COMMUNICATIONS IN SCIENCE AND TECHNOLOGY
Homepage: cst.kipmi.or.id

Heuristics miner for e-commerce visitor access pattern representation

Kartina Diah[a*], Wawan Yunanto[b]

[a]*Teknik Informatika, Politeknik Caltex Riau, Jl. Umbansari No.1, Pekanbaru 28265, Riau - Indonesia*
[b]*Sistem Informasi, Politeknik Caltex Riau, Jl. Umbansari No.1, Pekanbaru 28265, Riau - Indonesia*

Article history:
Received: 30 October 2016 / Received in revised form: 7 December 2016 / Accepted: 13 February 2017

Abstract

Click stream data from an e-commerce website can form a certain pattern that describes visitor behavior. This pattern can be used to determine alternative access sequence to surf the website. σ-Algorithm and Genetic Mining are two of the most common methods for pattern recognition that use frequent sequence item set approach. This study used heuristic miner algorithm, an advanced form of these methods, to discover the pattern of visitor behavior in e-commerce website. σ-Algorithm assumes that an activity in a website recorded in the data log is a complete sequence from start to finish, without any tolerance for incomplete data or data with noise. On the other hand, Genetic Mining is a method that tolerates incomplete data or data with noise, so it can generate a more detailed e-commerce visitor access pattern. In this study, the same sequence of events was obtained from six-generated patterns. The resulting pattern describes the sequence of how visitors access the e-commerce website. This sequence can be used to enhance the e-commerce website based on visitor behavior.

Keywords: Heuristic Miner, Visitor Behavior, Access Pattern

1. Introduction

E-commerce is a modern web based application that consists of buying process, sales process, and transfer or exchange goods and services, via the internet. E-commerce does not require the physical presence of the customers, and its transactions can be done at anytime, anywhere, regardless of distance [1,2]. In e-commerce, customers purchase goods or services using the facilities available in the e-commerce websites across the world [3]. E-commerce is defined as an activity of selling and buying products, provision of services and information via computer networks, especially the internet [4]. Furthermore, as Riggins states that e-commerce is not limited to trading activities, but includes a variety of processes in an organization that supports the objectives of the business. The number of transactions that occur using the e-commerce from around the world creates an enormous amount of data generated in a variety of forms and formats, ranging from transactions data, customer data, production data, sales data, consumers access data, web navigation data, and so on. With the fact that the data produced is huge, a wide range of technologies and methods are developed to find hidden information that can be derived from these data. E-commerce companies record user activities to gain actual information about their customers based on user behaviors when they perform certain procedures. The user activities are recorded in the form of log data [5]. The log data contain the identity of e-commerce users along with their browsing behavior on the website.

Heuristic Miner algorithm is an evolved form of σ-Algorithm. The two methods use frequent item set approach to do the mining process. σ-Algorithm assumes that an activity in data log is complete sequence from start to finish, without any tolerance for incomplete data or data with noise. To obtain the access pattern, Heuristic Miner needs log data which consists of timestamps, cases, and activities. Timestamps are used to determine activity sequences. The resulting sequence is a heuristic net that describes complex access pattern on a website. This pattern shows the process flow that happens in actual data [6].

Various mining techniques have been used to extract information from log data. As Senkul and Salin note [7], Web Usage Mining is used to provide the recommended pages for e-commerce website visitors based on profiles obtained from log data. Mishra conducted a research in extracting the log data using FP-Growth algorithm to gain the most frequent visitor access patterns. Classification method is also used to extract information from the log data, as conducted in [8], to gain access patterns of customers from India and outside India.

Based on these studies, Heuristic Miner algorithm is applied to e-commerce log data to predict customer browsing behavior. Then the prediction is compared to current e-commerce visitor's access data, so it can serve as the basis for functionality and appearance enhancement of e-commerce

* Corresponding author.
Email: diah@pcr.ac.id.

websites. The main purpose of this process is to provide services that better match the user's interests [9,10]. To conduct an analysis of web logs, the web log data first needs to be split into sessions. Sessions are defined as the period of continuous web browsing activity or web page display order. Sessions can also be seen as a sequence of user behavior such as visiting a website, doing the work, and then leaving or logging out from the website [11]. Furthermore, data preprocessing is performed to align the log data with the data needed at the pattern discovery stage.

This research proposes an alternative way to obtain visitor access pattern using heuristic miner algorithm on e-commerce website data log. The underlying idea of obtaining this pattern is to enhance the e-commerce website based on visitor behavior.

2. Materials and Methods

2.1. Data

This study uses 500 click stream data from 7 different e-commerce websites to gain consumers' access patterns. These collections of log data are downloaded from ECML/PKDD 2005 Discovery Challenge (lisp.vse.cz/challenge/CURRENT/). The log data consist of records of every user activity from 7 different e-commerce website from the first time they interact with the website until they leave or log out from the website

2.2. Method

Based on the type of the extracted data, web mining can be organized into 3 categories [12-17], namely:

1. Web Content Mining (WCM)
   A data mining technique used to generate information about the content of a web document. The web document can be text, images, audio, video, or structured records such as lists and tables.
2. Web Structure Mining (WSM)
   A data mining technique used to generate information about the structure of a website. This technique is divided into two parts: hyperlinks and document structures.
3. Web Usage Mining (WUM)
   A data mining technique used to find web usage patterns from web usage data in order to understand and provide better services to the web-based applications [18]. The resulting pattern is derived from the classification of web data, application server data, and application level data.

Each of this 3 categories can be applied separately or simultaneously [17].

Web Usage Mining (WUM) focuses on techniques used to predict the behavior of users as they interact with the world wide web. WUM collects data from web log records to find user access patterns as they browse web pages. The results of this analysis can be used for pages personalization, system enhancement, sites modification, business intelligence, or user characterization.

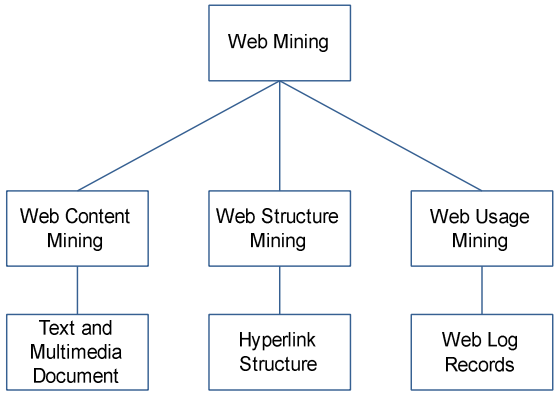


Fig. 1. Web mining categories and objects [17]

Table 1. Web mining categories [12]

Web Mining				
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR View	DB View		
View of Data	Unstructured	Semi Structure	Link Structure	Interactivity
	Structure	Web Site as DB		
Main Data	Text documents	-Hypertext Documents	Link Structure	Server Logs
	-Hypertext documents			Browser Logs
Represen-tation	Bag of words, n-gram Terms,	Edge Labeled Graph,	Graph	Relational Table
	Phrase, Concepts or ontology	Relational		Graph
	Relational			
Method	Machine Learning	Proprietary algorithms	Proprietary algorithms	Machine Learning
	Statistical (Include NLP)	Association rules		Statistical
				Association Rules
Applicatio n Category	Categorizatio n	Finding frequent sub structure	Categorizatio n	Site Construction
	Clustering	Web site schema discovery	Clustering	Adaptation and Management
	Finding extract rules			Marketing,
	Finding pattern in text			User Modeling

The steps undertaken in WUM is divided into the following 3 stages [17,19,20]:

1. Pre-processing
   The available data have a tendency to contain noise, incomplete and inconsistent. At this stage, the data will be processed to suit the needs of the next phase. This stage includes data cleansing, data integration, data transformation and data reduction.
2. Pattern discovery
   At this stage, several methods and algorithms such as statistic, data mining, machine learning and pattern recognition can be applied to get the pattern.

3. Pattern Analysis
   The pattern that has been found then analyzed and displayed using visualization and interpretation, in order to be easily understood by the user.

2.3. Heuristics Miner

Heuristic Miner is one of the algorithms in Mining Process used in the discovery phase. This algorithm focuses on measuring the frequency dependency between events and traces to build a process model. There are several steps that have to be implemented in this phase [6]:

- Building a Dependency Graph

Dependency Graph is a model that represents the dependency (causality) between events. There are 3 processes in building a dependency graph, and they are: the creation of matrix dependency, one loop dependency length, and two loop dependency length.

- Causal Matrix

In reality, a process can be done simultaneously (parallel), but in an event log, it is very difficult to determine whether the process runs sequentially or simultaneously. To avoid errors in the process model visualization, this Heuristic Miner uses Causal Matrix to represent the process model. Causal Matrix creation is conducted after the Dependency Graph is built. Whether an event has branches or not, it can be seen within this Dependency Graph. In this Causal Matrix there are two types of non-observal activities, that are AND and XOR. Non-observal AND activity states that a branching activity can be done in parallel or simultaneously, while the non-observal XOR activity states that a branching activity may only select one lane only.

3. Results and Discussion

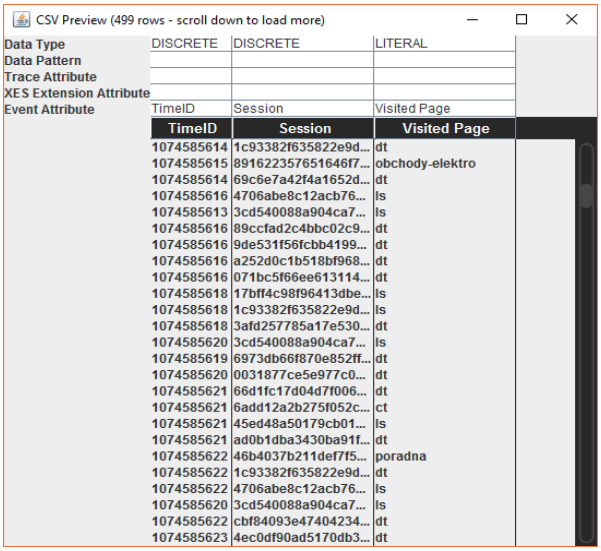
3.1. Log Data Filtering

Sequence of events was generated from log data, but sometimes log data could define a sequence of event appropriately. This problem arose because there was no specific standard on how to record visitor's activities, so the information stored in the log data varied. In order to get a sequence of event appropriately, the log data had to be preprocessed. In this study, 2 types of data were filtered: sessions and timestamps.

Session was defined as the unit that counted the amount of time of the event done by one particular web visitor. The length of time for a session varied. In e-commerce, commonly used standard for one session is 30 minutes, so that when a visitor visits an online store in 31 minutes, it is considered as a new session and a new visitor. This affected the sequence of events made by visitors in this study.

Timestamp is a marker of time documented by a computer when an event occurs. Timestamp can be used as a time functions in the smallest scope ranging from hours, minutes, seconds, down to milliseconds. In the same timestamp can occur several different events from different online shop visitors. A session may consist of a number of different timestamps. Fig. 2 is a click stream data table consisting of TimeID as timestamps, Sessions and Visited Pages.

Log data filtering conducted in this study provided a clear limit when an event initiated. An event was initiated when a visitor started accessing an online shop website, which was the home page. Every activity undertaken by the visitor, after an event was initiated, was considered as part of the event in the visitor's session. The closing of the event was not limited in order to get the habit patterns of online store's visitor in general (i.e. whether it would always end up until the checkout process or not).


CSV Preview (499 rows - scroll down to load more)

Data Type	DISCRETE	DISCRETE	LITERAL
Data Pattern			
Trace Attribute			
XES Extension Attribute			
Event Attribute	TimeID	Session	Visited Page

TimeID	Session	Visited Page
1074585614	1c93382f635822e9d...	dt
1074585615	891622357651646f7...	obchody-elektro
1074585614	69c6e7a42f4a1652d...	dt
1074585616	4706abe8c12acb76...	ls
1074585613	3cd540088a904ca7...	ls
1074585616	89ccfad2c4bbc02c9...	dt
1074585616	9de531f56fcbb4199...	dt
1074585616	a252d0c1b518bf968...	dt
1074585616	071bc5f66ee613114...	dt
1074585618	17bff4c98f96413dbe...	ls
1074585618	1c93382f635822e9d...	ls
1074585618	3afd257785a17e530...	dt
1074585620	3cd540088a904ca7...	ls
1074585619	6973db66f870e852ff...	dt
1074585620	0031877ce5e977c0...	dt
1074585621	66d1fc17d04d7f006...	dt
1074585621	6add12a2b275f052c...	ct
1074585621	45ed48a50179cb01...	ls
1074585621	ad0b1dba3430ba91f...	dt
1074585622	46b4037b211def7f5...	poradna
1074585622	1c93382f635822e9d...	dt
1074585622	4706abe8c12acb76...	ls
1074585620	3cd540088a904ca7...	ls
1074585622	cbf84093e47404234...	dt
1074585623	4ec0df90ad5170db3...	dt

Fig. 2. Click stream data of e-commerce website

In this study, log data filtering is conducted using Filter Log package in ProM. The filtering process is performed on several variables as follows: start event, and event, and event filters.

Table 2. Log data filtering results

Filtering	Start Event	End Event	Event Filter
1	90%	90%	90%
2	80%	80%	80%
3	90%	90%	100%
4	80%	80%	100%

3.2. Pattern Discovery

Pattern discovery was conducted to obtain e-commerce website visitor behavior patterns. The results from the log data filtering process in the previous stage was used in this stage. The values of variables resulting from filtering process were as follows: start event = 80%, end event = 80%, and event filter = 100%. In this study, the pattern discovery process was conducted using Heuristic Miner Algorithm tool in ProM. The variable dependency values used in heuristic miner algorithm was tested on several values, that is 80%, 90%, 100%, and without dependency value. Fig. 3 is a pattern resulting from heuristic miner with dependency value of 80%. This pattern describes the sequence of e-commerce website visitors behavior as follows:

1. home → product category → product sheet
2. home → product category → product sheet → detail of product → product sheet
3. home → fulltext search → detail of product → list of brand names → product sheet

4. home → fulltext search → detail of product → parameters based search → product sheet

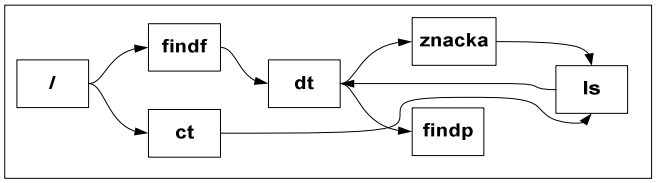


Fig. 3. Pattern discovery with dependency 80%

Fig. 4 is a pattern resulting from heuristic miner with dependency value of 90%. This pattern describes the sequence of e-commerce website visitors behavior as follows:

1. home → product category → product sheet → detail of product → parameters based search → product sheet
2. home → product category → product sheet → detail of product → list of brand names → product sheet
3. home → fulltext search → detail of product → list of brand names → product sheet
4. home → fulltext search → detail of product → parameters based search → product sheet

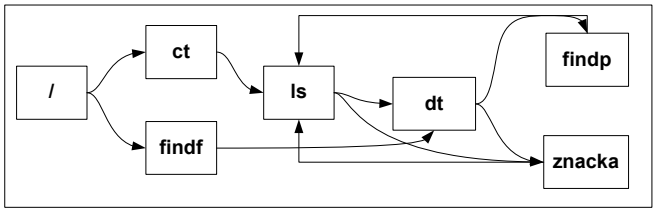


Fig. 4. Pattern discovery with dependency 90%

Fig. 5 is a pattern resulting from heuristic miner with dependency value of 100%. This pattern describes the sequence of e-commerce website visitors behavior as follows:

1. home → product category → product sheet → detail of product → parameters based search → product sheet
2. home → product category → product sheet → detail of product → list of brand names → product sheet
3. home → fulltext search → detail of product → list of brand names → product sheet
4. home → fulltext search → detail of product → parameters based search → product sheet

Fig. 6 is a pattern resulting from heuristic miner regardless of the dependency value. This pattern describes the sequence of e-commerce website visitors behavior as follows:

1. home → product category → product sheet
2. home → product category → product sheet → detail of product → product sheet
3. home → fulltext search → detail of product → list of brand names → product sheet
4. home → fulltext search → detail of product → parameters based search → product sheet

Based on the resulting pattern, causal matrix on heuristic miner algorithm indicated that an event can be done in parallel or sequentially. An event is done in parallel with the presence of branches in some events resulting from the patterns generated. This can be seen from the 4 resulting patterns. In the beginning of the home state, there are 2 branches, go to full text search or go to product category. It means that after visitor accesses the home page, they usually access the full text search or product category before they move to the next page.

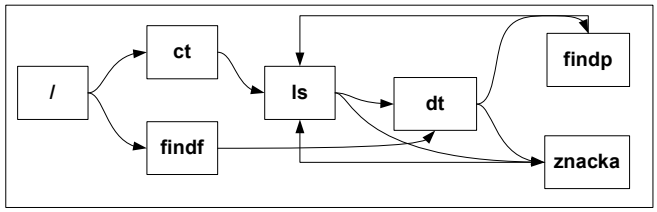


Fig. 5. Pattern discovery with dependency 100%

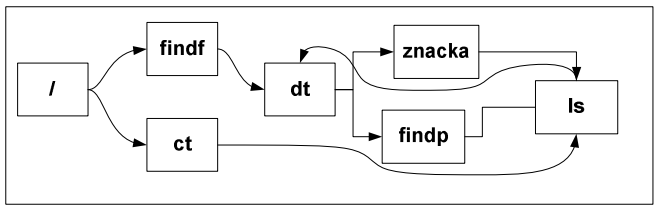


Fig 6 Pattern discovery with dependency ignored

The sequential pattern on the other hand is shown by the direction of the arrows depicting the sequence of events from one event to another event. As an example, depicted in Figure 3, the second access pattern is home → product category → product sheet → detail of product → product sheet. This shows that product sheet is the last event but not the last state because there is a reverse direction arrow to detail of product and then back to product sheet again. This access sequence describes the events passed by visitors.

4. Conclusion

Six patterns describing the e-commerce website visitor behavior are generated from the heuristic miner algorithm. A number of events that are used in the pattern discovery phase depends on the log data filtering variables used in pre-processing. In this study, the number of events used on heuristic miner algorithm were 7 events. Dependency variables used in heuristic miner algorithm affect the sequence of events and the end of event conducted by e-commerce website visitors. All of the resulting patterns are initiated from the home event, because at the log data filtering process, it is determined that when a visitor visits the home page, it will initiate the event, while the closing of the event is undetermined. The same sequence of event is obtained from the six generated patterns, namely that visitors are often access the home page and then the product category page or the home page and then the full text search page.

References

1. J. Lu and S. S. Gokhale, *Resource provisioning in an e-commerce application, 10th IEEE Conf. E-Commerce Technol. Fifth IEEE Conf. Enterp. Comput. E-Commerce E-Services, United State of America, 2008, pp. 209–214.*
2. N. H. Rawi, M. A. Bakar, R. Bahari, and A. M. Zin, *Development environment for layout design of e-commerce applications using block-based approach, Int. Conf. Electr. Eng. Informatics, Indonesia, 2011, pp.1-5.*

3. I. P. Tatsiopoulos, N. A. Panayiotou, and S. T. Ponis, *A modelling and evaluation methodology for e-commerce enabled BPR*, in *Comput. Ind. 49*, Denmark, 2002, pp. 107–121.
4. H. J. Wen and H. Chen, *E-commerce web site design : strategies and models*, Inf. Manag. Comput. Secur. 9 (2001) 5–12.
5. F. J. Riggins, *Toward a unified view of electronic commerce*, Commun. ACM. 41 (1998) 10-16.
6. P.Weber, B.Bordbar and Tino. P, *A principled approach to mining from noisy logs using heuristics mner*, IEEE (2013) 119–126.
7. P.Senkul and S.Salin, *Improving pattern quality in web usage mining by using semantic information*, Knowl. Inf. Syst. 3 (2012) 527–541.
8. Azam and N. Tabrez, *Data mining of web access logs using classification techniques*, Int. J. Res. Appl. Sci. Eng. Technol. 2 (2014) 1–5.
9. J. Vellingiri and S. C. Pandian, *A survey on web usage mining*, Glob. J. Comput. Sci. Technol. 11 (2011) 1-15.
10. K. B. Patel, J. A. Chauhan, and J. D. Patel, *Web mining in e-commerce : Pattern discovery , issues and applications*, Int. J. P2P Netw. Trends Technol. 1 (2011) 40–45.
11. L.Sun and X.Zhang, *Efficient frequent pattern mining on web log data*, *Adv. Web Technol. Appl. Sixth Asia-Pacific Web Conf. APWeb, Australia, 2004, pp. 1–24.*
12. D. M. Rathod, *A review on web mining*, Int. J. Eng. Res. Technol. 1 (2012) 108-113
13. N. Kaur, *Exploration of webminer system*, Int. J. Res. IT Manag. 2 (2012) 239–248.
14. R. Iváncsy and I. Vajk, *Frequent pattern mining in web log data*, Acta Polytech. Hungarica, (2006) 7– 90.
15. R. Kosala, B.Heverlee, and H. Blockeel, *Web mining research : A survey*, ACM SIGKDD Explor. Newsl. 2 (2000) 1-15.
16. Y. Wang, *Web mining and knowledge discovery of usage patterns*, CS748T Proj. 1 (2000) 1–25.
17. M. Gomes, *Web structure mining : An introduction*, in *IEEE Int. Conf. Inf. Acquis.*, China, 2005, pp. 590–595.
18. J. Srivastava, B. Mobasher, J. Han, and N. Jain, *Web Mining : Pattern discovery from world wide web transactions*, in *The 9th IEEE International Conference on Tools with AI (ICTAI,97), 1997, pp.1-11.*
19. R. Cooley, B. Mobasher, and J. Srivastava, *Data preparation for mining world wide web browsing patterns*, Knowl. Inf. Syst. 1 (2013) 5–32.
20. M. Han, J., Kamber, *Data mining: concept and techniques*. Morgan Kaufmann Publisher, 2000.