Plagiarism Checker X Originality Report

Similarity Found: 6%

Abstract. The use of e-commerce in companies or other types of business has supported them to develop and correspondingly cope with business pressures of high levels of competition.

More consumer information can be gathered based on the interactive nature of e-commerce technology. In an e-commerce competition, all information relating to consumer behavior, such as the knowledge of the visitor interests in a product marketed by e-commerce, is of value to e-commerce players. Users can use the Web Usage Mining techniques to explore these interests. This study aimed to compare three classification algorithms by using the dynamic mining approach of user interest navigation pattern.

The results of the study showed that the Decision Tree Classifier performed optimally in both the unbalanced data and independent or dependent data models.
1. Introduction E-commerce has been developing rapidly, along with the development of websites. Its rapid growth has led the company and its customers to new situations. However, while it is a struggle for a business to succeed with the further competition, it allows customers to choose between many products offered by different companies before finally selecting which product meets their needs.

The success of any online company depends on the potential to attract visitors. In e-commerce competition, all information relating to consumer behavior is of value to

e-commerce players. E-commerce websites provide customers with various navigation options and actions that enable the users to browse freely through a variety of product categories, follow a selection of navigation paths to visit a particular product, or use different techniques to purchase a product [1]. Therefore, a company must monitor customer behavior data through clickstream data. Clickstream data is the primary information for the company to adapt its services to its customers [2].

Clickstream data includes a variety of values that provide detailed information about the users of the website. The completion of this clickstream data collection helps the company to identify customer loyalty, encourage promotion effectiveness, and improve marketing strategies by recognizing consumer preferences. User interest is defined based on the document content read by the user [3].

Another study has found that the indicators of user interest are the overall time of access, the most regularly visited website page, or the most recently visited website page [4]. Several experts who analyzed user interests state that the user interest is frequently reflected in the page they visited and the posts they replied to [5]. In addition to the knowledge of user interest indicators, a pattern discovery with the correct algorithm is required to create a more accurate 2020 2nd International Conference on Science & Technology (2020 2nd ICoST) Journal of Physics: Conference Series 1844 (2021) 012025 IOP Publishing doi:10.1088/1742-6596/1844/1/012025 2 model of consumer behavior. Several methods, such as KNN [6], [7], Naïve Bayes [8], Decision Tree [9], SMO [10], have been used.

In contrast to previous studied, this paper aims at comparing the classification method based on dynamic mining of user interest using several indicators such as user ID, page browse time, the method used, and the page visited. Three classification algorithms are compared to find the patterns. Section II of this paper provides a review of the literature relevant to this paper. The approach to discovering the preferred navigation paths of the user is discussed in Section III, the findings of the experimental evaluation in Section IV, and the conclusion of the analysis in Section V. 2.

Related Works The advancement of the e-commerce market is accompanied by large-scale data growth and web data mining applications to investigate consumer desires and user interests. Several research studies have used consumer purchase patterns, web page visit numbers, and online browsing paths to create models designed to forecast customer preferences. Several user attributes, such as product rating [11],[12], purchase records, page series addressed, page length, and page browsing frequency, are analyzed to measure user preferences. Zheng et al.

[13] carried out a detailed study of user browse time for user preferences but did not consider the frequency and the order of visits. It's distinct from the analysis of Yong Li et al. [14] who used the timing and order of the visits. They introduced an algorithm to measure interest and show the efficiency and consistency of the algorithm. Kim et al. [15] presented a more systematic appraisal approach that defined the user interests based on user purchasing decisions and time spent on each page. Qingqing and Zhang et al.

[16] introduced another indicator to measure user interest, namely the combination of time bytes, weight time, and browse rates. In recent years, clickstream-related research has become more common in the field of web data analysis. Zaim et al. [17] used click-flow data and online consumer feedback to determine the user interest in addition to the value extracted from the website features. Lakshmi [18] used click-flow to propose a new paradigm for forecasting user navigation behaviour using a mathematical framework. It was found that the proposed approach using performance indicators was better than the conventional method.

In the meantime, several other studies use click-flow, such as user browsing behaviour, user reaction to website design, and how users travel through sites [5]. Pattern exploration is an indistinguishable part of click-flow statistical analysis. Several techniques used in clickstream data processing include the Bayesian method, collaborative filtering, fuzzy clustering, stochastic regression, fuzzy leader clustering, model-based bi-clustering, graph clustering, decision tree, association rule mining, K-Means clustering, and leader clustering [19]. Several experiments were performed using the Naive Bayes method [8] and the decision tree [9]. 3.

Background Web usage mining (WUM) is an exploration of useful weblog data patterns for a deeper comprehension of web users [2]. The WUM method consists of four main stages, namely pre- processing, influential factors on dynamic mining of the user interest, pattern discovery, and pattern analysis. 3.1. Data Pre-processing In the pre-processing stage, data are collected and then cleaned to remove unrelated objects such as graphic and multimedia entries.

There are three data sources to obtain a log row of data that can be used in web-based mining research, namely client log files, Web Proxy Server Log, and Log Files [20]. The data source applied in the experiment was the client log files derived from the e-commerce website komputermurahjogja.com. Visitor activities are recorded on the client-side of the client log files. The following are the stages of pre-processing performed in this experiment. _ Data cleaning. In the data-cleaning process, the web crawler robot, characterized by the absence of mouse scrolling and mouse movements,

is removed from the data log row.

3 _ Users Identification. A combination of IP addresses and other information, such as user agents and referrers, can be used to identify specific users [21] _ Session. A session is a series of pages viewed by the users during a specific visit [22]. For logs that span a long time, different users will likely use the same computer to access the server's web site _ Page-visit time. It is defined as a time difference between consecutive page requests, which is computed for each page.

For the last page of the user session, the time of the page is the mean of the page visit times for the page taken throughout all the sessions when a new IP address is encountered, or when the time of the visit exceeds 30 minutes for the same IP address [23]. 3.2. Influential Factor on Dynamic Mining of User Interest _ Browse Time. Browse Time reveals the user preference for desirable sites on which he/she spends more time. However, there may be a quick jump to another page. Thus, the size of the page can affect the page-visit time due to the short length of the web page.

The following is the Browse Time formula [24]: = (1) where is the total time the user accesses page I and Average is the average time the user accesses page . _ The method used: Generally, users spend much time on web pages and use the HTTP POST mode if they are interested in registering with web sites. However, they will not use the POST method if they are not interested in registering on websites [1]. A possible method of POST and GET can be obtained by: If (2) then else Where is the method used for page . _ Page visited: page visit indicates the average number of pages accessed on a site within a session.

If the user is interested or needs to visit the page, the intensity of the page will be even higher. It can be computed using the following [2]: (3) Where visit (i) shows many visits on page (i), and shows a collection of all pages that have been visited. _ Interest. Interest is built based on three factors: browse time, the method used, and frequency. We develop interest on every page by using: (4) Where is the user interest on page i. a, b and c are the weight of interest for , and . The sum of . This research used a , and .

4 3.4.

Pattern Analysis The last stage is pattern analysis, which determines the unattractive rules or patterns found at the discovery stage to be removed. 4. Methodology The data

collected included 91456 logs from komputermurahjogja.com, created by 243 users. Table 1 presents client log files format used. Table 1. Example of the client log files Field Meaning 1410926140537 Session Identification http://komputermurahjogja.com/peripherals/mainb oard The URL of the first request (original referrer) http://komputermurahjogja.com/peripherals/mainb oard/index.php The URL requested just before (referrer) http://komputermurahjogja.com/?term=&s=lga+77 5&post_type=product&taxonomy=product_cat The URL of current request Mousemove Mouse event s=lga+775 Item search by user GET The method of the request (method) 114.141.57.76 User• s IP proxy 108.162.208.32 User• s IP client Mozilla/5.0 (Windows NT 6.1; rv:33.0) Gecko/20100101 User• s agent Windows Firefox Browser and OS 2014-09-16 22:55:50 The date and time of the request The table shows the log data collected by recording visitor activities.

The log data retrieved were visitor identification, session identification, referrer, accessed pages, user event, date, time, visitor's IP address, and user agent. Each visitor had a different visitor ID, which remained unchanged when the visitor was browsing different web pages. Referrer was the web pages that the visitor previously accessed. User events recorded visitor behavior while browsing. Some examples of visitor behavior were loading, unloading, focusing, moving, and scrolling the mouse. The date and time represented a series of processes by which a visitor accessed a web page.

A visitor's IP address was the visitor's In the pattern discovery stage, learning algorithms as the output of the pre-processing stage are applied to mine for potential. Classification is the process of mapping a data item into one of several predefined classes. One involved in creating a profile of users belongs to a specific class or group in the web domain. In this study, three classification algorithms, namely Naïve Bayes, Decision Tree, and SVM are compared.

? Naïve Bayes is easy to construct without requiring any sophisticated iterative parameter estimation schemes and can be conveniently applied to large data sets. It may not be the best classifier available for all applications, but it can normally be counted on to be robust and to perform relatively well [8]. ? Decision Tree is a more efficient algorithm for analyzing the relationship between independent variables and dependent variables due to the tree searching schema [25]. ? SVM is a data classification technique with a training process.

The advantage of this method is that it correctly classifies the data into two classes and makes a small generalization error [10]. 3.3. Pattern Discovery 2020 2nd International Conference on Science & Technology (2020 2nd ICoST) Journal of Physics: Conference

computer address used to access the web page. User agents were the browsers that visitors used to access web pages [10]. The data from komputermurahjogja.com customer log files were then extracted, as illustrated in figure 1. Figure 1.

Classification Evaluation stages Data collection, feature selection, and data filtering were selected at the pre-processing stage. At the data collection stage, the data from each user was defined by session ID, user's IP client, the method used, date/time of the request, and the requested URL. The result of the pre-processing stage was a user session file. Interest was obtained by using equation (4). In new data, each of these activities became the dimension of each user. The data interest is presented in table 2. Table 2.

Example of interest data Indicator of user interest User ID Session ID Browse Time (mins) Method Used (POST/GET) Freq Interest U1 1411461871945 199 GET 13 69 U2 1411462065063 112 GET 9 39 U3 1411462149381 1007 GET 37 344 U4 1411462250414 57 GET 12 22 U5 1411462254528 3 GET 4 2 .... U109 1411463297310 3456 GET 60 887 .... U243 1411725486155 4500 GET 30 1494 Web usage mining techniques were implemented to find information from these log data. The data collected from pre-processing had not been labeled, so that data labeling was required. User clustering tended to establish a group of users exhibiting similar browsing patterns.

Such knowledge is especially useful for inferring user demographics to define market segmentation in E-commerce [3]. The K- means method was selected in the labeling process. K-means is a clustering algorithm using a 2020 2nd International Conference on Science & Technology (2020 2nd ICoST) Journal of Physics: Conference Series 1844 (2021) 012025 IOP Publishing doi:10.1088/1742-6596/1844/1/012025 6 partitioning method to analyze data that involves grouping data by partitioning. This algorithm will transform the object group o into the c position (c <= o).

The goal of the labeling process was to have two classes, namely potential user and casual user. Table 3 presents the result of the labeling process. Table 3. Labeling data with K-Means Cluster Items Cluster 0 146 Cluster 1 97 The table shows that 146 data are classified in cluster 0 and 97 data in cluster 1. It was found that there was data imbalance, proven by more data from casual users than potential users. The relationship between each attribute and the centroid point of each cluster is presented in figure 2.

Figure 2 shows that cluster 0 has characteristics with an Interest value of no more than 250, while cluster 1 is higher than cluster 0 such that cluster 0 is classified as casual users and cluster 1 as potential users. Figure 2. The relationship between the centroid points

of each cluster The characteristics of casual users and potential users are presented in table 4. Table 4. Example of data cluster Indicator of user interest User ID Session ID Browse Time POST/GET Freq Interest Class U1 141?.945 199 GET 13 69 Casual user U2 141?.063 112 GET 9 39 Casual user U3 141?.381 1007 GET 37 344 Casual user U4 141?414 57 GET 12 22 Casual user U5 141?528 3 GET 4 2 Casual user ?.. ?.. ?. ?. ?.. U149 141?310 3456 GET 60 887 Potential user ?.. ?? ?.. ?.. ?..

5. Result Table 5 describes the results of the comparison process of Naïve Bayes, Decision Tree, and SVM with k-fold cross-validation. It can be seen that the Naïve Bayes algorithm has the lowest accuracy value of 94.43%, and the Decision Tree has the highest accuracy value of 100%. Table 5 is the confusion matrix of the three algorithms.

It shows in the Naïve Bayes algorithm that 11 out of the 146 casual user classes are classified as false negative because they are considered to be a potential user class. Thus, the potential class data becomes 108, namely 97 true-positive data and 11 false-negative data. In this case, the Naïve Bayes algorithm, which has a characteristic of a very strong assumption of the independence of each condition or circumstance, reduces accuracy.

It is due to the correlation between one variable and the other but ignored by the Naïve Bayes algorithm. The advantage of Naïve Bayes is that it only requires a small number of training data to determine parameter estimation in the classification process [25] to precisely determine the potential class marked with several 0 false positives. It is different from the ==Decision Tree (DT) algorithm== that can classify data on an independent or dependent model so that it can perform classification optimally.

Another advantage of the DT classifier, in this case, ==is that it can== work on an imbalanced dataset, proven by the quantity of data on casual users that is more than potential users. Table 5. Final Result Classifier Accuracy Precision Recall Naïve Bayes 95.43% 90.27% 100% Decision 100% 100% 100% SVM 96.68% 93.03% 100% Table 6. Confusion Matrix of Classifier Actual Class Casual User Potential User Naïve Bayes Predicted Class Casual User 135 0 Potential User 11 97 Decision Tree Predicted Class Casual User 146 0 Potential User 0 97 SVM Predicted Class Casual User 138 0 Potential User 8 97 6. Conclusion This study aims to examine the performance of Classification algorithms by comparing the three types of classification.

To summarize, we have successfully created a DT Classifier that can classify user interest

with 100% accuracy, higher when compared to Naïve Bayes and SVM. This study also shows that the DT classifier approach can work optimally in imbalanced data case. The system may be further improved by increasing the number of features analyzed. The further work is required to handle imbalanced data and the effect of imbalanced data on the accuracy of the classification case. 7. References [1] S. Hernandez, P. Alvarez, J. Fabra, and J. Ezpeleta, "Analysis of Users' Behavior in Structured e- Commerce Websites," IEEE Access, vol. 5, pp. 11941–11958, 2017, doi: 10.1109/ACCESS.2017.2707600.

2020 2nd International Conference on Science & Technology (2020 2nd ICoST) Journal of Physics: Conference Series 1844 (2021) 012025 IOP Publishing doi:10.1088/1742-6596/1844/1/012025 8 [2] 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. [3] "User Modeling and User-Adapted Interaction: Editorial note," User Modelling and User-Adapted Interaction, vol. 15, no. 5, p. 507, 2005, doi: 10.1007/s11257-005-5133-7. [4] R. A. Gotardo, C. A. C. Teixeira, and S. D.

Zorzo, "An approach to recommender system applying usage mining to predict users' interests," 2008, pp. 113–116, doi: 10.1109/IWSSIP.2008.4604380. [5] 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. [6] D. A. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method," Applied Computing and Informatics, vol. 12, no. 1, pp. 90–108, 2016, doi: 10.1016/j.aci.2014.10.001. [7] V.

Subramaniyaswamy and R. Logesh, "Adaptive KNN based Recommender System through Mining of User Preferences," Wireless Personal Communications, vol. 97, no. 2, pp. 2229– 2247, 2017, doi: 10.1007/s11277-017-4605-5. [8] 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. [9] Y. H. Cho, J. K. Kim, and S. H. Kim, "A personalized recommender system based on web usage mining and decision tree induction," Expert Systems with Applications, vol. 23, no. 3, pp. 329–342, 2002, doi: 10.1016/S0957-4174(02)00052-0. [10] 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. [11] X. Zhao, Z. Niu, and W. Chen, "Interest before liking: Two-step recommendation approaches,"

Knowledge-Based Systems, vol. 48, pp. 46–56, 2013, doi: 10.1016/j.knosys.2013.04.009. [12] 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. [13] 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. [14] 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. [15] Y. S. Kim and B.-J. Yum, "Recommender system based on click stream data using association rule mining," Expert Systems with Applications, vol. 38, no. 10, pp. 13320–13327, 2011, doi: 10.1016/j.eswa.2011.04.154. [16] Q. Zhang, X. Wu, and T. Chen, "Research on adaptive recommendation algorithm emerging in user interest in an electronic commerce environment," International Journal of Computing Science and Mathematics, vol. 6, no. 5, pp. 425–433, 2015, doi: 10.1504/IJCSM.2015.072964. [17] 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. [18] P. D. Lakshmi, "A new model for prediction of user behavior based on large click stream data," Journal of Advanced Research in Dynamical and Control Systems, vol. 11, no. 5, pp. 374– 378, 2019. [19] M. Volk, A. E. Shareef, N. Jamous, and K. Turowski, "New E-Commerce User Interest Patterns," 2017, pp. 406–413, doi: 10.1109/BigDataCongress.2017.60. [20] S. Diwandari, A. E. Permanasari, and I.

Hidayah, "Performance analysis of Naïve Bayes, PART and SMO for classification of page interest in web usage mining," 2015, pp. 39–43, doi: 10.1109/ISITIA.2015.7219950. 2020 2nd International Conference on Science & Technology (2020 2nd ICoST) Journal of Physics: Conference Series 1844 (2021) 012025 IOP Publishing doi:10.1088/1742-6596/1844/1/012025 9 [21] S. D. Bernhard, C. K. Leung, V. J. Reimer, and J. Westlake, "Clickstream prediction using sequential stream mining techniques with markov chains," 2016, vol. 11-13-July-2016, pp. 24–33, doi: 10.1145/2938503.2938535. [22] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.

Berlin, Heidelberg: Springer-Verlag, 2009. [23] H. Ling, Y. Liu, and S. Yang, "An ant colony model for dynamic mining of users interest navigation patterns," 2008, pp. 281–283, doi: 10.1109/ICCA.2007.4376363. [24] 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. [25] C. F. Lin, Y.-C. Yeh, Y. H. Hung, and R. I. Chang,

"Data mining for providing a personalized learning path in creativity: An application of decision trees," Computers and Education, vol. 68, pp. 199–210, 2013, doi: 10.1016/j.compedu.2013.05.009. [26] Mehak, M. Kumar, and N.

Aggarwal, "Web usage mining: An analysis," Journal of Emerging Technologies in Web Intelligence, vol. 5, no. 3, pp. 240–246, 2013, doi: 10.4304/jetwi.5.3.240-246

INTERNET SOURCES:

1% - http://cscns2020.org/
<1% - http://www.ijstr.org/research-paper-publishing.php?month=oct2019
<1% - https://www.bizjournals.com/washington/news/2017/10/19/meet-the-90sera-businesses -on-our-fastest-growing.html
<1% - https://www.sciencedirect.com/science/article/pii/S0191261517304484
<1% - http://lyle.smu.edu/~mhd/8331f04/barsagada.doc
<1% - https://www.thefreelibrary.com/Influence+of+ratio+of+auxiliary+pages+on+the+pre-pr ocessing+phase+of...-a0428275920
<1% - https://www.prepressure.com/pdf/basics/page-boxes
<1% - https://www.keycloak.org/docs/4.8/server_development/
<1% - https://dl.acm.org/doi/proceedings/10.1145/3438872
<1% - https://www.naun.org/main/UPress/cc/2014/a022012-090.pdf
<1% - https://www.academia.edu/48068780/Introduction_and_Trends_to_Fuzzy_Logic_and_Fuz zy_Databases
<1% - https://onlineengineering.wordpress.com/2011/11/02/data-warehousing-data-mining-n otes-unit-iv/
<1% - https://jurnal.uns.ac.id/ijsascs/
<1% - https://www.sciencedirect.com/science/article/pii/S0957417405000588
<1% - https://content.iospress.com/articles/multiagent-and-grid-systems/mgs210343
<1% - https://link.springer.com/article/10.1007/s10489-020-01746-9
<1% - http://dns2.asia.edu.tw/~ysho/Trend%20Research%20Centre/Bibliometric%20Reference s/I-J.doc
<1% - https://www.sciencedirect.com/science/article/pii/S0098300415000722
<1% - https://www.sciencedirect.com/science/article/pii/S0012825220302713
<1% - https://www.sciencedirect.com/science/article/pii/S0306457309000259
<1% - https://link.springer.com/article/10.1007/s00607-021-00908-y

<1% - http://aktripathy.in/researchpapers.html

<1% - https://dl.acm.org/doi/abs/10.1016/j.elerap.2014.10.002

<1% - https://www.sciencedirect.com/science/article/pii/S0168169918311578

1% - https://dl.acm.org/doi/10.1504/IJCSM.2015.072964

1% - https://www.academia.edu/45631627/Estimation_of_posture_and_prediction_of_the_elderly_getting_out_of_bed_using_a_body_pressure_sensor

<1% - https://pt.scribd.com/document/109289143/jetwi0403