Journal of Engineering and Applied Sciences 12 (Special Issue 12): 9548-9553, 2017
ISSN: 1816-949X

A Study on Classification of Users Shopping Behavior Process Model Using Click Stream Data

[1]Dileep Kumar Padidem and [2]C. Nalini
[1]Department of CSE, CMR College of Engineering and Technology, Hyderabad, India
[2]Department of CSE, Bharath University, Chennai, India

Abstract: The dynamic nature of web creates large massive volumes of information in structured and semi structured nature. The dynamic 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 world. So many research results has been produced using web analytics study which observes customers behavior through 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 propose of applying Business Process Methodologies (BPM) to event logs of ecommerce websites to study the challenges and potential benefits of such an approach. The method of general web access pattern is extracted and analyzed using knowledge discovery techniques to understand the usage patterns of the customers. This study have a clear insight of process mining, observation of web usage by customers (click stream data) as sequence of tasks and analysis and study on classification of users four important shopping behavior as bargain shopper, surgical shopper, enthusiast shopper and power shopper. The workflow model of these four types of shoppers and their real time behavior are analyzed using process mining tool.

Key words: Web mining, click stream, process mining, alpha-algorithm, heuristic miner algorithm, shopper, behavior

INTRODUCTION

Web mining uses the data mining techniques to automatically discover and extract information (knowledge) from web documents and services. The method of general 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. Sequences of tasks are observed through click stream analysis and a business process model may be discovered in web structure mining. This thesis suggests a plan and proposal about how to considered the user behavior in e-commerce sites as process and to discover a process model which enhances business intelligence.

Two different approaches were taken in initially defining web mining. First was a "process-centric view" which defined web mining as a sequence of tasks (Etzioni 1996; Aalst, 2011). Second was a "data-centric view" which defined web mining in terms of the types of web data that was being used in the mining process (Cooley *et al*., 1997; Simeonova, 2014). There are many works on web mining in terms of data-centric view and this paper considers its study on web mining in terms of process centric view which defined web mining as a sequence of tasks. This study suggest its plan and view to consider the click stream data in ecommerce websites as sequence of tasks and framing different optimum business models.

Click stream data are the electronic 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, often very big in 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 visitor takes while navigating the web. This path reflects choices, often very large in number made by the user 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. An important unit of observation in clicks stream data is the page visit the recording of a user's visit

Corresponding Author: Dileep Kumar Padidem, Department of CSE, CMR College of Engineering and Technology, Hyderabad, India

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 single website and they are known as "Site-centric." Site-centric click streams can provide very detailed records of user's behavior that is about their navigation and interaction with a given site.

Click-stream data provides the opportunity for a detailed look at the decision making process itself and knowledge extracted from it can be used for optimizing, influencing the process, etc. Underhill has conclusively proven the value of process information in understanding user's behavior in traditional sites. Research needs to be carried out in extracting process models from usage data, understanding how different parts of the process model impact various web metrics of interest and how the process models change in response to various changes that are made, i.e., changing stimuli to the user.

Literature review: Process mining complements existing approach Business Process Management (BPM). BPM combines knowledge from management sciences and applies this to operational business processes (CMCC., 2017; Sharma, 2013). 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.

Aalst (2013) and Nithya (2013) 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 (Nithya, 2013). Simeonova (2014) and Herrouz *et al.* (2013) defined process mining as techniques that help to find, screen and enhance genuine procedures by concentrating learning from event logs. Data is gathered from assorted types of systems and examined to identify deviations from standard processes and see where the bottlenecks are. Process mining is based on fact-based data and starts with an analysis of data, followed by the creation of a process model.

The techniques of data mining and knowledge discovery could be applied efficiently on web sites or e-Commerce sites. This specific application of data mining on e-Commerce web pages called web mining and it has taken much attention of researches. A new research area was derived from web mining for guiding the solutions to its specific requirements. Some researchers has worked on mining 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 data needed to accomplish such tasks is derived normally from a web server log file-almost all e-Commerce applications are web based. Click stream files are generated in order to represent information that is specific to each web access attempt. Basically, a click stream contains, among other things, the IP address of origin site, the access time, the referring site, the URL of the target site (i.e., the web page or object accessed) the browser method and the protocol that was used. Nowadays, several commercial tools are available for click stream analysis and many more are accessible free on the internet.

Web usage mining is the application of data mining techniques to discover usage patterns from web data in order to understand and better serve the needs of web-based applications. The aim of web usage mining is to capture, model and analyze the behavioral patterns and profiles of users browsing with a web site. The extracted and discovered patterns are usually represented as collections of pages, objects or resources that are frequently accessed by groups of users with common needs or interests. Web usage mining contains three phases as follows:

Preprocessing: This stage deals with cleansing and partitioning of the click-stream data into a set of user transactions representing the activities of each user during different visits to the site. Preprocessing also deals with converting the usage, content and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery.

Pattern discovery: In this stage, statistical analysis, database analysis and machine learning operations are performed to obtain hidden patterns reflecting the typical behavior of users as well as summary statistics on web resources, sessions and users. This stage draws upon methods and algorithms such as statistical analysis,

association rules, clustering, classification, sequential pattern mining, dependency modeling and other machine learning operations.

Pattern analysis: In this stage, the discovered patterns and statistics are further processed, filtered, possibly resulting in aggregate user models that can be used as input to applications such as recommendation engines, visualization tools and web analytics and report generation tools. The main motivation is to filter out uninteresting rules or patterns from the set discovered in the pattern discovery stage.

Proposed work: Click stream analysis use click-stream data to conduct traffic analysis, e-Commerce market-based analysis and classification of customers based on their browsing history. The online shoppers are classified into four main types and their behavior is classified as the bargain shopper, the surgical shopper, the enthusiast shopper and the power shopper. The click-stream data is normally extracted from log files and cookies into the database and then analysts can make inferences using different business models. The online shoppers behavior and workflow pattern are explained as:

The bargain shopper (customer behavior): Bargain-hunting shoppers check for the deals or offers, compare costs extensively, sporting no brand loyalty but these shoppers are looking for the lowest price. Their shopping pattern will be check promotion-mail, connect web site. Search for promotional products, compare prices with other websites purchasing.

The surgical shopper (customer behavior): "Surgical" shoppers know exactly what they want before logging online and only purchase the required item. Typically they know the criteria on which they will base their decision, seek information to match against that criteria and purchase when they are confident they have found exactly the right product. Their shopping pattern will be connect web site, find stuff you want to buy, filtered according to conditions until the desired result is filtered purchasing decisions, check item details.

The enthusiast shopper (customer behavior): Enthusiast shoppers use shopping as a form of pastime and they purchase frequently and are the most adventurous shoppers. Their shopping pattern trends is connect web Site, most popular product eye shopping, add with list, compare other wish product, check item details, purchase decisions.

The power shopper (customer behavior): People shop out of necessity and they develop sophisticated shopping strategies to find what they want and do not want to waste time looking around. Their shopping trends will be connect web site, find stuff you want to buy, reviews confirmation, check best reviews of other site, check item details 6 purchase decisions.

In particular, process mining techniques are applied to business process insight platform to analyze web user behavior. The author experiment on custom click-stream logs from a large online shopping site. First, the web-clicks are compared with BPM events and then present a methodology to classify and transform URLs into events. The thesis evaluates traditional and custom process mining algorithms to extract business models from web data. The models resulting from analysis, present an abstracted view of the relation between pages, existing points and critical path taken by customers. The main motivation of the research, to use process mining technique to yield structured formal models of user behavior that can provide insights of prospective improvement to the site.

So, it is possible to provide easy and correct understanding of their user's real interaction patterns on the site and their development. The thesis claims to contribute in following three major areas:

Web clicks are transformed into tasks suitable for analysis and modeling with BPM tools. Then, the URLs are classified that correspond to web click logs into high level tasks that involve both manual and automatic classification techniques. Unlike most process mining algorithms that capture only the most common behavior in order to keep the resulting model simple enough, this thesis also addresses this issue with techniques such as saturating the dataset with low frequency behavior user tends to observer, clustering the process instances to extract pattern of behavior or using knowledge-based process mining algorithm. These algorithms are evaluated and the use of the knowledge-based mining algorithm under a variety of conditions and explaining it suitability to extract process models that abstract a complete over-view of user navigation from real, noisy data.

It is observed that web navigation shares characteristics with traditional BPM activities such as loops and parallel tasks. And also, sessions only span a few minutes on average and include no human intervention. In addition to this, the experiments results with discovery that any analysis of web logs requires the classification of URLs to higher logical tasks as the number of unique URLs become too big for human

consumption and tradition mining algorithms. Finally, it is shown that clustering algorithms can automatically classify URLs, requiring only that each cluster be named as different shoppers.

MATERIALS AND METHODS

Data extraction is the first step which is followed by data preprocessing and then using it for process discovery. Data extraction is one of the important step which includes getting event logs from the shopping site. User behaviors on the shopping website are retrieved and observed using analytics tool. Using its API, JSON event log files were extracted. The data consists of different entries like time, userCorrelationId, eventId, sessionBounce, browser, OS, deviceType, URL, refererURL, referrerHost, referrerHostClass and referrer SocialNetwork session-Start, sessionStop. The time is in UNIX format, which is converted into readable data format in data preprocessing stage. Preprocessing consists of converting the usage, content and structure information contained in the data sources into the data abstractions necessary for pattern discovery. This stage consists of converting the event logs into format suitable for process mining.

First, the event logs are extracted and it is converted aspetrinet model using alpha algorithm through the usage of tool PROM. Then petri net is analysed through heuristic miner and fuzzy miner algorithms. The real time event log, that's is observed model is compared with the original work flow pattern of the customer behavior and the different fitness parsing measure, foot print level conformance checking, precision level checking, structural appropriateness, behavioural appropriateness are measured to check the intended behavior of the different shopping model in the shopping site.

Alpha-algorithm: Alpha-algorithm used in process mining, aimed at reconstructing causality from a set of sequences of events and it constructs a workflow nets from event logs. It orders events sequentially such that each event refers to a case and activity. It has problem with noise, infrequent behavior and complex routing constructs. The 26 existing commercial tools such as 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 traced with alpha miner algorithm in PROM tool and the petri net model of the observed event logs with complexity and deviations in the control flow are identified.

Heuristic miner algorithm: Heuristic mining algorithm is a practically applicable mining algorithm that can deal with noise and it gives the principal behavior of the system, registered in an event log. Heuristic miner is the extension of alpha algorithm and it considers the frequency of 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 1 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 (Table 1-3).

Table 1: EventLog

Case-id	Event id	Time stamp	Activity	Resource
1	712023	12/12/2012 11:58:00 AM	Check promotion mail	Ravi
1	712024	12/12/2012 11:59:00 AM	Connect web site	Ravi
1	712025	12/12/2012 11:60:00 AM	Search for promotion products	Ravi
1	712026	12/12/2012 12:05:00 PM	Compare price with other web sites	Ravi
1	712027	12/12/2012 12:18:00 PM	Purchasing decision check	Ravi
2	712030	12/12/2012 12:18:00 PM	promotion mail	Rani
2	712033	12/13/2012 01:08:00 PM	Connect web site	Rani
2	712037	12/13/2012 01:28:00 PM	Compare price with other sites	Rani
2	712038	12/13/2012 01:32:00 PM	Purchasing decisions	Rani
3	712040	12/17/2012 11:28:00 PM	Check promotion mail	Akash
3	712041	12/17/2012 11:29:00 PM	Connect web site	Akash
3	712044	12/17/2012 11:32:00 PM	Search for promotional products	Akash
3	712045	12/17/2012 11:36:00 PM	Purchasing decisions	Akash
4	712047	12/17/2012 12:30:00PM	Connect web site	Sam
4	712048	12/17/2012 12:46:00 PM	Search for promotional products	Sam
4	712051	12/17/2012 12:54:00 PM	Compara price with other web sites	Sam

Table 2:Conformance checking using benchmark analysis

Benchmark metric/Item fitness parasing	Bargam shopper guess 1	Bargam shopper guess 2	Sargical shopper guess 1	Sargical shopper guess 2	Enthusiast shopper guess 1	Enthusiast shopper guess 2	Power shopper guess 1	Power shopper guess 2
Measure (PM)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Loken based fitness (f)	0.720	0.742	0.017	0.000	0.727	0.766	0.760	0.000
Fitness (PP) complete	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Behavioral appropriatness (aB')	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Behavioral Procision (BP)	0.619	0.693	0.812	0.801	0.691	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 appropriatness (aB')	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Structural Procision (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 Procision (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

Table 3: Benchmark metric per item

Bench mark metric/Item	Values
Token based fitness (f): Meaures the fitness of the model by replaying every type of trace	0 and 1
Fitness PF complete: It checks for the number of event that could be parsed without problems during replay	0 and 1
Behavioral appropriateness (aB'): It meaures how much behavior isallowed by the mdel 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 footprint: 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 doplicate tesk (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 relation 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
Doplicate Precision (DP): It is similar to precision. It checks how many duplicate tasks are mined	0 and 1
Doplicate Recall (DR): It is similar to recall. It checks how many duplicate tasks are in the referenced model	0 and 1

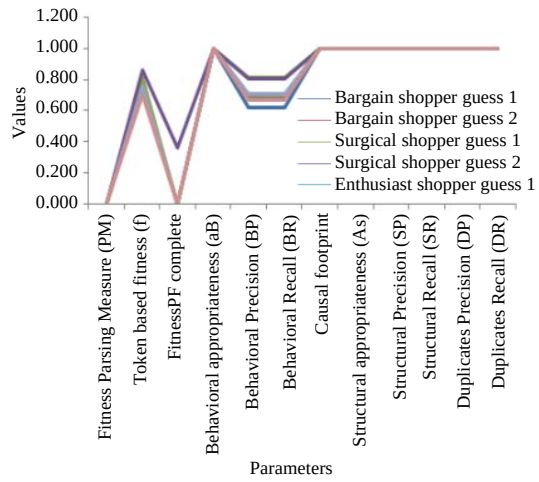

Fig. 1: Deviated level is observed in each categories of shoppers

RESULTS AND DISCUSSION

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

- Casual foot print
- Behavioural appropriateness
- Structural appropriateness (aS')
- Structural Precision (SP)
- Structural Recall (SR)
- Duplicates Precision (DP)
- Duplicates Recall (DR)

The above metrics are explained in Table 2. It is observed 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 which implies that there is a deviation in fitness of complete event log due to missing of sequential events token-based fitness (f) shows the value of 0.727 which implies that there is a deviation in the fitness of single event. It means the complete event log is separated as tokens and each token is tested for its sequential occurrence. From the results it is learnt that observed event log (process model) is deviated from the work flow model. The events are not taken place in the planned 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.

CONCLUSION

In this, 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 high deviation of fitness and behavioral appropriateness is measured between the original workflow model and the observed control flow constructs. It is suggested 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. This thesis identifies the amount of quantity and quality of observed model, amount of deviation from the work flow model in the business process management domain. This thesis identifies the place of deviation for all the group of shoppers in the ecommerce business. These results helps to identify the optimum behavior model in ecommerce business and market basket analysis, to improve product sales and to design better collaborative algorithms for recommender system.

RECOMMENDATION

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

REFERENCES

Aalst, V.D.W.M., 2013. Business process management: A comprehensive survey. ISRN. Software Eng., 2013: 1-37.

Aalst, W.M.P.V.D., 2011. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer, Berlin, Germany, ISBN:9783642193453, Pages: 352.

CMCC., 2017. E-Commerece customer behavior analysis. Centaur Media Commercial company, UK. https://econsultancy.com/blog/64704-25-effectived esign-patterns-for-ecommerce-site-search-results#i .1csntqn18pbf1h

Cooley, R., B. Mobasher and J. Srivastava, 1997. Web mining: Information and pattern discovery on the world wide web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence 1997, November 3-8, 1997, IEEE, Newport Beach, California, ISBN:0-8186-8203-5, pp: 558-567.

Etzioni, O., 1996. The world wide web: Quagmire or gold mine. Commun. ACM, 39: 65-68.

Herrouz, A., C. Khentout and M. Djoudi, 2013. Overview of web content mining tools. Intl. J. Adv. Res. Comput. Sci. Software Eng., 3: 375-385.

Nithya, T., 2013. Link analysis algorithm for web structure mining. Intl. J. Adv. Res. Comput. Commun. Eng., 2: 2950-2954.

Sharma, D.A., 2013. A study on E-commerce and online shopping: Issues and influences. Intl. J. Comput. Eng. Technol., 4: 364-376.

Simeonova, D., 2014. Big data and process mining. Directorate General for Informatics DIGIT, Belgium, Europe. https://ec.europa.eu/digit-ict/sites/digit-ict/ files/ictinterview.pdf.