INTELLIGENT WEB PROXY CACHING BASED ON SUPERVISED MACHINE LEARNING

WALEED ALI AHMED

UNIVERSITI TEKNOLOGI MALAYSIA

INTELLIGENT WEB PROXY CACHING BASED ON SUPERVISED MACHINE LEARNING

WALEED ALI AHMED

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

AUGUST 2012

DEDICATION

To my beloved parents, brothers and my sisters

ACKNOWLEDGMENTS

All praise and thanks are due to Allah, and peace and blessings of Allah be upon our prophet, Muhammad and upon all his family and companions .Thanks to Allah who give me good health in my life and thanks to Allah for everything. Without help of Allah, I was not able to achieve anything in this research.

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main supervisor, Prof. Dr. Siti Mariyam Shamsuddin, for encouragement, guidance, critics, advices and supports to complete this research. I really appreciate her ethics and great deal of respect with her students, which is similar to dealing between the mother, and her sons and daughters in the same family. I am also grateful to my co-supervisor Prof. Dr. Abdul Samad Ismail for his precious advices and comments.

In addition, I am extremely grateful to my brother Dr. Adel Ali for unlimited support and encouragement during this research. My sincere appreciation also extends to Soft Computing Research Group (SCRG) and all my colleagues for the support and incisive comments in making this study a success. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space.

ABSTRACT

Web proxy caching is one of the most successful solutions for improving the performance of web-based systems. In web proxy caching, the popular web objects that are likely to be revisited in the near future are stored on the proxy server, which plays the key roles between users and web sites by reducing the response time of user requests and saving the network bandwidth. However, the difficulty in determining the significant web objects that would be re-visited in the future is still a problem faced by the existing conventional web proxy caching techniques. In this study, three popular supervised machine learning techniques were used to enhance the performances of conventional web proxy caching policies: Least-Recently-Used (LRU), Greedy-Dual-Size (GDS), Greedy-Dual-Size-Frequency (GDSF) and Least-Frequently-Used-Dynamic-Aging (LFU-DA). A support vector machine (SVM), a naïve Bayes classifier (NB) and a decision tree (C4.5) were trained from web proxy logs files to predict the class of objects that would be re-visited. More significantly, the trained SVM, NB and C4.5 classifiers were intelligently incorporated with the conventional web proxy caching techniques to form novel intelligent caching approaches known as intelligent LRU, GDS, GDSF and DA approaches. For testing and evaluating the proposed proxy caching methods, the proxy logs files were obtained from several proxy servers located around the United States of the IRCache network, which are the most common proxy datasets used in the research of web proxy caching. The experimental results showed that SVM, NB and C4.5 achieved a better accuracy and a much faster than back-propagation neural network (BPNN) and adaptive neuro-fuzzy inference system (ANFIS). Furthermore, the proposed intelligent caching approaches were evaluated by trace-driven simulation and compared with the most relevant web proxy caching policies. The simulation results revealed that the proposed intelligent web proxy caching approaches substantially improved the performance in terms of hit ratio and byte hit ratio of the conventional techniques on a range of datasets. The average improvement ratios of hit ratio achieved by intelligent LRU, GDS, and DA approaches over LRU, GDS and LFU-DA increased by 32.60 %, 22.45 % and 35.458 %, respectively. In terms of byte hit ratio, the average improvement ratios achieved by intelligent LRU, GDS, GDSF, and DA approaches over LRU, GDS, GDSF and LFU-DA increased by 69.56 %, 229.14 %, 407.49 % and 69.074 %, respectively.

ABSTRAK

Cache proksi sesawang adalah salah satu penyelesaian yang paling berjaya untuk menambah baik prestasi sistem berasaskan sesawang. Dalam cache proksi sesawang, objek popular sesawang yang berkemungkinan dikunjungi semula dalam masa terdekat akan disimpan dalam pelayan proksi. Pelayan proksi memainkan peranan utama antara pengguna dengan sesawang bagi memendekkan masa tindak balas kepada permintaan pengguna dan menjimatkan rangkaian jalurlebar. Walau bagaimanapun kesukaran dalam menentukan objek sesawang yang ideal untuk dilawati semula pada masa hadapan masih menimbulkan masalah yang sering dihadapi oleh cache proksi sesawang dengan kaedah konvensional. Dalam kajian ini tiga teknik pembelajaran mesin yang terkenal digunakan untuk meningkatkan prestasi polisi cache proksi sesawang tradisi, iaitu *Least-Recently-Used* (LRU), *Greedy-Dual-Size* (GDS), *Greedy-Dual-Size-Frequency* (GDSF) dan *Least-Frequently-Used-Dynamic-Aging* (LFU-DA). Mesin Sokongan Vektor (MSV), Pengelas *Naive Bayes* (NB) dan Pepohon Keputusan (C4.5) dilatih daripada fail log proksi sesawang untuk meramal kelas objek yang akan dikunjungi semula. Lebih penting lagi pengelas MVS terlatih, NB dan C4.5 digabungkan dengan kaedah cache proksi sesawang tradisi untuk membentuk pendekatan cache novel pintar yang dikenali sebagai pendekatan LRU, GDS, GDSF dan DA pintar. Untuk menguji dan menilai kaedah cache proksi sesawang yang dicadangkan, fail log proksi diperoleh daripada beberapa pelayan proksi yang terletak di sekitar Amerika Syarikat melalui rangkaian IRCache, iaitu set data proksi yang paling lazim digunakan dalam kajian cache proksi sesawang. Keputusan ujikaji menunjukkan bahawa pengelas MVS, NB dan C4.5 mencapai ketepatan yang lebih baik daripada Perambatan Balik Rangkaian Neural (BPNN) dan Sistem Penyesuaian Taakulan Neuro-Fuzzy (ANFIS). Di samping itu pendekatan cache pintar yang dicadangkan telah dinilai oleh simulasi surih berpacu dan dibandingkan dengan polisi cache proksi sesawang yang paling relevan. Keputusan simulasi telah menunjukkan pendekatan cache proksi sesawang pintar yang dicadangkan telah menambah baik prestasi julat set data teknik konvensional berdasarkan nisbah capaian dan nisbah bait capaian. Peningkatan nisbah purata nisbah capaian yang dicapai oleh LRU dan GDS pintar dan pendekatan DA telah meningkat kepada 32.60%, 22.45% dan 35.458% setiap satunya. Dari segi nisbah bait capaian purata peningkatan yang dicapai oleh LRU, GDS dan GDSF pintar dan pendekatan DA telah meningkat kepada 69.56%, 229.14%, 407.49% dan 69.074% setiap satunya berbanding dengan kaedah LRU, GDS, GDSF dan LFU-DA.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	-	Artificial Intelligence
ANFIS	-	Adaptive Neuro-Fuzzy Inference System
ANN	-	Artificial Neural Network
AUC	-	Area Under ROC Curve
BHR	-	Byte Hit Ratio
BPNN	-	Back-propagation Neural Network
BPNNPCR	-	Back-propagation Neural Network Proxy Cache Replacement
BU	-	Boston University
C4.5	-	Decision Tree
C4.5-DA	-	Decision Tree-Dynamic-Aging
C4.5-GDS	-	Decision Tree-Greedy-Dual-Size
C4.5-GDSF	-	Decision Tree-Greedy-Dual-Size-Frequency
C4.5-LRU	-	Decision Tree-Least-Recently-Used
CCR	-	Correct Classification Ratio
CCR	-	Correct Classification Rate
CDN	-	Content Delivery Network
CM	-	Common Method
CPT	-	Conditional Probability Table
CPU	-	Central Processing Unit
DA	-	Dynamic-Aging
ErrR	-	Error Rate
FIFO	-	First-In-First-Out
FIS	-	Fuzzy Inference System
FN	-	False Negative
FNR	-	False Negative Rate
FP	-	False Positive
FPR	-	False Positive Rate

FS	-	Fuzzy System
FUNET	-	Finnish University and Research Network
GB	-	Gigabyte
GDM	-	Gradient Descent with Momentum
GDS	-	Greedy-Dual-Size
GDSF	-	Greedy-Dual-Size-Frequency
GM	-	Geometric Mean
GUI	-	Graphical User Interface
HR	-	Hit Ratio
HTML	-	Hypertext Markup Language
HTTP	-	Hypertext Transfer Protocol
ICP	-	Inter-Cache Protocol
ICWCS	-	Intelligent Client-side Web Caching Scheme
IR	-	Improvement Ratio
ISP	-	Internet Service Provider
LFU	-	Least Frequently Used
LFU-DA	-	Least-Frequently-Used-Dynamic-Aging
LR	-	Logistic Regression
LRU	-	Least Recently Used
LRV	-	Lowest Relative Value
LSR	-	Latency Saving Ratio
MB	-	Megabyte
MDL	-	Minimum Description Length
MF	-	Member Function
ML	-	Machine Learning
MLP	-	Multilayer Perceptron Network
MLR	-	Multinomial Logistic Regression
MSE	-	Mean Square Error
NB	-	Naïve Bayes
NB-DA	-	Naïve Bayes-Dynamic-Aging
NB-GDS	-	Naïve Bayes-Greedy-Dual-Size
NB-GDSF	-	Naïve Bayes-Greedy-Dual-Size-Frequency
NB-LRU	-	Naïve Bayes-Least-Recently-Used

NLANR	-	National Laboratory of Applied Network Research
NNPCR	-	Neural Network Proxy Cache Replacement
P	-	Precision
PM	-	Proposed Method
PSO	-	Particle Swarm Optimization
PV	-	Priority Value
RBF	-	Radial Basis Function
RMSE	-	Root Mean Square Error
SIZE	-	SIZE Algorithm
SV	-	Support Vectors
SVM	-	Support Vector Machine
SVM-DA	-	Support Vector Machine-Dynamic-Aging
SVM-GDS	-	Support Vector Machine-Greedy-Dual-Size
SVM-GDSF	-	Support Vector Machine-Greedy-Dual-Size-Frequency
SVM-LRU	-	Support Vector Machine-Least-Recently-Used
SWL	-	Sliding Window Length
TN	-	True Negative
TNR	-	True Negative Rate
TP	-	True Positive
TPR	-	True Positive Rate
TTL	-	Time–To-Live
URL	-	Uniform Resource Locator
WEKA	-	Waikato Environment for Knowledge Analysis
WWW	-	World Wide Web

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The World Wide Web (Web) is the most common and significant service on the Internet. The Web contributes greatly to our life in many fields such as education, entertainment, Internet banking, remote shopping and software downloading. This has led to rapid growth in the number of Internet users, which resulting in an explosive increase in traffic or bottleneck over the Internet performance (Kumar, 2009; Kumar and Norris, 2008; Patil and Pawar, 2011; Saha *et al.* , 2012; Soonthornsutee and Luenam, 2012). Consequently, this has resulted in problems during surfing some popular web sites; for instance, server denials, and greater latency for retrieving and loading data on the browsers (Kaya *et al.* , 2009; Patil and Pawar, 2011; Romano and ElAarag, 2012; Saha *et al.* , 2012; Soonthornsutee and Luenam, 2012) .

Since the early 1990's, many researchers have been working on improving Web performance. Currently, there are several techniques available at both hardware and software levels. The hardware-based solutions improve Internet bandwidth and device latencies but the user's expectations never end. Moreover, the hardware-based solution is not always the best solution due to the cost of hardware and other issues (location, network infrastructure, environment, and others) (Acharjee, 2006; Romano and ElAarag, 2012; Zeng et al. , 2011).

The most popular solution for improving Web performance is a web caching technology (Hu and Ding, 2010; Kaya *et al.* , 2009; Kumar, 2009; Kumar and Norris, 2008; Patil and Pawar, 2011; Saha *et al.* , 2012; Soonthornsutee and Luenam, 2012). The web caching technique is a very useful mechanism in reducing network bandwidth utilization, decreasing user-perceived delays, and reducing loads on the original servers.

1.2 Problem Background

Web caching is a well-known strategy for improving the performance of web-based system. Web objects that are likely to be used in the near future are kept in a location closer to the user. Web caching mechanisms are implemented at three levels: client level, proxy level and original server level (Chen, 2008; Chen, 2007). Proxy servers play key roles between users and web sites in reducing the response time of user requests and saving network bandwidth. In this study, much emphasis is focused on web proxy caching because it is still the most common strategy used for caching web pages (Kaya *et al.* , 2009; Kumar, 2009; Kumar and Norris, 2008; Romano and ElAarag, 2012; Romano and ElAarag, 2011, Sajeev and Sebastian, 2011).

Due to cache space limitations, an intelligent mechanism is required to manage the web cache contents efficiently. Cache replacement is the core or heart of web caching; hence, the design of efficient cache replacement algorithms is crucial for the success of caching mechanisms (Chen, 2007; Romano and ElAarag, 2012; Romano and ElAarag, 2011, Sajeev and Sebastian, 2011). In the proxy cache replacement, the proxy cache must effectively decide which objects are worth caching or replacing with other objects. The cache replacement algorithms are also known as web caching algorithms (Koskela *et al.* , 2003).

Most of the conventional web caching policies are not efficient enough in web caching. Romano and ElAarag (2011) stated that *"essentially, most strategies that are used in proxy cache software such as Squid are no longer ''good-enough''* *strategies today"*. This is because the conventional web caching approaches consider just one factor and ignore other factors that have an impact on the efficiency of the web caching (Ayani *et al.* , 2003; Cobb and ElAarag, 2008; Koskela *et al.* , 2003; Romano and ElAarag, 2012). In these caching policies, the most popular objects get the most requests, while a large portion of objects, which are stored in the cache, are never requested again. This is known as cache pollution problem.

In fact, a few important features or factors of web objects, such as recency, frequency, size, cost of fetching the object from its origin server and access latency of object, can influence the performance of web proxy caching (Chen, 2008; Kin-Yeung, 2006; Podlipnig and Böszörmenyi, 2003; Romano and ElAarag, 2012; Vakali, 2002). These factors can be incorporated into the replacement decision for better performance. Depending on these factors, web proxy policies can be classified into five categories: recency-based policies, frequency-based policies, size-based policies, function-based policies and randomized policies (Podlipnig and Böszörmenyi, 2003). Most of these methods use one or more of these factors for making decisions about caching. However, a combination of these factors to get wise replacement decision is not a simple task, because one factor in a particular environment may be more important in other environments (Chen, 2008; Kin-Yeung, 2006). Hence, there is a need for an effective and adaptive approach, which can effectively incorporate the significant factors into web caching decisions.

Several research works have developed intelligent approaches that are smart and adaptive to the web caching environment. These include adoption of supervised machine learning techniques (Ali and Shamsuddin, 2009a; Cobb and ElAarag, 2008; Farhan, 2007; Koskela *et al.* , 2003; Romano and ElAarag, 2011; Sulaiman *et al.* , 2011), fuzzy systems (Calzaross and Vall, 2003), and evolutionary algorithms (Tirdad *et al.* , 2009; Vakali, 2002; Yan *et al.* , 2004) in web caching, especially in web cache replacement. Availability of web proxy logs files that can be exploited as training data is the main motivation in adopting intelligent web caching approaches. In a web proxy server, web proxy logs file records activities of the users and can be considered as complete and prior knowledge of future access. The second motivation behind the development of an intelligent approach is the need for an efficient and adaptive web proxy caching approach based on web users' interests, which change and update continuously.

In recent years, several studies have proposed exploiting intelligent supervised machine learning techniques to cope with web caching problem (Ali and Shamsuddin, 2009a; Cobb and ElAarag, 2008; Farhan, 2007; Koskela *et al.* , 2003; Romano and ElAarag, 2011; Sajeev and Sebastian, 2011; Sulaiman *et al.* , 2008; Sulaiman *et al.* , 2011). Although the intelligent web caching approaches based on the supervised machine learning techniques can contribute in improving the performance of web caching, these approaches still have some limitations. Most of these studies utilize an artificial neural network (ANN) in web proxy caching although ANN training may consume more time and require extra computational overhead. More significantly, integration of an intelligent technique in web cache replacement is still a popular research subject.

Therefore, in this study, alternative supervised machine learning techniques are proposed for improving the performance of web proxy caching. Support vector machine (SVM), Naïve Bayes (NB) and decision tree (C4.5) are three popular supervised learning algorithms, which have been identified as three of the most influential algorithms in data mining (Wu *et al.* , 2008).

Support vector machine (SVM) is one of the most popular supervised learning algorithms, performing classifications faster and more accurately than most other algorithms in a wide range of applications such as text classification, web page classification, remote sensing, bioinformatics and medical applications (Chen and Hsieh, 2006; Liu, 2007; Mountrakis *et al.* , 2011, Sebastiani, 2002; Temko *et al.* , 2011, Yu *et al.* , 2010).

Bayesian networks are popular supervised learning algorithms that have great popularity in the medical field and other applications such as military, forecasting, control, modeling for human understanding, cognitive science, statistics, and philosophy (Bai, 2005; Darwiche, 2010; de Melo and Sanchez, 2008; Friedman *et al.*

, 1997; Goubanova and King, 2008; Lucas, 2001; Oliveira *et al.* , 2004; Van Koten and Gray, 2006). Naïve Bayes classifier is a simple Bayesian network classifier, which has been applied successfully in many domains. Despite the simplicity of the Naïve Bayes classifier and the restrictiveness of the independent assumptions among features, it is more effective compared with other more sophisticated classifiers (Hall, 2007). Therefore, it is not surprising that Naïve Bayes classifier has gained popularity in solving various classification problems (Fan *et al.* , 2009; Hall, 2007; Lu *et al.* , 2010).

Decision tree (C4.5) is also one of the most widely used and practical techniques for classification in many applications such as finance, marketing, engineering and medicine (Han and Kamber, 2001; Liu, 2007; Rokach and Maimon, 2008). The decision tree has several advantages. It is simple to understand and interpret. Besides, it is able to handle nominal and categorical data and performs well with large datasets in a short time (Huang *et al.* , 2011).

Since SVM, NB and C4.5 have been used successfully in a wide range of applications, they can be utilized to produce promising solutions for web proxy caching. In this study, we present new approaches that depend on the capability of SVM, NB and C4.5 to learn from proxy logs files and predict the class of objects that would be re-visited. The intelligent trained classifiers are utilized to improve the performance of web proxy cache replacement. In this study, the intelligent trained classifiers are incorporated effectively with traditional web proxy caching algorithm to present novel intelligent web proxy caching approaches with better performance in terms of hit ratio and byte hit ratio.

1.3 Problem Statement

Since the space apportioned to the cache is limited, the space must be utilized judiciously. The most common web proxy caching methods are not efficient enough and may suffer from a cache pollution problem since they consider just one factor and ignore other factors that have an impact on the efficiency of the web proxy caching (Cobb and ElAarag, 2008; Kaya *et al.*, 2009; Koskela *et al.* , 2003; Romano and ElAarag, 2011; Romano and ElAarag, 2012, Sajeev and Sebastian, 2011). Cache pollution means that a cache contains objects that are not frequently visited. This causes a reduction of the effective cache size and negatively affects the performance of web proxy caching. In other words, which web objects should be cached or replaced in order to make the best use of available cache space, improve hit rates, reduce network traffic, and alleviate loads on the original server (Chen, 2008; Cobb and ElAarag, 2008; Kaya *et al.* , 2009; Koskela *et al.* , 2003; Kumar and Norris, 2008; Romano and ElAarag, 2011; Romano and ElAarag, 2012).

Many research works have been proposed to resolve web caching problems. However, it is challenging to have an omnipotent policy that performs well in all environments or for all time due to the difficult combination of factors that can influence the performance of web proxy caching (Chen, 2008; Kin-Yeung, 2006; Romano and ElAarag, 2011; Sajeev and Sebastian, 2011). This is motivation to adopt intelligent techniques for solving web proxy caching problems**.**

Recent studies have shown that the intelligent web caching approaches are more efficient and adaptive to web caching environments compared to other approaches (Ali and Shamsuddin, 2009a; Cobb and ElAarag, 2008; Farhan, 2007; Koskela *et al.* , 2003; Romano and ElAarag, 2011; Sajeev and Sebastian, 2011; Sulaiman *et al.* , 2008; Sulaiman *et al.* , 2011). In the intelligent web caching approaches, ANN has been widely integrated in Least-Recently-Used (LRU) caching policy although ANN training may consume a considerable amount of time and require extra computational overheads. Moreover, employment of ANN in cache replacement decisions was not effective enough since they did not take into account the cost and size in replacement decisions. So far, the difficulty in determining which ideal web objects will be re-visited is still a major challenge faced by the existing web proxy caching techniques (Chen, 2008; Cobb and ElAarag, 2008; Kaya *et al.* , 2009; Koskela *et al.* , 2003; Kumar and Norris, 2008; Romano and ElAarag, 2011; Sajeev and Sebastian, 2011). More importantly, the integration of intelligent techniques in web cache replacement is still being researched.

1.4 Research Question

In order to overcome web proxy caching challenges, intelligent web proxy caching approaches based on popular machine learning algorithms are proposed in this research. SVM, NB and C4.5 learn from web proxy logs file to efficiently predict the ideal web objects that would be re-visited later. Consequently, the trained SVM, NB and C4.5 classifiers are effectively employed in web proxy caching policies to guide the cache replacement decision. Therefore, the main research question is:

> *How can the performance of web proxy caching be enhanced using supervised machine learning techniques?*

To answer the main research question stated above, the issues that need to be addressed in this study are as follows:

i. What are the most effective intelligent supervised machine learning techniques that can enhance performance of web proxy caching?
ii. How can the most effective intelligent machine learning techniques contribute in enhancing the performance of web proxy caching?
iii. How can the most effective intelligent machine learning techniques be integrated effectively into web proxy caching?
iv. How efficient are the proposed intelligent web proxy approaches compared to other works?

1.5 Research Aim

This research aims to enhance the performance of web proxy caching through intelligent web proxy caching approaches based on supervised machine learning techniques.

1.6 Research Objectives

In order to achieve the aim of the study, the objectives of this research are stated as follows:

i. To develop an intelligent approach based on SVM, NB and C4.5 classifiers for predicting the significant web objects demanded for web proxy caching.
ii. To develop new intelligent web proxy caching approaches based on SVM, NB and C4.5 for improving the performance of web proxy caching. This includes developing new intelligent Least-Recently-Used approaches, intelligent Greedy-Dual-Size approaches, intelligent Greedy-Dual-Size-Frequency approaches, and intelligent Dynamic Aging approaches.
iii. To evaluate, validate and compare the performance of the proposed intelligent web proxy caching approaches with the most common and relevant techniques.

1.7 Research Scope and Assumptions

In order to achieve the research objectives, the scope and assumptions of the study are stated as follows:

i. Web caching is applied on web proxy server.
ii. Data of the proxy traces used for testing and evaluating the proposed approaches are obtained from five proxy servers of the IRCache network located around the United States for a period of fifteen days (NLANR, 2010a).
iii. Hit ratio (HR) and byte hit ratio (BHR) are used to evaluate the performances of intelligent web proxy caching approaches since HR and BHR are two widely used metrics for evaluating the performance of web proxy caching policies (Ali and Shamsuddin, 2009a; Cobb and ElAarag,

2008; Kin-Yeung, 2006; Koskela *et al.* , 2003; Romano and ElAarag, 2011; Romano and ElAarag, 2012).

iv. Like several previous research works (Fernández *et al.* , 2009; Fernández *et al.* , 2008; Sajeev and Sebastian, 2011), correct classification rate (CCR), true positive rate (TPR), true negative rate (TNR), and geometric mean (GM) are used to evaluate the performance of machine learning techniques

v. MySQL Database is used to prepare the training datasets of the proxy logs files.

vi. MATLAB and WEKA are exploited for training the supervised machine learning algorithms.

vii. WebTraff trace-driven simulator (Markatchev and Williamson, 2002) is modified and used for evaluating the proposed intelligent web proxy caching approaches.

viii. It is assumed that an object that has the same URL but different size is the updated version of such object, as widely assumed by other researchers in web proxy caching (Abhari *et al.* , 2006; Cobb and ElAarag, 2008; Foong *et al.* , 1999; Romano and ElAarag, 2011).

1.8 Research Significance

Proxy servers play key roles in reducing the response time of user requests and saving the network bandwidth utilization since proxy servers are located between users and web servers. The proposed intelligent web proxy caching approaches can contribute to improving the performance of web proxy caching, web-based systems and Internet as follows**.**

Firstly, after training the SVM, NB and C4.5 as proposed in this study, the trained SVM, NB and C4.5 can effectively predict significant web objects. In addition to the utilization of the trained classifiers in the web cache replacement in this research, the trained classifiers can be utilized to improve the performance of web cache admission and web pre-fetching. Moreover, the trained classifiers would be helpful in web mining applications and web page prediction field.

Secondly, since the proposed intelligent web proxy caching approaches can successfully store the desired web object and remove the unwanted objects, cache pollution can be alleviated. Thus, cache usage can be optimized appropriately, and the hit ratio and/or the byte hit ratio can be considerably improved. Consequently, services demands on origin servers are reduced due to maximizing the cache hits, reducing connections to the origin servers. Hence, the proposed approaches can lower transit costs for accessing the origin servers.

Thirdly, unlike conventional caching approaches, the proposed intelligent web proxy approaches incorporate machine learning techniques, which form the basis for adaptive systems, to cope with web proxy caching issues. Therefore, the proposed approaches are more effective and more adaptive to web environment that changes and updates rapidly and continuously.

Finally, in the proposed intelligent web proxy approaches, requests of web objects are served well from the web proxy cache. This can reduce the amount of bandwidth used by clients. Thus, Internet network traffic can be reduced. Moreover, the user-perceived latency associated with obtaining web objects can be reduced accordingly.

1.9 Summary of Research Contributions

In this study, conventional web proxy caching approaches are extended using supervised machine learning to enable the algorithms to adapt intelligently over time. This study proposes a family of new intelligent cache replacement algorithms designed for use in web proxy cache. The core of the proposed approaches is to use machine learning techniques to predict whether web objects will be needed again in the future and to then incorporate this information into the methods determining what to remove from the proxy cache. In particular, SVM, NB and C4.5 are integrated with traditional web proxy caching algorithms, such as Least-Recently-Used (LRU), Greedy-Dual-Size (GDS), Greedy-Dual-Size-Frequency (GDSF) and Least-Frequently-Used-Dynamic-Aging (LFU-DA), to provide intelligent and more effective web proxy caching approaches.

The experimental results show that intelligent web proxy caching approaches outperform conventional caching techniques on a range of datasets. In particular, the major contributions in the field of web proxy caching can be summarized in the following aspects:

i. Intelligent proxy cache contents classification approach based on supervised machine learning techniques
ii. Intelligent LRU approaches known as SVM-LRU, NB-LRU and C4.5-LRU.
iii. Intelligent GDS approaches known as SVM-GDS, NB-GDS and C4.5- GDS.
iv. Intelligent GDSF approaches known as SVM-GDSF, NB-GDSF and C4.5-GDSF.
v. Intelligent DA approaches known as SVM-DA, NB-DA and C4.5-DA.

1.10 Thesis Outline

This thesis contains eight chapters and is organized as follows:

Chapter 1 provides a brief introduction of the study. It covers topics on problem background and motivations, problem statement, research objectives, research scope, significance of the research, summary of research contributions, and thesis outline.

Chapters 2 and 3 introduce a general overview of the literature review of this study. Chapter 2 reviews basic concepts of web proxy caching, locations of web proxy cache, web proxy cache replacement, and benchmarking of web proxy caching. More significantly, the conventional and intelligent web caching methods are analyzed and discussed in Chapter 2. Chapter 3 provides the basic concepts of supervised machine learning techniques used for intelligent web proxy caching approaches. Support vector machine (SVM), Naïve Bayes classifier (NB) and decision tree (C4.5) are presented in Chapter 3.

Chapter 4 describes in-depth the methodology used in this study. The research methodology is presented as a flow chart diagram that describes briefly how each step is carried out.

Chapter 5 illustrates an operational framework for the proposed intelligent web caching approaches. The framework consists of two functional components: offline component and online component. Chapter 5 explains in details how the offline component can train the SVM, NB and C4.5 for classification of web objects, either to be revisited or not. Moreover, this chapter describes the implementation of the methodology of web proxy cache contents classification approach based on machine learning techniques.

Chapter 6 presents how the trained machine learning techniques can be incorporated with traditional web proxy caching in online component in order to provide more effective web proxy caching approaches. This chapter provides detailed explanations of the intelligent web proxy caching approaches, which suggested for improving the performance of conventional web proxy caching algorithms. In addition, Chapter 6 describes the methodology of implementing the proposed intelligent web proxy caching approaches in simulation environment.

In Chapter 7, benchmarking for supervised machine learning classifiers is discussed in the first part of this chapter. In the second part, the proposed intelligent web proxy caching approaches are compared with the most common and relevant web proxy caching approaches, including conventional and intelligent web proxy caching techniques

Finally, Chapter 8 concludes the thesis and highlights the contributions and findings of the research work. In addition, Chapter 8 provides suggestions and recommendations for future study.

REFERENCES

Abe, S. (2010). *Support vector machines for pattern classification*: Springer-Verlag New York Inc.

Abhari, A., Dandamudi, S. P., and Majumdar, S. (2006). Web object-based storage management in proxy caches. *Future Generation Computer Systems.* 22(1-2), 16-31.

Abrams, M., Standridge, C. R., Abdulla, G., Fox, E. A., and Williams, S. (1996). Removal policies in network caches for World-Wide Web documents. *ACM.* 26(4), 293-305.

Acharjee, U. (2006). *Personalized and Artificial Intelligence Web Caching and Prefetching.* Master Thesis. University of Ottawa, Canada.

Ali, W., and Shamsuddin, S. M. (2009a). Intelligent Client-Side Web Caching Scheme Based on Least Recently Used Algorithm and Neuro-Fuzzy System. In W. Yu, H. He and N. Zhang (Eds.), *Advances in Neural Networks – ISNN 2009* (Vol. 5552, pp. 70-79): Springer Berlin / Heidelberg.

Ali, W., and Shamsuddin, S. M. (2009b). Integration of least recently used algorithm and neuro-fuzzy system into client-side web caching. *International Journal of Computer Science and Security, 3*(1), 1-15.

Arlitt, M., Cherkasova, L., Dilley, J., Friedrich, R., and Jin, T. (2000a). Evaluating content management techniques for web proxy caches. *ACM SIGMETRICS Performance Evaluation Review.* 27(4), 3-11.

Arlitt, M., Friedrich, R., and Jin, T. (2000b). Performance evaluation of Web proxy cache replacement policies. *Performance Evaluation.* 39(1-4), 149-164.

Ayani, R., Yong Meng, T., and Yean Seen, N. (2003). Cache pollution in Web proxy servers. *Parallel and Distributed Processing Symposium.* 22-26 April 2003. Sweden.

Bai, C.-G. (2005). Bayesian network based software reliability prediction with an operational profile. *Journal of Systems and Software.* 77(2), 103-112.

Barish, G., and Obraczke, K. (2000). World Wide Web caching: trends and techniques. *Communications Magazine, IEEE.* 38(5), 178-184.

Bin, W., and Kshemkalyani, A. D. (2004). *O*bjective-greedy algorithms for long-term Web prefetching. *Third IEEE International Symposium on the Network Computing and Applications* (NCA 2004). 30 Aug.-1 Sept, 2004. Cambridge, 61-68.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A. (2009). *WEKA Manual for Version 3-6-1*. Hamilton, New Zealand: The University of Waikato.

BU Web Trace. (1995). http://ita.ee.lbl.gov/html/contrib/BU-Web-Client.html.

Calzaross, M. C., and Vall, G. (2003). A Fuzzy Algorithm for Web Caching. *Simulation Series Journal.* 35(4), 630-636.

Cao, P., and Irani, S. (1997). Cost-Aware WWW Proxy Caching Algorithms. T*HE 1997 USENIX SYMPOSIUM ON INTERNET TECHNOLOGY AND SYSTEMS*. December 1997. Monterey, California, USA.

Chang, C. C., and Lin, C. J. (2001). LIBSVM: A library for support vector machines: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, H. T. (2008). *Pre-fetching and Re-fetching in Web caching systems: Algorithms and Simulation.* Master Thesis. TRENT UNIVESITY, Peterborough, Ontario, Canada.

Chen, R.-C., and Hsieh, C.-H. (2006). Web page classification based on a support vector machine using a weighted vote schema. *Expert Systems with Applications.* 31(2), 427-435.

Chen, T. (2007). Obtaining the optimal cache document replacement policy for the caching system of an EC website. *European Journal of Operational Research.* 181(2), 828-841.

Cherkasova, L. (1998). Improving WWW Proxies Performance with Greedy-Dual-Size-Frequency Caching Policy. *In HP Technical Report, Palo Alto*.

Cherkasova, L., and Ciardo, G. (2001). Role of Aging, Frequency, and Size in Web Cache Replacement Policies. *The 9th International Conference on High-*

Performance Computing and Networking. Springer-Verlag London, UK, 114-123.

Cobb, J., and ElAarag, H. (2008). Web proxy cache replacement scheme based on back-propagation neural network. *Journal of Systems and Software.* 81(9), 1539-1558.

Danzig, P., Mogul, J., Paxson, V., & Schwartz, M. 2000. The internet traffic archive. *URL: http://ita.ee.lbl.gov/html/traces.html*

Darwiche, A. (2010). Bayesian networks. *Commun. ACM.* 53(12), 80-90.

Datta, A., Dutta, K., Thomas, H., and VanderMeer, D. (2003). World Wide Wait: A Study of Internet Scalability and Cache-Based Approaches to Alleviate It. *Management Science.* 49(10), 1425-1444.

Davison, B. D. (2001). A Web caching primer. *Internet Computing, IEEE.* 5(4), 38-45.

Davison, B. D. (2002). *The design and evaluation of web prefetching and caching techniques.* Doctor of Philosophy. Rutgers, The State University of New Jersey.

de Melo, A. C. V., and Sanchez, A. J. (2008). Software maintenance project delays prediction using Bayesian Networks. *Expert Systems with Applications.* 34(2), 908-919.

Domenech, J., de la Ossa, B., Sahuquillo, J., Gil, J. A., and Pont, A. (2012). A taxonomy of web prediction algorithms. Expert Systems with Applications, 39(9), 8496-8502.

Domenech, J., Gil, J. A., Sahuquillo, J., and Pont, A. (2010a). Using current web page structure to improve prefetching performance. *Computer Networks.* 54(9), 1404-1417.

Domènech, J., Pont-Sanjuán, A., Sahuquillo, J., and Gil, J. A. (2010b). Evaluation, Analysis and Adaptation of Web Prefetching Techniques in Current Web. In J. Yao (Ed.), *Web-based Support Systems* (pp. 239-271): Springer London.

Domenech, J., de la Ossa, B., Sahuquillo, J., Gil, J. A., and Pont, A. (2012). A taxonomy of web prediction algorithms. Expert Systems with Applications, 39(9), 8496-8502.

ElAarag, H., and Romano, S. (2009). Improvement of the neural network proxy cache replacement strategy. *Proceedings of the 2009 Spring Simulation Multiconference*. San Diego, CA, USA.

Fan, L., Poh, K. L., and Zhou, P. (2009). A sequential feature extraction approach for naīve bayes classification of microarray data. *Expert Systems with Applications.* 36(6), 9919-9923.

Farhan. (2007). *Intelligent Web Caching Architecture.* Master Thesis. Faculty of Computer Science and Information System,UTM University,Johor,Malaysia.

Fayyad, U. M., and Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. *The 13th International Joint Conference on Artificial Intelligence (IJCAI-93)*. 1022-1027.

Fernández, A., del Jesus, M. J., and Herrera, F. (2009). Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. *International Journal of Approximate Reasoning.* 50(3), 561-577.

Fernández, A., García, S., del Jesus, M. J., and Herrera, F. (2008). A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Sets and Systems.* 159(18), 2378-2398.

Foong, A. P., Yu-Hen, H., and Heisey, D. M. (1999). Logistic regression in an adaptive Web cache. *Internet Computing, IEEE.* 3(5), 27-36.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning.* 29(2), 131-163.

Goubanova, O., and King, S. (2008). Bayesian networks for phone duration prediction. *Speech Communication.* 50(4), 301-311.

Hai, L., and Maobian, C. (2010). Evaluation of web caching consistency. Paper presented at the 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 2010.

Hall, M. (2007). A decision tree-based attribute weighting filter for naive Bayes. *Knowledge-Based Systems. 20*(2), 120-126.

Han, J., and Kamber, M. (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Hao, C., Ye, T., and Zhijun, W. (2002). Hierarchical Web caching systems: modeling, design and experimental results. *IEEE Journal on Selected Areas in Communications.* 20(7), 1305-1314.

Hsu, C. W., Chang, C. C., and Lin, C. J. (2009). A practical guide to support vector classification. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/libsvm

Hu, Z. H., and Ding, Y. S. (2010). An immune inspired co-evolutionary affinity network for prefetching of distributed object. *Journal of Parallel and Distributed Computing.* 70(2), 92-100.

Huang, C.-J., Wang, Y.-W., Huang, T.-H., Lin, C.-F., Li, C.-Y., Chen, H.-M. (2011). Applications of machine learning techniques to a sensor-network-based prosthesis training system. *Applied Soft Computing.* 11(3), 3229-3237.

Huang, Y. F., and Hsu, J. M. (2008). Mining web logs to improve hit ratios of prefetching and caching. *Knowledge-Based Systems. 21*(1), 62-69.

Jaeeun, J., Gunhoon, L., Haengrae, C., and Byoungchul, A. (2003). A prefetching Web caching method using adaptive search patterns. *IEEE Pacific Rim Conference on Communications, Computers and signal Processing(PACRIM).* 28-30 August, 2003. Victoria, 37-40.

Gawade, S., and Gupta, H. (2012). Review of Algorithms for Web Pre-fetching and Caching. International Journal of Advanced Research in Computer and Communication Engineering, 1(2), 62-65.

Kaya, C. C., Zhang, G., Tan, Y., and Mookerjee, V. S. (2009). An admission-control technique for delay reduction in proxy caching. *Decision Support Systems.* 46(2), 594-603.

Kin-Yeung, W. (2006). Web cache replacement policies: a pragmatic approach. *Network, IEEE.* 20(1), 28-34.

Koskela, T., Heikkonen, J., and Kaski, K. (2003). Web cache optimization with nonlinear model using object features. *Computer Networks.* 43(6), 805-817.

Krishnamurthy, B., and Rexford, J. (2001). *Web protocols and practice: HTTP/1.1, Networking protocols, caching, and traffic measurement* (Vol. 108): Addison-Wesley.

Kumar, C. (2009). Performance evaluation for implementations of a network of proxy caches. *Decision Support Systems. 46*(2), 492-500.

Kumar, C., and Norris, J. B. (2008). A new approach for a proxy-level web caching mechanism. *Decision Support Systems. 46*(1), 52-60.

Liu, B. (2007). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.* Springer Verlag.

Lu, S. H., Chiang, D. A., Keh, H. C., and Huang, H. H. (2010). Chinese text classification by the Naīve Bayes Classifier and the associative classifier with multiple confidence threshold values. *Knowledge-Based Systems.* 23(6), 598-604.

Lucas, P. (2001). Bayesian networks in medicine: a model-based approach to medical decision making. *The EUNITE workshop on intelligent systems in patient care*. Vienna.

Markatchev, N., and Williamson, C. (2002). WebTraff: A GUI for Web Proxy Cache Workload Modeling and Analysis. *Proceedings of the 10th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems.* IEEE Computer Society, p. 356.

Mountrakis, G., Im, J., and Ogole, C. (2011). Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing, 66(3), 247-259.

Nagaraj, S. (2004). *Web caching and its applications.* (Vol. 772): Springer Netherlands.

NLANR. (2005). National Lab of Applied Network Research (NLANR). Sanitized Access Logs: http://www.ircache.net.

NLANR. (2010a). National Lab of Applied Network Research(NLANR). Sanitized access logs: collected between 21st August and 4th September, 2010 from: http://www.ircache.net/.

NLANR. (2010b). National Lab of Applied Network Research(NLANR). *Sanitized access logs: http://www.ircache.net/*(last accessed in January 2010).

Oliveira, L. S. C., Andreão, R. V., and Sarcinelli-Filho, M. (2010). The Use of Bayesian Networks for Heart Beat Classification. In A. Hussain, I. Aleksander, L. S. Smith, A. Kardec Barros, R. Chrisley and V. Cutsuridis (Eds.), *Brain Inspired Cognitive Systems 2008* (Vol. 657, pp. 217-231): Springer New York.

Patil, J., and Pawar, B. (2011). Improving Performance on WWW using Intelligent Predictive Caching for Web Proxy Servers. International Journal of Computer Science Issues, 8(1), 402-408.

Pernkopf, F. (2005). Bayesian network classifiers versus selective k-NN classifier. *Pattern Recognition.* 38(1), 1-10.

Podlipnig, S., and Böszörmenyi, L. (2003). A survey of Web cache replacement strategies. *ACM Comput. Surv.* 35(4), 374-398.

Qiang, Y., and Zhang, H. H. (2003). Web-log mining for predictive Web caching. *IEEE Transactions on Knowledge and Data Engineering.* 15(4), 1050-1053.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning*: Morgan kaufmann.

Rokach, L., and Maimon, O. Z. (2008). *Data mining with decision trees : theroy and applications*. Singapore ; Hackensack, NJ: World Scientific.

Romano, S., and ElAarag, H. (2011). A neural network proxy cache replacement strategy and its implementation in the Squid proxy server. *Neural Computing and Applications. 20*(1), 59-78.

Romano, S., and ElAarag, H. (2012). A quantitative study of Web cache replacement strategies using simulation. *SIMULATION. 88(5), 507-541.*

Saha, A. K., Deb, P. P., Kar, M., and Rudrapal, D. (2012). An Optimization Technique of Web Caching using Fuzzy Inference System. International Journal of Computer Applications, 43(17), 20-23.

Sajeev, G., and Sebastian, M. (2011). A novel content classification scheme for web caches. *Evolving Systems.* 2(2), 101-118.

Samanta, B. (2004). Gear fault detection using artificial neural networks and support vector machines with genetic algorithms. Mechanical Systems and Signal Processing, 18(3), 625-644.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR). 34*(1), 1-47.

Songwattana, A. (2008). Mining Web Logs for Prediction in Prefetching and Caching. *Third International Conference on Convergence and Hybrid Information Technology (ICCIT '08).* 11-13 November, 2008. Busan, 1006-1011.

Soonthornsutee, R., and Luenam, P. (2012). Web Log Mining for Improvement of Caching Performance. *Proceedings of the International MultiConference of Engineers and Computer Scientists*.

Starr, C., and Shi, P. (2004). An introduction to bayesian belief networks and their applications to land operations. *Network. DSTO Systems Sciences Laboratory*. Edinburgh South Australia.

Sulaiman, S., Shamsuddin, S. M., Forkan, F., and Abraham, A. (2008). Intelligent Web Caching Using Neurocomputing and Particle Swarm Optimization Algorithm. *Second Asia International Conference on Modeling and Simulation(AICMS 08)*. 13-15 May, 2008. Kuala Lumpur, 642-647.

Sulaiman, S., Shamsuddin, S. M., and Abraham, A. (2011). INTELLIGENT WEB CACHING USING MACHINE LEARNING METHODS. Neural Network World, 5, 429-452.

Temko, A., Thomas, E., Marnane, W., Lightbody, G., and Boylan, G. (2011). EEG-based neonatal seizure detection with Support Vector Machines. Clinical Neurophysiology, 122(3), 464-473.

Tian, W., Choi, B., and Phoha, V.V. (2002). An Adaptive Web Cache AccessPredictor Using Neural Network. *Lecture Notes In Computer Science, Proceedings of the 15th international conference on Industrial and engineering applications of artificial intelligence and expert systems: developments in applied artificial intelligence*. 2358, 450-459. Published by: Springer- Verlag London, UK.

Tirdad, K., Pakzad, F., and Abhari, A. (2009). Cache replacement solutions by evolutionary computing technique. *Proceedings of the 2009 Spring Simulation Multiconference*. San Diego, California, 1-4.

Vakali, A. (2002). Evolutionary Techniques for Web Caching. *Distrib. Parallel Databases. 11*(1), 93-116.

Van Koten, C., and Gray, A. R. (2006). An application of Bayesian network for predicting object-oriented software maintainability. *Information and Software Technology.* 48(1), 59-67.

Vapnik, V. (1995). The nature of statistical learning theory. (2nd edition). New York: Springer.

Venketesh, P., and Venkatesan, R. (2009). A survey on applications of neural networks and evolutionary techniques in web caching. IETE Technical Review, 26(3), 171.

Wang, B., Wu, Y. W., and Zheng, W. M. (2011). Web Caching Replacement Based on User's Visiting Action. Applied Mechanics and Materials, 52, 25-30.

Wessels, D. (2001). *Web caching*: O'Reilly Media.

Wessels, D., and Claffy, K. (1997). *Application of internet cache protocol (ICP).* USA: RFC Editor.

Wolman, A., Voelker, M., Sharma, N., Cardwell, N., Karlin, A., and Levy, H. M. (1999). On the scale and performance of cooperative web proxy caching. *ACM SIGOPS Operating Systems Review. 33*(5), 16-31.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems. 14*(1), 1-37.

Yan, C., Zeng-Zhi, L., and Zhi-Wen, W. (2004). *A GA-based cache replacement policy.* Paper presented at the Proceedings of 2004 International Conference on Machine Learning and Cybernetics, 2004.

Yang, Q., Huang, J. Z., and Ng, M. (2003). A Data Cube Model for Prediction-Based Web Prefetching. *Journal of Intelligent Information Systems.* 20(1), 11-30.

Yu, W., Liu, T., Valdez, R., Gwinn, M., and Khoury, M. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making, 10(1), 16.

Zeng, Z., Veeravalli, B., and Li, K. (2011). A novel server-side proxy caching strategy for large-scale multimedia applications. Journal of Parallel and Distributed Computing, 71(4), 525-536.

Zhao, Y., and Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. *Advances in Space Research.* 41(12), 1955-1959.

Zhijie, B., Zhimin, G., and Yu, J. (2009). A survey of Web prefetching. *Journal of computer research and development. 46*(2), 202-210.