THE APPLICATION OF DATA MINING TECHNIQUES TO PREDICT INTERNET USAGE CONSUMPTION FOR PERSONAL OBJECTIVES IN THE WORK PLACE

Boonyavee Boonyamanop, Siripun Sanguansintukul, Chidchanok Lursinsap
Advanced Virtual and Intelligent Computing (AVIC) Research Center
Department of Mathematics, Faculty of Science, Chulalongkorn University, Bangkok, 10330, Thailand.
dr_tatae@hotmail.com
Siripun.s@chula.ac.th, Chidchanok.l@chula.ac.th

ABSTRACT

Internet usage by employees for personal or inappropriate purposes can directly impact the productivity and efficiency of the organization. This translates to lost time, opportunity and money. In this research we use a data mining technique to build an internet usage consumption model by applying two different methods to web server log data: 1) decision trees based upon a C4.5 Algorithm and 2) Multilayer perceptrons. The overall results obtained indicate that multilayer perceptrons with the cross validation have higher performance in classifying and predicting employee web browsing habits than decision trees. This data mining technique can therefore be a good candidate for helping organizations make more effective evaluation of their human and computer resources.

Index Terms— Data mining techniques, C4.5, multilayer perceptrons, personal web usage in the work place, web usage log

1. INTRODUCTION

The emerge of internet applications as the main media of communication inside and outside the organization has given rise to the use of internet extensively for business activities [2]. Nowadays the Internet is more cost-effective and faster than other methods of communication, making it easier for employers to coordinate the global activities of customers, suppliers, and employees [1].

Although the Internet has benefit for improving communications in the organization, sometimes employees use the Internet for their personal interest. This may cause the loss of time and money for the company due to decreasing the efficiency of work. Therefore, the Internet usage consumption pattern of employees will be useful for predicting the loss of money due to Internet abuse during working hours [16].

Many have warned about the risk of employee Internet usage where lost productivity, waste of time and other resources, becomes a legal liability. For example, Greenfield and Davis [15] show that excessive use of Internet can be a result of addiction.

Today, one of the tools used for analyzing Internet abuse in the workplace is web tracking. However recent research still does not use data mining techniques to build prediction models for forecasting Internet usage consumption.

Data mining techniques can help managers to analyze and predict the consumption of Internet usage from a large web usage log of employee groups in a short period of time. In this paper, we build a new system that applies two such methods of data mining techniques for that purpose: 1) decision trees based on Weka's implemented C4.5 algorithm and 2) Multilayer perceptrons.

2. RELATED WORK

Related work in this study can be classified in 2 categories as follows:

1. The research that applies data mining techniques to build prediction models from web usage tracking data logs. Zhenguo Chen [13] applied data mining techniques for building a model that predicts a user's future URL requests. His experiment consists of 5 input attributes: 1) pattern 2) age 3) gender 4) http version 5) request time and 1 output attribute: class.

In comparison, we conclude that the difference between Chen's research and our research is that Chen's research uses web log data to build a user's future URL request model but our research uses web log data to build an internet usage consumption model.

2. Research that studies the internet usage behavior of employees in the workplace, such as that by Jeffrey J. Johnson [6]. However, this research doesn't use data mining techniques to build an Internet usage consumption model.

3. DATA MINING SYSTEMS

Data mining is concerned with the extraction of interesting and predictive patterns from interesting data. These patterns

can be used to gain insight into process at work in the data and to predict outcomes for future situations.

Data mining system may have 6 major components as follows [3]:

- Database repository. Today, this component refers to a set of data warehouses, databases or other kinds of information repositories. It may use data cleaning to remove noise and inconsistent data and data integration techniques to combine data.
- Database server. A computer program that is responsible for providing database services such as fetching data based on the request of users. Today the Database server is an essential component of E–Commerce systems [4].
- Knowledge base. This is a centralized repository for information such as a public library, a database of related information about a particular subject. Knowledge bases are not a static collection of information. They are dynamic resources that have the capacity to learn as part of an expert system [5].
- Data mining engine. A set of functional modules for data mining tasks, such as: classification, estimation, segmentation, prediction and association [3].
- Pattern evaluation module. Employs interestingness measures and interacts with data mining modules to focus the search toward interesting patterns.
- User interface. Handles the communications between users and the data mining system and helps users to do data mining tasks more efficiency. This includes guiding users to explore database schemas and visualize the result in different forms.

4. METHODS

Two different data mining techniques (C4.5 Decision Tree and Multilayer perceptrons) were investigated and compared. Each model analyzed the same data in order to predict the Internet usage consumption of employees in the workplace.

4.1. Decision Trees.

Decision Trees are supervised learning algorithms for data mining that use class-labeled training tuples to classify data [3]. The algorithm and concept of decision trees was developed by J. Ross Quinlan [8, 9]. The major decision tree algorithm that we use in this paper is C4.5. The algorithm for constructing the C4.5 decision tree is shown as follows

Algorithm: C4.5 Tree Construction

```
//T is the set of cases associated at the node.
//N is a decision node.
//A is attribute.
ConstructDecisionTree (T)
1. ComputeFrequency(T);
2. If OneClass or FewClass
        return a leaf;
        create a decision node N;
3. For Each Attribute A
        ComputeGain(A);
4. N.test = AttributeWithBestGain;
5. if N.test is continuous
        find Threshold;
6. ForEach T' in the splitting of T
7. If T' is Empty
        Child of N is a leaf
        else
8. Child of N = ConstructDecisionTree (T');
9. ComputeErrors of N;
        return N
```

The major advantages of using decision trees to build predictive models are that they are easy to understand and are easily converted to a set of production rules. Decision trees can classify both numerical and categorical data, but the output attribute must be categorical [11].

4.2. Multilayer Perceptrons.

A MLP is a network of simple neurons called perceptrons. The basic concept of a single perceptron was introduced by Rosenblatt in 1958. The perceptron computes a single output from multiple real-valued inputs by forming a linear combination according to its input weights, usually putting the output through some nonlinear activation function. Mathematically, this can be written as

$$y = \varphi(\sum_{i=1}^{n} w_i x_i + b) = \varphi(W^T X + b) \qquad (1)$$

Where w denotes the vector of weights, x is the vector of inputs, b is the bias and φ is the activation function [10, 12]. Nowadays, and especially in multilayer networks, the activation function is often chosen to be the logistic sigmoid [11] defined as

$$f(x) = \frac{1}{(1+e^{-x})} \qquad (2)$$

These functions are used because they are mathematically convenient and are close to linear near origin while saturating rather quickly as they move away from the origin. This allows MLP networks to model well both strongly and mildly nonlinear mappings.

A typical multilayer perceptron (MLP) network consists of a set of source nodes forming the input layer, one or more hidden layers of computation nodes, and an output layer of

nodes. The input signal propagates through the network layer-by-layer.

The computations performed by such a feed forward network with a single hidden layer with nonlinear activation functions and a linear output layer can be written mathematically as [10]

$$x = f(s) = B\varphi(As + a) + b \qquad (3)$$

Where s is a vector of inputs and x is a vector of outputs. A is the matrix of weights of the first layer, a is the bias vector of the first layer. B is the weight matrix and b is the bias vector of the second layer. The function φ denotes an element wise nonlinearity.

MLP networks are typically used in supervised learning problems. The supervised learning problem of the MLP can be solved with the back propagation algorithm. The algorithm consists of two steps. In the forward pass, the predicted outputs corresponding to the given inputs are evaluated as in Equation (3). In the backward pass, partial derivatives of the cost function with respect to the different parameters are propagated back through the network. The chain rule of differentiation gives very similar computational rules for the backward pass for those in the forward pass. The network weights can then be adapted using any gradient-based optimization algorithm. The whole process is iterated until the weights have converged.

To apply the gradient descent procedure, the error function is to be minimized by adjusting weights. We use the squared error loss function because it is the most widely used [11] and it defined by

$$E = \frac{1}{2}(y - f(x))^2 \qquad (4)$$

Where $f(x)$ is the prediction output from the network while y is the instance class label.

There are many advantages of using multilayer perceptrons to build prediction model: they have an ability to learn how to do complex tasks based on the data given for training or initial experience; they can also be used to extract patterns and detect trends that are too complex to be noticed by humans [10, 11].

5. WEB USAGE LOG AND DATASET PREPARATION

In this paper we use the Internet tracking software called "Track4Win" to capture the log file of employees in a small-sized software company. We then analyzed and converted the web usage log to the appropriate format for our experiments. The class attributes that we used in the final test are list as shown in Table 1.

Table 1: Example of a web usage log table generalized in attributes

Data / Attributes	Record 1	Record 2
Browser	IE	Fire fox
Age	20 ~ 25	31 ~ 35
Position	Programmer	DBA
Gender	Male	Male
Days of Week	Monday	Tuesday
URL Type	working related	inappropriate
Cost lost per day	$0	$4.38 ~ $8.76

From Table 1, we defined 6 input attributes: browser, age, position, gender, days of week, URL type, and 1 calculated output attribute that is cost lost per day.

- Browser. This attribute defines type of the browser. It consists of 2 browsers: Internet explorer and Fire fox.
- Age. We classify age into 4 classes according to the group of users: 1) 20 ~ 25, 2) 26 ~ 30, 3) 31 ~ 35 and 4) 36 ~ 40.
- Position. In this paper, we classified position of users into 5 different positions: programmer, database administrator, tester, system analyst and accountant.
- URL Type. We classify URL type into 4 main groups: working related site, search engine site, knowledge site and inappropriate site as shown in Table 2 [6].

Table 2: URL Type table and its sub URL Type

URL Type	Category of web site
Working related site	Web site of organization
	Web application
	Web mail of organization
Search Engine site	Google
	Yahoo
	AltaVista
Knowledge site	Academic
	Technical support
Inappropriate site	Entertainment
	Gambling
	Games
	Humor
	Dating/Social
	Nudity
	Provocative Attire
	Pornography
	Personal Pages
	Profanity
	Sports

- Cost lost. We measure cost lost per day by using the equation [16]:

$$C = S \times T \qquad (5)$$

Where C represents cost lost per day, S represents salary rates per hour [7] and T represents the time of accessing inappropriate sites in each day. For example, if employee A access inappropriate site between 1 hour ~ 2 hours per day and the salary rates per hour of employee A is $3.44, then the cost lost per day is within $3.44 ~ $6.88.

6. CASE STUDY – INTERNET USAGE CONSUMPTION DATA MINING SYSTEM

6.1. Experiment Process.

The design architecture of the data mining system we use is based on the concept of Jiawei Han and Micheline Kamber [3] as shown in Figure 1 and consisting of 7 components:

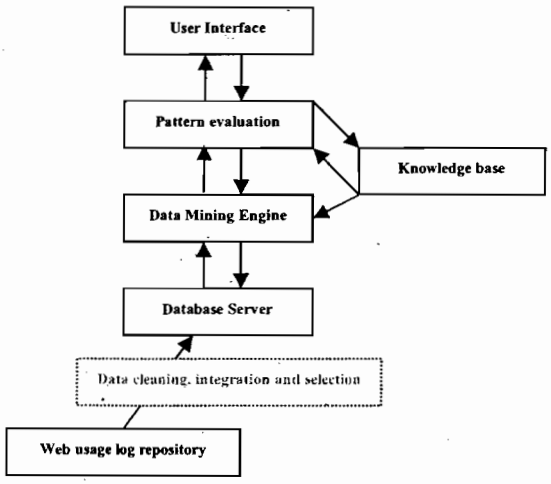

Figure 1. Architecture of Data Mining System

1) Web usage log repository. The log file was tracked by Track4Win [17] software for 1 month. It consisted of 96,749 records in which those having attributes inappropriate for this study were removed. The resulting filtered log file has the following attributes: User Name, Application, URL, and Active time. An example is shown in table 3:

Table 3: Example of captured log file

Data / Attributes	Record 1	Record 2
User Name	Win2006	Extreme
Application	IE	Fire fox
URL	10.4.9.120/cmos	javaworld.com
Active Time (second)	16	20

From table 3, Log file has 4 attributes: User Name, Application, URL and Active Time.

2) Data cleaning, integration and selection. In this process, we use Java language to construct program for convert log file from Table 3 to the appropriated format as show in Table 1.

3) Database Server. Log file that converted to the appropriated format had store in database server.

4) Data Mining Engine. We use Weka tools [14] for classified web log data in CSV format.

5) Pattern evaluation. This component measures the performance of the data mining techniques.

6) Knowledge base. The model generated from the data mining engine is stored in the knowledge base.

7) User Interface. Used for communication between the user and the data mining system.

6.2. Experiment Results.

We select 885 records randomly. The data were split into a training dataset of 531 records and a test dataset of 354 records.

Table 4. Measured Performance Results

Methods / Performance Measures	C4.5	MLP
Correctly Classified Instances (%)	95.1977	94.9153
MAE, mean absolute error	0.0087	0.0092
RMSE, root mean square error	0.0717	0.0729

In this paper, data mining technique performance are compared between the C4.5 algorithm and the multilayer perceptrons for classification of employee Internet usage consumption. We then use Weka software [14] to generate an Internet usage consumption model.

The Result is shown in Table 4. We can see the data mining technique based on the C4.5 algorithm has higher performance.

From table 4, we use three performance measures for comparing the efficiency of each method:

- Correctly classified instances (%). It is the percentage of correctly classified instances when we input test data into each method [11]. For example, when we apply multilayer perceptrons to the classified instance, the rate of correct classification is 94.92 %. It means that when we use 354 instances for testing the prediction model, there are 336 correctly classified instances and 18 that are incorrectly classified.

- Mean absolute error. It can be calculated by an equation as follows:

$$MAE = \frac{|p_1 - a_1| + ... + |p_n - a_n|}{n} \quad (6)$$

Where p is the predicted value and a is the actual value. From this experiment, the mean absolute error of the C4.5 algorithm is lower than the mean absolute error of a multilayer perceptron.

- Root mean square error. It can be calculated by an equation as follows:

$$RMSE = \sqrt{\frac{(p_1 - a_1)^2 + ... + (p_n - a_n)^2}{n}} \quad (7)$$

Where p is the predicted value and a is the actual value. From this experiment, the mean absolute error of the C4.5 algorithm is lower than the mean absolute error of a multilayer perceptron.

An example of the decision tree that generated from the C4.5 algorithm is shown in figures 2 and 3.

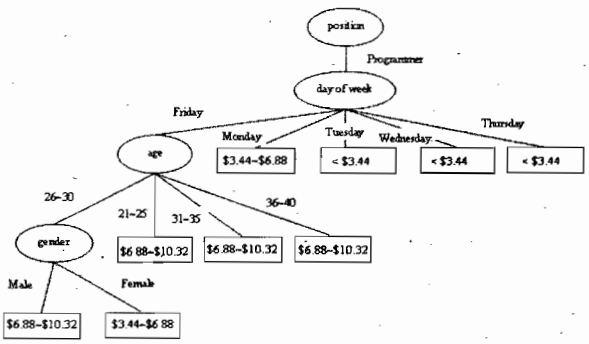

Figure 2. Example of decision tree of programmer that use Internet for personal objectives

From figure 2 we can see that personal Internet consumption is highest on Friday for programmers.

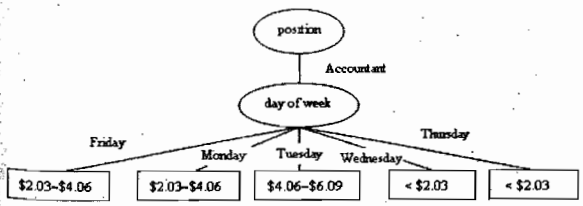

Figure 3. Example of decision tree of accountant that use Internet for personal objectives

From figure 3 we can see that personal Internet consumption is highest on Tuesday for accountants.

When we use Multilayer perceptrons for building an Internet usage consumption model, we can plot graph between predicted values and actual values as show in figure 4.

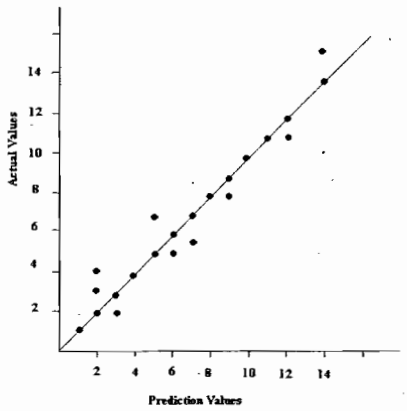

Figure 4. Prediction values vs. Actual values in the experiment of Multilayer perceptrons

From figure 4 it can be seen that most points are line along the 45 degrees diagonal line. The points above the line can be interpreted as false positive, which means that the predicted values are smaller than the actual value. The points below the diagonal line imply that the predicted value is more than the actual value.

6.3. Experiment Results after applying the cross validation to a multilayer perceptron.

In this paper we make several different divisions of the observed data into training set and testing set. This is called cross validation [18]. K-fold cross validation is used to measure the performance of a multilayer perceptron. Here, a 10-fold partition of the data set was created. Split data to 10-fold, hold out successive blocks of observations as test sets, for example, 100 observations, observations 1 through 10, then observations 11 through 20, until to reach 100, and so on.

Each fold is held out in turn and learning scheme trained on the remaining nine-tenths, then the error rate is calculated on the holdout set. Thus the learning procedure is executed a total of 10 times on different training sets. Finally, the 10 error estimates are averaged to show an overall error estimate.

The Result is shown in Table 5. We can see that the data mining technique based on a Multilayer perceptron has higher performance after applying the cross validation concept.

Table 5. Measured Performance Results

Methods / Performance Measures	C4.5	MLP (cross validation)
Correctly Classified Instances (%)	95.1977	95.2542
MAE, mean absolute error	0.0087	0.0086
RMSE, root mean square error	0.0717	0.0693

7. CONCLUSIONS AND FUTURE WORK

In this paper we have applied different data mining techniques for predicting Internet usage consumption by employees for personal objectives in the work place. Two different machine learning algorithms: a C4.5 decision tree and a multilayer perceptron, have been applied to the web usage log dataset of 885 instances. As a result, a multilayer perceptron with the cross validation has achieved a higher performance than the decision tree based on all measurements we used.

Although, we have developed different data mining techniques for building an Internet usage consumption model, this work can further be extended in the following directions. Firstly, we can use the application log instead of web usage log to classify the behavior of employees. Secondly, we can apply web usage log data to associate new performance measurement of employees in the work place.

8. REFERENCES

[1] Gupta, Jatinder N D, "Improving workers' productivity and reducing Internet abuse", *The Journal of Computer Information Systems*, Thursday, January 1 2004.

[2] Murugan Anandarajan and Claire A. Simmers, *Personal web usage in the workplace: A guide to effective human resources management*, Information Science Publishing, United States of America, 2004

[3] Jiawei Han, Micheline Kamber, *Data Mining: Concepts and Techniques second edition*, Morgan Kaufmanm, San Francisco, 2006.

[4] Fujian Liu, Yanping Zhao, Wenguang Wang and Dwight Makaroff, "Database Server Workload Characterization in an E-commerce Environment", *MASCOTS'04*, IEEE, 2004.

[5] Chitta Baral, Sarit Kraus and Jack Minker, "Combining Multiple Knowledge Bases", IEEE Transactions on knowledge and data engineering, Vol 3, No. 2, June 1991.

[6] Jeffrey J. Johnson and Zsolt Ugray, "Employee Internet abuse: Policy versus reality", Issues in information Systems, Vol 8, No. 2, 2007.

[7] Kelly Services Staffing & Recruitment (Thailand) Co Ltd, "Thailand Salary Guide 2007".

[8] J. Ross Quinlan, *C4.5: programs for machine learning*, Morgan Kaufmann, USA, 1993.

[9] Salvatore Ruggieri, "Efficient C4.5", *IEEE Transactions on knowledge and data engineering*, Vol. 14, No. 2, march/april 2002.

[10] Simon Haykin, *Neural Networks: A Comprehensive foundation second edition*, Pearson Prentice Hall, Delhi India, 2005.

[11] Ian H. Witten and Eibe Frank, *Data mining: practical machine learning tools and techniques second edition*, Morgan Kaufmann, San Francisco USA, 2005

[12] Christopher Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995

[13] Zhenguo Chen, "Web Log Mining Based On Fuzzy Immunity Clonal Selection Neural Network", Service Systems and Service Management, 2007 International Conference on Volume, Issue, 9-11 Pages: 1 – 4, June 2007.

[14] University of Waikato, Weka for Windows. Online: http://www.cs.waikato.ac.nz/ml/weka.

[15] D. N. Greenfield and R. A Davis, "Lost in Cyberspace: The Web @ Work", CyberPsychology & Behavior, Vol. 5, No. 4, Mary Ann Leibert, Inc., August 2002.

[16] Annie Wynn and Paris Trudeau, *Internet Abuse at Work: Corporate Networks are paying the price,* SurfControl co Ltd, Online: http://www.suftcontrol.com.

[17] Sepama software co Ltd, Track4Win. Online: http://www.track4win.com.

[18] Kohavi, Ron , "A study of cross-validation and bootstrap for accuracy estimation and model selection", Morgan Kaufmanm, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2*, San Mateo, 1995.