บทที่ 4

การวิเคราะห์ผลการทำงาน

การวัดผลการทำงานของโปรแกรมการจัดทำดัชนีที่ออกแบบขึ้นนั้น ข้อมูลที่ใช้ในการวัดผลมีอยู่ 4 ประเภท ดังนี้

1. โคลงกลอน ลักษณะข้อมูลประเภทนี้จะมีคำที่เกิดจากการแผลงคำ การตัดคำ (ลดรูป) การเติมคำ ฯลฯ

2. ข่าว โดยเฉพาะข่าวต่างประเทศ จะมีคำที่ทับศัพท์ คำที่เป็นชื่อคน ชื่อสถานที่ต่างๆ ที่ไม่มีอยู่ในพจนานุกรม

3. เนื้อเพลง คำต่างๆ ในเนื้อเพลงมักจะเป็นคำที่มีอยู่ในพจนานุกรมทั้งสิ้น การนำมาวัดผลจึงเหมาะสำหรับการเปรียบเทียบ กับอัลกอริทึมการจัดทำดัชนีที่คำเหล่านั้นอยู่ในพจนานุกรมทั้งหมด

4. ข้อสอบเข้ามหาวิทยาลัยสายวิทยาศาสตร์ มีคำศัพท์เฉพาะ ศัพท์ทางวิทยาศาสตร์ รวมถึงวิชาภาษาไทย จะมีคำศัพท์ที่ไม่มีในพจนานุกรม

ขนาดของข้อมูลในแต่ละประเภทที่ใช้ในการทดสอบ มีขนาดดังตารางที่ 4.1

ประเภทของข้อมูล	ขนาดของข้อมูล (Bytes)
Poem	1950970
News	1572730
Lyric	1626092
Entrance	2360350

ตารางที่ 4.1 ขนาดของข้อมูลที่ใช้ในการทดสอบ

แต่ละประเภทของข้อมูลที่นำมาทดสอบนั้น จะมีรูปแบบของคำแตกต่างกัน โดยจะทำการทดสอบดังต่อไปนี้

1. จำนวนคำหลักที่ได้ เมื่อผ่านขั้นตอนการทำงานแต่ละขั้นตอน ของอัลกอริทึมการจัดทำดัชนีที่สร้างขึ้น

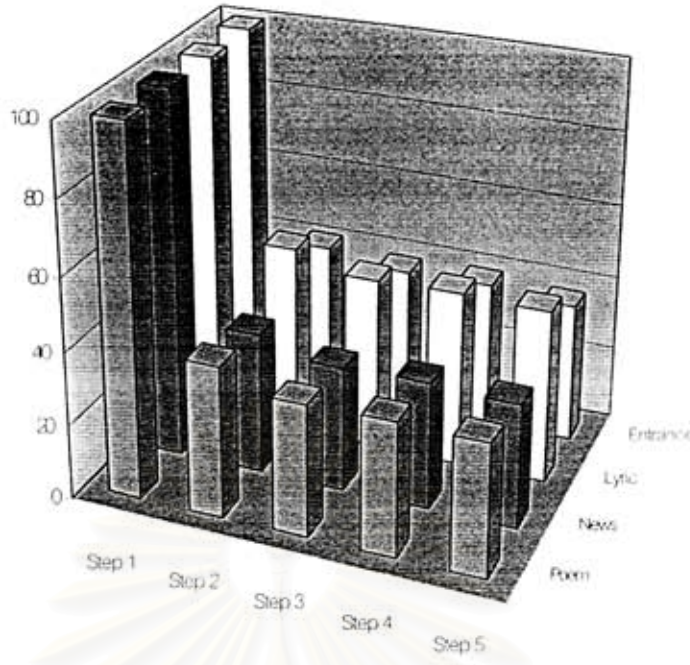
2. จำนวนของค่าน้ำหนักที่ได้ ในขั้นตอนที่ 2 ของอัลกอริทึมการจัดทำดัชนี

3. จำนวนคำหลักที่ได้ จากการใช้อัลกอริทึมการจัดทำดัชนีแบบต่างๆการเปรียบเทียบ

จำนวนคำหลักในแต่ละขั้นตอนที่ได้จากอัลกอริทึมการจัดทำดัชนี

	Step 1	Step 2	Step 3	Step 4	Step 5
Poem	100	41.01	35.76	36.40	35.99
News	100	38.10	33.75	34.62	33.70
Lyric	100	50.96	46.62	47.55	46.98
Entrance	100	41.41	38.62	39.37	37.61

ตารางที่ 4.2 ผลที่ได้จากการทดลอง แสดงเป็นเปอร์เซนต์
โดยให้ขั้นตอนแรกเท่ากับ 100 เปอร์เซนต์



รูปที่ 4.1 กราฟแสดงจำนวนเปอร์เซ็นต์ของคำหลัก
ที่ได้จากแต่ละขั้นตอนของอัลกอริทึมการจัดทำดัชนี

จากรูป 4.1

แกน x แสดงขั้นตอนของอัลกอริทึมการจัดทำดัชนีที่สร้างขึ้น

แกน y แสดงจำนวนเปอร์เซ็นต์ของคำหลักที่เกิดขึ้น

แกน z แสดงข้อมูลที่นำมาทดสอบ

ขั้นตอนที่ 1 แสดงถึงจำนวนของคำที่ได้จากการดึงคำที่มีอยู่ในพจนานุกรมออกมาทั้งหมด โดยกำหนดให้เป็นตัวอ้างอิงสำหรับขั้นตอนอื่นๆ

ขั้นตอนที่ 2 แสดงถึงจำนวนคำหลักที่ได้จากการทำงานของอัลกอริทึมการจัดทำดัชนีที่สร้างขึ้นในขั้นตอนแรกคือการหาคำที่ยาวที่สุดในแต่ละตำแหน่ง ผลที่ได้คือข้อมูลประเภทเนื้อเพลงจะมีจำนวนเปอร์เซนต์ของคำที่หาได้มากที่สุด เนื่องจากข้อมูลที่เป็นเนื้อเพลงจะเป็นข้อมูลที่ประกอบด้วยคำที่อยู่ในพจนานุกรมเกือบทั้งหมด จึงทำให้จำนวนเปอร์เซนต์ของคำหลักที่ได้ในขั้นตอนนี้มีจำนวนมากกว่าข้อมูลประเภทอื่นๆ

ขั้นตอนที่ 3 แสดงถึงจำนวนคำหลักที่ได้จากการหาค่าเส้นทางที่มีค่าน้ำหนักน้อยที่สุด ดังนั้นผลที่ได้จากการทดลอง ข้อมูลทุกประเภทจะมีจำนวนเปอร์เซ็นต์น้อยลง สังเกตข้อมูลประเภทเนื้อเพลง จะมีจำนวนเปอร์เซ็นต์มากที่สุด ซึ่งเป็นผลมาจากข้อมูลประเภทนี้ประกอบด้วยคำที่อยู่ในพจนานุกรมเป็นส่วนใหญ่นั้นเอง

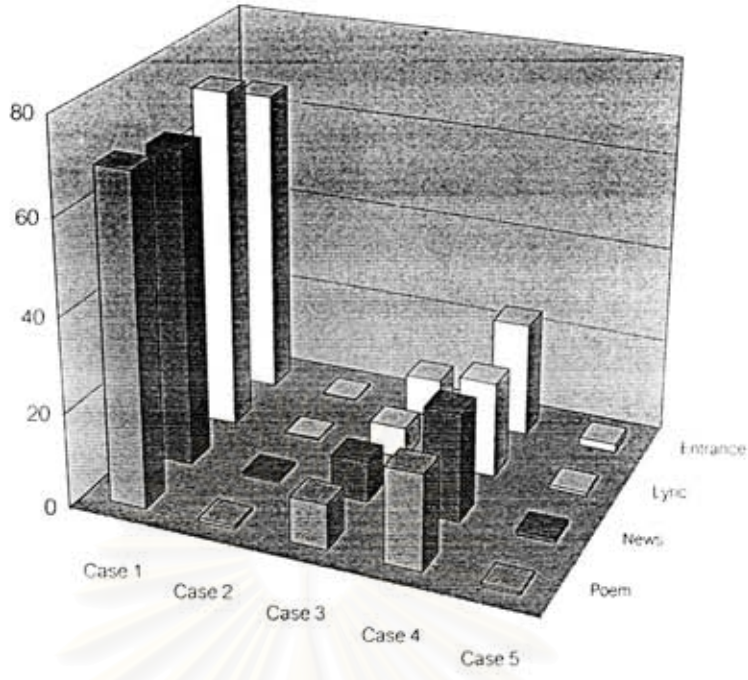
ขั้นตอนที่ 4 แสดงถึงจำนวนคำหลักที่ได้จากการเพิ่มคำที่จำเป็นต้องเก็บเพิ่ม รวมกับจำนวนคำหลักที่ได้จากขั้นตอนที่ 3

ขั้นตอนที่ 5 แสดงถึงจำนวนเปอร์เซ็นต์ของผลลัพธ์ที่ได้จากขั้นตอนสุดท้าย (ผลสุดท้ายของอัลกอริทึมการจัดทำดัชนี)

การเปรียบเทียบจำนวนของค่าน้ำหนักที่ได้ในขั้นตอนที่ 2 ของอัลกอริทึมการจัดทำดัชนี

	Case 1	Case 2	Case 3	Case 4	Case 5
Poem	69.59	0.66	9.39	19.85	0.51
News	67.07	0.72	8.54	22.76	0.91
Lyric	72.42	0.35	6.48	20.08	0.66
Entrance	65.46	0.41	8.22	24.27	1.64

ตารางที่ 4.3 ข้อมูลที่ได้จากการทดลอง แสดงเป็นเปอร์เซนต์ของคำที่เกิดขึ้นในแต่ละกรณี



รูปที่ 4.2 กราฟแสดงจำนวนค่าของน้ำหนักที่ได้จากขั้นตอนที่ 2 ของอัลกอริทึมการจัดทำดัชนี

จากรูป 4.2

แกน x แสดงค่าน้ำหนักในกรณีต่างๆ

แกน y แสดงจำนวนเปอร์เซ็นต์ของจำนวนที่เกิดขึ้น

แกน z แสดงข้อมูลที่นำมาทดสอบ

กรณี 1 คือกรณีที่คำสองคำต่อกันพอดี มีค่าน้ำหนักเท่ากับ 1

กรณี 2 คือกรณีที่คำสองคำทับกัน ส่วนที่ไม่ทับกันก็เป็นคำที่อยู่ในพจนานุกรมด้วย หรือคำสองคำที่นำมาพิจารณา สามารถดึงคำไปหลายรูปแบบ มีค่าน้ำหนักเท่ากับ 10

กรณี 3 คือกรณีที่คำสองคำทับกัน โดยมีส่วนที่ไม่ทับกันข้างหนึ่งข้างใด ที่เป็นคำที่อยู่ในพจนานุกรม มีค่าน้ำหนักเท่ากับ 100

กรณี 4 คือกรณีที่คำสองคำทับกัน และไม่มีส่วนที่ไม่ทับกันอยู่ในพจนานุกรม มีค่าน้ำหนักเท่ากับ 1000

กรณี 5 คือเกิดช่องว่างระหว่างคำสองคำที่พิจารณา

ข้อสังเกตจากการทดลอง

ในกรณีที่ 1 คือกรณีที่คำสองคำอยู่ติดกัน จะเป็นกรณีที่มีเปอร์เซ็นต์มากที่สุดที่เกิดขึ้นกับข้อมูลทุกแบบ

กรณีที่ 2 จะเป็นกรณีที่มีโอกาสเกิดขึ้นน้อยที่สุด

กรณีที่ 2, 3, 4 และ 5 จะเป็นกรณีที่ทำให้ต้องเพิ่มจำนวนของคำหลัก

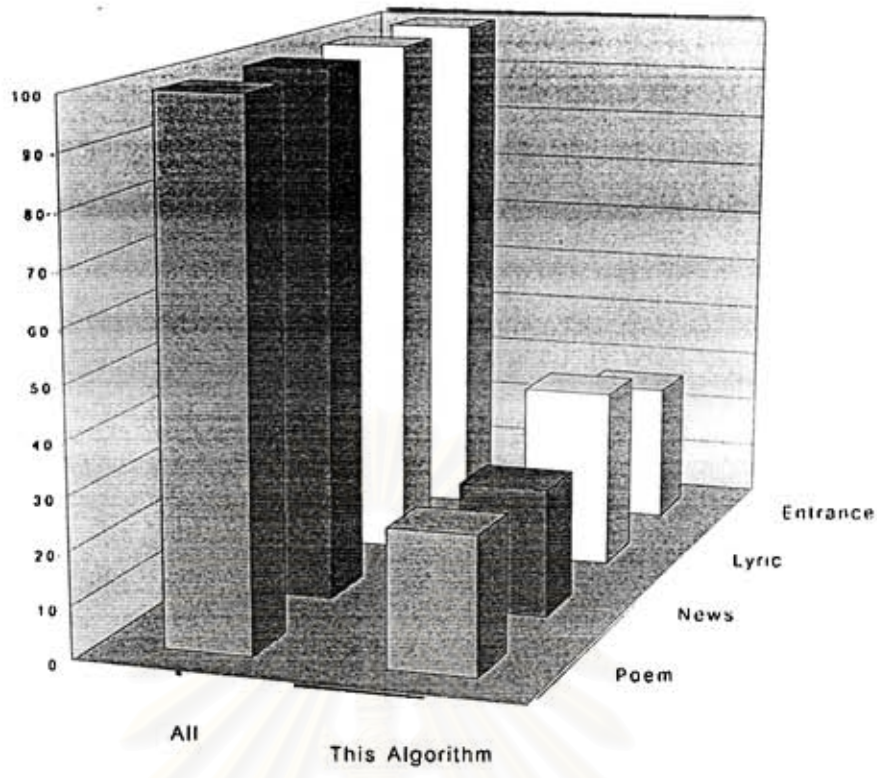
กรณีที่ 4 และ 5 แสดงถึงจำนวนคำที่ไม่มีอยู่ในพจนานุกรม ถ้าทั้งสองกรณีนี้มีจำนวนมาก หมายถึงข้อมูลนั้นๆ มีคำที่ไม่มีอยู่ในพจนานุกรมมาก

ข้อมูลประเภทเนื้อเพลง เป็นข้อมูลที่ประกอบไปด้วยคำที่มีอยู่ในพจนานุกรมมากกว่าข้อมูลประเภทอื่นๆ สังเกตได้จากกราฟกรณีที่ 4 และ 5 รวมกันมีจำนวนน้อยที่สุด

การเปรียบเทียบจำนวนคำหลักที่ได้จากอัลกอริทึมต่างๆ

	All	This Algorithm
Poem	100	25.79
News	100	24.27
Lyric	100	34.09
Entrance	100	26.77

ตารางที่ 4.4 ผลที่ได้จากการทดลอง แสดงเป็นเปอร์เซนต์
โดยให้การดึงคำทุกคำมีค่าเท่ากับ 100 เปอร์เซนต์



รูปที่ 4.3 กราฟแสดงจำนวนคำหลักที่ได้จากอัลกอริทึมต่างๆ

จากรูป 4.3

แกน x แสดงถึงรูปแบบของอัลกอริทึมที่ใช้

แกน y แสดงจำนวนเปอร์เซ็นต์ของจำนวนคำหลักที่ได้

แกน z แสดงข้อมูลที่นำมาทดสอบ

อัลกอริทึมที่ 1 คือการดึงคำทุกคำที่เป็นไปได้ รวมกับกลุ่มของตัวอักษรที่ไม่มีอยู่ในพจนานุกรม

อัลกอริทึมที่ 2 คืออัลกอริทึมที่คิดค้นขึ้น

ข้อสังเกตจากการทดลอง

กราฟที่แสดงผลจะให้จำนวนคำหลักที่ได้จากอัลกอริทึมที่ 1 เป็นตัวเปรียบเทียบกับอัลกอริทึมอื่นๆ

ข้อมูลประเภทเนื้อร้อง เป็นข้อมูลที่มีจำนวนเปอร์เซนต์ของคำที่มีอยู่ในพจนานุกรมมากกว่าข้อมูลประเภทอื่นๆ จึงทำให้จำนวนเปอร์เซนต์ที่ได้ในอัลกอริทึมที่ 2 มีค่ามากกว่าข้อมูลประเภทอื่นๆ ด้วย

จำนวนคำหลักที่ได้จากอัลกอริทึมที่ 2 จะมีจำนวนน้อยกว่าอัลกอริทึมที่ 1 เพราะเมื่อเกิดช่วงของกลุ่มตัวอักษรที่ไม่สามารถดึงคำได้อย่างถูกต้อง จะใช้กฎในการแบ่งพยางค์ช้าช่วยทำให้จำนวนคำหลักที่ได้ลดลง ไม่จำเป็นต้องเก็บคำจำนวนมากๆ อย่างในอัลกอริทึมที่ 1

ประสิทธิภาพในแต่ละขั้นตอนของอัลกอริทึม

ขั้นตอนที่ 1 มี n sistrings ของ T (n คือจำนวนตัวอักษรของ T) เพราะว่าพจนานุกรมให้โครงสร้างข้อมูลแบบ trie ในการเก็บ ดังนั้นในขั้นตอนนี้ใช้เวลา $O(n*k)$ โดย k คือขนาดของคำในพจนานุกรมที่มีความยาวมากที่สุด

ขั้นตอนที่ 2 สร้างกราฟ $G = (V, E)$ ใช้เวลา $O(|V| + |E|)$ กรณีที่ไม่ดีที่สุดมีค่าเท่ากับ $O(n^2)$

ขั้นตอนที่ 3 ค้นหากราฟที่มีค่าน้ำหนักน้อยที่สุด มีค่าเท่ากับ $O(n^2)$

ขั้นตอนที่ 4 ค้นหากลุ่มตัวอักษรที่ไม่รู้จัก มีค่าเท่ากับ $O(n)$

ดังนั้นอัลกอริทึมการจัดทำดัชนีที่นำเสนอในกรณีที่ไม่ดีที่สุด ใช้เวลาเท่ากับ $O(n*k + n^2)$ หรือ $O(n^2)$ เมื่อ $k < n$ สำหรับข้อความยาว