Website Clickstream Data Visualization Using Improved Markov Chain Modelling In Apache Flume

Amjad Jumaah Frhan

PhD Candidate, Department of Telecommunication and Information Technology, University Politehnica of Bucharest, Romania.

Abstract- Clickstream data analysis is considered as the process of collecting, analysing and reporting the aggregate data about the web pages a visitor clicks. Visualizing the clickstream data has gained significant importance in many applications like web marketing, customer prediction, product management, etc. Most existing works employ different tools for visualizing along with techniques like Markov chain modelling. However the accuracy of the methods can be improved when the shortcomings are resolved. Markov chain modelling has problems of occlusion and unable to provide clear display of data visualizing. These issues can be resolved by improving the Markov chain model by introducing a heuristic method of Kolmogorov– Smirnov distance and maximum likelihood estimator for visualizing. These concepts are employed between the underlying distribution states to minimize the Markov distribution. The proposed model named as WebClickviz is performed in Hadoop Apache Flume which is a highly advanced tool. The clickstream data visualization accuracy can be improved when Apache Flume tools are used. The performance evaluation are made on a specific website clickstream data which shows the proposed model of visualization has better performance than existing models like VizClick.

Keywords- Clickstream data, VizClick, WebClickviz, Apache Flume, Markov chain, Kolmogorov–Smirnov distance.

1 Introduction

Click data analytics [1] devices to mine websites, social media and online transactions are helping companies maximize customer interactions. A clickstream is a series of page requests; every page requested generates a flag [2]. These signs can be graphically represented for clickstream reporting. The principle purpose of clickstream taking after is to give webmasters understanding into what guests on their site are doing. There are two levels of clickstream investigation, traffic analytics and e-commerce analytics. Traffic analytics [3] operates at the server level and tracks what number of pages is served to the user, to what extent it takes each page to stack [4], how often the user hits the browser's back or stop catch and how much data is transmitted before the user moves on [5]. E-commerce-based examination [6] uses clickstream data to determine the effectiveness of the site as a channel-to-market. It's concerned with what pages the shopper lingers on, what the shopper puts in or takes out of a shopping basket, what items the shopper purchases, whether or not the shopper belongs to a dependability program and uses a coupon code and the shopper's preferred method of payment [7].

Because an extremely large volume of data can be gathered through clickstream investigation, numerous e-businesses rely on enormous data analytics and related apparatuses [8], for example, Hadoop [9] to help interpret the data and generate reports for specific areas of interest. Clickstream investigation is considered to be best when used in conjunction with other, more standard, market evaluation resources. Inaugurating clickstream or snap way data must be gleaned from server log files. Because human and machine traffic were not differentiated, the investigation of human snaps required a considerable effort. Subsequently, Javascript technologies [10] were developed which use a taking after cookie to generate a series of signs from browsers.

Analysing the information of clients that visit an organization website can be imperative in order to remain competitive [11]. This analysis can be used to generate two discoveries for the organization, the first being an analysis of a user's clickstream while utilizing a website to reveal usage patterns, which thus gives a heightened understanding of customer behaviour [12]. This use of the analysis creates a user profile that guides in understanding the types of people that visit an organization's website [13]. Clickstream analysis can be used to predict whether a customer is likely to purchase from an e-commerce website. Clickstream analysis can also be used to improve customer fulfilment with the website and with the organization itself [14]. This can generate a business advantage, and be used to assess the effectiveness of advertising on a web page or site. Clickstreams can likewise be used to enable the user to see where they have been and enable them to easily return to a page they have already visited, a capacity that is already incorporated in many browsers.

Unauthorized clickstream information collection is considered to be spyware. However, authorized clickstream information collection comes from associations that use select in panels to generate market research utilizing panelists who agree to share their clickstream information with other companies by downloading and introducing specialized clickstream collection agents. VizClick [16] attempted to visualize the website clickstream data using a systematic approach which was performed on www.adobe.com to analyse the market behaviour of customers. However this model does provide only nominal clarity in clickstream data visualization. Hence this paper developed improved Markov chain based clickstream data visualization model named as WebClickviz, which is explained in the following sections. The proposed visualization model utilizes a heuristic determination method in general Markov chain to overcome the issues of display clarity and occlusion. The remainder of the article is organized as: Section 2 discusses some the most related research works. The improved markov chain modelling is discussed in Section 3. Section 4 focuses on the webclickviz visualization methodology while section 5 presents the visualization performance and evaluation results. Finally, Section 6 explains a conclusion about the proposed work.

2 Related Works

Website clickstream data visualization is a step by step procedure by which the user propagation is tracked from the server log files and clickstream files. In [17], an extensive survey has been made to clickstream data analysis. This work discussed about the scientific visualization and information visualization creates graphical models on the KDD process. More than offering resources for interactive visual exploration of databases, visual mapping techniques are presently being used to enhance user interpretation of mining errands and furthermore as an integrated some portion of expository DM calculations. Many mining techniques require user intervention at different stages and representation is beginning to be used to bolster the decision processes involved in making such interventions.

In [18], Moe has proposed an empirical two-stage choice model with the varying decision rules of the clickstream data. The author proposes and applies an empirical two-stage choice model to Internet clickstream information that captures observed choices for two choice stages: items viewed and items purchased. The model takes into account interdependences between choices inside a stage and the use of changing decision rules in each stage. The author accommodates heterogeneity in preferences and in decision rules. The proposed model uses observed choices to infer both attribute preference evaluations and criterion attributes, examinations and criterion attributes.

In [19], the authors proposed a practical methodology for the prediction of demographic web site guest profiles that can be used for web advertising targeting purposes. The methodology involves the change of web site guests' clickstream patterns to a set of features and the preparation of Random Forest classifiers that generate predictions for gender, age, educational level and occupation category. These demographic predictions can bolster online advertisement targeting (i) as an extra contribution to personalized advertising or behavioral targeting, in order to restrict promotion targeting to demographically defined target gatherings, or (ii) as a contribution for aggregated demographic web site guest profiles that bolster marketing managers in selecting web sites and achieving an ideal correspondence between target gatherings and web site audience piece.

In [20], the authors employed a big data approach to discover the user interests in e-commerce. The authors of [21] also employed similar approach to extract customer shopping types from online sites. In [22], the authors introduced VisMOOC, a visual analytic system to help analyse user learning behaviours by using video clickstream data from Massive Open Online Courses (MOOC) platforms. They work closely with the instructors of two Coursera courses to understand the data and collect task analysis requirements. In [23], the authors applied some standard algorithms to CFA prediction in this setting, and showed how one type of behavioural data collected about students – video-watching clickstream events – can be used as learning features to improve prediction quality. This can be taken as motivation for the future researches of clickstream data. Though there have been various techniques been utilized successfully for data analysis, most techniques relied on standard Markov chain. As stated earlier, the drawbacks in standard Markov chain reduces visualization quality and hence this research model focuses on eliminating them.

3 Improved Markov Chain Modelling

The shortcomings of standard Markov chain [24] for the website clickstream data visualization led to the development of the Improved Markov chain. This improved version overcomes the occlusion and display problems by heuristic determination of the grid spacing distributions. The Kolmogorov– Smirnov distance and maximum likelihood estimator are used between the underlying distribution states to minimize the Markov distribution. Considering the probability space $(\Omega, \mathbb{F}, \mathbb{P})$, equipped with a filtration $\mathbb{F} = \{\mathcal{F}(t): t \geq 0\}$. Let the continuous stochastic process $X(t) = \{X_t, t \geq 0\}$ be the solution of the univariate jump-diffusion process

$$dX_t = \mu(X_t;\vartheta)dt + \sigma(X_t;\vartheta)dW_t + \int_{\zeta(X_t)} \eta(X_t, v;\vartheta)P(X_t, dt, dv;\vartheta) \qquad (1)$$

with an preliminary value $X_0 = x_0$, where ϑ denotes the unknown parameter set; $\mu(.)$ and $\sigma(.)$ define the drift and diffusion functions; W_t is the Wiener process; $P(.)$ represents a Poisson random measure with intensity $\mu(X_t;\vartheta)$. Given a mark set ζ, the jump coefficient η has a mark density $\phi\zeta(v, X_t)$.

For a continuous time Markov chain with a finite support, the grid elements are assumed to be monotonically increasing. Let h denotes the grid spacing between two adjacent grid elements on a n grid points Markov chain while I be the unit matrix.. Define a n×n rate generator matrix by $Q = (q_{i,j}: i \neq j)$, with the rate elements $q_{i,j}$ subject to the conditions: $q_{i,i} \leq 0$, $q_{i,j} \geq 0$ and $\sum_j q_{i,j} = 0$. The transition probability from state x_i^h to x_j^h in time t, for a homogeneous continuous time Markov chain, is obtained by

$$P(t) = \left(p_{i,j}(t)\right) = e^{tQ} = \sum_{k=0}^{\infty} \frac{(tQ)^k}{k!} =$$

$$I + \sum_{k=1}^{\infty} \frac{(tQ)^k}{k!} \qquad (2)$$

For the jump-diffusion in Eqn.(1), because of the freedom of the continuous parts from the hop parts, we can compose the comparing rate generator network Q as $Q = Q^c + Q^j$. Q^c and Q^j signify the generator framework that approximates the continuous part $\mu(.)dt + \sigma(.)dWt$ and jump part $\int_\zeta \eta(.)P(.)$ individually. Since in continuous time, a stochastic differential condition is completely portrayed by its mean and fluctuation, a very much characterized Q-rate network will coordinate the chain's first and second prompt minutes to those of the fundamental procedure. The approximation for Q^c matrix for univariate diffusion and the rate elements are given by

$$q_{i,i-1} = \frac{1}{2h^2}\sigma^2(x_i^h) + \frac{1}{h}\mu^-(x_i^h), q_{i,i=-} \frac{1}{h^2}\sigma^2(x_i^h) - \frac{1}{h}|\mu(x_i^h)|,$$

$$q_{i,i+1} = \frac{1}{2h^2}\sigma^2(x_i^h) + \frac{1}{h}\mu^+(x_i^h) \; q_{i,j=}0 \; \forall j \neq i, i-1, i+1 \qquad (3)$$

where the ± denotes the respective absolute value. However, when the grid spacing is too coarse, the proposed rate matrix formula exhibits an approximation error of $h|\mu(x_i^h)|$ in matching the second moment. Hence the corrected formula is presented to address this error

$$q_{i,i-1} = \frac{1}{2h^2}\sigma^2(x_i^h) - \frac{1}{2h}\mu(x_i^h), q_{i,i=-} \frac{1}{h^2}\sigma^2(x_i^h),$$

$$q_{i,i+1} = \frac{1}{2h^2}\sigma^2(x_i^h) + \frac{1}{2h}\mu(x_i^h) \qquad (4)$$

subject to the necessary condition of

$$h < \frac{\sigma^2(x_i^h)}{|\mu(x_i^h)|} \qquad (5)$$

Considering the empirical distribution of the data, the generalized Q^c formula is needed to accommodate a non-equidistant grid setting while satisfying the local consistency condition. For a n-state non-equidistant Markov chain with n −1 associated grid spacing of h, the Q^c is given by

$$q_{i,i-1} = \frac{1}{h_i}\mu^{+}\left(x_i^h\right) + \frac{\sigma^2\left(x_i^h\right) - \left(h_{i-1} \times \mu^{-}\left(x_i^h\right) + h_i \times \mu^{+}\left(x_i^h\right)\right)}{h_{i-1}(h_{i-1} + h_i)},$$

$$q_{i,i} = -q_{i,i-1} - q_{i,i+1}$$

$$q_{i,i+1} = \frac{1}{h_i}\mu^{+}\left(x_i^h\right) + \frac{\sigma^2\left(x_i^h\right) - (h_{i-1} \times \mu^{-}\left(x_i^h\right) + h_i \times \mu^{+}\left(x_i^h\right))}{h_i(h_{i-1} + h_i)}$$

$$q_{i,j=}0 \; \forall j \neq i, i-1, i+1 \qquad (6)$$

The following condition is needed to be satisfied for a well-defined probability matrix to be guaranteed.

$$sup\ (h) \leq \frac{\sigma^2(x_i^h)}{|\mu(x_i^h)|} \qquad (7)$$

Then the jump part is approximated, in which the matrix elements for the jump-diffusion generator matrix are given by

$$q_{i,j} = \lambda(x_i)\phi_{\zeta}(x_i; \zeta(x_i) \cap (x_j - x_i - h_{i-1}/2, x_j - x_i + h_{i+1}/2]), \qquad for\ j \neq 1, i, n,$$

$$q_{i,j} = \lambda(x_i)\phi_{\zeta}(x_i; \zeta(x_i) \cap (-\infty, x_1 + h_1/2]),$$

$$q_{i,j} = \lambda(x_i)\phi_{\zeta}(x_i; \zeta(x_i) \cap (x_n - h_{n-1}/2, \infty)),$$

$$q_{i,i} = -\textstyle\sum_{j \neq i} q_{j,i.} \qquad (8)$$

This setting can have a state-subordinate jump force and a jump distribution, which considers a conduct back translation, is hard to fuse with conventional numerical strategies. The execution of a model will be touchy to matrix separating and the lower and upper limits of the lattice. The benefits of the non-equidistant (non-uniform) lattice have been recorded in the exploration territory of finite difference methodology (FDM) and partial differential equations (PDE).

The improved model acquaints a heuristic approach with examining the matrix components for a n-states Markov chain, to such an extent that the Kolmogorov–Smirnov distance between the first distribution function G(X) and the Markov distribution function $\widetilde{\mathrm{G}}(\mathrm{X}^{\mathrm{h}})$ is limited. In such a case, we show that a non-equidistant Markov model can accomplish more elevated amount of exactness than an equidistant Markov display. The subsequent network x^{h} as for the Kolmogorov–Smirnov distance is given by

$$x_i^h = G^{-1}(\tfrac{2_i - 1}{2_n}) \qquad (9)$$

A repercussion of the Markov chain move likelihood network is the semi-explanatory log-likelihood function, which can be utilized to align the parameters of a jump-dispersion. The maximum likelihood estimator (MLE) of Improved Markov chain is characterized by

$$\hat{\vartheta}MCA := arg\ max\mathcal{L}(\vartheta), \qquad (10)$$

where $\mathcal{L}(\vartheta)$ is the log-likelihood. Given m discretely checked time arrangement data $\mathrm{x}_{\mathrm{t1}}, \mathrm{x}_{\mathrm{t2}}, \ldots \mathrm{x}_{\mathrm{tm}}$, the log-likelihood value produced by a period homogeneous transition probability matrix is given by

$$\mathcal{L}(\vartheta) = \textstyle\sum_{i=1}^{m-1} ln\left(e'(t_i)P(t_{i+1} - t_i)e(t_{i+1})\right) \qquad (11)$$

This principle of the improved Markov chain model can significantly enhance the visualization performance.

4 WebClickviz Visualization Methology

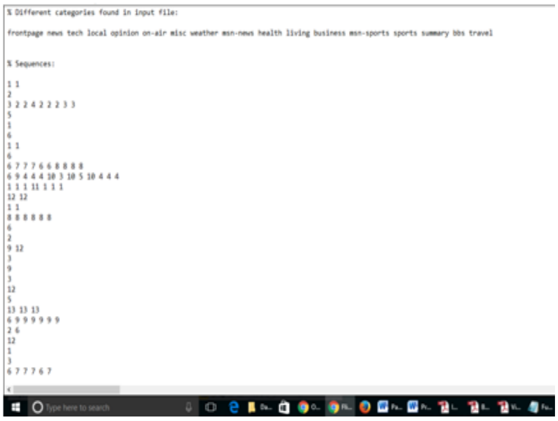
4.1 Tool and Data

The analysis of the website clickstream data has been carried out worldwide using many tools. Google analytics is one of the famous tools which have the basic functionality of clickstream data visualization. In this paper, Apache Flume is utilized to load, analyse the clickstream data and visualize it. Apache Flume is a distributed, reliable, and available service for productively gathering, aggregating, and moving a lot of streaming data into the Hadoop Distributed File System (HDFS). It has a straightforward and adaptable engineering in light of streaming data streams; and is robust and fault tolerant with tunable dependability instruments for failover and recuperation. Apache flume ingests the streaming data from multiple sources into the Hadoop storage and analysis and then insulates the buffer storage.

Flume utilizes channel-based transactions to ensure reliable message delivery. At the point when a message moves starting with one operator then onto the next, two transactions are begun, one on the specialist that conveys the occasion and the other on the specialist that gets the occasion. This guarantees ensured delivery semantics. The data used to load Apache Flume is the data that describes the page visits of users who visited msnbc.com [25]. Visits are recorded at the level of URL

DOI: 10.1051/matecconf/201712504025

category and are recorded in time order. The data comes from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com for an entire day. The categories are "frontpage", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "msn-news", and "msn-sports". A total of 989818 users have been recorded with average visits of 5.7 per user. Fig.1 shows the sample view from the input data collected from msnbc.com.

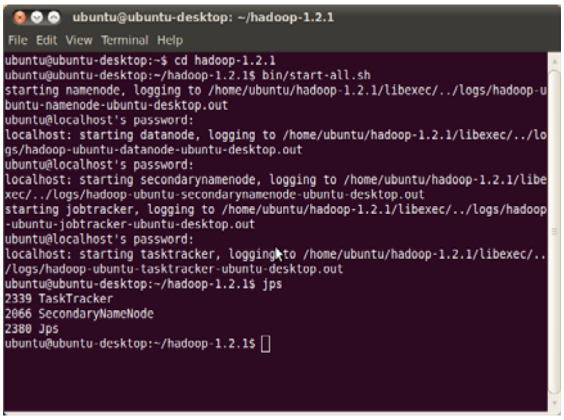


```
% Different categories found in input file:

frontpage news tech local opinion on-air misc weather msn-news health living business msn-sports sports summary bbs travel


% Sequences:

1 1
2
3 2 2 4 2 2 2 3 3
5
1
6
1 1
6
6 7 7 7 6 6 8 8 8 8
6 9 4 4 4 10 3 10 5 10 4 4 4
1 1 1 11 1 1 1
12 12
1 1
8 8 8 8 8 8
6
2
9 12
3
9
3
12
5
13 13 13
6 9 9 9 9 9 9
2 6
12
1
3
6 7 7 7 6 7
```

Fig.1. Input data

4.2 Loading Data

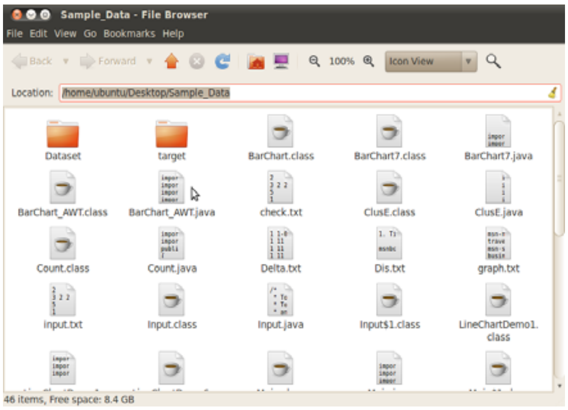
Each sequence in the dataset corresponds to page views of a user during that twenty-four hour period. Each event in the sequence corresponds to a user's request for a page. Requests are not recorded at the finest level of detail--that is, at the level of URL, but rather, they are recorded at the level of page category (as determined by a site administrator). Fig.2a & 2b shows how the data are loaded into Apache flume.



ubuntu@ubuntu-desktop: ~/hadoop-1.2.1
File Edit View Terminal Help

```
ubuntu@ubuntu-desktop:~$ cd hadoop-1.2.1
ubuntu@ubuntu-desktop:~/hadoop-1.2.1$ bin/start-all.sh
starting namenode, logging to /home/ubuntu/hadoop-1.2.1/libexec/../logs/hadoop-u
buntu-namenode-ubuntu-desktop.out
ubuntu@localhost's password:
localhost: starting datanode, logging to /home/ubuntu/hadoop-1.2.1/libexec/../lo
gs/hadoop-ubuntu-datanode-ubuntu-desktop.out
ubuntu@localhost's password:
localhost: starting secondarynamenode, logging to /home/ubuntu/hadoop-1.2.1/libe
xec/../logs/hadoop-ubuntu-secondarynamenode-ubuntu-desktop.out
starting jobtracker, logging to /home/ubuntu/hadoop-1.2.1/libexec/../logs/hadoop
-ubuntu-jobtracker-ubuntu-desktop.out
ubuntu@localhost's password:
localhost: starting tasktracker, logging to /home/ubuntu/hadoop-1.2.1/libexec/..
/logs/hadoop-ubuntu-tasktracker-ubuntu-desktop.out
ubuntu@ubuntu-desktop:~/hadoop-1.2.1$ jps
2339 TaskTracker
2066 SecondaryNameNode
2380 Jps
ubuntu@ubuntu-desktop:~/hadoop-1.2.1$ 
```

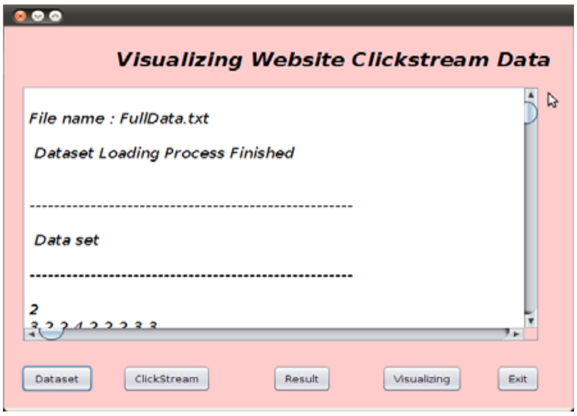
Fig.2a



Location: /home/ubuntu/Desktop/Sample_Data

46 items, Free space: 8.4 GB

Fig.2b



Visualizing Website Clickstream Data

File name : FullData.txt

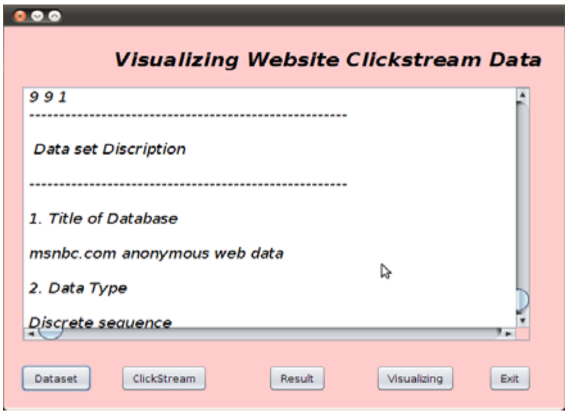
Dataset Loading Process Finished

Data set

2
3 2 2 4 2 2 2 3 3

Fig.3. View Loaded data

Fig.3 shows the loaded data while the Fig.4 shows the aggregated data. The data is loaded by means of loaddata() command which asks for the folder location of the data. When given, the data is loaded into the tool and can be viewed.



Visualizing Website Clickstream Data

9 9 1

Data set Discription

----

1. Title of Database

msnbc.com anonymous web data

2. Data Type

Discrete sequence

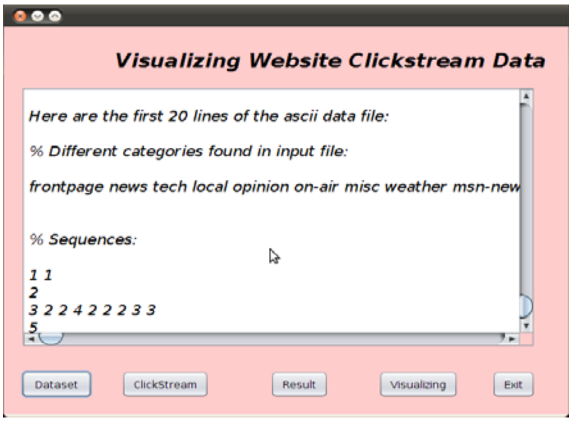
Fig.4. Aggregation of CRM data

Any page requests served via a caching mechanism were not recorded in the server logs and, hence, not present in the data. Fig.5 shows the visualized categories.

DOI: 10.1051/matecconf/201712504025

The clickstream process is executed once the data are loaded. This includes the aggregation and categorization view.

An implementation of Flume's RpcClient interface encapsulates the RPC mechanism supported by Flume. The user's application can simply call the Flume Client SDK's append(Event) or appendBatch(List<Event>) to send data and not worry about the underlying message exchange details.

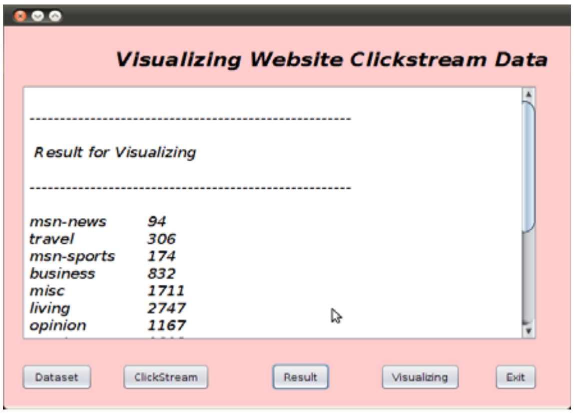


```
Here are the first 20 lines of the ascii data file:
% Different categories found in input file:
frontpage news tech local opinion on-air misc weather msn-new

% Sequences:
1 1
2
3 2 2 4 2 2 2 3 3
5
```

Fig.5. Categories of msmbc.com data

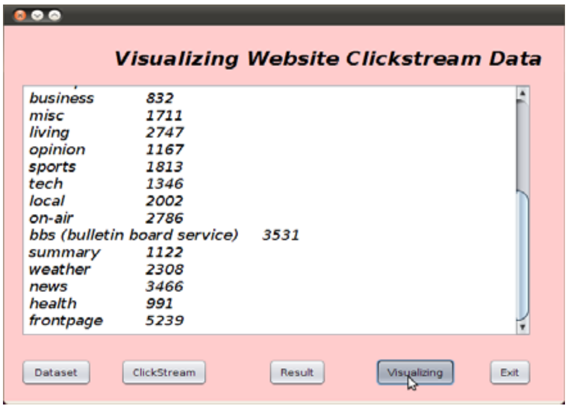
The user can provide the required Event arg by either directly implementing the Event interface, by using a convenience implementation such as the SimpleEvent class, or by using EventBuilder's overloaded withBody() static helper methods.

Data visualization helps to optimize the website and improve the business sales and values. Fig.6a & 6b shows the Visualized results. It can be seen that the categorization is accurately completed and the visits of the users are recorded as shown.



```
-----------------------------------------------------

 Result for Visualizing

-----------------------------------------------------

msn-news      94
travel        306
msn-sports    174
business      832
misc          1711
living        2747
opinion       1167
```

Fig.6a. Visualization Result 1



```
business      832
misc          1711
living        2747
opinion       1167
sports        1813
tech          1346
local         2002
on-air        2786
bbs (bulletin board service)    3531
summary       1122
weather       2308
news          3466
health        991
frontpage     5239
```

Fig.6b. Visualization Result 2

4.3 Geographic Representation

The visualization is complete only when the data are visualized either in graphical or association representation. Fig.7a shows the global representation of the clickstream data while Fig.7b shows the graphical representation in USA specifically.

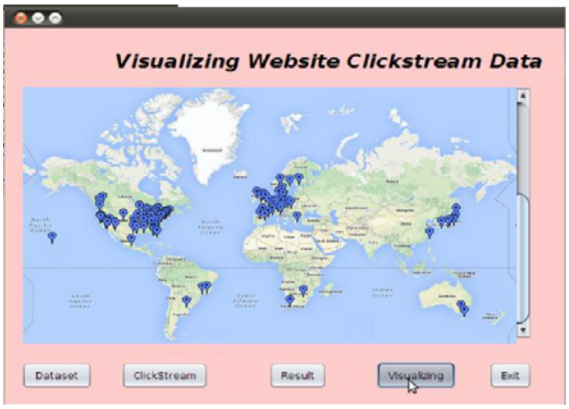


Fig.7a. Geographical Visualization 1

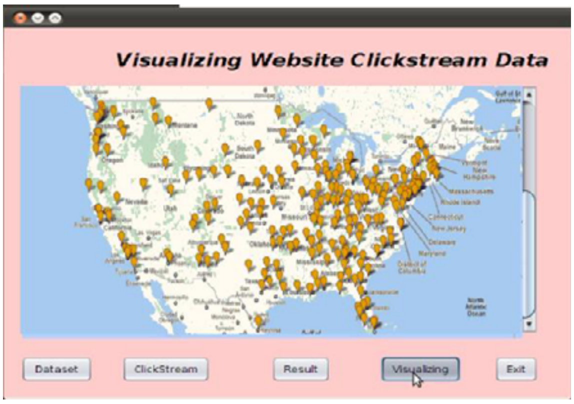


Fig.7b. Geographical Visualization 2

5 Visualization Performances

The performance of the WebClickviz is visualized in the charts given below. The charts are generated for the sample set of the msnbc.com website clickstream data.

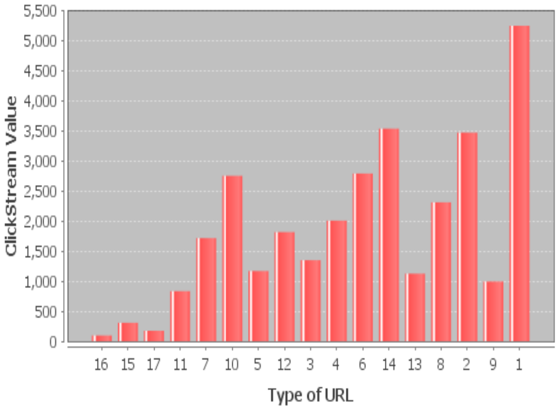


Fig.8. Clickstream value vs. type of URL

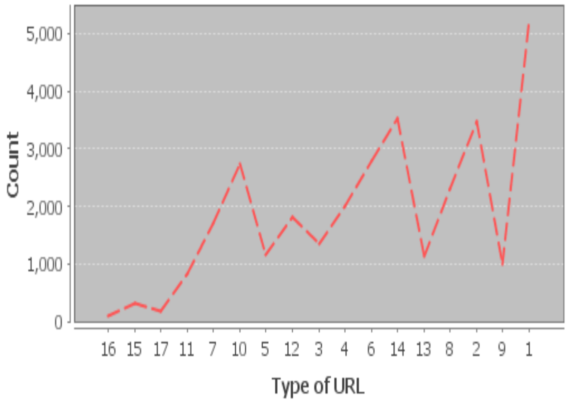


Fig.9. Count vs. type of URL

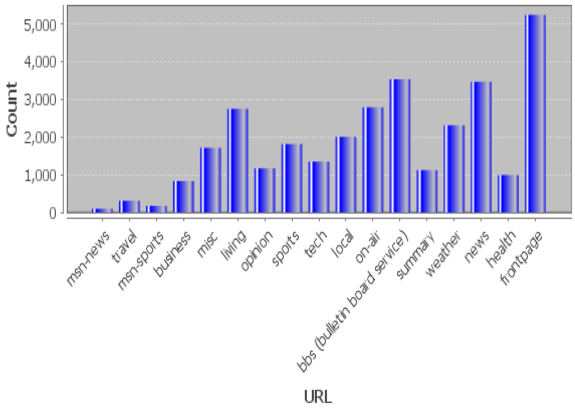


Fig.10. Count vs. URL specified

From the figures 8, 9 & 10, it can be seen that the charts clearly indicate the performance of the proposed model in visualization of clickstream data. As the visualization is highly improved in Hadoop Apache Flume, this work can serve as the initial step in making sense out of large web analytics data.

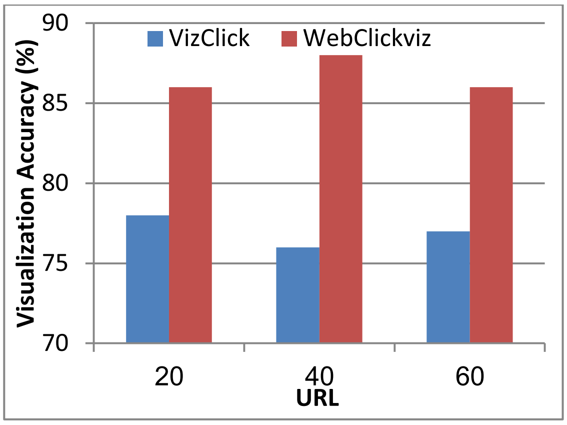


Fig.11. Visualization Accuracy

Fig.11 shows the comparison of visualization accuracy of existing VizClick versus the proposed WebClickViz. It is seen that the accuracy is higher in the proposed model at all counts of urls. The fundamental reason for existing was to examine measurable methodologies on clickstream information, as the accumulated arrangement of site visit demands executed by a specific client, and other client route components, can give understanding into their expectations, particularly as for purchase engagement and real-time purchase likelihood prediction. This can enhance the web analytics techniques by employing different strategies.

6 Conclusions

As stated in this article, these results are very encouraging as new methods of targeting customers could be derived from this solution. The proposed model consisting of the Improved Markov chain based visualization (WebClickviz) improves the web analytics by providing accurate visualization of the website clickstream data. This article suggested the method of interactive visualization in order to utilize these results in the analysis of data for different applications. In the field of clickstream data research is still in its earliest stages, much research still should be finished. With the rebellion of new and speedier innovation, the idea of big data is exceptionally hot right now, particularly on the grounds that companies can, more than ever, make an interpretation of customer data into higher revenue. In the future researches, it will be analysed how to utilize these results for different applications. Likewise the use of new learning algorithms to fit clickstream data, namely, by introducing other models such as neural networks, support vector machines [26], genetic algorithms, etc will be investigated.

REFERENCES

1. Farney, T. A. Click analytics: Visualizing website use data. Information Technology and Libraries, **30**(3), 141, (2011)
2. Kimball, R., & Merz, R. The data webhouse toolkit. Wiley (2000).
3. Phippen, A., Sheppard, L., & Furnell, S. (2004). A practical evaluation of Web analytics. Internet Research, **14**(4), 284-293.
4. Gonçalves, B., & Ramasco, J. J. (2008). Human dynamics revealed through Web analytics. Physical Review E, 78(2), 026123.
5. Plaza, B. Monitoring web traffic source effectiveness with Google Analytics: An experiment with time series. In Aslib Proceedings (Vol. **61**, No. 5, pp. 474-482), (2009). Emerald Group Publishing Limited.
6. Kohavi, R., Rothleder, N. J., & Simoudis, E. Emerging trends in business analytics. Communications of the ACM, **45**(8), 45-48, (2002).
7. Hasan, L., Morris, A., & Probets, S. Using Google Analytics to evaluate the usability of e-commerce sites. Human centered design, 697-706, (2009).
8. Kohavi, R., Mason, L., Parekh, R., & Zheng, Z. Lessons and challenges from mining retail e-commerce data. Machine Learning, **57**(1), 83-113, (2004).
9. White, T. Hadoop: The definitive guide. "O'Reilly Media, Inc.", (2012).
10. Flanagan, D. JavaScript: the definitive guide. "O'Reilly Media, Inc.", (2006).
11. Bucklin, R. E., & Sismeiro, C. Click here for Internet insight: Advances in clickstream data analysis in marketing. Journal of Interactive Marketing, **23**(1), 35-48, (2009).
12. Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. Modeling online browsing and path analysis using clickstream data. Marketing science, **23**(4), 579-595, (2004).
13. Moe, W. W., & Fader, P. S. Capturing evolving visit behavior in clickstream data. Journal of Interactive Marketing, **18**(1), 5-19, (2004).
14. Van den Poel, D., & Buckinx, W. Predicting online-purchasing behaviour. European journal of operational research, **166**(2), 557-575, (2005).
15. Danaher, P. J., Mullarkey, G. W., & Essegaier, S. Factors affecting web site visit duration: a cross-domain analysis. Journal of Marketing Research, **43**(2), 182-194, (2006).
16. Kateja, R., Rohith, A., Kumar, P., & Sinha, R. VizClick visualizing clickstream data. In Information Visualization Theory and Applications (IVAPP), 2014 International Conference on (pp. 247-255). IEEE, (2014).
17. De Oliveira, M. F., & Levkowitz, H. From visual data exploration to visual data mining: a survey. IEEE Transactions on Visualization and Computer Graphics, **9**(3), 378-394, (2003).
18. Moe, W. W. An empirical two-stage choice model with varying decision rules applied to internet clickstream data. Journal of Marketing Research, **43**(4), 680-692, (2006).
19. De Bock, K., & Van den Poel, D. Predicting website audience demographics forweb advertising targeting using multi-website clickstream data. Fundamenta Informaticae, **98**(1), 49-70, (2010).
20. Chen, L., & Su, Q. Discovering user's interest at E-commerce site using clickstream data. In Service systems and service management (ICSSSM), 2013 10th international conference on (pp. 124-129). IEEE, (2013).
21. Schellong, D., Kemper, J., & Brettel, M. Clickstream data as a source to uncover con-sumer shopping types in a large-scale online setting, (2016).
22. Shi, C., Fu, S., Chen, Q., & Qu, H. VisMOOC: Visualizing video clickstream data from massive open online courses. In Visualization Symposium (PacificVis), 2015 IEEE Pacific (pp. 159-166). IEEE, (2015).
23. Brinton, C. G., & Chiang, M. Mooc performance prediction via clickstream data and social learning networks. In Computer Communications (INFOCOM), 2015 IEEE Conference on (pp. 2299-2307). IEEE, (2015).
24. Gilks, W. R., Richardson, S., & Spiegelhalter, D. (Eds.), Markov chain Monte Carlo in practice. CRC press, (1995).
25. http://www.msnbc.com
26. Steinwart, I., & Christmann, A. Support vector machines. Springer Science & Business Media, (2008).