AN ARTIFICIAL NEURAL NETWORK CLASSIFICATION APPROACH FOR IMPROVING ACCURACY OF CUSTOMER IDENTIFICATION IN E-COMMERCE

Nader Sohrabi Safa[1], Norjihan Abdul Ghani[2], and Maizatul Akmar Ismail[3]

[1, 2, 3] Department of Information System
Faculty of Computer Science & Information Technology
University of Malaya
50603 Kuala Lumpur, MALAYSIA

[1] sohrabisafa@um.edu.my,[2] norjihan@um.edu.my,[3] maizatul@um.edu.my

ABSTRACT

With the advancesin Web-based oriented technologies, experts are able to capture user activities on the Web. Users' Web browsing behavior is used for user identification. Identifying users during their activities is extremely important in electronic commerce (e-Commerce)as it has the potential to prevent illegal transactions or activities particularly for users who enter the system through the use of unknown methods.In addition, customer behavioral pattern identification provides a wide spectrum of applications such as personalized Web pages, product recommendations and present advertisements. In this research, a framework for users' behavioral profiling formation is presented and customer behavioral patternsare used for customer identification in the e-Commerce environment. Based on activity control, policies such as user restriction or blockingcan be applied.The neural network classification and the measure of similarity among behavioral patterns are two approaches applied in this research. The results of multi-layer perceptron with a back propagation learning algorithm indicate that there is less error and up to 15.12% more accuracy on average.The results imply that the accuracy of the neural network approach in customer pattern behavior recognition increases when the number of customers grows.In contrast, the accuracy of the similarity of pattern method decreases.

Keywords*:*Customer identification, Behavioral pattern, Profile, e-Commerce.

1.0 INTRODUCTION

Information technology has provided an enormous volume of business data for experts during the last decade. Security or in other words, confidentiality, integrity and accurate availability of data are extremely important in the e-Commerce environment[1].Users play an important role in this domain. Different methods of user identification have attracted expert attention in recent years. Information about users is collected through Web systems and surveys, and information about Web users' activitiesissurreptitiously collected forcertainpurposes. Users' profilescontainconsiderable data about their demographics and activities. The Web system, e-Commerce systems in particular, collects user data during online registration in order to personalize information or to offer product recommendation [2]. User name, gender, e-mail, telephone, address, job, education and even hobbies comprise the information that is usually collected for marketing. User activities and transactions can be recorded and categorized as business information. This information can be about the users' popular brand sites, the frequency of user visits, the averageamount of money spent per E-purchase and the services or products that customers buy the most[3]. The user profile can contain information that comes from the results of data mining, statistical analysis or other processes[4]. The cookies, login and keys are even used for user identification by experts [5].

In this research, we recorded each user'sactivities in a profileto identifythe user. The user profilescontained information about user activities that werecollectedfrom an anonymous Web session. Then, the user pattern behavior wasmatched with the existing user patterns. The match score identifiedwhether the user wasan authorized user. Biometric user identification, such as finger or iris identification is more accurate in comparing user behavioral identification, because there is a high match score between user and fingerprint or iris. Therefore, identification

based on user behavior can be applied together with other methodsto increase the reliability of user identification or for predicting illegal activity in the e-Commerce environment[6]. For instance, a sudden change in activity, such as heavy network trafficking or large amounts of money being transferred or more frequent transfersof money can indicate illegal activities in the system byanonymous users who enter the system through the use of unknown methods.The Federal Financial Institutions Examination advisesthe use of multi-channel authentication methodology to increase the reliability of user identification. User behavioral pattern identification is one of the methods that can be used in thismulti-approach to user identification[7].

Customerbehavioral patternsencompassimportant items; for example, 40% of users visit yahoo.com as part of their daily activities, 80% of user activities on the Web take more than one hour, users usually visit 5 to 10 pages during their Web activities. Most Web browsing starts with checking one'sYahoo and Gmail email [4]. These repetitivemanner can buildusers' behavioral patterns, and the recurring nature of a customer's manner on the Internet can be used for customer identification. Customer activities make customer behavioral patterns, and the amount of repeat behavior by customers shows the strength of the behavioral pattern[8].Users' profilesencompasstheir behavioral patterns and these patternscontain behavioral itemswhich include measurements based on previous customer activities. These profilescanbe used for identification ataparticular time in the future.

The remainder of this paper is organized as follows. Section 2 describes the previous research in this domain. Section 3 demonstrates the approaches that have been applied to customer identification in e-Commerce. Section 4 describes the data that were used in this research. An evaluation of the proposed methods is presented in section 5, and the implications and conclusion of this research arediscussed in sections 6 and 7, respectively.

2.0 RELATED WORKS

The Internet has become an integral part of human life and has brought significant advantages for people and firms. Many businesses cannot operate well without using this technology. However, security is still the main concern for users and enterprises. User identification based on behavior is an important issue that has attracted the attention of scholars in recent years[9, 10]. Recommendation of products and services, and advertisements are the other applications of this approach. The user behavioral model has beendeveloped based on dynamic and static information onthe interaction of userswith the system. The explicit and implicit approaches are the two main approachesforcollectinga user's behavioral data to form a user'sbehavioral pattern. Psychometric instruments, questionnaires and Web registration forms on the Internet are the main toolsused in explicit user data gathering through which users enter their information into the system consciously[11]. Yang (2010) applied the implicit approach and extracted the characteristics of users' behavior by tracking users' Web navigation during their interaction with the system[4]. This information include the user's purchasing history such as the user'smost frequented e-Commercewebsite, most purchased product category, rating of product, price or average money spent for purchasing, time or session life, start time, finish time, search items, first site and the number of pages that the customer usually visited, user network traffic, mouse click and keyword strokes[12].

2.1. Customer purchase behavior

The nature of human behavior on the Web is repetitive. This is the characteristic that experts use for creating a user profile and for identification. Yang and Su[13] investigated the recurring nature of user behavior with respect to web system activities and considered repeated behavior as the strength of the behavioral pattern. The behavioral characteristics with the measure of their strength were stored in a user profile. A user profilewas created based on the history of customer activitiesin the e-Commerce environment and the profile wasthen used for customer identification. Previous investigations showed that user identification based on behavioral patterns can be applied to small and medium sized websites when enough distinctive information about the e-Customers is available. Chen, Kuo, Wu and Tang [14] proposed Sequential Pattern Mining (SPM) as a useful method to identify customers' purchasing patterns over time. A novel algorithm was presented by Ha, Bae and Park [15] based on the Recency, Frequency and Monetary (RFM) concept to define the sequential pattern of customers' purchasing. Their pattern

segmentation method generates information based on customer purchasing behavior for decision-making by management.

2.2. Applications of customer pattern behavior

User centric and site centric are two methods that are appliedin gathering customers' behavioral data. e-Commerce sitescan capture information about the behavioral manner of customers and consequently, sites can adopt certain methods of profiling to calculate the strength of a user's behavioral pattern[16]. Identifying users or customers who are not explicitly signed in has some advantages for the system. For example, based on behahavioral patterns, particular recommendations or Web personalization can be given[17]. The most important advantage of this approach is fraud detection. Abnormal activities that do not usually match with the other behavioral patterns can be a symbol of fraud. The system sendsan alarm to the administrator once it sensesa large amount of money transactions or network trafficking. The system then tries to identify usersundertaking these kinds of activity or even blocks them pendingfurther investigation[18].

User centric data gathering is another approach for creating a customer profile.Affiliated websites such as Windows Live or Microsoft Passport share users' information among associated websites. Different affiliated sites help to coordinate products or service recommendations[19]. Downloadable client-side software has the potential of tracking client-side activity to build user profiles, and consequently,carry out user identification. Fig. 1 shows the overall view of customer identification process in this research.

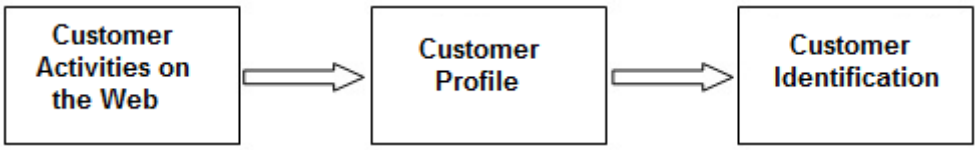

Fig.1.Overall view of customer identification process

Accurate and comprehensive profiles of customers' Web activitiesarethe key issue in customer identification based on behavioral patterns. The personalization of Web content, recommendations based on user profile and fraud detection are important usages of this method. A behavioral profile contains variables that are used by association or conjunctive rules [20]. Some experts have investigated user profiles as a group of rules. They focus on rule validation based on the relationship between users' activities. These profiles can be applied across a group of customers who have similar interests. The time that the customers spend on a page thathas been assigned to a particular product or brand depicts their interest.

The use of profilesin fraud detection has increased theattention ofexperts who endeavor to improve the efficiency of these methods for real time fraud detection.This is because, it is not possible to analyze all profile records immediately in order to identify fraud in behavior patterns. Profiles have been used for intrusion or fraud detection in the study of networks and telecommunications. There are two approaches to fraud or intrusion detection based on customer profile –the anomaly detection technique and signature based detection. In anomaly detection, the profile for each entity is compared with other recent events to produce a profile, which is compared with the typical behavior of the customer to detect abnormal activities. Sudden changes in a user'sWeb manner can be a symbol of fraud. To determine changes in a user's behavior, the traffic of user activities is compared with the individual profile. In the signature based technique, the users' recent traffic is compared with aberrant attack profiles that have been stored in a signature database[21].

3.0 PROPOSED METHOD

The recurring nature of human behavior is an important characteristic that helps us to identify users on a website. Customers enter the web system at different times. A user's activities determine his/her behavioral patterns, and the amount of repeated behavior by a customer shows the strength of that behavior pattern. The system captures the behavior. Therefore, we have several behavioral patterns for acustomer. We need the pattern of the user (strongest pattern for each customer) to compare with the patterns that we have in our database. The strongest behavioral pattern of an active user is considered his or her behavioral pattern, which is compared with the patterns in our database. In this research, two approaches were applied for identifying customers based on their behavioral patterns. First, the similaritiesbetween user behavior and the patterns that are in our database were investigated. Similarity was calculated based on the distance that existedbetween the main items in terms of their behavioral patterns. Secondly, neural network classification was carried out based on the behavioral patterns that were collected before time,*t*. Patterns with a high similarity based on the learning process were in the class that belonged to a customer. Seventy percent of user behavioral patterns were used inthe learning process and thirty percent of the patterns were used inthe testing process[22]. The difference betweenthe testing results and the real data shows the accuracy of the method**.**

3.1. Customer pattern strength

Customer activities are captured from the moment that the customer logs in to the system and the session is created. User pattern strength plays an important role in the accuracy of this method. In other words, the similarity inthe behavioral patterns of a customer leads to the high accuracy of this method. In the next step, the behavioralsimilarity of a certain customer is calculated to identify the customer. Supposed*k* is the number of customers and *p* contains the patterns of customers behavior, then

Customers:$\{\boldsymbol{c_1}, \boldsymbol{c_2}, \dots, \boldsymbol{c_k}\}$ (1)

P:$\{\boldsymbol{p_1}, \boldsymbol{p_2}, \dots, \boldsymbol{p_m}\}$ (2)

$\boldsymbol{s_i}$ is the number of sessions for $\boldsymbol{c_i}$ (3)

$\boldsymbol{s_{ip}}$is the number of sessions for $\boldsymbol{c_i}$ that contain the behavioral pattern of $\boldsymbol{p_i}$ (4)

The similar behaviors of a customer or the support of a pattern among all the patterns of the customer's patternsare defined as the strength of the customer behavioral pattern. Patterns before time *t* are considered for customer identification based on their manner.

Sessions for customer 1 before time *t*: $\{\boldsymbol{e_1}^{\boldsymbol{1}}, \boldsymbol{e_1}^{\boldsymbol{2}}, \dots, \boldsymbol{e_1}^{\boldsymbol{m}}\}$ (5)

Sessions for customer 2 before time *t*: $\{\boldsymbol{e_2}^{\boldsymbol{1}}, \boldsymbol{e_2}^{\boldsymbol{2}}, \dots, \boldsymbol{e_2}^{\boldsymbol{n}}\}$ (6)

Sessions for customers before time *t*: $\{\boldsymbol{e_1}^{\boldsymbol{1}}, \boldsymbol{e_1}^{\boldsymbol{2}}, \dots, \boldsymbol{e_1}^{\boldsymbol{m}}, \boldsymbol{e_2}^{\boldsymbol{1}}, \boldsymbol{e_2}^{\boldsymbol{2}}, \dots, \boldsymbol{e_2}^{\boldsymbol{n}}, \dots, \boldsymbol{e_z}^{\boldsymbol{1}}, \boldsymbol{e_z}^{\boldsymbol{2}}, \dots, \boldsymbol{e_z}^{\boldsymbol{y}}\}$ (7)

The strongest pattern for each customer is considered the candidate pattern for the next step. In other words, the output of this pattern selection is a strong pattern, which belongs to a particular customer. The customer strength of a behavioral pattern is: $\frac{\boldsymbol{s_i}}{\boldsymbol{s_{ip}}}$ (8)

Selected patternsthathavestrong patterns are considered for building customer profiling. Yang and Padmanabhan [23] suggestedthat a combination of the strongest patterns for creating customer profiling leads to more accuracy in customer identification,because, a pattern should be strong enough to identify at least one customer out of a number of others.

3.2. Behavioral pattern similarity

In this step, the customer behavioral profiles that werepreviously created are used to identify the owner of a pattern in future customer activities. Once a customer logs into the system, a session is created. The candidate patterns for customer identification are the patterns that aregenerated based on all user sessions before time *t*. The strength of new or anonymous customer patternsis calculated based on his/her behavioral patterns for the future (equation 1 to 8). To identify the anonymous customer, his/her behavioral pattern is compared with the patterns that have been provided in the previous steps. Similarity between the patterns of an anonymous customer with the previous patterns is calculated by the distance between the two profiles.

Supposed a customer has a$\boldsymbol{P}^{+}$ top Pattern and we want to find a similar pattern. The structure of the patterns is the same. However, the measures of attributes are different.

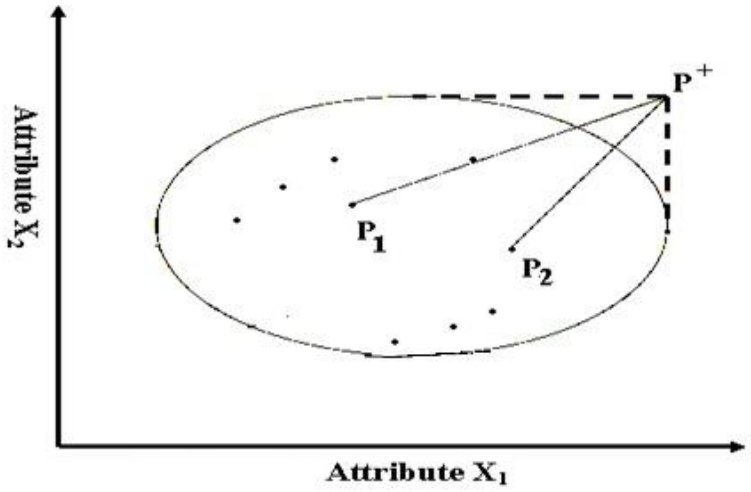

Fig.2.Patterns distances[24]

The matrix below shows the m patterns and n attributes forcustomers;$\boldsymbol{x_{ij}}$shows the measure of attribute *j* in the pattern*i.*

$$\boldsymbol{D} = \begin{matrix} \boldsymbol{p_1} \\ \boldsymbol{p_2} \\ . \\ . \\ . \\ \boldsymbol{p_m} \end{matrix} \begin{bmatrix} \boldsymbol{x_{1,1}} & \boldsymbol{x_{1,2}} & \boldsymbol{x_{1,3}} & \cdots & \boldsymbol{x_{1,n}} \\ \boldsymbol{x_{2,1}} & \boldsymbol{x_{2,2}} & \boldsymbol{x_{2,3}} & \cdots & \boldsymbol{x_{2,n}} \\ . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ \boldsymbol{x_{m,1}} & \boldsymbol{x_{m,2}} & \boldsymbol{x_{m,3}} & \cdots & \boldsymbol{x_{m,n}} \end{bmatrix} \quad (9)$$

To simplify the calculation and transform dimensional measures to non-dimensional ones, the measures are normalized as below:

$$r_{i,j} = \frac{x_{i,j}}{\sqrt{\sum_{i=1}^{m} (x_{i,j})^2}} (0<i\leq m \text{ and } 0<j\leq n) \qquad (10)$$

Some characteristics play a more important role in customer identification. For instance, [25] investigated the differences between male and female motivation and utilization of time spent online. The results of their research showed that females generally use the Internet more for communication than males. Women also use online video calls, text messaging,and social media more than men. In other research,Hwang [26] discussed the effect of perceived enjoyment and social norms onthe intention to adopt e-Commerce systems based on gender classification. The outcomes of his study revealed that the effect of enjoyment is stronger in the men's group and the effect of social norms such as media, friends and family are stronger in the women's group in terms of e-Commerce adoption. Colley and Maltby [27] investigated the effect of the Internet on the lives of females and males. Their research revealed that women use the Internet to find new friends, renew old friendships, shop and book tickets, access information, meet parents, and study online while men use the Internet to find jobs and play games. Park and Chang [28]who studied customer behavior in order to personalize products, revealed that customer behaviorsare different in terms of purchases, basket insertion, field of interest and clicks. They analyzed the features of product and customer behavior and suggested a recommendation system based on the relationship between them. Akman and Mishra [29] also studied the Internet behavior of 200 employees from private and public sector organizations. The results of their survey indicated that there is a difference between genders in terms of daily time spent on the Internet for chatting, emailing, downloading, entertainment and accessing information. The results also showed the relationship between age and time spent daily, on downloading, entertainment, and time spent on information access.Users agedunder 40 years old spent three hours more than users over 40 years old. Aproximately 94% of younger (aged under 40) and 93% of older (aged over 40) respondents reported spending less than one hour using electronic services. They also mentioned that female employees usedthe Internet more than males.

The aforementioned literature indicates that some behavioral characteristics can be applied in classification of customer behavior based on gender, age, time spent on e-Commerce websites, and products in their basket etc. The behaviors that play a greater role in customer identification, such as the time that the customers spend on the e-Commerce website, start and end time, and the products that are usually in their baskets can have more weight. w_i is the weight that is considered for every attribute.

$$V = \begin{bmatrix} v_{1,1} & v_{1,2} & v_{1,3} \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & v_{2,3} \cdots & v_{2,n} \\ . & . & . \cdots & . \\ . & . & . \cdots & . \\ . & . & . \cdots & . \\ v_{m,1} & v_{m,2} & v_{m,3} \cdots & v_{m,n} \end{bmatrix} = \begin{bmatrix} w_1 r_{1,1} & w_2 r_{1,2} & w_3 r_{1,3} & \cdots w_n r_{1,n} \\ w_1 r_{2,1} & w_2 r_{2,2} & w_3 r_{2,3} & \cdots w_n r_{2,n} \\ . & . & . & \cdots . \\ . & . & . & \cdots . \\ . & . & . & \cdots . \\ w_1 r_{m,1} & w_2 r_{m,2} & w_3 r_{m,3} & \cdots w_n r_{m,n} \end{bmatrix} \qquad (11)$$

We can calculate the measure of closeness of the patterns to an ideal pattern to determine the similarity between them.

$$S_i^+ = \sqrt{\sum_{j=1}^{n} \left(v_{i,j} - v_j^+\right)^2} \qquad (i=1,2,3,..,m \text{ and } j=1,2,3,..,n) \qquad (12)$$

The above method is frequently used by experts in different areas of science. The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) is one of the methods used in Multi Criteria Decision Making (MCDM) to assess, evaluate and rankalternatives. In TOPSIS, the distance between a particular alternative with positive and negative alternatives is calculated and the ranking of alternatives is based on their closeness and farness [30]. In this research, the distance between a particular customer's behavioral pattern and other patternsis used to determinethe similarity of the patterns.

In the above method, old and recent customer behavior has equal weight, whereas, in fact, the behavior of customers change over time[31, 32]. To fill this gap in the proposed method, we assigneda higher weight value to recent customer behavioral patterns and less weight value to old customer behavioral patterns[33]. The exponentially Weighted Moving Average (EWMA) is a common method that can be used for this purpose. In EWMA, the weighting for older customer behavior decreases exponentially, but never reaches zero[34].

$$\boldsymbol{F_{t+1}} = \alpha.\boldsymbol{A_t} + \alpha.(1-\alpha).\boldsymbol{A_{t-1}} + \boldsymbol{\alpha}.(\mathbf{1}-\boldsymbol{\alpha})^2.\boldsymbol{A_{t-2}} + \ldots$$
$$= \alpha.\boldsymbol{A_t} + (1-\alpha)(\alpha.\boldsymbol{A_{t-1}} + \boldsymbol{\alpha}.(\mathbf{1}-\boldsymbol{\alpha}).\boldsymbol{A_{t-2}} + \ldots)$$
$$= \alpha.\boldsymbol{A_t} + (1-\alpha)\,\boldsymbol{F_t}$$

(Where α changes between 0 and 1). Finally we get $\boldsymbol{F_t} = \alpha.\boldsymbol{A_t} + (1-\alpha)\,\boldsymbol{F_{t-1}}$ (13)

$\boldsymbol{F_t}$ shows that EWMA at time *t* and $\boldsymbol{A_t}$ is the most recent pattern. $\boldsymbol{F_{t-1}}$ shows the immediate preceding. α =0 means that we ignore the most recent behavior pattern and α =1 means we do not consider the old behavioral manner. α with a measure close to 1 means that recent behavior patterns are more important and measures close to 0 mean the old behavioral pattern is important for us.

As can be seen, recent purchasingbehavior (equation 13), frequency of customer behavior in purchasing (equation 7 and 8) and monetaryhave been considered in this method. Recency, Frequency and Monetary (RFM) are the characteristics of this model.

3.3. Neural network classification approach

Classification and prediction are two main applications of neural networks. A connected group of artificial neurons that use a computational or mathematical method for data processing according to the connection approach to calculation is defined as an artificial neural network. Scholars in different domains try to adopt a research model based upon external and internal information with ANN to acquire better results in terms of accuracy and less error. ANN encompasses neurons, units, nodes and cells. Neurons relate to each other with an associated weight. The training algorithm determines the weights directly from the data, which are considered for training without any assumptions about the statistical distribution of the data [35]. A Multilayer Perceptron (MLP) with error back propagation learning algorithms is one of the most applied kinds of ANN. The MLP contains the three following main parts:

Input layer: this layer includes the variables that contain traits of the input data and releases them to the hidden layers.

Hidden layer: this layer merges the weights with inputs. Each neural node in the hidden layer gets data from neurons, which are in the layer after the input layer. The values change through applied weights and adjust the output value by an activation function.

Output layer: this layer determines the output nodes of the variables.

The error back propagation learning algorithm optimizes the connected weights in the MLP structure. Error back propagation uses the basic principle of the gradient or steepest decent method to reduce the error. The iterative training process starts when training samples are presented to the network and leads to adjusting weights between two neurons that are in adjacent layers [36]. In the end, comparing the calculated result with the real data (target) determines the accuracy of the ANN model. From a classification point of view, when the weights and output classes are normalized (between 0 and 1), the MLP achieves the performance of the maximum a posteriori receiver. The MLP is believed to be capable of the approximation arbitrary function. In the study of nonlinear dynamic and other function mapping problems, the MLP is widely utilized [37]. The backpropagation algorithm is normally used for training in MLP. This means that the desired response to the system must be known. The backpropagation rule distributes the errors through the network and allows adaptation of the hidden processing elements (PEs). In the MLP, any element of a given layer feeds all the elements of the next layer. Massive interconnectivity and nonlinear PEs are two important characteristics of MLP. Error correction learning in PEs at iteration n is based on the equation below:

$$\boldsymbol{e_i}(n)=\boldsymbol{d_i}(n) - \boldsymbol{y_i}(n) \qquad (14)$$

where,
$\boldsymbol{e_i}(n)$ is an instantaneous error
$\boldsymbol{d_i}(n)$ is the desired response for a given input pattern

Based on the theory of gradient descent learning each weight in the network can be adapted by correcting the present value of the weight with a term that is proportional to the present input and error at that weight.

$$\boldsymbol{w_{i,j}}(n+1)= \boldsymbol{w_{i,j}}(n) + \eta\ \boldsymbol{\delta_i}(n)\boldsymbol{x_i}(n) \qquad (15)$$

In the backpropagation algorithm, the legal error $\boldsymbol{\delta_i}(n)$ can be computed from $\boldsymbol{e_i}(n)$ at the output PE or as a wighted sum of errors at the internal PEs (η is step size). This procedure is called the backpropagation algorithm.

The sensitivity of a cost functional is computed by the backpropagation function with respect to each weight in the network. The beauty of this procedure is that it can be implemented with local information and only requires a few multiplications per weight which is very efficient. This is due to the gradient descent procedure, which uses local information that can be caught in local minima. Momentum learning is an improvement to the straight gradient descent in the sense that a memory term (the past increment to the weight) is used to speed up and stabilize convergence. In momentum learning, the equation to update the weights becomes

$$\boldsymbol{w_{i,j}}(n+1)= \boldsymbol{w_{i,j}}(n) + \eta\ \boldsymbol{\delta_i}(n)\boldsymbol{x_i}(n)+ \alpha\ (\boldsymbol{w_{i,j}}(n)- \boldsymbol{w_{i,j}}(n+1)) \qquad (16)$$

where α is the momentum and should be between 0.1 and 0.9.

Two approaches are applied in training. A pattern can be presented and the weights adapted based on the pattern or we can present all the patterns in an input file (epoch), accumulate the weight updates and then update the weights with the average weight update (batch learning).
To start backpropagation, an initial value should be determined for each weight (it can be a small random value) and proceeds until some stopping criterion is met (to cap number iterations, to threshold the output mean square or to use cross validation). Cross validation stops the training at the point of best generalization and is more popular. In any iteration training procedure, the progress in learning is fundamental. The mean square error shows the learning curve. When the learning curve is flat, the step size should be increased to speed up learning. On the other hand, when the learning curve oscillates up and down, the step size should be decreased. When the error goes steadily up, it shows that learning is unstable. In this situation, the network should be reset. To decrease the training times and for better performance, the training data were normalized and tangent hyperbolic (tanh) nonlinearity was used instead of the logistic function[36].

4.0 RESULTS AND DISCUSSION

4.1. Data

User centric data collection helps in the understanding of user online behavior by capturing customers' activities on e-Commerce websites. However, the site centric data collection approach is unable tocapture customer activities on external sites. We need to capture the entire history of customer activities for each customer. The data set for this research was provided by a commercial data vendor. However, due to certain drawbacks such as insufficient number of records, incomplete records and the incompatible structure of the data set withour research the research group made the decision to simulate the customers' behavior based on the trends of previous data. Simulation led to a sufficient number of records and items toform customer behavioral patterns and consequently increased the reliability of the results. The incomplete records in the vendor's dataset were omitted and all complete records in the vendor's dataset were used in the experiment.Customer profile updates were based on customer activities without looking at the sessions before time *t.* The pattern in time *t*was incorporated into the old profiles to generate the updated user profiles at time *t1*[38]. After data screening, one hundred users with a minimum of one hundered sessions were considered the final data set for testing the methods[4, 23]. The start time of customer activity, session life time, end time, purchased product category, amount, price, number of pages visited, number of mouse clicks and keyboard strokes were considered in creating customers' behavioral patterns based on the aforementioned literature. The research method was applied tothe entire data set and also for a data set with a different number of records to study the effect of the method on customer identification. In the matching pattern approach, the patterns of customer behavior in the current session were compared with the other customer patterns in the data set. The customer data comprisedof two parts;1)data manner before time *t,* and2) data manner in the current session. The research method was applied to match these two manners.In the classification approach, based on the training process, the system determinedwhich class or user the pattern belonged to. A brief description of the data set is presented in Table 1.

Table 1. Dataset description

Minimum sessions for each user	100
Number of classes	100
Trining dataset	30% of all records
Testing dataset	70% of all records
Number of users	2,5,10,50,75,100
Session slites	1,10,20,30,..,100

4.2. Results

A review of the literature revealed that the manner of users in different environments has particular characteristics [39]. For instance, in the educational environment, some students use the Internet at a particular time of day in a particular manner that can be recognized based on the time that they start to study, duration of their study, and the subject in which they are most interested. Similarly, these characteristics exist in the e-Commerce environment. Different weights were considered for different behavioral characteristics and the effects of these weights on the accuracy of customer identification were explored. Start time, end time, the duration time of their activities on the e-Commerce website, the product categories that they usually boughtwere the characteristics that we considered for identifying customers based on the above method and the accuracy was calculated. Table 2 shows the accuracy of both methods clearly.

In this research,the minimum and maximum of sessions for a customer were 1 and 100, respectively. The customers were divided into different groups and the accuracy number of the methods was tested. Table 2 depicts the accuracy of pattern similarity and the MLP classification with a different number of customers and sessions. The first row in each cell shows the accuracy of MLP and the second row reflectsthe accuracy of pattern similarity. As shown in Table 2, the accuracy of MLP significantly dominates the pattern similarity method when the size of sessions gets larger and the number of customers increases.

As was mentioned before, the manner of customers in this environment encompasses start time of customer activity, session life time, end time, purchased product category, amount, price, number of pages visited, number of mouse clicks and keyboard strokes. As shown in Table 3, the MLP classification method has significantly greater accuracy compared to the pattern similarity method. Fig. 3 presents the trend of accuracy in the MLP and similarity methods. It is found that the trend of accuracy improves in the MLP classification once the number of sessions and customers increase. In contrast, the accuracy of customer identification decreases in the pattern similarity method when the number of sessions and customers increases.

Table 2.Accuracy of methods (similarity of patterns and MLP classification)

Number of user	**Number of sessions**										
	1	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
2	92.56	93.39	94.48	94.15	96.47	94.83	94.68	95.74	94.97	94.32	95.97
	86.06	90.32	89.52	85.77	82.17	79.76	75.33	71.45	69.86	67.45	65.33
5	91.72	91.61	91.88	92.81	94.65	92.72	92.38	93.91	94.65	94.12	95.31
	89.34	89.78	87.99	83.29	81.63	78.91	73.67	70.58	69.01	66.98	64.97
10	90.91	88.83	88.97	90.68	90.71	90.94	91.03	91.03	92.56	93.21	92.55
	86.64	85.45	85.48	81.39	80.18	77.29	72.57	69.57	68.48	66.02	64.21
20	88.54	86.31	87.14	87.57	83.22	89.34	87.56	90.22	90.11	90.05	90.18
	75.65	83.67	83.91	80.77	79.49	75.73	71.43	68.39	67.99	65.89	63.96
50	86.19	82.42	84.63	85.53	87.65	87.51	87.67	89.21	88.21	89.21	89.11
	70.16	81.59	80.23	78.37	78.39	74.36	69.29	67.78	67.01	64.87	63.21
75	83.46	81.06	86.80	85.03	85.56	86.33	86.84	88.12	87.64	88.23	88.52
	64.96	79.53	78.11	77.23	75.36	73.46	68.37	65.49	66.87	64.12	62.78
100	79.67	79.56	84.02	83.64	82.62	84.18	86.23	87.84	85.27	87.32	87.32
	62.11	78.69	76.36	75.31	74.22	72.66	67.11	64.39	65.57	63.89	62.03

In each cell, the first number is the accuracy ofthe neural network (MLP) classification methodand the second number isthe pattern similarity method.

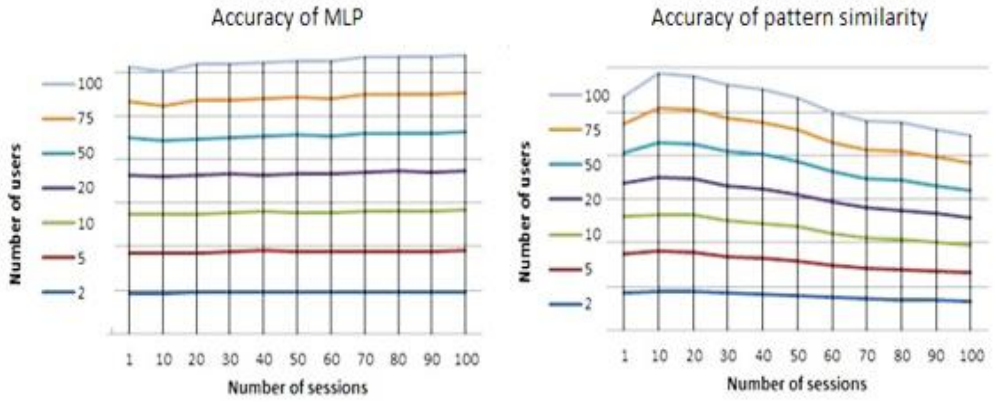

Fig.3.The trend of accuracy in both methods

Table 3 shows the improvement inaccuracy in the MLP and pattern similarity methods. This improvement varies between 0.83 and 30.64 which is significant. The difference between the accuracy of the two methods is not high when the number of customers and sessions is small. However, the advantage that the neural network classification has over the pattern similarity method is that it improves significantly as the size of sessions and customers grows (right side of Table 3).

Table 3.Accuracy improvement by MLP

Number	**Number of sessions**										
of user	**1**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
2	6.5	3.07	4.96	8.38	14.3	15.07	19.35	24.29	25.11	26.87	30.64
5	2.38	1.83	3.89	9.52	13.02	13.81	18.71	23.33	25.64	27.14	30.34
10	4.27	3.38	3.49	9.29	10.53	13.65	18.46	21.46	24.08	27.19	28.34
20	12.89	2.64	3.23	6.8	3.73	13.61	16.13	21.83	22.12	24.16	26.22
50	16.03	0.83	4.4	7.16	9.26	13.15	18.38	21.43	21.2	24.34	25.9
75	18.5	1.53	8.69	7.8	10.2	12.87	18.47	22.63	20.77	24.11	25.74
100	17.56	0.87	7.66	8.33	8.4	11.52	19.12	23.45	19.7	23.43	25.29

4.3. Implications

e-Commerce websites collect and store customer behavior data to improve their systems in order to gain more benefit from their customers. This study focuses on customer behavior in e-Commerce for customer identification, and finally, controlling their access. Firstly, this study presented a simple and efficient method for user behavioral pattern formation in ane-Commerce environment. Customer behavioral pattern includes users' activity on the Web and users' profile includes user behavioral patterns. Customer behavior encompasses items,such as session life time, end time, purchase product category, amount, price, number of pages visited, number of mouse clicks and keyboard strokes. Secondly, this research presenteda method and framework for user behavior profiling and user identification. In addition to the mathematical method, Recency (equation 13), Frequency(equation 8) and Monetary (RFM) are the most important characteristics of the method presented. Thirdly, in this study the accuracy of user identification wasexplored based on the different weights of items in customers' behavioral patterns.This research endeavors to identify customers based on the pattern similarity and MLP classification approaches. The presented method has the potential for application in fraud detection, product recommendation, Web customization and advertisements based on customer behavior.

5.0 CONCLUSION AND FUTURE WORKS

The advantages of Web-based technology are significant in the lives of people today.e-Commerce companies strive to identify the behavior of their customers in order to gaingreater business advantages. e-Commerce websites collect and store customer behavior data both explicitlyand implicitly to improve their systems and gain more benefit from their customers. Users in different environments, such as the educational, health care and commercialenvironments have different behaviors, and the characteristics of these spaces are different[40]. Users in the e-Learning environment are students and teachers, in e-Health care comprise patients and doctors and in e-Commerceare customers and sellers. In the e-Learning environment,the course, subject, and number of units, etc. are the main itemsof student behaviorwhile fore-Commerce, product category, price, amount and period of purchasing are important aspects of customer behavior. The repetitive nature of human behavior is an important characteristic that helps us to identify customers in e-Commerce.

This research focused on customer behavior and identification based on customeractivities. Two approaches were applied to identify customers in the e-Commerce environment. Our experiements show that pattern similarity and neural network classification approaches can be effective and efficient. As mentioned before, customer identification based on behavioral patterns has considerable potential in fraud detection, advertising and product recommendation.

We used the recurring nature of human behavior to calculate customer pattern strength (Frequency, equation 8). Human behavioral patterns change over time; Exponentially Weighing Moving Average (EWMA) was applied to assign more weight values to recent and less weight values to old customer behavioral patterns (Recency, equation 13).The results revealed that although the neural network approach is time consuming, its accuracy with a different number of sessions is higher than the pattern similarity method.

The present study faces a number of limitations. The first limitation refers to situations where there are not many repeat visitors. On sites where occasional users dominate, and the majority of users do not login, it might be more beneficial to assign users to various groups based on their behaviour instead of attempting to uniquely identify them. The second limitation relates to time frames in the user identification process. The sessions we observed after time *t*were from the same anonymous user. There needs to be methods (IP address or cookies) to connect consecutive sessions. In cases where this assumption does not hold, identification can only be done during fairly short periods of Web activity, which can be quite difficult. The third limitation concerns the difficulty in situationswhere the user logs in using someone else's identity, in such intances we cannot directly test the effectiveness of the method on user detection.

In response to the thriving development in e-Commerce, many retailers have developed e-Commerce systems on the Internet to gain a greater competitive advantage. Despite the high development in this domain, some areas such as the application of customer behavior in the recommendation system, the customization of Web pages based on customer interests and previous behavior, fraud detection and customer identification have been neglected. Men and women have different behaviorsin the e-Commerce environment. Their interests are different, they buy different products on the Web, and the duration of their activities on the Web is different. This can provide a clue for future research to increase the accuracy of customer identification based on gender recognition as thefirst step and customer identification as the second step. Using GPS as well as customer identification has been mentioned in other research as a method to improve the accuracy of customer identification. This can also be a subject for future study**.**

Acknowledgment.

This research was conducted by members of the Faculty of Computer Science and Information Technology, University of Malaya and supported by the Institute of Research Management and Monitoring (RG105/12ICT). The authors would like to express their appreciation to all who supported them during the different stages of this research**.**

REFERENCES

[1] Ward, P. and C.L. Smith, "The Development of Access Control Policies for Information Technology Systems*",Computers & Security*,Vol. 21,No. 4, 2002, p. 356-371.

[2] Raghu, T.S., P.K. Kannan, H.R. Rao, and A.B. Whinston, "Dynamic profiling of consumers for customized offerings over the Internet: a model and analysis*",Decision Support Systems*,Vol. 32,No. 2, 2001, p. 117-134.

[3] Reinbach, H.C., D. Giacalone, L.M. Ribeiro, W.L.P. Bredie, and M.B. Frøst, "Comparison of three sensory profiling methods based on consumer perception: CATA, CATA with intensity and Napping*",Food Quality and Preference*,Vol. 32, Part B,No. 0, 2014, p. 160-166.

[4] Yang, Y., "Web user behavioral profiling for user identification*",Decision Support Systems*,Vol. 49,No. 3, 2010, p. 261-271.

[5] Mangipudi, K. and R. Katti, "A Secure Identification and Key agreement protocol with user Anonymity (SIKA)*",Computers & Security*,Vol. 25,No. 6, 2006, p. 420-425.

[6] Ferraiolo, D., V. Atluri, and S. Gavrila, "The Policy Machine: A novel architecture and framework for access control policy specification and enforcement",*Journal of Systems Architecture*,Vol. 57,No. 4, 2011, p. 412-424.

[7] Sharma, S.K. and J. Sefchek, "Teaching information systems security courses: A hands-on approach",*Computers & Security*,Vol. 26,No. 4, 2007, p. 290-299.

[8] Belk, M., E. Papatheocharous, P. Germanakos, and G. Samaras, "Modeling users on the World Wide Web based on cognitive factors, navigation behavior and clustering techniques",*Journal of Systems and Software*,Vol. 86,No. 12, 2013, p. 2995-3012.

[9] Martinez-Moyano, I.J., S.H. Conrad, and D.F. Andersen, "Modeling behavioral considerations related to information security",*Computers & Security*,Vol. 30,No. 6–7, 2011, p. 397-409.

[10] Parsons, K., A. McCormac, M. Butavicius, M. Pattinson, and C. Jerram, "Determining employee awareness using the Human Aspects of Information Security Questionnaire (HAIS-Q)",*Computers & Security*,Vol. 42,No. 0, 2014, p. 165-176.

[11] Zigoris, P. and Y. Zhang, "Bayesian Adaptive User Profiling with Explicit & Implicit Feedback", in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, 2006, Arlington, Virginia, USA.

[12] Aggarwal, C.C., Z. Sun, and P.S. Yu, "Fast algorithms for online generation of profile association rules",*IEEE Transactions on Knowledge and Data Engineering*,Vol. 14,No. 5, 2002, p. 1017–1028.

[13] Yang, Z. and X. Su, "Customer Behavior Clustering Using SVM",*Physics Procedia*,Vol. 33,No. 0, 2012, p. 1489-1496.

[14] Chen, Y.-L., M.-H. Kuo, S.-Y. Wu, and K. Tang, "Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data",*Electronic Commerce Research and Applications*,Vol. 8,No. 5, 2009, p. 241-251.

[15] Ha, S.H., S.M. Bae, and S.C. Park, "Customer's time-variant purchase behavior and corresponding marketing strategies: an online retailer's case",*Computers & Industrial Engineering*,Vol. 43,No. 4, 2002, p. 801-820.

[16] Safa, N.S. and M.A. Ismail, "A customer loyalty formation model in electronic commerce",*Economic Modelling*,Vol. 35,No. 0, 2013, p. 559-564.

[17] Adomavicius, G. and A. Tuzhilin, "Expert-Driven Validation of Rule-Based User Models in Personalization Applications",*Data Mining and Knowledge Discovery*,Vol. 5,No., 2001, p. 33–58.

[18] Li, S.-H., D.C. Yen, W.-H. Lu, and C. Wang, "Identifying the signs of fraudulent accounts using data mining techniques",*Computers in Human Behavior*,Vol. 28,No. 3, 2012, p. 1002-1013.

[19] Cantador, I. and P. Castells, "Extracting multilayered Communities of Interest from semantic user profiles: Application to group modeling and hybrid recommendations",*Computers in Human Behavior*,Vol. 27,No. 4, 2011, p. 1321-1336.

[20] Germanakos, P., N. Tsianos, Z. Lekkas, C. Mourlas, and G. Samaras, "Capturing essential intrinsic user behaviour values for the design of comprehensive web-based personalized environments",*Computers in Human Behavior*,Vol. 24,No. 4, 2008, p. 1434-1451.

[21] Edge, M.E. and P.R. Falcone Sampaio, "A survey of signature based methods for financial fraud detection",*Computers & Security*,Vol. 28,No. 6, 2009, p. 381-394.

[22] Lee, S. and J.Y. Choeh, "Predicting the helpfulness of online reviews using multilayer perceptron neural networks",*Expert Systems with Applications*,Vol. 41,No. 6, 2014, p. 3041-3046.

[23] Yang, Y. and B. Padmanabhan, "Toward user patterns for online security: Observation time and online user identification",*Decision Support Systems*,Vol. 48,No. 4, 2010, p. 548-558.

[24] Liu, S., F.T.S. Chan, and W. Ran, "Multi-attribute group decision-making with multi-granularity linguistic assessment information: An improved approach based on deviation and TOPSIS",*Applied Mathematical Modelling*,Vol. 37,No. 24, 2013, p. 10129-10140.

[25] Kimbrough, A.M., R.E. Guadagno, N.L. Muscanell, and J. Dill, "Gender differences in mediated communication: Women connect more than do men",*Computers in Human Behavior*,Vol. 29,No. 3, 2013, p. 896-900.

[26] Hwang, Y., "The moderating effects of gender on e-commerce systems adoption factors: An empirical investigation",*Computers in Human Behavior*,Vol. 26,No. 6, 2010, p. 1753-1760.

[27] Colley, A. and J. Maltby, "Impact of the Internet on our lives: Male and female personal perspectives",*Computers in Human Behavior*,Vol. 24,No. 5, 2008, p. 2005-2013.

[28] Park, Y.-J. and K.-N. Chang, "Individual and group behavior-based customer profile model for personalized product recommendation",*Expert Systems with Applications*,Vol. 36,No. 2, Part 1, 2009, p. 1932-1939.

[29] Akman, I. and A. Mishra, "Gender, age and income differences in internet usage among employees in organizations",*Computers in Human Behavior*,Vol. 26,No. 3, 2010, p. 482-490.

[30] Behzadian, M., S. Khanmohammadi Otaghsara, M. Yazdani, and J. Ignatius, "A state-of the-art survey of TOPSIS applications",*Expert Systems with Applications*,Vol. 39,No. 17, 2012, p. 13051-13069.

[31] Huang, T.C.-K., "Mining the change of customer behavior in fuzzy time-interval sequential patterns",*Applied Soft Computing*,Vol. 12,No. 3, 2012, p. 1068-1086.

[32] Chen, M.-C., A.-L. Chiu, and H.-H. Chang, "Mining changes in customer behavior in retail marketing",*Expert Systems with Applications*,Vol. 28,No. 4, 2005, p. 773-781.

[33] Shaikh, R.A., K. Adi, and L. Logrippo, "Dynamic risk-based decision methods for access control systems",*Computers & Security*,Vol. 31,No. 4, 2012, p. 447-464.

[34] Chou, C.-Y., J.-C. Cheng, and W.-T. Lai, "Economic design of variable sampling intervals EWMA charts with sampling at fixed times using genetic algorithms",*Expert Systems with Applications*,Vol. 34,No. 1, 2008, p. 419-426.

[35] Wong, T.C., K.M.Y. Law, H.K. Yau, and S.C. Ngan, "Analyzing supply chain operation models with the PC-algorithm and the neural network",*Expert Systems with Applications*,Vol. 38,No. 6, 2011, p. 7526-7534.

[36] Hruschka, E.R. and N.F.F. Ebecken, "Extracting rules from multilayer perceptrons in classification problems: A clustering-based approach",*Neurocomputing*,Vol. 70,No. 1–3, 2006, p. 384-397.

[37] Charalampidis, D. and B. Muldrey, "Clustering using multilayer perceptrons",*Nonlinear Analysis: Theory, Methods & Applications*,Vol. 71,No. 12, 2009, p. e2807-e2813.

[38] Zhang, J. and M. Shukla, "Rule-Based Platform for Web User Profiling", in *Proceedings of the Sixth International Conference on Data Mining (ICDM'06)*, 2006.

[39] Georgakis, A. and H. Li, "User behavior modeling and content based speculative web page prefetching",*Data & Knowledge Engineering*,Vol. 59,No. 3, 2006, p. 770-788.

[40] Loyola, P., P.E. Román, and J.D. Velásquez, "Predicting web user behavior using learning-based ant colony optimization",*Engineering Applications of Artificial Intelligence*,Vol. 25,No. 5, 2012, p. 889-897.