บทที่ 3
แนวทางการวิจัย

การวิจัยจะดำเนินการทดลองเข้ารหัสกับข้อมูลยูอาร์แอลด้วยอัลกอริทึมที่กำหนดในขอบเขตการวิจัยเพื่อวัดค่าดังนี้

1. **ความเร็วในการเข้ารหัส** (Speed)

ความเร็วในการเข้ารหัสเป็นสิ่งที่วัดประสิทธิภาพของอัลกอริทึมที่ใช้ได้ชัดเจนที่สุด การวัดความเร็วในการเข้ารหัสจะวัดจากเวลาที่ใช้ในการเข้ารหัส ถ้าหากใช้เวลาในการเข้ารหัสน้อยแสดงว่ามีความเร็วสูง ในการวัดค่าจะจับเวลาในการเข้ารหัสยูอาร์แอลแต่ละยูอาร์แอลในหน่วยไมโครวินาที (microsecond) หลังจากนั้นจะรวมเวลาที่ใช้ทั้งหมดเข้าด้วยกัน และนำมาหาค่าเฉลี่ยของเวลาที่ใช้ในการเข้ารหัสยูอาร์แอล 1 ยูอาร์แอล โดยการนำเวลารวมหารด้วยจำนวนยูอาร์แอลทั้งหมด และใช้ค่านี้ในการเปรียบเทียบประสิทธิภาพ

สาเหตุที่ต้องใช้ค่าเฉลี่ยที่ได้จากการจับเวลามากกว่า 1 รอบ ก็เพื่อลดผลกระทบอันเนื่องมาจากความแปรปรวนของการจับเวลา ที่เป็นผลมาจากการทำงานของระบบปฏิบัติการ หรือโปรแกรมอื่นๆ ที่ทำงานอยู่ซึ่งทำให้ผลการทดลองน่าเชื่อถือมากยิ่งขึ้น

2. **ขนาด หรือ ความยาวรหัสที่ได้**

ขนาดของรหัสหมายถึงความยาวของรหัสที่ได้ หน่วยเป็น บิต ไบต์ หรือเป็นตัวอักษร การวัดขนาดของรหัสที่ได้จะวัดในหน่วยเดียวกัน ขนาดของรหัสที่สั้นกว่าถือว่ามีประสิทธิภาพมากกว่า เพราะใช้เนื้อที่ในการเก็บข้อมูลน้อยกว่า เช่น อัลกอรทึมที่ได้รหัสจากยูอาร์แอลขนาด 2 ไบต์จะมีประสิทธิภาพในด้านขนาด มากกว่าอัลกอริทึมที่ได้รหัสขนาด 4 ไบต์ เป็นต้น ในการเปรียบเทียบขนาดของรหัสในการวิจัย จะเปลี่ยนหน่วยความยาวรหัสให้เป็นหน่วยเดียวกัน เนื่องจากผลลัพธ์จากอัลกอริทึมทั้งหมดมิได้ออกมาอยู่ในรูปแบบข้อมูลเดียวกัน เช่น อัลกอริทึมในกลุ่ม MD ได้รหัสที่มีความยาวในหน่วยเป็นบิต ส่วนในกลุ่ม Simple Hash Function ให้ผลออกในรูปของไบต์ เป็นต้น ดังนั้นในการเปรียบเทียบผลจะพิจารณาในระดับต่ำสุดของข้อมูลซึ่งก็คือหน่วยบิต

3. **ปริมาณการชนกันของข้อมูล** (Collision)

การชนกันของข้อมูล หมายถึง การเกิดการซ้ำกันของรหัส ซึ่งได้จากข้อมูลก่อนเข้ารหัสที่แตกต่างกัน รหัสที่ซ้ำกันจึงไม่สามารถระบุได้ว่าเกิดจากข้อมูลก่อนเข้ารหัสตัวใด การซ้ำกันของรหัสเกิดจากข้อมูลต้นฉบับมากกว่า 2 ก็ได้

การชนกันของรหัสส่งผลต่อความถูกต้องในการนำรหัสไปใช้งานในการระบุถึงข้อมูลต้นฉบับ ดังนั้นอัลกอริทึมซึ่งไม่เกิดการชนกันของรหัสที่ได้เลยถือว่ามีความถูกต้องสูงสุดและมีประสิทธิภาพสูงสุด แต่ในความเป็นจริงยิ่งรหัสสั้นกว่าข้อมูลจริงมากเท่าใด ย่อมมีโอกาสที่จะเกิดการชนกันของข้อมูลสูงมากขึ้นเท่านั้น เช่น อัลกอริทึมที่ให้รหัสที่มีความยาว 16 บิต ย่อมเกิดโอกาสชนกันของรหัสมากกว่าอัลกอริทึมที่ให้รหัสที่มีความยาว 32 บิต เป็นต้น

ในการวัดประสิทธิภาพของการชนกัน จะวัดจากปริมาณของการเกิดการชนกันของรหัสที่ได้ โดยนับจำนวนการซ้ำกันของรหัสที่เกิดขึ้นจากข้อมูลก่อนเข้ารหัสที่ไม่มีการซ้ำกันของข้อมูลเลย จำนวนครั้งของการซ้ำกัน จะนับจำนวนครั้งที่เกิดการซ้ำของข้อมูลดังนี้

สมมติว่ามียูอาร์แอลดังต่อไปนี้

```
http://disney.go.com/worldsofdisney/Tarzan/
http://1-888-watch.com/htmlos/0045.1.9224752747
http://members.dencity.com/Ram98/bows_y.gif
http://www.lovetop.com/pic/hellosex.wav
http://www.ba.cmu.ac.th/images/staff_____.jpg
http://www.geocities.com/Hollywood/Highrise/1212/oh15.htm
```

ซึ่งถ้าหากเข้ารหัสด้วยอัลกอริทึม CRC-16 จะได้ผลลัพธ์ดังนี้ คือ 0151, 5075, 0151, 0151, 0011, และ 0011 ตามลำดับ ดังนั้นจะนับว่ามีการชนของข้อมูลจำนวน 3 ครั้งซึ่งมาจากยูอาร์แอลที่ 3,4 และ 6 ตามลำดับ หรือมีการชนกันของรหัส 50% จากข้อมูลทั้งหมด หลังจากที่นับจำนวนครั้งที่มีการชนกันของข้อมูลแล้วจะนำมาคำนวนเปอร์เซ็นต์ของจำนวนครั้งที่มีการชนกันต่อจำนวนยูอาร์แอลทั้งหมดที่ไม่ซ้ำกันเลย แล้วนำค่าเปอร์เซ็นต์การชนกันนั้นมาเปรียบเทียบกันแต่ละอัลกอริทึม

3.1 ขั้นตอนในการทำวิจัย

3.1.1 ศึกษาการเข้ารหัสข้อมูลด้วยอัลกอริทึมตามที่กำหนดในขอบเขตของงานวิจัย

3.1.2 พัฒนาโปรแกรมในแต่ละอัลกอริทึมเพื่อใช้ในการเข้ารหัสยูอาร์แอล

3.1.3 เก็บข้อมูลยูอาร์แอลจากแหล่งข้อมูลที่พิจารณาเลือก

3.1.4 ศึกษาการจัดข้อมูลตามรูปแบบ squid log file และพัฒนาโปรแกรมเพื่อคัดยูอาร์แอลออกจาก log file

3.1.5 ทดลองเข้ารหัสด้วยอัลกอริทึมต่างๆ และบันทึกผลการทดลอง

3.1.6 วิเคราะห์ผลการทดลอง

3.1.7 สรุปผลการทดลองและข้อเสนอแนะ

3.1.8 จัดทำรายงาน

3.2 การพิจารณาเลือกใช้ข้อมูล

ยูอาร์แอลที่จะนำมาใช้ในการวิจัยจะต้องมาจากข้อมูลการใช้เว็บที่เกิดจากเรียกขอเข้ามายังเว็บพร็อกซี่ที่มีปริมาณมากพอ เพื่อให้ยูอาร์แอลที่ใช้ในการวิจัยสามารถใช้แทนยูอาร์แอลที่เกิดขึ้นในสภาวะการใช้งานจริงของเว็บพร็อกซี่โดยทั่วไปได้

จากการพิจารณาปริมาณการใช้เว็บของสำนักเทคโนโลยีสารสนเทศ จุฬาลงกรณ์มหาวิทยาลัย พบว่ามีปริมาณการเรียกขอเฉลี่ยมากกว่า 2 ล้านครั้งต่อวัน ซึ่งเป็นปริมาณการใช้ที่มากพอ จึงเลือกข้อมูลการใช้เว็บดังกล่าวช่วงเวลาระหว่างวันที่ 21-26 พฤศจิกายน 2542 ซึ่งเป็นช่วงเวลาเปิดเทอมวันจันทร์ถึงวันศุกร์ ปริมาณการเรียกขอในแต่ละวันแสดงในตารางที่ 3.1

Log file	ปริมาณการเรียกขอ (ครั้ง)
22 พ.ย. 2542	2,238,692
23 พ.ย. 2542	2,934,809
24 พ.ย. 2542	1,900,981
25 พ.ย. 2542	2,978,209
26 พ.ย. 2542	2,879,071

ตารางที่ 3.1 ปริมาณการเรียกขอของข้อมูลที่นำมาใช้ในการวิจัย

3.2.1 ลักษณะเฉพาะของยูอาร์แอล

ยูอาร์แอลเป็นสตริงที่มีลักษณะเฉพาะ ดังนี้

1. เป็นตัวอักษรภาษาอังกฤษ
2. ขึ้นต้นด้วย http://www., http://, ftp://
3. มีความยาวไม่แน่นอน (ไม่มีการกำหนดขนาดสูงสุดไว้)
4. ยูอาร์แอลของเว็บที่มีความนิยมสูงจะปรากฏมากเป็นพิเศษ

คุณสมบัติของยูอาร์แอลตามที่กล่าวมาอาจส่งผลกระทบต่อประสิทธิภาพในการเข้ารหัสและการพิจารณาความยาวของยูอาร์แอล เพื่อใช้ในการกำหนดความยาวของรหัสสำหรับบางอัลกอริทึม ดังนั้นในการวิจัยจึงมีการสร้างฮิสโตแกรมเพื่อวิเคราะห์การกระจายความยาวของยูอาร์แอล

3.2.2 การกำจัดข้อมูลที่มีการซ้ำกัน

จากหัวข้อที่แล้ว ยูอาร์แอลมีลักษณะเฉพาะประการหนึ่งก็คือ ยูอาร์แอลที่มีความนิยมสูงจะปรากฏยูอาร์แอลดังกล่าวซ้ำกันมากเป็นพิเศษ ดังนั้นในการเตรียมข้อมูลก่อนจะนำไป

เข้ารหัสจะต้องมีการกำจัดยูอาร์แอลที่ซ้ำกันออกให้เหลือเฉพาะยูอาร์แอลที่ไม่ซ้ำกันเลย (unique) เพื่อให้ผลการเข้ารหัสสามารถตรวจสอบได้ว่ามีการชนกันของผลลัพธ์หรือไม่ ซึ่งถ้าหากเกิดการซ้ำกันของผลลัพธ์ ก็แสดงว่าเกิดจากยูอาร์แอลที่แตกต่างกัน และเกิดการชนกันของผลลัพธ์ขึ้นอย่างแน่นอน

3.2.3 การคัดเลือกยูอาร์แอล

ยูอาร์แอลที่จะนำไปเข้ารหัสมีการคัดเลือกข้อมูลที่แตกต่างกัน 2 ลักษณะ คือ ในการทำการเข้ารหัสเพื่อทดสอบความเร็วในการเข้ารหัส จะเลือกเฉพาะยูอาร์แอลที่สามารถแคชได้เท่านั้น ส่วนยูอาร์แอลที่จะนำไปทดสอบการชนกันของรหัส นอกจากจะเป็นยูอาร์แอลที่แคชได้แล้ว ยังตัดยูอาร์แอลที่ซ้ำกันออกดังหัวข้อที่แล้วด้วย

ในการคัดเลือกยูอาร์แอลที่แคชได้นั้น จะพิจารณาจากสิ่งต่างๆ ต่อไปนี้

- นามสกุล หรือ Extension ของไฟล์ที่ปรากฏในยูอาร์แอล ไฟล์ที่มีนามสกุล .cgi, .jsp, .exe, .pl, .php, .dll, .cfm, .class ถือว่าเป็นยูอาร์แอลที่ไม่สามารถแคชได้ (non-cacheable)
- มีเครื่องหมาย ? ปรากฏในส่วนที่ไม่ได้เป็นชื่อโฮสต์ของยูอาร์แอล เครื่องหมาย ? นี้ แสดงว่ายูอาร์แอลดังกล่าวเป็นคิวรี่ (Query)

3.3 วิธีทำการทดลอง

การทดลองเข้ารหัสยูอาร์แอล (3.1.5) มีวิธีการดังต่อไปนี้

3.3.1 ขยายข้อมูลที่ได้จากสำนักเทคโนโลยี จุฬาลงกรณ์มหาวิทยาลัย ได้ผลลัพธ์ในรูปแบบของ Squid log file

3.3.2 คัดเฉพาะยูอาร์แอลออกจาก log file (squid log file) ซึ่งได้จากข้อมูลการใช้เว็บของสำนักเทคโนโลยีสารสนเทศ จุฬาลงกรณ์มหาวิทยาลัย

3.3.3 วิเคราะห์ข้อมูลยูอาร์แอลเพื่อหาการกระจายตัวของความยาวยูอาร์แอล บันทึกผลการทดลอง และสร้างฮิสโตแกรมของความยาวยูอาร์แอล

3.3.4 นำข้อมูลยูอาร์แอลที่ได้มาคัดเฉพาะยูอาร์แอลที่สามารถทำการแคชได้เท่านั้น

3.3.5 นำยูอาร์แอลที่ได้จากข้อ 3.3.4 ไปเข้ารหัสด้วยอัลกอริทึมต่างๆ ที่กำหนดในขอบเขตการวิจัย บันทึกเวลาที่ใช้ในการเข้ารหัสและความยาวของรหัสที่ได้

3.3.6 กรณีของ Division Method, Midsquare Method และ Huffman Coding เนื่องจากรหัสที่ได้มีความยาวไม่คงที่ ดังนั้นจะเก็บรหัสที่ได้เพื่อใช้ในการวิเคราะห์ความยาวรหัสต่อไป

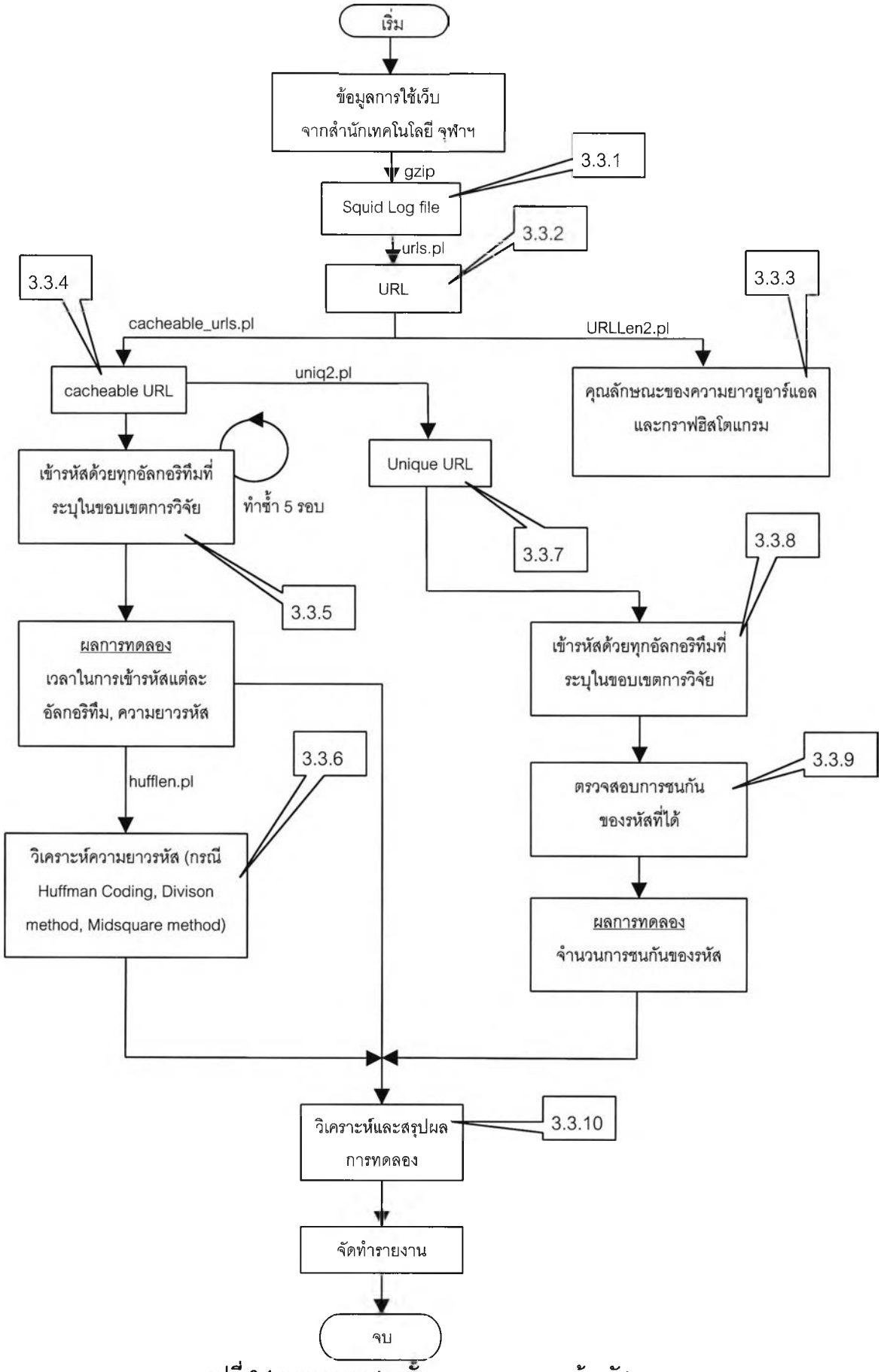
3.3.7 นำข้อมูลยูอาร์แอลในข้อ 3.3.4 มาตัดยูอาร์แอลที่ซ้ำกันออก

3.3.8 นำข้อมูลจาก 3.3.7 ไปเข้ารหัสด้วยอัลกอริทึมดังที่กำหนดไว้ในขอบเขตการวิจัย เก็บรหัสที่ได้ไว้

3.3.9 นำรหัสที่ได้ในข้อ 3.3.8 มานับจำนวนครั้งที่เกิดการชนกันของรหัส (Collision) และบันทึกผล

3.3.10 นำผลที่ได้จาก 3.3.9 และ 3.3.6 ไปวิเคราะห์และสรุปผลการทดลอง

ขั้นตอนในการทำการทดลองดังกล่าวสามารถสรุปเป็นแผนภาพได้ ดังรูปที่ 3.1



รูปที่ 3.1 แผนภาพแสดงขั้นตอนการทดลองเข้ารหัส

3.4 การพิจารณาเลือกภาษาโปรแกรม

ภาษาโปรแกรมที่ใช้ในการพัฒนาโปรแกรมเข้ารหัสทั้งหมดคือ ภาษาซี เนื่องจากเป็นภาษาโปรแกรมที่มีคุณสมบัติตรงตามความต้องการของงานวิจัย ดังนี้

1. ทุกอัลกอริทึมต้องใช้ภาษาเดียวกันในการพัฒนา เพื่อให้สามารถเปรียบเทียบผลได้
2. เป็นภาษาที่สามารถเข้าถึงและจัดการข้อมูลระดับต่ำได้อย่างมีประสิทธิภาพ
3. ภาษาที่ใช้ต้องทำงานได้เร็ว มีโอเวอร์เฮดที่เกิดจากตัวแปลภาษาน้อย เนื่องจากมีการพิจารณาเวลาที่ใช้ในการทำงาน
4. เป็นภาษาระดับสูงที่รู้จักกันดี พัฒนาได้เร็ว

เนื่องจากการวิจัยนี้ต้องการให้อัลกอรทึมเป็นตัวกำหนดความเร็วในการทำงานเท่านั้น ดังนั้นจึงต้องควบคุมปัจจัยอื่นให้ส่งผลกระทบต่อความเร็วในการทำงานของโปรแกรมน้อยที่สุดเท่าที่จะทำได้ ดังนั้นนอกจากการพิจารณาเลือกใช้ภาษาโปรแกรมแล้ว ทุกอัลกอริทึมจะมีโปรแกรมที่มีลักษณะคล้ายคลึงกันแตกต่างเฉพาะในส่วนของการเข้ารหัสเท่านั้น

สำหรับโปรแกรมวิเคราะห์ความยาวและตัดยูอาร์แอลใช้ภาษาเพิร์ล (Perl) ในการพัฒนา เนื่องจากเป็นภาษาที่ง่ายในการจัดการไฟล์ตัวอักษร และความเร็วในการทำงานไม่ส่งผลใดๆ ต่อผลการทดลอง

3.5 การทำการทดลองซ้ำ

ในการทดลอง นอกจากตัวโปรแกรมที่มีผลต่อเวลาที่ใช้ในการทำงานแล้ว ปัจจัยอื่นๆ เช่น สภาพแวดล้อมของเครื่องคอมพิวเตอร์ที่ใช้ทำการทดลอง เช่น การทำงานของระบบปฏิบัติการ การทำงานของโปรแกรมหรือดีมอน (daemon) อื่นๆ บนเครื่องที่ทำการทดลอง มีผลกระทบต่อเวลาที่ใช้ในการทำงานของโปรแกรมเข้ารหัสทั้งสิ้น เพื่อตัดการแปรปรวนจากปัจจัยต่างๆ และเพื่อให้ผลการทดลองที่ได้มีความน่าเชื่อถือมากขึ้น จึงต้องทำการทดลองมากกว่า 1 ครั้ง และใช้ค่าเฉลี่ยเป็นผลการทดลอง

จากผลการทดลองพบว่าไม่มีความแปรปรวนในแต่ละรอบมากนัก จึงกำหนดจำนวนครั้งในการทำซ้ำไว้ที่ 5 ครั้งในแต่ละอัลกอริทึม

3.6 ผลการทดลอง

ผลการทดลองส่วนที่เป็นเวลาที่ใช้ในการเข้ารหัส จะใช้ข้อมูลยูอาร์แอลเฉพาะที่สามารถทำการแคชได้ ตามการเรียกขอจริง (อาจมียูอาร์แอลที่ซ้ำกัน) ส่วนผลการทดลองที่เป็นความยาวของรหัส และปริมาณการชนกันของรหัสจะใช้ยูอาร์แอลที่สามารถแคชได้และตัดยูอาร์แอลที่ซ้ำกันออกแล้ว ผลการทำงานของโปรแกรมเข้ารหัสยูอาร์แอลเป็นดังนี้

1. *เวลาที่ใช้ในการเข้ารหัสข้อมูลขาเข้าทั้งไฟล์* ซึ่งได้จากการรวมเวลาที่ใช้ในการเข้ารหัสแต่ละยูอาร์แอลเข้าด้วยกัน (การจับเวลาในโปรแกรมจะจับเวลาในการเข้ารหัสแต่ละยูอาร์แอล)
2. *เวลาที่มากที่สุดที่ใช้ในการเข้ารหัสยูอาร์แอล 1 ยูอาร์แอล* ได้มาจากการเปรียบเทียบเวลาที่ใช้ในการเข้ารหัสแต่ละยูอาร์แอล และใช้ค่าเวลาที่มากที่สุด
3. *เวลาที่น้อยที่สุดที่ใช้ในการเข้ารหัสยูอาร์แอล 1 ยูอาร์แอล* ทำนองเดียวกับเวลามากที่สุด แต่ใช้ค่าเวลาที่น้อยที่สุดแทน
4. *เวลาเฉลี่ยที่ใช้ในการเข้ารหัสยูอาร์แอล 1 ยูอาร์แอล* ได้มาจากการนำผลลัพธ์ในข้อ 1 หารด้วยจำนวนยูอาร์แอลทั้งหมดที่เข้ารหัสในคราวนั้น
5. *จำนวนยูอาร์แอลที่ทำการเข้ารหัสทั้งหมด* ได้มาจากการนับจำนวนยูอาร์แอลที่ถูกเข้ารหัส และใช้ค่านี้เป็นตัวการในการหาค่าเฉลี่ยในข้อที่ 4
6. *รหัสของยูอาร์แอล* ที่เกิดจากอัลกอริทึมนั้นๆ

ผลการทดลองจะได้มาจากการทดลองซ้ำทั้งสิ้น 5 ครั้ง เพื่อลดผลจากปัจจัยอื่นๆ ดังที่ได้กล่าวไว้ในข้อ 3.5 จากนั้นจะหาค่าเฉลี่ยของผลการทดลองที่เป็น ค่าเวลาที่ใช้ทั้งหมด ค่าเวลาเฉลี่ยที่ใช้ต่อยูอาร์แอล ส่วนเวลามากที่สุดและเวลาน้อยที่สุด จะใช้เวลามากที่สุดและน้อยที่สุดจากการทดลองทั้ง 5 รอบ (ไม่ใช้ค่าเฉลี่ย)

ผลการทดลองจะนำเสนอในบทที่ 4 ซึ่งอยู่ในรูปตารางแสดงผลการทดลอง