A Model of Website Usage Visualization Estimated on Clickstream Data with Apache Flume Using Improved Markov Chain Approximation

AMJAD JUMAAH FRHAN
Department of Telecommunication and Information Technology
University Politehnica of Bucharest
ROMANIA.

Abstract: - Visualization of the website clickstream data has been a pivotal process as it aids in defining the user preferences. It includes the processes of gathering, investigating and reporting about the web pages that are being viewed by the users. Clickstream visualization is primarily employed by organizations which focuses on gaining the user preferences and improve their products or services towards achieving maximum satisfaction of users. Most existing visualization tools come up short in helping the organizations achieve this goal. Markov chain model is the commonly utilized method for developing data visualization tools. However the issues such as occlusion and inability to provide clear data visualization display makes the tools volatile. This paper aims at developing a visualization tool named as WebClickviz by resolving the above mentioned issues by improving the Markov chain modelling. A heuristic method of Kolmogorov– Smirnov distance and maximum likelihood estimator is introduced for improving the clear display of visualization. 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. Through the experiments conducted on evaluation dataset, it can be shown that the proposed model outperforms the existing models with higher visualization accuracy.

Key-Words: - Clickstream data, Data Visualization, Hadoop, WebClickviz, Apache Flume, Markov chain, Kolmogorov– Smirnov distance, maximum likelihood estimator, heuristic approximation.

1 Introduction

Website clickstream data visualization is a step by step method by which the user propagation is tracked from the server log files and clickstream files. 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.

Analyzing 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 Modeling

3.1 Clickstream Data Collections

Customarily, clickstream data could be gathered by keeping point by point Web server logs, maybe increased by a cookie. A "cookie" is a constrained data question transported in factor length fields inside headers of Hypertext Transfer Protocol ("HTTP") ask for messages (utilized while asking for items) and reaction messages (utilized while giving the asked for objects). Cookies are regularly put away on the customer, either for the length of a session—e.g. all through a client's electronic shopping associations with an on-line trader—or for all time. A cookie stores certain data that the server application needs to recollect about a specific customer. This could incorporate customer distinguishing proof, session parameters, client inclinations, session state data, or nearly whatever else an application essayist can consider to incorporate inside the constraints of the cookie detail.

Gathering and dissecting clickstream data utilizing Web server logs and discretionary cookie data is called "logfile examination". Numerous merchants are utilizing this system, including WebTrends (which advertises an item called "WebTrends Log Analyzer") and Net.Genesis (which showcases an item called "net.Analysis"). In any case, as Web locales increment in size and many-sided quality, regularly spreading over the globe with several servers, gathering and breaking down clickstream data utilizing logfiles turns out to be progressively troublesome. A couple of organizations are beginning to utilize a more customer situated strategy by inserting references to "clear GIFs" (regularly called "Web Bugs") in Web pages. The expression "clear GIF" alludes to utilization of a little realistic picture, regularly just 1 pixel by 1 pixel in estimate when shown, which is ordinarily certain or straightforward. Contingent upon how the reasonable GIF reference is encoded in the Web page definition, and relying upon the client's activities when seeing the Web page, a demand message will be issued from the client's gadget (i.e. the customer) to recover the document containing the reasonable GIF. Due to its little size and straightforwardness, a reasonable GIF that is rendered on the client's show is moderately unpretentious as far as the client gadget's stockpiling limit and the client's view of the Web page. Frequently, the reasonable GIF (alluded to hereinafter as a Web Bug) utilizes executable code written in Javascript to give an account of the substance of the individual Web page (i.e. by communicating something specific with data about the specific page inside which the reasonable GIF was asked). The HTTP ask for header which demands conveyance of the reasonable GIF additionally supplies certain sorts of data about the customer, for example, the client operator (i.e. program) being used at the time, what sorts of encoding this client operator backings, et cetera (as is known in the craftsmanship). When utilizing this Web Bug system, the client's program sends clickstream data straightforwardly to a website investigation application.

These conventional methods for getting clickstream data are destined to disappointment later on for various reasons, including:

- the failure to recognize a particular client by the client's Internet Protocol ("IP") address and port number (which is because of various variables, for example, utilization of Network Address Translation or "NAT", virtual space facilitating, et cetera, which are examined in more detail in the related applications), along these lines anticipating relationship of the gathered data with a specific client in light of these parameters;

- objections to utilization of Web Bugs by different security gatherings, and development to drive bolster for quit arrangements (like existing cookie quit arrangements) which, if actualized, may debilitate the capacity for locales to gather essential operational data;

- the expanding utilization of dynamic Web pages whose substance can't be derived by the individual address, or Uniform Resource Locator ("URL"), from the HTTP ask for if Javascript is not empowered, in this manner averting full comprehension of what the end client was really seeing when the clickstream data was gathered;

- the utilization of network caches to serve static substance (wherein demands for Web pages are here and there blocked and served without making a full network round-trip, to such an extent that the solicitations don't

bring about log messages that achieve the website examination area), in this way forestalling accumulation of a total picture of what was conveyed to the end client;

- the inclination of owning Internet Service Providers ("ISPs") to design their store servers to overlook reserving rules in the HTTP ask for headers originating from customer gadgets, in this manner keeping the storing rules from rendering already stored data stale, and serving data from reserve in any case so as to streamline purpose of-presence ("POP") transfer speed prerequisites—which likewise avoids gathering of a total picture of what was conveyed to the end user;

- privacy concerns and programmer action have made various users arrange their program as well as individual firewall intermediary to cripple many Web highlights (counting such things as treats, Javascript, VBscript, ActiveX, Java™ program bolster, and diligent storage);

- the reality that a given Web page comprises of an assortment of HTTP asks for and comparing answers that might be served by different servers as well as reserves, making it hard to collect and break down clickstream data in view of receipt of a demand at a specific system area; and

- pervasive gadgets and individual advanced collaborator ("PDA") programs may channel the sorts of substance to acknowledge, including GIFs or pictures (e.g, because of the powerlessness of the gadgets to store a lot of data), and henceforth the reasonable GIF asset utilized for a Web Bug would not be recovered—in which case the Web Bug strategy would be totally ineffectual for clickstream data gathering.

In perspective of these weaknesses of current methodologies, what is required is an enhanced system for gathering clickstream data. The present innovation characterizes a clickstream data gathering method which empowers accumulation of granular execution, hit, and user navigation knowledge. Data is gathered with respect to an end user's navigation way among a progression of Web archives, and a stock of the articles contained in the rendered records is likewise gathered.

The lessons of the main related development empower a server liking to be characterized for a specific customer (without requiring the customer's IP deliver to be exceptional), along these lines bypassing load adjusting for related messages inside that partiality, and re-establish Web applications' capacity to depend on the nearness of cookies (with no presumption on the capacity of a specific customer to help cookie usefulness). Rather, any cookies embedded into an outbound HTTP header are separated preceding conveyance of the reaction message to the customer, and are put away in server-side stockpiling alluded to as a "cookie jostle". In the wake of separating the cookies, if the outbound reaction incorporates a mark-up dialect archive, at that point any implanted URLs inside that mark-up dialect report alluding to or in respect to the session's server are revamped in a discernible way to incorporate what is characterized as a "sticky directing token". A sticky directing token is a customer special URL that shows entomb alia, where in the system the cookie shake for this specific customer session dwells. The revised URL organize is straightforward to the customer. In the event that one of the reworked URLs is thusly referenced (e.g. by the user tapping on a connection to that URL, or rendering a page which incorporates the URL), at that point the sticky directing token is naturally returned on the request message which is sent to recover the substance of that URL. The principal creation additionally instructs assessing the HTTP header of an inbound request message after touching base at the server side, hunting down a sticky steering token. If one is found (and if its substance are legitimate and not stale), at that point it is expelled from the header. (Something else, another sticky steering token is made.) A key is extricated from the sticky directing token (or other comparable stockpiling), where this key recognizes the cookie jostle in which this present session's cookies are put away. The cookies are then recovered, embedded into the HTTP request header, and sent to the application for use inside the server-side condition as important to benefit the approaching request.

The second related development expands on this establishment, and includes extra fields into the sticky steering token to make a "URL token" (i.e. a markup string for utilize when changing URLs). These extra fields give nature of administration

parameters that are to be utilized while directing related messages for a specific exchange.
The second related creation along these lines empowers exchange particular nature of administration data, (for example, relative transmission need and accessible last-mile transfer speed) to be resolved and conveyed through the system throughout the exchange, and enables a steady treatment to be utilized for conveyance of all items transmitted inside that exchange (counting objects that might be referenced from a rendered record and additionally recovered when the end user taps on a URL referencing a protest). The second related innovation additionally characterizes a strategy whereby the cookie jolt rationale might be put on at least one of a Web application server, a server-webpage edge server or surrogate, and an in-organize branch side edge server or storing intermediary.
At the point when a cache server actualizing the present innovation gets a URL token containing a non-zero counter (demonstrating that clickstream data is being gathered) and can serve the separate question from its cache, at that point it reports the got URL token with suitable time stamps to the site analyzer handle. This site analyzer can utilize the URL token to find the cookie esteems announced before by the cookie shake preparing rationale at the hub containing the separate cookie jolt. Upstream servers (that is, those servers back of the hub containing the cookie jolt) may keep on operating as they did before the present development (with the exception of that the application or application server is expanded to produce clickstream cookies and the site analyzer must have the capacity to get to the ordinary application server logs to finish its perspective of the separately checked page streams), in that they make log sections utilized by the site analyzer prepare. Since these log passages will now contain the particular cookies which were reestablished from the downstream cookie jostle handling, the site analyzer prepare has all the data it needs to put these log sections in the general stream alongside the cache-created log passages.
On the off chance that the request touched base at a server or cache that does not bolster the present innovation (for instance, the neighborhood program cache), the protest's URL would seem, by all accounts, to be a one of a kind page reference (that is, the primary use in a page would not relate to any cached question); along these lines, the cache would forward the request to an upstream cache or application server, in all likelihood bringing about it being served by a cache or edge server which implements the present creation or by the application server itself. This implies the site analyzer process would see the separate reference. The present development additionally takes into account a pecking order of edge servers, for instance one at the edge of a server group and one at a branch area close to the customers. By testing URLs to check whether they were reworked by an upstream hub supporting the present innovation, clickstream data accumulation data can be passed between servers without requiring this help on all servers.
An essential objective of the present creation is to permit particular gathering of clickstream data for both dynamic and static pages, including same-site objects referenced by such pages (scripts, text styles, outlines, shapes, pictures, and so forth.) in a way that keeps away from the issues related with the conventional
clickstream data accumulation approach. As will be clarified in more detail thus, this is refined by embeddings clickstream data accumulation data inside URL tokens in all URL references to a specific server site in a way that is not reliant on a program, intermediary/passage, or unavoidable gadget's capacity to help pictures, cookies, Javascript, Java, VBscript, ActiveX, relentless capacity, or even reserve control headers. The correlator esteem transmitted in a clickstream cookie is ideally created by the application as a special identifier that can be utilized by site examination rationale to reassemble log sections for a specific exchange.

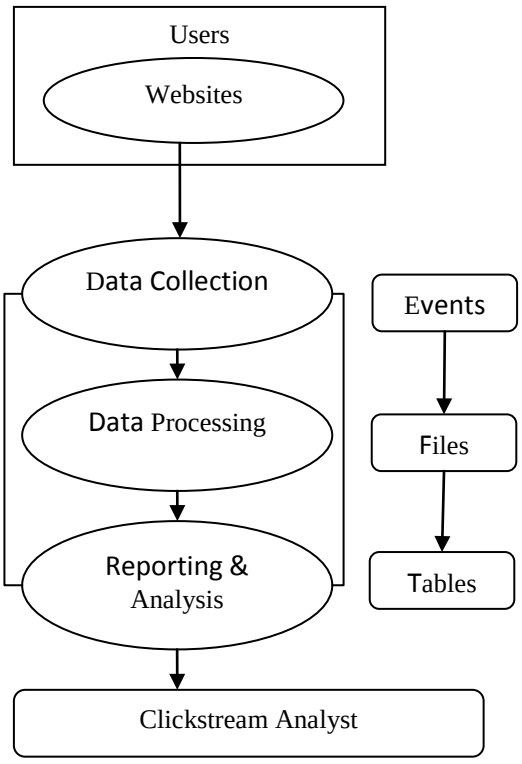

Fig.1. Clickstream analysis

As expressed over, a zero or invalid correlator esteem ideally shows that data gathering is crippled, and a non-zero or non-invalid correlator esteem demonstrates that data accumulation is empowered; or, then again, a different banner might be added to the cookie for this reason.

In a further discretionary part of the present development, a URL-based govern assessment strategy might be upheld which permits accumulation of clickstream data for a specific exchange (or for singular messages of an exchange) to be empowered in view of managerial arrangement data. Thus the website clickstream data can be collected more effectively.

3.2. Markov chain Model

In Web Usage Mining [24], sequential patterns techniques [25]) attempt to discover standard events of a same arrangement of components (designs) for each Web session under supervision, for example, the nearness of a thing set taken after by other thing in an objective gathering of page sees. Utilizing these systems it is conceivable to foresee future visit designs that could be utilized to dispatch arranged commercials, practical notices, or basically to advise Web users about new things that they could be intrigued. Fundamentally, we mean to locate some applicable data about Web users' conduct, so as to give more alluring website. A standard navigation handle over a normal Web website starts by choosing an underlying hyperlink and getting the page see inside a traditional Web program. This initially page see characterizes the underlying condition of another Web session. Next, the navigation procedure keeps choosing another hyperlink that will characterize another phase for the present Web session. This progressive choice of hyperlinks is constantly recorded, click after snap, in particular clickstream documents. Last mentioned, and with appropriated procedures, the investigations of these log documents will uncover all the navigation ways (arrangements of pages sees) for a Web website that its users have done their sessions [24]. As alluded some time recently, a standout amongst the most applicable ways to deal with delineate navigation ways is to utilize Markov chains.

With a Markov affix it is conceivable to demonstrate which is the following page will be requested by a user in view of its present area and on its past navigation sessions. Additionally, when created in view of Web exchanges, the chains give us the navigation ways taken after by a user amid a particular timeframe, giving us the likelihood to recognize the most incessant page grouping that the user will presumably follow in a future Web session. Speaking to Web navigation ways with Markov chains we can get profitable data about the webpage in examination and its future users navigation tendencies– for example, through the investigation of a Markov chain, it is conceivable to anticipate what pages see succession an uncommon sort of user will do in a daily paper, and get ready as needs be some ad spots that go towards the portrayal of the navigation profile. Markov fastens are particularly valuable to manufacture expectation models [25], taking into account the foundation of future user conduct while users are cooperating with the destinations. This is finished with the examination of beforehand users conduct with comparative interests. We can utilize Markov models to locate the more successive trails (navigation ways) trailed by users in their navigation forms, which intends to locate the most incessant groupings of pages that the users visit amid their navigation sessions.

Markov chains could be viewed as a customary diagram in which hubs (stages) speak to went by pages and edges (moves) the likelihood identified with some Web user going from a page to another. For the most part, edges‟ probabilities are computed checking de number of users that go starting with one page then onto the next, contemplating the quantity of visits for every inception page. However this model does provide only nominal clarity in clickstream data visualization.

3.3 Proposed Model

The shortcomings of standard Markov chain [26] 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

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

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 $\mu(X_t;\theta)$. 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

$$P(t)=\left(p_{i,j}(t)\right)=e^{tQ}=\sum_{k=0}^{\infty}\frac{(tQ)^k}{k!}=I+\sum_{k=1}^{\infty}\frac{(tQ)^k}{k!} \quad (2)$$

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

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

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

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

subject to the necessary condition of

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

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

$$q_{i,i-1}=\frac{1}{h_i}\mu^+(x_i^h)+\frac{\sigma^2(x_i^h)-\left(h_{i-1}\times\mu^-(x_i^h)+h_i\times\mu^+(x_i^h)\right)}{h_{i-1}(h_{i-1}+h_i)},$$
$$q_{i,i}=-q_{i,i-1}-q_{i,i+1}$$
$$q_{i,i+1}=\frac{1}{h_i}\mu^+(x_i^h)+\frac{\sigma^2(x_i^h)-(h_{i-1}\times\mu^-(x_i^h)+h_i\times\mu^+(x_i^h))}{h_i(h_{i-1}+h_i)}$$
$$q_{i,j=}0\ \forall j\neq i,i-1,i+1 \quad (6)$$

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

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

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

$$q_{i,j}=\lambda(x_i)\phi_\zeta(x_i;\zeta(x_i)\cap(x_j-x_i-h_{i-1}/2,x_j-x_i+h_{i+1}/2]),\quad \text{for } j\neq 1,i,n,$$
$$q_{i,j}=\lambda(x_i)\phi_\infty(x_i;\zeta(x_i)\cap(-\infty,x_1+h_1/2]),$$
$$q_{i,j}=\lambda(x_i)\phi_\zeta(x_i;\zeta(x_i)\cap(x_n-h_{n-1}/2,\infty)),$$
$$q_{i,i}=-\sum_{j\neq i}q_{j,i}. \quad (8)$$

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 $\tilde{G}(X^h)$ 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

$$x_i^h=G^{-1}\left(\frac{2_i-1}{2_n}\right) \quad (9)$$

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

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

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

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

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 [27].

Visits are recorded at the level of URL 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.2 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
```

Fig.2. 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.3a & 3b shows how the data are loaded into Apache flume.

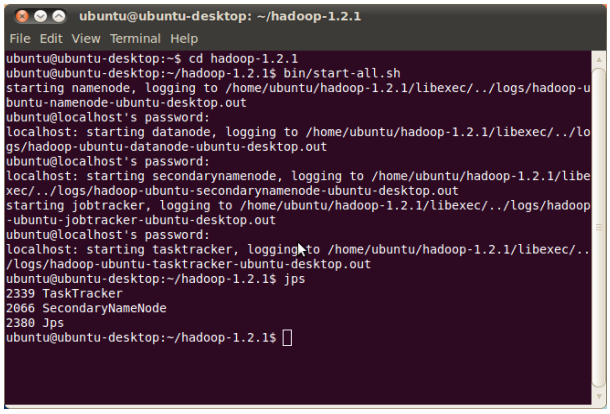

Fig.3a

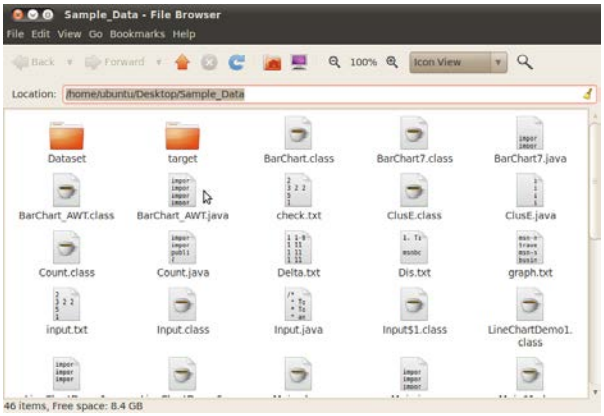

Fig.3b

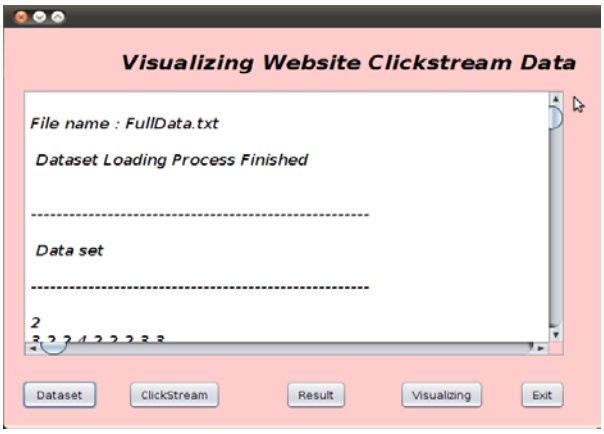

Fig.4. View Loaded data

Fig.4 shows the loaded data while the Fig.5 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.

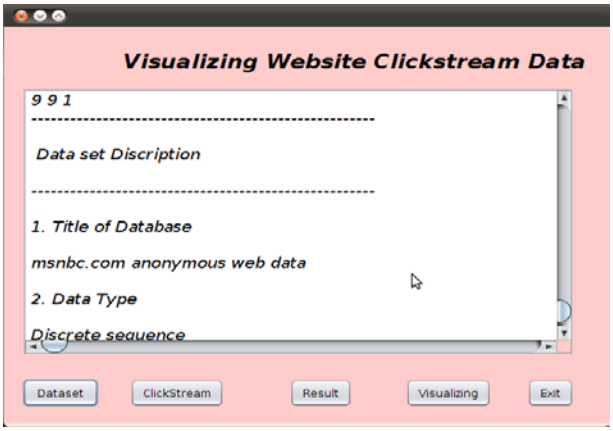

Fig.5. Aggregation of CRM data

Any page requests served via a caching mechanism were not recorded in the server logs and, hence, not present in the data. Fig.6 shows the visualized categories. The clickstream process is executed once the da8ta 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.

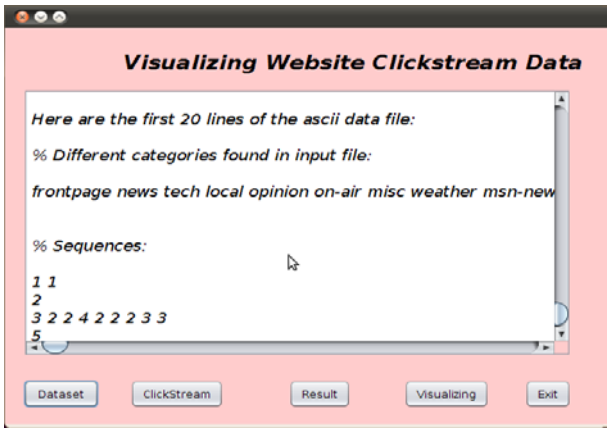

Fig.6. Categories of msmbc.com data

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.

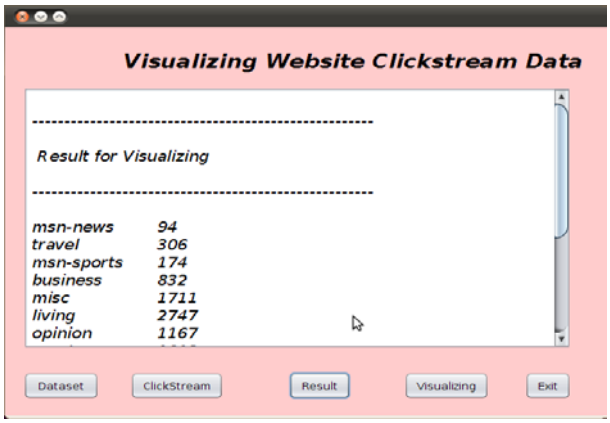

Fig.7a. Visualization Result 1

Data visualization helps to optimize the website and improve the business sales and values. Fig.7a & 7b shows the Visualized results. It can be seen that the categorization is accurately completed and the visits of the users are recorded as shown.

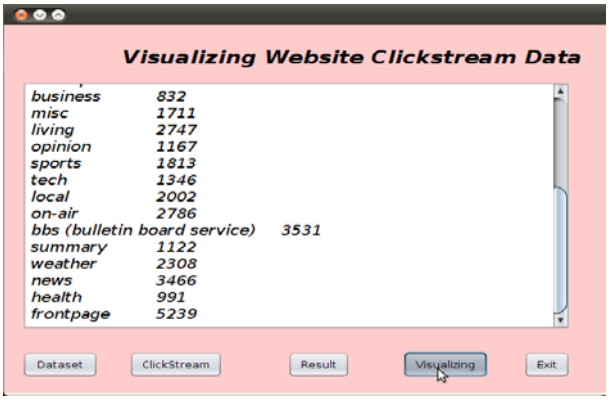

Fig.7b. Visualization Result 2

4.3 Geographic Representation

The visualization is complete only when the data are visualized either in graphical or association representation.

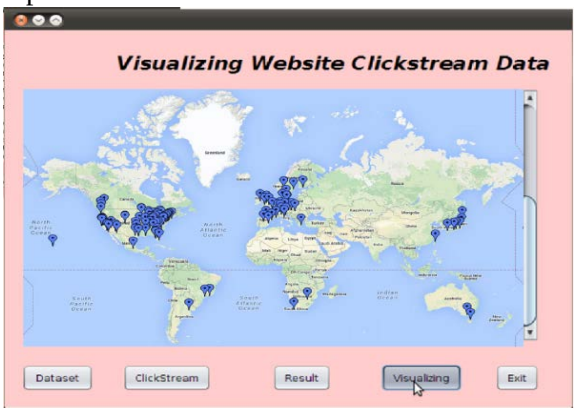

Fig.8a. Geographical Visualization 1

Fig.8a shows the global representation of the clickstream data while Fig.8b shows the graphical representation in USA specifically.

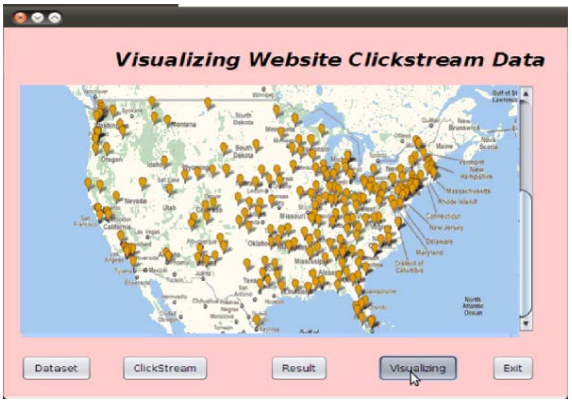

Fig.8b. Geographical Visualization 2

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.

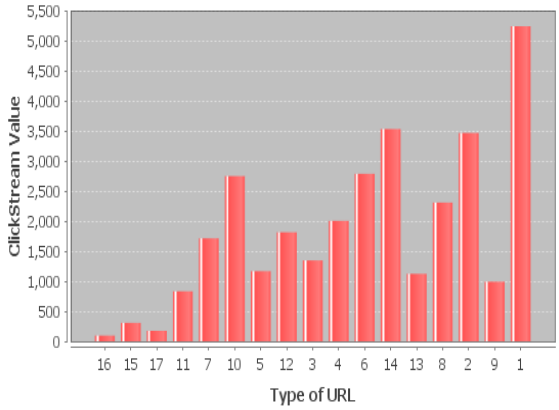

Fig.9. Clickstream value vs. type of URL

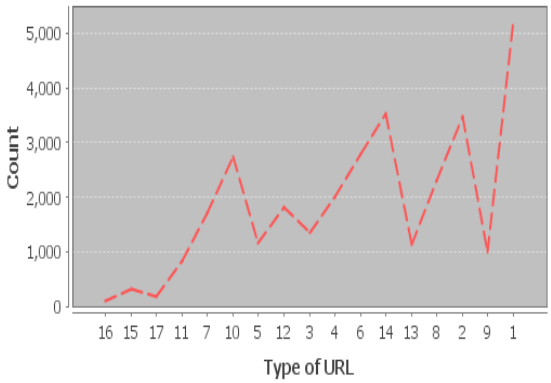

Fig.10. Count vs. type of URL

From the figures 9, 10 & 11, 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.

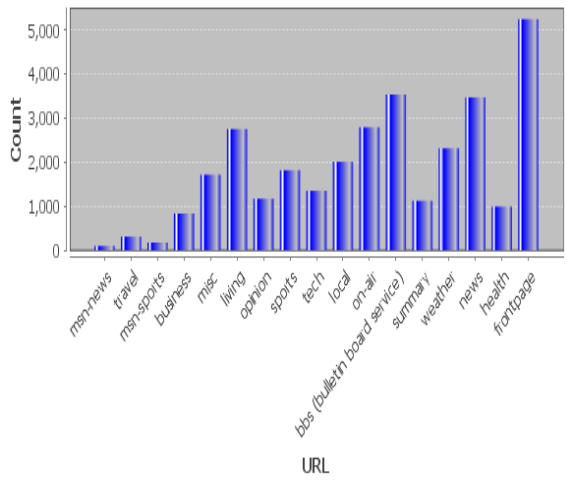

Fig.11. Count vs. URL specified

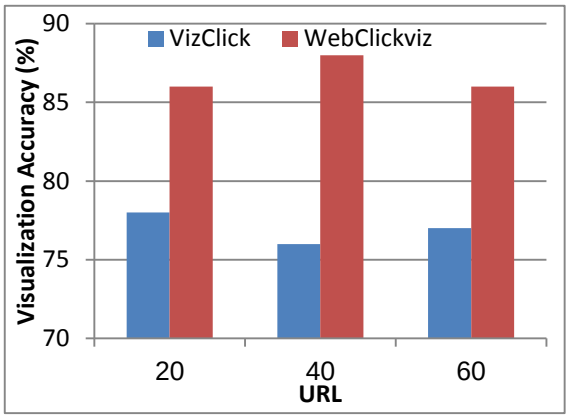

Fig.12. Visualization Accuracy

Fig.12 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 Conclusion

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 analyzed 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 [28], 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. A practical evaluation of Web analytics. *Internet Research*, *14*(4), 284-293, 2004.

[4] Gonçalves, B., & Ramasco, Human dynamics revealed through Web analytics. *Physical Review E*, *78*(2), 026123, 2008.

[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). Emerald Group Publishing Limited, 2009.

[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. Modelling 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 consumer 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.
[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.
[24] Srivastava, J., Cooley, R., Deshpande, M. and Tan, P.N. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. ACM SIGKDD Explorations Newsletter, 1(2), 12-23. 2000.
[25] Esmaeili, M., Gabor, F., Finding Sequential Patterns from Large Sequence Data. IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, 2010.
[26] Gilks, W. R., Richardson, S., & Spiegelhalter, D. (Eds.). *Markov chain Monte Carlo in practice*. CRC press, 1995.
[27] http://www.msnbc.com
[28] Steinwart, I., & Christmann, A. Support vector machines. Springer Science & Business Media, 2008.