IJITEB
Vol. 3
No. 2
2021

International Journal of Information Technology and Business

http://ejournal.uksw.edu/ijiteb

cANT WUM : Web User Classification using Ant Colony Optimization Algorithm

Abdurrahman[a,*], Riyanto, B.T[b], Govindaraju, R[c], Mandala, R.[b]

[a] STMIK BANDUNG, Jl. 1 STMIK BANDUNG, Jl. Cikutra 113, Bandung, Indonesia

[b] School of Electrical Engineering & Informatics, ITB, Jl. Ganesha 10 Bandung, Indonesia Ganesha 10 Bandung, Indonesia

[c]Faculty of Industrial Technology, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung, Indonesia

Keywords: web usage mining, classification of web users, ant colony optimization, Ant-Miner algorithm, cAnt-Miner algorithm, heuristic functions, preprocessing, accuracy of rules,

Abstract : Web Usage Mining (WUM) is the use of data mining methods to extract knowledge from web usage data. One function of WUM is to support Business Intelligence (BI) purpose in which one of the important information needed is the classification of web users that can be used for acquisition, penetration, and user retention activity. There are two main problems encountered in conducting the classification of web users. The first is the determination of antecedent attributes as a term of classification rules, which is a major problem in data mining classification function in general. The second problem is the preprocessing activity which involves preparing the supporting data for the web users' classification need which is the most difficult stage in WUM.

For the web user classification method, we propose a classification method based on ant colony optimization method (ACO) as a distributed intelligent system using heuristic function which is in line with the problem areas. We proposed a heuristic functions for web user classification based on web usage data that uses entropy of antecedent candidate, information gain from attribute of total number of web user access and average of access duration of web user. For preprocessing purpose, a method of data preparation that can support the needs of web users' classification is proposed. The data used consists of web access log data, web user profile data and web transaction data. The preprocessing activity consists of parsing, data cleansing, and extraction of the web user sessions using heuristic method concerning web page access timeout and differences in web browser agent.

Testing is done by comparing the performance of the proposed algorithm with Ant-Miner algorithm, cAnt-Miner algorithm, and the Continuous Ant-Miner algorithm. The results of testing of four web data shows that the performance of the proposed algorithm is better in terms of accuracy of rules and simplification of rules.

Keywords: web usage mining, classification of web users, ant colony optimization, Ant-Miner algorithm, cAnt-Miner algorithm, heuristic functions, preprocessing, accuracy of rules, simplification of rules.

*Corresponding authors
e-mail addresses : mr.indonesian@gmail.com , briyanto@lskk.ee.itb.ac.id, rajesri_g@mail.ti.itb.ac.id, rila@informatika.org

1. Introduction

Web user interaction (web users) to access the web produces a huge amount of web access data for a certain time period which is stored in the web access log file on the server. This data is expected to provide valuable information for the web management in order to market its web presence and sell products. In this context, web usage mining (WUM) has a role in discovering knowledge (knowledge discovery) from the web usage data. In addition to web access log data, data composed from user interaction with e-commerce website is a user profile data and transaction data.

2. Literature Review

Business intelligence (BI) is one of WUM function to support the marketing activities of the web presence and the sold products (Abraham, 2003). To support the BI function in WUM, web usage patterns information is required. Web usage patterns can be analyzed from the access period aspect which is how long users interact with the web, frequency which is how often users access the web, and intensity aspect which is the total amount of transactions conducted through the web user [3]. WUM is the use of techniques of data mining for knowledge discovery (knowledge discovery) from the web usage data.

One of the knowledge that can be discovered in WUM is the rules of web users' classification. To extract the rules for web users' classification a combination of web access log data, user profile data, and transactions data can be used. In WUM research domain would use combines the three types of data has been done and most of the research has used only web access log data.

Classification of web users is expected to support the marketing activities of web and its products that include new customer acquisition activities, customer retention, and penetration of the customer in terms of increasing the value of sales transactions [6]. In the domain of data mining, classification in general is an attempt to classify the data into a class or group based on certain criteria. Various methods have been developed to obtain classification rules with a high degree of accuracy, computational efficiency, and ease of understanding. One of the problems in the classification function is to determine the attributes as a potential antecedent in a certain rule of classification.

Ant-Miner algorithm which is based on Ant Colony Optimization (ACO) provides better performance in terms of level of accuracy and ease of understanding of the rules generated, compared to other classification methods (algorithms) in data mining such as C4.5 and CN.2 [6].

Ant-Miner algorithm uses the same heuristic function as used by the algorithm C4.5 decision tree which measures the entropy of a certain term as a candidate of antecedent [8]. The main difference between them is in the context of the use of heuristic functions. The decision tree calculates overall entropy of attributes, so that all the attributes are selected to develop the tree, whereas Ant-Miner algorithm calculates the entropy of just a single pair of attributes so that only a pair of attribute and its values is selected and used to develop the rules. In the decision tree algorithm the entropy is calculated at tree development phase, whereas in Ant-Miner the entropy calculation is also used to update the pheromone. This makes the development process of the rules is more robust [6].

Based on that an algorithm with a specific heuristic function is proposed which is called cAnt- WUM. cAnt-WUM algorithm is a method for web users' classification in web usage mining using discrete attributes and numerical (continuous). The Minimum Description Length (MDL) which was applied in cAnt-Miner and modification of pheromones using Probability Density Function (PDF) which was applied in the algorithm Continuous Ant-Miner are used to form discrete numerical attributes [2] [5] [7] [11].

The proposed algorithm can generate the classification rules with discrete input data and numerical attributes, using the minimum description length (MDL) for discretization numerical attributes and implement reform-based pheromone PDF (Probabilistic Density Function) [5][7][11].

Heuristic function developed in this research is the heuristics based on web usage which is very specific to the WUM problem. This function uses three heuristics, i.e. are term entropy heuristic of the rule antecedent candidates, heuristic of access number information gain and heuristic of average access time information gain. This function uses weight for the three heuristics, to provide a relative priority of one heuristic to other heuristics.

In addition, the study reported here also proposed a method of data preparation that supports the classification of web users in the WUM. The data preparation method uses a combination of web access log data, profile data and transaction data.

3. Research Method

A method of data preparation as a *preprocessing* step was developed in this study. This method developed to support WUM for the need of the classification of web users in the domain of *web usage mining* by making modifications heuristic function. The development of the proposed method is based on the following analysis:

1. One type of data used in WUM is *a web access log* data which is unstructured so that it requires a method to extract the information needed for web users classification [17]. In addition to that, to get web usage data that supports *business intelligence* needs it is required to map the web log data with transaction data and web user profile data.

2. In *data mining* in general, ACO algorithm has been proven to produce classification rules with a higher accuracy level and simpler rules, compared to other meta-heuristics algorithms [6]. This is the reason why this algorithm is applied. To improve the performance of the algorithm, modifications is made to the heuristic function taking into account the WUM problems.

In Fig 1 the framework of the research is described.

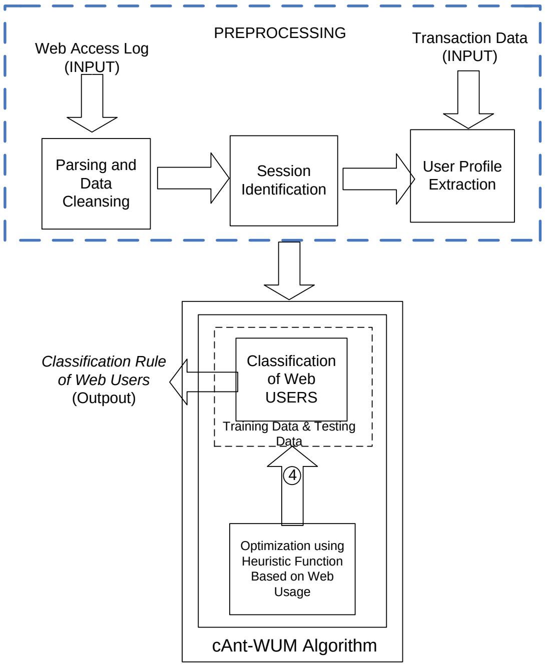

PREPROCESSING
Web Access Log (INPUT)
Transaction Data (INPUT)
Parsing and Data Cleansing
Session Identification
User Profile Extraction
Classification Rule of Web Users (Outpout)
Classification of Web USERS
Training Data & Testing Data
4
Optimization using Heuristic Function Based on Web Usage
cAnt-WUM Algorithm

Fig 1 Framework of the Research

In the *preprocessing stage*, the proposed development of methods that can prepare the data for the classification of the functions of web users by using AcO-based algorithm. This *preprocessing* stage as illustrated in Fig 1 above can be explained as follows:

a. Parsing of *web access logsfrom the* *. *log* formats into a table with the provisions of page URLs not containing an image extension (*. png, *. jpeg, *. gif, etc.) and connection status taken are only with ones with code 200 (access to successful web pages) and 301 (performing *login* or *logout)*

b. Performing *session* data extraction from the tables containing *parsing* results using a *timeout* heuristic and *agent web browser*

c. Performing extraction on the *session* data formed in point (b) to produce the following attributes: number of access, duration of access, access average, the number of *logins,* the duration of time between *login* to *logout,* and the amount of data successfully delivered to the user's web *server.* Total Access done by users within a certain period *(A)* is calculated using Equation 1 as follows:

$$A(t) = \sum_{p=1}^{p=n} t_p, t_p < t_{out} \qquad (1)$$

where p indicates the web page accessed by the user, t_p denote the access time per page and t_{out} defined *timeout* defined by the user.

Length of access by users within a certain period *(D)* is calculated by Equation 2 as follows:

$$D = A(t)_1 + A(t)_2 + \ldots\ldots + A(t)_n \qquad (2)$$

where *A(t)* indicates the amount of time used by the user in one *session.*

The average time a user access within a certain period is calculated using Equation 3 as follows:

$$\bar{D} = \frac{D}{\sum A(t)} \qquad (3)$$

where D represents access duration and $A\ (t)$ indicates the number of access within a certain period.

LOGIN session web users can be determined with Equation 4 as follows:

$$L(t) = \sum_{p=1}^{p=n} t_p, t_p < t_{out}, p_n = "LOGIN" \qquad (4)$$

where p indicates the web page, t_p denotes a web page access time, t_{out} states *timeout* defined by the user and P_n = *"LOGIN"* stated the same web page *LOGIN page.*

The average time a user login is calculated from the total time *logged in* as defined in Equation 5 divided by the number of users *logging* in within the timeframe defined by the user which is calculated using Equation 6 as follow:

$$TotL = L(t)_1 + L(t)_2 + ...+ L(t)_n \tag{5}$$

$$\overline{L} = \frac{TotL}{\sum L(t)} \tag{6}$$

Where TotL states the number of web user *login session*, *L (t)* $_{1\text{-}n}$ indicates *login session* 1 to *n*, and $\overline{L}$ states the average duration of *login* access.

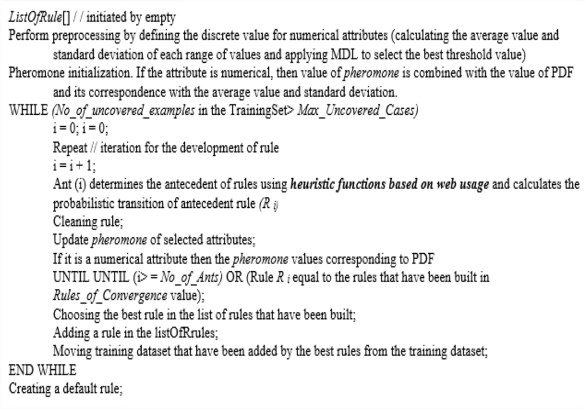
Mapping the list of IP addresses extracted from *web access logs* with IP addresses in the data transaction. In this case of the list of IP addresses in the data transaction is part of the list of IP addresses as the result of of *web access log.*

4. Disscusion

The developed algorithm can generate classification rules with combination of discrete input data and numeric data. In the algorithm, heuristic function used is unique for WUM and is developed using *sequential covering approach*. The format of the classification rules to be built with *cAnt-WUM* algorithm is:

IF < *term* 1 AND *term* 2 AND...> THEN <class>. IF <*term* 1 AND *term* 2 AND ...> THEN <class>.

Each *term* consists of three parts (attribute, operator, value) where value is the value possessed by an attribute. Part of the operator is the operator interface "=" for discrete attributes, and "<" for numerical attributes, where the *term* is the rule antecedent and *class* is the consequent rule.



```
ListOfRule[] / / initiated by empty
Perform preprocessing by defining the discrete value for numerical attributes (calculating the average value and
        standard deviation of each range of values and applying MDL to select the best threshold value)
Pheromone initialization. If the attribute is numerical, then value of pheromone is combined with the value of PDF
        and its correspondence with the average value and standard deviation.
WHILE (No_of_uncovered_examples in the TrainingSet> Max_Uncovered_Cases)
        i = 0; i = 0;
        Repeat // iteration for the development of rule
        i = i + 1;
        Ant (i) determines the antecedent of rules using heuristic functions based on web usage and calculates the
        probabilistic transition of antecedent rule (R ij
        Cleaning rule;
        Update pheromone of selected attributes;
        If it is a numerical attribute then the pheromone values corresponding to PDF
        UNTIL UNTIL (i> = No_of_Ants) OR (Rule R i equal to the rules that have been built in
        Rules_of_Convergence value);
        Choosing the best rule in the list of rules that have been built;
        Adding a rule in the listOfRrules;
        Moving training dataset that have been added by the best rules from the training dataset;
END WHILE
Creating a default rule;
```

Fig 2 Pseudo Code

3.1 Numerical Attributes Preprocessing

Numerical attribute value ranges is developed using C4.5 method [8]. The method of MDL (minimum description length) in the cannot Miner algorithm is also used in the process. The MDL method used is represented in Equation 7, 8, and 9 as follow [2] :

$$Gain\ (y_i, v; S) > \frac{log_2(|S| - 1}{|S|} + \frac{\Delta(y_i, v; S)}{|S|} \tag{7}$$

$$\begin{aligned} Gain(y_i, v; S) = Entropy(S) \\ - \frac{|S_{y_i<v}|}{|S|} . entropy\ (S_{y_i<v}) \\ - \frac{|S_{y_i\geq v}|}{|S|} . entropy\ (S_{y_i\geq v}) \end{aligned} \tag{8}$$

$$\begin{aligned} \Delta(y_i, v; S)\ =\ log_2(3^k - 2) - \\ [k. entropy(S) - k_{y_i<v}. entropy(S_{y_i<v}) - \\ k_{y_i\geq v}. entropy(S_{y_i\geq v})] \end{aligned} \tag{9}$$

where k, $kyi$where k, kyi
where k, kyi<v dan kyi≥v
is the number of different values in the class and if the MDL criterion is defined in Equation 8 are met, then the threshold value v for numerical attribute y i received, and if not then rejected.

3.2. Heuristic Function for Determining the Rule Antecedent

The proposed heuristic function is the function of the use of web-based heuristic (heuristic based on web usage). This function is based on the analysis that identifying or determining the classification of web users in the group of potential or not is done by studying the pattern of usage or interaction of web users with the web through the access duration, frequency of access and the number of transactions [3]. Heuristic function developed is specific for WUM and can not be used for datasets that do not contain the attribute "access number" and "access average ". The heuristic function proposed is as follows:

$$\eta_{ij} = \frac{log_2 k - W_a H\ (C\ |A_i = V_{ij})}{(Entropy\ (S)x\ 2) - ((W_b G_{TotA}) + (W_c G_{AR}))} \tag{10}$$

where ηij is a heuristic value of the candidate term ij, G TotA is the information gain (Eq. 8) of the attribute "Total Access" and G AR stated information gain of attribute "Average Access". $H\ (C\ |Ai = Vij)$ is the value of the entropy of the candidate term ij, k is the number of classes in a dataset, Entropy (S) is the value of the entropy of a dataset, W a is a weighting value for $H\ (C\ |Ai = Vij\ )$, W b is the weight value of G AR, and W c is the weight value of GAR. Entropy (S) is determined by Equation 11 below [4]:
where n c is the number of data for class c, k is the number of classes, and N is the number of data items

$$Entropy\ (S) = -\sum_{c=1}^{k} \frac{n_c}{N} log_2 \frac{n_c}{N} \tag{11}$$

in the dataset. For determining the entropy term ij as a rule antecedent candidates the following Eq 12 [6] is used.

$$H\ (C\ |A_i = V_{ij}) = -\sum_{c=1}^{k}(P(c\ |A_i = V_{ij}).\ log_2 P(c\ |A_i = V_{ij})), \qquad (12)$$

where c is the class attributes, k is the number of class attribute values, A i is the i-th attribute and V ij is value of j-th attribute of i-th attribute, and A i = V ij is the possibility of c class covered by A i = V ij.

This web-usage heuristic function uses three heuristics which can be explained as follow:

1. Heuristics entropy term ij is used to measure the level of uniformity (homogeinity) in corresponding with its class [4]. If the entropy of the data is great, then the data is not uniform in the corresponding to the class or the data scattered in various classes, while when the entrophy is small, then the data is more uniform in corresponding with its class. In this greater entropy term ij the smaller the chances to be selected as the rule antecedent and the smaller the entropy term ij the greater the chances to be selected as the antecedent rule [6].
2. Information gain is used to measure the expected reduction of entropy caused by the dataset partition attribute data in a dataset [4]. Information gain is used to determine the most relevant attributes for the classification, in the decision tree algorithm. The greater the information gain of an attribute, the greater the chances of the attributes are selected as attributes for the classification, and vice versa. Inherited this nature, the greater the information gain of attribute access number and the average access time the better the chances term ij to be selected as the antecedent rules.
3. The weight value indicates how important a heuristic relative to the other heuristics.

3.3 Pheromone Updates and Counting Probabilistic

Transition For discrete attributes, pheromone level for each attribute value is initialized with 1/n in which n is the total number of attribute value for all attributes including numerical attributes. For numerical attributes, updates pheromone values is done using composite kernel with probability density function (PDF) as presented in Equation 13 [11] as follows:

$$\tau_{ij} = \frac{\sum_{x=1}^{h} G(x)}{k\ j} \qquad (13)$$

where τ ij shows the level of pheromone in the range of values to a numeric attribute j at the position i, j k states range normalization constant to the value-j, G (x) states combined kernel PDFs for numerical attributes, l stated range minimum value to the valuei and h shows the maximum value range of the value of i. The function G (x) is determined by the following Equation 14 [11].

$$P(x)G\ (x, \omega, \mu, \sigma) = \sum_{j=1}^{k} \omega \cdot g(x, \mu_j, \sigma_j) \qquad (14)$$

where ω states weight vector associated with the mixture, µ stated the average vector, and σ states the standard deviation vector. The function of k i is determined by the following Equation 15 [11].

$$k_i = \frac{\sum_{x=l}^{h} G(x)}{\frac{1}{n}} \qquad (15)$$

where ki shows the constant normalization for the range of values of the i-numerical attributes, n is the number of attribute values for all attributes including numerical attributes, G (x) is the combined kernel PDFs for numerical attributes, l indicates the minimum value of the value range i, and h show the maximum value of the value range i. Having defined heuristic function and pheromone value, the next step is to determine probabilistic value of the term to be included in the rules using the following Equation 16 [6].

$$P_{ij} = \frac{\eta_{ij}.\tau_{ij}\ (t)}{\sum_{i\ =\ 1}^{a} x_i\ \sum_{j=1}^{b_i}(\eta_{ij}.\tau_{ij}(t))} \qquad (16)$$

where P ij is the probability term ij which is selected as antecedent rules with a range of values [0.1], ηij= value of heuristic function for term ij , τij (t)= the amount of pheromone is associated with at iteration t corresponds to the amount of pheromone which is available on the existing track position i, j, followed by the existing ant, a = number of attributes, xi = 1 if the attribute value had not been used by the existing ants, and the value 0 otherwise, and bi = Number of values in the domain attribute i.

3.4 Adding Rule Consequent

This stage is the process of adding the rule consequent by taking a class predictor of the majority of data covered by the rule antecedent, so the rules are fully completely developed consisting of antecedents and consequents.

3.5 Rule Cleaning

This process is intended to improve the quality of antecedent rules by simplifying the rules. This is done repeatedly by moving the terms of the rules at a time when the quality of the rule is increased using the principle of sensitify/spesificity as represented in the following Equation 17 [6].

$$Q = \frac{TP}{TP + FN} \cdot \frac{TN}{FP + TN} \qquad (17)$$

where TP (True Positive) is the amount of data covered by the antecedent rules (attribute-value) and also in accordance with the rule consequent (class), FP (False Positive) is the amount of data covered by the antecedent rules but not in accordance with the consistent rule, FN (False Negative) is the amount of data that are not covered by the antecedent rules but in accordance with consistent principles, and TN (True Negative) is the amount of data that are not covered by the antecedent rules and not in accordance with the rules consistently. After the cAnt-WUM algorithm produces a set of web users' classification rule in the process of repeat-UNTIL iteration, then the best rule would be included in the list of classification rules

5. Conclusion

cAnt-WUM algorithm proposed in this research can produce a web user with a high degree of accuracy and lower number of rules (more simple). Heuristics function based on web-usage data proposed is applicable to all web usage dataset containing the access number and the average access time as attributes that indicates the frequency and potential levels of access and cannot be applied to datasets that do not have these two attributes. In development of methods in the WUM preprocessing stage to reduce or eliminate duplicated IP address mapping as the result of web access log and transaction data extraction

6. References

[1] Abraham, A. (2003) : Business Intelligence From Web usage mining, Journal of Information & Knowledge Management, Vol. 2, No. 4, 375-390.

[2] Fayyad, U., and Irani, K. (1993): Multi-interval Discretization of Continuous Valued Attributes for Classification Learning, in Thirteenth International Joint Conference on Artifical Inteligence, Morgan Kaufmann, 1022–1027.

[3] Kimpball, R. (2000): The Data Webhouse Toolkit, Wiley Computer Publishing.

[4] Mitchell, T.M. (1997): Machibe Learning, McGraw Hills.

[5] Otero, F., Freitas A, and Johnson C.G. (2008): cAnt-Miner: an Ant colony Classification Algorithm to Cope with Continuous Attributes, in Ant Colony Optimization and Swarm Intelligence (Proc. ANTS 2008),LNCS 5217. Springer, 48–59.

[6] Parpinelli, R.S., Lopes, H.S., and Freitas, A.A. (2002): Data mining with an Ant Colony Optimization Algorithm, IEEE Transaction on Evolutionary Computation, special issue on Ant Colony Algorithm,Vol.6,321-332.

[7] Padmajavalli, R. (2006): An Overview of Data Pre-Processing in Web Usage Mining, The ICFAI Journal of Information Technology, Vol. 2, No. 3, 55-66.

[8] Quinlan, J.R. (1993): C4.5: Programs for Machine Learning, San Francisco, CA: Morgan Kaufmann.

[9] Srivastava, J. (2000): Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data mining) Explorations, January.

[10] Spiliopoulou, M. (2000): Web Usage Mining for Web Site Evaluation, Communications of The ACM, Vol. 43 No. 68.

[11] Swaminathan, S. (2006): Rule Induction Using Ant Colony Optimization For Mixed Variable Attributes, Tesis Computer Science Texas Tech University.

[12] Xie, X. (2004): Classification Rule Induction With Ant Colony Optimization Algorithm, Tesis Computer Science Texas Tech University.

[13] Abraham, A. (2005): Web Usage Mining Using Artificial Ant Colony Clustering and Linear Genetic Programming, http://cs.okstate.edu, downloaded on February 27th 2007.

[14] Nicholas, H. (2007): Web Page Classification with an Ant Colony Optimization, http://www.kent.ac.uk, downloaded on January 12, 2009.

[15] Ramakrishnan, N. (2007): Data Mining, Lecture 10-September 19, 2007, http://people.cs.vt.edu/~ramakris/Courses, downloaded on September 8th, 2009.

[16] Srivastava, J. (2004): Web Mining-Concepts, Applications & Research Direction, http://cs.umn.edu, downloaded on October 5th, 2004.

[17] Patel, K.B., Patel, A.R.: Process of Web Usage Mining to find Interesting Patterns from Web Usage Data. In: International Journal of Computers & Technology Volume 3(1), August 2012.

[18] Mitharam, M.D.: Preprocessing in Web Usage mining. International Journal of Scientific & Engineering Research 3(2), 1 (2012) ISSN 2229-5518