การแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล

นายสัณห์ชัย นักรบ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2549

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

VISUALIZATION BITMAPS FOR DIGITAL DOCUMENT COLLECTION

Mr. Sunchai Nakrob

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in Computer Science

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2006

Copyright of Chulalongkorn University

หัวข้อวิทยานิพนธ์	การแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล
โดย	นายสัณห์ชัย นักรบ
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษา	อาจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์

(ศาสตราจารย์ ดร.ดิเรก ลาวัณย์ศิริ)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร.พรศิริ หมื่นไชยศรี)

..... อาจารย์ที่ปรึกษา

(อาจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ)

..... กรรมการ

(อาจารย์ ดร.อติวงศ์ สุชาโต)

..... กรรมการ

(อาจารย์ ธงชัย โรจน์กังสดาล)

สัณห์ชัย นักรบ : การแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล. (VISUALIZATION BITMAPS FOR DIGITAL DOCUMENT COLLECTION) อาจารย์ ที่ปรึกษา : อ.ดร.โชติรัตน์ รัตนามหัทธนะ, 62 หน้า.

วิทยานิพนธ์นี้มีวัตถุประสงค์ในการแสดงผลภาพสำหรับข้อมูลเอกสารดิจิทัล โดยทำการแปลงข้อมูลในเอกสารจากตัวอักษรให้เป็นรูปภาพ เพื่อช่วยในการพิจารณาเปรียบเทียบความเหมือนและความแตกต่างของประเภทหรือหมวดหมู่เอกสาร ทำให้ผู้ใช้สามารถ จัดการ และจำแนกรูปแบบหรือประเภทของเอกสารได้ง่ายและรวดเร็วมากยิ่งขึ้น โดยไม่จำเป็นต้องเข้าไปพิจารณาเนื้อความในเอกสาร โดยการแสดงผลภาพมีแนวทางในการพัฒนาจากแนวคิดของทฤษฎีเคออสเกม ประยุกต์ร่วมกับการแสดงผลภาพบิตแม็บของข้อมูลอนุกรมเวลาโดยใช้วิธีการแบบแซ็ค

งานวิจัยนี้ได้ทำการวิเคราะห์รูปแบบและลักษณะต่างๆของเอกสาร โดยการปรับข้อมูลในเอกสารและกำหนดพารามิเตอร์ที่สำคัญต่างๆ เพื่อให้การแสดงผลภาพบิตแม็บจากข้อมูลในเอกสารมีความชัดเจนและมีประสิทธิภาพ ซึ่งได้มาจากการทดลองด้วยข้อมูลจริง นอกจากนี้ยังได้ทำการทดสอบประสิทธิภาพของการแสดงผลภาพจากการพิจารณาเปรียบเทียบภาพบิตแม็บของข้อมูลเอกสาร ทั้งจากการสังเกตและใช้วิธีการจัดกลุ่มภาพบิตแม็บโดยอัตโนมัติ ซึ่งได้ผลสรุปจากการทดสอบว่า การแสดงผลภาพสำหรับข้อมูลเอกสารดิจิทัล สามารถช่วยในการพิจารณาเปรียบเทียบความเหมือนและความแตกต่างของประเภทหรือหมวดหมู่เอกสารดิจิทัลได้อย่างมีประสิทธิภาพ

ภาควิชา.....วิศวกรรมคอมพิวเตอร์..... ลายมือชื่อนิสิต.....สัณห์ชัย นักรบ.....
สาขาวิชา.....วิทยาศาสตร์คอมพิวเตอร์..... ลายมือชื่ออาจารย์ที่ปรึกษา.....
ปีการศึกษา.....2549.....

4871543121 : MAJOR COMPUTER SCIENCE

KEY WORD: BITMAP VISUALIZATION / DOCUMENT BITMAPS / VISUALIZATION

SUNCHAI NAKROB : VISUALIZATION BITMAPS FOR DIGITAL DOCUMENT COLLECTION. THESIS ADVISOR : CHOTIRAT RATANAMAHATANA, Ph.D., 62 pp.

The objective of this research is to visualize digital documents by converting text data in the digital documents to a bitmap image to help compare the similarities and differences of document types or categories so that the document can be easily and more conveniently clustered and managed. Users do not need to read details in the document. This visualization technique combines together the advance in Chaos Game Theory and SAX representation in Time Series bitmap visualization.

By experimenting with real data, this research analyzes the feature and format of digital documents and later adjusts document data and defines important parameters so that bitmap visualization of the document data is well-defined and effective. Moreover, this research also tests the visualization efficiency by comparing the bitmaps of the digital document through both users' observation and automatic clustering. The result shows that the bitmap visualization technique for digital document data can effectively help differentiate the documents types or categories.

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

Department.....Computer Engineering..... Student's signature.....

Field of study.....Computer Science..... Advisor's signature.....

Academic year .....2006.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยความอนุเคราะห์ และความช่วยเหลืออย่างยิ่ง จาก อ.ดร.โชติรัตน์ รัตนามหัทธนะ อาจารย์ที่ปรึกษา ซึ่งให้ข้อคิด แนวทาง และคำปรึกษา ตลอดจนเป็น ผู้ตรวจทานแก้ไข ทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วง ขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณ รศ.ดร.พรศิริ หมื่นไชยศรี อ.ดร.อติวงศ์ สุชาโต และ อ.ธงชัย โรจน์กังสดาล ประธานกรรมการและกรรมการสอบวิทยานิพนธ์ ที่กรุณาให้คำแนะนำในการแก้ไขวิทยานิพนธ์ให้มีคุณภาพยิ่งขึ้น ขอขอบพระคุณคณาจารย์ในภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัยทุกท่านที่ประสิทธิ์ประสาทความรู้อันมีค่ายิ่งแก่ผู้วิจัย

ที่สำคัญที่สุดขอบพระคุณ คุณพ่อ คุณแม่ และเพื่อนๆ ที่เป็นแรงผลักดัน เป็นกำลังใจ ที่สำคัญให้ตลอดการศึกษาครั้งนี้

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ฉ

สารบัญ ..... ช

สารบัญตาราง ..... ญ

สารบัญภาพ ..... ฎ

บทที่

1 บทนำ 1

1.1 ความเป็นมาและความสำคัญของปัญหา 1

1.2 วัตถุประสงค์ของการวิจัย 1

1.3 ขอบเขตงานวิจัย 1

1.4 ขั้นตอนและวิธีการดำเนินงานวิจัย 2

1.5 ประโยชน์ที่คาดว่าจะได้รับ 2

2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 3

2.1 ลักษณะและวิธีการแสดงผลภาพ 3

2.1.1 การแสดงผลแบบเคออสเกม 3

2.1.2 การแสดงผลภาพแบบวงแหวน 3

2.1.3 การแสดงผลภาพแบบบิตแม็บ 5

2.2 ทฤษฎีเคออสเกม 5

2.3 การแปลงข้อมูลอนุกรมเวลาเป็นสัญลักษณ์หรืออักขระ 8

2.4 ทฤษฎีการจัดกลุ่มแบบเคมีน 9

2.5 งานวิจัยที่เกี่ยวข้อง 11

2.5.1 ภาพบิตแม็บของข้อมูลอนุกรมเวลา: เครื่องมือการแสดงผลภาพสำหรับการทำงานกับฐานข้อมูลของข้อมูลอนุกรมเวลาขนาดใหญ่ 11

2.5.2 สัญรูปอัจฉริยะ: การทำเหมืองข้อมูลขนาดย่อม และทำการแสดงผลภาพสู่ระบบปฏิบัติการแบบส่วนต่อประสานด้วยภาพกับผู้ใช้ 11

3 การออกแบบวิธีการแสดงผลภาพสำหรับข้อมูลเอกสารดิจิทัล 13

3.1 การแปลงข้อมูลจากเอกสารดิจิทัลไปเป็นข้อมูลอนุกรมเวลา 13

หน้า

3.1.1 การวิเคราะห์และปรับแต่งเอกสารดิจิทัล 13
3.1.2 การแปลงตัวอักษรไปเป็นตัวเลข 15
3.1.3 การปรับข้อมูลอนุกรมเวลาที่ได้จากข้อมูลเอกสารดิจิทัล 16
3.2 การแปลงข้อมูลอนุกรมเวลาไปเป็นอักขระ 17
3.2.1 การลดขนาดหรือมิติของข้อมูลโดยวิธีลดสัดส่วนจำนวนเฉลี่ย 17
3.2.2 การแปลงข้อมูลให้อยู่ในรูปการกระจายแบบเกาส์เซียน 18
3.2.3 การกำหนดจำนวนอักขระ 18
3.2.4 ตารางการกระจายข้อมูลของเกาส์เซียน 19
3.2.5 การแปลงข้อมูลเลขจำนวนจริงไปเป็นอักขระ 20
3.3 การแปลงอักขระไปเป็นภาพบิตแม็บ 21
3.3.1 การกำหนดระดับและรูปแบบอักขระของตารางเมทริกซ์ 21
3.3.2 การนับความถี่ของสายอักขระ 22
3.3.3 ค่าบรรทัดฐานมากที่สุดและน้อยที่สุด 23
3.3.4 การกำหนดระดับขั้นของแถบสีอาร์จีบี 24
3.3.5 การสร้างภาพบิตแม็บ 24
4 การทดลองและผลการทดลอง 26
4.1 ข้อมูลที่ใช้ในการทดลอง 26
4.1.1 ข้อมูลดีเอ็นเอ 26
4.1.2 ข้อมูลคลื่นหัวใจ 28
4.1.3 ข้อมูลเอกสารดิจิทัล 28
4.2 การเลือกใช้ค่าพารามิเตอร์ที่เหมาะสม 31
4.2.1 ค่าสัดส่วนจำนวนเฉลี่ย 31
4.2.2 ค่าเฉลี่ยเคลื่อนที่ 34
4.2.3 ขนาดความยาวของเอกสาร 34
4.3 ผลการทดลอง 35
4.3.1 การทดลองเพื่อสนันสนุบแนวทางการวิจัย 35
4.3.2 การทดลองเพื่อหาพารามิเตอร์และข้อมูลที่เหมาะสม 38

หน้า

5 ผลภาพบิตแม็บและการวัดผลการแสดงผลภาพสำหรับข้อมูลเอกสารดิจิทัล.....44
5.1 ผลภาพบิตแม็บจากข้อมูลเอกสารปะเภทเดียวกัน.....44
5.2 ผลภาพบิตแม็บจากข้อมูลเอกสารต่างประเภทกัน.....45
5.3 การพิจารณาภาพบิตแม็บโดยการจัดกลุ่มด้วยวิธีเคมีน 46
6 สรุปผลการวิจัยและข้อเสนอแนะ 49
6.1 สรุปผลการวิจัย 49
6.1.1 ผลการทดสอบกับข้อมูลดีเอ็นเอและข้อมูลอนุกรมเวลา.....49
6.1.2 ผลการทดสอบกับข้อมูลเอกสารดิจิทัลด้วยการจัดกลุ่มด้วยวิธีเคมีน.....50
6.2 ปัญหาที่พบจากการวิจัย 50
6.3 ข้อเสนอแนะ.....51
รายการอ้างอิง 52

ภาคผนวก 54
ภาคผนวก ก รายการคำที่ไม่มีนัยสำคัญ.....55
ภาคผนวก ข รายการอักษรพิเศษ.....58
ภาคผนวก ค ระยะห่างระหว่างเอกสารจากการคำนวนแบบแมนฮัทตัน.....59
ภาคผนวก ง ภาพบิตแม็บจากข้อมูลเอกสารที่นำมาทดลอง.....60
ประวัติผู้เขียนวิทยานิพนธ์ 62

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญตาราง

หน้า

ตารางที่ 3.1 ตัวอย่างการเปรียบเทียบระหว่างเลขจำนวนจริงและตัวอักษรจากมาตรฐานแอสกี 16
ตารางที่ 3.2 การเปรียบเทียบค่าสีอาร์จีบีกับค่าความถี่แบบบรรทัดฐานมากที่สุดและน้อยที่สุด 25
ตารางที่ 4.1 รายละเอียดข้อมูลดีเอ็นเอที่นำมาใช้ในการทดลอง 27
ตารางที่ 4.2 รายละเอียดข้อมูลคลื่นหัวใจ (ECG) ชุดข้อมูล “ANSI/AAMI EC13 Test Waveforms” 29
ตารางที่ 4.3 ข้อมูลเอกสารดิจิทัลที่นำมาใช้ในการทดลอง 30
ตารางที่ 4.4 ประโยคและความยาวตัวอักษรที่ทำการแบ่งแยกด้วยโปรแกรม RSTTool 32
ตารางที่ 5.1 สรุปผลการทดสอบการจัดกลุ่มเอกสารด้วยวิธีเคมีน 47

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญภาพ

หน้า

รูปที่ 2.1 การแสดงผลภาพแบบวงแหวน ซึ่งเป็นข้อมูลที่มีความสัมพันธ์กับช่วงเวลาทุกๆวัน ในเวลา 24 ชั่วโมง 4
รูปที่ 2.2 ผลของการทำซ้ำจากอัลกอริธึมของเคออสเกม ที่เลือกจุดเริ่มต้น 3 จุด 6
รูปที่ 2.3 ขั้นตอนการกำหนดจุดตามอัลกอริธึมของเคออสเกมกับข้อมูลดีเอ็นเอ "GAATTC" 7
รูปที่ 2.4 ภาพการประยุกต์ใช้อัลกอริธึมของเคออสเกมกับข้อมูลดีเอ็นเอขนาดความยาว 73,357 ตัวอักษร 8
รูปที่ 2.5 ตารางเมทริกซ์คุณสมบัติของวัตถุและเมทริกซ์ความไม่คล้าย 9
รูปที่ 2.6 อัลกอริธึมของวิธีการการจัดกลุ่มแบบเคมีน 10
รูปที่ 3.1 ตัวอย่างการปรับตัวอักษรให้เป็นตัวพิมพ์ใหญ่ 14
รูปที่ 3.2 ตัวอย่างการกำจัดคำที่ไม่มีนัยสำคัญ หรือ คำหยุด 14
รูปที่ 3.3 ตัวอย่างการกำจัดอักษรพิเศษ (Special Character) 15
รูปที่ 3.4 ตัวอย่างการคำนวณค่าเฉลี่ยเคลื่อนที่แบบพื้นฐาน 17
รูปที่ 3.5 การลดขนาดของข้อมูลโดยสัดส่วนจำนวนเฉลี่ย 18
รูปที่ 3.6 ตารางเมทริกซ์ที่กำกับด้วยอักขระและตัวอย่างของภาพบิตแม็บ 19
รูปที่ 3.7 ตารางการกระจายข้อมูลแบบไม่ต่อเนื่องของเกาส์เซียน 20
รูปที่ 3.8 การแปลงข้อมูลอนุกรมเวลาเป็นอักขระโดยกำหนดจุดขั้นจำนวน 4 จุด 20
รูปที่ 3.9 ลักษณะตารางเมทริกซ์ในระดับที่ 1 และระดับที่ 2 21
รูปที่ 3.10 การนับความถี่ของคู่อักขระ "aa" 22
รูปที่ 3.11 ตารางเมทริกซ์กับผลการนับความถี่ของข้อมูล 22
รูปที่ 3.12 ตารางเมทริกซ์เปรียบเทียบค่าที่ได้มาจากการนับความถี่ของข้อมูล และค่าความถี่ที่ได้มาจากการหาค่าสัดส่วนแบบปรับค่าบรรทัดฐานมากที่สุดและน้อยที่สุด 23
รูปที่ 3.13 ระดับแถบสีอาร์จีบีเปรียบเทียบกับค่าความถี่ 24
รูปที่ 3.14 ตารางเมทริกซ์ที่กำหนดค่าความถี่ที่ได้มาจากค่าความถี่แบบบรรทัดฐานมากที่สุดและน้อยที่สุด กับรูปภาพบิตแม็บหลังจากทำการแปลงค่าความถี่เทียบกับแถบสีอาร์จีบี 25
รูปที่ 4.1 โปรแกรม RSTTool เวอร์ชัน 3.45 31
รูปที่ 4.2 ผลการทดลองจากข้อมูลดีเอ็นเอ 35

หน้า

รูปที่ 4.3 ผลการทดลองจากข้อมูลอนุกรมเวลา..... 37
รูปที่ 4.4 ผลภาพบิตแม็บเมื่อกำหนดค่าสัดส่วนจำนวนเฉลี่ยขนาด 120 39
รูปที่ 4.5 ผลภาพบิตแม็บเมื่อกำหนดค่าสัดส่วนจำนวนเฉลี่ยขนาด 90 39
รูปที่ 4.6 ผลภาพบิตแม็บเมื่อกำหนดค่าสัดส่วนจำนวนเฉลี่ยขนาด 60 35
รูปที่ 4.7 ผลภาพบิตแม็บเมื่อกำหนดค่าสัดส่วนจำนวนเฉลี่ยขนาด 30 40
รูปที่ 4.8 ผลภาพบิตแม็บเมื่อกำหนดค่าเฉลี่ยเคลื่อนที่ขนาด 60 41
รูปที่ 4.9 ผลภาพบิตแม็บเมื่อกำหนดค่าเฉลี่ยเคลื่อนที่ขนาด 30 42
รูปที่ 4.10 ผลภาพบิตแม็บเมื่อกำหนดค่าเฉลี่ยเคลื่อนที่ขนาด 0..... 35
รูปที่ 4.11 ภาพบิตแม็บที่ได้มาจากการประมวลผลที่ความยาวของเอกสารขนาดต่างๆ 43
รูปที่ 5.1 ภาพเอกสารบิตแม็บของเอกสารที่อยู่กลุ่มเดียวกัน..... 44
รูปที่ 5.2 ภาพเอกสารบิตแม็บของเอกสารที่อยู่ต่างกลุ่มกัน 45
รูปที่ 5.3 อัลกอริธึมของวิธีการการจัดกลุ่มแบบเคมีนที่ทำการปรับเพิ่มเติม..... 48
รูปที่ 5.4 ผลการจัดกลุ่มของภาพเอกสารบิตแม็บด้วยวิธีเคมีน..... 49

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เนื่องจากโลกได้พัฒนาเข้าสู่ยุคโลกาภิวัตน์ เป็นยุคของเทคโนโลยีซึ่งมีการติดต่อสื่อสารกันระหว่างประเทศมากยิ่งขึ้น ทำให้มีเอกสารหรือข้อมูลต่างๆเกิดขึ้น ไม่ว่าจะเป็น ข่าวสาร นิตยสาร หนังสืออ้างอิง ข้อมูลทางการเงิน เป็นต้น ซึ่งการติดต่อสื่อสารนี้ทำให้จำนวนของเอกสารและประเภทของเอกสารในรูปแบบต่างๆ มีปริมาณเพิ่มขึ้นมาก เช่น เอกสารอิเล็กทรอนิกส์ (e-documents) เอกสารเสียง (audio documents) รูปภาพดิจิทัล (digital images) เป็นต้น การจัดเก็บและแยกหมวดหมู่ของเอกสารจำนวนมากมายดังกล่าวกลายเป็นสิ่งที่ยากลำบาก เพราะต้องอาศัยเวลาและบุคลากรเป็นจำนวนมาก เนื่องจากเอกสารบางชนิดต้องพิจารณาถึงเนื้อความภายในเอกสาร และยังอาจจะต้องใช้ผู้มีความรู้ความชำนาญเฉพาะด้านในการแยกแยะเอกสารอีกด้วย จึงทำให้เกิดแนวคิดที่จะหากลวิธีในการช่วยจัดการเอกสารดังกล่าวได้ง่าย และรวดเร็วยิ่งขึ้น

งานวิจัยนี้จึงนำเสนอการแสดงผลภาพบิตแม็บ ที่ช่วยให้การแยกแยะเอกสารจำนวนมากสามารถทำได้รวดเร็วมากขึ้น โดยการแปลงข้อมูลจากตัวอักษรในเอกสารให้เป็นรูปภาพ ซึ่งช่วยให้สามารถแบ่งประเภทและจัดการเอกสารได้สะดวกยิ่งขึ้น

1.2 วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีจุดประสงค์เพื่อพัฒนาการแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล ที่ใช้ในการแปลงข้อมูลเอกสารเป็นรูปภาพบิตแม็บ เพื่อช่วยให้สามารถพิจารณาความเหมือนความแตกต่างของชนิดหรือหมวดหมู่ ของเอกสารเป็นจำนวนมากในเบื้องต้นได้ โดยไม่จำเป็นต้องเข้าไปพิจารณาเนื้อความในเอกสาร ทำให้สามารถแบ่งแยก จัดการ และจำแนกรูปแบบหรือประเภทของเอกสารได้อย่างรวดเร็ว

1.3 ขอบเขตของการวิจัย

1. ศึกษาและออกแบบขั้นตอนการดำเนินการวิจัย เพื่อใช้ในการจัดกลุ่ม และจำแนกเอกสารจากกลุ่มเอกสารในเบื้องต้น
2. มุ่งเน้นการวิเคราะห์ และแปลงเอกสารพื้นฐาน (Plain Text) ซึ่งเป็นข้อมูลตัวอักษร ตัวเลข และอักษรพิเศษ ที่อยู่ในมาตรฐานแอสกี (ASCII)
3. ทำการวิเคราะห์เอกสารในภาษาอังกฤษเป็นหลัก

4. ใช้วิธีการแบบแซ็ค (Symbolic Aggregate approXimation - SAX) ในการแปลงข้อมูลอนุกรมเวลาไปเป็นข้อมูลสัญลักษณ์หรืออักขระ

1.4 ขั้นตอนและวิธีดำเนินการวิจัย

1. ศึกษารูปแบบและลักษณะของเอกสารในแต่ละชนิด หมวดหมู่ หรือประเภทของเอกสาร เพื่อค้นหาลักษณะที่แตกต่างของเอกสาร ที่สามารถนำมาเป็นแนวทางในการพัฒนาการวิเคราะห์เอกสารออกมาเป็นภาพบิตแม็บ

2. ศึกษารูปแบบและทฤษฎีที่เกี่ยวข้อง กับการนำข้อมูลในรูปแบบต่างๆ มาสรุป วิเคราะห์และแสดงข้อมูลเป็นรูปภาพ ซึ่งง่ายต่อการทำความเข้าใจและตีความหมาย

3. ออกแบบและพัฒนาการแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล เพื่อแสดงเป็นภาพบิตแม็บ ตามรูปแบบและทฤษฎีที่ได้ทำการศึกษา

4. ประเมินผลการแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล

5. สรุปผลการวิจัยและจัดทำรายงานวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้วิธีการที่ใช้ในการแปลงเอกสารออกมาเป็นภาพบิตแม็บ เพื่อช่วยในการพิจารณาเปรียบเทียบความเหมือนและความแตกต่างของเอกสาร ตามประเภทหรือหมวดหมู่

2. สามารถใช้การแสดงผลภาพบิตแม็บ เพื่อช่วยในการ จัดกลุ่ม จำแนกประเภท หรือหมวดหมู่ ของเอกสารในเบื้องต้นได้

3. เป็นแนวทางในการพัฒนาการแสดงผลภาพบิตแม็บ เพื่อแปลงข้อมูลประเภทอื่นๆ เป็นภาพบิตแม็บ หรือการแสดงผลสรุปกับข้อมูลเอกสารในรูปแบบอื่นๆ ต่อไป

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ลักษณะและวิธีการแสดงผลภาพ (Visualization Techniques)

การแสดงผลภาพจากข้อมูล เป็นการสรุปรวมข้อมูลให้แสดงเป็นภาพในรูปแบบต่างๆ เช่น รูปภาพกราฟแท่ง ภาพกราฟเส้น หรือภาพวงกลม เพื่อให้ง่ายในการพิจารณาและทำความเข้าใจ ด้วยเหตุนี้เองทำให้การแสดงผลภาพจากข้อมูล [1][2][3] ถูกพัฒนาและนำเสนอในรูปแบบอื่นๆ มากขึ้น เพื่อให้เหมาะสมกับข้อมูลแต่ละประเภท โดยเฉพาะข้อมูลที่มีความซับซ้อนและมีปริมาณมาก จนไม่สามารถพิจารณาหรือทำความเข้าใจได้โดยง่าย [2] การแสดงผลภาพจากข้อมูลดังกล่าว ได้มีการนำเสนอในหลากหลายวิธี โดยมีลักษณะและวิธีการแสดงผลภาพที่น่าสนใจและเป็นแนวทางในงานวิจัยนี้ ยกตัวอย่างเช่น

2.1.1 การแสดงผลแบบเคออสเกม

การแสดงผลแบบเคออสเกม (Chaos Game Representations) ถูกประยุกต์มาจากทฤษฎีเคออสเกม [1][4] ที่มีอัลกอริธึมในการแสดงผลภาพจากการเกิดขึ้นของจุดโดยการทำซ้ำแบบสุ่ม การแสดงผลแบบเคออสเกมมีหลักการมาจากวิธีการกำหนดจุดเริ่มต้น เป็นเสมือนมุมของภาพและกำหนดจุดที่เกิดขึ้นต่อไปบริเวณตรงกลางระหว่างจุดสองจุดใดๆ ที่มีอยู่แล้วโดยการสุ่ม ซึ่งรายละเอียดจะกล่าวในหัวข้อ 2.2 ต่อไป

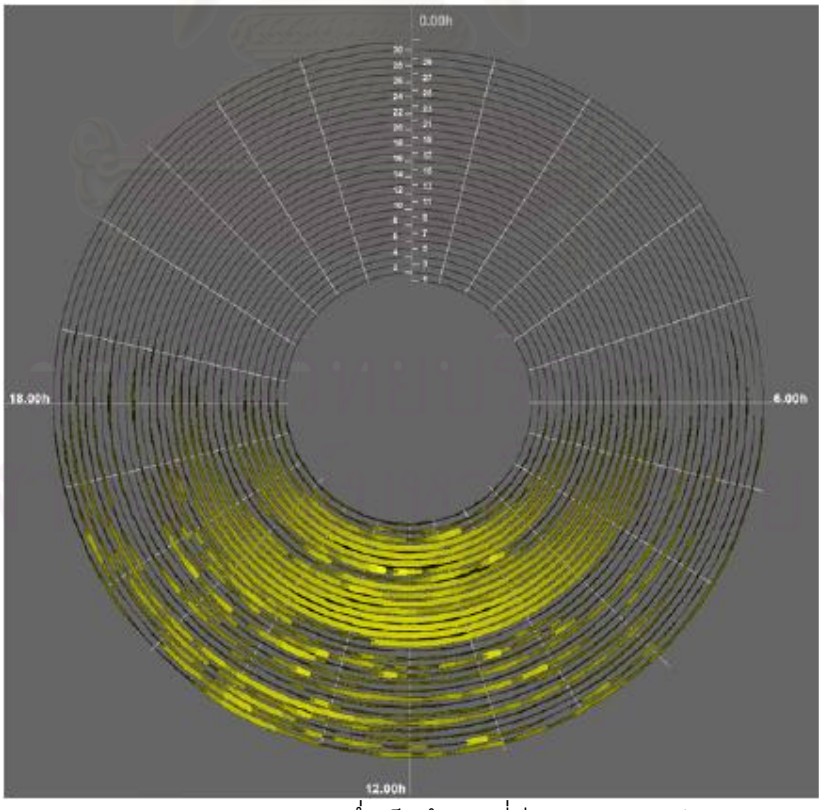
การกำหนดจุดแต่ละจุดจะเกิดโดยการทำซ้ำไปเรื่อยๆ จนได้ภาพที่มีลักษณะเฉพาะเกิดขึ้น การแสดงผลแบบเคออสเกมได้มีการนำไปใช้ในการศึกษารูปแบบของชุดข้อมูลดีเอ็นเอ [1] โดยระบุจุดเริ่มต้นขอบเขต 4 จุด เป็นตัวอักษรที่กำกับด้วยอักขระในข้อมูลดีเอ็นเอ แล้วใช้ข้อมูลของดีเอ็นเอสร้างจุดต่างๆขึ้นมา ต่อมาการแสดงผลแบบเคออสเกมได้ถูกพัฒนาจากการกำหนดจุดมาเป็นการนับความถี่ของอักขระของข้อมูลดีเอ็นเอที่เกิดขึ้นแทน ซึ่งการแสดงผลแบบเคออสเกม ของข้อมูลโครงสร้างของดีเอ็นเอหรือหน่วยพันธุกรรมแบบนับความถี่ เป็นการแสดงผลภาพจากข้อมูลที่มีประสิทธิภาพในการวิเคราะห์โครงสร้างและการจัดกลุ่ม ทำให้สามารถทำความเข้าใจในข้อมูล ได้ง่ายและรวดเร็วขึ้น [4]

2.1.2 การแสดงผลภาพแบบวงแหวน

การแสดงผลภาพแบบวงแหวน (Spiral Representation) เป็นการนำเสนอวิธีการในการวิเคราะห์ข้อมูลอนุกรมเวลา (Time Series) ที่มีปริมาณมากโดยเฉพาะ [3] โดยส่วนมากการแสดงผลภาพแบบวงแหวน ถูกนำมาใช้ในการวิเคราะห์ข้อมูลเชิงวิศวกรรมศาสตร์ วิทยาศาสตร์ หรือแม้แต่สาขาอื่นๆ ที่มีข้อมูลเป็นแบบอนุกรมเวลา และมีขนาดข้อมูลที่ต้องการวิเคราะห์เป็น

จำนวนมาก การแสดงผลภาพแบบวงแหวนอาศัยหลักการสร้างรูปภาพที่มีลักษณะเป็นวงแหวน ซึ่งเกิดจากการนำข้อมูลในแต่ละช่วงเวลามาวางเรียงกันขดเป็นวงอย่างต่อเนื่อง ทำให้มีลักษณะเหมือนเป็นวงแหวนที่ซ้อนกันอยู่หลายวง

การแสดงภาพแบบวงแหวน มีลักษณะการจัดวางข้อมูลแต่ละช่วงในวง ให้เกิดความสัมพันธ์กับทุกๆส่วนของวง ซึ่งนำเสนอเพื่อเปรียบเทียบรายละเอียดของข้อมูลในช่วงเวลาที่ต่างกันในแต่ละคาบของข้อมูล เช่น การเปรียบเทียบปริมาณการใช้ไฟฟ้าในรอบวัน หรือสัปดาห์ ช่วงเวลาที่มีการใช้ไฟฟ้ามากจะปรากฏช่วงของวงแหวนที่มีความเข้มมาก เป็นต้น ซึ่งความเข้มนั้นจะเกิดจากความถี่ที่มีปริมาณของข้อมูลเป็นจำนวนมาก นอกจากนี้การแสดงผลภาพแบบวงแหวนยังสามารถเพิ่มเติมคุณลักษณะพิเศษเพื่อใช้แสดงรายละเอียดต่างๆ ได้มากยิ่งขึ้น เช่น การขยายส่วนเฉพาะจุด (Zooming) เป็นการเลือกเข้าไปดูข้อมูลโดยละเอียดเฉพาะส่วนใดส่วนหนึ่ง โดยเลือกดูข้อมูลเฉพาะช่วงที่สนใจบางส่วนได้ [2] หรือการเน้นและเชื่อมโยงข้อมูล คือ สามารถเลือกเอาข้อมูลบางช่วงที่น่าสนใจหลายๆช่วง แยกออกมาวิเคราะห์และหาความสัมพันธ์กัน วิธีการแสดงผลภาพแบบวงแหวนยังเป็นวิธีการหนึ่ง ที่สามารถแสดงข้อมูลที่มีขนาดใหญ่และมีปริมาณข้อมูลมาก ให้สามารถแสดงผลอยู่ในพื้นที่ที่จำกัดได้ (Information Mural) [5] แสดงตัวอย่างของการแสดงผลภาพแบบวงแหวน ดังรูปที่ 2.1



รูปที่ 2.1 การแสดงผลภาพแบบวงแหวน ซึ่งเป็นข้อมูลที่มีความสัมพันธ์กับช่วงเวลาทุกๆวัน ในเวลา 24 ชั่วโมง (ที่มา: Weber, M., Alexa, M., and Mueller, W.) [3]

2.1.3 การแสดงผลภาพแบบบิตแม็บ

การแสดงผลภาพบิตแม็บ (Bitmap Representation) เป็นวิธีการแสดงผลภาพแบบดิจิทัล ที่มีข้อมูลหรือโครงสร้างเป็นลักษณะรูปสี่เหลี่ยมที่เกิดจากจุดของสี ซึ่งแสดงอยู่บนจอคอมพิวเตอร์ หรือในอุปกรณ์แสดงผลต่างๆ ที่เรารู้จักกันโดยทั่วไป โดยจุดสีที่เกิดขึ้นมาแต่ละจุดเกิดจากค่าที่ประกอบด้วยจำนวนบิตจำนวน 3 ค่า ซึ่งเป็นไปตามรูปแบบการแสดงสีแบบอาร์จีบี (RGB Color Space) กล่าวคือ ประกอบไปด้วยค่าบิตสีจำนวน 3 แบบ คือ บิตที่แสดงค่าสีแดง สีเขียว และสีน้ำเงิน การเพิ่มจำนวนบิตต่อหนึ่งจุดการแสดงสีสามารถทำให้การแสดงสีต่อจุดมีได้มากขึ้น แต่จำเป็นต้องใช้ขนาดของหน่วยความจำในการเก็บข้อมูลต่อจุดของภาพบิตแม็บ หรือภาพดิจิทัลมากขึ้นตามไปด้วย ในทางกลับกันถ้าต้องการแสดงสีของภาพออกมาเป็นเพียงสเกลสีเทา (Grayscale) จะอาศัยจำนวนบิตข้อมูลเพียง 2 บิต หรือหากต้องการแสดงผลข้อมูลเพียงภาพสีขาวดำ จะใช้จำนวนบิตข้อมูลเพียง 1 บิตเท่านั้น

การแสดงผลภาพบิตแม็บบนหน้าจอ มีการแสดงค่าความละเอียดและสีสันจะขึ้นอยู่กับลักษณะความกว้างและความยาวของภาพ ประกอบกับจำนวนบิตและจุดสี (Pixel) ที่ใช้ซึ่งบ่งบอกถึงจำนวนของสีที่สามารถแสดงได้ ในปัจจุบันการแสดงภาพบิตแม็บโดยทั่วไปจะใช้ขนาดของบิตจำนวน 24 บิตต่อหนึ่งจุดสี เพื่อทำให้สามารถแสดงสีให้ได้เพียงพอต่อภาพเสมือนจริง ซึ่งในบางครั้งอาจมากเกินกว่าที่สายตามนุษย์สามารถรับรู้ได้ และทำให้ต้องใช้หน่วยความจำขนาดใหญ่ในการเก็บข้อมูล รวมถึงต้องใช้เวลาในการประมวลผลมากอีกด้วย

การแสดงผลภาพดังกล่าวข้างต้น เป็นเพียงส่วนหนึ่งของการแสดงผลภาพที่เป็นที่รู้จักและถูกนำไปใช้กันอย่างกว้างขวาง อย่างไรก็ตามยังมีการแสดงผลภาพวิธีอื่นๆ อีกเป็นจำนวนมาก ซึ่งถูกนำเสนอในลักษณะต่างๆ ขึ้นอยู่กับรูปแบบของข้อมูลและจุดประสงค์ที่จะนำไปใช้ [2]

2.2 ทฤษฎีเคออสเกม

เคออสเกม (Chaos Game) เป็นอัลกอริธึมที่ใช้ในการสร้างรูปภาพจากจุด ที่เกิดจากการทำซ้ำโดยการสุ่ม [1][4] เคออสเกมถูกนำมาใช้ในการแสดงรูปแบบของข้อมูลดีเอ็นเอ หรือโครงสร้างหน่วยพันธุกรรม ซึ่งมีอัลกอริธึม [1] ดังนี้

อัลกอริธึมของเคออสเกมเริ่มจากการกำหนดจุดเริ่มต้น จากนั้นจะสุ่มกำหนดจุดต่อๆไปเพื่อสร้างเป็นภาพขึ้นมา โดยมีขั้นตอนดังนี้

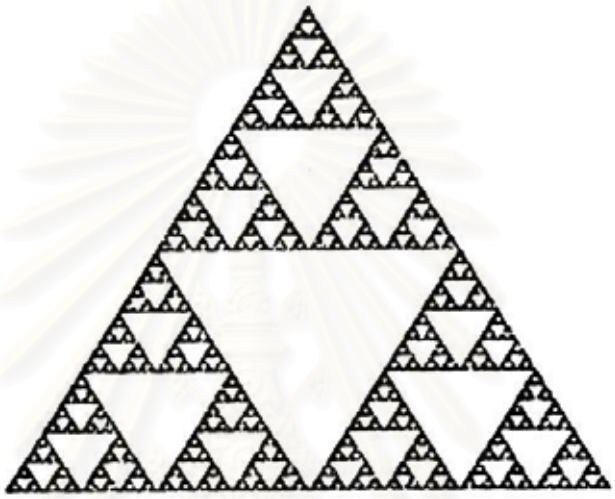
ขั้นตอนที่ 1 กำหนดจุดเริ่มต้น อย่างน้อย 3 จุด ซึ่งจะกำหนดที่ตำแหน่งใดก็ได้ แต่จุดที่กำหนดนั้นต้องไม่เป็นตำแหน่งที่สามารถลากผ่านได้โดยเส้นตรงเส้นเดียวพร้อมกัน ทั้ง 3 จุด

ขั้นตอนที่ 2 กำหนดชื่อของแต่ละจุดเริ่มต้นนั้น เป็นเลขจำนวนนับที่ต่อเนื่อง

ขั้นตอนที่ 3 เลือกจุดตั้งต้น 2 จุด จากจุดเดิมที่เคยกำหนดมาแล้วโดยการสุ่ม จะกำหนดจุดที่เกิดขึ้นมาใหม่จากบริเวณตรงกลางระหว่างจุด 2 จุดที่เลือกนั้น

ขั้นตอนที่ 4 ดำเนินการแบบขั้นตอนที่ 3 ไปเรื่อยๆ โดยยึดหลักการแบบเดิม

เมื่อทำซ้ำจากขั้นตอนข้างต้นเป็นจำนวนหลายพันครั้ง ผลที่ได้ออกมาจากการเลือกจุดแบบสุ่มที่กำหนดจุดเริ่มต้นเป็น 3 จุด เป็นภาพสามเหลี่ยมซ้อนกัน ดังรูปที่ 2.2

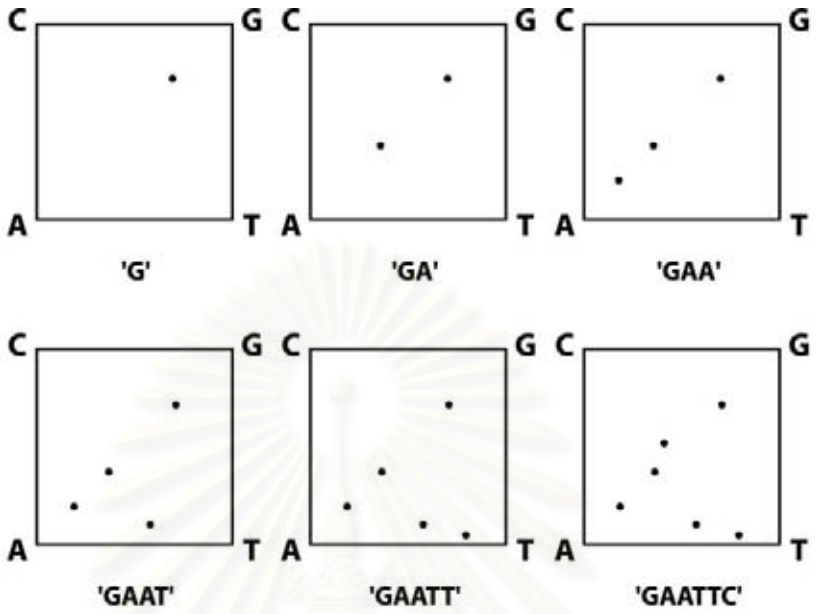


รูปที่ 2.2 ผลของการทำซ้ำจากอัลกอริธึมของเคออสเกม ที่เลือกจุดเริ่มต้น 3 จุด
(ที่มา: Jeffrey, H. J.) [1]

จากรูปที่ 2.2 เป็นผลจากหลักการทางคณิตศาสตร์ที่ได้มีการค้นพบมานานแล้ว และเป็นที่รู้จักกันดีในชื่อของ สามเหลี่ยมเซอร์ปินส์กี (Sierpinski Triangle) นอกจากนี้อัลกอริธึมจากหลักการของเคออสเกมยังสามารถกำหนดจุดเริ่มต้นที่มีจำนวนมากกว่า 3 จุดได้ ซึ่งให้ผลออกมาในลักษณะคล้ายกัน เช่น การกำหนดจุดเริ่มต้นเป็น 5 จุด 6 จุด หรือ 7 จุด จะทำให้เกิดรูปหลายเหลี่ยมตามจำนวนจุดเริ่มต้นที่กำหนด แต่สิ่งที่น่าสนใจในอัลกอริธึมของเคออสเกม คือ กรณีกำหนดจุดเริ่มต้นเป็น 4 จุด ผลที่ได้จะแตกต่างออกไปจากกรณีกำหนดจุดเริ่มต้นเป็น 3 จุด และมากกว่า 4 จุดขึ้นไป ซึ่งผลที่เกิดขึ้นจากการกำหนดจุดเริ่มต้น 4 จุดนั้น จะได้ภาพที่ไม่มีรูปแบบเฉพาะ โดยเกิดเป็นรูปสี่เหลี่ยมที่มีลักษณะไม่แน่นอนขึ้นอยู่กับจำนวนการเลือกจุดสุ่ม

ประโยชน์จากการเกิดรูปแบบภาพที่ไม่แน่นอน ที่เกิดจากการสุ่มโดยอัลกอริธึมของเคออสเกมนั้น ถูกนำไปประยุกต์ใช้ในการแสดงข้อมูลของดีเอ็นเอ เนื่องจากข้อมูลของดีเอ็นเอเป็นข้อมูลที่มีโครงสร้างเฉพาะ ซึ่งสามารถแปลงให้เกิดลักษณะของภาพที่มีลักษณะเฉพาะได้ อีกทั้งข้อมูลพื้นฐานของดีเอ็นเอประกอบไปด้วยตัวอักษร 4 ตัว คือ A C G และ T พอดีกับการนำไปกำหนดเป็นจุดเริ่มต้นที่มี 4 จุดได้

ดังนั้นเมื่อประยุกต์ใช้อัลกอริธึมของเคออสเกมกับข้อมูลของดีเอ็นเอ ยกตัวอย่างข้อมูลเช่น “GATTC” สามารถแสดงรายละเอียดขั้นตอนการกำหนดจุดได้ ดังรูปที่ 2.3



รูปที่ 2.3 ขั้นตอนการกำหนดจุดตามอัลกอริธึมของเคออสเกมกับข้อมูลดีเอ็นเอ “GAATTC”

จากรูปที่ 2.3 แสดงการเกิดจุดขึ้นอย่างต่อเนื่อง โดยมีลักษณะเหตุการณ์ดังต่อไปนี้

รูป ‘G’ เกิดจุดจากข้อมูล ‘G’ ซึ่งจะเกิดจุดที่บริเวณตรงกลาง ระหว่างจุดบริเวณกึ่งกลางของภาพสี่เหลี่ยมกับมุมสี่เหลี่ยมด้าน ‘G’

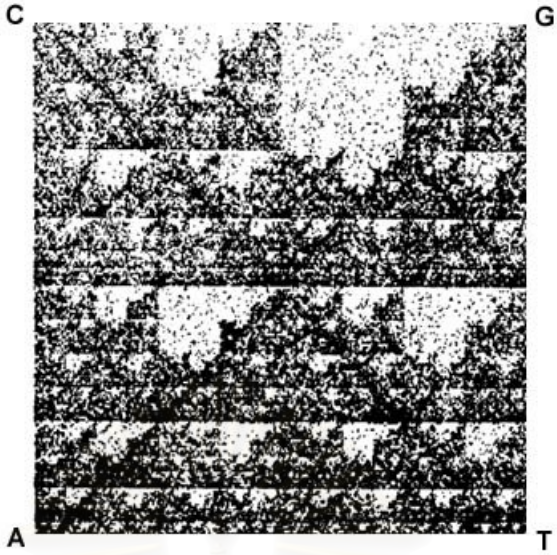
รูป ‘GA’ เกิดจุดจากข้อมูล ‘A’ ซึ่งจะเกิดจุดที่บริเวณตรงกลาง ระหว่างจุดที่สร้างก่อนหน้านี้ ซึ่งคือจุด ‘G’ กับมุมสี่เหลี่ยมด้าน ‘A’

รูป ‘GAA’ เกิดจุดจากข้อมูล ‘A’ ซึ่งจะเกิดจุดที่บริเวณตรงกลาง ระหว่างจุดที่สร้างก่อนหน้านี้ ซึ่งคือจุด ‘GA’ กับมุมสี่เหลี่ยมด้าน ‘A’

รูป ‘GAAT’, ‘GAATT’ และ‘GAATTC’ จะเกิดขึ้นจากเหตุการณ์ลักษณะเดียวกันไปเรื่อยๆ

เมื่อใช้อัลกอริธึมของเคออสเกมกับข้อมูลดีเอ็นเอขนาดความยาว 73,357 ตัวอักษร ซึ่งเป็น ดีเอ็นเอที่มาจาก HUMHBB [1] จะได้ภาพที่เกิดมาจากการกำหนดจุด ดังรูปที่ 2.4

ทฤษฎีเคออสเกมจัดเป็นแนวคิดพื้นฐานที่สำคัญของงานวิจัยเป็นจำนวนมาก สามารถนำไปประยุกต์และพัฒนาต่อเนื่องได้หลากหลาย งานวิจัยนี้จึงนำเอาแนวคิดดังกล่าวมาเป็นแนวทางในการทำวิจัยเช่นกัน เพราะทฤษฎีเคออสเกมเป็นแนวทางที่ทำให้เข้าใจถึงหลักการพื้นฐานของการแสดงผลภาพจากข้อมูลที่มีปริมาณมาก และสามารถนำไปเป็นแนวคิดในการวิจัยต่อเนื่องได้เป็นอย่างดี



รูปที่ 2.4 ภาพการประยุกต์ใช้อัลกอริธึมของเคออสเกมกับข้อมูลดีเอ็นเอขนาดความยาว 73,357 ตัวอักษร (HUMHBB) (ที่มา: Jeffrey, H. J.) [1]

2.3 การแปลงข้อมูลอนุกรมเวลาเป็นสัญลักษณ์หรืออักขระ

การแปลงข้อมูลอนุกรมเวลาเป็นสัญลักษณ์หรืออักขระ (Symbolic Time Series Representations) เป็นวิธีการนำข้อมูลอนุกรมเวลาที่มีปริมาณมาก และมีลักษณะข้อมูลเป็นเลขจำนวนจริง แปลงให้เป็นสัญลักษณ์หรืออักขระ เพื่อที่จะสามารถใช้ข้อมูลลักษณะดังกล่าวมาทำการสรุปรวมข้อมูล แบ่งแยก จำแนก หาลักษณะที่ผิดแปลกของข้อมูล หรือนำไปใช้ในงานต่อเนื่องอื่นๆ [6] จุดเด่นที่สำคัญของการแปลงข้อมูลอนุกรมเวลาเป็นสัญลักษณ์หรืออักขระ คือ การทำให้ข้อมูลที่เป็นเลขจำนวนจริงให้มาอยู่ในรูปแบบของสัญลักษณ์หรืออักขระ เพราะการประมวลผลกับข้อมูลอนุกรมเวลาโดยตรงในบางกรณี จะยากต่อการใช้อัลกอลริธึมในการจัดการรูปแบบข้อมูล เพราะอัลกอริธึมบางประเภทไม่รองรับการทำงานกับข้อมูลแบบต่อเนื่อง จำเป็นต้องทำให้ข้อมูลนั้นอยู่ในรูปแบบไม่ต่อเนื่องก่อน

การแปลงข้อมูลอนุกรมเวลาเป็นสัญลักษณ์หรืออักขระ มีการวิจัยและนำมาใช้อยู่หลายวิธี อาทิเช่น วิธีสัญลักษณ์นิยม (Symbolizing) วิธีโทเค็น (Tokenizing) หรือวิธีการแจงหน่วย (Quantizing) [7][6] แต่ในการวิจัยนี้ได้ศึกษาและเลือกใช้วิธี การแปลงจากข้อมูลอนุกรมเวลาไปเป็นข้อมูลสัญลักษณ์หรืออักขระที่ได้มาจากการหาค่าเฉลี่ยโดยรวม หรือแซ็ค (Symbolic Aggregate approXimation - SAX) เนื่องจากวิธีการแบบแซ็คเป็นวิธีการที่ไม่ซับซ้อน ประมวลผลเร็ว และข้อมูลจากการแปลงจะไม่เสียคุณสมบัติจากข้อมูลเดิม [2]

สายสัญลักษณ์หรืออักขระของข้อมูลอนุกรมเวลาที่ได้มาจากการหาค่าเฉลี่ยโดยรวม ทำให้ข้อมูลอนุกรมเวลาขนาดความยาว *n* เปลี่ยนไปเป็นข้อมูลที่ลดขนาดลงเหลือ *w* ได้ โดยที่ขนาดของ *w* จะน้อยกว่าหรือเท่ากับ *n* เสมอ โดยปกติจะน้อยกว่าขนาดของ *n* มากๆ ซึ่งจะน้อยลงมากเพียงใดขึ้นกับการกำหนดขนาดสัดส่วนจำนวนเฉลี่ย (Piecewise Aggregate Approximation - PAA) ที่เป็นตัวกำหนดช่วงของข้อมูลเพื่อหาค่าเฉลี่ย ทำให้เกิดสายอักขระใหม่ที่มีขนาดเท่ากับ *w* ส่งผลให้เกิดการลดขนาดของข้อมูลจำนวนมาก ทำให้สามารถพิจารณาข้อมูลโดยรวมทั้งหมดได้ง่ายและข้อมูลไม่เสียรูปแบบจากเดิม [6]

วิธีการแปลงข้อมูลอนุกรมเวลาออกมาเป็นสัญลักษณ์หรืออักขระแบบแซ็ค มีหลักการที่สำคัญอยู่ 2 ประการ คือ การลดขนาดหรือมิติของข้อมูลโดยสัดส่วนจำนวนเฉลี่ย (PAA Dimensionality Reduction) และวิธีการแปลงข้อมูลจำนวนจริงของข้อมูลอนุกรมเวลา (Time Series Data) ไปเป็นข้อมูลอักขระ ซึ่งจะอธิบายรายละเอียดและวิธีการในหัวข้อ 3.2 ต่อไป

2.4 ทฤษฎีการจัดกลุ่มแบบเคมีน

กลุ่ม (Cluster) คือคอลเลคชัน (Collection) ของวัตถุ ซึ่งมีคุณสมบัติคือ วัตถุที่อยู่ในกลุ่มเดียวกันจะคล้ายกัน แต่แตกต่างจากวัตถุในกลุ่มอื่น การจัดกลุ่ม (Clustering) จึงเป็นการจำแนกประเภทของวัตถุที่มีความคล้ายกันให้อยู่ในกลุ่มเดียวกัน และจัดแยกวัตถุที่แตกต่างกันให้อยู่ต่างกลุ่มกัน ซึ่งเป็นการจำแนกโดยที่ไม่ทราบจำนวนกลุ่มและประเภทของกลุ่มล่วงหน้า [8]

การจัดกลุ่มที่ดีจะผลิตกลุ่มที่มีคุณภาพสูง ซึ่งเป็นกลุ่มที่วัตถุภายในกลุ่มเดียวกันมีความคล้ายกันสูง ขณะเดียวกันวัตถุที่อยู่ต่างกลุ่มจะมีความคล้ายกันต่ำ คุณภาพของผลลัพธ์การจัดกลุ่มขึ้นอยู่กับมาตรวัดความคล้ายคลึงที่ใช้และการนำวิธีการไปใช้ให้เกิดผล วิธีการจัดกลุ่มที่ดีควรจะสามารถค้นพบรูปแบบที่ซ่อนอยู่ในข้อมูลบางส่วนหรือทั้งหมด

โดยทั่วไปมาตรวัดค่าความคล้ายหรือไม่คล้ายกันของวัตถุ อยู่ในรูปแบบฟังก์ชันระยะห่าง *d*(*i*,*j*) ซึ่งจัดเก็บในเมทริกซ์ความ(ไม่)คล้าย เปรียบเทียบกับข้อมูลของวัตถุ *n* ตัว แต่ละตัวจะมีคุณลักษณะจำนวน *p* คุณลักษณะ ดังรูปที่ 2.5

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & xi_{f} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & xn_{f} & \dots & x_{np} \end{bmatrix} \qquad \begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \dots & \dots & \dots & \dots & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

รูปที่ 2.5 ตารางเมทริกซ์คุณสมบัติของวัตถุและเมทริกซ์ความไม่คล้าย

นิยามฟังก์ชันระยะห่าง $d(i,j)$ จะมีความแตกต่างกันไปตามประเภทของข้อมูล และวิธีการวิเคราะห์ข้อมูล ทำให้ฟังก์ชันการหาระยะห่างที่ใช้วัดความคล้ายหรือไม่คล้ายระหว่างวัตถุ 2 ตัว มีหลายฟังก์ชัน ซึ่งฟังก์ชันที่ไม่ซับซ้อนและนิยมใช้กันอย่างแพร่หลายคือ การคำนวณระยะทางแบบแมนฮัทตัน (Manhattan Distance) ดังสมการที่ (2.1)

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}| \tag{2.1}$$

วิธีการจัดกลุ่มแบบเคมีน (*k*-Means Clustering) ใช้หลักการการตัดแบ่งของวัตถุ *n* ตัว ออกเป็นจำนวน *k* กลุ่ม (เมื่อทราบค่า *k*) อัลกอริธึมของเคมีน จะทำการตัดแบ่งของวัตถุเป็น *k* กลุ่ม โดยการแทนแต่ละกลุ่มด้วยค่าเฉลี่ยของกลุ่ม ซึ่งใช้เป็นจุดศูนย์กลางของกลุ่มในการวัดระยะห่างของตัวอย่างในกลุ่มเดียวกัน โดยมีอัลกอริธึมดังรูปที่ 2.6

อัลกอริทึมการจัดกลุ่มด้วยวิธีเคมีน

<u>ข้อมูลเริ่มต้น</u>: จำนวนกลุ่ม *k* และวัตถุที่นำมาจัดกลุ่ม

<u>ผลลัพธ์</u>: ชุดของวัตถุในกลุ่ม จำนวน *k* กลุ่ม

<u>วิธีการ</u>:

(1) ทำการเลือกวัตถุจำนวน *k* ตัว เพื่อเป็นตัวแทนกลุ่มเริ่มต้น

(2) ทำซ้ำ

(2.1) ทำการเลือกวัตถุในกลุ่ม จากระยะทางของวัตถุกับตัวแทนกลุ่ม

(2.2) ทำการเลือกตัวแทนกลุ่มใหม่ จากค่าเฉลี่ยของวัตถุในกลุ่ม

(2.3) ทำซ้ำต่อไป จนกระทั่งสมาชิกในกลุ่มไม่มีการเปลี่ยนแปลง

รูปที่ 2.6 อัลกอริธึมของวิธีการการจัดกลุ่มแบบเคมีน [8]

วิธีการจัดกลุ่มแบบเคมีน อาจมีความแตกต่างกันไปบ้าง ในประเด็นของการคัดเลือกค่าเริ่มต้น การคำนวณค่าความคล้ายหรือไม่คล้ายของวัตถุ และวิธีการที่ใช้ในการคำนวณค่าเฉลี่ยของกลุ่ม

ข้อดีของวิธีเคมีน คือ เป็นวิธีการจัดกลุ่มที่มีประสิทธิภาพ เวลาที่ใช้ในการทำงานของอัลกอริธึม คือ $O(tkn)$ โดยที่ *n* คือจำนวนวัตถุ *k* คือจำนวนกลุ่ม และ *t* คือจำนวนรอบที่ทำซ้ำ ซึ่งปกติแล้วค่าของ *k* จะน้อยกว่าค่าของ *n* และ *t* มากๆ ($k << n,t$) แต่วิธีแบบเคมีนมีข้อด้อยที่การทำการจัดกลุ่มข้อมูลจำเป็นต้องกำหนดค่า *k* หรือจำนวนกลุ่มล่วงหน้า สามารถทำการจัดกลุ่มกับวัตถุที่สามารถหาค่าเฉลี่ยได้เท่านั้น และไม่เหมาะกับการจัดกลุ่มที่มีข้อมูลรบกวนหรือข้อมูลผิดปกติเป็นจำนวนมาก

2.5 งานวิจัยที่เกี่ยวข้อง

2.5.1 **ภาพบิตแม็บของข้อมูลอนุกรมเวลา: เครื่องมือการแสดงผลภาพสำหรับการทำงานกับฐานข้อมูลของข้อมูลอนุกรมเวลาขนาดใหญ่** (Time-series Bitmaps: A Practical Visualization Tool for working with Large Time Series Databases) [2]

งานวิจัยนี้นำเสนอเครื่องมือที่ทำให้ผู้ใช้งานสามารถแยกแยะ จัดการ และบริหารข้อมูลอนุกรมเวลาได้ง่ายขึ้น โดยใช้วิธีการจัดการข้อมูลอนุกรมเวลาให้แสดงออกมาเป็นลักษณะรูปภาพบิตแม็บ ซึ่งนำเอาลักษณะต่างๆของข้อมูลอนุกรมเวลา แสดงออกมาเป็นสีสันในภาพบิตแม็บ ทำให้ผู้ใช้งานสามารถที่จะทำการจัดกลุ่ม หาความแตกต่าง จัดการกับชุดของข้อมูลได้อย่างรวดเร็ว และง่ายขึ้น เครื่องมือดังกล่าวนี้ยังสามารถนำไปใช้ร่วมกับระบบปฏิบัติการที่สนับสนุนการต่อประสานด้วยภาพกับผู้ใช้ (Graphical User Interfaces) เช่น ไมโครซอฟต์วินโดวส์ (Microsoft Windows) อควา (Aqua) และ เอ็กซ์วินโดวส์ (X-windows) เป็นต้น ผู้ใช้เครื่องมือนี้จะสามารถจำแนกข้อมูลอนุกรมเวลาจากสีของรูปภาพบิตแม็บ แทนการพิจารณาข้อมูลดิบที่เป็นเลขอนุกรมจำนวนมาก

งานวิจัยนี้ได้นำเอาแนวคิดจากทฤษฎีของเคออสเกม ที่นำไปใช้ในการแสดงผลภาพจากข้อมูลดีเอ็นเอ มาประยุกต์ใช้ในการแสดงผลภาพให้เข้าใจง่าย และดีขึ้น โดยการนำเอาความถี่ของการเกิดสายอักขระต่างๆในข้อมูลดีเอ็นเอมาใช้แทนการกำหนดจุด และยังพัฒนานำไปประยุกต์ใช้กับข้อมูลอนุกรมเวลาที่ยากแก่การพิจารณาข้อมูลโดยรวม ที่มีปริมาณมากได้ดีอีกด้วย แต่อย่างไรก็ตามเครื่องมือนี้ยังจำเป็นต้องใช้ผู้เชี่ยวชาญในการกำหนดตัวแปรสำคัญ ที่มีผลต่อการแสดงผลภาพของแต่ละชนิดและรูปแบบของข้อมูลที่แตกต่างกันออกไป ซึ่งไม่ตรงกับจุดประสงค์ของผู้ทำการวิจัย เช่น การกำหนดขนาดของช่วงแบ่งข้อมูลในวิธีการแบบแซ็ค หรือการกำหนดระดับขั้นของการแสดงภาพบิตแม็บกับลักษณะของผลลัพธ์ที่ต้องการ อีกทั้งการจำแนกประเภทและการแบ่งกลุ่ม จำเป็นต้องใช้ผู้เชี่ยวชาญที่มีความรู้ทั้งทางด้านวิธีการแบบแซ็ค และทางด้านลักษณะของผลลัพธ์ที่ต้องการ จึงจะสามารถกำหนดตัวแปรที่เหมาะสมและแสดงผลลัพธ์ออกมาได้อย่างมีประสิทธิภาพ

2.5.2 **สัญรูปอัจฉริยะ: การทำเหมืองข้อมูลขนาดย่อม และทำการแสดงผลภาพสู่ระบบปฏิบัติการแบบส่วนต่อประสานด้วยภาพกับผู้ใช้** (Intelligent Icons: Integrating Lite-Weight Data Mining and Visualization into GUI Operating Systems) [9]

งานวิจัยนี้นำเสนอแนวคิดที่จะแทนสัญรูป (Icon) มาตรฐาน ที่ปรากฏอยู่ในระบบปฏิบัติการที่สนับสนุนการต่อประสานด้วยภาพกับผู้ใช้ (Graphical User Interfaces) เช่น ไมโครซอฟต์วินโดวส์ (Microsoft Windows) โอเอสเอกซ์ (OS X) หรือลีนุกซ์ (Linux) ด้วยสัญรูปที่

ถูกสร้างขึ้นมาโดยอัตโนมัติ โดยลักษณะของสัญรูปใหม่จะมีความสัมพันธ์กับคุณสมบัติในข้อมูลนั้นๆ ซึ่งงานวิจัยนี้เรียกว่า สัญรูปอัจฉริยะ (Intelligent Icon) นอกจากนี้ยังสามารถทำการจัดกลุ่มข้อมูลจากความเหมือน หรือความแตกต่างของสัญรูปได้อีกด้วย

งานวิจัยนี้มีแนวทางมาจากวิชวลไอดีส์ (VisualIDs) [10] ที่มีแนวความคิดที่ว่าการค้นหาและจดจำภาพทำให้สามารถจดจำและค้นหาได้ดีและมีประสิทธิภาพ กว่าการค้นหาและการจำเป็นประโยค ซึ่งวิชวลไอดีส์จะทำการสร้างภาพสัญรูป (Icon) ที่เกิดมาจากขั้นตอนวิธีแบบแฮช (Hashing) กับชื่อของแฟ้มข้อมูล (File) ด้วยเหตุนี้ทำให้แฟ้มข้อมูลที่มีชื่อคล้ายกันจะได้สัญรูปที่มีลักษณะใกล้เคียงกัน สามารถทำให้ผู้ใช้สามารถเห็นความเหมือนหรือแตกต่างกันของแฟ้มข้อมูลได้ง่ายขึ้น ซึ่งสัญรูปอัจฉริยะนำเอาแนวคิดนี้มาประยุกต์โดยการนำเอาข้อมูลของแฟ้มข้อมูลนั้นมาทำการสร้างสัญรูปแทนที่จะใช้เพียงชื่อของแฟ้มข้อมูล

ถึงแม้ว่างานวิจัยนี้จะนำเอาสัญรูปอัจฉริยะ มาทำการทดลองกับข้อมูลในรูปแบบต่างๆ ซึ่งได้ผลการทดลองที่น่าสนใจ แต่อย่างไรก็ตามผลการทดลองของข้อมูลหน่วยพันธุกรรมและข้อมูลอนุกรมเวลาได้เคยถูกนำเสนอในงานวิจัยอื่นๆมาแล้ว [2][3][5][6] สำหรับข้อมูลเอกสารและข้อมูลที่เป็นวีดีโอเกม ที่นำเสนอในงานวิจัยนี้ไม่ได้ระบุถึงขั้นตอนที่จำเป็นในการนำเอาข้อมูลมาทำให้เป็น สัญรูปอัจฉริยะ เช่นหลักการแปลงข้อมูลตัวอักษรของเอกสารมาเป็นสัญรูป ซึ่งในงานวิจัยได้กล่าวถึงเพียงหลักการพิจารณาคุณลักษณะของเอกสารเพียงเบื้องต้นเท่านั้น อีกทั้งไม่มีหลักการและเหตุผลในการกำหนดรูปแบบ (Template) ที่แตกต่างกันของสัญรูปอัจฉริยะสำหรับข้อมูลแต่ละประเภท

บทที่ 3
การออกแบบวิธีการแสดงผลภาพสำหรับข้อมูลเอกสารดิจิทัล

งานวิจัยนี้เป็นการพัฒนาวิธีแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล ซึ่งอาศัยแนวคิดจากทฤษฎีเคออสเกม (Chaos Game) [1] ประยุกต์ร่วมกับ การแสดงผลภาพบิตแม็บของข้อมูลอนุกรมเวลา โดยใช้วิธีแบบแซ็ค (SAX Symbolic Representation) ในการกำหนดอักขระที่เกิดขึ้น [2][6] ซึ่งวิธีดังกล่าวจะนำเอาความถี่ของสายอักขระที่เป็นข้อมูลอนุกรมเวลารวมถึงข้อมูลจำพวกสายอักขระของดีเอ็นเอ แปลงออกมาเป็นรูปภาพบิตแม็บ ซึ่งข้อมูลที่นำมาศึกษาในลักษณะนี้เป็นข้อมูลที่มีขนาดใหญ่และมีปริมาณมาก ยากแก่การนำมาพิจารณาจัดหมวดหมู่ และต้องอาศัยความเข้าใจในรายละเอียดของข้อมูลดิบนั้นๆ

จากแนวคิดในข้างต้น งานวิจัยนี้จึงนำเสนอวิธีจัดการข้อมูลประเภทข้อความเอกสารพื้นฐาน (Plain Text) นอกเหนือจากข้อมูลที่เป็นสายอักขระดีเอ็นเอ (DNA String) และข้อมูลอนุกรมเวลา (Time Series Data) ซึ่งจุดประสงค์ของงานวิจัยนี้ต้องการแปลงข้อมูลเอกสารให้สามารถแสดงเป็นรูปภาพบิตแม็บ (Bitmap Representation) ตามคุณลักษณะนั้นๆ ของเอกสาร

ขั้นตอนออกแบบและการดำเนินการดังกล่าวจะเริ่มจากการแปลงข้อมูลเอกสาร ให้อยู่ในรูปข้อมูลอนุกรมเวลา (Time Series Data) แล้วนำไปผ่านกระบวนการเพื่อแปลงข้อมูลออกมาเป็นอักขระ [2] เพื่อที่จะนำไปสร้างรูปภาพบิตแม็บ [2] รายละเอียดของแนวคิดและขั้นตอนสำหรับการพัฒนาการแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัลมีดังนี้

3.1 การแปลงข้อมูลจากเอกสารดิจิทัลไปเป็นข้อมูลอนุกรมเวลา

การแปลงข้อมูลจากเอกสารดิจิทัลไปเป็นข้อมูลอนุกรมเวลานั้น จำเป็นต้องมีการวิเคราะห์และปรับข้อมูลในเอกสารก่อน เพื่อลดข้อแตกต่างรวมถึงสิ่งผิดปกติต่างๆ ที่อาจเกิดขึ้นกับเอกสารซึ่งมีผลทำให้การแสดงผลภาพของเอกสารออกมาคลาดเคลื่อนได้ และต้องเลือกใช้วิธีที่เหมาะสมในการแปลงข้อมูลตัวอักษรไปเป็นข้อมูลแบบเลขจำนวนจริง อีกทั้งต้องทำการปรับข้อมูลอนุกรมเวลาที่ได้มา เพื่อลดความแปรปรวนของข้อมูล ซึ่งมีรายละเอียดตามขั้นตอนดังกล่าวดังนี้

3.1.1 การวิเคราะห์และปรับแต่งเอกสารดิจิทัล

เนื่องจากข้อมูลในเอกสารโดยทั่วไปมีการใช้ตัวอักษร และเครื่องหมายพิเศษแตกต่างกันไป ตามแต่ลักษณะและแนวการเขียนของเอกสารนั้นๆ งานวิจัยนี้จึงพยายามปรับเอกสารโดยการลดข้อแตกต่างที่อาจมีผลต่อการแสดงผลภาพเอกสารบิตแม็บ ดังนี้

3.1.1.1 การปรับตัวอักษรให้เป็นตัวพิมพ์ใหญ่

การปรับตัวอักษรในเอกสารเป็นตัวพิมพ์ใหญ่ทุกตัว (Uppercase Letter) เพื่อลดการเกิดความแตกต่างของตัวอักษรในการวิเคราะห์เอกสาร โดยทำการปรับตัวอักษร คำ หรือ ประโยค ที่อยู่ในเอกสารให้เป็นตัวพิมพ์ใหญ่ทั้งหมด แสดงตัวอย่างการปรับตัวอักษรให้เป็นตัวพิมพ์ใหญ่ ดังรูปที่ 3.1

Now when he came near to Egypt, he said to Sarai, his wife, Truly, you are a fair woman and beautiful to the eye. เมื่อผ่านการปรับตัวอักษรให้เป็นตัวพิมพ์ใหญ่ทั้งหมดแล้ว จะได้ผลเป็น NOW WHEN HE CAME NEAR TO EGYPT, HE SAID TO SARAI, HIS WIFE, TRULY, YOU ARE A FAIR WOMAN AND BEAUTIFUL TO THE EYE.

รูปที่ 3.1 ตัวอย่างการปรับตัวอักษรให้เป็นตัวพิมพ์ใหญ่

3.1.1.2 การกำจัดคำที่ไม่มีนัยสำคัญ

การกำจัดคำที่ไม่มีนัยสำคัญ หรือ คำหยุด (Stop Word) เป็นการกำจัดกลุ่มคำที่อาจไม่ส่งผล หรือไม่ได้ทำให้การพิจารณาเอกสารเกิดข้อแตกต่าง เพื่อกรองคำที่ไม่สื่อถึงความสำคัญของเอกสารออกไป โดยทำการตัดคำด้วยช่องว่าง (Space Character) ในเอกสารเพื่อเปรียบเทียบกับคำที่ไม่มีนัยสำคัญจำนวน 331 คำ ดังตัวอย่างในรูปที่ 3.2 (แสดงรายการคำที่ไม่มีนัยสำคัญทั้งหมดในภาคผนวก ก)

Now when he came near to Egypt, he said to Sarai, his wife, Truly, you are a fair woman and beautiful to the eye. เมื่อผ่านการกำจัดคำที่ไม่มีนัยสำคัญ หรือ คำหยุด (Stop Word) จะได้ผลเป็น he came near Egypt, he said Sarai, his wife, Truly, you fair woman beautiful the eye.

รูปที่ 3.2 ตัวอย่างการกำจัดคำที่ไม่มีนัยสำคัญ หรือ คำหยุด (Stop Word)

3.1.1.3 การกำจัดอักษรพิเศษ

การกำจัดอักษรพิเศษ (Special Character) เป็นการกำจัดตัวอักษรพิเศษต่างๆ ที่เกิดขึ้นในเอกสารทั่วๆ ไป โดยทำการตัดคำด้วยช่องว่าง (Space Character) ในเอกสารเพื่อเปรียบเทียบกับอักษรพิเศษจำนวน 33 อักษร ดังตัวอย่างในรูปที่ 3.3 (แสดงรายการอักษรพิเศษทั้งหมดในภาคผนวก ข)

Now when he came near to Egypt, he said to Sarai, his wife, Truly, you are a fair woman and beautiful to the eye. เมื่อผ่านการกำจัดอักษรพิเศษ (Special Character) จะได้ผลเป็น Now when he came near to Egypt he said to Sarai his wife Truly you are a fair woman and beautiful to the eye

รูปที่ 3.3 ตัวอย่างการกำจัดอักษรพิเศษ (Special Character)

3.1.1.4 การกำหนดขีดจำกัดของตัวอักษร

การกำหนดขีดจำกัดของตัวอักษรในเอกสาร หรือกำหนดจำนวนตัวอักษรที่นำมาวิเคราะห์ เป็นการกำหนดขอบเขตการพิจารณาขนาดของเอกสาร ในการประมวลผลภาพเอกสารบิตแม็บ เพื่อให้การแสดงผลภาพเอกสารดิจิทัลมีความรวดเร็วและมีประสิทธิภาพ โดยจะทำการพิจารณาเอกสารตามค่าขีดจำกัดที่กำหนดไว้

การวิเคราะห์และปรับเอกสารดิจิทัลดังที่กล่าวในข้างต้น สามารถเลือกใช้การวิเคราะห์และปรับเอกสารดิจิทัลทั้งหมดหรือบางส่วนได้ ซึ่งขึ้นกับชนิดของเอกสาร และความสัมพันธ์ของผลการทดลองที่จะกล่าวต่อไปในบทที่ 4

3.1.2 การแปลงตัวอักษรไปเป็นตัวเลข

การแปลงข้อมูลจากตัวอักษรของเอกสารไปเป็นตัวเลข หรือชุดข้อมูลอนุกรมเวลามีได้หลายวิธี เช่น การกำหนดค่าเลขจำนวนจริงกับตัวอักษรเฉพาะ การเทียบตัวอักษรกับค่าเลขจำนวนจริงที่มาจากการเข้ารหัส แต่ในงานวิจัยนี้เลือกใช้วิธีของมาตรฐานแอสกี (ASCII) ในการแปลงข้อมูลจากตัวอักษรไปเป็นตัวเลข ซึ่งเป็นวิธีที่ได้รับความนิยม สะดวกในการใช้งาน และเป็นมาตรฐานที่เป็นที่ยอมรับโดยทั่วไป แสดงตัวอย่างการแปลงข้อมูลจากตัวอักษรไปเป็นตัวเลขตามมาตรฐานแอสกี ดังตารางที่ 3.1

ตารางที่ 3.1 ตัวอย่างการเปรียบเทียบระหว่างเลขจำนวนจริงและตัวอักษรจากมาตรฐานแอสกี

Decimal Number	Character
64	@
65	A
66	B
67	C
…	…

การใช้มาตรฐานแอสกีในการแปลงตัวอักษรให้เป็นเลขจำนวนจริงนี้ รวมถึงการแปลงช่องว่างและอักขระพิเศษต่างๆด้วย ผลที่ได้จะเป็นชุดของข้อมูลเลขจำนวนจริง ซึ่งจัดได้ว่าเป็นข้อมูลอนุกรมเวลาที่มีขนาดจำนวนตัวเลขเท่ากับขนาดจำนวนตัวอักษรของเอกสาร

3.1.3 การปรับข้อมูลอนุกรมเวลาที่ได้จากข้อมูลเอกสารดิจิทัล

การปรับข้อมูลอนุกรมเวลามีจุดประสงค์เพื่อให้ การแสดงแนวโน้มและรูปแบบของการเคลื่อนไหวของข้อมูลที่ต้องการมีความชัดเจนมากขึ้น อีกทั้งยังช่วยลดความแปรปรวน สัญญาณรบกวน และการเปลี่ยนแปลงที่เกิดขึ้นอย่างฉับพลันของข้อมูลได้ ซึ่งทำให้เห็นพฤติกรรมการเปลี่ยนแปลงของตัวแปรที่ผันไปพร้อมๆกับเวลา การปรับข้อมูลในงานวิจัยนี้จะใช้การปรับเรียบ (Smooth) ข้อมูล โดยการคำนวณค่าเฉลี่ยเคลื่อนที่ (Moving Average) ซึ่งเป็นวิธีที่ไม่มีความซับซ้อน และนิยมใช้ในการวิเคราะห์ข้อมูลอนุกรมเวลา การกำหนดค่าขีดแบ่ง (Threshold) ของการคำนวณค่าเฉลี่ยเคลื่อนที่นั้น ขึ้นอยู่กับความต้องการที่จะพิจารณาแนวโน้มของข้อมูลในระยะเวลาขนาดเท่าใด

ในงานวิจัยนี้ทำการปรับเรียบข้อมูล โดยใช้วิธีโดยการคำนวณค่าเฉลี่ยเคลื่อนที่แบบพื้นฐาน (Simple Moving Average) ซึ่งไม่มีการกำหนดค่าน้ำหนัก (Weight) ของแต่ละตัวแปร การคำนวณค่าเฉลี่ยเคลื่อนที่แบบพื้นฐานจะทำการหาค่ากลาง (Mean) ของข้อมูลอนุกรมเวลาจำนวน *n* ข้อมูลล่วงหน้า ดังสมการที่ (3.1)

$$SimpleMovingAverage = \frac{P_m + P_{m+1} + P_{m+2} + \ldots + P_{m+n}}{n} \tag{3.1}$$

โดยที่ *P* คือ ข้อมูลอนุกรมเวลาที่เป็นเลขจำนวนจริง ณ ตำแหน่งใดๆ

m คือ ค่าตำแหน่งใดๆ ของข้อมูลอนุกรมเวลา

n คือ ขนาดของค่าขีดแบ่ง (จำนวนข้อมูลล่วงหน้า)

การกำหนดขนาดของค่าขีดแบ่ง สามารถกำหนดได้ตามความเหมาะสมขึ้นกับรูปแบบของข้อมูล และจุดประสงค์ที่ต้องการนำไปใช้ เช่น รูปแบบการเคลื่อนที่ของข้อมูล ระยะการมองแนวโน้มของข้อมูลที่อาจเป็นระยะสั้น ระยะกลาง หรือระยะยาว แสดงตัวอย่างการคำนวณค่าเฉลี่ยเคลื่อนที่แบบพื้นฐาน ดังรูปที่ 3.4

ตัวอย่างชุดข้อมูล 12, 36, 25, 44, 13, 16, 45, 49, 85, 59, 63, 12, 10, 2, 78, 32, 38, 89, 65, 48 เมื่อกำหนด n = 5 จะได้ชุดข้อมูลใหม่หลังทำการคำนวณค่าเฉลี่ยเคลื่อนที่แบบพื้นฐาน คือ 26, 26.8, 28.6, 33.4, 41.6, 50.8, 60.2, 53.6, 45.8, 29.2, 33, 26.8, 32, 47.8, 60.4, 54.4, 38, 89, 65, 48

รูปที่ 3.4 ตัวอย่างการคำนวณค่าเฉลี่ยเคลื่อนที่แบบพื้นฐาน (Simple Moving Average)

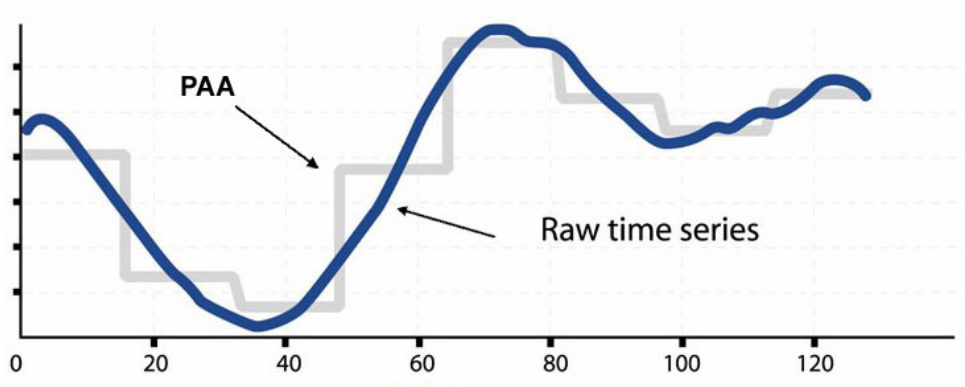
อย่างไรก็ตามการคำนวณค่าเฉลี่ยเคลื่อนที่แบบพื้นฐาน มีข้อจำกัดอยู่ที่ไม่สามารถคำนวนหาค่าเฉลี่ยของชุดข้อมูล n - 1 ตัวสุดท้ายได้ ทำให้ไม่สามารถทำการปรับเรียบข้อมูลช่วง n-1 สุดท้ายได้

3.2 การแปลงข้อมูลอนุกรมเวลาไปเป็นอักขระ

งานวิจัยนี้เลือกใช้การแปลงข้อมูลอนุกรมเวลาไปเป็นข้อมูลสัญลักษณ์หรืออักขระ ที่ได้มาจากการหาค่าเฉลี่ยโดยรวม หรือแซ็ค (Symbolic Aggregate approXimation - SAX) เนื่องจากวิธีการแบบแซ็คเป็นวิธีการที่ไม่ซับซ้อน ประมวลผลเร็ว และข้อมูลจากการแปลงจะไม่เสียคุณสมบัติจากข้อมูลเดิม [2] การแปลงข้อมูลเอกสารเป็นภาพบิตแม็บนี้ พัฒนาโดยอาศัยแนวทางวิธีการแบบแซ็ค ซึ่งมีรายละเอียดและขั้นตอนวิธีดังนี้

3.2.1 การลดขนาดหรือมิติของข้อมูลโดยวิธีสัดส่วนจำนวนเฉลี่ย

การลดขนาดหรือมิติของข้อมูลโดยสัดส่วนจำนวนเฉลี่ย (PAA Dimensionality Reduction) เป็นวิธีการลดขนาดของข้อมูลอนุกรมเวลาแบบแบ่งเป็นช่วงๆ โดยข้อมูลแต่ละช่วงจะเป็นส่วนของข้อมูลที่ถูกแบ่งด้วยขนาดใดๆที่กำหนด ซึ่งข้อมูลเลขจำนวนจริงในแต่ละส่วนจะถูกหาค่าเฉลี่ยจากทุกๆข้อมูลในช่วงนั้น กล่าวคือข้อมูลอนุกรมเวลาจะถูกลดขนาดลงโดยการนำไปหาค่าเฉลี่ยในแต่ละช่วงของข้อมูล ดังรูปที่ 3.5



รูปที่ 3.5 การลดขนาดของข้อมูลโดยสัดส่วนจำนวนเฉลี่ย (ที่มา: Lin et al.) [6]

3.2.2 การแปลงข้อมูลให้อยู่ในรูปการกระจายแบบเกาส์เซียน

เป็นการนำเอาข้อมูลที่ได้จากการลดขนาดหรือมิติในหัวข้อ 3.2.1 มาทำให้เป็นบรรทัดฐานโดยการแปลงข้อมูลอยู่ในรูปการกระจายแบบเกาส์เซียน ซึ่งเป็นขั้นตอนในการแปลงข้อมูลอนุกรมเวลาไปเป็นข้อมูลสัญลักษณ์แบบแซ็ค ดังสมการที่ (3.2)

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{3.2}$$

โดยที่ Z คือ ค่าบรรทัดฐานของการกระจายแบบเกาส์เซียน

x คือ ข้อมูลเลขจำนวนจริง ณ ตำแหน่งใดๆ ของข้อมูลอนุกรมเวลา

μ คือ ค่าเฉลี่ยของชุดข้อมูล

σ คือ ค่าความเบี่ยงเบนมาตรฐาน

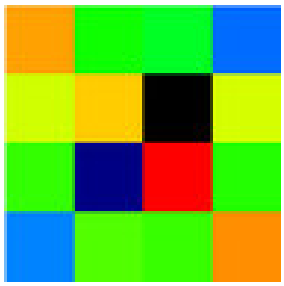
การทำให้ข้อมูลอนุกรมเวลา อยู่ในรูปการกระจายแบบเกาส์เซียนนั้น เพื่อที่จะสามารถนำข้อมูลนั้นไปเปรียบเทียบกับตารางการกระจายของข้อมูลแบบเกาส์เซียน ซึ่งจะกล่าวในหัวข้อ 3.2.4 ต่อไป

3.2.3 การกำหนดจำนวนอักขระ

การกำหนดจำนวนอักขระ คือการกำหนดจำนวนของอักขระที่มีโอกาสที่เกิดขึ้นจากการแปลงข้อมูลอนุกรมเวลา ซึ่งวิธีการกำหนดจำนวนอักขระจะพิจารณาจากรูปแบบการนำเอาไปใช้ของข้อมูล เนื่องจากงานวิจัยนี้เป็นการสร้างภาพบิตแม็บจากข้อมูลเอกสาร ที่ภาพบิตแม็บมีลักษณะเป็นรูปสี่เหลี่ยมจัตุรัส ซึ่งมีการแบ่งโครงสร้างภายในเป็นส่วนต่างๆ ที่มีความสมมาตรกันเป็นแบบตารางเมทริกซ์ นอกจากนี้ยังต้องพิจารณาการเลือกใช้ระดับของเมทริกซ์ (กล่าวรายละเอียดในหัวข้อ 3.3.1) ที่ส่งผลต่อการเลือกกำหนดจำนวนอักขระด้วย ดังนั้นการกำหนดจำนวนอักขระเพื่อให้สามารถนำเอาอักขระไปกำกับตามช่องต่างๆ ในตารางเมทริกซ์ได้

เหมาะสม ในงานวิจัยจึงเลือกกำหนดอักขระจำนวน 4 อักขระ คือ อักขระ a b c และ d เพราะสามารถรองรับการกำหนดระดับของตารางเมทริกซ์ได้ทุกระดับ แสดงตัวอย่างของตารางเมทริกซ์ที่ใช้ในการแสดงภาพบิตแม็บจากเอกสาร ดังรูปที่ 3.6

aa	ab	ba	bb
ac	ad	bc	bd
ca	cb	da	db
cc	cd	dc	dd



รูปที่ 3.6 ตารางเมทริกซ์ที่กำกับด้วยอักขระและตัวอย่างของภาพบิตแม็บ

นอกจากนี้การกำหนดอักขระจำนวน 4 อักขระทำให้ข้อมูลอนุกรมเวลาถูกแบ่งได้ออกเพียง 4 ช่วงเท่านั้น ทำให้ช่วงการแปลงข้อมูลอนุกรมเวลาไม่แคบจนเกินไป และไม่ทำให้เกิดการกระจายของข้อมูลที่มากเกินไปเช่นกัน

โดยสีที่แสดงออกมาในช่องของเมทริกซ์รูปสี่เหลี่ยม เกิดขึ้นจากความถี่ของตัวอักษรที่มาจากข้อมูลของเอกสาร ซึ่งกล่าวในหัวข้อ 3.3.2 ต่อไป

3.2.4 ตารางการกระจายข้อมูลของเกาส์เซียน

ตารางการกระจายข้อมูลของเกาส์เซียน เป็นตารางบอกค่าทางสถิติของการกระจายข้อมูลแบบไม่ต่อเนื่อง ซึ่งตารางจะแสดงช่วงตัวเลขที่เป็นไปตามลำดับขั้นที่ถูกแบ่งแต่ละช่วงด้วยจุดขั้น (Breakpoint) โดยแต่ละจุดขั้นจะเป็นขีดแบ่งเขตพื้นที่ใต้กราฟจำนวน *n* ส่วนที่มีขนาดเท่าๆกัน [11] เช่น แสดงจุดขั้นตั้งแต่ 3 จุดจนถึง 10 จุดขั้น ดังแสดงในรูปที่ 3.7

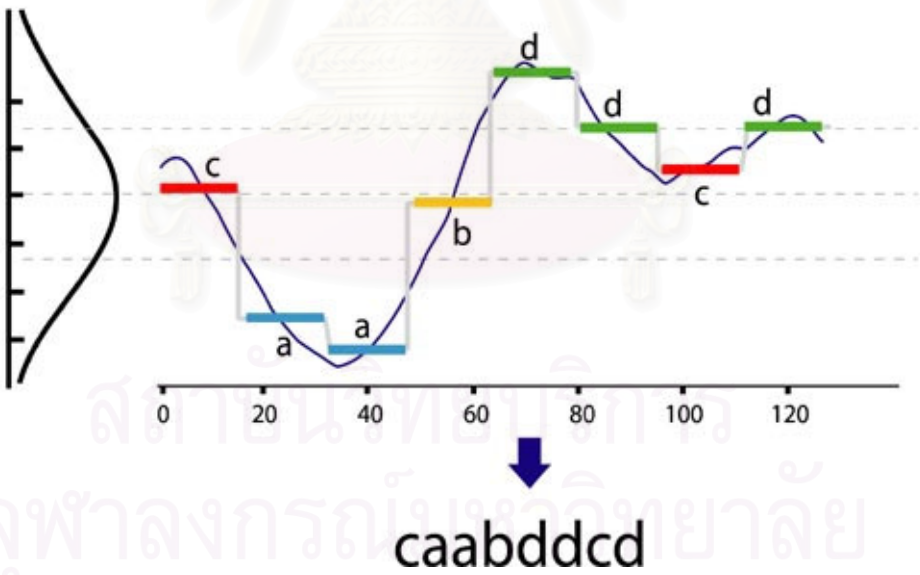
เนื่องจากข้อมูลอนุกรมเวลาที่ได้มาจากการแปลงข้อมูลเอกสาร ที่เป็นข้อมูลเลขจำนวนจริงและถูกทำข้อมูลให้เป็นบรรทัดฐาน ในหัวข้อ 3.2.2 ซึ่งจึงจัดว่าเป็นข้อมูลที่อยู่ในรูปแบบการกระจายแบบเกาส์เซียน ส่งผลให้ข้อมูลดังกล่าวมีลักษณะเป็นข้อมูลแบบไม่ต่อเนื่อง (Discrete Representation) ทำให้สามารถพิจารณาการแปลงข้อมูลเป็นอักขระโดยวิธีการของข้อมูลแบบไม่ต่อเนื่องได้ [12] ซึ่งการเลือกใช้ขนาดของจุดขั้นจะขึ้นกับจำนวนอักขระที่ต้องการแปลง ในหัวข้อ 3.2.3 ได้กำหนดจำนวนอักขระที่ต้องการแปลงไว้ 4 อักขระ ทำให้สามารถนำเอาค่าเขตแบ่งจากตารางค่าทางสถิติของการกระจายข้อมูลแบบไม่ต่อเนื่องแบบเกาส์เซียน ที่มีจุดขั้น 4 จุด คือค่าช่วงแบ่ง -0.67 0 และ 0.67 เพื่อนำไปใช้ในขั้นตอนการแปลงข้อมูลเลขจำนวนจริงไปเป็นอักขระในหัวข้อ 3.2.5 ต่อไป

β_i \ a	3	4	5	6	7	8	9	10
β_1	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15	-1.22	-1.28
β_2	0.43	0	-0.25	-0.43	-0.57	-0.67	-0.76	-0.84
β_3		0.67	0.25	0	-0.18	-0.32	-0.43	-0.52
β_4			0.84	0.43	0.18	0	-0.14	-0.25
β_5				0.97	0.57	0.32	0.14	0
β_6					1.07	0.67	0.43	0.25
β_7						1.15	0.76	0.52
β_8							1.22	0.84
β_9								1.28

รูปที่ 3.7 ตารางการกระจายข้อมูลแบบไม่ต่อเนื่องของเกาส์เซียน

3.2.5 การแปลงข้อมูลเลขจำนวนจริงไปเป็นอักขระ

เป็นการแปลงข้อมูลเลขจำนวนจริงไปเป็นอักขระ โดยการเปรียบเทียบกับค่าช่วงแบ่งที่ได้จากตารางการกระจายข้อมูลแบบไม่ต่อเนื่องของเกาส์เซียนกับอักขระ ซึ่งได้ทำการกำหนดจำนวนอักขระที่จะทำการแปลงข้อมูล และขนาดจำนวนจุดขั้นในตารางค่าทางสถิติไว้ในหัวข้อ 3.2.3 และ 3.2.4 ซึ่งแสดงตัวอย่างการแปลงเลขจำนวนจริงไปเป็นอักขระ ดังรูปที่ 3.8



รูปที่ 3.8 การแปลงข้อมูลอนุกรมเวลาเป็นอักขระโดยกำหนดจุดขั้นจำนวน 4 จุด

จากรูปที่ 3.8 แสดงตัวอย่างการแปลงข้อมูลอนุกรมเวลาเป็นอักขระ โดยมีการกำหนดจุดขั้นกับข้อมูลเป็น 4 ช่วง และกำหนดช่วงแรกเป็นอักขระ a ช่วงที่สองเป็น b ช่วงที่สามและช่วงสุดท้ายเป็นอักขระ c และ d ตามลำดับ ซึ่งได้สายอักขระที่แปลงจากข้อมูลอนุกรมเวลาเป็น “caabddcd”

3.3 การแปลงอักขระไปเป็นภาพบิตแม็บ

การแปลงข้อมูลอักขระไปเป็นภาพบิตแม็บ เป็นแนวคิดที่ประยุกต์มาจากการแสดงผลภาพแบบเคออสเกม (Chaos Game Representations) [1][4] ซึ่งเป็นแนวคิดที่นิยมนำมาใช้กันมากในงานวิจัยที่เกี่ยวข้องกับการเรียนรู้ข้อมูลของดีเอ็นเอ (DNA) และถูกนำมาพัฒนาประยุกต์ใช้กับข้อมูลอนุกรมเวลา งานวิจัยนี้จึงนำเอาแนวคิดดังกล่าวมาปรับใช้กับข้อมูลที่เป็นข้อความเอกสาร โดยมีขั้นตอนในการแปลงข้อมูลอักขระไปเป็นภาพบิตแม็บ ดังนี้

3.3.1 การกำหนดระดับและรูปแบบอักขระของตารางเมทริกซ์

ระดับของตารางเมทริกซ์ ที่ประยุกต์มาจากการแสดงรูปภาพแบบเคออสเกมมีรูปแบบของตารางได้หลายระดับขั้น ขึ้นอยู่กับความต้องการในการแสดงข้อมูลที่ซับซ้อนและละเอียดมากเพียงใด ซึ่งมีความสัมพันธ์กับการแทนค่าอักขระในช่องของเมทริกซ์ด้วย แสดงตัวอย่างของตารางเมทริกซ์ ดังรูปที่ 3.9

a	b
c	d

ระดับที่ 1

aa	ab	ba	bb
ac	ad	bc	bd
ca	cb	da	db
cc	cd	dc	dd

ระดับที่ 2

รูปที่ 3.9 ลักษณะตารางเมทริกซ์ในระดับที่ 1 และระดับที่ 2

รูปที่ 3.9 แสดงรูปแบบของตารางเมทริกซ์ในระดับที่ 1 และระดับที่ 2 ที่ถูกกำกับด้วยอักขระจำนวน 4 อักขระ ในแต่ละช่องของตารางเมทริกซ์ ซึ่งเป็นโครงสร้างในการสร้างภาพบิตแม็บจากข้อมูลเอกสาร

งานวิจัยนี้เลือกการแสดงภาพบิตแม็บจากข้อมูลเอกสาร โดยใช้ตารางเมทริกซ์ระดับที่ 2 เนื่องจากภาพบิตแม็บที่ได้มีลักษณะเป็นตารางขนาด 16 ช่อง ซึ่งแต่ละช่องถูกแทนค่าด้วยความถี่ของคู่อักขระ ที่สามารถเกิดจากอักขระจำนวน 4 อักขระ ที่ได้กำหนดไว้ในหัวข้อ 3.2.3 และภาพบิตแม็บที่ได้ มีความชัดเจนเพียงพอต่อการแยกแยะ เพราะมีช่องแสดงสีในเมทริกซ์ที่สามารถแสดงสีที่แตกต่างกันได้ถึง 16 ช่อง นอกจากนี้วัตถุประสงค์ของงานวิจัยพยายามที่จะทำการวิเคราะห์ข้อมูลเอกสารดิจิทัลที่มีความยาวจำกัด เพื่อให้การแสดงผลภาพบิตแม็บมีความสามารถในการประมวลผลได้รวดเร็ว และทำงานได้อย่างมีประสิทธิภาพ แต่หากใช้

ตารางเมทริกซ์ในระดับมากกว่า 2 มีผลให้จำเป็นต้องใช้ความยาวของข้อมูลเอกสารที่เพิ่มขึ้นเพื่อให้เพียงพอต่อความถี่ของแต่ละอักขระที่มีเพิ่มมากขึ้น ซึ่งส่งผลให้การประมวลผลช้าและไม่มีประสิทธิภาพได้

3.3.2 การนับความถี่ของสายอักขระ

เนื่องจากการกำหนดระดับและรูปแบบของตารางเมทริกซ์ในระดับที่ 2 และเลือกจำนวนอักขระที่ใช้ในการแปลงข้อมูลจำนวน 4 อักขระ การนับความถี่ของสายอักขระจะทำการนับผลรวมความถี่ของอักขระเป็นคู่ๆ ตั้งแต่สายอักขระตัวแรกของชุดข้อมูลถึงตัวสุดท้าย เริ่มตั้งแต่อักขระ aa ab ac จนถึงข้อมูลในตารางคู่สุดท้ายคือ dc และ dd โดยการนับจำนวนคู่ของอักขระนั้นๆ ว่าเกิดขึ้นจำนวนเท่าใด แล้วนำค่าความถี่ที่ได้ไปกำกับในช่องของตารางเมทริกซ์ ซึ่งแสดงตัวอย่างการนับความถี่ของสายอักขระ ดังรูปที่ 3.10 และแสดงตัวอย่างการนำค่าความถี่ไปกำกับในช่องของตารางเมทริกซ์ ดังรูปที่ 3.11

ตัวอย่างข้อมูล เช่น

bdddcdabbdbcccbabdbbdcbdadcdddaddccdabcabddbaddcadcccdcacdaacddd

aadddbaddaad

นับความถี่ของ “aa” ได้เท่ากับ 3 ซึ่งมาจาก

bdddcdabbdbcccbabdbbdcbdadcdddaddccdabcabddbaddcadcccdcacd<u>aa</u>cddd

<u>aa</u>dddbadd<u>aa</u>d

รูปที่ 3.10 การนับความถี่ของคู่อักขระ “aa”

aa	ab	ba	bb
ac	ad	bc	bd
ca	cb	da	db
cc	cd	dc	dd

78	63	57	33
38	84	30	70
45	30	83	63
38	66	74	67

รูปที่ 3.11 ตารางเมทริกซ์กับผลการนับความถี่ของข้อมูล

3.3.3 ค่าบรรทัดฐานมากที่สุดและน้อยที่สุด

ค่าบรรทัดฐานมากที่สุดและน้อยที่สุด เป็นการปรับชุดของข้อมูลตัวเลขกลุ่มหนึ่งให้มีค่ามากที่สุดและน้อยที่สุดในชุดข้อมูลเป็นค่าที่อยู่ในช่วงที่ต้องการ เพื่อนำไปใช้ประโยชน์ในการเปรียบเทียบชุดข้อมูลกับมาตรฐาน หรือมาตรวัดบางชนิด

ในงานวิจัยนี้ต้องการปรับค่าความถี่ที่ได้มาจากหัวข้อ 3.3.2 เพื่อให้สามารถนำไปเปรียบเทียบกับค่าที่กำกับในแถบสีอาร์จีบี (กล่าวต่อไปในหัวข้อ 3.3.4) โดยนำค่าความถี่นั้นมาหาค่าสัดส่วนแบบค่าบรรทัดฐานมากที่สุดและน้อยที่สุด (Min-Max Normalization) ซึ่งกำหนดค่ามากที่สุดเป็น 1 และ ค่าน้อยที่สุดเป็น 0 เพื่อทำให้ค่าความถี่กำหนดในตารางเมทริกซ์ อยู่ในช่วงค่ามากที่สุดในเมทริกซ์เป็น 1 และค่าน้อยที่สุดในเมทริกซ์เป็น 0 ซึ่งคำนวณจากสมการที่ (3.3)

$$v' = \frac{v - min_A}{max_A - min_A}(new_max_A - new_min_A) + new_min_A \qquad (3.3)$$

โดยที่ v' คือ ค่าความถี่ใหม่หลังจากผ่านการหาค่าสัดส่วนแบบค่าบรรทัดฐานมากที่สุดและน้อยที่สุด

v คือ ค่าความถี่เดิม

min_A คือ ค่าความถี่ที่น้อยที่สุดในชุดข้อมูล

max_A คือ ค่าความถี่ที่มากที่สุดในชุดข้อมูล

new_min_A คือ ค่าน้อยที่สุดที่กำหนด

new_max_A คือ ค่ามากที่สุดที่กำหนด

เมื่อทำการปรับค่าสัดส่วนกับความถี่ แบบค่าบรรทัดฐานมากที่สุดและน้อยที่สุดกับชุดข้อมูลค่าความถี่แล้ว ทำให้ได้ค่าความถี่ในเมทริกซ์ออกมาใหม่ แสดงตัวอย่างการปรับค่าสัดส่วนกับความถี่แบบค่าบรรทัดฐานมากที่สุดและน้อยที่สุด ดังรูปที่ 3.12

78	63	57	33
38	84	30	70
45	30	83	63
38	66	74	67

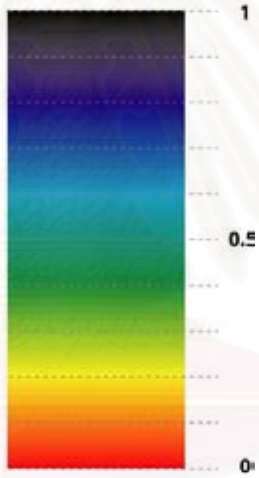
0.889	0.611	0.5	0.056
0.148	1	0	0.741
0.278	0	0.981	0.611
0.148	0.667	0.815	0.685

รูปที่ 3.12 ตารางเมทริกซ์เปรียบเทียบค่าที่ได้จากการนับความถี่ของข้อมูล และค่าความถี่ที่ได้จากการหาค่าสัดส่วนแบบปรับค่าบรรทัดฐานมากที่สุดและน้อยที่สุด

3.3.4 การกำหนดระดับขั้นของแถบสีอาร์จีบี

สีอาร์จีบีเป็นคุณลักษณะของการแสดงสีแบบหนึ่ง ที่ประกอบไปด้วยค่าสี 3 แบบ คือ สีแดง สีเขียว และสีน้ำเงิน ซึ่งเป็นไปตามรูปแบบมาตรฐานการแสดงสีแบบอาร์จีบี (RGB Color Space) แถบสีที่เกิดขึ้นมาจากมาตรฐานการแสดงแบบสีอาร์จีบีเกิดจากการผสมด้วยค่าสีต่างๆ เริ่มตั้งแต่สีแดงเข้มซึ่งเกิดจากค่าสีแดงเพียงอย่างเดียว จนเป็นสีน้ำเงินเนื่องจากมีค่าสีน้ำเงินเพียงอย่างเดียว และเป็นสีดำเมื่อไม่มีค่าสีใดเลย

การกำหนดระดับขั้นของแถบสีอาร์จีบี จึงนำเอารูปแบบแถบสีที่เกิดขึ้นดังกล่าวมาทำการแบ่งช่วงสีให้ได้ 10 ช่วง ตามค่าบรรทัดฐานมากที่สุดและน้อยที่สุดที่ได้กำหนดค่าไว้ตั้งแต่ 0 ถึง 1 ทำให้แถบสีอาร์จีบีซึ่งมีสีแดงอยู่ส่วนล่างสุดของแถบสีถูกกำหนดให้มีค่าเท่ากับ 0 และสีดำซึ่งอยู่ส่วนบนสุดของแถบสีถูกกำหนดให้มีค่าเท่ากับ 1 โดยมีความกว้างของช่วงในแถบสีช่วงละ 0.1 หน่วย ดังรูปที่ 3.13



รูปที่ 3.13 ระดับแถบสีอาร์จีบีเปรียบเทียบกับค่าความถี่

3.3.5 การสร้างภาพบิตแม็บ

การสร้างภาพบิตแม็บจากข้อมูลสายอักขระที่ได้มาจากข้อมูลเอกสารดิจิทัล เป็นการเลือกค่าสีในแถบสีอาร์จีบีที่ตรงกับความถี่ของสายอักขระ มาแทนค่าในแต่ละช่องของเมทริกซ์ซึ่งมีค่าอยู่ระหว่าง 0 จนถึง 1 ที่มาจากการปรับค่าความถี่แบบค่าบรรทัดฐานมากที่สุดและน้อยที่สุด

การเลือกใช้ค่าสีที่จะปรากฏในช่องของตารางเมทริกซ์ เกิดจากหลักการของการเกิดสีในภาพบิตแม็บ ที่ประกอบไปด้วยค่าสี 3 แบบ คือ สีแดง สีเขียว และสีน้ำเงิน ซึ่งสีแต่ละสีจะถูกกำหนดด้วยค่าตั้งแต่ 0 หน่วยจนถึง 255 หน่วย ตามมาตรฐานของสีอาร์จีบี โดยสัดส่วนความ

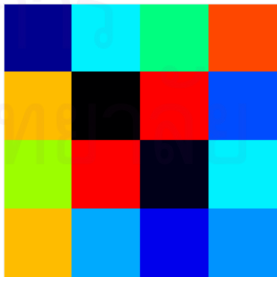
เข้มของสีปรากฏตั้งแต่ไม่มีค่าสีนี้อยู่เลย จนกระทั่งมีความเข้มของค่าสีนี้มากที่สุด สีอื่นๆที่เกิดขึ้นยกเว้นสีแดง สีเขียว และสีน้ำเงิน ซึ่งเกิดมาจากการผสมกันของค่าสี 3 สี ด้วยความเข้มของค่าสีที่แตกต่างกัน การเปรียบเทียบค่าสีอาร์จีบีกับค่าความถี่ของสายอักขระจะเป็นไปตาม ตารางที่ 3.2

ตารางที่ 3.2 การเปรียบเทียบค่าสีอาร์จีบีกับค่าความถี่แบบบรรทัดฐานมากที่สุดและน้อยที่สุด

ค่าความถี่ (n)	**ค่าสีแดง** (R)	**ค่าสีเขียว** (G)	**ค่าสีน้ำเงิน** (B)
0.00 – 0.09	255	0 + [(*n* – 0.00) x 18]	0
0.10 – 0.19	255	127 + [(*n* – 0.10) x 18]	0
0.20 – 0.29	255 - [(*n* – 0.20) x 18]	255	0
0.30 – 0.39	127 - [(*n* – 0.30) x 18]	255	0
0.40 – 0.49	0	255	0 + [(*n* – 0.40) x 18]
0.50 – 0.59	0	255	127 + [(*n* – 0.50) x 18]
0.60 – 0.69	0	255 - [(*n* – 0.60) x 18]	255
0.70 – 0.79	0	127 - [(*n* – 0.70) x 18]	255
0.80 – 0.89	0	0	255 - [(*n* – 0.80) x 18]
0.90 – 0.99	0	0	127 - [(*n* – 0.90) x 18]
1	0	0	0

จากตารางที่ 3.2 แสดงการเปรียบเทียบค่าของความถี่แบบบรรทัดฐานมากที่สุดและน้อยที่สุดกับค่าสีอาร์จีบี โดยเลือกค่าสีอาร์จีบีจากช่วงของค่าความถี่ที่เกิดขึ้น ที่มีความละเอียดระดับทศนิยม 2 ตำแหน่ง แล้วทำการปรับค่าสีที่แตกต่างกันจำนวน 18 หน่วยต่อความถี่ 0.01 หน่วย ทำให้ได้ค่าสีอาร์จีบีที่นำไปแทนที่ค่าความถี่แต่ละช่องในตารางเมทริกซ์ ดังรูปที่ 3.14

0.889	0.611	0.5	0.056
0.148	1	0	0.741
0.278	0	0.981	0.611
0.148	0.667	0.815	0.685



รูปที่ 3.14 ตารางเมทริกซ์ที่กำหนดค่าความถี่ที่ได้มาจากค่าความถี่แบบบรรทัดฐานมากที่สุดและน้อยที่สุด กับรูปภาพบิตแม็บหลังจากทำการแปลงค่าความถี่เทียบกับแถบสีอาร์จีบี

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

- <u>แบคทีเรีย Chlamydophila Pneumoniae สายพันธุ์ TW-183</u>

เป็นแบคทีเรียที่ทำให้เกิดโรคปอดบวมในมนุษย์ โดยแต่ละข้อมูลเป็นดีเอ็นเอของ Chlamydophila pneumoniae ที่ถูกแบ่งออกเป็น 3 ส่วน

- <u>แบคทีเรีย Mycobacterium</u>

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

- <u>แบคทีเรีย Francisella Tularensis</u>

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

- <u>ลิง Macaca Mulatto</u>

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

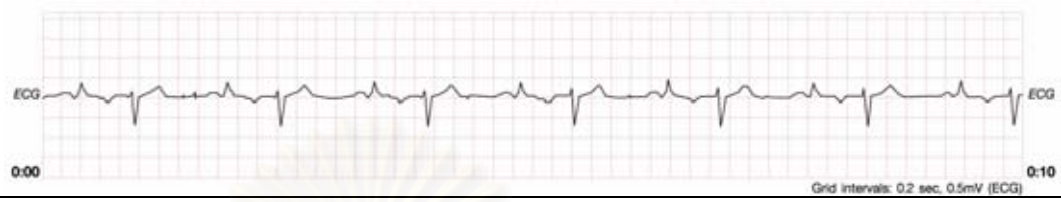
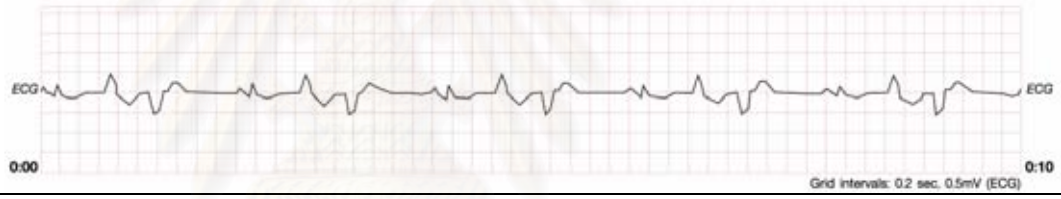
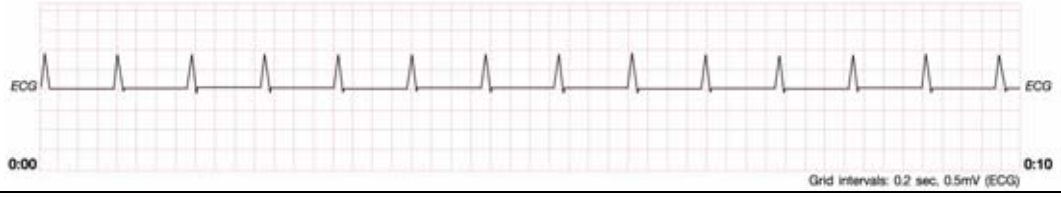
ข้อมูลคลื่นหัวใจที่นำมาใช้ในงานวิจัยนี้ คือข้อมูลของผู้ป่วยที่มีอาการโรคหัวใจในภาวะต่างๆ เช่น ภาวะกล้ามเนื้อหัวใจตีบตัน ภาวะลิ้นหัวใจรั่ว เป็นต้น ซึ่งเป็นข้อมูลที่นำมาจากแหล่งข้อมูลของฟิซิโอเน็ต (PhysioNet - http://www.physionet.org) ซึ่งเป็นองค์กรสนับสนุนการวิจัยที่ไม่แสวงผลกำไร โดยมีรายละเอียดและลักษณะของข้อมูล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

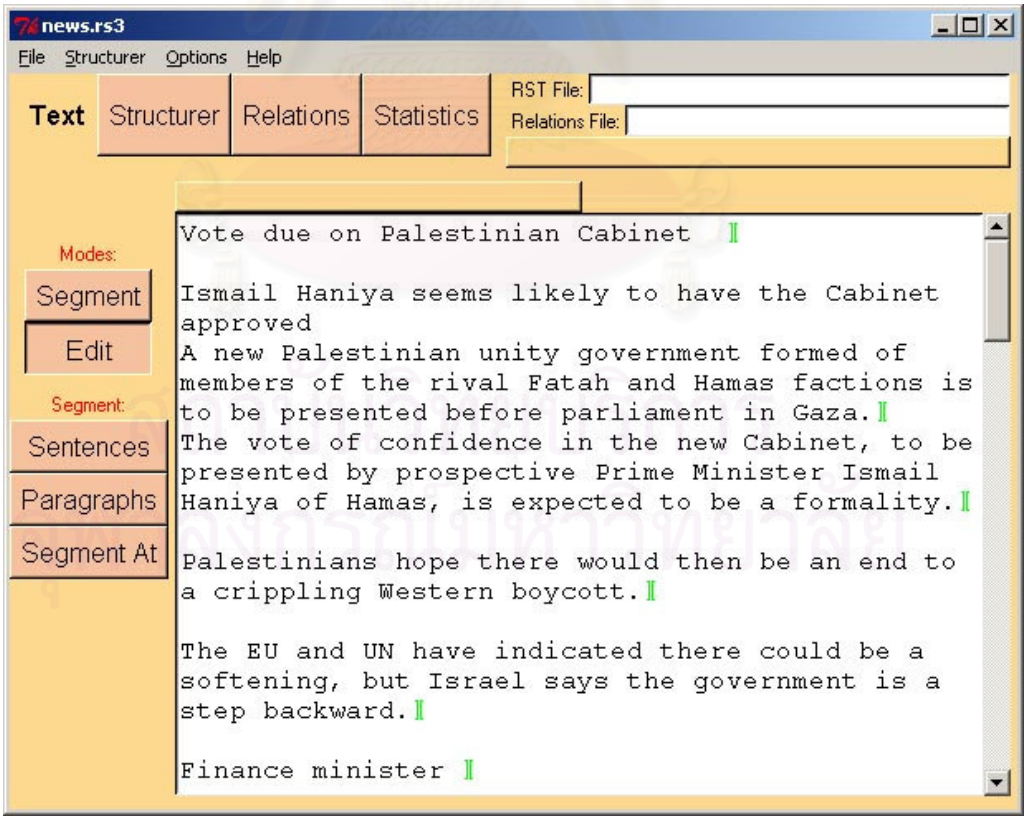
ลำดับที่	รายการเอกสาร	จำนวน (เอกสาร)
1	พระคัมภีร์ไบเบิลส่วนที่ 1 ถึง 4 (ที่มา: http://www.o-bible.com)	4
2	บทละครเรื่องเฟรนด์ (Friends Scripts) ตอนที่ 1 ถึง 10 (ที่มา: http://www.geocities.com/vspramod/links/friends/friends.htm)	5
3	บทละครเรื่องวิลล์และเกรซ (Will & Grace Script) ตอนที่ 1 ถึง 8 (ที่มา: 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

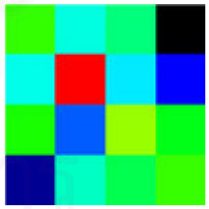
4.3 ผลการทดลอง

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

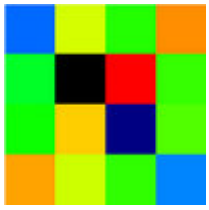


Macaca Mulatta clone CH250-253A7

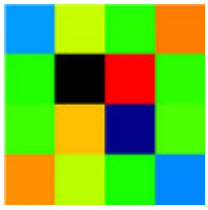
Macaca Mulatta clone CH250-155N20

Macaca Mulatta clone CH250-190E12

รูปที่ 4.2 ผลการทดลองจากข้อมูลดีเอ็นเอ

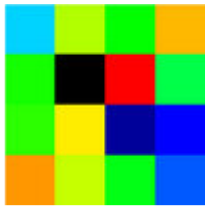


Mycobacterium Bovis AF2122/97



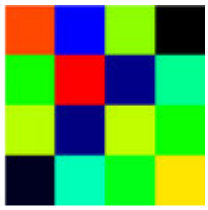
Mycobacterium Tuberculosis H37Rv

Mycobacterium Paratuberculosis K-10

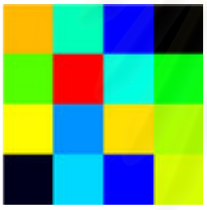


Mycobacterium Tuberculosis CDC1551

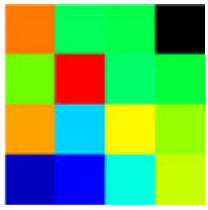
Francisella Tularensis Holarctica strain LVS



Francisella Tularensis Tularensis strain FSC



Chlamydophila Pneumoniae (TW-183 P1/4)



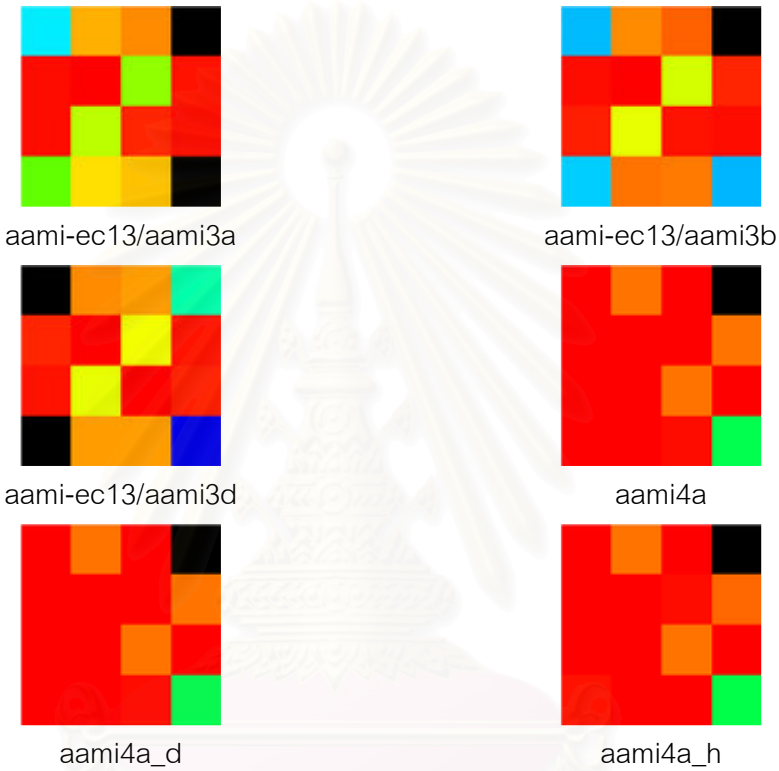
Chlamydophila Pneumoniae (TW-183 P 3/4)

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 แสดงผลการของภาพบิตแม็บจากข้อมูลชุดนี้ ดังรูปที่ 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

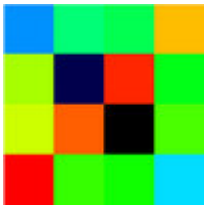
4.3.2.1 ค่าเหมาะสมของค่าสัดส่วนจำนวนเฉลี่ย

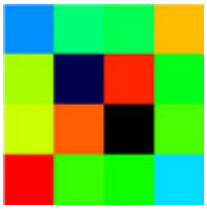
การทดลองเพื่อค่าเหมาะสมของค่าสัดส่วนจำนวนเฉลี่ย ทำการทดลองแสดงผลภาพบิตแม็บของข้อมูลเอกสารจำนวน 3 กลุ่มเอกสาร ประกอบไปด้วยเอกสารทั้งหมดจำนวน 6 เอกสาร โดยแสดงตัวอย่างผลการทดลองการแสดงภาพบิตแม็บจากข้อมูลเอกสารที่ขนาดความยาวของค่าสัดส่วนจำนวนเฉลี่ยเป็น 120 90 60 และ 30 ตามลำดับ ดังรูปที่ 4.4 ถึง 4.7

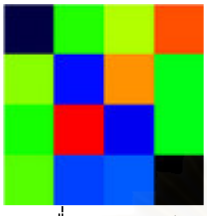
จากการทดลองการแสดงผลภาพบิตแม็บของเอกสาร ที่ค่าสัดส่วนจำนวนเฉลี่ยขนาดต่างๆ พบว่า เมื่อกำหนดค่าสัดส่วนจำนวนเฉลี่ยขนาด 60 ให้ผลภาพบิตแม็บที่สามารถแสดงความเหมือนและความแตกต่างตามกลุ่มเอกสารได้ชัดเจนที่สุด ดังนั้นจึงสรุปได้ว่าพารามิเตอร์ของค่าสัดส่วนจำนวนเฉลี่ยที่เหมาะสม คือ 60

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

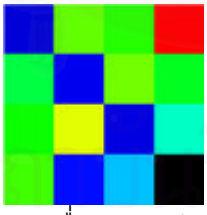
เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 1

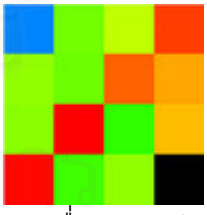
เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 2

รูปที่ 4.4 ผลภาพบิตแม็บเมื่อกำหนดค่าสัดส่วนจำนวนเฉลี่ยขนาด 120

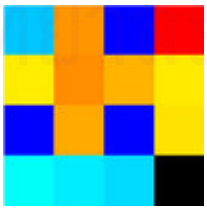
พระคัมภีร์ไบเบิลส่วนที่ 1

พระคัมภีร์ไบเบิลส่วนที่ 3


บทละครเรื่องเฟรนด์ชุดที่ 2

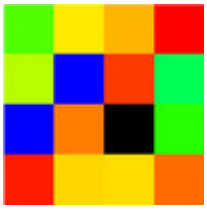

บทละครเรื่องเฟรนด์ชุดที่ 3

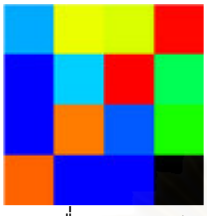
เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 1

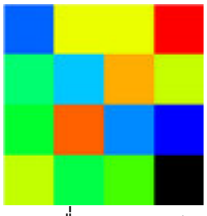

เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 2

รูปที่ 4.5 ผลภาพบิตแม็บเมื่อกำหนดค่าสัดส่วนจำนวนเฉลี่ยขนาด 90

พระคัมภีร์ไบเบิลส่วนที่ 1


พระคัมภีร์ไบเบิลส่วนที่ 3


บทละครเรื่องเฟรนด์ชุดที่ 2


บทละครเรื่องเฟรนด์ชุดที่ 3

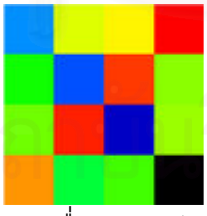
เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 1

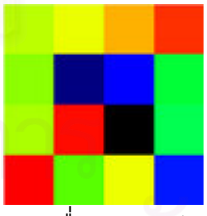
เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 2

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

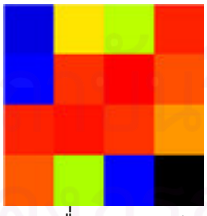
4.3.2.3 **ค่าเหมาะสมของความยาวของเอกสาร**

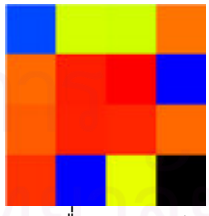
ความยาวของเอกสารที่เหมาะสม สามารถพิจารณาได้จากการนำเอกสารมาทดลองประมวลผลภาพบิตแม็บที่ความยาวของเอกสารต่างๆกัน โดยพิจารณาผลของภาพบิตแม็บรวมถึงพิจารณาเวลาในการประมวลผลของความยาวเอกสารที่แตกต่างกัน

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

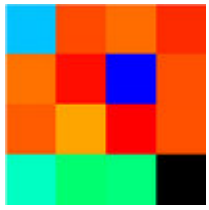
พระคัมภีร์ไบเบิลส่วนที่ 1

พระคัมภีร์ไบเบิลส่วนที่ 3


บทละครเรื่องเฟรนด์ชุดที่ 2

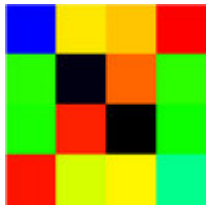

บทละครเรื่องเฟรนด์ชุดที่ 3

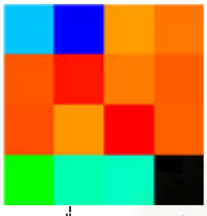
เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 1

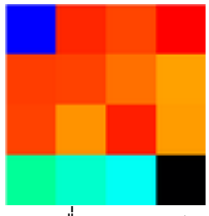

เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 2

รูปที่ 4.8 ผลภาพบิตแม็บเมื่อกำหนดค่าเฉลี่ยเคลื่อนที่ขนาด 60

พระคัมภีร์ไบเบิลส่วนที่ 1


พระคัมภีร์ไบเบิลส่วนที่ 3


บทละครเรื่องเฟรนด์ชุดที่ 2


บทละครเรื่องเฟรนด์ชุดที่ 3

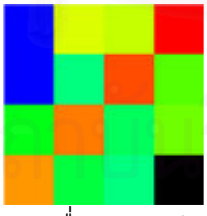
เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 1

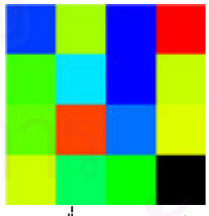
เอกสารวิชาการคอมพิวเตอร์เรื่องที่ 2

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

บทที่ 5
ผลภาพบิตแม็บและการทดสอบการแสดงผลภาพสำหรับข้อมูลเอกสารดิจิทัล

ในบทนี้จะนำเสนอผลภาพบิตแม็บจากข้อมูลเอกสารดิจิทัล ทั้งในข้อมูลเอกสารที่มาจากกลุ่มเอกสารประเภทเดียวกัน และต่างประเภทกัน อีกทั้งแสดงแนวทางการพิจารณาภาพบิตแม็บในเบื้องต้น รวมถึงการทดสอบของการแสดงผลภาพสำหรับข้อมูลเอกสารดิจิทัล โดยการพิจารณาภาพบิตแม็บจากอัตโนมัติ จากการจัดกลุ่มด้วยวิธีเคมีน (*k*-Means Clustering) ซึ่งเป็นการจัดกลุ่มข้อมูลจากระยะห่างของภาพบิตแม็บ ทำให้สามารถแยกแยะความเหมือนและความแตกต่างของภาพบิตแม็บได้ชัดเจนและน่าเชื่อถือ มากกว่าการพิจารณาผลภาพบิตแม็บจากข้อมูลเอกสารด้วยการมองด้วยสายตาเพียงอย่างเดียว

5.1 ผลภาพบิตแม็บจากข้อมูลเอกสารประเภทเดียวกัน

เป็นการแสดงผลภาพบิตแม็บที่มาจากเอกสารประเภทเดียวกัน ซึ่งภาพบิตแม็บที่ออกมาควรเป็นภาพที่มีรูปแบบและลักษณะของการแสดงสีที่ใกล้เคียงกัน สามารถพิจารณาจากลักษณะของภาพได้ค่อนข้างชัดเจน ซึ่งแสดงตัวอย่างผลการแสดงภาพบิตแม็บที่มาจากเอกสารในกลุ่มพระคัมภีร์ไบเบิลที่ทำการตัดแบ่งเป็น 4 ส่วน โดยแต่ละส่วนของพระคัมภีร์ไบเบิลไม่มีบทหรือเนื้อหาในเอกสารที่ซ้ำกันเลย ดังรูปที่ 5.1

พระคัมภีร์ไบเบิลส่วนที่ 1

พระคัมภีร์ไบเบิลส่วนที่ 2

พระคัมภีร์ไบเบิลส่วนที่ 3

พระคัมภีร์ไบเบิลส่วนที่ 4

รูปที่ 5.1 ภาพเอกสารบิตแม็บของเอกสารที่อยู่กลุ่มเดียวกัน

จากรูปที่ 5.1 สังเกตได้ว่าภาพบิตแม็บที่ได้ มีรูปแบบและลักษณะของการแสดงสีที่ใกล้เคียงกันค่อนข้างชัดเจน และมีโครงสร้างของการแสดงสีแต่ละช่องที่เป็นไปในทางเดียวกัน

โดยมีการแสดงสีน้ำเงินเข้มและดำบริเวณตรงกลางภาพ และแสดงสีเขียวและตามด้วยสีส้มแดงบริเวณแนวตั้งด้านขวาของภาพ คล้ายกันทั้ง 4 ภาพ

5.2 ผลภาพบิตแม็บจากข้อมูลเอกสารต่างประเภทกัน

เป็นการแสดงผลภาพบิตแม็บที่มาจากเอกสารที่อยู่ต่างประเภทกัน ซึ่งภาพบิตแม็บที่ออกมาควรเป็นภาพที่มีรูปแบบและลักษณะของการแสดงสีที่ต่างกัน สามารถพิจารณาจากลักษณะของภาพที่แตกต่างกันได้ค่อนข้างชัดเจน การแสดงผลภาพบิตแม็บนี้นำเอาภาพบิตแม็บจากเอกสารทั้ง 4 กลุ่มมาทำการเปรียบเทียบ ซึ่งประกอบไปด้วย

- เอกสารพระคัมภีร์ไบเบิล
- เอกสารบทละครโทรทัศน์
- เอกสารนิยายวิทยาศาสตร์
- เอกสารวิชาการทางคอมพิวเตอร์

แสดงตัวอย่างภาพบิตแม็บของเอกสารแต่ละกลุ่ม ดังรูปที่ 5.2

รูปที่ 5.2 ภาพเอกสารบิตแม็บของเอกสารที่อยู่ต่างกลุ่มกัน

รูปที่ 5.2 แสดงภาพบิตแม็บของเอกสารแต่ละกลุ่มที่งานวิจัยนี้นำมาพิจารณา สังเกตได้ว่าภาพบิตแม็บของเอกสารที่อยู่ในกลุ่มบทละคร พระคัมภีร์ไบเบิล และนิยายวิทยาศาสตร์ มีรูปแบบและลักษณะของการแสดงสีที่แตกต่างกันอย่างชัดเจน โดยเฉพาะเอกสารวิชาการคอมพิวเตอร์ที่มี

พื้นที่ส่วนใหญ่ในภาพเป็นสีแดงส้ม และนิยายวิทยาศาสตร์ที่มีการแสดงสีของสีเขียวเหลืองที่เด่นชัด แต่ถ้าพิจารณาภาพบิตแม็บของเอกสารเฉพาะในกลุ่มบทละคร คือ บทละครเรื่องเฟรนด์ และบทละครเรื่องวิลล์และเกรซ พบว่ามีรูปแบบและลักษณะของการแสดงสีที่คล้ายกันมาก แม้เป็นบทละครที่ต่างเรื่องกัน โดยแสดงสีน้ำเงินตัดกับสีแดงส้มตามแนวทแยงของภาพ และแสดงสีเขียวเหลืองตามขอบของภาพบิตแม็บ

จากผลการแสดงภาพบิตแม็บในหัวข้อ 5.1 และ 5.2 แสดงให้เห็นว่าการแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล สามารถช่วยให้การเปรียบเทียบ แยกแยะ และจำแนกเอกสารโดยการพิจารณาภาพบิตแม็บด้วยสายตาได้เป็นอย่างดี (แสดงผลภาพบิตแม็บของเอกสารทั้งหมดในภาคผนวก ง) อย่างไรก็ตามเพื่อเป็นการเพิ่มความน่าเชื่อถือของผลการทดลองมากยิ่งขึ้นและสามารถวัดผลได้ งานวิจัยนี้จึงทำการทดลองจัดกลุ่มเอกสารจากข้อมูลภาพบิตแม็บด้วยวิธีเคมีน ซึ่งเป็นวิธีการที่สามารถวิเคราะห์ถึงคุณสมบัติจากระยะทางระหว่างภาพบิตแม็บ ที่สามารถบอกถึงความแตกต่างของข้อมูลเอกสารได้

5.3 การพิจารณาภาพบิตแม็บโดยการจัดกลุ่มด้วยวิธีเคมีน

เป็นการพิจารณาเปรียบเทียบความเหมือนและความแตกต่างของภาพบิตแม็บ โดยอาศัยการจัดกลุ่มด้วยวิธีเคมีน ซึ่งเป็นจำแนกประเภทของข้อมูลที่มีความคล้ายกันให้อยู่ในกลุ่มเดียวกันและจัดแยกข้อมูลที่แตกต่างกันให้อยู่คนละกลุ่มกัน โดยการวิเคราะห์จากระยะทางของข้อมูลด้วยการคำนวณระยะทางแบบแมนฮัทตัน (Manhattan Distance) ซึ่งแสดงรายละเอียดการจัดกลุ่มด้วยวิธีเคมีนในหัวข้อ 2.4

งานวิจัยนี้เลือกการจัดกลุ่มโดยวิธีเคมีน เนื่องจากเป็นวิธีการจัดกลุ่มที่มีประสิทธิภาพประมวลผลเร็ว เหมาะสมกับการพิจารณาข้อมูลที่ทราบจำนวนกลุ่มก่อนหน้า และผลจากการจัดกลุ่มภาพบิตแม็บของเอกสารโดยวิธีเคมีน เป็นวิธีที่ใช้ในการวัดผลของการแสดงผลภาพบิตแม็บจากเอกสารดิจิทัลได้เป็นอย่างดี เพราะเป็นวิธีที่วิเคราะห์ถึงคุณสมบัติจากระยะทางระหว่างภาพบิตแม็บ ทำให้สามารถพิจารณาเปรียบเทียบความเหมือนและความแตกต่างของภาพบิตแม็บได้ชัดเจนมากขึ้น

การทดสอบการจัดกลุ่มด้วยวิธีเคมีนของภาพบิตแม็บจากข้อมูลเอกสาร จะใช้เอกสารทุกเอกสารมาทำการทดสอบ ซึ่งมีเอกสารทั้งหมดจำนวน 22 เอกสาร แบ่งออกได้เป็น 4 กลุ่ม โดยใช้ค่าความถี่ของการนับอักขระที่กำกับอยู่ในตารางเมทริกซ์ มาทำการคำนวณค่าระยะทางระหว่างเอกสารด้วยวิธีแบบแมนฮัทตัน (แสดงระยะทางระหว่างคู่เอกสารทั้งหมดใน ภาคผนวก ค) และทำการทดสอบการจัดกลุ่มของภาพบิตแม็บด้วยวิธีเคมีนจำนวน 20 ครั้ง

ผลการทดสอบการจัดกลุ่มในเบื้องต้นพบว่า สามารถทำการจัดกลุ่มได้ถูกต้องเฉลี่ย 78.18 เปอร์เซ็นต์ ซึ่งสรุปได้ว่าการแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล ยังทำการจัดกลุ่มเอกสารผิดพลาดอยู่บ้าง เนื่องจากได้สมาชิกกลุ่มที่ไม่แน่นอนจากการจัดกลุ่มในแต่ละครั้ง

จากการศึกษาเพื่อหาสาเหตุดังกล่าวพบว่า ภาพบิตแม็บของเอกสารมีระยะทางที่แตกต่างกันไม่มาก และจำนวนเอกสารที่นำมาทดลองมีจำนวนน้อย ซึ่งจากการศึกษาเพิ่มเติมพบว่าสาเหตุดังกล่าวเป็นข้อจำกัดของวิธีเคมีน เนื่องจากวิธีเคมีนจะทำการเลือกตัวแทนกลุ่มเริ่มต้นจากข้อมูลทั้งหมดโดยการสุ่ม หากข้อมูลมีความแตกต่างของระยะทางไม่มาก และมีจำนวนข้อมูลน้อย ส่งผลให้การเลือกตัวแทนกลุ่มเริ่มต้นอาจจะอยู่ในกลุ่มข้อมูลเดียวกัน ทำให้ผลของการจัดกลุ่มเกิดความผิดพลาด

เพื่อทำให้การแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล สามารถทำการจัดกลุ่มเอกสารให้มีความถูกต้องมากยิ่งขึ้น งานวิจัยนี้จึงทำการปรับอัลกอริธึมของเคมีนเพิ่มเติม ในขั้นตอนของการเลือกตัวแทนกลุ่มเริ่มต้น โดยทำการตรวจสอบระยะห่างของตัวแทนกลุ่มเริ่มต้นในแต่ละกลุ่ม ให้มีความแตกต่างกันไม่น้อยกว่าค่าระยะห่างที่มากที่สุดของเอกสารในกลุ่มเดียวกัน ทำให้การเลือกตัวแทนกลุ่มเริ่มต้นสามารถทำได้ดีและมีประสิทธิภาพมากยิ่งขึ้น สามารถป้องกันไม่ให้เกิดการเลือกตัวแทนกลุ่มเริ่มต้นที่ไม่ดี นอกจากนี้ยังเพิ่มขั้นตอนการเปรียบเทียบสมาชิกที่ได้จากการจัดกลุ่มที่ต้องได้สมาชิกในกลุ่มเหมือนกันอย่างน้อย 2 ครั้ง

ผลการทดสอบการจัดกลุ่มหลังทำการปรับอัลกอริธึมของเคมีนเพิ่มเติม พบว่าสามารถทำการจัดกลุ่มได้ถูกต้องเฉลี่ย 92.72 เปอร์เซ็นต์ โดยให้ผลการจัดกลุ่มที่ถูกต้องเพิ่มขึ้นจากเดิม 14.54 เปอร์เซ็นต์ ซึ่งสรุปรายละเอียดของผลการทดสอบการจัดกลุ่มจากเอกสารทั้ง 22 เอกสารด้วยอัลกอริธึมของเคมีน โดยแสดงผลการจัดกลุ่มเฉลี่ยจากการจัดกลุ่มจำนวน 20 ครั้ง ดังตารางที่ 5.1

ตารางที่ 5.1 สรุปผลการทดสอบการจัดกลุ่มเอกสารด้วยวิธีเคมีน

ประเภทการจัดกลุ่ม	จำนวนเอกสารที่จัดกลุ่มถูกต้อง (เฉลี่ย)	เปอร์เซ็นต์ความถูกต้อง (เฉลี่ย)
การจัดกลุ่มในเบื้องต้น	17.2	78.18
การจัดกลุ่มหลังทำการปรับอัลกอริธึมของเคมีนเพิ่มเติม	20.4	92.72
เพิ่มขึ้น		14.54

การปรับอัลกอริธึมของวิธีเคมีน นอกจากจะทำให้ผลการจัดกลุ่มข้อมูลเอกสารที่มีจำนวนน้อย และมีระยะห่างระหว่างเอกสารไม่มาก มีความถูกต้องเพิ่มขึ้นแล้ว ยังสามารถนำปรับไปใช้ในการจัดกลุ่มข้อมูลเอกสารในชีวิตประจำวันได้อีกด้วย เนื่องจากจำนวนของเอกสารของแต่ละคนโดยปกติมักมีจำนวนไม่มาก จึงถือเป็นประโยชน์ที่การแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัลสามารถที่จะทำการจัดกลุ่มกับเอกสารเหล่านี้ได้ถูกต้องมากยิ่งขึ้น แสดงอัลกอริธึมของวิธีการการจัดกลุ่มแบบเคมีนที่ทำการปรับเพิ่มเติม ดังรูปที่ 5.3

อัลกอริทึมการจัดกลุ่มด้วยวิธีเคมีนที่ทำการปรับเพิ่ม

<u>ข้อมูลเริ่มต้น</u>: วัตถุที่นำมาจัดกลุ่ม ระยะห่างของตัวแทนกลุ่มเริ่มต้น d และจำนวนกลุ่ม *k*

<u>ผลลัพธ์</u>: ชุดของกลุ่มจำนวน *k* กลุ่ม

<u>วิธีการ</u>:

(1) ทำซ้ำ

(1.1) ทำการเลือกวัตถุจำนวน *k* ตัว เพื่อมาเป็นตัวแทนกลุ่มเริ่มต้น

(1.2) ทำการตรวจสอบระยะห่างของตัวแทนกลุ่มเริ่มต้น

(1.3) เริ่มทำ (1.1) อีกครั้ง จนกระทั่งระยะห่างของตัวแทนกลุ่มเริ่มต้นไม่น้อยกว่า *d*

(2) ทำซ้ำ

(2.1) ทำการเลือกวัตถุในกลุ่ม จากระยะทางของวัตถุกับตัวแทนกลุ่ม

(2.2) เริ่มทำ (2.1) อีกครั้ง จนกระทั่งสมาชิกในกลุ่มไม่มีการเปลี่ยนแปลง

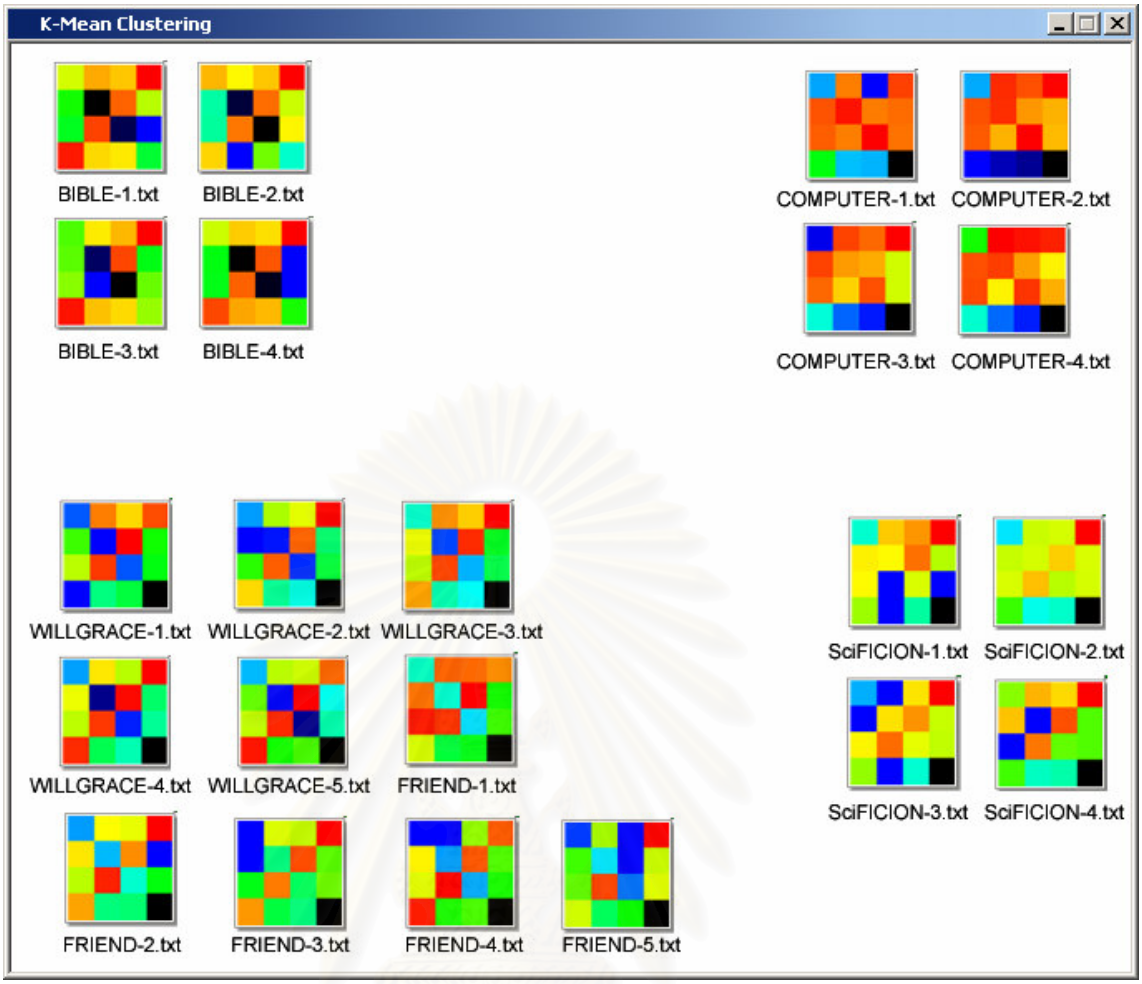
(4) เก็บข้อมูลสมาชิกในกลุ่มทุกกลุ่ม

(5) เริ่มทำ (1) อีกครั้ง จนกระทั่งมีการเก็บข้อมูลสมาชิกในกลุ่มทุกกลุ่มจำนวน 2 ครั้ง

(6) เริ่มทำ (1) อีกครั้ง จนกระทั่งข้อมูลสมาชิกในกลุ่มจากการจัดกลุ่มทั้ง 2 ครั้งไม่แตกต่างกัน

รูปที่ 5.3 อัลกอริธึมของวิธีการการจัดกลุ่มแบบเคมีนที่ทำการปรับเพิ่มเติม

จากผลการทดลองสามารถสรุปได้ว่า การแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล สามารถทำการวิเคราะห์ข้อมูลเอกสารได้ตรงตามคุณลักษณะของเอกสาร ซึ่งสามารถแยกแยะประเภทเอกสารได้โดยการจัดกลุ่มแบบวิธีเคมีน แสดงผลการจัดกลุ่มของเอกสารทั้งหมด ดังรูปที่ 5.4



รูปที่ 5.4 ผลการจัดกลุ่มของภาพเอกสารบิตแม็บด้วยวิธีเคมีน

บทที่ 6
สรุปผลงานวิจัยและข้อเสนอแนะ

6.1 สรุปผลการวิจัย

งานวิจัยนี้เป็นการพัฒนาการแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล โดยการแปลงข้อมูลเอกสารเป็นภาพบิตแม็บ เพื่อช่วยให้ผู้ใช้สามารถพิจารณาความเหมือนและความแตกต่างของชนิดหรือหมวดหมู่ของเอกสารเป็นจำนวนมากในเบื้องต้นได้ ซึ่งผู้ใช้ไม่จำเป็นต้องเข้าไปพิจารณาในเนื้อความของเอกสาร โดยงานวิจัยนี้ได้แนวคิดมาจากทฤษฎีเคออสเกม ที่ถูกนำมาประยุกต์ใช้ในการแสดงผลภาพของข้อมูลดีเอ็นเอ และข้อมูลอนุกรมเวลา เพื่อทำการสรุปรวมข้อมูลที่ซับซ้อนและมีปริมาณมากให้สามารถแสดงผลเป็นภาพ ที่ง่ายแก่การพิจารณาเปรียบเทียบความเหมือนและความแตกต่างข้อมูล

ขั้นตอนการพัฒนาการแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล เริ่มจากการแปลงข้อมูลจากเอกสารดิจิทัลไปเป็นข้อมูลอนุกรมเวลาด้วยมาตรฐานแอสกี ซึ่งต้องมีการวิเคราะห์และปรับข้อมูลในเอกสาร แล้วทำการปรับข้อมูลอนุกรมเวลาที่ได้มา เพื่อลดความแปรปรวนของข้อมูล จากนั้นจะทำการแปลงข้อมูลอนุกรมเวลาไปเป็นอักขระ ด้วยวิธีการแบบแซ็ค (Symbolic Aggregate approXimation) แล้วทำการแปลงข้อมูลอักขระไปเป็นภาพบิตแม็บ ซึ่งภาพบิตแม็บที่ได้จะมีความสัมพันธ์กับลักษณะและเนื้อความในเอกสาร นอกจากนี้ยังได้ทำการทดลองกับข้อมูลเอกสารชุดตัวอย่างเพื่อหาค่าพารามิเตอร์ที่เหมาะสม เพื่อทำให้การแสดงผลภาพบิตแม็บ สามารถแสดงผลภาพบิตแม็บจากข้อมูลเอกสารได้ชัดเจนและมีประสิทธิ ภาพมากยิ่งขึ้น

งานวิจัยนี้ได้ทำการวัดผลการแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล โดยทำการวัดผล ทั้งกับข้อมูลดีเอ็นเอและข้อมูลอนุกรมเวลา เพื่อทดสอบความเป็นไปได้ของแนวทางการวิจัย และทำการทดสอบกับข้อมูลเอกสาร เพื่อทดสอบประสิทธิภาพของการแสดงผลภาพบิตแม็บ ซึ่งสรุปผลการทดสอบได้ดังนี้

6.1.1 ผลการทดสอบกับข้อมูลดีเอ็นเอและข้อมูลอนุกรมเวลา

ภาพบิตแม็บของข้อมูลดีเอ็นเอและข้อมูลอนุกรมเวลา ที่ได้จากการแสดงผลภาพบิตแม็บ สามารถสังเกตความเหมือนและแตกต่างจากลักษณะการแสดงสี และโครงสร้างของภาพบิตแม็บที่เกิดจากข้อมูลตามกลุ่มได้อย่างชัดเจน ทำให้สรุปได้ว่ามีความเป็นได้ที่จะพัฒนาการแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล

6.1.2 ผลการทดสอบกับข้อมูลเอกสารดิจิทัลโดยการจัดกลุ่มด้วยวิธีเคมีน

ผลจากการจัดกลุ่มภาพบิตแม็บจากข้อมูลเอกสารดิจิทัลด้วยวิธีเคมีน สามารถทำการจัดกลุ่มเอกสารจากภาพบิตแม็บได้ถูกต้องเฉลี่ย 78.18 เปอร์เซ็นต์ แต่หลังจากมีการปรับอัลกอริธึมของเคมีนในการเลือกตัวแทนกลุ่มเริ่มต้น โดยทำการตรวจสอบระยะห่างของตัวแทนกลุ่มเริ่มต้นในแต่ละกลุ่ม ให้มีระยะห่างกันไม่น้อยกว่าค่าระยะห่างที่มากที่สุดของเอกสารในกลุ่มเดียวกัน และเพิ่มขั้นตอน การเปรียบเทียบสมาชิกที่ได้จากการจัดกลุ่มที่ต้องได้สมาชิกในกลุ่มเหมือนกันอย่างน้อย 2 ครั้ง ทำให้สามารถทำการจัดกลุ่มได้ถูกต้องเฉลี่ย 92.72 เปอร์เซ็นต์

สรุปผลจากการทดสอบการแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล พบว่าการแสดงผลภาพบิตแม็บนี้สามารถนำไปใช้เพื่อช่วยในการพิจารณา จัดกลุ่ม จำแนก และแยกแยะข้อมูลเอกสารดิจิทัลได้จริงและมีประสิทธิภาพ

6.2 ปัญหาที่พบจากการวิจัย

ปัญหาที่พบจากการทำการวิจัยบางประการ ที่น่าจะเป็นประโยชน์และสามารถนำไปเป็นแนวทางในการแก้ไขปัญหาในงานวิจัยที่ใกล้เคียงอื่นๆต่อไป ได้แก่

1) ปัญหาในการเลือกพิจารณาเนื้อความในเอกสารดิจิทัล ที่เกิดจากข้อมูลในเอกสารมีข้อความบางส่วนที่ไม่มีความสัมพันธ์กับเนื้อความหลักอยู่ โดยเฉพาะช่วงต้นของเอกสาร เช่น คำนำ สารบัญ หรือบทนำ เป็นต้น ซึ่งส่งผลให้การประมวลผลข้อมูลเอกสารของการแสดงผลภาพบิตแม็บเกิดความผิดพลาด ทำให้ได้ภาพบิตแม็บที่มีลักษณะโครงสร้างและการแสดงสีผิดแปลกออกไป ซึ่งการแก้ไขเบื้องต้นสามารถทำได้โดยเข้าไปทำการเลือกตัดข้อมูลในเอกสาร ที่ไม่เกี่ยวข้องกับเนื้อความหลักของเอกสารด้วยมือ แต่ในกรณีทั่วไปสำหรับเอกสารที่มีความยาวมาก สามารถทำการแก้ไขปัญหานี้ได้ โดยการเลือกจุดเริ่มต้นการประมวลผลข้อมูลเอกสารด้วยการสุ่ม โดยคำนวณช่วงการสุ่มจากขนาดความยาวของเอกสาร ที่ต้องการประมวลผลกับขนาดความยาวทั้งหมดของเอกสาร

2) ปัญหาของการจัดกลุ่มภาพบิตแม็บของเอกสารด้วยวิธีเคมีน ที่ให้ผลการจัดกลุ่มเอกสารผิดพลาดอยู่บ้าง เนื่องจากได้สมาชิกกลุ่มที่ไม่แน่นอนจากการจัดกลุ่มในแต่ละครั้ง ซึ่งจากการศึกษาเพื่อหาสาเหตุดังกล่าวพบว่า ภาพบิตแม็บของเอกสารมีระยะทางที่แตกต่างกันไม่มากและจำนวนเอกสารที่นำมาทดลองมีจำนวนน้อย ซึ่งเป็นข้อจำกัดอย่างหนึ่งของการจัดกลุ่มด้วยวิธีเคมีน ดังนั้นเพื่อให้จัดกลุ่มภาพบิตแม็บมีความถูกต้องมากขึ้น จึงทำการปรับอัลกอริธึมของเคมีนเพิ่มเติม โดยขั้นตอนการตรวจสอบระยะห่างของตัวแทนกลุ่มเริ่มต้นในแต่ละกลุ่ม ให้มีความแตกต่างกันไม่น้อยกว่าค่าระยะห่างที่มากที่สุดของเอกสารที่อยู่ในกลุ่มเดียวกัน เพื่อป้องกัน

ไม่ให้เกิดการเลือกตัวแทนกลุ่มเริ่มต้นที่ไม่ดี นอกจากนี้ยังเพิ่มขั้นตอนในการเปรียบเทียบสมาชิกที่ได้จากการจัดกลุ่ม ที่ต้องมีสมาชิกในกลุ่มเหมือนกันจากการทำการจัดกลุ่มอย่างน้อย 2 ครั้ง เพื่อเป็นการตรวจสอบสมาชิกกลุ่มให้มีความถูกต้องมากยิ่งขึ้น

6.3 ข้อเสนอแนะ

1) อาจพัฒนาการแสดงผลภาพบิตแม็บต่อเนื่อง เพื่อให้รองรับการประมวลผลภาพบิตแม็บ จากข้อมูลเอกสารในหลากหลายรูปแบบ เช่น เอกสารเอชทีเอ็มแอล (HTML) เอกสารจากโปรแกรมไมโครซอฟต์เวิร์ด หรือ เอกสารอโครแบต (Acrobat Document)

2) หากสามารถรองรับการประมวลผลภาพบิตแม็บ จากข้อมูลรูปภาพที่ปรากฏอยู่ในเอกสารด้วย น่าจะทำให้ผลของภาพบิตแม็บจากเอกสารมีความชัดเจนมากขึ้น

3) สามารถนำเอาแนวทางการประมวลผลภาพบิตแม็บจากข้อมูลเอกสาร ไปพัฒนาในการประมวลผลภาพบิตแม็บเพื่อแยกแยะลักษณะหรือประเภท กับข้อมูลประเภทอื่นๆ ยกตัวอย่าง เช่น ข้อมูลแบบไบนารี และข้อมูลโปรแกรมที่ได้จากตัวแปรโปรแกรม (Compiler) เป็นต้น

รายการอ้างอิง

[1] Jeffrey, H. J. Chaos game representation of gene structure. Nucleic Acids Research. 18 (1990).

[2] Kumar, N., Lolla, N., Keogh, E., Lonardi, S. and Ratanamahatana, C. Time-series Bitmaps: A Practical Visualization Tool for working with Large Time Series Databases. In Proceedings of SIAM International Conference on Data Mining (SDM '05), Newport Beach, CA, Apr 2005.

[3] Weber, M., Alexa, M. and Mueller, W. Visualizing Time-Series on Spirals, In Proceedings of the IEEE Symposium on Information Visualization, 2001.

[4] Joseph, J. and Sasikumar, R. Chaos game representation for comparison of whole genomes. BMC Bioinformatics, 2006.

[5] Jerding, D. and Stasko, T. The Information Mural: A Technique for Displaying and Navigating Large Information Spaces. In Proceedings of IEEE Symposium on Information Visualization, Atlanta, GA,1995 : 43-50.

[6] Lin, J., Keogh, E., Lonardi, S. and Chiu, B. Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In Proceedings of 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Jun 2003.

[7] Huang, Y. and Yu, P.S. Adaptive Query Processing for Time- Series Data. In Proceedings of 5th Int'l Conference on Knowledge Discovery and Data Mining, San Diego, CA, Aug 1999 : 282-286.

[8] Han, J. and Kamber, M. Data Mining Concept and Techniques. Morgan Kaufmann Publishers, 2001.

[9] Keogh, E., Wei, L., Xiaopeng, X., Lonardi, S., Shieh, J. and Sirowy, S. Intelligent Icons: Integrating Lite-Weight Data Mining and Visualization into GUI Operating Systems. ICDM, 2006.

[10] Lewis, J.P., Rosenholtz, R., Fong, N. and Neumann, U. VisualIDs: automatic distinctive icons for desktop interfaces, In Proceedings of the 2004 SIGGRAPH Conference, ACM Transactions on Graphics (TOG), volume 23, issue 3, 2004 : 416-423.

[11] Larsen, R. J. and Marx, M.L. An Introduction to Mathematical Statistics and Its Applications, Prentice Hall, Englewood, Cliffs, N.J. 2nd Edition, 1986.

[12] Apostolico, A., Bock, M.E. and Lonardi, S. Monotony of Surprise and Large-Scale Quest for Unusual Words, In proceedings of the 6th International Conference on Research in Computational Molecular Biology, Washington, DC, April 18-21, 2001 : 22-31.

[13] Donnell, M.O. RSTTool. (Online). Available from: http://www.wagsoft.com [8 January 2007].

[14] Buja, A., McDonald J.A., Michalak, J and Stützle, W. Interactive Data Visualization Using Focusing and Linking. In Proceedings of Visualization IEEE Computer Society Press, Los Alamitos, 1991 : 156 – 163.

[15] Andre-Jonsson, H. and Badal, D. Using Signature Files for Querying Time-Series Data In Proceedings of Principles of Data Mining and Knowledge Discovery, 1st European Symposium. Trondheim, Norway, Jun 1997 : 211-220.

[16] Barnsley, M.F. and Rising, H. Fractals Everywhere. second edition. Academic Press, 1993.

[17] Yi, K.B. and Faloutsos, C. Fast Time Sequence Indexing for Arbitrary Lp Norms. In Proceedings of 26th International Conference on Very Large Data Bases, Cairo, Egypt., Sep 2000 : 385-394.

ภาคผนวก

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

ภาคผนวก ก
รายการคำที่ไม่มีนัยสำคัญ

A	ABOUT	ABOVE	ACROSS
AFTER	AFTERWARDS	AGAIN	AGAINST
ALL	ALMOST	ALONE	ALONG
ALREADY	ALSO	ALTHOUGH	ALWAYS
AM	AMONG	AMONGST	AMOUNGST
AMOUNT	AN	AND	ANOTHER
ANY	ANYHOW	ANYONE	ANYTHING
ANYWAY	ANYWHERE	ARE	AROUND
AS	AT	BACK	BE
BECAME	BECAUSE	BECOME	BECOMES
BECOMING	BEEN	BEFORE	BEFOREHAND
BEHIND	BEING	BELOW	BESIDE
BESIDES	BETWEEN	BEYOND	BILL
BOTH	BOTTOM	BUT	BY
CALL	CAN	CANNOT	CANT
CO	COMPUTER	CON	COULD
COULDNT	CRY	DE	DESCRIBE
DETAIL	DO	DONE	DOWN
DUE	DURING	EACH	EG
EIGHT	EITHER	ELEVEN	ELSE
ELSEWHERE	EMPTY	ENOUGH	ETC
EVEN	EVER	EVERY	EVERYONE
EVERYTHING	EVERYWHERE	EXCEPT	FEW
FIFTEEN	FIFY	FILL	FIND
FIRE	FIRST	FIVE	FOR
FORMER	FORMERLY	FORTY	FOUND
FOUR	FROM	FRONT	FULL

FURTHER	GET	GIVE	GO
HAD	HAS	HASNT	HAVE
HE	HENCE	HER	HERE
HEREAFTER	HEREBY	HEREIN	HEREUPON
HERS	HERSELF	HIM	HIMSELF
HIS	HOW	HOWEVER	HUNDRED
I	IE	IF	IN
INC	INDEED	INTEREST	INTO
IS	IT	ITS	ITSELF
KEEP	LAST	LATTER	LATTERLY
LEAST	LESS	LTD	MADE
MANY	MAY	ME	MEANWHILE
MIGHT	MILL	MINE	MORE
MOREOVER	MOST	MOSTLY	MOVE
MUCH	MUST	MY	MYSELF
NAME	NAMELY	NEITHER	NEVER
NEVERTHELESS	NEXT	NINE	NO
NOBODY	NONE	NOONE	NOR
NOT	NOTHING	NOW	NOWHERE
OF	OFF	OFTEN	ON
ONCE	ONE	ONLY	ONTO
OR	OTHER	OTHERS	OTHERWISE
OUR	OURS	OURSELVES	OUT
OVER	OWN	PART	PER
PERHAPS	PLEASE	PUT	RATHER
RE	SAME	SEE	SEEM
SEEMED	SEEMING	SEEMS	SERIOUS
SEVERAL	SHE	SHOULD	SHOW
SIDE	SINCE	SINCERE	SIX
SIXTY	SO	SOME	SOMEHOW
SOMEONE	SOMETHING	SOMETIME	SOMETIMES

SOMEWHERE	STILL	SUCH	SYSTEM
TAKE	TEN	THAN	THAT
THE	THEIR	THEM	THEMSELVES
THEN	THENCE	THERE	THEREAFTER
THEREBY	THEREFORE	THEREIN	THEREUPON
THESE	THEY	THICK	THIN
THIRD	THIS	THOSE	THOUGH
THREE	THROUGH	THROUGHOUT	THRU
THUS	TO	TOGETHER	TOO
TOP	TOWARD	TOWARDS	TWELVE
TWENTY	TWO	UN	UNDER
UNTIL	UP	UPON	US
VERY	VIA	WAS	WE
WELL	WERE	WHAT	WHATEVER
WHEN	WHENCE	WHENEVER	WHERE
WHEREAFTER	WHEREAS	WHEREBY	WHEREIN
WHEREUPON	WHEREVER	WHETHER	WHICH
WHILE	WHITHER	WHO	WHOEVER
WHOLE	WHOM	WHOSE	WHY
WILL	WITH	WITHIN	WITHOUT
WOULD	YET	YOU	YOUR
YOURS	YOURSELF	YOURSELVES	

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

ภาคผนวก ข
รายการอักษรพิเศษ

เป็นรายการอักษรพิเศษที่มาจากมาตรฐานแอสกี (ASCII-American Standard Code for Information Interchange) ซึ่งมีทั้งหมด 32 ตัวอักษร

!	)	;	\
"	*	<	]
#	+	=	^
$	,	>	_
%	-	?	{
&	.	@	|
'	/	`	}
(	:	[	~

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

ภาคผนวก ค

ระยะทางระหว่างเอกลักษณ์ของการจัดการเรียนรู้ในชั้นเรียน

B1 - B4 แทนพระคัมภีร์ไบเบิล ตั้งแต่ส่วนที่ 1 ถึงส่วนที่ 4

C1 - C4 แทนของเอกสารวิชาการคอมพิวเตอร์ ตั้งแต่เรื่องที่ 1 ถึงเรื่องที่ 4

S1 - S5 แทนของนิยายวิทยาศาสตร์ ตั้งแต่เรื่องที่ 1 ถึงเรื่องที่ 4

W1 - W5 แทนบทละครเรื่องวิลล์และเกรซ ตั้งแต่ชุดที่ 1 ถึงชุดที่ 5

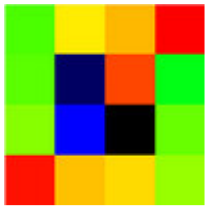
F1 - F4 แทนของบทละครเรื่องเฟรนด์ ตั้งแต่ชุดที่ 1 ถึงชุดที่ 5

	B1	B2	B3	B4	C1	C2	C3	C4	S1	S2	S3	S4	W1	W2	W3	W4	W5	F1	F2	F3	F4	F5
B1	0																					
B2	1.21	0																				
B3	0.95	1.76	0																			
B4	0.38	1.28	0.83	0																		
C1	5.53	5.23	5.61	5.76	0																	
C2	6.24	5.94	6.31	6.47	1.08	0																
C3	5.68	5.47	5.75	5.91	1.21	0.96	0															
C4	5.65	5.35	5.72	5.88	1.20	0.96	0.90	0														
S1	3.98	3.68	4.00	4.15	1.61	2.32	1.82	2.08	0													
S2	4.32	3.98	4.31	4.46	1.81	2.40	1.81	2.25	0.78	0												
S3	4.25	3.97	4.29	4.39	1.50	2.18	1.62	2.14	0.52	0.54	0											
S4	3.75	3.70	3.72	3.81	2.09	2.80	2.25	2.22	1.04	1.28	1.21	0										
W1	2.97	3.02	2.88	3.11	3.38	4.19	3.51	4.12	2.21	2.41	2.26	2.23	0									
W2	3.02	2.90	2.87	3.06	3.67	4.39	3.82	4.37	2.42	2.26	2.26	2.45	1.02	0								
W3	2.94	3.17	2.65	3.06	3.92	4.65	4.07	4.62	2.56	2.63	2.52	2.60	1.15	0.85	0							
W4	2.80	3.02	2.50	2.94	4.53	5.33	4.78	5.24	3.21	3.12	3.17	3.26	1.34	1.23	0.94	0						
W5	3.05	3.17	2.85	3.20	3.22	3.94	3.38	3.70	1.93	2.18	2.05	1.87	0.89	0.90	0.92	1.59	0					
F1	3.18	3.17	3.15	3.27	3.01	3.70	3.12	3.67	1.65	1.74	1.59	1.80	0.98	1.10	1.13	1.68	0.90	0				
F2	3.15	3.05	3.32	3.14	3.26	3.98	3.14	3.96	1.88	1.82	1.84	1.99	1.48	1.45	1.84	2.15	1.74	1.13	0			
F3	2.86	2.58	3.06	3.04	3.22	3.93	3.22	3.91	1.88	1.81	1.82	2.37	1.03	1.22	1.75	1.82	1.64	1.09	1.15	0		
F4	2.99	3.21	2.87	3.13	3.70	4.51	3.89	4.44	2.41	2.44	2.38	2.47	1.25	1.48	1.30	1.17	1.42	0.93	1.58	1.30	0	
F5	3.58	3.63	3.41	3.74	2.83	3.54	3.01	3.22	1.92	2.53	2.31	1.77	1.27	2.08	1.95	2.16	1.37	1.46	2.03	1.79	1.62	0

ภาคผนวก ง
ภาพบิตแม็บจากข้อมูลเอกสารที่นำมาทดลอง

พระคัมภีร์ไบเบิลส่วนที่ 1

พระคัมภีร์ไบเบิลส่วนที่ 2


พระคัมภีร์ไบเบิลส่วนที่ 3

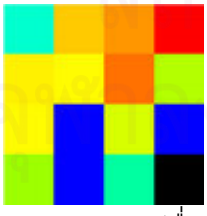
พระคัมภีร์ไบเบิลส่วนที่ 4

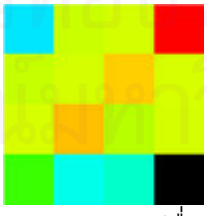
เอกสารวิชาการคอมพิวเตอร์
เรื่องที่ 1

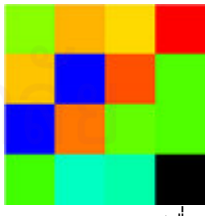
เอกสารวิชาการคอมพิวเตอร์
เรื่องที่ 2

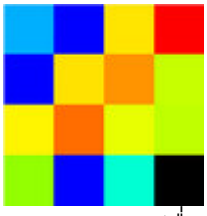
เอกสารวิชาการคอมพิวเตอร์
เรื่องที่ 3

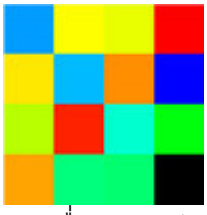
เอกสารวิชาการคอมพิวเตอร์
เรื่องที่ 4


นิยายวิทยาศาสตร์เรื่องที่ 1

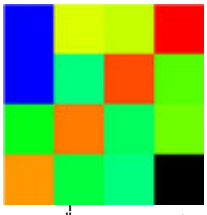

นิยายวิทยาศาสตร์เรื่องที่ 2


นิยายวิทยาศาสตร์เรื่องที่ 3

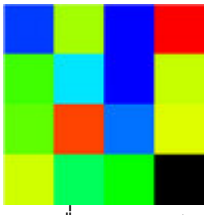

นิยายวิทยาศาสตร์เรื่องที่ 4



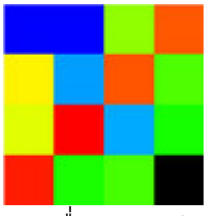
บทละครเรื่องเฟรนด์ชุดที่ 1



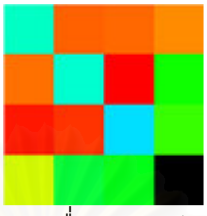
บทละครเรื่องเฟรนด์ชุดที่ 2



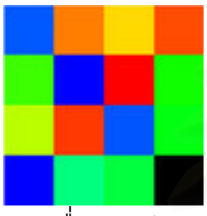
บทละครเรื่องเฟรนด์ชุดที่ 3



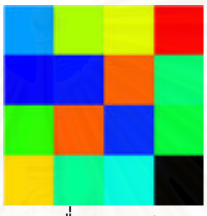
บทละครเรื่องเฟรนด์ชุดที่ 4



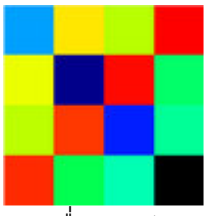
บทละครเรื่องเฟรนด์ชุดที่ 5



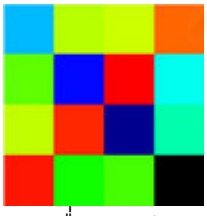
บทละครเรื่องวิลล์และเกรซ ชุดที่ 1



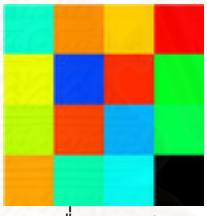
บทละครเรื่องวิลล์และเกรซ ชุดที่ 2



บทละครเรื่องวิลล์และเกรซ ชุดที่ 3



บทละครเรื่องวิลล์และเกรซ ชุดที่ 4



บทละครเรื่องวิลล์และเกรซ ชุดที่ 5

ประวัติผู้เขียนวิทยานิพนธ์

นายสัณห์ชัย นักรบ เกิดเมื่อวันที่ 3 พฤษภาคม พ.ศ. 2522 ที่จังหวัดปราจีนบุรี สำเร็จการศึกษาระดับปริญญาตรี วิทยาศาสตรบัณฑิต จากภาควิชาคณิตศาสตร์ สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2543 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2548 ปัจจุบันทำงานอยู่บริษัทหลักทรัพย์ เคจีไอ (ประเทศไทย) จำกัด (มหาชน)

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย