DEVELOPMENT OF INTELLIGENT WEB PROXY CACHE REPLACEMENT ALGORITHMS BASED ON ADAPTIVE WEIGHT RANKING POLICY VIA DYNAMIC AGING

BY

DUA'A MAHMOUD AL-QUDAH

A dissertation submitted in fulfilment of the requirement for the degree of Master of Science (Computer and Information Engineering)

Kulliyyah of Engineering
International Islamic University Malaysia

March 2018

ABSTRACT

Nowadays, the World Wide Web plays an essential role in our lives. It has become a great useful tool for people in all facets of life. The vast usage of the World Wide Web leads to an increase in network traffic and create a bottleneck over the internet performance. For most people, the accessing speed or the response time is the most critical factor when using the internet. Web proxy cache technique reduces response time by storing copies of pages between client and server sides. If requested pages are cached in the proxy, there is no need to access the server. Due to the limited size and excessive cost of cache as compared to other storages, cache replacement algorithm is used to determine evict page when the cache is full. On the other hand, the conventional algorithms for replacement such as Least Recently Use (LRU), First in First Out (FIFO), Least Frequently Use (LFU), Randomised Policy and etc. may discard important pages just before its use. Furthermore, using conventional algorithm cannot be well optimized since it requires some decision to intelligently evict a page before replacement. Hence, this research proposes integrated Adaptive Weight Ranking Policy (AWRP) with intelligent classifiers based on Naïve Bayes (NB), J48 decision via dynamic aging factor to form intelligent replacement algorithms called NB-AWRP-DA and J48-AWRP-DA that improves the performance of AWRP based on hit rate (HR) and byte hit rate (BHR) over LRU, LFU and FIFO algorithms. In order to enhance classifier's power of prediction by rising classifier accuracy before integrating them with AWRP, this research proposes using automated wrapper feature selection methods to choose the best subset of features that are relevant and influence classifiers prediction accuracy. The results present that by using wrapper feature selection methods namely: Best First (BFS), Incremental Wrapper subset selection (IWSS) embedded NB and particle swarm optimization (PSO) reduce number of features and have a good impact on reducing computation time. However, Using PSO enhances NB classifier accuracy by 1.1%, 0.43% and 0.22% over training NB with all features, using BFS and using IWSS embedded NB respectively. PSO rises J48 accuracy by 0.03%, 1.91% and 0.04% over using J48 classifier with all features, using IWSS-embedded NB and using BFS respectively. While using IWSS embedded NB fastest NB and J48 classifiers are much more than BFS and PSO. However, it reduces computation time of NB by 0.1383 seconds and reduce computation time of J48 by 2.998 seconds. Moreover, experimental result of intelligent replacement algorithms shows that NB-AWRP-DA enhances the performance of web proxy cache a cross multi proxy datasets by 4.008%,4.087% and 14.022% over LRU, LFU and FIFO in terms of HR. Moreover J48-AWRP-DA increases HR by 0.483%, 0.563% and 10.497% over LRU, LFU and FIFO respectively. Meanwhile, BHR rises by 0.991%,1.008% and 11.584% over LRU, LFU and FIFO respectively using NB-AWRP-DA. And by 0.0204%, 0.0379% and 10.614% for LRU, LFU, FIFO respectively using J48-AWRP-DA.

ABSTRACT IN ARABIC

تلعب الشبكة العنكبوتية العالمية دورا أساسيا في حياتنا في الوقت الحاضر. حيث تسهل وتساعد الناس في شتى مجالات الحياة. حياة معظم الناس أصبحت متعلقة باستخدام الانترنت من حيث الدراسة والعمل والترفيه والتسوق وغير ذلك. لكن هذا الاستخدام الهائل للإنترنت أدى إلى زيادة تدفق البيانات عبر الشبكة العنكبوتية مما أدى الى ظهور مشكلة عنق الزجاجة التي أثرت بدورها سلبا على أداء الشبكة. يهتم معظم مستخدمي الشبكة العنكبوتية بسرعة الاتصال بالشبكة وسرعة استجابة الشبكة لطلباتهم. لذلك بدا استخدام ذاكرة التخزين لمؤقت أمرا ضروريا لتقليل وقت الاستجابة وتسريع عملية الرد على العميل . حيث انها تقوم بالاحتفاظ بنسخة من صفحات الويب بين الخادم والعميل وعليه لا داعي للعودة الى الخادم إذا توفرت الصفحة على ذاكرة التخزين المؤقت عند طلبها. ولكن ونظرا لصغر حجم ذاكرة التخزين المؤقت وتكلفتها المرتفعة مقارنة بوحدات التخزين الأخرى فان استخدام خوارزميات الاستبدال عند امتلاء ذاكرة التخزين المؤقت بدا أمرا ضروريا. لكن استخدام خوارزميات الاستبدال التقليدية مثل خوارزمية الأقل استخداما مؤخرا وخوارزمية الأقل تكرارا وخوارزمية الذي يدخل أولا يخرج أولا قد يستبدل صفحات مهمه من الممكن أن يتم استخدامها في الوقت القريب. لذلك فإن دمج المصنفات الذكية مع الخوارزميات التقليدية يحسن من أداء الشبكة. لذلك، يقدم هذا البحث مقترحا لاستخدام طريقة الاختيار الآلي مع المصنفات الذكية لتحسين أدائها قبل دمجها مع خوارزميات الاستبدال. حيث ان لديها القدرة على اختيار أفضل عينة من العناصر التي لها تأثير فعال وقوي على أداء المصنفات. حيث تشير النتائج الى ان استخدام عمليات اختيار العناصر الألية مثل آلية الأفضل أولا وآلية الاختيار التدريجي متضمنة المصنف NBوآلية تحسين سرب الجسيمات قد أدى الى تقليل عدد العناصر المستخدمة وعليه تقليل الوقت اللازم لتدريب المصنفات. وأثبتت النتائج ان آلية تحسين سرب الجسيمات قد رفعت أداء المصنف NB بنسبة 1.1%، و 0.43%، و 0.22 % مقارنة بتدريب المصنف بناء على كل العنصر الموجودة، و تدريبه اعتمادا على العناصر المختارة باستخدام آلية الأفضل أولا، وبناء على اختيار العناصر باستخدام آلية الاختيار التدريجي على التوالي. بالإضافة إلى تحسين أداء المصنف J48 ورفع أدائه بنسبة 0.03%، و 1.19%، و0.04% مقارنة باستخدام المصنف مع كل العناصر الموجودة، أو استخدامه مع العناصر المختارة بآلية الأفضل أولا أو الاختيار التدريجي حسب التتابع. كما أثبتت النتائج أن استخدام آلية الاختيار التدريجي تقلل الوقت المستغرق في تدريب المصنف NB بمعدل 0.138 ثانيه في حين انه يقلل الوقت المستغرق في تدريب ال J48 بمعدل 2.998 ثانية. يقترح هذا البحث دمج الخوارزمية المعتمدة على الوزن مع

المصنفات NBو J48 لإنتاج خوارزميات استبدال ذكية تعتمد على الوزن وهي NB-AWRP-DA و J48-AWRP-DA . حيث تشير نتائج تقييم NB-AWRP-DA الى ان الخوارزمية رفعت أداء ذاكرة التخزين اعتمادا على نسبة عدد الطلبات بمعدل 4.008% مقارنة مع خوارزمية الأقل استخداما مؤخرا. وبمعدل 4.087% مقارنة مع خوارزمية الأقل تكرارا. وبمعدل 14.022% مقارنة مع خوارزمية الذي يدخل أولا يخرج أولا. كما توضح النتائج ان نسبة معدل التحسين بناءا على نسبة عدد الطلبات باستخدام J48-AWRP-DA هي كالتالي 0.483% مقارنة بخوارزمية الأقل استخداما مؤخرا، و 0.563% مقارنة ب الأقل تكرار ، و 10.497% مقارنة ب الذي يدخل أولا يخرج أولا. كما أن NB-AWRP-DAحسنت الأداء اعتمادا على نسبة حجم الطلبات بالبايت بمعدل 0.991%، و 1.008% و 11.584% مقارنة بخوارزمية الأقل استخداما مؤخرا وخوارزمية الأقل تكرارا و خوارزمية الذي يدخل أولا يخرج أولا على التوالي. كما ان J48-AWRP-DAحسنت نسبة حجم الطلبات بمعدل 0.0204% و 0.0379% و 10.613% مقارنة بخوارزمية الأقل استخداما مؤخرا وخوارزمية الأقل تكرارا و خوارزمية الذي يدخل أولا يخرج أولا على التوالي.

APPROVAL PAGE

I certify that I have supervised and read this study and that in my opinion, it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master of Science (Computer and Information Engineering).

……………………………..
Rashidah Funke Olanrewaju
Supervisor

……………………………..
Amelia Wong Azman
Co-Supervisor

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master of Science (Computer and Information Engineering).

……………………………..
Mashkuri Yaacob
Internal Examiner

……………………………..
Khairul Azami Sidek
Internal Examiner

This thesis was submitted to the Department of Electrical and Computer Engineering and is accepted as a fulfilment of the requirement for the degree of Master of Science (Computer and Information Engineering).

……………………………..
Mohamed Hadi Habaebi
Head, Department of Electrical and Computer Engineering

This thesis was submitted to the Kulliyyah of Engineering and is accepted as a fulfilment of the requirement for the degree of Master of Science (Computer and Information Engineering).

……………………………..
Erry Yulian Triblas Adesta
Dean, Kulliyyah of Engineering

DECLARATION

I hereby declare that this dissertation is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Dua'a Mahmoud Mohammad Al-Qudah

Signature .... Date ....

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

DEVELOPMENT OF INTELLIGENT WEB PROXY CACHE REPLACEMENT ALGORITHMS BASED ON ADAPTIVE WEIGHT RANKING POLICY VIA DYNAMIC AGING

I declare that the copyright holders of this dissertation are jointly owned by the student and IIUM.

No part of this unpublished research may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the copyright holder except as provided below

1. Any material contained in or derived from this unpublished research may be used by others in their writing with due acknowledgement.

2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purposes.

3. The IIUM library will have the right to make, store in a retrieved system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Dua'a Mahmoud Al-Qudah

……..…………………….. ……………………….

Signature Date

TO MY LATE FATHER

AL Haj Mahmoud Mohammad Al-Qudah

Who returned to Allah on 13th December 2010,

May Allah have mercy on his soul, forgive his shortcomings, admit him into the Jannatul Firdaus, and protect him from the punishment of the grave and the torment of the Fire Amin.

You are the one who taught me how to face challenges and the concept of insistence and determination. You always prayed for me to succeed. Thanks a lot, my dear father.

ACKNOWLEDGEMENTS

All glory is due to Allah, the Almighty, whose Grace and Mercies have been with me throughout the duration of my study. Although, it has been challenging, His Mercies and Blessings on me ease the herculean task of completing this thesis.

I am most indebted to my supervisors, Asst. Prof. Dr. Rashidah Funke Olanrewaju, whose enduring disposition, kindness, promptitude, thoroughness and friendship have facilitated the successful completion of my work. I put on record and appreciate her detailed comments, useful suggestions and inspiring queries which have considerably improved this thesis. Her brilliant grasp of the aim and content of this work led to their insightful comments, suggestions and queries which helped me a great deal. Despite her commitments, she took time to listen and attended to me whenever requested. The moral support she extended to me is in no doubt a boost that helped in building and writing the draft of this research work. I am also grateful to my co-supervisor, Assoc. Prof. Dr. Amelia, whose support and cooperation contributed to the outcome of this work.

In addition, it is the biggest pleasure to express my gratitude to my beloved mother Na'amah Yousef Al-Momani, brothers and sisters back in Jordan, as their never ending loving support, prayers and encouragement led me to finish my study and have a changed life. I would also like to express thanks to my father in law Mustafa Hijazi, mother in law Ferial Al-Momani, brothers and sisters of my husband, as they have supported me to complete my study successfully.

Last but not least, my gratitude goes to my beloved husband Musab Hijazi and my lovely daughters Jood Hijazi and Jana Hijazi; for their patience, love, prayers, understanding and endurance me. Without them, this work will never have been completed. Thank you Allah for blessing me with a wonderful family.

Once again, we glorify Allah for His endless mercy on us one of which is enabling us to successfully round off the efforts of writing this thesis. Alhamdulillah.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

ANF	Adaptive Neuro Fuzzy
ANFIS	Adaptive Neuro Fuzzy Inference System
ANN	Artificial Neural Network
AWRP	Adaptive Weight Ranking Policy
AWRP-DA	Adaptive Weight Ranking Policy Via Dynamic Aging Factor
BFS	Best First Method
BHR	Byte Hit Rate
BO2	Boulder
BSS	Backward Sequential Selection
CAR	Clock Adaptive Replacement
CCR	Correct Classification Rate
DCPR	Dual Cache Replacement Policy
DT	Decision Tree
FIFO	First Input First Output
FIS	Fuzzy Inference System
FN	False Negative
FP	False Positive
FSS	Forward Selection Method
GD*	Greedy Dual*
GDS	Greedy Dual SIZE
GDSF	Greedy-Dual-Size-Frequency

Gmean	Geometric Mean
GUI	Graphical User Interface
HR	Hit Rate
ICWCS	Intelligent Client-Side Web Caching Scheme
IR_Frequency	Internal Request Frequency
IWSS	Incremental Wrapper Feature Subset
Embedded NB	Selection with Naïve Bayes
J48-AWRP-	Decision Tree (J48) Adaptive Weight
DA	Ranking Policy Via Dynamic Aging
LFU	Least Frequently Used
LNC-R-W3	Least Normalized Cache Replacement on the Web
LRU	Least Recently Used
LUV	Least Unified Value
MDL	Minimum Description Length
NB	Naïve Bayes
NB-AWRP-	Naïve Bayes Adaptive Weight Ranking
DA	Policy Via Dynamic Aging
NNPCR	Neural Network Proxy Cache Replacement
PA	Palo Alto
PSO	Particle Swarm Optimization
RFT	Random Forest Tree Classifier
SD	San Diego
SJ	San Jose

SV	Silicon Valley
SVM	Support Vector Machine
SWL	Sliding Window
TANB	Tree Augmented Naïve Bayes Classifier
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
TSP	Taylor Series Prediction
UC	Urbana-Champaign
WFS	Without Feature Selection
WRCP	Weighted-Rank Cache Replacement Policy
WRP	Weighting Replacement Policy
WWPCS	Windows Web Proxy Caching Simulation

CHAPTER ONE

INTRODUCTION

1.1. BACKGROUND OF THE STUDY

The large usage of internet leads to increase in network traffic and page access latency (Sarhan, Elmogy, & Ali, 2014). By having web cache, web objects can be kept closer to the client. A web proxy cache saves web object that can be revisited in the near future closer to the user which leads to decrease the number of web object requested from the server, hence, reducing the volume of data that is transmitted over the network as well as the delay while accessing the web page (Ali, Shamsuddin, & Ismail, 2012a; Sathiyamoorthi & Bhaskaran, 2012). Consequently, the overall performance of web system can be improved. These web objects are commonly placed in three locations; 1) browser caching, 2) proxy caching and; 3) server side proxy caching (KumarSaha, Pratim Dev, Kar, & Rudrapal, 2012). Figure 1.1 shows various types of web cache.

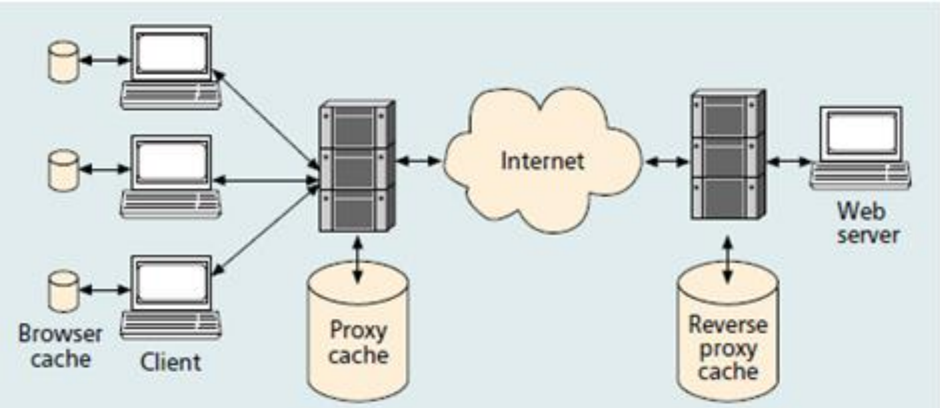

Figure 1. 1 Web cache types (KumarSaha, Pratim Dev, Kar, & Rudrapal, 2012)

There are three significant factors that are the pivotal roles in cache performance. which are the cache replacement algorithms, cache consistency and cache algorithm (passive or active). Cache replacement algorithm is a process of evicting pages from cache to make room for new requested page where there is not enough space for it. which have significant effect in the web caching (Muralidhar & Geethanjali, 2012), and are classified into five categories, namely: frequency based, recency based, frequency/recency based, function based and randomized based (ElAarag, 2013). However, most algorithms that used are Least Recently Used(LRU), Least Frequently Used(LFU), Size and Greedy Dual SIZE(GDS)- that are called conventional algorithms are not suitable for web caching since the algorithms consider one or two features while ignoring the others(Ali et al., 2012a). Conventional algorithms also suffer from many issues like cache pollution in which the cache saved objects are not requested frequently. There are many algorithms that combine two or more features to enhance the performance of the replacement algorithms. However, the selection of features are not a straightforward task because an important factor in one environment may be not important in another environment (Ali et al., 2012a).

Adaptive Weight Ranking Policy (AWRP) attempts to give a weight to each page depending on three factors: recency, frequency and total number of access to be made to improve performance of cache and overcome traditional algorithms problem such as least recently used LRU, First Input First Output FIFO and Clock Adaptive Replacement CAR (Debabala Swain, 2011). Intelligent replacement algorithms were proposed to overcome shortcomings of conventional replacement algorithms and to enhance the performance of web cache. It predicts which web objects will be revisited in the future or not. Machine learning algorithms such as Fuzzy logic, Artificial Neural Network, Adaptive Neuro Fuzzy, Naïve Bayes, Decision Tree, and Support Vector

Machine have been used for the prediction procedure. Since the intelligent algorithms are predictive, hence, it can be combined with one or more with one or more of the conventional algorithms to improve their performance mainly in terms of hit ratio and byte hit ratio (Ali, Shamsuddin, & Ismail, 2011).

1.2. PROBLEM STATEMENT AND ITS SIGNIFICANCE

As the rapid growth in network usage and significant role for network in our life, there is a need for good service, enough bandwidth and fast response. Web caching mainly depends on storing copies of web objects closer to the user. In web caching, if the end user requests a page that is found in the cache, it is sent directly. Meanwhile, if it is not in the cache, fetching page from the origin server is required. However, transferring of web object over the networks tends to increase the network traffic. Thus, web caching helps in decreasing network traffic and reduce loads on the origin server.

Meanwhile, due to limited cache size, appropriate cache replacement algorithms are required to manage web caching. Although there are conventional replacement algorithms such as LRU, LFU, and SIZE, but they are not effective enough in web cache because of the selectiveness criteria of choosing features and ignore others which may have a good impact on web cache performance. Furthermore, those algorithms suffer from cache pollution issues where objects will be stored in the cache which may not be revisited in future hence causing waste of limited resources. Intelligent replacement algorithms are required to effectively utilize the limited resources (web cache) as well as enhancing its performance by integrating multiple features at once.

Meanwhile, features that have an important impact on web caching are: recency, frequency, size, cost, access latency, expiration time and modification time. Chosen features for training intelligent system will strongly influence prediction process. Most strategies that used intelligent systems chose features manually while automated feature selection enhances prediction process by choosing the optimal and best subset of the features that have a significant impact on intelligent systems. Hence, intelligent adaptive cache ranking policy via dynamic aging (AWRP-DA) is proposed to enhance web proxy cache performance by using wrapper feature selection method to choose the best subset of features for training C4.5, NB intelligent classifiers, as well as prevent pollution problem.

1.3. RESEARCH OBJECTIVES

The main objective of this study is to develop an Intelligent Web Proxy Cache Replacement Algorithms Based on Adaptive Weight Ranking Policy (AWRP) via Dynamic Aging. This will be done by achieved following sub objectives.

1. To study intelligent web proxy cache replacement algorithms.
2. To develop intelligent AWRP algorithms by incorporating dynamic aging factor to enhance web cache performance.
3. To enhance intelligent classifiers accuracy and reduce computation time via using wrapper feature selection methods.
4. To evaluate and verify the newly developed algorithms by comparing them with traditional web cache replacement algorithms (LRU, LFU and FIFO) in terms of Hit Rate and Byte Hit Rate.

1.4. RESEARCH METHODOLOGY

In order, to achieve research objectives, the methodology adopted set of states. Literature review help in clarification on how the proposed algorithms can be achieved. In addition, to achieve the first objective, NB and J48 classifiers are trained and evaluated based on set of features extracted using WEKA 3.8 software. After that, NB and J48 trained classifiers are integrated with AWRP algorithm. In addition, dynamic aging factor is added to Adaptive Weight Ranking Policy (AWRP) for enhancing its performance and overcoming cache pollution issue. However, trace driven simulator has been built to simulate and evaluate NB-AWRP-DA and J48-AWRP-DA. Comparison among NB-AWRP-DA, J48-AWRP-DA and LRU, LFU and FIFO algorithms are done in terms of HR and BHR. Figure 1.2 present methodology states to achieve research objectives.

1.5. SCOPE OF RESEARCH

This study was limited to web proxy cache replacement algorithms either traditional and intelligent. In addition, it focusses on how to improve prediction power of intelligent system by using automated wrapper feature selection methods. It works with proxy log files from different servers. The proposed algorithms depend on frequency, recency, dynamic aging and a total number of access factors only. which does not need to deal with other features. Moreover, the measures which are used to evaluate classifiers are CCR, TPR, TNR and Gmean while other measures do not consider in this research. However, web proxy cache algorithms are evaluated in terms of HR and BHR.

1.6. DISSERTATION ORGANIZATION

The rest of the thesis is organized as follows: Chapter Two reviewed the literatures in the field of cache replacement algorithms either conventional algorithms or the proposed algorithms by researchers. In addition, this chapter reviewed intelligent web cache replacement algorithms.

In Chapter Three, research methodology is explained in detail, so it includes dataset preparation, methods that had been used, simulation and evaluation procedure. Simulation result and evaluation discussed in chapter four. Finally, the thesis concluded in chapter five.

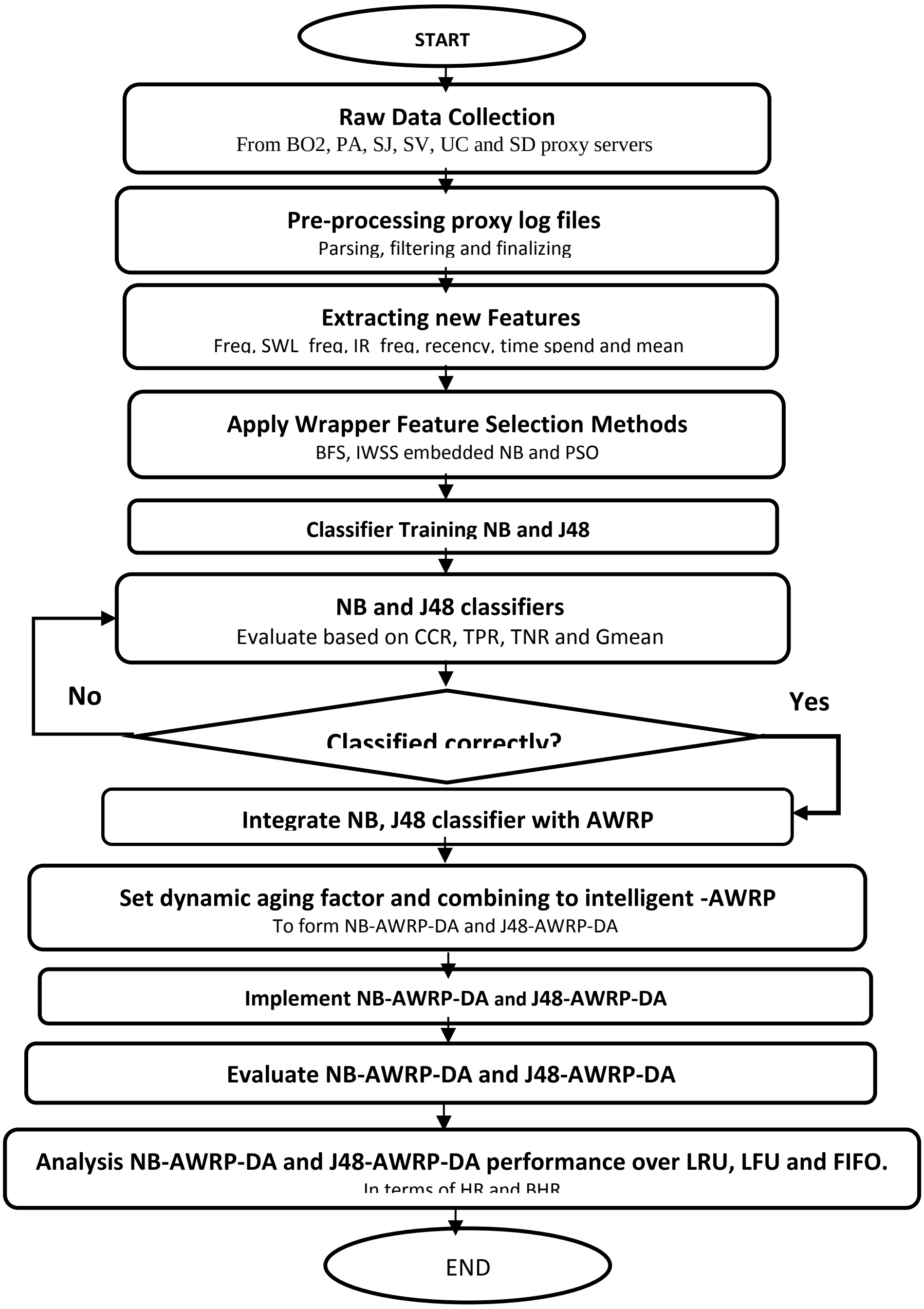

Figure 1.2: Intelligent Adaptive Weight Ranking Policy via dynamic aging replacement algorithms flowchart.

CHAPTER TWO

LITERATURE REVIEW

2.1 INTRODUCTION

Reviewing related works enable the researchers to have deep understanding about a specific area of study. This chapter presents a literature review of: conventional cache replacement algorithms, cache replacement algorithms, web cache replacement algorithms, intelligent web cache replacement algorithms and finally presents summary of the literature review.

2.2 CONVENTIONAL CACHE REPLACEMENT ALGORITHMS

Conventional cache replacement algorithms are the most used algorithms. They are applicable in CPU caches and virtual memory system but, they are not efficient in web caching area. Related works in (Ali et al., 2011; ElAarag, 2013; ElAarag & Romano, 2009), they explained most commonly algorithms are: Least Recently Used(LRU), Least Frequently Used (LFU), SIZE, Greedy-Dual-Size (GDS) and Greedy-Dual-Size-Frequency (GDSF).

LRU is the most common and simplest algorithm. It removes the least recently accessed object to make space for new object. However, it is not good in web caching because it does not take in to account object's size and download latency. LFU is considered as a base algorithm for frequency based algorithm class. It removes object that are least frequently used. However, it causes cache pollution with large reference accounts since they will not be evicted, even if they are not re-accessed.

In SIZE, largest object will be evicted first when the cache is full. Moreover, it depends on LRU algorithm when object's size is equal. Thus, objects with small size