บทที่ 4
การทดลองและผลการทดลองของการแสดงผลภาพบิตแม็บ

การทดลองและผลการทดลองที่นำเสนอในบทนี้ เป็นการทำการทดลองเพื่อทดสอบความเป็นไปได้ของแนวทางการวิจัย และเป็นการทดลองเพื่อหาค่าพารามิเตอร์ที่เหมาะสมกับลักษณะของข้อมูลเอกสารที่นำมาทดลอง เพื่อทำให้การแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัลสามารถทำงานได้อย่างมีประสิทธิภาพ

โดยบทนี้กล่าวถึงข้อมูลชนิดต่างๆ ที่นำมาใช้ในการทดลอง การวิเคราะห์ข้อมูลเอกสารและทำการทดลองเพื่อหาตัวแปรหรือพารามิเตอร์สำคัญที่เหมาะสม กับการแสดงผลภาพบิตแม็บ และผลการทดลองกับข้อมูลชนิดต่างๆ ที่เลือกนำมาทำการทดลอง โดยมีรายละเอียดของแต่ละขั้นตอน ดังนี้

4.1 ข้อมูลที่ใช้ในการทดลอง

ข้อมูลที่นำมาทดลองในงานวิจัยประกอบไปด้วยข้อมูล 3 ประเภท คือ ข้อมูลดีเอ็นเอ หรือข้อมูลหน่วยพันธุกรรม ข้อมูลคลื่นหัวใจ และข้อมูลเอกสารดิจิทัล ซึ่งข้อมูลทั้งหมดที่นำมาทดลองเป็นข้อมูลที่เผยแพร่ทางอินเทอร์เน็ต ซึ่งรายละเอียดและแหล่งที่มาของข้อมูลที่ใช้ในการทดลองมีดังนี้

4.1.1 ข้อมูลดีเอ็นเอ

ดีเอ็นเอ (DNA) หรือข้อมูลหน่วยพันธุกรรม มีชื่อวิทยาศาสตร์ว่า กรดดีออกซีไรโบนิวคลีอิก (Deoxyribonucleic Acid) ที่พบในเซลล์ของสิ่งมีชีวิตทุกชนิด ได้แก่ คน สัตว์ พืช เชื้อรา แบคทีเรีย และไวรัส เป็นต้น ดีเอ็นเอบรรจุข้อมูลทางพันธุกรรมของสิ่งมีชีวิตชนิดนั้นไว้ โดยลักษณะข้อมูลดีเอ็นเอประกอบไปด้วยอักขระที่แทนสัญลักษณ์เบสจำนวน 4 ชนิด คือ A C G และ T ซึ่งข้อมูลทางพันธุกรรมในสิ่งมีชีวิตชนิดต่างๆ เกิดขึ้นจากการเรียงลำดับของเบสในดีเอ็นเอนั่นเอง

ข้อมูลดีเอ็นเอ หรือข้อมูลหน่วยพันธุกรรม ที่นำมาใช้ในงานวิจัยนี้ประกอบไปด้วยข้อมูลดีเอ็นเอของสัตว์ชนิดต่างๆ จำนวน 4 ชนิด 13 สายพันธุ์ ซึ่งนำมาจากธนาคารฐานข้อมูลดีเอ็นเอของดีดีบีเจ ประเทศญี่ปุ่น (DDBJ-DNA Data Bank of Japan) โดยมีรายละเอียดของข้อมูลดีเอ็นเอของสัตว์ชนิดต่างๆ ดังตารางที่ 4.1

ตารางที่ 4.1 รายละเอียดข้อมูลดีเอ็นเอที่นำมาใช้ในการทดลอง
(ที่มา: http://www.ddbj.nig.ac.jp)

ลำดับที่	ประเภทข้อมูล	ความยาว (อักขระ)
1	Chlamydophila Pneumoniae (TW-183 Section 1)	300,901
2	Chlamydophila Pneumoniae (TW-183 Section 3)	310,070
3	Chlamydophila Pneumoniae (TW-183 Section 4)	335,572
4	Mycobacterium Bovis AF2122/97	354,484
5	Mycobacterium Tuberculosis H37Rv	326,586
6	Mycobacterium Paratuberculosis K-10	4,829,781
7	Mycobacterium Tuberculosis CDC1551	4,403,837
8	Francisella Tularensis Holarctica strain LVS	1,959,194
9	Francisella Tularensis Tularensis strain FSC	1,892,616
10	Macaca Mulatta clone CH250-135N17	385,507
11	Macaca Mulatta clone CH250-253A7	450,117
12	Macaca Mulatta clone CH250-155N20	356,991
13	Macaca Mulatta clone CH250-190E12	374,468

ตารางที่ 4.1 แสดงรายละเอียดข้อมูลดีเอ็นเอของสัตว์ชนิดต่างๆ ที่นำมาใช้ในงานวิจัยนี้ ซึ่งสัตว์แต่ละชนิดเป็นสัตว์ที่มีความแตกต่างกันอย่างชัดเจน โดยมีรายละเอียดของสัตว์แต่ละชนิดดังนี้

- แบคทีเรีย Chlamydophila Pneumoniae สายพันธุ์ TW-183

เป็นแบคทีเรียที่ทำให้เกิดโรคปอดบวมในมนุษย์ โดยแต่ละข้อมูลเป็นดีเอ็นเอของ Chlamydophila pneumoniae ที่ถูกแบ่งออกเป็น 3 ส่วน

- แบคทีเรีย Mycobacterium

เป็นแบคทีเรียที่เป็นสาเหตุหลักที่ทำให้เกิดการติดเชื้อในสัตว์เลี้ยงลูกด้วยนม ซึ่งเลือกมาจำนวน 4 สายพันธุ์ คือ AF2122/97 H37Rv K-10 และ CDC1551 โดยแต่ละสายพันธุ์ถูกค้นพบตามสถานที่และสิ่งแวดล้อมที่แตกต่างกัน

- แบคทีเรีย Francisella Tularensis

เป็นแบคทีเรียที่เป็นสาเหตุทำให้เกิดโรคทูละรีเมีย (Tularemia) หรือ ไข้กระต่าย โดยมีอาการไข้เป็นพักๆ ถูกค้นพบทางตอนเหนือของประเทศสหรัฐอเมริกา ซึ่งเลือกมาจำนวน 2 สายพันธุ์ คือ สายพันธุ์ LVS และ FSC

- ลิง Macaca Mulatto

เป็นลิงสายพันธุ์ที่มีชื่อว่า Macaca Mulatto สามารถพบได้ในบริเวณประเทศแอฟกานิสถานทางตะวันตก ประเทศอินเดีย และทางตอนเหนือของประเทศไทย โดยแบ่งได้เป็นหลายสปีชีส์ (Species) ซึ่งเลือกมาจำนวน 4 สปีชีส์ มีรหัสที่ถูกกำหนดไว้ทางชีววิทยา คือ CH250-135N17 CH250-253A7 CH250-155N20 และ CH250-190E12

4.1.2 ข้อมูลคลื่นหัวใจ

ข้อมูลคลื่นหัวใจ (ECG - Electrocardiograms) เป็นข้อมูลสัญญาณไฟฟ้าที่เกิดจากการเต้นของหัวใจที่มีข้อมูลเกิดขึ้นตามช่วงเวลา คลื่นหัวใจถูกแสดงเป็นภาพกราฟที่มีลักษณะต่อเนื่อง ซึ่งข้อมูลภาพกราฟดังกล่าวคือข้อมูลที่ประกอบไปด้วยเลขจำนวนจริงที่เกิดขึ้นตามช่วงเวลา โดยทั่วไปจะเกิดข้อมูลที่เป็นเลขจำนวนจริงทุกๆ 0.20 วินาที ที่หัวใจทำงาน

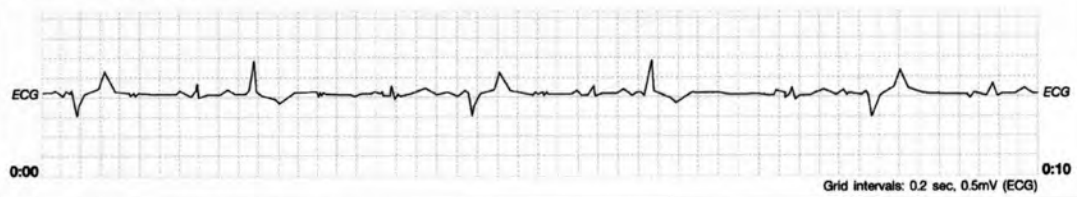
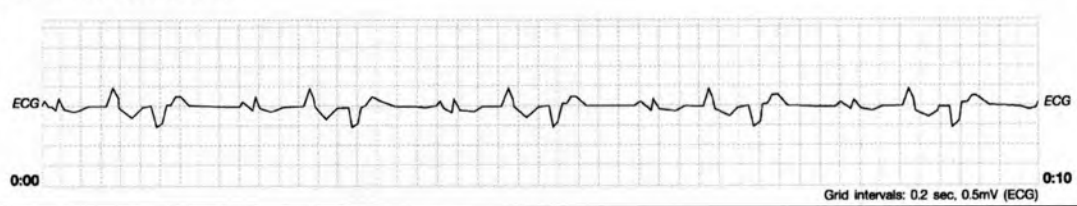
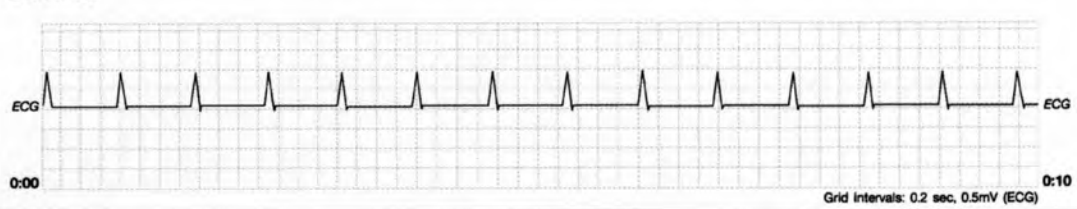
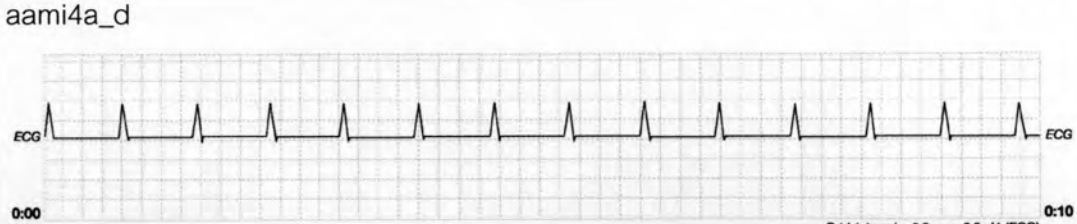
ข้อมูลคลื่นหัวใจที่นำมาใช้ในงานวิจัยนี้ คือข้อมูลของผู้ป่วยที่มีอาการโรคหัวใจในภาวะต่างๆ เช่น ภาวะกล้ามเนื้อหัวใจตีบตัน ภาวะลิ้นหัวใจรั่ว เป็นต้น ซึ่งเป็นข้อมูลที่นำมาจากแหล่งข้อมูลของฟิซิโอเน็ต (PhysioNet - http://www.physionet.org) ซึ่งเป็นองค์กรสนับสนุนการวิจัยที่ไม่แสวงผลกำไร โดยมีรายละเอียดและลักษณะของข้อมูล ดังตารางที่ 4.2

4.1.3 ข้อมูลเอกสารดิจิทัล

ข้อมูลเอกสารดิจิทัลจะประกอบไปด้วยข้อมูลตัวอักษร ตัวเลข และอักขระพิเศษต่างๆ ตามมาตรฐานของแอสกี (ASCII) โดยไม่รวมถึงรูปภาพหรือวัตถุอื่นๆ ซึ่งเป็นไปตามวัตถุประสงค์ของงานวิจัย ที่เน้นการวิเคราะห์ข้อมูลแบบข้อความพื้นฐานเป็นหลัก ข้อมูลเอกสารดิจิทัลทุกเอกสาร ถูกรวบรวมมาจากแหล่งข้อมูลที่เผยแพร่ทางอินเทอร์เน็ต โดยจำแนกข้อมูลเอกสารดิจิทัลได้เป็น 4 กลุ่ม คือ เอกสารพระคัมภีร์ไบเบิล เอกสารบทละครโทรทัศน์ เอกสารนิยายวิทยาศาสตร์ และ เอกสารวิชาการทางคอมพิวเตอร์ ซึ่งเห็นได้ชัดเจนว่ากลุ่มเอกสารดังกล่าวมีความแตกต่างกันอย่างชัดเจน โดยมีรายละเอียดของข้อมูลเอกสารดิจิทัลทั้งหมด ดังตารางที่ 4.3

โดยเอกสารพระคัมภีร์ไบเบิลถูกแบ่งออกเป็น 4 ส่วน ซึ่งเป็นเนื้อหาของพระคัมภีร์ไบเบิลเล่มเดียวกัน ส่วนบทละครเรื่องเฟรนด์และเรื่องวิลล์และเกรซ ถูกนำมาจากตอนต่างๆ หลายๆ ตอนรวมกันและตัดแบ่งเป็นเอกสารจำนวนเรื่องละ 5 เอกสาร

ตารางที่ 4.2 รายละเอียดข้อมูลคลื่นหัวใจ (ECG) ชุดข้อมูล "ANSI/AAMI EC13 Test Waveforms"

ลำดับที่	ชุดข้อมูล
1	aami-ec13/aami3a 
2	aami-ec13/aami3b 
3	aami-ec13/aami3d 
4	aami4a 
5	aami4a_d 
6	aami4a_h 

ตารางที่ 4.3 ข้อมูลเอกสารดิจิทัลที่นำมาใช้ในการทดลอง

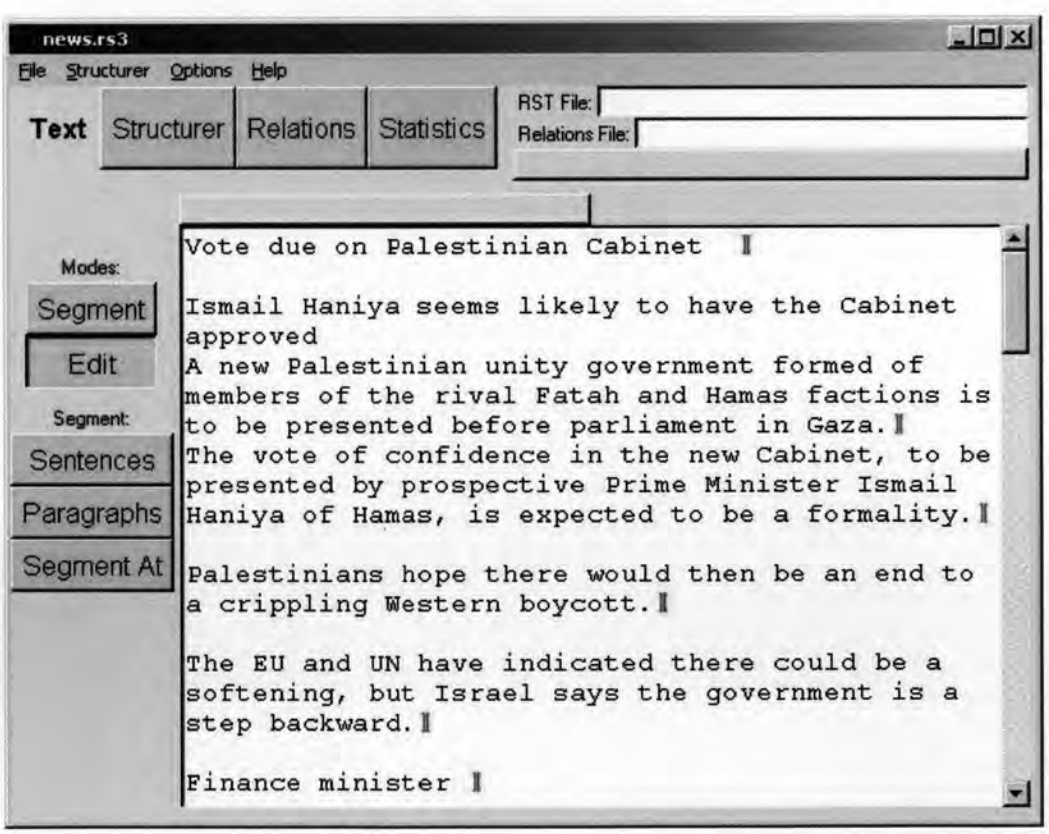
ลำดับที่	รายการเอกสาร	จำนวน (เอกสาร)
1	พระคัมภีร์ไบเบิลส่วนที่ 1 ถึง 4 (ที่มา: http://www.o-bible.com)	4
2	บทละครเรื่องเฟรนด์ (Friends Scripts) ตอนที่ 1 ถึง 10 (ที่มา: http://www.geocities.com/vspramod/links/friends/friends.htm)	5
3	บทละครเรื่องวิลล์และเกรซ (Will & Grace Script) ตอนที่ 1 ถึง 8 (ที่มา: http://www.durfee.net/will/)	5
4	เอกสารวิชาการคอมพิวเตอร์เรื่อง The Art of Assembly Language (ที่มา: http://www.planetpdf.com)	1
5	เอกสารวิชาการคอมพิวเตอร์เรื่อง Object-oriented programming with ANSI-C (ที่มา: http://www.planetpdf.com)	1
6	เอกสารวิชาการคอมพิวเตอร์เรื่อง Thinking in C++ (ที่มา: http://www.planetpdf.com)	1
7	เอกสารวิชาการคอมพิวเตอร์เรื่อง Thinking in Java (ที่มา: http://www.planetpdf.com)	1
8	นิยายวิทยาศาสตร์เรื่อง Arturius - A Quest For Camelot (ที่มา: http://www.legendofkingarthur.com)	1
9	นิยายวิทยาศาสตร์เรื่อง The Apocalypse Troll (ที่มา: http://www.webscription.net)	1
10	นิยายวิทยาศาสตร์เรื่อง Crusade (ที่มา: http://www.webscription.net)	1
11	นิยายวิทยาศาสตร์เรื่อง VirtuallyReal (ที่มา: http://www.angiehulme.com)	1
	รวม	22

4.2 การเลือกใช้ค่าพารามิเตอร์ที่เหมาะสม

ขั้นตอนในการประมวลผลภาพบิตแม็บจากข้อมูลเอกสาร จำเป็นต้องเลือกใช้พารามิเตอร์ที่สามารถทำให้ผลของภาพบิตแม็บที่ถูกประมวลผลมาจากข้อมูลเอกสาร มีประสิทธิภาพ มีความชัดเจน และใช้เวลาในการประมวลผลน้อยที่สุด การกำหนดและเลือกใช้พารามิเตอร์ดังกล่าวจึงเป็นตัวแปรที่สำคัญสำหรับการแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล ซึ่งมีแนวคิดและขั้นตอนการได้มาของพารามิเตอร์ที่สำคัญดังนี้

4.2.1 ค่าสัดส่วนจำนวนเฉลี่ย

ค่าสัดส่วนจำนวนเฉลี่ย คือ ค่าที่ใช้ในการลดขนาดหรือมิติ (Dimensionality Reduction) ของข้อมูลอนุกรมเวลา ดังรายละเอียดในหัวข้อ 3.2.1 ซึ่งการกำหนดค่าสัดส่วนจำนวนเฉลี่ยที่เหมาะสม มาจากแนวคิดในการวิเคราะห์รูปแบบข้อมูลจากประโยคของเอกสาร จึงทำการศึกษาลักษณะและรูปแบบของประโยค จากการประมาณขนาดความยาวตัวอักษรเฉลี่ยในแต่ละประโยคของเอกสาร งานวิจัยนี้เลือกใช้โปรแกรมการตัดแบ่งประโยค RSTTool [13] เวอร์ชัน 3.45 ซึ่งเป็นโปรแกรมที่สามารถทำการแบ่งแยกประโยคในเอกสารภาษาอังกฤษ และยังสามารถทำการปรับแต่งผลการแบ่งแยกประโยคเพิ่มเติมจากโปรแกรมได้เองอีกด้วย ดังรูปที่ 4.1



รูปที่ 4.1 โปรแกรม RSTTool เวอร์ชัน 3.45

(ที่มา: http://www.wagsoft.com)

โปรแกรม RSTTool เวอร์ชัน 3.45 ทำงานโดยการอ่านข้อความในเอกสาร แล้วทำการตัดแบ่งประโยค ซึ่งโปรแกรมทำการพิจารณาตัดแบ่งประโยคจาก สัญลักษณ์หรือเครื่องหมายจบประโยค ยกตัวอย่างเช่น เครื่องหมาย . ? และ ! เป็นต้น โปรแกรมทำการแสดงผลการตัดแบ่งประโยคโดยการสร้างเครื่องหมายขั้นระหว่างประโยคในหน้าจอแสดงผล ทำให้สามารถสังเกตผลการตัดแบ่งประโยคและทำการแก้ไขเพิ่มเติมได้ทันที

เมื่อใช้โปรแกรม RSTTool ทำการตัดแบ่งประโยคจากเอกสารตัวอย่าง เพื่อหาความยาวเฉลี่ยของตัวอักษรในแต่ละประโยค โดยคัดเลือกประโยคที่เป็นหัวข้อเรื่อง และข้อความที่ไม่มีลักษณะเป็นประโยคออก เพราะอาจทำให้การหาความยาวเฉลี่ยของประโยคเกิดความคลาดเคลื่อนได้ ซึ่งแสดงประโยคและความยาวของแต่ละประโยค ดังตารางที่ 4.4

ตารางที่ 4.4 ประโยคและความยาวตัวอักษรที่ทำการแบ่งแยกด้วยโปรแกรม RSTTool

ตัวอย่างประโยค	ความยาว (ตัวอักษร)
Ismail Haniya seems likely to have the Cabinet approved A new Palestinian unity government formed of members of the rival Fatah and Hamas factions is to be presented before parliament in Gaza.	192
The vote of confidence in the new Cabinet, to be presented by prospective Prime Minister Ismail Haniya of Hamas, is expected to be a formality.	143
Palestinians hope there would then be an end to a crippling Western boycott.	76
The EU and UN have indicated there could be a softening, but Israel says the government is a step backward.	107
The parliament is due to meet at 110 local time (900 GMT) to hear a speech from Palestinian Authority President Mahmoud Abbas, who is also the leader of Fatah.	159
Mr Haniya will then present his planned Cabinet and read a policy speech before the vote of confidence.	103
Israel has indicated it will deal with only Mahmoud Abbas If ratified, the ministers will then be sworn in at Mr Abbas's office.	128
The Palestinian economy has been badly hit by the international embargo.	72
It was imposed after the election victory in January last year of Hamas, which rejects international calls for it to recognise Israel and renounce violence.	156

ตารางที่ 4.4 (ต่อ) ประโยคและความยาวตัวอักษรที่ทำการแบ่งแยกด้วยโปรแกรม RSTTool

ตัวอย่างประโยค	ความยาว (ตัวอักษร)
The BBC's Matthew Price in Jerusalem says the new government contains a cross section of Palestinian parties, including some ministers who recognise Israel.	156
The US has also indicated it may leave the door open to some contact with the proposed finance minister.	104
Salam Fayyad is a Western-backed economist who is thought to be respected by the Bush administration.	101
One US official said Washington would not deal with him officially but might consider unofficial contacts.	106
Israel, however, said it would shun the new administration.	59
Deputy Defence Minister Ephraim Sneh said on Friday that Israel should try to deal with only Mr Abbas as a means to "drive Hamas out of power".	143
Although there have been signs of a softening in the international stance towards the new government, particularly by France and Russia, there are no guarantees the international boycott will end.	196
Britain has said it will only have diplomatic contact with non-Hamas members of the government.	95
The US says the administration must accept Israel's right to exist, renounce violence and conform to past peace deals but has otherwise reserved judgment.	154
The new administration was forged after several months of fighting between the Hamas and Fatah factions left more than 140 people dead.	135
Saturday's vote comes amid increasing lawlessness in the Gaza Strip.	68
There has been a series of abductions over recent months of Western aid workers and journalists.	96
Intensive efforts are continuing to find missing BBC Gaza correspondent Alan Johnston, who is feared kidnapped.	111
ค่าเฉลี่ย	120.58

จากตารางที่ 4.4 แสดงประโยคที่ทำการตัดแบ่งมาจากเอกสาร และขนาดความยาวของตัวอักษรในแต่ละประโยคที่นับรวมช่องว่าง และอักขระพิเศษต่างๆ โดยมีความยาวเฉลี่ยประมาณ 120 ตัวอักษรต่อประโยค ซึ่งเป็นค่าเริ่มต้นที่สามารถใช้เป็นแนวทางในการกำหนดค่าสัดส่วนจำนวนเฉลี่ยที่เหมาะสมได้

เนื่องจากความยาวเฉลี่ยของประโยคในเอกสารบางชนิด อาจมีความแตกต่างกันบ้าง ขึ้นกับรูปแบบ บริบท และลักษณะของเอกสาร งานวิจัยนี้จึงต้องทำการพิจารณาความยาวที่น้อยกว่าค่าความยาวเฉลี่ยที่ได้ด้วย ดังนั้นการเลือกค่าสัดส่วนจำนวนเฉลี่ยที่เหมาะสมที่สุดจึงต้องทำการทดลองลดขนาดข้อมูลที่ค่าสัดส่วนจำนวนเฉลี่ยต่างๆ ที่มีขนาดไม่เกิน 120 และดูผลลัพธ์ภาพบิตแม็บของเอกสาร โดยงานวิจัยนี้เลือกทำการทดลองค่าสัดส่วนจำนวนเฉลี่ยที่มีขนาดต่างๆ กัน ซึ่งแสดงรายละเอียดและผลการทดลองในหัวข้อ 4.3

4.2.2 ค่าเฉลี่ยเคลื่อนที่

ค่าเฉลี่ยเคลื่อนที่ (Moving Average) ใช้ในการลดความแปรปรวน สัญญาณรบกวน และการเปลี่ยนแปลงที่เกิดขึ้นอย่างฉับพลันของข้อมูลเอกสาร ซึ่งต้องกำหนดค่าขีดแบ่ง (Threshold) ของการคำนวณค่าเฉลี่ยเคลื่อนที่ โดยพยายามหาค่า n ที่เหมาะสม ดังปรากฏในสมการที่ (3.1)

แนวทางในการกำหนดค่าขีดแบ่งที่ใช้ในการทำค่าเฉลี่ยเคลื่อนที่ที่เหมาะสมนั้น มีลักษณะเดียวกันกับการกำหนดค่าสัดส่วนจำนวนเฉลี่ย ในหัวข้อ 4.2.1 เนื่องจากงานวิจัยนี้มีแนวคิดในการวิเคราะห์รูปแบบข้อมูลจากประโยคของเอกสาร ดังนั้นค่าขีดแบ่งที่เหมาะสมสามารถพิจารณาจากค่าเฉลี่ยของขนาดความยาวตัวอักษรในแต่ละประโยคได้เช่นเดียวกัน ซึ่งการเลือกค่าขีดแบ่งที่เหมาะสมที่สุด ต้องมาจากการทดลองปรับเรียบข้อมูลที่ค่าขีดแบ่งต่างๆ ที่มีขนาดไม่เกิน 120 และดูผลลัพธ์ภาพบิตแม็บของเอกสาร โดยงานวิจัยนี้เลือกทำการทดลองค่าขีดแบ่งที่มีขนาดต่างๆ กัน ซึ่งแสดงรายละเอียดและผลการทดลองในหัวข้อ 4.3

4.2.3 ขนาดความยาวของเอกสาร

ขนาดความยาวของเอกสาร ส่งผลทางด้านเวลาและประสิทธิภาพโดยตรงกับการประมวลผลของการแสดงผลภาพบิตแม็บ เพราะถ้าการแสดงผลภาพบิตแม็บทำการประมวลเอกสารที่มีความยาวมาก ทำให้เวลาในการสร้างภาพบิตแม็บต้องใช้เวลานาน หากเอกสารมีความยาวที่น้อยไป อาจทำให้ผลภาพบิตแม็บมีความไม่ชัดเจน ไม่สามารถแยกแยะคุณสมบัติของเอกสารได้ ดังนั้นเพื่อให้การแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล

สามารถแสดงผลได้อย่างมีประสิทธิภาพ รวดเร็ว และได้ภาพเอกสารบิตแม็บที่ชัดเจนมีคุณภาพ จึงต้องพิจารณาหาขนาดความยาวของเอกสารที่เหมาะสมในการนำไปประมวลผล

แนวทางในการหาขนาดความของเอกสาร ที่เหมาะสมต่อการประมวลผลของการแสดงผลภาพบิตแม็บ คือ การทดลองประมวลผลที่ขนาดความยาวเอกสารต่างๆกัน แล้วทำการสังเกตผลการทดลอง รวมถึงทำการจับเวลาในการประมวลของการแสดงผลภาพบิตแม็บ ซึ่งแสดงรายละเอียดและผลการทดลองในหัวข้อ 4.3

4.3 ผลการทดลอง

4.3.1 การทดลองเพื่อสนับสนุนแนวทางการวิจัย

เนื่องจากงานวิจัยนี้ นำแนวคิดมาจากงานวิจัยที่มีการประมวลผลภาพบิตแม็บจากข้อมูลดีเอ็นเอ [1][4] และข้อมูลอนุกรมเวลา [2][9][10] ดังนั้นเพื่อเป็นการสนับสนุนแนวคิดของงานวิจัยนี้ จึงทำการทดลองกับข้อมูลดีเอ็นเอ และข้อมูลอนุกรมเวลาชนิดต่างๆ ที่ทราบลักษณะและความสัมพันธ์ของข้อมูลอยู่แล้ว

การทดลองกับข้อมูลดีเอ็นเอ โดยนำเอาข้อมูลดีเอ็นเอของสัตว์ต่างๆ จำนวน 4 ชนิด 13 สายพันธุ์ ซึ่งมีรายละเอียดดีเอ็นเอของสัตว์แต่ละชนิดดังตารางที่ 4.1 ภาพบิตแม็บที่ออกมาจากผลการทดลองควรมีความแตกต่างกันหากเป็นดีเอ็นเอของสัตว์ที่ต่างชนิดกัน และภาพบิตแม็บ ควรมีความคล้ายหรือใกล้เคียงกัน หากเป็นดีเอ็นเอของสัตว์ชนิดเดียวกัน แสดงผลการทดลอง ดังรูปที่ 4.2

Macaca Mulatta clone CH250-135N17

Macaca Mulatta clone CH250-253A7

Macaca Mulatta clone CH250-155N20

Macaca Mulatta clone CH250-190E12

รูปที่ 4.2 ผลการทดลองจากข้อมูลดีเอ็นเอ

Mycobacterium Bovis AF2122/97

Mycobacterium Tuberculosis H37Rv

Mycobacterium Paratuberculosis K-10

Mycobacterium Tuberculosis CDC1551

Francisella Tularensis Holarctica strain LVS

Francisella Tularensis Tularensis strain FSC

Chlamydophila Pneumoniae (TW-183 P1/4)

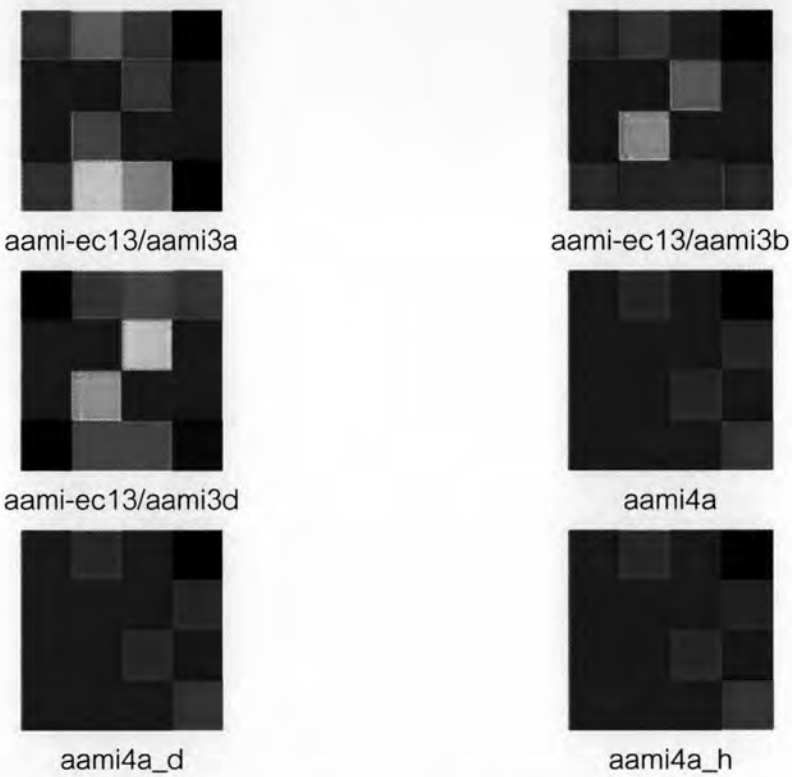
Chlamydophila Pneumoniae (TW-183 P 3/4)

Chlamydophila Pneumoniae (TW-183 P 4/4)

รูปที่ 4.2 (ต่อ) ผลการทดลองจากข้อมูลดีเอ็นเอ

รูปที่ 4.2 แสดงผลภาพบิตแม็บจากข้อมูลดีเอ็นเอของสัตว์จำนวน 4 ชนิด 13 สายพันธุ์ ซึ่งสังเกตได้ชัดเจนว่า ภาพบิตแม็บของสัตว์ชนิดเดียวกันมีลักษณะคล้ายกันถึงแม้ว่าจะมีสายพันธุ์ที่แตกต่างกัน แต่เมื่อเปรียบเทียบกับภาพบิตแม็บที่ของข้อมูลดีเอ็นเอของสัตว์ต่างชนิดกัน สังเกตได้ว่ามีลักษณะที่แตกต่างกันอย่างชัดเจน เช่น ภาพบิตแม็บของลิงสายพันธุ์ Macaca Mulatto กับแบคทีเรีย Mycobacterium

นอกจากการทดลองกับข้อมูลดีเอ็นเอแล้ว งานวิจัยนี้ได้ทำการทดลองกับข้อมูลอนุกรมเวลาด้วย ซึ่งเป็นข้อมูลคลื่นหัวใจ (Electrocardiograms) ของผู้ป่วยโรคหัวใจอาการต่างๆ ที่มาจากแหล่งข้อมูลของฟิซิโอเน็ต ที่ถูกใช้เป็นข้อมูลในการทดสอบความถูกต้องของอุปกรณ์วัดคลื่นหัวใจ มีชื่อชุดข้อมูลว่า "ANSI/AAMI EC13" ประกอบไปด้วยข้อมูลจำนวน 6 ชุด แบ่งออกได้เป็นข้อมูล 2 กลุ่มที่มีลักษณะคล้ายกันภายในกลุ่ม แสดงผลการของภาพบิตแม็บจากข้อมูลชุดนี้ดังรูปที่ 4.3



รูปที่ 4.3 ผลการทดลองจากข้อมูลอนุกรมเวลา

รูปที่ 4.3 แสดงภาพบิตแม็บของข้อมูลคลื่นหัวใจ ซึ่งเป็นข้อมูลเลขจำนวนจริงที่เกิดขึ้นตามช่วงเวลา จากผลภาพบิตแม็บแสดงให้เห็นได้ชัดเจนว่ากราฟคลื่นหัวใจที่มีลักษณะแนวโน้มที่เหมือนกันหรือคล้ายกัน มีลักษณะของภาพบิตแม็บที่คล้ายกัน และกราฟคลื่นหัวใจที่มีลักษณะแตกต่างกันมีลักษณะของภาพบิตแม็บที่แต่กต่างกัน ซึ่งแสดงกราฟคลื่นหัวใจตามลักษณะของข้อมูล ดังตารางที่ 4.2

ผลการทดลองการแสดงภาพบิตแม็บของข้อมูลดีเอ็นเอ และข้อมูลอนุกรมเวลาข้างต้น แสดงให้เห็นว่าสามารถนำวิธีการวิจัยไปประยุกต์ใช้กับการแสดงผลภาพบิตแม็บกับข้อมูลประเภทอื่นได้ หากแต่ต้องมีการศึกษาและวิเคราะห์เพื่อหาคุณลักษณะของข้อมูล และวิธีการจัดการข้อมูลที่เหมาะสมก่อนจะนำมาแสดงผลภาพบิตแม็บ จึงเป็นเหตุผลที่สามารถสนับสนุน

แนวทางของงานวิจัยในการนำเอาข้อมูลเอกสารมาศึกษา และหาวิธีในการสร้างรูปภาพบิตแม็บได้เป็นอย่างดี

4.3.2 การทดลองเพื่อหาพารามิเตอร์และข้อมูลที่เหมาะสม

การทดลองเพื่อหาพารามิเตอร์และข้อมูลที่เหมาะสม เป็นการทดลองเพื่อหาค่าพารามิเตอร์ที่สำคัญ ที่ส่งผลต่อประสิทธิภาพของการแสดงภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล ซึ่งมีพารามิเตอร์ที่สำคัญจำนวน 3 พารามิเตอร์ ประกอบไปด้วย ค่าสัดส่วนจำนวนเฉลี่ย (PAA Dimensionality Reduction) ค่าเฉลี่ยเคลื่อนที่ (Moving Average) และ ขนาดความยาวของเอกสาร

การทดลองนี้ทำการเลือกข้อมูลเอกสารดิจิทัลบางส่วน ที่ปรากฏในหัวข้อ 4.1 เพื่อนำมาเป็นข้อมูลชุดทดสอบ โดยเลือกเอกสารจำนวน 2 ชุดจากทุกกลุ่มเอกสาร ซึ่งมีรายละเอียดและผลการทดลองดังนี้

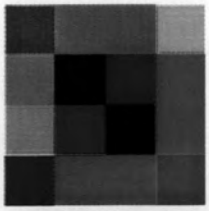
4.3.2.1 ค่าเหมาะสมของค่าสัดส่วนจำนวนเฉลี่ย

การทดลองเพื่อค่าเหมาะสมของค่าสัดส่วนจำนวนเฉลี่ย ทำการทดลองแสดงผลภาพบิตแม็บของข้อมูลเอกสารจำนวน 3 กลุ่มเอกสาร ประกอบไปด้วยเอกสารทั้งหมดจำนวน 6 เอกสาร โดยแสดงตัวอย่างผลการทดลองการแสดงภาพบิตแม็บจากข้อมูลเอกสารที่ขนาดความยาวของค่าสัดส่วนจำนวนเฉลี่ยเป็น 120 90 60 และ 30 ตามลำดับ ดังรูปที่ 4.4 ถึง 4.7

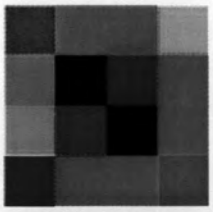
จากการทดลองการแสดงผลภาพบิตแม็บของเอกสาร ที่ค่าสัดส่วนจำนวนเฉลี่ยขนาดต่างๆ พบว่า เมื่อกำหนดค่าสัดส่วนจำนวนเฉลี่ยขนาด 60 ให้ผลภาพบิตแม็บที่สามารถแสดงความเหมือนและความแตกต่างตามกลุ่มเอกสารได้ชัดเจนที่สุด ดังนั้นจึงสรุปได้ว่าพารามิเตอร์ของค่าสัดส่วนจำนวนเฉลี่ยที่เหมาะสม คือ 60

4.3.2.2 ค่าเหมาะสมของค่าเฉลี่ยเคลื่อนที่

ค่าเฉลี่ยเคลื่อนที่ที่เหมาะสม เป็นค่าที่มีความสัมพันธ์กับค่าสัดส่วนจำนวนเฉลี่ย เนื่องจากค่าเฉลี่ยเคลื่อนที่ถูกใช้กำหนดขนาดในการปรับเรียบข้อมูล จึงไม่ควรมีขนาดเกินค่าความยาวเฉลี่ยของแต่ละประโยค อีกทั้งไม่ควรมีค่าเกินค่าสัดส่วนจำนวนเฉลี่ยด้วย การทดลองเพื่อหาค่าเฉลี่ยเคลื่อนที่ จะใช้เอกสารชุดเดียวกับการทดลองในการหาค่าสัดส่วนจำนวนเฉลี่ยในหัวข้อ 4.3.2.1 โดยแสดงตัวอย่างผลการทดลองการแสดงภาพบิตแม็บจากข้อมูลเอกสารที่ขนาดค่าเฉลี่ยเคลื่อนที่เป็น 60 30 และ 0 ตามลำดับ ดังรูปที่ 4.8 ถึง 4.10

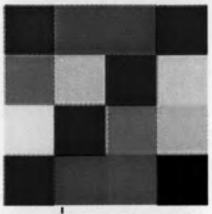


พระคัมภีร์ไบเบิลส่วนที่ 1



พระคัมภีร์ไบเบิลส่วนที่ 3

บทละครเรื่องเฟรนด์ส่วนที่ 2



บทละครเรื่องเฟรนด์ส่วนที่ 3

เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 1

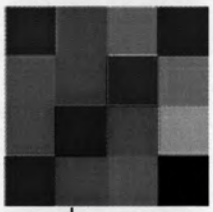
เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 2

รูปที่ 4.4 ผลภาพบิตแม็บเมื่อกำหนดค่าสัดส่วนจำนวนเฉลี่ยขนาด 120

พระคัมภีร์ไบเบิลส่วนที่ 1

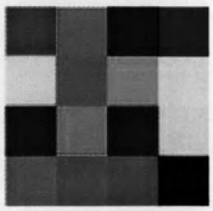
พระคัมภีร์ไบเบิลส่วนที่ 3

บทละครเรื่องเฟรนด์ชุดที่ 2



บทละครเรื่องเฟรนด์ชุดที่ 3

เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 1



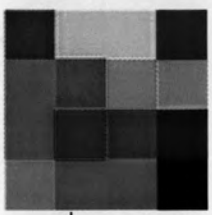
เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 2

รูปที่ 4.5 ผลภาพบิตแม็บเมื่อกำหนดค่าสัดส่วนจำนวนเฉลี่ยขนาด 90

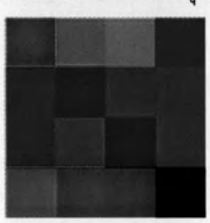
พระคัมภีร์ไบเบิลส่วนที่ 1

พระคัมภีร์ไบเบิลส่วนที่ 3

บทละครเรื่องเฟรนด์ชุดที่ 2



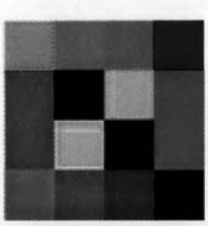
บทละครเรื่องเฟรนด์ชุดที่ 3



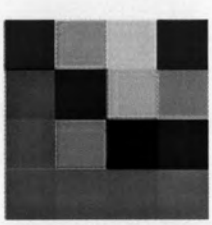
เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 1

เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 2

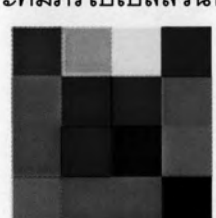
รูปที่ 4.6 ผลภาพบิตแม็บเมื่อกำหนดค่าสัดส่วนจำนวนเฉลี่ยขนาด 60



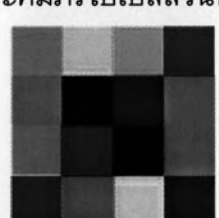
พระคัมภีร์ไบเบิลส่วนที่ 1



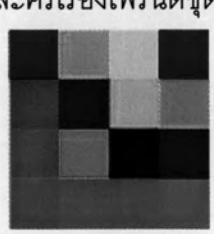
พระคัมภีร์ไบเบิลส่วนที่ 3



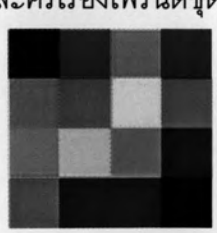
บทละครเรื่องเฟรนด์ชุดที่ 2



บทละครเรื่องเฟรนด์ชุดที่ 3



เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 1



เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 2

รูปที่ 4.7 ผลภาพบิตแม็บเมื่อกำหนดค่าสัดส่วนจำนวนเฉลี่ยขนาด 30

รูปที่ 4.8 ถึงรูปที่ 4.10 แสดงผลการทดลองกำหนดค่าเฉลี่ยเคลื่อนที่ขนาดต่างๆ ซึ่งกำหนดค่าสัดส่วนจำนวนเฉลี่ยที่เหมาะสมเป็น 60 สังเกตเห็นว่าเมื่อมีการกำหนดค่าเฉลี่ยเคลื่อนที่ทำให้ภาพบิตแม็บของเอกสารในกลุ่มเดียวกันมีลักษณะที่เหมือนกันมากขึ้น แต่อย่างไรก็ตามการกำหนดค่าเฉลี่ยเคลื่อนที่ ทำให้ภาพบิตแม็บของเอกสารที่อยู่ต่างกลุ่มกันมีลักษณะที่แตกต่างกันน้อยลงเช่นกัน ซึ่งส่งผลให้การแยกแยะเอกสารต่างกลุ่มทำได้ยากมากขึ้น จากผลการทดลองจึงสรุปได้ว่า การปรับเรียบข้อมูลโดยใช้ค่าเฉลี่ยเคลื่อนที่ส่งผลเสียต่อการแสดงผลภาพบิตแม็บของข้อมูลเอกสาร งานวิจัยนี้จึงเลือกที่จะไม่ทำการปรับเรียบข้อมูล หรือกำหนดค่าเฉลี่ยเคลื่อนที่เป็น 0 ในการประมวลผลภาพบิตแม็บจากเอกสาร

4.3.2.3 ค่าเหมาะสมของความยาวของเอกสาร

ความยาวของเอกสารที่เหมาะสม สามารถพิจารณาได้จากการนำเอกสารมาทดลองประมวลผลภาพบิตแม็บที่ความยาวของเอกสารต่างๆกัน โดยพิจารณาผลของภาพบิตแม็บรวมถึงพิจารณาเวลาในการประมวลผลของความยาวเอกสารที่แตกต่างกัน

แสดงตัวอย่างภาพบิตแม็บ ที่ได้จากการทดลองประมวลผลข้อมูลเอกสารที่ความยาวจำนวน 100,000 200,000 300,000 400,000 500,000 และ 600,000 ตัวอักษร ตามลำดับ ดังรูปที่ 4.11

พระคัมภีร์ไบเบิลส่วนที่ 1

พระคัมภีร์ไบเบิลส่วนที่ 3

บทละครเรื่องเฟรนด์ชุดที่ 2

บทละครเรื่องเฟรนด์ชุดที่ 3

เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 1

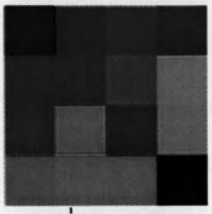
เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 2

รูปที่ 4.8 ผลภาพบิตแม็บเมื่อกำหนดค่าเฉลี่ยเคลื่อนที่ขนาด 60

พระคัมภีร์ไบเบิลส่วนที่ 1

พระคัมภีร์ไบเบิลส่วนที่ 3

บทละครเรื่องเฟรนด์ชุดที่ 2



บทละครเรื่องเฟรนด์ชุดที่ 3

เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 1

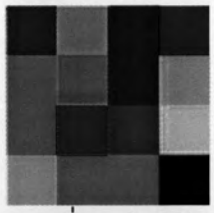
เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 2

รูปที่ 4.9 ผลภาพบิตแม็บเมื่อกำหนดค่าเฉลี่ยเคลื่อนที่ขนาด 30

พระคัมภีร์ไบเบิลส่วนที่ 1

พระคัมภีร์ไบเบิลส่วนที่ 3

บทละครเรื่องเฟรนด์ชุดที่ 2



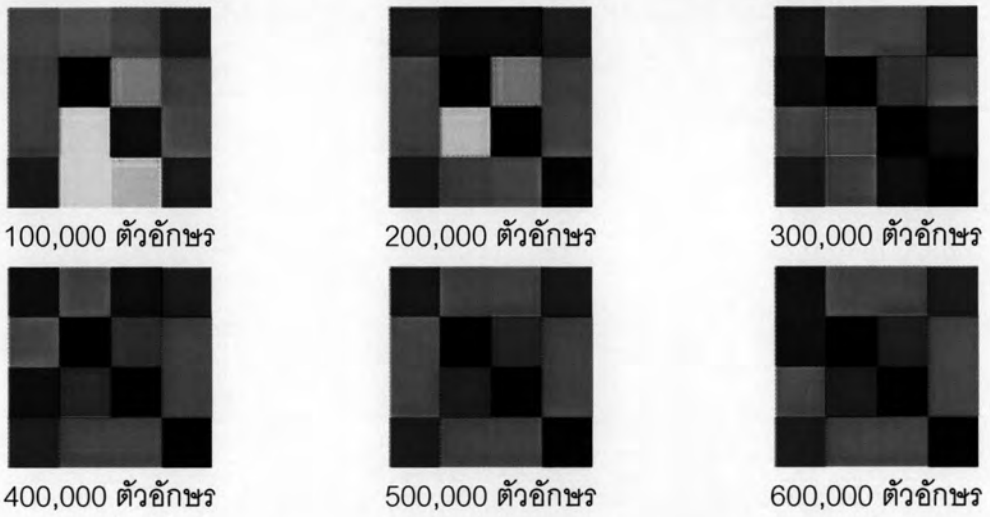
บทละครเรื่องเฟรนด์ชุดที่ 3

เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 1

เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 2

รูปที่ 4.10 ผลภาพบิตแม็บเมื่อกำหนดค่าเฉลี่ยเคลื่อนที่ขนาด 0

จากรูปที่ 4.11 สังเกตได้ว่าลักษณะของภาพบิตแม็บเปลี่ยนแปลงไป เมื่อการประมวลผลที่ขนาดของเอกสารมีความยาวมากขึ้น การเปลี่ยนแปลงของสีและรูปแบบภาพบิตแม็บจะเริ่มคงที่หรือเปลี่ยนแปลงน้อยมาก เมื่อการประมวลผลของเอกสารมีขนาดความยาวที่ประมาณ 300,000 ตัวอักษรขึ้นไป นอกจากนี้เวลาที่ใช้ในการประมวลผลภาพบิตแม็บของข้อมูลเอกสารที่ขนาดความยาวต่างๆ มีความแตกต่างกันน้อยมากจนไม่ส่งผลกระทบต่อเวลาในการประมวลผลของการแสดงผลภาพบิตแม็บแต่อย่างใด ดังนั้นจากการผลการทดลองจึงสรุปได้ว่าพารามิเตอร์ที่เหมาะสมของค่าความยาวของเอกสารคือ 300,000 ตัวอักษร



รูปที่ 4.11 ภาพบิตแม็บที่ได้มาจากการประมวลผลที่ความยาวของเอกสารขนาดต่างๆ