

การประยุกต์เทคนิคเหมืองข้อมูลเพื่อทำนายค่าใช้จ่ายเนื่องจากการใช้อินเทอร์เน็ตเพื่อจุดประสงค์
ส่วนบุคคลในสำนักงาน



นายบุญยวีร์ บุญขมานพ

ศูนย์วิทยทรัพยากร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

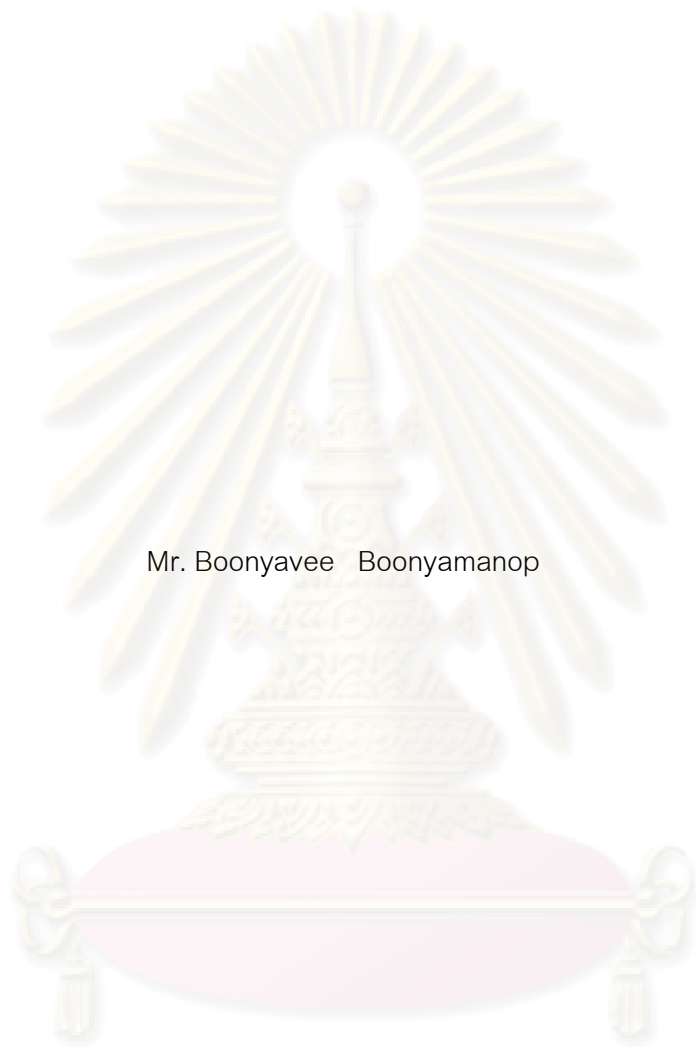
สาขาวิชาวิทยาการคอมพิวเตอร์และสารสนเทศ ภาควิชาคณิตศาสตร์

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2551

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

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



Mr. Boonyavee Boonyamanop

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science and Information

Department of Mathematics

Faculty of Science

Chulalongkorn University

Academic Year 2008

Copyright of Chulalongkorn University

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


By Mr. Boonyavee Boonyamanop

Field of Study Computer Science and Information

Advisor Siripun Sanguansintukul, Ph.D.

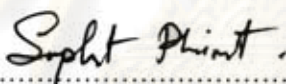
Co-Advisor Professor Chidchanok Lursinsap, Ph.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial
Fulfillment of the Requirements for the Master's Degree

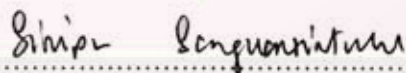


..... Dean of the Faculty of Science
(Professor Supot Hannongbua, Dr.rer.nat.)

THESIS COMMITTEE



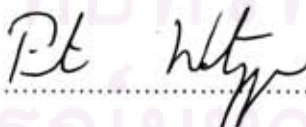
..... Chairman
(Suphakant Phimoltares, Ph.D.)



..... Advisor
(Siripun Sanguansintukul, Ph.D.)



..... Co-Advisor
(Professor Chidchanok Lursinsap, Ph.D.)



..... External Examiner
(Assistant Professor Pirawat Watanapongse, Ph.D.)

บุญยวีร์ บุญยามานพ : การประยุกต์เทคนิคเหมืองข้อมูลเพื่อทำนายค่าใช้จ่ายเนื่องจาก
การใช้อินเทอร์เน็ตเพื่อจุดประสงค์ส่วนบุคคลในสำนักงาน. (APPLICATION OF DATA
MINING TECHNIQUES TO PREDICT INTERNET USAGE CONSUMPTION FOR
PERSONAL OBJECTIVES IN THE WORK PLACE) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: อ.
ดร. สิริพันธ์ สงวนสินธุกุล, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม : ศ. ดร. ชิดชนก เหลือสินทรัพย์,
62 หน้า.

การใช้ Internet เพื่อจุดประสงค์ส่วนบุคคลของพนักงานในสำนักงานได้ส่งผลกระทบ
โดยตรงต่อผลผลิตและประสิทธิภาพการทำงานภายในหน่วยงาน ขณะที่พนักงานใช้ Internet เพื่อ
จุดประสงค์ส่วนตัวนั้น จะก่อให้เกิดการสูญเสียทางด้านเวลาและเงินซึ่งจะลดผลผลิตโดยรวมใน
หน่วยงาน หากเวลาที่พนักงานสูญเสียเนื่องจากการเข้าใช้ website ที่ไม่เหมาะสมยิ่งมาก
ค่าใช้จ่ายก็จะยิ่งมากตามไปด้วย ในงานวิจัยนี้ เราใช้เทคนิคทำเหมืองข้อมูลเพื่อสร้าง
classification model ของค่าใช้จ่ายเนื่องจากการใช้ Internet เพื่อจุดประสงค์ส่วนบุคคลของ
พนักงานในองค์กรจาก web log โดยการใช้อย่างแรก 1. อัลกอริทึมต้นไม้ตัดสินใจ C4.5 และ 2. Multilayer
perceptron ผลการทดลองที่ได้ทั้งหมดบ่งชี้ได้ว่า multilayer perceptrons โดยการใช้อย่างแรก
cross validation มีประสิทธิภาพในการจำแนกและทำนายพฤติกรรมของการเข้าใช้ website ได้ดีกว่า
อัลกอริทึมต้นไม้ตัดสินใจ C4.5 เทคนิคการทำเหมืองข้อมูลในงานวิจัยนี้สามารถช่วยให้องค์กร
ประเมินประสิทธิภาพของพนักงานและแหล่งข้อมูลทางคอมพิวเตอร์ได้เป็นอย่างดี

ศูนย์วิทยทรัพยากร

จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา คณิตศาสตร์.....

สาขาวิชา วิทยาการคอมพิวเตอร์และสารสนเทศ

ปีการศึกษา 2551.....

ลายมือชื่อนิติศ Boonyavee Boonyamanop

ลายมือชื่อ.ที่ปรึกษาวิทยานิพนธ์หลัก Sirin Sirumrithe

ลายมือชื่อ.ที่ปรึกษาวิทยานิพนธ์ร่วม C. Lu

5073632023 : MAJOR COMPUTER SCIENCE AND INFORMATION

KEYWORDS: DATA MINING TECHNIQUES / C4.5 / MULTILAYER PERCEPTRONS / PERSONAL WEB USAGE IN THE WORK PLACE / WEB USAGE LOG

BOONYAVEE BOONYAMANOP : APPLICATION OF DATA MINING TECHNIQUES TO PREDICT INTERNET USAGE CONSUMPTION FOR PERSONAL OBJECTIVES IN THE WORK PLACE. ADVISOR: SIRIPUN SANGUANSINTUKUL, Ph.D., CO-ADVISOR : PROF. CHIDCHANOK LURSINSAP, Ph.D., 62 pp.

Internet usage by employees for personal of 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.

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

Department : Mathematics.....

Field of Study : Computer Science and Information

Academic Year : 2008.....

Student's Signature : *Boonyavee Boonyamanop*

Advisor's Signature : *Siripun Sanguansintukul*

Co-Advisor's Signature : *C. Cox*

Acknowledgments

I am deeply indebted to my thesis advisor, Dr. Siripun Sanguansintukul and Professor Chidchanok Lursinsap, for their valuable guidance, great encouragement and untiring help. Without their constant support and attention, this thesis would have never been written.

I am also grateful to the thesis committee, Assistant Professor Pirawat Wattanipong, and Dr. Suphakant Phimoltares, for their constructive criticism and invaluable advise.

I would like to thank all my teachers for their great contributions and my friends for their encouragement and support during my study.

Finally, words are insufficient to express my gratitude towards my parents who always are a source of unconditional love and support for me.



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

Contents

Abstract in Thai.....	iv
Abstract in English.....	v
Acknowledgements.....	vi
Contents.....	vii
List of Tables.....	ix
List of Figures.....	x
CHAPTER	
I INTRODUCTION.....	1
1.1 Motivation and Problem Description.....	1
1.2 The Objective of Research.....	2
1.3 The Scope of Study.....	2
II BACKGROUND KNOWLEDGE AND RELATED WORK.....	3
2.1 Background on The Problem of Internet Abuse in The Workplace.....	3
2.1.1 Personal Web Page Usage in Organizations.....	3
2.1.2 Monitoring Strategies for Internet Technologies.....	5
2.1.3 Internet Abuse in the Workplace.....	7
2.2 Background on Data Mining.....	9
2.2.1 Data Mining Concept.....	9
2.2.2 Decision Tree based upon a C4.5 Algorithm.....	11
2.2.3 Multilayer perceptron.....	14
2.3 Related Work.....	17
III IMPLEMENTATION.....	20
3.1 The Architecture of Internet Usage Consumption Data Mining System.....	20
3.2 Web Usage Log and Dataset Preparation.....	27
IV EXPERIMENT RESULTS.....	30
4.1 Experiment Results without Applying The Cross Validation.....	30
4.1.1 Experiment Results of Predicting Cost Lost per Day.....	30
4.1.2 Experiment Results of Predicting Time Lost per Day.....	36
4.2 Experiment Results after Applying The Cross Validation to a Multilayer Perceptron...	40

4.2.1 Compare Cost Lost between C4.5 Algorithm and A Multilayer Perceptron with The Cross Validation.....	40
4.2.2 Compare Time Lost between C4.5 Algorithm and A Multilayer Perceptron with The Cross Validation.....	41
V CONCLUSION AND FUTURE WORK.....	43
5.1 Conclusion.....	43
5.2 Discussion.....	43
5.3 Future Work.....	45
REFERENCES.....	46
APPENDICES.....	48
Appendix A Track4Win Software.....	49
Appendix B Weka Software.....	51
VITAE.....	62



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

List of Tables

2.1	A web log table in Zhenguo Chen's experiment.....	17
2.2	Web sites in "Inappropriate" Categories from the experiment of Jeffrey J. Johnson.....	18
3.1	Example of captured log file.....	21
3.2	The attributes of data in appropriate format.....	23
3.3	Table web_usage_log.....	24
3.4	Example of a web usage log table generalized in attributes.....	27
3.5	URL Type table and its sub URL Type.....	28
4.1	Measured Performance Results of cost lost per day.....	32
4.2	Measured Performance Results of time lost per day.....	37
4.3	Measured Performance Results of cost lost per day.....	40
4.4	Measured Performance Results of time lost per day.....	41

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

List of Figures

2.1	Research model of Zoonky Lee.....	4
2.2	A signal-flow graph.....	14
2.3	A signal-flow of a network with one hidden layer.....	15
3.1	Architecture of Data Mining System.....	20
3.2	Flowchart for data cleaning, integration and selection process.....	22
3.3	The flow diagram of data mining technique.....	25
4.1	Decision tree based on C4.5 algorithm on cost lost constructed by WEKA software.....	33
4.2	Example of decision tree of system analysts that use Internet for personal objective.....	34
4.3	Prediction values vs. Actual values.....	35
4.4	Decision tree based on C4.5 algorithm of time lost.....	38
4.5	Prediction values vs. Actual values.....	39
A.1	Example of Track4win interface.....	49
B.1	The WEKA Explorer.....	51
B.2	The interface of Preprocessing.....	52
B.3	Attributes Explorer.....	53
B.4	The interface of preprocess.....	55
B.5	The Classification explorer.....	56

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

CHAPTER I

INTRODUCTION

1.1 Motivation and Problem Description

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 [1]. 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 [2].

Although the Internet has benefit for improving communications in the organization, sometimes employees use the Internet for their personal interest. Personal web use is defined as voluntary online web behavior during normal working hours using any of the organization's resources for activities outside current customary job and work requirements. Such activities can include reading news, making travel arrangements, online purchases, downloading files and music, and searching for jobs. Work inefficiency on the other hand means bad work performance such as low productivity, sloppy work, and lateness. Employees' efficiency could suffer when employees use their Internet access for personal reasons which in turn leads to serious loss of productivity and clogged networks. 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 [3].

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 [4] 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. Web tracking software can monitor all Internet activity used by employee

in the working hour and automatically track visited website addresses includes access time on each website. Managers can use the information from web tracking to analyze Internet usage consumption. However recent research still does not use data mining techniques to build prediction models for forecasting Internet usage consumption.

Data mining is the process of sorting through large amounts of data and picking out relevant information. It is normally used by large corporations employing Business Intelligence integrated with an ERP system to help make managerial decisions based on the patterns and forecasts generated from the data collected. 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 thesis, 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.

1.2 The Objective of Research

In this thesis, the focus is on building Internet usage consumption model to predict cost lost cause from accessing inappropriate websites by applying two different data mining techniques to web server log data: 1) decision trees based upon a C4.5 Algorithm and 2) Multilayer perceptrons. The objective of this thesis is comparing the performance of each method for building Internet usage consumption model.

1.3 The Scope of Study

In this research, we build a new system that applies two different methods of data mining techniques for constructing an Internet usage consumption model and an Internet time consumption model: 1) decision trees based on Weka's implemented C4.5 algorithm and 2) Multilayer perceptrons. We use the Internet tracking software called "Track4Win" to capture the log file of employees and converted the web usage log to the appropriate format by using our java programming.

CHAPTER II

BACKGROUND KNOWLEDGE AND RELATED WORK

This chapter provides a summary of important theoretical backgrounds that are required in this research. It contains three main topics: the problem of Internet abuse in the workplace, data mining and related work.

The first topic gives an elementary introduction to the problem of Internet abuse in the workplace and explains how it decreases overall productivity of employees when they are accessing inappropriate web site.

The second topic provides an introduction to the concept of data mining. It also introduces the data mining techniques such as decision tree based on C4.5 and a multilayer perceptron.

The third topic provides the background of the previous research of Internet abuse and data mining techniques that use in web usage mining. The previous research of this study can be classified in 2 categories: 1) the research that applies data mining techniques to build prediction models from web usage tracking data logs and 2) the research that studies the Internet usage behavior of employees in the workplace.

2.1 Background on the Problem of Internet Abuse in the Workplace

Internet abuse problems are concerned with the use of Internet accessing to inappropriate site of the employees in the workplace. Many organizations indicated that accessing to inappropriate site will affect the productivity of employee and make lost of time.

2.1.1 Personal Web Page Usage in Organizations

Personal web usage during working time is a pervasive behavior observed in the work place. Conlin [5] reported that 37% of working hours are spends to personal web usage. An Internet usage survey noted that 68% of the surveyed companies have Internet

usage policies, that 31% of U.S. companies have spent money on Internet-based activity monitoring and filtering systems [1].

Figure 2.1 shows research model of Zoonky Lee [1]. He developed a comprehensive model based on theories from several areas. Different types of moral attitudes toward personal web usage were used along with resource facilitating conditions, denial of responsibility, social influences and moral obligations.

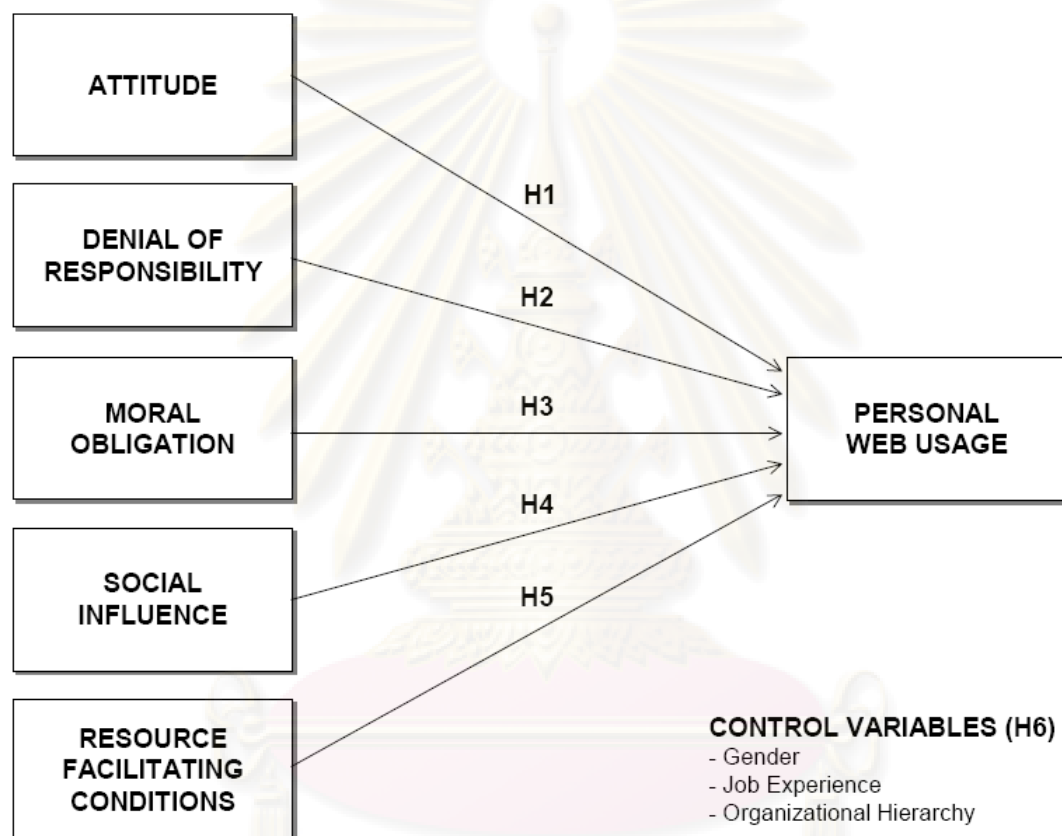


Figure 2.1: Research model of Zoonky Lee.

The research model of Zoonky Lee contains seven components:

1. Personal Web Usage. Zoonky Lee defines personal web usage as extensive personal use of the Internet at work on the grounds that most individuals use the Internet for personal purpose and that whether it becomes unethical behavior is a matter of frequency and time spent. For his analysis purpose, he considers non work related Internet use for more than 30 minutes a day as extensive personal use since companies

are adopting an Internet policy that any extra usage over 30 minutes should be approved by supervisors .

2. Attitude. An attitude toward a behavior is defined as the degree to which the person has a favorable or unfavorable evaluation of the behavior in question.

3. Denial of Responsibility. The denial of responsibility is defined as people's tendency to ascribe responsibility to one or to diffuse and depersonalize it to others, is related to rationalizing the consequences of one's behavior.

4. Moral Obligation. Moral obligation is defined as an individual's perception of the moral correctness or incorrectness of performing a behavior [1].

5. Social Influence. Social influence is defined as the social pressure to perform or not to perform the behavior. Social influence was found to be an important factor in explaining human intention in the social psychology field, information technology adoption, and computer-mediated communications.

6. Resource Facilitating Conditions. Whether resources are easily available or not is considered an important factor that governs an individual's behavior. Ajzen [1] argued that judgment of resource accessibility and opportunity for completing unethical behavior successfully, as well as the perceived power of each facilitator or inhibitor of the behavior, would direct human intention.

7. Control Variables. Loe, Ferrell, and Mansfield [1] performed analysis of individual differences related to ethical decision making and found that individual characteristics such as age, education, gender and work experience affect personal web usage's decisions.

2.1.2 Monitoring Strategies for Internet Technologies

There are several different control mechanisms that an organization might use, but they are generally grouped into one of two categories: 1) managerial and 2) technical. The managerial techniques for monitoring are similar to ways that monitoring of employees has been done for decades is walking around and keeping one's eyes open. Managers

may not feel a need to monitor an employee that is not causing any problems. For example, if work is getting done on time, then no complaints about their. When a manager starts to wonder about an employee's performance, which might be when the manager starts to pay more attention to that employee's work habits.

Electronic monitoring functions can keep a watchful eye on all systems in the network at all hours of the day and night. Records can then be kept and offered as proof of an employee's misgivings as related to using organizational computing equipment and network time. There are two main ways that an organization can accomplish electronic monitoring of personal Internet usage: 1) logging at the gateway and 2) sniffing at the client [1].

1. Logging at the Gateway. When a computer tries to make a connection to another computer, it first checks to see if the destination is on the same local network as it is. If the destination is not on the same subnet, then the packet must be routed outside the network through what is commonly referred to as a gateway. The router that functions as the gateway is essentially the virtual in-out door from the organization's network to the rest of the world. Many logging technologies are then designed to capture and record all of the packets that enter and leave the organization, or at least the header information that indicates the sender, recipient, and content of the message. Gateway logging can be a useful tool in that it provides a central point of control for the network.

Moreover, gateway logging can quite often be defeated by the use of encryption tools. A recent case involving the Scarfo family [1] in the Philadelphia organized crime scene was using PGP (a freely available encryption program) to code computer files which contained family business. Gateway logging did the FBI little good in identifying the contents of the messages, even though they had a search warrant.

2. Sniffing at the Client. When gateway logging is not sufficient, another means of electronically monitoring connections is to monitor them at the source, or make a record at the client's machine. In the Scarfo case, the FBI did exactly that. They installed a keystroke logging program on Scarfo's computer. It recorded all of the keystrokes that

he used, including the ones that made up his pass phrase (a series of words used in PGP, much longer than a password). Once the FBI had his pass phrase, they could decode his messages and then had the evidence to make the arrest.

Client sniffing programs are excellent at recording exactly what the user is doing with the computer at any given time. Many will not only record all of the keystrokes that the user makes, but also will calculate mouse movements and active windows, allowing the reconstruction of the entire computing session. Moreover, they capture undesirable activity that may not be directly network related, such as playing games and typing job application letters. However, these programs are not without their own faults. First of all, the manager must install the program on the user's computer, which may not be as easy as it sounds, especially with laptop and other mobile computers. Second, the program must not be detectable (and thus able to be compromised) by the monitored employees. Third, the program must work on a variety of operating systems, including Windows, UNIX, Linux, and Macintosh, in order to work with all computers. This is not a limitation of gateway logging, as network protocols such as TCP/IP tend to be device independent. Next, the manager has to actually get access to the data captured by the program. Finally, the manager must be able to sift through the mountains of generated data to determine whether or not there is any untoward activity, or enough of it to warrant further investigation. This all being said, there are products available which meet the above concerns to varying degrees, and the next section will discuss some of those products.

2.1.3 Internet Abuse in the Workplace

Internet abuse involves the use of the Internet during work hours in which other non-work-related activities are done (for example: online gambling, online shopping, and online travel booking.). This may be one of the most common forms of Internet abuse in the workplace.

There are many factors which makes Internet abuse in the workplace seductive. It is clear from research in the area of communication that virtual environments have the potential to provide short-term comfort, excitement, and distraction [6]. These reasons

provide compelling reasons why employees may engage in non-work-related Internet use. There are also other reasons as following list [1].

1. Opportunity and access. The Internet is now commonplace and widespread, and is almost integral to most workplace environments. It is not surprising that the development of Internet abuse appears to be increasing across the population [1]. Research into other socially acceptable but potentially problematic behaviors have demonstrated that increased accessibility leads to increased uptake and that this eventually leads to an increase in problems.

2. Affordability. Given the wide accessibility of the Internet, it is now becoming cheaper and cheaper to use the online services on offer. Furthermore, for almost all employees, Internet access is totally free of charge and the only costs will be time and the financial costs of some particular activities (for example: online sexual services, online gambling.).

3. Convenience. Interactive online applications such as e-mail, chat rooms, newsgroups, or role-playing games provide convenient mediums to meet others without having to leave one's work desk. Online abuse will usually occur in the familiar and comfortable environment of home or workplace, thus reducing the feeling of risk and allowing even more adventurous behaviors.

4. Longer working hours. All over the world, people are working longer hours and it is perhaps unsurprising that many of life's activities can be performed from the workplace Internet. Dating via the desktop may be a sensible option for workaholic professionals. It is effectively a whole new electronic "singles bar" which, because of its text-based nature, breaks down physical prejudices. For others, Internet interaction takes away the social isolation that we can all sometimes feel. There are no boundaries of geography, class, or nationality. It opens up a whole new sphere of relationship-forming.

Internet abuse in the workplace has effect the productivity of the organization as following list [1]:

1) Wasting time and money. When employees use the Internet for their personal interest, this may cause the loss of time and money for the organization due to decreasing the efficiency of work.

2) Lost business opportunities. Employees are not only hired to do a job, but to return multiples of their salaries. These would include sales people, of course, but also marketing and public relations personnel and often managers, product developers, and even CEOs. Time wasted surfing the Internet is time not spent selling, marketing or developing a product.

3) Productivity and network performance. Many employees need a fast and reliable network to maximize their own job-related speed and reliability. If some employees are using the Internet for their personal objective, the network performance can be significantly diminished, and this will delay the work of others.

4) Malware and productivity. Social networking, pornography, and even anonymous proxy sites can be the sources of malware and viruses. These can diminish network performance and spread throughout an organization to disrupt or crash computers.

2.2 Background on Data Mining

2.2.1 Data Mining Concept

Data mining has attracted a great deal of attention in the computer industry and in the organization. The information and knowledge gained can be used for application ranging from market analysis, fraud detection, and customer retention, to production control and science exploration [4]. Data mining can be viewed as a result of the evolution of information technology. The database system industry has witnessed an evolutionary path in the development of the following functionalities: data collection and database creation, data management and advanced data analysis. The early development of data collection and database creation mechanisms served as a prerequisite for later development of effective mechanisms for data storage and retrieval, and query and transaction processing.

Data can now be stored in many different kinds of databases and information repositories. Data warehouse technology includes data cleaning, data integration, and on-line analytical processing (OLAP), that is, analysis techniques with functionalities such as summarization, consolidation, and aggregation as well as the ability to view information from different angles. Although OLAP tools support multidimensional analysis and decision making, additional data analysis tools are required for in-depth analysis, such as data classification, clustering, and the characterization of data changes over time. In addition, huge volumes of data can be accumulated beyond databases and data warehouses. Typical examples include the World Wide Web and data streams, where data flow in and out like streams, as in applications like video surveillance, telecommunication, and sensor networks. The effective and efficient analysis of data in such different forms becomes a challenging task [4].

The abundance of data, coupled with the need for powerful data analysis tools, has been described as a data rich but information poor situation. The fast-growing, tremendous amount of data, collected and stored in large and numerous data repositories, has far exceeded our human ability for comprehension without powerful tools.

Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other data source and has the following major components [4]:

1. Database. This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.
2. Database server. The database or data warehouse server is responsible for fetching the relevant data.
3. Knowledge base. This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can

include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction.

4. Data mining engine. This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as classification, prediction, and cluster analysis.

5. Pattern evaluation module. This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search toward interesting patterns. It may use interestingness thresholds to filter out discovered patterns.

6. User interface. This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results.

Data mining involves an integration of techniques from multiple disciplines such as database and data warehouse technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial or temporal data analysis [4].

2.2.2 Decision Tree based upon a C4.5 algorithm

Decision Trees are supervised learning algorithms for data mining that use class-labeled training tuples to classify data [4]. The algorithm and concept of decision trees was developed by J. Ross Quinlan [7, 8]. The major decision tree algorithm in this experiment is C4.5.

Among classification algorithms, decision tree based upon a C4.5 algorithm deserves a special mention for several reasons. It represents the result of research in machine learning that traces back to the ID3 system [9]. The result of T.S. Lim [10] shows that the C4.5 tree-induction algorithm provides good classification accuracy and

is the fastest among the compared main-memory algorithms for machine learning and data mining.

The C4.5 algorithm constructs the decision tree with a divide and conquer strategy. In C4.5 algorithm, each node in the tree is associated with a set of cases. Cases are assigned weights to take into account unknown attribute values. Firstly, only the root is present and associated with the whole training set \mathcal{T} and with all case weights equal to 1.0. Secondly, the divide and conquer algorithm is executed at each node and trying to exploit the locally best choice as shown in the following algorithm:

Program 1: Pseudo code of the C4.5 Tree-Construction Algorithm FormTree(T)

1. *ComputeClassFrequency(T);*

2. *if OneClass or FewCases*

return a leaf;

create a decision node N;

3. *ForEach Attribute 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 = FormTree(T');*

9. *ComputeErrors of N;*

return N

Let T be the set of cases associated at the node. In step 1, the weighted frequency $\text{freq}(D_i, T)$ is computed of cases in T whose class is D_i for $i \in [1, N\text{Class}]$. In step 2, if all cases in T belong to a same class C_j then the node is a leaf with associated class C_j . The classification error of the leaf is the weighted sum of the cases in T whose class is not C_j .

In Step 3, If T contains cases belonging to two or more classes then the information gain of each attribute is calculated. For discrete attributes, the information gain is relative to the splitting of cases in T into sets with distinct attribute values. For continuous attributes, the information gain is relative to the splitting of T into two subsets.

In step 4, the attribute with the highest information gain is selected for the test at the node. Moreover, in case a continuous attribute is selected and the threshold is computed in step 5 as the greatest value of the whole training set that is below the local threshold.

In step 6, a decision node has s children if T_1, \dots, T_s are the sets of the splitting produced by the test on the selected attribute. Obviously, $s = 2$ when the selected attribute is continuous and $s=h$ for discrete attributes with h known values. In step 7, for $l = [1, s]$, if T_l is empty, the child node is directly set to be a leaf, with associated class the most frequent class at the parent node and classification error 0.

In step 8, if T_l is not empty, the divide and conquer approach consists of recursively applying the same operations on the set consisting of T_l plus those cases in T with an unknown value of the selected attribute. Note that cases with an unknown value of the selected attribute are replicated in each child with their weights proportional to the proportion of cases in T_l over cases in T with a known value of the selected attribute.

In step 9, the classification error of the node is calculated as the sum of the errors of the child nodes. If the result is greater than the error of classifying all cases in T as belong with the most frequent class in T then the node is set to be a leaf and all subtrees are removed.

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 [4].

2.2.3 Multilayer Perceptron

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 and then possibly putting the output through some nonlinear activation function. Mathematically this can be written as [11]

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

Where w denotes the vector of weights; x is the vector of inputs; b is the bias and φ is the activation function. A signal-flow graph of this operation is shown in Figure 2.2 [11].

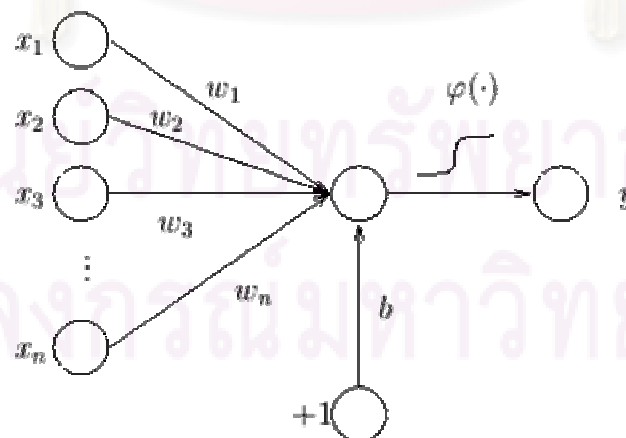


Figure 2.2: A signal-flow graph.

Nowadays, and especially in multilayer networks, the activation function is often chosen to be the logistic sigmoid defined as [12]

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

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

A single perceptron is not very useful because of its limited mapping ability. No matter what activation function is used, the perceptron is only able to represent an oriented ridge-like function. The perceptrons can be used as building blocks of a larger, much more practical structure. 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 signal-flow of such a network with one hidden layer is shown in Figure 2.3 [11].

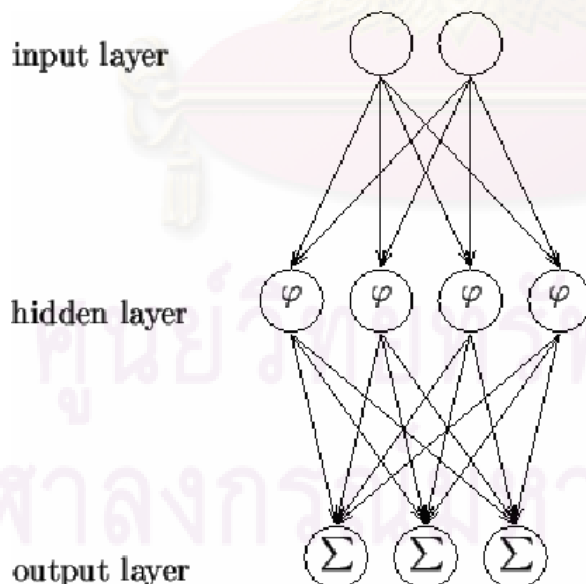


Figure 2.3: A signal-flow of a network with one hidden layer.

The computations performed by such a feedforward network with a single hidden layer with nonlinear activation functions and a linear output layer can be written mathematically as

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

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.

While single-layer networks composed of parallel perceptrons are rather limited in what kind of mappings they can represent, the power of an MLP network with only one hidden layer is surprisingly large.

MLP networks are typically used in supervised learning problems. This means that there is a training set of input-output pairs and the network must learn to model the dependency between them. The training here means adapting all the weights and biases to their optimal values for the given pairs $(s(t), x(t))$. The criterion to be optimized is typically the squared reconstruction error: $\sum_t \|f(s(t)) - x(t)\|^2$.

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 [12] and it defined by

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

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

There have many advantages of using multilayer perceptrons to build prediction model: 1) they have an ability to learn how to do complex tasks based on the data given for training or initial experience and 2) they can also be used to extract patterns and detect trends that are too complex to be noticed by humans [12, 11].

2.3 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 as following table.

Pattern	age	gender	http version	Request time	Class
P1	20~25	Male	1.1	PM3	-1
P2	25~30	Female	1.2	AM7	1

Table 2.1: A web log table in Zhenguo Chen's experiment.

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 [14]. He analyzed Internet log file from a medium sized international company with several locations in North America, Europe, Asia and the Pacific. The company is in the business of manufacturing bio medical and scientific instruments. From his experiment, He categorizes inappropriate web site with frequency of GET calls and Number of Users as shown in following table.

Code	Category	Frequency of GET calls	Number of Users
cs	criminal skills	474	20
et	entertainment	293,561	351
gb	gambling	11,312	49
gm	games	52,019	147
hm	humor	41,718	95
mm	dating/Social	43,899	139
nd	nudity	19,107	49
pa	provocative attire	43,824	108
sm	sexual material	31,329	62
sx	pornography	70,379	89
pp	personal pages	58,714	250
sp	sports	171,753	198
tb	tobacco	556	9
tg	gruesome content	19,472	58
vi	violence	1,317	20

Table 2.2: Web sites in "Inappropriate" Categories from the experiment of Jeffrey J.

Johnson

From his experiment, he concludes that accessing inappropriate web site will make lost of overall productivity. However, this research doesn't use data mining techniques to build Internet usage consumption model for calculating cost lost.



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

CHAPTER III

IMPLEMENTATION

This chapter describes the implementation of the proposed method. It consists of two parts: 1) the architecture of Internet usage consumption data mining system and 2) web usage log and dataset preparation. The first part describes in the detail about the design and analysis of data mining system that bases on the concept of Jiawei Han and Micheline Kamber [4]. The second part explains the detail and attributes of web log data that was captured from web tracking software.

3.1 The Architecture of Internet Usage Consumption Data Mining System

The design architecture of the data mining system in this experiment based on the concept of Jiawei Han and Micheline Kamber [4] as shown in Figure 3.1. It consists of seven components: 1) web usage log repository 2) data cleaning, integration and selection 3) database server 4) data mining engine 5) pattern evaluation 6) knowledge base and 7) user interface.

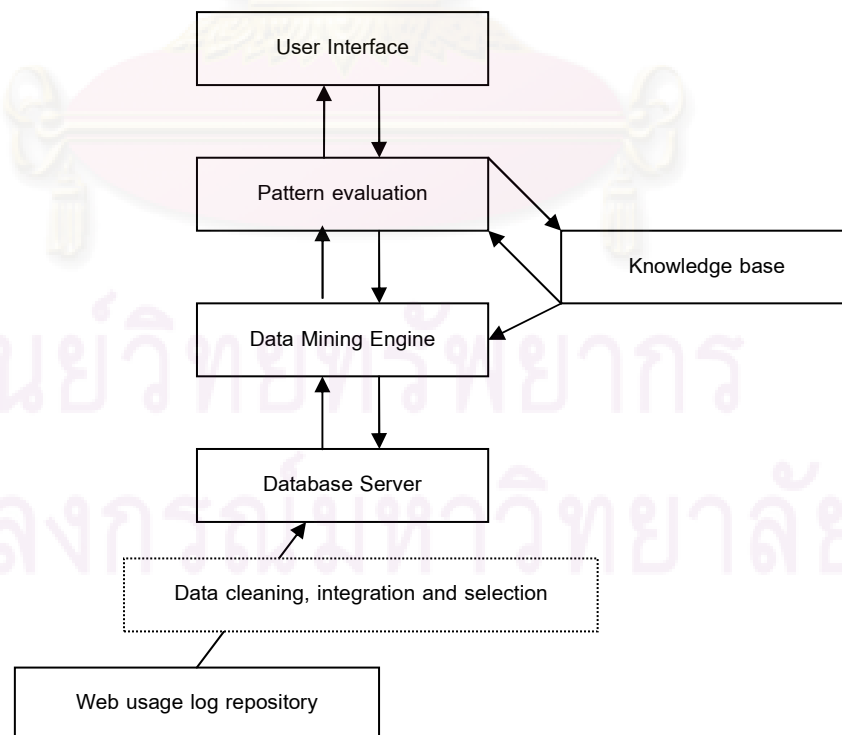


Figure 3.1: Architecture of Data Mining System [4].

3.1.1 Web usage log repository

The log file was tracked by Track4Win [15] 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: 1) User Name 2) Application 3) URL and 4) Active time. An example of captured log file is show in table 3.1:

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

Table 3.1: Example of captured log file.

From table 3.1, Log file has 4 attributes:

1. User Name. It refers to the person who uses a computer system in the company. Users may need to identify themselves for accounting, security, logging, resource management and accessing Internet. The system uses user name to identify employees who use computer in working activity.
2. Application. It refers to the application software that employees use in computer activity. It is any tool that functions and is operated by a computer with the purpose of supporting or improving the software user's work. In this research, it consists of two application program: 1) Internet explorer version 6.0 and 2) Firefox version 2.
3. Uniform Resource Locator (URL). It is a type of Uniform Resource Identifier (URI) that specifies where an identified resource is available and the mechanism for retrieving it. In this research, URL refers to as a web address that an employee accesses in the work place.
4. Active Time. It refers to the time that an employee accesses Internet in each website.

3.1.2 Data cleaning, integration and selection.

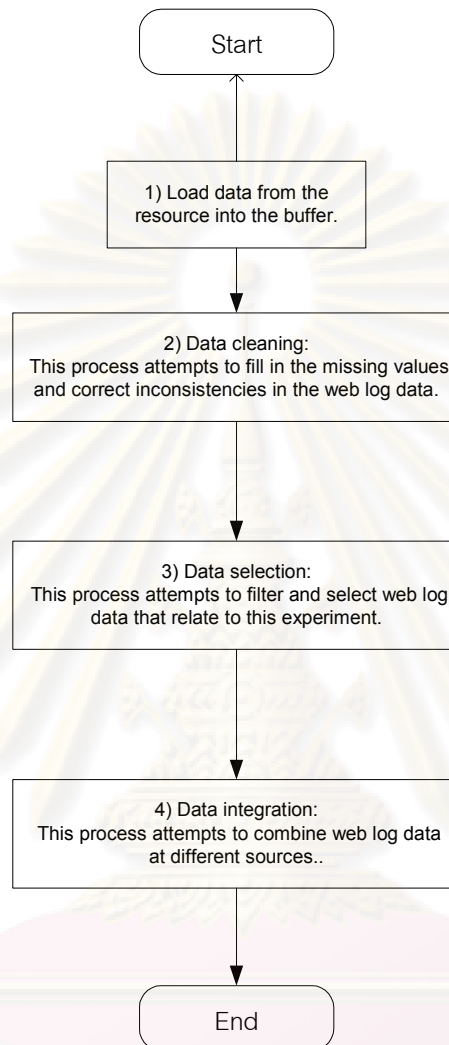


Figure 3.2: Flowchart for data cleaning, integration and selection process.

This process consists of four steps as shown in Figure 3.2.

1. Load data from the resource into the buffer. The system loads captured log data (inappropriate format) from the resource and stores in a buffer in this step.
2. Data cleaning. The missing attributes (age, gender, job position and date) were captured by surveying from employees in the small-size company.
3. Data selection. In this step, the system filters and chooses the data that were captured on Monday, Tuesday, Wednesday, Thursday and Friday.

4. Data integration. In this step, the system calculates the value of 2 output attributes: 1) cost lost per day and 2) time lost per day (the formulas are demonstrated in topic 3.2).

The remaining attributes (age, gender, job position and date) were captured by interviewing from employees. The output from the whole process is the data in the appropriate format that contains 6 input attributes and 2 output attributes as show in table 3.2. The details of attributes will be discussed in topic 3.2.

Attribute name	Type of attribute
Browser	input
Age	input
Position	input
Gender	input
days of week	input
URL type	input
cost lost per day	output
time lost per day	output

Table 3.2: The attributes of data in appropriate format.

3.1.3 Database Server.

Database server is a computer program that provides database services to other computer programs or computers. In this experiment, it refers to application program that stores the data with appropriate format and inappropriate format. The database server in this experiment is MySQL server version 4.1 because it is freeware. The design table for storing the web log data is shown in table 3.3.

Attribute Name	Type
Browser	varchar(100)
Age	varchar(2)
Position	varchar(100)
Gender	varchar(1)
days of week	varchar(1)
URL type	varchar(1)
Cost lost per day	int(100)
Time lost per day	int(100)

Table 3.3: Table web_usage_log.

Table 3.3 is 'web_usage_log' which is created to store web log data in the appropriate format. The first column of table is the attribute name used for training. The second column displays the data type of each attribute. For example, 'varchar' (variable character) is the character string whereas 'int' represents integer values.

3.1.4 Data Mining Engine.

Data mining engine is an application program that stores data mining techniques. A Weka tool [12] is used to classify web log data in CSV format. In the experiment, two different methodologies are compared: 1) Decision tree based on C4.5 algorithm and 2) multilayer perceptrons.

The steps for running the experiment are illustrated in Figure 3.3.

1. Load web log data from resource. In this step, web log data are loaded from a CSV file by Weka and stored in the buffer of computer memory.
2. Setting parameter for C4.5 algorithm and MLP algorithm. In this step, the parameters for both algorithms are set for training data by Weka. The values of parameters are demonstrated in chapter 4.

3. Run training process by using C4.5 algorithm and MLP algorithm. In this step, the function of C4.5 and MLP are processed by Weka.

4. Comparing the result between two techniques. In this step, data mining technique performance are compared between the C4.5 algorithm and the multilayer perceptrons.

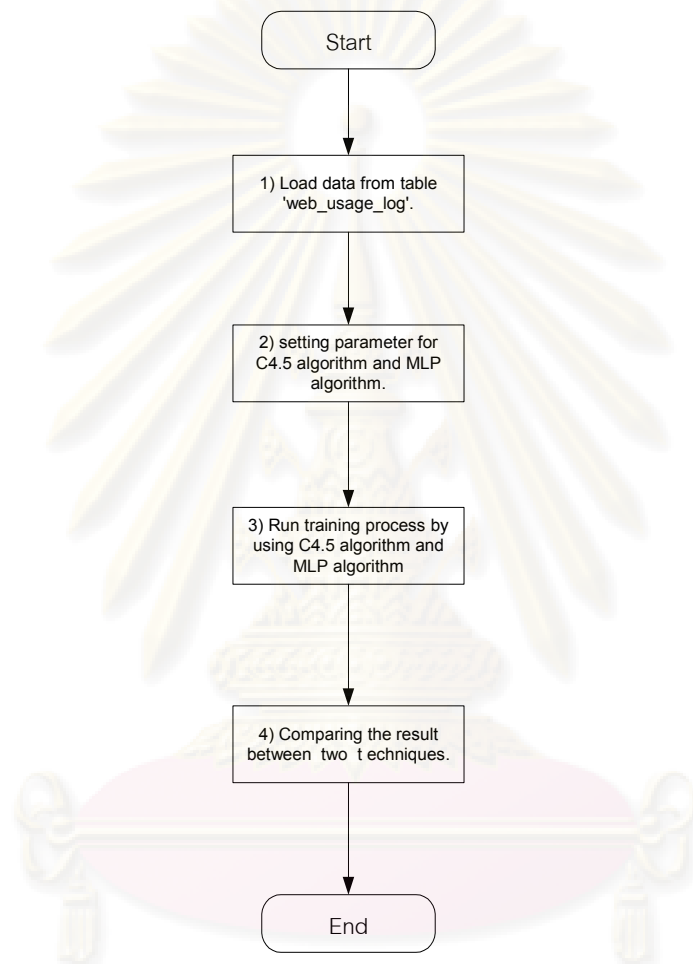


Figure 3.3: The flow diagram of data mining technique.

3.1.5 Pattern Evaluation.

This component measures the performance of the data mining techniques. In this experiment, we use three performance measures for comparing the efficiency of each method:

1) Correctly classified instances (%). It is the percentage of correctly classified instances when we input test data into each method [12]. 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.

2) Mean absolute error. The mean absolute error is a quantity used to measure the efficiency of prediction or forecasts in data mining application. It can be calculated by an equation as follows:

$$MAE = \frac{|p_1 - a_1| + \dots + |p_n - a_n|}{n} \quad (3.1)$$

Where p is the predicted value, a is the actual value and n is the number of data records starting from 1 to n .

3) Root mean square error. The root mean square error is a quantity used to measure the efficiency for prediction or forecasts in data mining application. It is a measure of total error defined as the square root of the sum of the variance and the square of the bias. It can be calculated using the following equation:

$$RMSE = \sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n}} \quad (3.2)$$

Where p is the predicted value, a is the actual value and n is the number of data records starting from 1 to n .

3.1.6 Knowledge Base.

A knowledge base is a special kind of database for knowledge management. It provides the means for the computerized collection and retrieval of knowledge. In this experiment, the model is generated by the data mining engine and stored in the knowledge base. In this experiment, two file directories were built: 1) the knowledge base of cost lost data and 2) the knowledge base of time lost data.

3.1.7 User Interface.

This component communicates between the data mining system and users. It allows users to explore the detail of web log data and the result of experiment. Weka provides GUI (graphic user interface) modules for communicating between human and computer.

3.2 Web Usage Log and Dataset Preparation

The log file was captured from a small-sized software company by using the Internet tracking software called “Track4Win”. The web usage log was analyzed and converted to the appropriate format by the Java programming. The class attributes in the final test are list as shown in Table 3.4.

Attributes Data	Browser	Age	Position	Gender	Days of Week	URL Type	Cost lost per day	Time lost per day
Record 1	IE	20~25	Programmer	Male	Monday	Working related	\$0	0 hour
Record 2	Fire fox	31~35	DBA	Male	Tuesday	Inappropriate	\$4.38~\$8.76	1hour~2hours
...

Table 3.4: Example of a web usage log table generalized in attributes

From Table 3.4, it contains 6 input attributes: browser, age, position, gender, days of week, URL type, and 2 output attributes that are cost lost per day and time lost per day.

1. Browser. This attribute defines type of the browser. It consists of two browsers: Internet explorer and Fire fox.
2. Age. Age in the experiment is classified into four classes according to the group of users: 1) 20 ~ 25, 2) 26 ~ 30, 3) 31 ~ 35 and 4) 36 ~ 40.
3. Position. Position of user is classified into five different positions: programmer, database administrator, tester, system analyst and accountant.
4. Gender. Gender of user can be classified into two groups: male and female.
5. Days of week. This attribute is consists of five values: Monday, Tuesday, Wednesday, Thursday and Friday.

6. URL Type. URL type can be classified into four main groups as shown in Table 3.5 [12]. The classification of URL type that used in this experiment is based on the policy of company.

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

Table 3.5: URL Type table and its sub URL Type.

From table 3.5, working related site is the URL type that related to employee's work and it consists of three sub category: 1) web site of organization 2) web application and 3) web mail of organization. Search engine site is the URL type that employees used for searching the knowledge that related to their work. Search engine site consists of three sub category: 1) Google 2) Yahoo and 3) AltaVista.

Knowledge site is the URL type that employees used for improved their knowledge such as programming skill. Knowledge site consists of two sub category: 1) academic and 2) technical support. Inappropriate site is the URL type that an employee does not have the permission to access it because it will make lost of productivity such as lost of time and money.

7) Cost lost per day. We measure cost lost per day by using the equation :

$$C = S \times T \quad (3.3)$$

Where C represents cost lost per day, S represents salary rates per hour [15] 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.

8) Time lost per day. Time lost per day is the time of accessing inappropriate sites in each day. For example, employee C access inappropriate site between 1 hour and 2 hours per day.



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

CHAPTER IV

EXPERIMENT RESULTS

This chapter describes the result from the experiments. The content of the chapter is divided into two parts:

- Experiment results without applying the cross validation. The experiment in this part does not apply the cross validation. The setting parameters in the C4.5 algorithm and a multilayer perceptron were set to maximize their efficiencies. Two more experiments are involved:

- Experiment results for predicting cost lost per day.
- Experiment results for predicting time lost per day.

- Experiment results after applying the cross validation to a multilayer perceptron. The cross validation is applied to a multilayer perceptron and compared the result with the C4.5 algorithm. Two more experiments are involved.

- Compare cost lost between C4.5 algorithm and a multilayer perceptron with the cross validation.
- Compare time lost between C4.5 algorithm and a multilayer perceptron with the cross validation.

4.1 Experiment Results without Applying The Cross Validation

4.1.1 Experiment Results for Predicting Cost Lost per Day

Two parameters related to C4.5 algorithm training are 'minNumObj' and 'confidenceFactor'.

1. The 'minNumObj' value is set to 2. This parameter is the minimum number of instances per leaf.

2. The 'confidenceFactor' that shows the best result is equals to 0.25. This parameter is the confidence factor used for pruning (smaller values incur more pruning). Pruning is a variant of decision tree learning. It attempts to resolve the problem of overfitting. Decision tree pruning modifies the standard learning algorithm for decision trees so that it does not split on attributes that may be irrelevant [12].

The Multilayer perceptrons consists of five main parameters as following list:

1. The 'learningRate' value is equal to 0.01. It is a parameter that used for update the amount of weights.
2. The 'momentum' value is equal to 0.1. It is the momentum applied to the weights during updating.
3. The 'trainingTime' value is equal to 20,000. It is the number of epochs to train in the experiment.
4. The 'validationThreshold' value is equal to 20. It is the parameter that Used to terminate validation testing. Its value dictates how many times in a row the validation set error can get worse before training is terminated.
5. The 'hiddenLayers' that shows the best result is equals to 17. This parameter defines the number of hidden layers of the neural network.

The data were selected from 885 records randomly and then split into a training dataset of 531 records and a test dataset of 354 records.

In this experiment, data mining technique performances are compared between the C4.5 algorithm and the multilayer perceptrons to classify cost lost of employees. The experiments were run on WEKA software.

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

Table 4.1: Measured Performance Results of cost lost per day.

Table 4.1 shows the comparison of different performance measurements between C4.5 algorithm and multilayer perceptrons approach. The following lists describe each measurement:

1) Correctly classified instances (%). It is the percentage of correctly classified instances when we input test data into each method [12]. The percentage of correctly classified instances of C4.5 algorithm and multilayer perceptrons are 95.1977, 94.9153 respectively. We can see that C4.5 gives the higher percentage than multilayer perceptrons.

2) Mean absolute error. The mean absolute error for C4.5 algorithm and multilayer perceptrons are 0.0087, 0.0092 respectively. We can see that the mean absolute error of the C4.5 algorithm is lower than multilayer perceptrons.

3) Root mean square error. The root mean square error for C4.5 algorithm and multilayer perceptrons are 0.0717, 0.0729 respectively. We can see that the root mean square error of the C4.5 algorithm is lower than multilayer perceptrons.

The decision tree generated from the C4.5 algorithm is shown in figures 4.1.

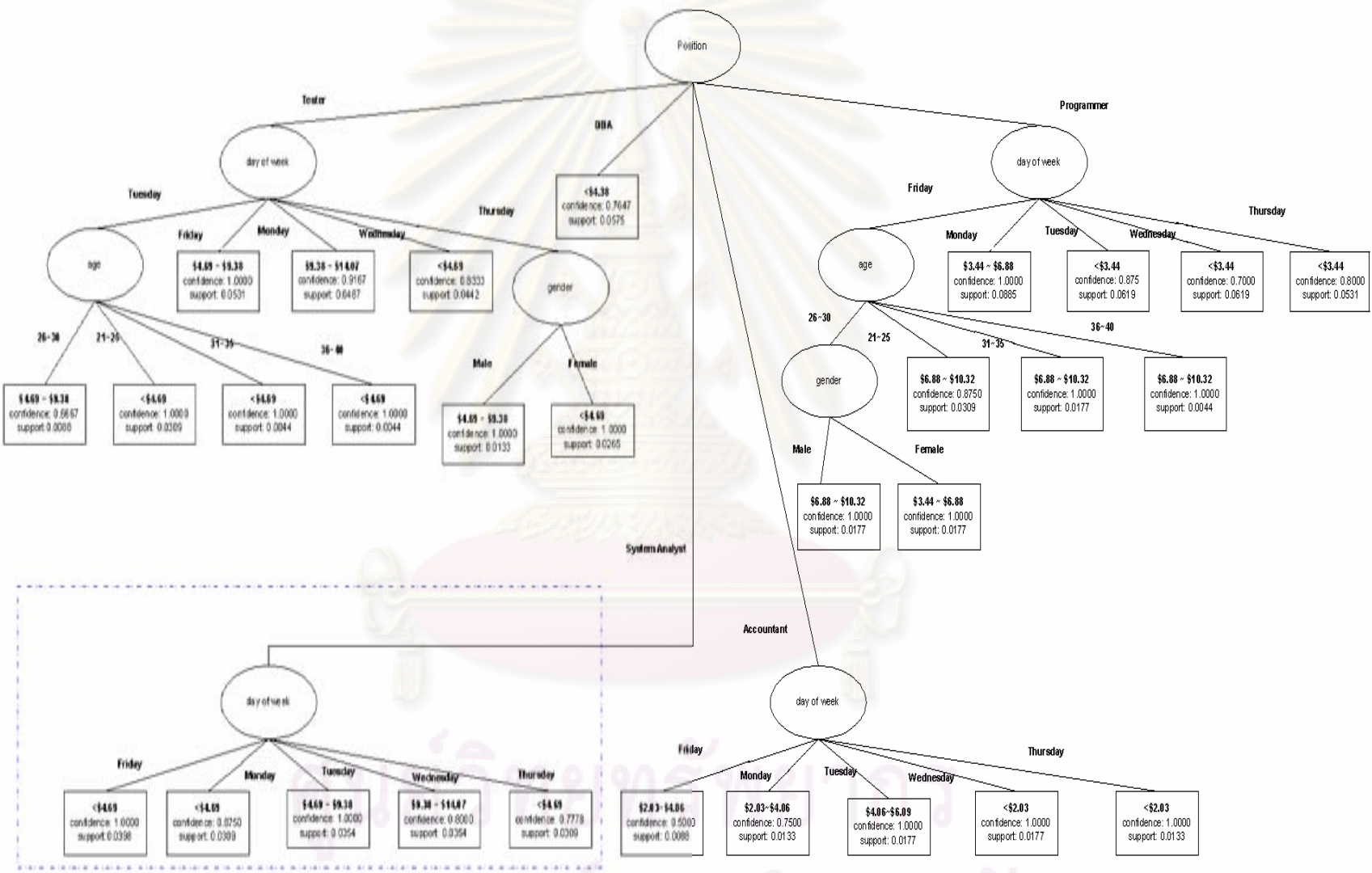


Figure 4.1: Decision tree based on C4.5 algorithm on cost lost constructed by WEKA software.

Figure 4.1 illustrates the tree structure generated by Weka on cost lost. For example, the top most level is the root node. In this case, the 'Position' attribute is the root. There are five different positions namely tester, database administrator, system analyst, accountant and programmer. For the purpose of clear demonstration, different positions might not be on the same level. The second node of ramification is based the 'day of week' attribute. It demonstrates that a system analysis has higher cost lost than other positions. It also shows that the cost lost of database administrator is less than \$4.38 everyday. Some examples of interpretation of the decision tree's branches are the followings:

"If employee's position is a system analyst, the cost lost for this position is between \$9.38 and \$14.07 on Wednesday".

"If employee's position is a database administration, the cost lost for this position is less than \$4.38 in everyday of the week".

The confidence is the percentage value that shows how frequently the rule occurs among all the groups containing the rule body. The confidence value indicates how reliable this rule is. The higher the value, the more often this set of items is associated together. The support is the percentage of groups that contain all of the items listed in that rule. The support value is calculated from among all the groups that were considered and it shows how often the joined rule body and rule head occur among all of the groups that were considered.

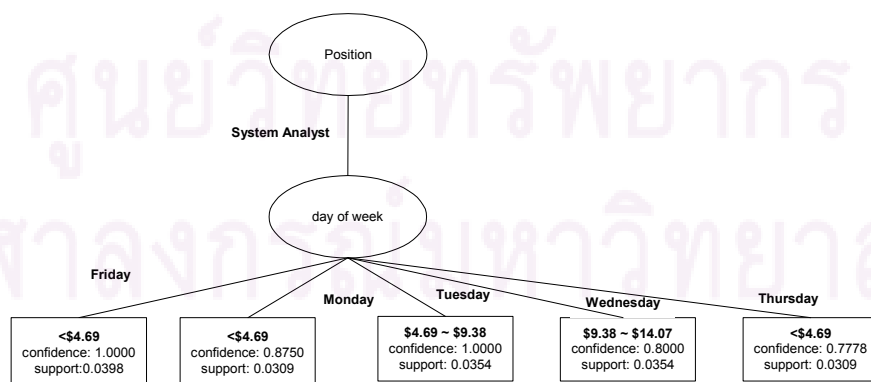


Figure 4.2: Example of decision tree of system analysts that use Internet for personal objective

Figure 4.2 is a part of the tree on Figure 4.1. This decision tree shows that a system analyst accessed inappropriate web site the most on Wednesday for everyday of the week and the cost lost is in the range \$9.38 ~ \$14.07. From child node on Wednesday, A confidence of 0.8 means that if system analysts work on Wednesday, there is an 80% chance that the company will lost \$9.38 ~ \$14.07 per person by Internet abuse and A support of 0.0354 means 3.54% of all of the transactions under analysis showed that a system analyst will make cost lost \$9.38 ~ \$14.07 on Wednesday.

Multilayer perceptron was applied to predict cost lost. The actual values and prediction values are the range of cost lost consisting of 15 different ranges, for example: \$0, \$3.44~\$6.88, \$6.88 ~ \$10.32. Each range is represented by numerical values from 1 to 15. There are altogether 354 points (equal to 354 testing data records). Some points are overlapped. The graph between actual values and prediction values is shown in Figure 4.3.

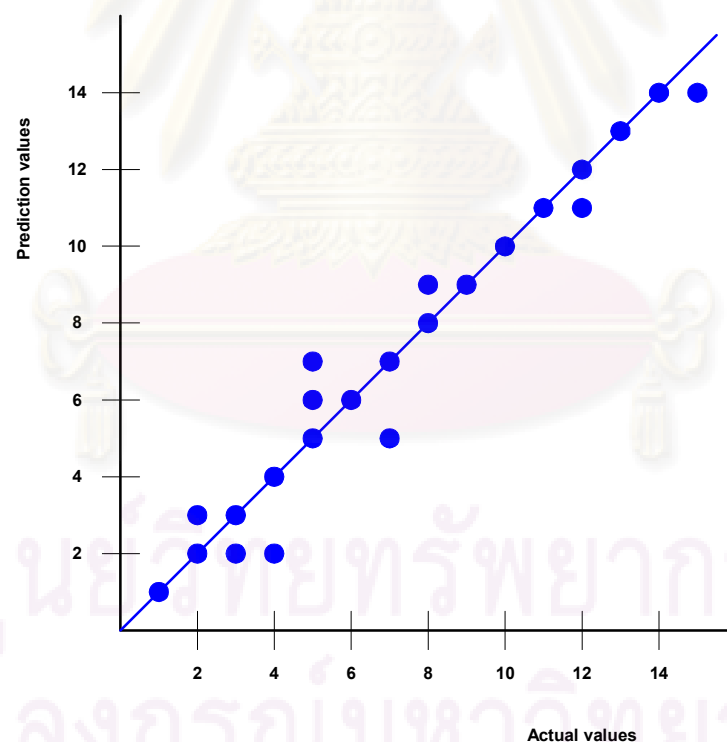


Figure 4.3: Prediction values vs. Actual values.

From figure 4.3, it can be seen most points are line along the 45 degree diagonal line. The points below the line can be interpreted as false positive which means that the

predicted values are smaller than the actual value. The points above the diagonal line imply that the predicted value is more than the actual value.

4.1.2 Experiment Results for Predicting Time Lost per Day

Two parameters related to C4.5 algorithm training are 'minNumObj' and 'confidenceFactor'.

1. The 'minNumObj' value is set to 2. This parameter is the minimum number of instances per leaf.
2. The 'confidenceFactor' that shows the best result is equals to 0.25. This parameter is the confidence factor used for pruning.

The Multilayer perceptrons consists of the following five parameters:

1. The 'learningRate' value is equal to 0.01. It is a parameter that used for update the amount of weights.
2. The 'momentum' value is equal to 0.1. It is the momentum applied to the weights during updating.
3. The 'trainingTime' value is equal to 20,000. It is the number of epochs to train in the experiment.
4. The 'validationThreshold' value is equal to 20. It is the parameter that Used to terminate validation testing. Its value dictates how many times in a row the validation set error can get worse before training is terminated.
5. The 'hiddenLayers' that shows the best result is equals to 12. This parameter defines the number of hidden layers of the neural network.

The data contains 885 records and split into a training dataset of 531 records and a test dataset of 354 records (training: testing = 60:40).

The performance measures on time lost per day between C4.5 and Multilayer perceptrons are compared. The experiments are performed by WEKA.

Performance Measures	C4.5	MLP
Correctly Classified Instances (%)	94.9153	94.6328
MAE (mean absolute error)	0.0353	0.0299
RMSE (root mean square error)	0.145	0.1503

Table 4.2: Measured Performance Results of time lost per day.

Table 4.2 shows the comparison of different performance measurements between C4.5 algorithm and multilayer perceptrons approach. The different performances are discussed below:

1) Correctly classified instances (%). It is the percentage of correctly classified instances when we input test data into each method [12]. The percentage of correctly classified instances of C4.5 algorithm and multilayer perceptrons are 94.9153, 94.6328 respectively. We can see that C4.5 gives the higher percentage than multilayer perceptrons.

2) Mean absolute error. The mean absolute error for C4.5 algorithm and multilayer perceptrons are 0.0353, 0.0299 respectively. We can see that the mean absolute error of the C4.5 algorithm is lower than multilayer perceptrons.

3) Root mean square error. The root mean square error for C4.5 algorithm and multilayer perceptrons are 0.145, 0.1503 respectively. We can see that the root mean square error of the C4.5 algorithm is lower than multilayer perceptrons.

The decision tree of time lost generated from the C4.5 algorithm is shown in figure 4.4.

Figure 4.4: Decision tree based on C4.5 algorithm of time lost.

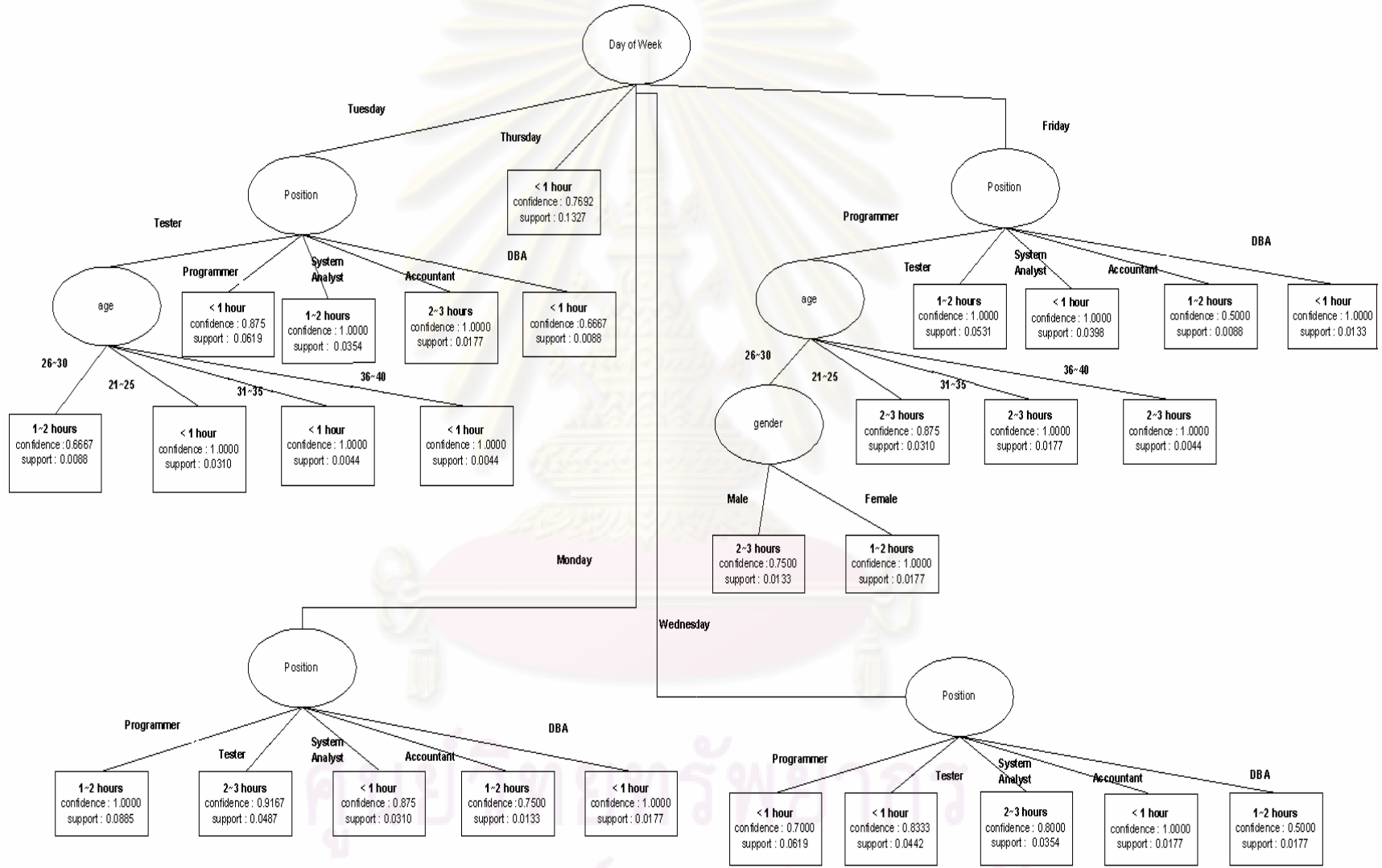


Figure 4.4 illustrates the tree structure generated by WEKA on time lost. In this case, the 'Day of Week' attribute is the root. There are five different days: 1) Monday 2) Tuesday 3) Wednesday 4) Thursday and 5) Friday. For the purpose of clear demonstration, the level of days may not be arranged at the same level. The second node is based on the 'position' attribute. It illustrates that time lost is the highest on Wednesday for system analyst's position. This decision tree shows that on Wednesday, a system analyst accessed inappropriate web site more often than other position. Examples of interpretation of the decision tree's branches:

“On Tuesday an accountant accesses inappropriate web site more often than other position”.

“On Friday a male programmer that age between 26 and 30 accesses inappropriate 2~3 hours per day”.

Multilayer perceptrons were applied for predicting time lost. The actual values and prediction values are the range of time lost that consists of four different values and represent by label: the 'o' stands for no working time lost, the 'a' stands for working time lost less than 1 hour, the 'b' stands for working time lost between 1 hour and 2 hours, the 'c' stands for working time lost between 2 hours and 3 hours. There are altogether 354 points (equal to 354 testing data records) and some points are overlapped. The graph between actual values and prediction values is shown in Figure 4.5.

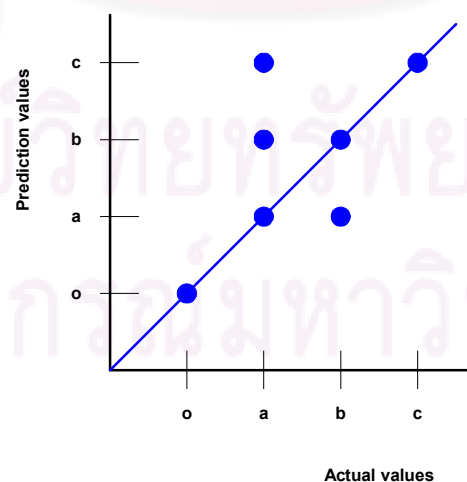


Figure 4.5: Prediction values vs. Actual values.

From figure 4.5, it can be seen that most points are line along the 45 degree diagonal line. The points below the line can be interpreted as false positive which means that the predicted values are smaller than the actual value. The points above the diagonal line imply that the predicted value is more than the actual value.

4.2 Experiment Results after applying the cross validation to a multilayer perceptron

4.2.1 Compare cost lost between C4.5 algorithm and a multilayer perceptron with the cross validation

The observed data were training set and testing set. This is called cross validation [16]. K-fold cross validation is used to measure the performance of a multilayer perceptron. A 10 fold partition of the data set was created. Split data into 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 [17].

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.

Performance Measures	C4.5	MLP	MLP (cross validation)
Correctly Classified Instances (%)	95.1977	94.9153	95.2542
MAE (mean absolute error)	0.0087	0.0092	0.0086
RMSE (root mean square error)	0.0717	0.0729	0.0693

Table 4.3: Measured Performance Results of cost lost per day.

Table 4.3 demonstrates the comparison of different performance measurements: C4.5 algorithm, multilayer perceptrons without applying cross validation and multilayer

perceptrons with cross validation. The results show that a Multilayer perceptron with cross validation has higher performance than C4.5 algorithm and multilayer perceptrons without applying the cross validation. Three different measurements are compared:

1) Correctly classified instances (%). The percentage of correctly classified instances of C4.5 algorithm, MLP without applying the cross validation and MLP with the cross validation are 95.1977, 94.9193, and 95.2542 respectively. We can see that multilayer perceptrons after applying the cross validation gives the higher percentage than C4.5 algorithm.

2) Mean absolute error. The mean absolute error of C4.5 algorithm, MLP without applying the cross validation and MLP with the cross validation are 0.0087, 0.0092 and 0.0086 respectively. It can be seen that the mean absolute error of multilayer perceptrons with the cross validation is lower than the C4.5 algorithm.

3) Root mean square error. The root mean square error of C4.5 algorithm, MLP without applying the cross validation and MLP with the cross validation are 0.0717, 0.0729 and 0.0693 respectively. We can see that the root mean square error of multilayer perceptrons with the cross validation is lower than the C4.5 algorithm.

4.2.2 Compare time lost between C4.5 algorithm and a multilayer perceptron with the cross validation

Performance Measures	C4.5	MLP	MLP (cross validation)
Correctly Classified Instances (%)	94.9153	94.6328	96.0452
MAE (mean absolute error)	0.0353	0.0299	0.0281
RMSE (root mean square error)	0.145	0.1503	0.141

Table 4.4: Measured Performance Results of time lost per day.

Table 4.4 displays the performance between C4.5, MLP without applying the cross validation and MLP with the cross validation on time lost. Three different measurements are compared:

1) Correctly classified instances (%). The percentage of correctly classified instances of C4.5 algorithm, MLP without applying the cross validation and MLP with the cross validation are 94.9153, 94.6328 and 96.0452 respectively. From the result, the performance of multilayer perceptrons with the cross validation is higher than that of C4.5.

2) Mean absolute error. The mean absolute error of C4.5 algorithm, MLP without applying the cross validation and MLP with the cross validation are 0.0353, 0.0299 and 0.0281 respectively. The MAE of multilayer perceptrons with the cross validation is lower than C4.5.

3) Root mean square error. The root mean square error of C4.5 algorithm, MLP without applying the cross validation and MLP with the cross validation are 0.145, 0.1503 and 0.141 respectively. The RMSE of multilayer perceptrons with the cross validation is lower than C4.5.



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

CHAPTER V

CONCLUSION AND FUTURE WORK

5.1 Conclusion

In this thesis, performance measurements of employees in the work place were studied and analyzed using two methods: 1) a decision trees based on a C4.5 algorithm and 2) Multilayer perceptrons. The Internet tracking software called “Track4Win” was used to capture the log files of employees in a small-sized software company which less than 50 persons [18].

1) Internet usage consumption (cost lost) by employees. The experiments were performed on WEKA. Two methodologies namely a C4.5 and a multilayer perceptron were applied to web usage log dataset of 885 instances. The result showed that a multilayer perceptron with the cross validation has a higher performance than the C4.5 algorithm based on three measurements we used.

2) Internet time consumption (time lost) by employees. A C4.5 decision tree and a multilayer perceptron have been applied to the web usage log dataset of 885 instances. The result shows that a multilayer perceptron with the cross validation has achieved a higher performance than the C4.5 algorithm based on three measurements used in the experiment.

5.2 Discussion

Due to the increase in Internet abuse, employees monitoring has become more widespread and much easier with the use of new and cheaper technologies such as Internet monitoring software. The web usage log that was captured by Internet tracking system can be used for analyzing performance measures such as Internet time usage and cost lost. Although Internet tracking system is an efficiency strategy for monitoring and analyzing the activity of employees during working hour, there are at least two issues that employers should pay more attentions [19].

1) Trust. The Internet tracking mechanism may decrease trust between employees and a manager. There is trust from an employee to his manager and there is also trust from the manager to his employee. Therefore, both entities concurrently take on the role of trustor and trustee in the working environment [1]. After employees access an inappropriate web site for their objectives, trust of managers to their employees will decrease and it also decreases the relationship between them.

2) Ethical. Both employers and employees are concerned with the ethical implications of web access monitoring. While employers use monitoring software to keep track of their employees' actions and productivity, their employees feel that too much monitoring is an invasion of their privacies. Employers want to make sure their employees are doing a good work, but employees feel that their Internet activities are exposed and being watched. This is not a healthy and ethical work environment for employees.

To ensure high ethical employee behavior, every level of management and non-management employees must fully understand the ethical implications of their decisions as it relates to their personal and professional values. Organizations need to implement a business code of ethics and review with all employees. Also, an excellent tool for learning is case studies [20]. The key in this learning is to make the code accessible and position it as a helpful tool for all employees. It is also recommended that all managers display the code on their desks in a healthy manner. Real world learning and the negative results of unethical behavior or actions should be showcased to support this venture [20].

In this experiment, although the results indicated that a multilayer perceptron without the cross validation has lower performance than C4.5 algorithm based on three measurements, it can further be improved by configuring the appropriated parameters to a multilayer perceptron, so the result of a multilayer perceptron will be increased and its performance may be better than C4.5.

5.3 Future Work

Although different data mining techniques were developed for building an Internet usage consumption model, this work can further be extended by applying the application log instead of web usage log to classify the behavior of employees. The application log can be tracked by using computer tracking software, for example: Track4Win Enterprise Edition and Spector Pro Software [21, 22].



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

References

- [1] Murugan Anandarajan and Claire A. Simmers. Personal web usage in the workplace: A guide to effective human resources management. **Information Science Publishing** (2004).
- [2] Gupta, Jatinder N D. Improving workers' productivity and reducing Internet abuse. **The Journal of Computer Information Systems** (2004): 59-65.
- [3] Fujian Liu, Yanping Zhao, Wenguang Wang and Dwight Makaroff. Database Server Workload Characterization in an E-commerce Environment. **Proceedings of the IEEE Computer Society's 12th Annual International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems** (2004).
- [4] Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques second edition. **Morgan Kaufmann** (2006).
- [5] Conlin, M. Workers, surf at your own risk. **Business Week** (June 2000): 105-106.
- [6] Griffiths, M.D. Excessive Internet use: Implications for sexual behavior. **CyberPsychology and Behavior** (2000): 537-552.
- [7] J. Ross Quinlan. C4.5: programs for machine learning. **Morgan Kaufmann** (1993).
- [8] Salvatore Ruggieri. Efficient C4.5. **IEEE Transactions on knowledge and data engineering** (April 2002).
- [9] T.S. Lim, W.Y. Loh, and Y.S. Shih. A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty - Tree Old and New Classification Algorithms. **Machine Learning** 40 (2000): 203-228.
- [10] J.R. Quinlan. Induction of Decision Trees. **Machine Learning** 1 (1996): 81-106.
- [11] Simon Haykin. Neural Networks: A Comprehensive foundation second edition. **Pearson Prentice Hall** (2005).
- [12] Ian H. Witten and Eibe Frank. Data mining: practical machine learning tools and techniques second edition. **Morgan Kaufmann** (2005).
- [13] Zhenguo Chen. Web Log Mining Based On Fuzzy Immunity Clonal Selection Neural Network. **2007 International Conference on Service Systems and Service Management** (June 2007): 1-4.

- [14] Jeffrey J. Johnson and Zsolt Ugray. Employee Internet abuse: Policy versus reality. **Issues in information Systems** (2007).
- [15] Sepama Software Co Ltd: Track4win software [on line]. Available from: <http://www.track4win> [2008, March 1].
- [16] Young, K. Psychology of computer use: XL. Addictive use of the Internet: A case that breaks the stereotype. **Psychological Reports** (1996): 899-902.
- [17] Kohavi, Ron. A study of cross-validation and bootstrap for accuracy estimation and model selection. **Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence** (1995).
- [18] Microsoft Corporation: Microsoft Solution Finder [on line]. Available from: Available from: <https://solutionfinder.microsoft.com> [2008, November 1].
- [19] Annie Wynn and Paris Trudeau: Internet Abuse at Work: Corporate Networks are paying the price [on line]. Available from: <http://www.suftcontrol.com> [2008, March 1].
- [20] Mujtaba Bahaudin. Ethical Implications of Employee Monitoring: What Leaders Should Consider. **Journal of Applied Management and Entrepreneurship** (July 2003).
- [21] SpectorSoft Corporation: Spector software [on line]. Available from: <http://www.netbus.org/computer-tracking-software.html> [2008, November 1].
- [22] Kelly Services Staffing & Recruitment (Thailand) Co Ltd: Thailand Salary Guide 2007 [on line]. Available from: <http://www.kellyservices.co.th> [2008, March 1].



APPENDICES

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

Appendices A

Track4Win Software

Track4win is application software that can monitor all computer activity and Internet use. It can automatically track visited website addresses, and log work time on each application. Its features are as following list:

1. Time tracking. It can captures and calculates how much time that employees spend on a specific application, such as Internet Explorer, Word, Photoshop, Dream weaver, JBuilder as shown in figure A.1.

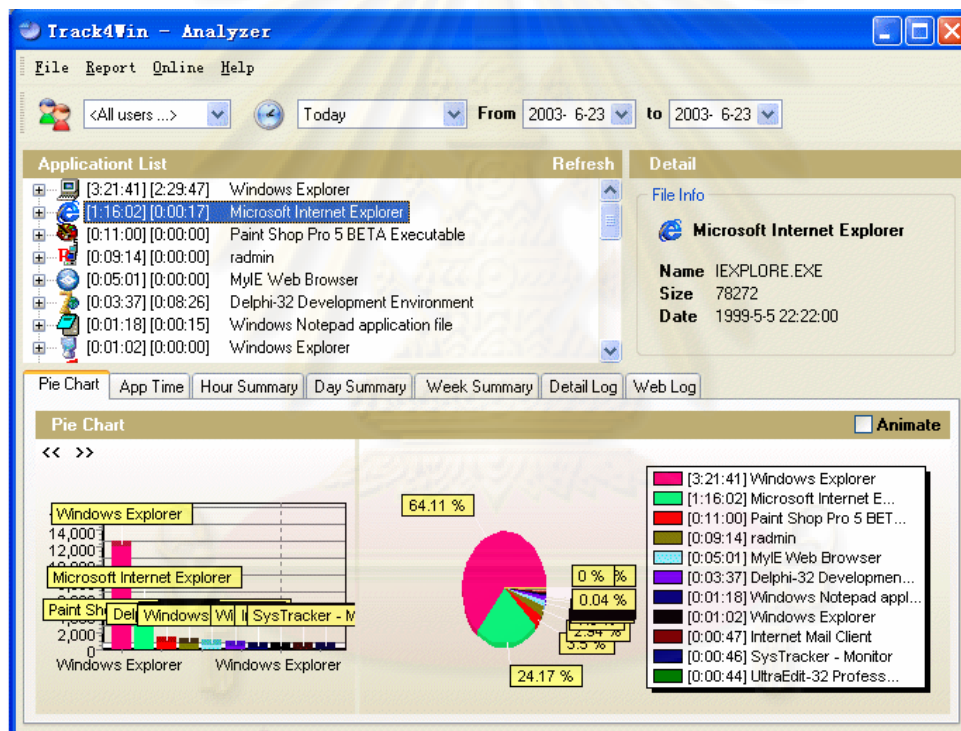


Figure A.1: Example of Track4win interface.

2. Employee monitoring and project time tracking. It monitors employee daily computer usage and asks that employee who is overdoing the personal use back into work immediately. It can estimate project time, calculates project cost, tracks work time and analyzes work process.

3. Internet and computer usage tracking. It prevents Internet abuse. It can tracks web sites Internet use, and records how much time is spent on writing e-mail, visiting

chat room and surfing Internet. Track4Win can monitor internet usage through LAN and WAN. It tracks computer use and analyze the computer usage rate in the company.

4. File Access monitoring and security. It keeps track of file related activities real time, including: 1)create 2) copy 3) modify 4) delete 5) rename and 6) move Immediately identify the file access type, the file location, the action date and time, the user, and the workstation.



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

Appendices B

Weka Software

WEKA (Waikato Environment for Knowledge Analysis) is a programming suite of machine learning software written in Java programming language and distributed under the terms of the GNU General Public License. WEKA was developed at the University of Waikato in New Zealand. It contains a uniform interface to many different learning algorithms with methods for preprocessing and for evaluating the result of learning schemes on any given dataset.

1. The WEKA Explorer.

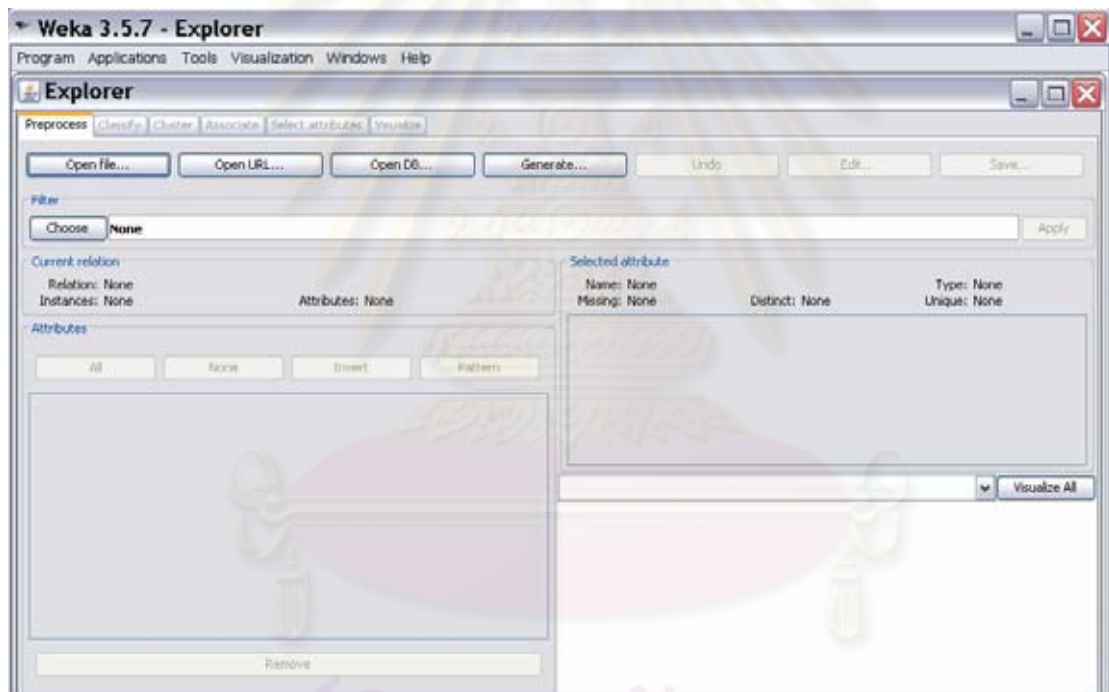


Figure B.1: The WEKA Explorer.

Figure B.1 demonstrates a row of tabs in the WEKA explorer. When the Explorer is started only the first tab is active, the others are grayed out. This is because it is necessary to open a data set before starting to explore the data. The tabs are as follows:

1. Preprocess. This tab used for Choosing and modifying the data being acted on.
2. Classify. This tab used for learning classify from the data.

3. Cluster. This tab used for learning clusters from the data.
 4. Associate. This tab used for learning association rules from the data.
 5. Select attributes. This tab used for selecting the most relevant attributes in the data.
 6. Visualize. Users can view an interactive 2D plot of the data from this tab.
2. Preprocessing.

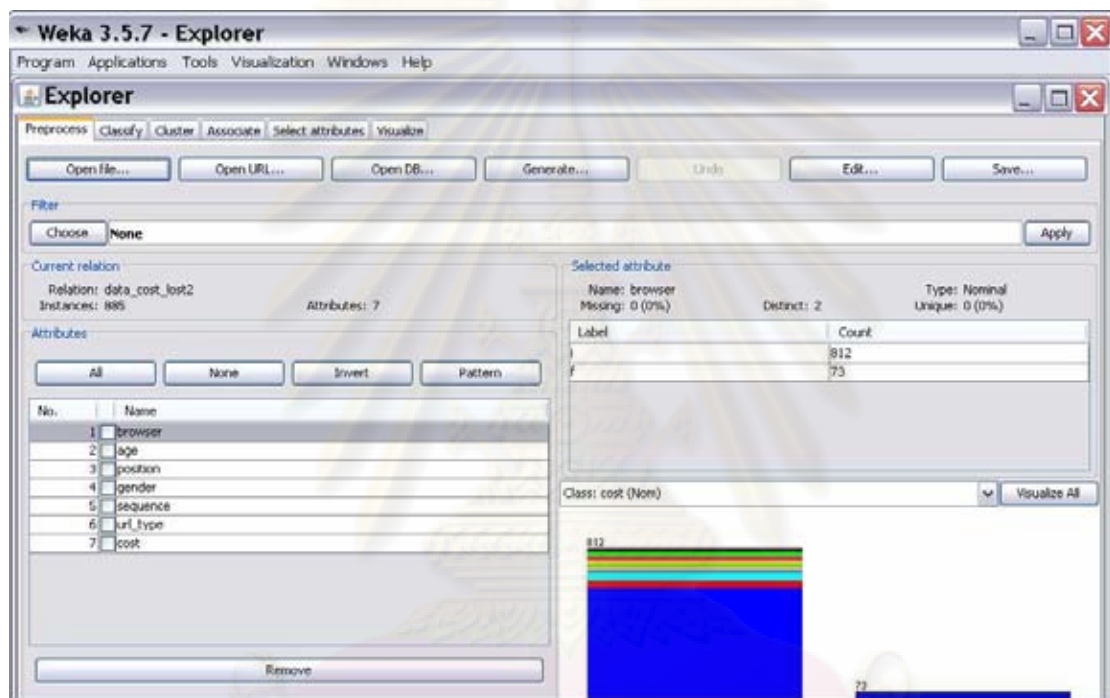


Figure B.2: The interface of Preprocessing.

2.1 Loading Data.

Figure B.2 shows the interface of preprocessing. The first four buttons at the top of the preprocess section enable users to load the data into WEKA:

1. Open file. It brings up a dialog box allowing users to browse for the data file on the local file system.
2. Open URL. It asks for a Uniform Resource Locator address for where the data is stored.
3. Open DB. It reads data from a database.
4. Generate. It enables users to generate artificial data from a variety of data generators.

Using the 'Open file' button users can read files in a variety of formats: WEKA's ARFF format, CSV format, C4.5 format, or serialized Instances format. ARFF files typically have a .arff extension, CSV files have a .csv extension, C4.5 files have a .data and .names extension, and serialized Instances objects have a .bsi extension.

2.2 The Current Relation.

Once some data has been loaded, the preprocess panel shows a variety of information. The current relation box has three entries:

1. Relation. It is the name of the relation, as given in the file it was loaded from. Filters modify the name of a relation.
2. Instances. It is the number of instances in the data.
3. Attributes. It is the number of attributes in the data.

2.3 Working with Attributes.

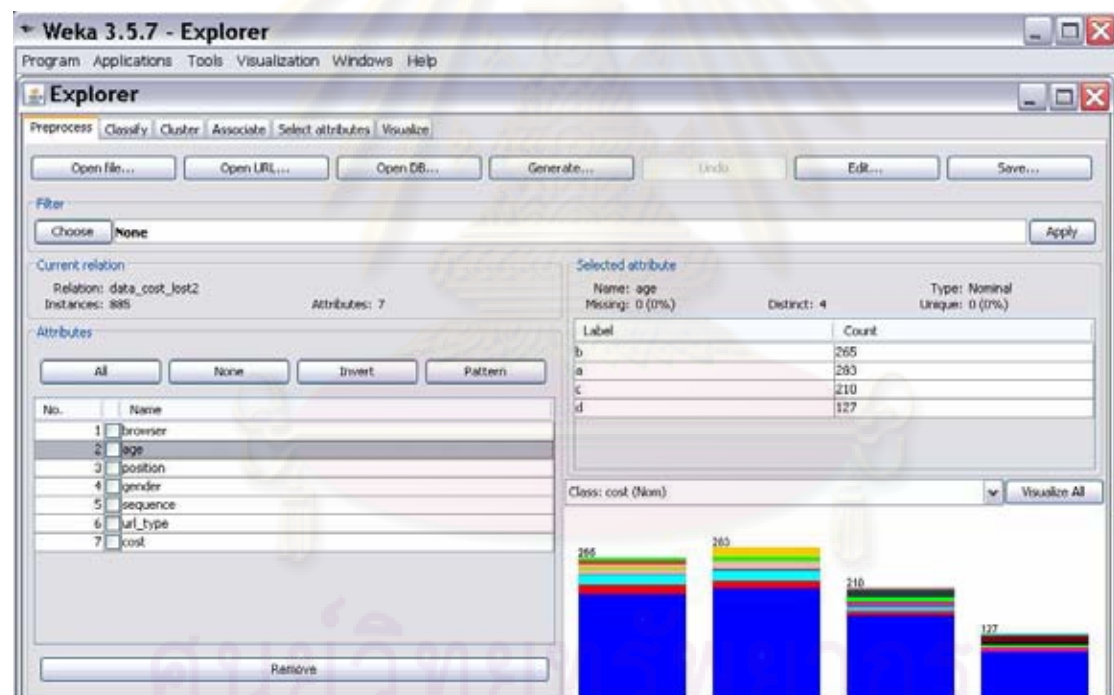


Figure B.3: Attributes Explorer.

Below the current relation box is a box titled attributes as shown in Figure B.3. There are four buttons and beneath them is a list of the attributes in the current relation.

The list has three columns:

1. No. It is a number that identifies the attribute.

2. Selection tick boxes. These components allow user select which attributes are present in the relation.
3. Name. This column shows the name of the attribute, as it was declared in the data file. When users click on different rows in the list of attributes as shown in Figure B.3, the fields change in the box to the right titled selected attribute. This box displays the characteristics of the currently highlighted attribute in the list:
 1. Name. It is the name of the attribute, the same as that given in the attribute list.
 2. Type. It is the type of attribute, most commonly nominal or numeric.
 3. Missing. It is the number of instances in the data for which this attribute is missing (unspecified).
 4. Distinct. It is the number of different values that the data contains for this attribute.
 5. Unique. It is the number of instances in the data having a value for this attribute that no other instances have.

Below these statistics is a list showing more information about the values stored in this attribute, which differ depending on its type. If the attribute is nominal, the list consists of each possible value for the attribute along with the number of instances that have that value. If the attribute is numeric, the list gives four statistics describing the distribution of values in the data: the minimum, maximum, mean and standard deviation.

From the attribute list, to begin with all the tick boxes are unchecked. They can be toggled off by clicking on them individually. The four buttons above can also be used to change the selection:

1. All. All boxes are ticked.
2. None. All boxes are cleared (unchecked).
3. Invert. Boxes that are ticked become unchecked and vice versa.
4. Pattern. It enables the user to select attributes.

Once the desired attributes have been selected, they can be removed by clicking the Remove button below the list of attributes. Note that this can be undone by clicking the Undo button, which is located next to the Edit button in the top-right corner of the Preprocess panel.

2.4 Working with Filters.

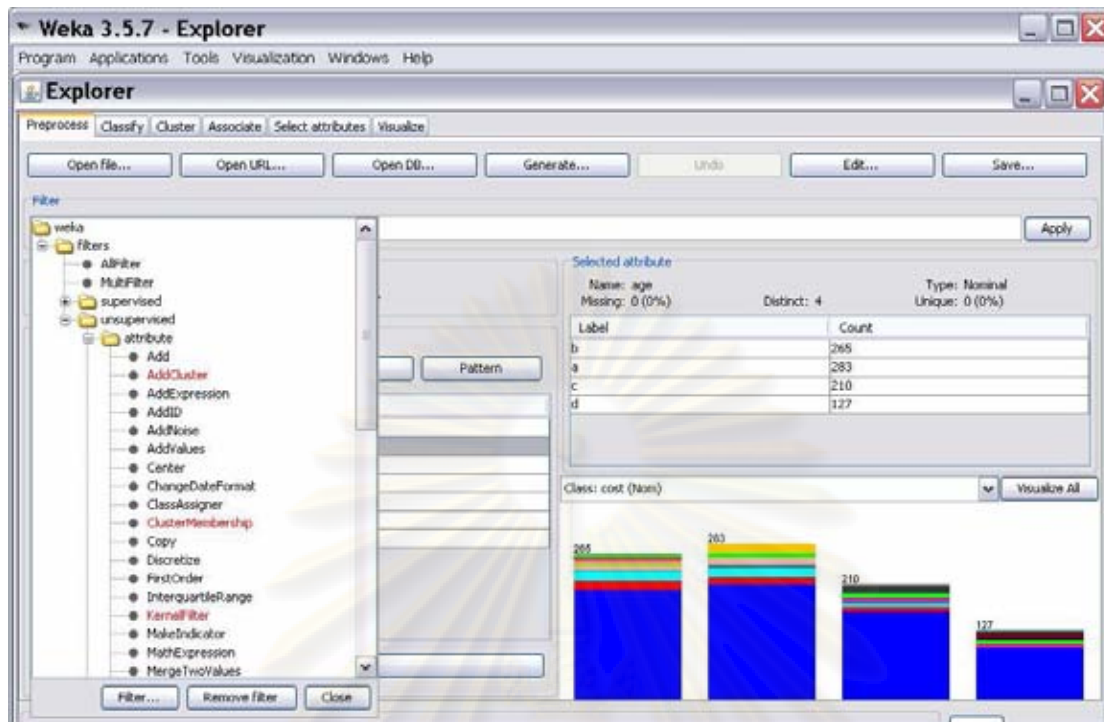


Figure B.4: The interface of preprocess.

Figure B.4 demonstrates the interface of preprocess. The filter box is used to set up the filters that are required.

At the left of the filter box is a choose button. By clicking this button, it is possible to select one of the filters in WEKA. Once a filter has been selected, its name and options are shown in the field next to the choose button. Clicking on this box with the left mouse button brings up a 'GenericObjectEditor' dialog box.

A click with the right mouse button (or Alt+Shift+left click) brings up a menu where users can choose, either to display the properties in a 'GenericObjectEditor' dialog box, or to copy the current setup string to the clipboard.

The 'GenericObjectEditor' dialog box lets users configure a filter. The same kind of dialog box is used to configure other objects, such as classifiers. The fields in the window reflect the available options. Right-clicking on this field will bring up a popup menu, listing the following options:

1. Show properties. It has the same effect as left-clicking on the field.
2. Copy configuration to clipboard.
3. Enter configuration. It is the 'receiving' end for configurations that got copied to the clipboard earlier on. In this dialog users can enter a classname followed by options (if

the class supports these). This also allows users to transfer a filter setting from the preprocess panel to a FilteredClassifier used in the classify panel.

Once users have selected and configured a filter, they can apply it to the data by pressing the apply button at the right end of the Filter panel in the Preprocess panel. The Preprocess panel will then show the transformed data. The change can be undone by pressing the Undo button. Users can also use the 'Edit' button to modify their data manually in a dataset editor. Finally, the 'Save' button at the top right of the Preprocess panel saves the current version of the relation in file formats that can represent the relation, allowing it to be kept for future use.

3. Classification.

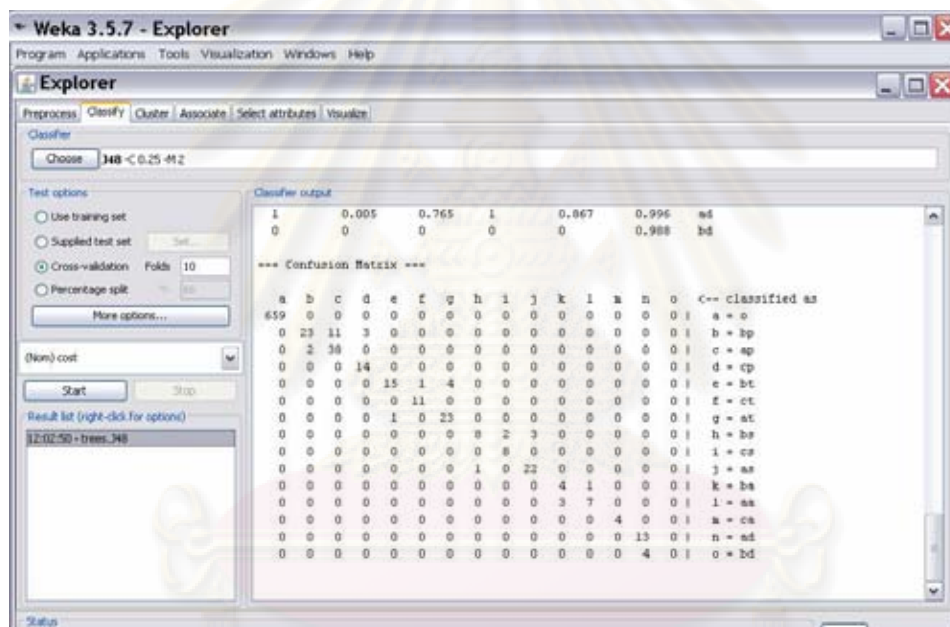


Figure B.5: The classification explorer.

3.1 Selecting a Classifier.

Figure B.5 shows the classification explorer. At the top of the classify section is the classifier box. This box has a text field that gives the name of the currently selected classifier, and its options. Clicking on the text box with the left mouse button brings up a GenericObjectEditor dialog box, just the same as for filters that users can use to configure the options of the current classifier. With a right click (or Alt+Shift+left click) users can once again copy the setup string to the clipboard or display the properties in a GenericObjectEditor dialog box. The Choose button allows users to choose one of the classifiers that are available in WEKA.

3.2 Test Options.

The result of applying the chosen classifier will be tested according to the options that are set by clicking in the test options box. There are four test modes:

1. Use training set. The classifier is evaluated on how well it predicts the class of the instances it was trained on.
2. Supplied test set. The classifier is evaluated on how well it predicts the class of a set of instances loaded from a file. Clicking the 'Set' button brings up a dialog allowing users to choose the file to test on.
3. Cross validation. The classifier is evaluated by the cross validation, using the number of folds that are entered in the folds text field.
4. Percentage split. The classifier is evaluated on how well it predicts a certain percentage of the data which is held out for testing. The amount of data held out depends on the value entered in the % field.

The model that is output is always the one build from all the training data. Further testing options can be set by clicking on the 'More options' button:

1. Output model. The classification model on the full training set is output so that it can be viewed, visualized, etc. This option is selected by default.
2. Output per class stats. This option is also selected by default.
3. Output entropy evaluation measures. Entropy evaluation measures are included in the output. This option is not selected by default.
4. Output confusion matrix. The confusion matrix of the classifier's predictions is included in the output. This option is selected by default.
5. Store predictions for visualization. The classifier's predictions are remembered so that they can be visualized. This option is selected by default.
6. Output predictions. The predictions on the evaluation data are output.
7. Output additional attributes. If additional attributes need to be output alongside the predictions, for example: an ID attribute for tracking misclassifications, then the index of this attribute can be specified here.
8. Cost sensitive evaluation. The errors are evaluated with respect to a cost matrix. The 'Set' button allows users to specify the cost matrix used.

9. Random seed for 'xval / % Split'. This specifies the random seed used when randomizing the data before it is divided up for evaluation purposes.
10. Preserve order for % Split. This suppresses the randomization of the data before splitting into train and test set.
11. Output source code. If the classifier can output the built model as Java source code, users can specify the class name here. The code will be printed in the "Classifier output" area.

3.3 The Class Attribute.

The classifiers in WEKA are designed to be trained to predict a single 'class' attribute, which is the target for prediction. Some classifiers can only learn nominal classes. By default, the class is taken to be the last attribute in the data. Users can click on the box below the 'test options' box to open a drop-down list of attributes.

3.4 Training a Classifier.

The learning process is started by clicking on the 'Start' button. Users can stop the training process at any time by clicking on the 'Stop' button.

When training is complete, the Classifier output area to the right of the display is filled with text describing the results of training and testing. A new entry appears in the Result list box.

3.5 The Classifier Output Text.

The text in the Classifier output area has scroll bars allowing users to browse the results. The output contains several sections:

1. Run information. A list of information giving the learning scheme options, relation name, instances, attributes and test mode that were involved in the process.
2. Classifier model (full training set). It is a textual representation of the classification model that was produced on the full training data.
3. Summary. It is a list of statistics summarizing how accurately the classifier was able to predict the true class of the instances under the chosen test mode.
4. Detailed Accuracy By Class. It shows detailed per-class break down of the classifier's prediction accuracy.

5. Confusion Matrix. It shows how many instances have been assigned to each class. Elements show the number of test examples whose actual class is the row and whose predicted class is the column.

6. Source code. This section lists the Java source code if users chose 'Output source code' in the 'More options' dialog.

3.6 The Result List.

After training several classifiers, the result list will contain several entries. Right-clicking an entry invokes a menu containing these items:

1. View in main window. It shows the output in the main window.
2. View in separate window. It opens a new independent window for viewing the results.
3. Save result buffer. It opens a dialog that allows users to save a text file containing the textual output.
4. Load model. It loads a pre-trained model object from a binary file.
5. Save model. It saves a model object to a binary file. Objects save in Java 'serialized object' form.
6. Re-evaluate model on current test set. It takes the model that has been built and tests its performance on the data set that has been specified with the 'Set' button under the supplied test set option.
7. Visualize classifier errors. It opens a visualization window that plots the results of classification. Correctly classified instances are represented by crosses, whereas incorrectly classified ones show up as squares.
8. Visualize tree. It opens a graphical representation of the structure of the classifier model. The graph visualization option only appears if a Bayesian network classifier has been built. In the tree visualizer, users can bring up a menu by right-clicking a blank area, pan around by dragging the mouse, and see the training instances at each node by clicking on it.
9. Visualize margin curve. It generates a plot illustrating the prediction margin. The margin is defined as the difference between the probability predicted for the actual class and the highest probability predicted for the other classes.
10. Visualize threshold curve. Generates a plot illustrating the trade-offs in predictions that are obtained by varying the threshold value between classes. For example, with the

default threshold value of 0.5, the predicted probability of 'positive' must be greater than 0.5 for the instance to be predicted as 'positive'. The plot can be used to visualize the precision/recall trade-off, for ROC curve analysis and for other types of curves.

11. Visualize cost curve. It generates a plot that gives an explicit representation of the expected cost.

12. Plug-in. This menu item only appears if there are visualization plug-in available.

4. Visualizing.

4.1 The scatter plot matrix.

When users select the visualize panel, it shows a scatter plot matrix for all the attributes, color coded according to the currently selected class. It is possible to change the size of each individual 2D plot and the point size, and to randomly the data (to uncover obscured points). It also possible to change the attribute used to color the plots, to select only a subset of attributes for inclusion in the scatter plot matrix, and to sub sample the data.

4.2 Selecting Instances.

There may be situations where it is helpful to select a subset of the data using the visualization tool. A special case of this is the User Classifier in the Classify panel, which lets users build their own classifier by interactively selecting instances. Below the y-axis selector button is a drop-down list button for choosing a selection method. A group of data points can be selected in four methods:

1. Select Instance. Clicking on an individual data point brings up a window listing its attributes. If more than one point appears at the same location, more than one set of attributes is shown.

2. Rectangle. Users can create a rectangle, by dragging, that selects the points inside it.

3. Polygon. Users can build a free-form polygon that selects the points inside it. The

Polygon will always be closed off by connecting the first point to the last.

4. Polyline. Users can build a Polyline that distinguishes the points on one side from those on the other. The resulting shape is open (as opposed to a polygon, which is always closed).

Once an area of the plot has been selected using Rectangle, Polygon or Polyline, it turns grey. At this point, clicking the 'Submit' button removes all instances from the plot except those within the grey selection area. Clicking on the 'Clear' button erases the selected area without affecting the graph.



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

Vitae

Boonyavee Boonyamanop was born in February 23, 1983, in Bangkok. He obtained his Bachelor's degree in Electrical Engineering from the Faculty of Engineer, Thammasat University in 2005.



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย