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.