicity or whether there are any cycles in the chain
- Presence of an Invariant Probability Distribution

(Ish-Shalom & Hansen)

A click stream is created when a visitor visits a web page. As he moves on to web pages, one after the other, a sequence of "clicks" is created, known as a Clickstream. A Click stream is created when a visitor visits a web page. As he moves on to web pages, one after the other, a sequence of "clicks" is created, known as a Clickstream. Click stream data can show the web pages browsed along with visiting length, retrieval words, ISP, countries and exploration, throwing light on the user's personal preferences and choices. Thus Click Stream Analysis can enable marketing analysts study patterns and gather data on customer requirements and personalize product offerings, reduction of operating costs (by streamlining results, based on customer goals) and improvement of Customer Satisfaction Levels.

Statistical Models and their challenges: A statistical model would estimate relationships between relevant variables using the concepts of causality and correlation. However, the complexity of data and noise and the concept of causality that the click stream must follow a continued path and no loops, creates a significantly large margin of error.
(https://en.wikipedia.org/wiki/Click_path)

Scholz, 2016 in his paper has said that each click event is a character, click streams for a particular session could be modelled as a vector. A collection of clickstream could be mop0delled as a list in R. The package clickstream is an S3 class for storing lists of vectors. An example is depicted in the Figure 3 below.

```
R> cls <- list(Session1 = c("P1", "P2", "P1", "P3", "P4", "Defer"),
+     Session2 = c("P3", "P4", "P1", "P3", "Defer"),
+     Session3 = c("P5", "P1", "P6", "P7", "P6", "P7", "P8", "P7", "Buy"),
+     Session4 = c("P9", "P2", "P11", "P12", "P11", "P13", "P11", "Buy"),
+     Session5 = c("P4", "P6", "P11", "P6", "P1", "P3", "Defer"),
+     Session6 = c("P3", "P13", "P12", "P4", "P12", "P1", "P4", "P1", "P3",
+       "Defer"),
+     Session7 = c("P10", "P5", "P10", "P8", "P8", "P5", "P1", "P7", "Buy"),
+     Session8 = c("P9", "P2", "P1", "P9", "P3", "P1", "Defer"),
+     Session9 = c("P5", "P8", "P5", "P7", "P4", "P1", "P6", "P4", "Defer"))
R> class(cls) <- "Clickstreams"
```

Figure 3 *Source: (Scholz, 2016)*

If Click Stream data is represented as a hierarchical tree in which nodes, links and tree-depth represent web pages, transitions and click number respectively. As tree depth increases, the Markov Assumption (given a certain state, no additional information is needed to predict the next state) viz: P(Xn=xn/ Xn-1=xn-1. Xn-2=xn-2, …, X1=x1) = P(Xn = xn/ Xn-1=xn-1) could be applied. (https://en.wikipedia.org/wiki/Markov_chain)

To convert Click Stream Data to Transition Matrix for computation and fitting of av Markov Model we need to

1. Clean data: This is mostly done by filtering out observations with only one page visit.
2. Compute a matrix with counts of transition from State I to State j
3. Normalize the above mentioned Matrix with Row Sums to form the transition matrix

The said conversion is depicted in the Figure4 below

Bigram Count Matrix

	Frontpage	News	Tech
Frontpage	6	3	1
News	3	3	4
Tech	1	1	8

Normalize

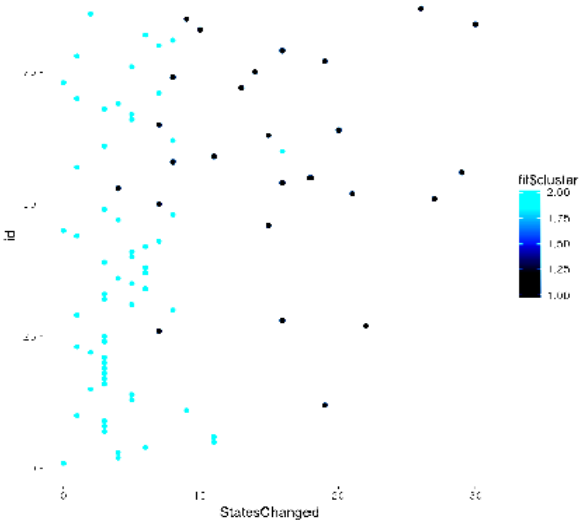
Transition Matrix

	Frontpage	News	Tech
Frontpage	0.6	0.3	0.1
News	0.3	0.3	0.4
Tech	0.1	0.1	0.8

Figure 4 *Source: (Ish-Shalom & Hansen)*

We need yo group together similar Clickstreams and user profiles, thus finding customer segments and identifying communities of users with similar interests. (Markou)

A typical graphic representation of clusters is given in Figure 5 below.



The y-axis represents a unique identifier for each session while the x-axis corresponds to the total number of states changed during each session.

Figure 5 *Source: (Markou)*

Clickstreams are heterogeneous and there is uncertainty during the clustering, making it difficult to partition the data. In their study, Wei, Shen, Sundaresan, & Ma, 2012, felt the need for an interactive data exploration environment and they designed a visual analytic system to support the same. They derived the Self Organizing Map (SOM) to map and cluster the Clickstreams. The SOM is a Neural Network model for high-dimensional data mapping and clustering. It consists of nodes or neurons arranged on a 2 dimensional grid. A node is associated with a vector prototype, which represents the cluster of input data. The SOM gets trained by the competitive learning method, updates the vector prototypes iteratively. The neighborhood relation h(i,j) between two prototypes i and j are determined by their geometric relation between the points (xi, y_i) and (x_j, y_j). The most commonly used neighborhood relation is the Gaussian Function depicted in the below:

$$h(i,j) = \frac{1}{\sqrt{2\pi}\delta} exp(-\frac{||o_i - o_j||^2}{2 \times \delta^2})$$

While using Markov Models, we could have vector prototypes that have the same dimensionality as the input data. The similarity between an input data and a prototype is measured by a pre defined similarity metric such as the Euclidean distance. (Wei, Shen, Sundaresan, & Ma, 2012)

4. Findings

The steps that one could follow to use a Markov Model for the analysis of Clickstream data are

1. Preliminarily check user visits; E.g.: User starts in homepage, views some product information, then bookmarks configuration page for later
2. Define research objectives and label thrusts; E.g.: Dynamic prediction, dynamic classification, clustering, control
3. Data gathering, cleaning and preparation; Identify patterns with cookies
4. Check out complications
5. Content categorization: Summarize URL data using a small number of categories
6. Detect patterns in Training Data
7. Represent sequences of fixed dimension vectors of features
8. Define a distance measure between sequences
9. Find transition matrix and fit an appropriate Markov Mixture model (Bertsimas, Mersereau, & Patel)

5. Conclusions

In a world where online retailing is common, Clickstream analysis can throw enough light of consumer preferences. Though Statistical Models have been used, the complexities of Clickstream data are often handled with finer ease using the Markov Model.

6. References

1. Bertsimas, D., Mersereau, A., & Patel, N. (n.d.). http://ebusiness.mit.edu/sponsors/common/2002-june-wksp-datam/bertsimas.pdf. Retrieved Nov 30, 2019, from http://ebusiness.mit.edu.
2. https://en.wikipedia.org/wiki/Click_path. (n.d.). Retrieved Nov 30, 2019, from https://en.wikipedia.org.
3. https://en.wikipedia.org/wiki/Markov_chain. (n.d.). Retrieved Nov 30, 2019, from https://en.wikipedia.org.
4. Ish-Shalom, S., & Hansen, S. (n.d.). https://pdfs.semanticscholar.org/b9bc/0aeeb511c87620a6e97390981820352b6ba8.pdf. Retrieved Oct 25, 2019, from https://pdfs.semanticscholar.org.
5. Markou, E. (n.d.). https://www.blendo.co/blog/clickstream-data-mining-techniques-introduction/. Retrieved Nov 30, 2019, from https://www.blendo.co.
6. Scholz, M. (2016). R Package clickstream: Analyzing Clickstream Data with Markov Chains. Journal of Statistical Software, 74 (4).
7. Soni, D. (n.d.). https://towardsdatascience.com/introduction-to-markov-chains-50da3645a50d. Retrieved Nov 30, 2019, from https://towardsdatascience.com.
8. Wei, J., Shen, Z., Sundaresan, N., & Ma, K.-L. (2012). Visual Cluster Exploration of Web Clickstream Data. 2012 IEEE Conference on Visual Analytics Science and Technology (VAST). Seattle, WA, USA: Institute of Electrical and Electronic Engineers.