การวิเคราะห์เปรียบเทียบกลวิธีในการลดความยาวของ URL

นายประเสริฐ วิชชุโอภาส

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2544
ISBN 974-03-0381-1

A COMPARATIVE STUDY OF ALGORITHMS FOR REDUCING URL LENGTH

Mr. Prasert Vitchu-o-pas

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2001

ISBN 974-03-0381-1

หัวข้อวิทยานิพนธ์	**การวิเคราะห์เปรียบเทียบกลวิธีในการลดความยาวของ URL**
โดย	**นายประเสริฐ วิชชุโอภาส**
สาขาวิชา	**วิทยาศาสตร์คอมพิวเตอร์**
อาจารย์ที่ปรึกษา	**อาจารย์ ดร. ณัฐวุฒิ หนูไพโรจน์**

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร. สมศักดิ์ ปัญญาแก้ว)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(อาจารย์ ดร. ยรรยง เต็งอำนวย)

.....อาจารย์ที่ปรึกษา
(อาจารย์ ดร. ณัฐวุฒิ หนูไพโรจน์)

.....กรรมการ
(รองศาสตราจารย์ ดร. วันชัย ริ้วไพบูลย์)

.....กรรมการ
(อาจารย์ ดร. ทวิตีย์ เสนีวงศ์ ณ อยุธยา)

ประเสริฐ วิชชุโอภาส : การวิเคราะห์เปรียบเทียบกลวิธีในการลดความยาวของยูอาร์แอล (THE COMPARATIVE STUDY OF ALGORITHMS FOR REDUCING URL LENGTH)
อ. ที่ปรึกษา : ดร. ณัฐวุฒิ หนูไพโรจน์, 68 หน้า, ISBN 974-03-0381-1

การทำงานของพร็อกซี่แคชเกี่ยวข้องกับการประมวลผลยูอาร์แอลเป็นจำนวนมาก ทั้งการระบุว่าข้อมูลที่เครื่องลูกข่ายต้องการอยู่ในแคชของตนหรือไม่ และการสอบถามข้อมูลระหว่างพร็อกซี่แคชด้วยกัน ล้วนแล้วแต่ต้องประมวลผลยูอาร์แอลทั้งสิ้น ถ้าหากสามารถเพิ่มประสิทธิภาพโดยรวมของการประมวลผลยูอาร์แอลก็จะเป็นการเพิ่มประสิทธิภาพพร็อกซี่แคชด้วยเช่นกัน จากการศึกษาในเบื้องต้นพบว่า การเข้ารหัสยูอาร์แอลที่ใช้งานในการทำงานพร็อกซี่แคช มีความต้องการคุณสมบัติของวิธีการเข้ารหัสที่แตกต่างกัน พร็อกซี่แคชในปัจจุบันมีการใช้วิธีการเข้ารหัสที่ชื่อว่า MD5 ในการย่อยูอาร์แอลก่อนนำไปประมวลผลหรือสอบถามข้อมูลระหว่างพร็อกซี่แคชด้วยกัน แต่ยังไม่มีงานวิจัยใดที่ทำการศึกษาวิธีการเข้ารหัสอื่นๆ ดังนั้นในงานวิจัยนี้จึงได้เลือกวิธีการเข้ารหัสแบบต่างๆขึ้นมาจำนวนหนึ่ง ทำการเข้ารหัสยูอาร์แอลเพื่อเปรียบเทียบ เวลาที่ใช้ในการเข้ารหัส ความยาวรหัส และปริมาณการชนกันของรหัส ของแต่ละวิธีการ และนำมาใช้ในการเสนอแนะแนวทางในการเลือกใช้วิธีการเข้ารหัสยูอาร์แอลที่มีประสิทธิภาพ สำหรับพร็อกซี่แคช

ข้อมูลยูอาร์แอลที่ใช้ในการทดสอบนำมาจากข้อมูลการใช้เว็บของจุฬาลงกรณ์มหาวิทยาลัย ข้อมูลนี้ถูกนำมาเข้ารหัสด้วยวิธีการเข้ารหัสต่างๆ ซึ่งเลือกขึ้นมา 12 วิธีการ โดยวิธีการเข้ารหัสเหล่านั้นเป็นที่รู้จักกันดี รวมทั้งวิธีการ MD5 ผลการทดลองทำให้ทราบว่าวิธีการเข้ารหัสที่มีความซับซ้อนน้อยใช้เวลาในการเข้ารหัสน้อยกว่าวิธีการที่มีความซับซ้อนมากและมีโอกาสที่เกิดการชนของรหัสมากกว่า วิธีการเข้ารหัสที่มีขนาดของรหัสที่สั้นมีโอกาสเกิดการชนกันของผลลัพธ์มากกว่า วิธีการที่เลือกมาทดสอบในการวิจัยไม่มีวิธีการเข้ารหัสใดที่ดีที่สุด การพิจารณาว่าวิธีการใดเหมาะสมจะพิจารณาจากความต้องการของแอปพลิเคชั่นเป็นหลัก โดยแอปพลิเคชั่นที่ต้องการความเร็วในการเข้ารหัสและรหัสที่สั้น ควรเลือกวิธีการ CRC-16, Digit Analysis method, Folding method แอปพลิเคชั่นที่ต้องการความเร็วในการเข้ารหัสและไม่ต้องการให้มีการชนกันของรหัสเลย ควรเลือกวิธีการในกลุ่ม MD หรือ Huffman Coding แต่ถ้าหากยอมให้มีการชนกันของรหัสได้บ้าง ควรเลือกวิธีการ CRC-32 หรือ Folding method และสำหรับแอปพลิเคชั่นที่ต้องการรหัสที่สั้นและมีปริมาณการชนกันของรหัสน้อย ควรเลือกวิธีการ CRC-32, Division method และ Folding method นอกจากนี้ ยังพบว่าวิธีการ CRC-32 และ Folding method ใช้เวลาในการเข้ารหัสน้อย รหัสที่ได้มีขนาดสั้น และมีการชนกันของรหัสน้อยมาก จึงอาจปรับปรุงโดยการรวมวิธีการทั้งสองเข้าด้วยกันเพื่อหาวิธีการใหม่ที่มีประสิทธิภาพมากขึ้นได้

ภาควิชา วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต .....
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์	ลายมือชื่ออาจารย์ที่ปรึกษา .....
ปีการศึกษา 2544	ลายมือชื่ออาจารย์ที่ปรึกษาร่วม .....

4171455821 : MAJOR COMPUTER SCIENCE
KEY WORD: WWW / PROXY CACHING / MESSAGE DIGEST ALGORITHM / URL
PRASERT VITCHU-O-PAS : THE COMPARATIVE STUDY OF ALGORITHMS FOR REDUCING URL LENGTH. THESIS ADVISOR : NATAWUT NUPAIROJ, Ph. D. 68 pp.
ISBN 974-03-0381-1

Webs Proxy Caches usually process large amounts of URLs to identify if requested pages are in the caches and to communicate among shared Web Caches. Improving URL processing can greatly enhance the performance of Web Caches. The preliminary study of this thesis shows that the requirements of URL processing for web caching application are varied.

Typically, many Web Proxy Caches use a hashing function, MD5, to encode URL. However, there is no study of a suitable algorithm for URL. In this study, we choose some well-known encoding or hashing algorithms, use them to encode URL in web access log and compare their encoding time, encoded length, and number of collisions. The comparison of the result is used as a guideline to select suitable URL encoding algorithm.

In our experiment, 12 well known coding and hashing algorithms include MD5 are selected to encode the cacheable URL from web access log of Office of Information Technology, Chulalongkorn University. We found that complicated algorithms use more time to encode URL but their encoded keys have less number of collisions. In addition, we found that algorithms that generate shorter length encoded keys lead to more number of collisions. Our studies indicate that there is no algorithm that can be considered the best for every Web Cache Applications. The algorithm that is suitable for each Web Cache Application is determined by their core requirements. For example, applications that speed and key length are critical should use CRC-16, Digit Analysis method, or Folding method as their main coding algorithm. Applications that require no collision and focus on fast speed should use a MD2, MD4, or MD5 algorithm. However, if some collisions are allowed, CRC-32 and Folding method are suitable. For the applications that require short length key and tolerate some collisions, CRC-32, Division method and Folding method should be selected. We also found that the CRC-32 and Folding method are faster than other algorithms and their encoded keys have very low collision rates. Therefor, combining these 2 algorithms may create a new powerful algorithm for Web Cache Application.

Department Computer Engineering	Student's signature
Field of study Computer Science	Advisor's signature
Academic year 2001	Co-advisor's signature

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความช่วยเหลืออย่างดียิ่งของ ดร. ณัฐวุฒิ หนูไพโรจน์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้ให้คำแนะนำและข้อคิดเห็นต่างๆในการวิจัยมาด้วยดีตลอด

ท้ายนี้ ผู้วิจัยใคร่ขอขอบพระคุณ บิดา – มารดา เพื่อนร่วมงาน ซึ่งให้ข้อคิดเห็น และสนับสนุนผู้วิจัยจนสำเร็จการศึกษา

สารบัญ

หน้า

สารบัญ (ต่อ)

สารบัญ (ต่อ)

บทที่ หน้า

สารบัญตาราง

หน้า