Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
15/1 (2022), 33-38. DOI: http://dx.doi.org/10.21609/jiki.v15i1.1024

Predicting Analysis of User's Interest from Web Log Data in e-Commerce using Classification Algorithms

Saucha Diwandari, Ahmad Tri Hidayat

[1,2]Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Ringroad Utara, Sleman, Yogyakarta, Indonesia

E-mail: saucha.diwandari@staff.uty.ac.id, ahmad.tri.h@uty.ac.id

Abstract

The accelerated development of e-commerce has been a concern for businesspeople. Businesspeople should be able to gain customer interest in a variety of ways so that their companies can compete with others. Analyzing click-flow data will help organizations or firms assess customer loyalty, provide advertising privileges, and develop marketing strategies through user interests. By understanding consumer preferences, clickstream data analysis may be used to determine who is participating, assist companies in evaluating customer contentment, boost productivity, and design marketing strategies. This research was performed by defining experimental user interests using Dynamic Mining and Page Interest Estimation methods. The findings of this analysis, using three algorithms at the pattern discovery page, demonstrated that the Decision Tree method excelled in both methods. It indicated that the operational performance of the Decision Tree performed well in the assessment of user interests with two different approaches. The findings of this experiment can be used as a proposal for researching the field of web usage mining, collaborating with other approaches to achieve higher accuracy values.

Keywords: *user interest; web usage mining; classification; e-commerce; web log data*

1. Introduction

The increasing rise of e-commerce is the primary driver of faster and more convenient commercial transactions. Users always expect the applications to provide whatever they want, quick and convenient e-commerce features for consumers to access, and information about products relevant to consumer interests. This information is precious for building loyal consumers and can improve the business strategies of e-commerce. Many businesses have understood the value of offering goods and services tailored according to customer needs. It can be achieved by defining the interests of the user. The document's substance also determines the user's interest that the user has read [1]. Clickstream data analysis can also select the people involved, help businesses recognize customer satisfaction, encourage productivity, and develop marketing campaigns by identifying consumer preferences. Clickstream data is the essential piece of information for firms looking to tailor their offerings to their clients[1]. It has a lot of value and may provide you with a lot of information about the people who visit your website. The document's substance also determines the user's interest that the user has read. In other research, user interest may also be established by overall access time, pages on the website that are most often viewed, or the pages that have been accessed the most recently. Several other specialists looked into user interests and discovered that they are expressed in the posts they visit and the ones they respond to [2]. To develop a more trustworthy model of consumer behaviour, pattern discovery with the proper methodology is also necessary. Several techniques have been employed., such as KNN[2], Naïve Bayes[3], Decision Tree[4]–[6]

Large-scale data expansion and the use of network data mining to explore customer preferences and user interests accompanied the e-commerce company's growth. Several studies have created models to forecast consumer preferences based on user purchase habits, web page visit counts, and web surfing paths. Several user attributes are used to measure market interest[6][7] the order in which pages are discussed, as well as the length of time spent on each page and the frequency with which they are seen, are all factors to. Zheng *et al*[9] conducted a comprehensive analysis of user browsing time with user interaction. However, they did not take into account the number

of visits or the order in which they occurred. It is different from the research of Yong Li *et al*[8] that used the time visit and the sequence visit. This study proposes an interest stickiness measurement algorithm and reveals the reliability and accuracy of the algorithm. Kim *et al*[10] was utilizing the PIE (Page Interest Estimation) to perform a research to determine user interest approach, which measures user interest by using information related to browsing time, frequency of visits, and pages visited. The findings of this analysis suggest that the PIE approach can be used to assess user interest. The dynamic Mining method is used to measure consumer interest[11]. In several studies, it is reported that the Dynamic Mining method is capable of determining user interest. In a variety of studies, it is reported that the Dynamic Mining approach is capable of determining user interest. In this paper, we will measure user interest by using PIE and Dynamic Mining approaches and compare the two methods that are best used in the case study that we have.

This article uses two techniques to characterize user interest based on numerous indications such as user id, page-time browsing, tool used, and page visited, namely, Dynamic Mining and Page Interest Estimation to predicted potential user and not potential. The potential users are users who have a high-interest value. The high-interest value used in this article for the PIE approach if user interest > 2 and in Dynamic Mining user interest > 600.

The information regarding potential users can be used as part of product marketing targets for the company. We have performed comparisons with multiple classification algorithms to identify patterns.

2. Material And Method

2.1 Page Interest Estimation

Examining the user's browsing activity and evaluating the user's interest in the web page resulted in an implicit estimate of user interest in the website. Many studies have been conducted to gauge user interest in the website. Saucha *et al*[3] used this approach to assess user interest in products sold on the e-commerce market. Yan LI *et al*[12] devised a technique for estimating page interest that builds an accurate user profile without disturbing the user and does not require personal information. The length of the reference, i.e. the time spent on the accessible page, including the hours lost on the page seen, the frequency of visits, and the time between the first and final visit, are all thought to represent user interest. According to the above viewpoint, the interest page may be assessed as follows:

$$Interest_i = RL_i' \cdot (1 + C_F \cdot Fre_i' + C_R \cdot VT_i') \quad (1)$$

Where $1 \le i \le m$, m denotes the sum of pages visited by visitors, and m is the total number of visitors.

$$Fre' = \frac{Fre_i}{\max_{1 \le j \le m} \langle Fre_j \rangle} \quad (2)$$

$$RL' = \frac{RL_i \big/ \langle CT.Bytes_i \rangle}{\max_{1 \le j \le m} \langle RL_j \big/ (CT.Bytes_j) \rangle} \quad (3)$$

$$VT_i' = \frac{VT_i}{VT_{\log}} = \frac{LastDate_i - Firstdate_i}{LogEndDate - LogStartDate} \quad (4)$$

Where RL is the total amount of time spent on the website by the user throughout the given access log period, byte denotes the size of the website page, and Fre' denotes the number of visits. For HTML files, CF and CR are weight coefficients of 1 and CT coefficients of 1.

2.2 Dynamic Mining

1) Browse Time: The user's choice for alluring locales on which they spend more time is uncovered by Browse Time. There may, in any case, be a brief switch to another page. In this manner, the page estimate is influenced by pageview time.

$$BrowTime(i) = \frac{TotalTime_{(i)}}{Averange_{i \varepsilon tabu_k} TotalTime_{(i)}} \quad (5)$$

where $TotalTime_{(i)}$ is the total amount of time the user spends on page i, and The average amount of time a person spends on a page i is called average.

2) The strategy utilized: Clients who are interested in enrolling with web destinations should, for the most part, spend a lot of time on web pages and use the HTTP POST method. They will not make advantage of the POST technique if they have no desire to register on websites.

A method POST and GET can be established by:

If $MU(i) = POST$ (6)

then $MU(i) = 1$

else $MU(i) = 0$

Where $MU(i)$ is the method used for page i.

3) Page Visited: Page visits represent the average number of pages accessed on a location during a session. The page's escalation will be much larger if the client is interested in or needs to visit the page. The formulas below can be used to compute it[1].

$$Freq\ (i) = \frac{visit(i)}{\sum_{t \epsilon tabu_k} visit(i)} \quad (7)$$

Where visit (i) represents a collection of all pages that have been viewed (i), and $t \epsilon tabu_k$ represents a collection of all pages that have been visited.

4) *Interest*: Three factors influence the level of interest : browse time, the strategy utilized, and recurrence. We create intrigued on each page by using:

$$Interest(i) = a.BrowTime(i) + b.MU(i) + c.Freq(i) \quad (8)$$

Where $Interest(i)$ is the weight of interest on page *i*. a, b and c represents the user's interest to $BrowTime$, MU and $Freq$. The total of $a + b + c = 1$. Therefore, this research used a $= 0{,}33$, $b = 0{,}33$ and $c = 0{,}33$. The choice of weight values for parameters a, b, and c is 0.33 because of the consideration to divide the three parameters *BrowTime, MU*, and *Freq*, equally.

2.3 Pattern Recognition

Within the pattern recognition, learning calculations as the result of the pre-processing organization was connected to mine for possibilities within the pattern finding in pattern recognition. The process of classifying data into one of a few predefined groups is known as classification. One who is included in the creation of a user profile belongs to a certain course or group inside the online domain. In this stage, three classification calculations, to be specific Naïve Bayes, Decision Tree, and SVM, were compared.

1) Naive Bayes is simple to build and can be used for massive data sets without the need for complicated iterative parameter estimation techniques. Although it may not be the ideal classifier for every circumstance, it is usually trustworthy and successful.

2) A Decision Tree may be a more effective calculation for the purpose of examining the link between free factors and subordinate factors due to the tree looking pattern.

3) SVM is a data classification method that includes a pre-processing phase. The benefit of this technique is that it properly categorizes the data into two groups while also making a minor generalization error.

3. Result And Discussion

The dataset used was a web log file from the komputermurahjogja.com which has moved to ascomputer.co.id e-commerce website. The number of weblog files included in this analysis was 91456 log files, created by 243 user. The data collected has complete information, consisting of Number of Session Identification, Original Referrer, referrer page, current page, mouse event, users IP, user agent, and browser. Table 1 shows an example of the log file format that was used.

Table 1. Example of Web Log File

Field	Meaning
1410926140537	Number of Session Identification
http://komputermurahjogja.com/peripherals/mainboard/	Original Referrer
http://komputermurahjogja.com/peripherals/mainboard/index.php	The next URL (referrer)
http://komputermurahjogja.com/?term=&s=lga+775&post_type=product&taxonomy=product_cat 18	The current URL
mousemove	An event involving the mouse
s=lga+775	User-defined search
GET	The method of the request (method)
114.141.xx.xx	Users' ip proxy
108.162.xxx.xx	Users' ip client
Mozilla/5.0 (Windows NT 6.1; rv:33.0) Gecko/20100101 Firefox/33.0	Users' agent
Windows Firefox	Browser and OS
2020-09-16 20:50:58	The request's date and time

3.1 Data Pre-processing

The pre-processing phase of the Dynamic Mining system was carried out in several stages, namely :

a. Cleaning of information: Data cleansing removes the web crawler robot. Examples of web crawler robots are Googlebot, Bingbot, or Soguo.
b. Identification of the user. Integration of IP addresses and additional information, such as user agents or referral websites, can identify users and their sessions correctly.

c. Browsing time. It indicates the user's preferences for a certain page of interest on which the user spends more time.
d. Method used. If a person is interested in registering on the website, they spend a lot of time on the website and use the HTTP POST method.
e. Page visited: pages/visits that indicate the average number of pages visited on the web during one session.

In comparison to the Dynamic Mining method, the pre-processing phase of the Page Interest Estimation consists of four stages, namely:

a. Data cleaning. The data cleaning process is the same as the data cleaning process in the Dynamic Mining method
b. Page Identification. Identification of what pages visited by users
c. Page size: measure the size of each page accessed by the user
d. Identify the session id

3.2 Application of Dynamic Mining and Page Interest Estimation Methods

The pre-processing phase in Dynamic Mining generated user activity, as presented in Tables 2 and 3.

Table 2. User Activity of Dynamic Mining

User Interest Indicator					
User id	***Session id***	***Browsing Time (mins)***	***Method Used (POST/ GET)***	***No of pages visited***	***Interest***
U1	1411461871945	199	GET	13	69
U2	1411462065063	112	GET	9	39
U3	1411462149381	1007	GET	37	344
U4	1411462250414	17	GET	11	22
U5	1411462254528	24	GET	20	2
…..	…..	….	….	…..	
U109	1411463297310	988	GET	34	887
…..	……	…..	…..	…..	….
U243	1411725486155	135	GET	38	1494

Table 3. User Activity of Page Interest Estimation

The user: 105.23.203.132 (Mozilla/5.0 (Windows NT 5.1; rv:33.0) Gecko/20100101 Firefox/33.0)				
Page	**Interest**	**RL**	**VT**	**Fre**
Lenovo	2.0023640	1.000	0.000	1
Asus	0.1490880	0.065	0.000	1
Samsung	0.1890000	0.035	0.000	1

In Table 2, the formation of user activity is influenced by the browsing time and the number of pages visited by each visitor. In contrast to the case in Table 3, in the formation of user activity, the influencing factors are the frequency, the time of the visit, which is recapitulated from the first visit to the end in one session, as well as the time taken spent by each visitor on each visit. In Table 4, descriptive statistics of the dataset used.

Table 4. Descriptive Statistics of Dataset

Method		**Min**	**Max**	**Avg**	**Std**
Page Interest Estimation	Freq	0.2	1	0.924	0.190
	RL	0.002	1.342	0.631	0.421
	VT	0	1.101	0.099	0.221
	Interest	0.004	2.84	1.266	0.853
Dynamic Mining	BrowTime	0	7890	1366	1556
	Number of visited	1	897	33.97	82.024
	Method used	0	0	0	0
	Interest	0	2625	461.67	523.21

3.3 Comparison Dynamic Mining and Page Interest Estimation

In analyzing pattern discovery, the researcher used the Rapid Miner application to process data with the algorithms that had been described. The cross-validation technique (k-fold = 10) used by the process can evaluate the performance of SVM, Decision Tree, and Naive Bayes. The results of this stage are presented in the table below. All datasets totaling 243 rows were processed using k-fold = 10, and the average accuracy, precision, and recall using k-fold = 10 folds were calculated. The results are presented in tables 5 and 6.

Table 5. Accuracy Result of Three Algorithms Using Dynamic Mining Approach

Classifier	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Naïve Bayes	95.43 %	90.27%	100%	0.94
Decision Tree	100%	100%	100%	1
SVM	96.68%	93.03%	100%	0.94

Table 6. Accuracy Result of Three Algorithms Using Page Interest Estimation Approach

Classifier	Accuracy	Precision	Recall	F1 Score
Naïve Bayes	99.64 %	100%	99.38%	0.99
Decision Tree	100%	100%	100%	1
SVM	98.60%	97.77%	100%	0.97

Dynamic Mining and Page Interest Estimation methods have quite different parameters for determining user interest. The Dynamic Mining method identifies a user by session-id, browsing time, the technique used, and the number of pages visited. The analysis result found that the amount of time and the number of visits greatly affected the determination of user interest. However, the parameter method used did not influence the decision of user interest because the data used in this study had the same method used, namely the GET method. The findings could be different since different approaches to the data were used.

Other parameters were found relevant to pages viewed by users, which needed to be further investigated to improve the accuracy of the algorithm in assessing user interest. In contrast to the PIE method, the parameters used to determine the user interest were more varied, not only in terms of the pages visited but also the frequency of visits and the time spent by the user visiting the page. Determining user interest in PIE was focused on the identity of the pages visited by the user. In addition to the frequency of page visits, the size of each page visited was also one of the parameters in determining user interest.

Both approaches used different parameters for gauging user interest, but the decision tree algorithm in the Pattern discovery phase had the same accuracy value. Tables 5 show that of the three algorithms utilized, namely nave Bayes, decision tree, and SVM, the Naive Bayes method has a lower level of accuracy than the other two. In comparison, the decision tree approach has a greater level of accuracy. Unlike in Table 6, the accuracy of the Naive Bayes method is higher than that of the SVM algorithm. However, suppose you look at tables 4 and 5. In that case, the decision tree algorithm has the highest accuracy in the Dynamic Mining and Page Interest Estimation approach, reaching 100% in the determination of user interest. It can also be seen if the decision tree algorithm has a better performance in determining user interest in the Dynamic Mining or Page Interest Estimation approach. Several variables impact the decision tree algorithm's performance, including that a decision tree requires less data preparation during preprocessing and does not require data normalization. Creating a decision tree is unaffected by missing values in the data. The lower level of accuracy in the SVM & Naïve Bayes was influenced by the amount of data that is not much, as it is known if SVM and Naïve Bayes are an algorithm that can produce a high level of accuracy if the dataset used is in large quantities while in this study after going through the data preprocessing process, the data user formed is 243.

4. Conclusion

The rapid development of e-commerce is a challenge for business people. Business people should be able to gain customer interest in a variety of ways so that their companies can compete with others. Analyzing user behavior when dealing with e-commerce will help business people improve their competitive advantage. Businesses and corporations may utilize clickstream data to analyze client loyalty and promotion effectiveness, as well as build marketing campaigns by detecting user interests. In this paper, experiments related to user interest detection using Dynamic Mining and Page Interest Prediction approaches have been performed. Dynamic Mining and Page Interest Estimation approaches are considered to have quite different criteria for evaluating user interest. The Dynamic Mining method identifies a user with the session id, browsing time, method used, and the number of pages visited. The research result shows that the Dynamic Mining method, the amount of time, and the number of visits greatly affect the determination of user interest. However, the parameter method used does not influence the determination of user interest because the data used in this study have the same method used, namely the GET method. The findings might be different if there were different approaches applied. By using three algorithms at the pattern discovery stage, it is established that the decision tree method excels in all methods. It shows that the Decision Tree algorithm performs satisfactorily in evaluating consumer interest with two different approaches.

Finally, this approach is suitable for use to find user interest in e-commerce in another case study. However, different performances can occur because of differences in several parameters such as time, the number of visits and frequency of visits.

References

[1] A. V. Bharathi, J. M. Rao, and A. K. Tripathy, "Click Stream Analysis in E-Commerce Websites-a Framework," presented at the Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018, 2018. doi: 10.1109/ICCUBEA.2018.8697475.

[2] Q. Su and L. Chen, "A method for discovering clusters of e-commerce interest patterns using click-stream data," *Electronic Commerce Research and Applications*, vol. 14, no. 1, pp. 1–13, 2015, doi: 10.1016/j.elerap.2014.10.002.

[3] S. Diwandari, A. E. Permanasari, and I. Hidayah,

"Research methodology for analysis of E-commerce user activity based on user interest using web usage mining," *Journal of ICT Research and Applications*, vol. 12, no. 1, pp. 54–69, 2018, doi: 10.5614/itbj.ict.res.appl.2018.12.1.4.

[4] M. Khosravi and M. J. Tarokh, "Dynamic mining of users interest navigation patterns using naive Bayesian method," 2010, pp. 119–122. doi: 10.1109/ICCP.2010.5606453.

[5] J. Bhavithra and A. Saradha, "Personalized web page recommendation using case-based clustering and weighted association rule mining," *Cluster Computing*, vol. 22, pp. 6991–7002, 2019, doi: 10.1007/s10586-018-2053-y.

[6] S. Diwandari and A. T. Hidayat, "Comparison of Classification Performance Based on Dynamic Mining of User Interest Navigation Pattern in e-Commerce Websites," *Journal of Physics: Conference Series*, vol. 1844, no. 1, p. 012025, Mar. 2021, doi: 10.1088/1742-6596/1844/1/012025.

[7] S. Cleger-Tamayo, J. M. Fernández-Luna, and J. F. Huete, "Top-N news recommendations in digital newspapers," *Knowledge-Based Systems*, vol. 27, pp. 180–189, 2012, doi: 10.1016/j.knosys.2011.11.017.

[8] Y. Li, B. Liu, and C. Wang, "Study of the Evolution of Online User Interest Behavior," 2019, pp. 166–171. doi: 10.1109/CIS.2019.00043.

[9] L. Zheng, S. Cui, D. Yue, and X. Zhao, "User interest modeling based on browsing behavior," 2010, vol. 5, pp. V5455–V5458. doi: 10.1109/ICACTE.2010.5579511.

[10] H. Zaim, A. Haddi, and M. Ramdani, "A novel approach to dynamic profiling of E-customers considering clickstream data and online reviews," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 1, pp. 602–612, 2019, doi: 10.11591/ijece.v9i1.pp.602-612.

[11] X. Wei, Y. Wang, Z. Li, T. Zou, and G. Yang, "Mining Users Interest Navigation Patterns Using Improved Ant Colony Optimization," *Intelligent Automation and Soft Computing*, vol. 21, no. 3, pp. 445–454, 2015, doi: 10.1080/10798587.2015.1015778.

[12] Y. Li and B.-Q. Feng, "Page interest estimation model considering user interest drift," 2009, pp. 1893–1896. doi: 10.1109/ICCSE.2009.5228238.