A Brief Review on Process Mining Procedure Using the Click Stream Analysis for the Shopping behavior of Customer in E-Commerce Websites

Dr.DileepKumar Padidem
Prof.Dept of CSE, Chadalawada Ramanamma Engineering College,A.P, India
T.Lakshmi Prasanna
Asst.Professor,M.C.A Dept, Chadalawada Ramanamma Engineering College, A .P,India

ABSTRACT

The effective nature of web online data creates huge massive volumes of required knowledge data in structured & fully pattern and unstructured data. Knowledge discovering data from these knowledge data are said to be web mining and this web mining process can be focused on three different categories of the information available. They are Web Content Mining, Web Structure Mining, and Web Usage Mining. Web mining is used in distinct areas & domains and researches are carried out in concentration in all the three categories with different methodologies. In this paper, research their work on web mining application in the fields of online transaction E-commerce. The Effective nature of web and its growing importance as an economic platform is in need of new methods and tools to improve business efficiency in this ecommerce website. There are many research analysis outputs has been produced using web analytics study which observes customers behavior through online click stream behavior and market basket analysis which will not provide critical path of site visitors behavior and abstracted view of underlying customer processes. We given idea of applying data cleaning Business Process Methodologies (BPM) to event logs of online business websites to study the challenges and potential benefits of such an approach.

Key words: web mining, click stream, e-record (electronic record),process mining, Alpha-algorithm, Heuristic Miner Algorithm

1.Introduction

Web mining uses the data mining techniques to extract Knowledge information from Web documents and services. The Procedure of web access pattern is extracted and analyzed using knowledge discovery techniques to understand the patterns. This usage mining can be done from the click stream data of the website by the users with the help of weblogs. This tasks are observed through click stream analysis and a business process model may be discovered in web structure mining. This paper suggests a plan and information about how to considered the user behavior in E commerce sites as general process and to discover a process model which enhances business intelligence.

Two distinct procedures were taken in initially defining web mining. First was a "process-centric view," states that web mining as a sequence of tasks (Etzioni 1996)[4]. Second was a "data-centric view," states web mining in terms of the types of web data that was being used in the mining process (Cooley, Srivastava, and Mobasher 1997)[5]. This paper considers its study on web mining in terms of process centric view which defined web mining as a sequence of tasks. This paper gives plan view to consider the Click stream data in ecommerce websites as sequence of tasks and setting different optimum business models.

Click stream data are the e- record of a user's behavior on the web sites.This data trace the path a visitor takes while navigating the Web and this path reflects Choices, Even though in more number, made by the user both within and across websites. For example, the data set of a click stream might include a record of every website and every page Click stream data are defined as the electronic record of a user's activity on the Internet.

Thus, the data trace the path a shopper takes while navigating the Web. This path reflects Choices, often very large in number, made by the visitor both within and across websites. For example, a click stream dataset might include a record of every website and every page visited, the time user spent on each site and the order the sites and pages were visited. Consider unit of observation in clicks stream data is the page

visit– the recording of a user's visit to a given website page. Technically, the assembly of a "page view" from the user's perspective can involve numerous "hits" to the Web server. These reflect the downloading of various page elements before they are assembled in the User's Internet browser window. Click stream data is automatically aggregated from hits to page views but in some cases (e.g., raw server log files), the analyst may need to perform this step.

Raw click stream data can be captured by server log files maintained by a website can record all the requests and information transferred between the server and the user's computer system. The data are collected from a one website and they are known as "site-centric." Site-centric click streams can provide very detailed records of customer's behavior that is about their navigation and interaction with a given site.

Click-stream data provides the change for an in depth checkup on the choice creating method itself, and data extracted from it may be used for maximum, influencing the method, etc. (Ong and Keong 2003) Underhill (2000) has once and for all well-tried the worth of model data in understanding user's behavior in ancient sites. analysis has to be allotted in (1)getting method models from usage information(2)getting however totally different components of the method model impact varied internet metrics of interest and(3)however the method models, amendment in response to varied changes that area unit created i.e, ever changing stimuli to the customer.

2. Literature survey

Process mining complements existing approach Business Process Management (BPM). BPM combines knowledge from management sciences and applies this to operational business processes [8, 10]. BPM can be seen as an extension of Workflow Management (WFM) and it focuses on the automation of business processes. Process mining is close to BPM life-cycle.

W. M. P. vander Aalst [6] stated that there is currently a missing link between business processes and the real processes with information systems. Process mining has arisen as a new scientific discipline to provide a link between process models and event data [6]. Simeonova [7] defined process mining as facilitate to search out, screen and extend real process by concentrating learning from event logs. Knowledge is getting from different varieties of systems and examined to spot deviations from normal processes and see where the bottlenecks are. Process mining relis on actual based data and starts with an analysis of data, followed by the creation of a process model.

The procedure of data mining and knowledge data discovery will applied efficiently on web sites. This particular procedure of data mining on e-commerce web pages called Web Mining and it has consider more concerating area of researches. A noval research area was emerged from Web Mining for giving the solutions to its specific requirements. Some researchers scholars has worked on extracting the contents of a web site in web content mining, mean while others has decided to study the structure of a web site in web structure mining or analyze the usage of a web site (web usage mining).

The information expected to achieve such errands is gotten regularly from an Internet server log document − all web based business applications are Electronic. Snap stream records are produced keeping in mind the end goal to speak to data that is particular to each Internet get to endeavor. Fundamentally, a tick stream contains, in addition to other things, the IP address of root site, the entrance time, the alluding site, the URL of the referring site, the program technique, and the convention that was utilized. These days, a few business instruments are accessible for click stream examination and numerous more are available free on the web.

Web usage mining is the utilization of information mining methods to find use designs from Web information, keeping in mind the end goal to comprehend and better serve the necessities of Online applications .The point of web use mining is to catch, demonstrate, and break down the behavioral examples and profiles of clients perusing with a Site. The extricated and found examples are generally spoken to as accumulations of pages, items, or assets that are oftentimes gotten to by gatherings of clients with basic needs or interests .Web use mining contains three stages follows :

1. Initial processing This stage manages purifying and dividing of the snap stream information into an arrangement of client exchanges speaking to the exercises of every client amid various visits to the site. Preprocessing likewise manages changing over the use, substance and structure data contained in the different accessible information sources into the information reflections important for design disclosure.

2. Structure Revelation: In this stage, measurable investigation, database examination, and machine learning operations are performed to get shrouded designs mirroring the commonplace conduct of clients, and additionally outline estimations on Web resources, sessions, and clients. This stage draws upon techniques and calculations, for example, factual investigation, affiliation rules, bunching, characterization, successive example

mining, reliance demonstrating and other machine learning operations.

3. Structure Investigation: In this stage, the found examples and measurements are additionally handled, sifted, potentially bringing about total client models that can be utilized as contribution to applications, for example, suggestion motors, perception instruments, and Web examination and report age apparatuses. The primary inspiration is to sift through uninteresting standards or examples from the set found in the example disclosure arrange.

3. Proposed Work

Click stream analysis use click-stream data to conduct traffic analysis, Online-business market-based analysis and classification of customers based on their browsing history. The online customers are divided into four types and their behavior is classified as the Bargain Shopper, the Surgical Shopper, the Enthusiast Shopper and the Power Shopper. The click-stream information is typically separated from log documents and treats into the database and after that experts can make inductions utilizing distinctive plans of action. The online customers conduct and work process design are clarified as underneath

(a)The Bargain Shopper (shopper behavior)

Bargain shoppers check for the arrangements or offers, analyze costs broadly, Brandishing no brand devotion, yet these customers are searching at the most minimal cost. Their shopping example will be1.Check Promotion-mail,

2.ConnectWeb Site3.Search for Promotional Products,4.Compare prices with other websites Purchasing

(b)The Surgical Shopper (Customer behavior)

"Surgical" customer know precisely what they need before logging on the web and just buy the required thing. Normally they know the criteria on which they will base their choice, look for data to coordinate against that criteria, and buy when they are certain they have discovered precisely the correct item. Their shopping example will be Their shopping pattern will be 1.Connect Web Site, 2.Find stuff you want to buy, 3Filtered according to conditions Until the desired result is filtered 4.Purchasing decisions,5.Check item details

(c) The Enthusiast Shopper (Customer behavior)

Enthusiast customer use purchasinging as a form of previous and they shopping frequently and are the most adventurous shoppers. Their shopping pattern trends is 1 .Connect Web Site , 2.Most Popular Product Eye Shopping, 3.Add With List 4.Compare Other Wish Product 5. Check item details 6. Purchase decisions.

(d)The Power Shopper (Customer behavior)

People shop out of required necessity and they develop comportable shopping strategies to find what they want, will not want to spent time looking around. Their shopping trends will be 1.Connect Web Site 2. Find stuff you want to buy 3. Reviews confirmation 4.Check best reviews of other site 5 check item details 6 purchase decisions.

Process mining procedure are applied to Business Process Insight platform to gether web user behavior. In this paper,we experiment on custom click-stream logs from a substantial internet shopping website. To begin with the Internet clicks are contrasted and BPM occasions and after that present a procedure to order and change URLs into occasions. The theory assesses customary and custom process mining calculations to separate plans of action from web information. The models coming about because of examination, exhibit a disconnected perspective of the connection between pages, existing focuses and basic way taken by clients. The primary inspiration of the exploration, to utilize process mining procedure to yield organized formal models of client conduct that can give bits of knowledge of forthcoming change to the site.

Along these lines, it is conceivable to give simple and right comprehension of their clients' genuine communication designs on the site and their advancement. The proposal cases to contribute in following three noteworthy areas:

Web clicks are changed into errands appropriate for investigation and demonstrating with BPM devices. At that point the URLs are arranged that compare to web click sign into abnormal state errands that include both manual and programmed characterization strategies. Dissimilar to most process mining calculations that catch just the most widely recognized conduct so as to keep the subsequent model sufficiently straightforward, this theory likewise addresses this issue with

methods, for example, immersing the dataset with low recurrence conduct client watches out for spectator, grouping the procedure occurrences to separate example of conduct or utilizing learning based process mining calculation. These calculations are assessed and the utilization of the learning based mining calculation under an assortment of conditions and disclosing it appropriateness to separate process models that dynamic a total outline of customer route from genuine, unwanted data.

It seemed that web route imparts qualities to customary BPM exercises, for example, circles and parallel tasks. And furthermore, sessions just traverse a couple of minutes by and large and incorporate

no human mediation. Likewise, the tests comes about with revelation that, any investigation of web logs requires the characterization of URLs to higher legitimate undertakings, as the quantity of one of a kind URLs turn out to be too huge for human utilization and convention mining calculations. At long last, it is demonstrated that bunching calculations can naturally characterize URLs, requiring just that each group be named as various customers.

4. Methodology

Information extraction is the initial step, which is trailed by information preprocessing and afterward utilizing it for process revelation. Information extraction is one of the vital advance which incorporates getting event logs from the shopping site. Customer practices on the shopping site are recovered and watched utilizing investigation apparatus. Utilizing its Programming interface, JSON event log documents were removed. The information comprises of various passages like time, user CorrelationId, eventId, sessionBounce, program, OS, deviceType, URL, refererUrl, referrerHost,referrer HostClass and referrer SocialNetwork session-Begin, sessionStop. The time is in UNIX arrange, which is changed over into discernable information organize in information preprocessing stage. Preprocessing comprises of changing over the use, substance and structure data contained in the information sources into the information deliberations fundamental for design revelation. This stage comprises of changing over the occasion sign into organize reasonable for process mining.

To begin with the event logs are removed and it is changed over as petrinet display utilizing algorithm through the usage of tool PROM. At that point Petrinet is examined through heuristic miner and fuzzy miner calculations. The continuous event log, that is watched show is contrasted and the first work process example of the customer conduct fitness parsing measure, foot print level conformance checking , precision level checking ,structural appropriateness, behavioural appropriateness are measured to check the proposed conduct of the diverse shopping model in the shopping site.

(a)Alpha-algorithm

Alpha-algorithm,used at reconstructing causality from an arrangement of groupings of occasions and it builds a work process nets from event logs. It orders occasions successively, with the end goal that every event refers to a case and movement. It has issue with commotion, occasional conduct and complex directing develops.Existing business devices, for example, Perceptive Process Mining and Fluxicon Disco, and academic tools, such as Inductive Visual Miner, mainly focus on the control-flow perspective, and provide for data-aware process exploration. So the event logs are followed with alpha miner algorithm in PROM instrument and the Petri net model of the watched event logs with intricacy and deviations in the control stream are recognized. **(b) Heuristic Miner Algorithm**

Heuristic Mining algorithm that can deal with unwanted, and it gives the procedure for systyem, enrolled in an event log. Heuristic Miner is the extension of alpha algorithm and it considers the regular traces in the log. To get the process model it considers the sequence of the events within a case. Control flow perspective of The Heuristic Miner plug-in is used to find the deviation in the observed event sequences to form the observed work flow model using the PROM tool. Figure 11 and 12 presents deviated control flow from the original control flow constructs. This is achieved through Heuristic Miner algorithm and its conversion plug-ins. This control flow provides a long distance dependency of events and their fitness measures.

Table1.EventLog

case_id	event_id	time_stamp	activity	resource
1	71202	12/12/2012 11:58:00 AM	check promotion mail	ravi
1	71202	12/12/2012 11:59:00 AM	connect web site	ravi
1	71202	12/12/2012 11:60:00 AM	search for promotional products	ravi
1	71202	12/12/2012 12:05:00 PM	compare price with other websites	ravi
1	71202	12/12/2012 12:18:00 PM	purchasing decisions	ravi
2	71203	12/12/2012 12:18:00 PM	check promotion mail	rani
2	71203	12/13/2012 01:08:00 PM	connect web site	rani
2	71203	12/13/2012 01:28:00 PM	compare price with other websites	rani
2	71203	12/13/2012 01:32:00 PM	purchasing decisions	rani
3	71204	12/17/2012 11:28:00 AM	check promotion mail	akash
3	71204	12/17/2012 11:29:00 AM	connect web site	akash
3	71204	12/17/2012 11:32:00 AM	search for promotional products	akash
3	71204	12/17/2012 11:36:00 AM	purchasing decisions	akash
4	71204	12/17/2012 12:30:00 PM	connect web site	sam
4	71204	12/17/2012 12:46:00 PM	search for promotional products	sam
4	71205	12/17/2012 12:54:00 PM	compare price with other websites	sam

CONFORMANCE CKECKING USING CONTROL FLOW BENCHMARK ANALYSIS

Table2.Conformance checking using BenchmarkAnalysis

Benchmark metric / Item	Bergen Shopper Guess 1	Bergen Shopper Guess 2	Surgical Shopper Guess 1	Surgical Shopper Guess 2	Enthusiast Shopper Guess 1	Enthusiast Shopper Guess 2	Power Shopper Guess 1	Power Shopper Guess 2
Fitness Parsing Measure PM	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Token based fitness f	0.720	0.741	0.917	0.919	0.727	0.761	0.759	0.699
Fitness PF complete	0.000	0.000	0.000	0.260	0.000	0.000	0.000	0.000
Behavioral Appropriateness aB'	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Behavioral Precision BP	0.619	0.683	0.812	0.804	0.694	0.700	0.706	0.670
Behavioral Recall BR	0.619	0.683	0.812	0.804	0.694	0.700	0.706	0.670
Causal Footprint	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Structural Appropriateness aS'	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Structural Precision SP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Structural Recall SR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Duplicates Precision DP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Duplicates Recall DR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table3.Benchmark Metric per Item

Bench Mark Metric / Item	
Token-based Fitness (f) - It measures the fitness of the model by replaying every type of trace.	0 and 1
Fitness PF Complete – It checks for the number of events that could be parsed without problems during replay.	0 and 1
Behavioral Appropriateness (aB') – It measures how much behavior is allowed by the model which is not present in the log.	0 and 1
Behavioral Precision (BP) – It checks whether enabled activities in the model actually correspond to observed executions in the log.	0 and 1
Behavioral Recall (BR) – It is used in pattern recognition and information retrieval, through the construction of a confusion matrix	0 and 1
Causal Foot Print - A footprint is a matrix showing causal dependencies between activities	0 and 1
Structural Appropriateness (aS') – To express the presence of same behavior in the process model which results in complex model due to duplicate task (Transition with the same label, invisible task, transition without a label or label T)	0 and 1
Structural Precision (SP) – It assesses how many causality relations the mined model has that are not in the original model	0 and 1
Structural Recall (SR) – It checks how many causality relations from the original model are not included in the mined model.	0 and 1
Duplicates Precision (DP) – It is similar to precision. It checks how many duplicate tasks are mined	0 and 1
Duplicates Recall (DR) - It is similar to Recall. It checks how many duplicate tasks are in the referenced model.	0 and 1

5 Results and Discussion

The following metrics to measure the flow control of original event logs to coinside with the work flow model produces control flow deviations as 1 which implies the flow is not deviated in terms of following measures:

1.Casual Foot Print

2.Behavioural Appropriateness

3. Structural Appropriateness aS’

4. Structural Precision SP

5. Structural Recall SR

6. Duplicates Precision DP

7 Duplicates Recall DR

The above metrics are shown in table no 2. It is seemed that the metric Fitness Parsing Measure shows the value 0.000 which implies that there is a complete deviation for the observed model from the work flow model planned. It means the real time event traces are recorded not as planned in the work flow model. Fitness PF Complete shows the value of 0.000 implies there is a deviation in fitness of complete event log due to missing of sequential events Token-based Fitness (f) gives the value of 0.727 implies that there is a deviation in the fitness of single event.So,the complete event log is separated as tokens and each token is tested for its sequential occurrence. From the results it observed event log (process model) is deviated from the work flow model. The events are not taken place in the procedure work flow in the business environment. Most of the times the events are not taken place in the planned sequential order. The results are useful for future improvement to avoid deviation in the work flow in the business environment.

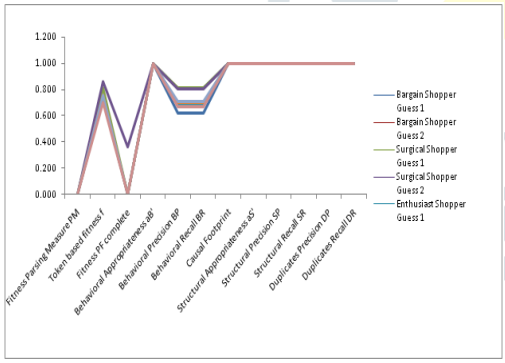

Fig1:Deviated level is observed in each categories of shoppers

6. CONCLUSION

In this paper, the metrics of fitness, behavioral appropriateness, token based fitness, quality, relevant event traces, structural appropriateness and structural quality are measured for a set of event logs. And there is a more deviation of fitness and behavioral appropriateness is measured between the original workflow model and the observed control flow metrics. It is informed that the control flow constructs of the observed event logs have to follow the workflow model more closely to improve the fitness and the appropriateness.

In this paper,we identifies the amount of quantity and quality of observed model, amount of deviation from the work flow model in the business process management domain and also identifies the place of deviation for all the group of shoppers in the ecommerce business.

The outcomes is used to identify the maximum behavior model in ecommerce business and market basket analysis, to improve product sales, and to design better collaborative algorithms for recommender system .

Future research may be carried out in complexity metrics of the same event logs, and analysis can be carried out with different algorithms.

REFERENCES

[1] Etzioni, 0. (1996). The World Wide Web: Quagmire or gold mine. Communications of the ACM, 39(ll), 65-68.
[2] Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web mining: information and pattern discovery on the World Wide Web. Proceedings of the 9th ZEEE International Conference on Tools with Artificial Intelligence, 558-567
[3] W. M. P. vander Aalst, "Business Process Management: A Comprehensive Survey," *ISRN Softw. Eng.*, 2013, pp. 1–37.
[4] W. M. P. vander Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, vol. 136. 2011, pp. 80.
[5] D. Simeonova, "BIG DATA AND PROCESS MINING," *DG DIGIT*, 2014. [Online]. Available: https://ec.europa.eu/digit-ict/sites/digit-ict/files/ictinterview.pdf. [Accessed: 13-Dec-2014].
[6] Nithya, T. (2013), 'Link Analysis Algorithm for Web Structure Mining', *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2, No. 8, pp. 2950-2954
[7] Herrouz, A., Khentout, C., Djoudi, M. (2013), 'Overview of Web Con-tent Mining Tools', *The International Journal of Engineering And Science*, Vol. 2, No. 6
[8] E-Commerece customer behavior analysis https://econsultancy.com/blog/64704-25-effectivedesign-patterns-for-ecommerce-site-search-results#i.1csntqn18pbf1h
[9] Malpani Radhika S and Dr.Sulochana Sonkamble, A Data Mining Approach to Avoid Potential Biases. International Journal of Computer Engineering and Technology, 6 (7), 2015, pp. 27-34.
[10] Dr .Anukrati Sharma, A Study on E – Commerce and Online Shopping: Issues and Influences. *International Journal of Computer Engineering and Technology*, 4(1), 2013, pp. 364–376.