บทที่ 3
การออกแบบวิธีการแสดงผลภาพสำหรับข้อมูลเอกสารดิจิทัล

งานวิจัยนี้เป็นการพัฒนาวิธีแสดงผลภาพบิตแม็บสำหรับข้อมูลเอกสารดิจิทัล ซึ่งอาศัยแนวคิดจากทฤษฎีเคออสเกม (Chaos Game) [1] ประยุกต์ร่วมกับ การแสดงผลภาพบิตแม็บของข้อมูลอนุกรมเวลา โดยใช้วิธีแบบแซ็ค (SAX Symbolic Representation) ในการกำหนดอักขระที่เกิดขึ้น [2][6] ซึ่งวิธีดังกล่าวจะนำเอาความถี่ของสายอักขระที่เป็นข้อมูลอนุกรมเวลา รวมถึงข้อมูลจำพวกสายอักขระของดีเอ็นเอ 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 ที่อาจเกิดขึ้นกับเอกสาร ซึ่งมีผลทำให้การแสดงผลภาพของเอกสารออกมาคลาดเคลื่อนได้ 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)

$$SimpleMovingAverage = \frac{P_m + P_{m+1} + P_{m+2} + ... + P_{m+n}}{n} \quad (3.1)$$

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)

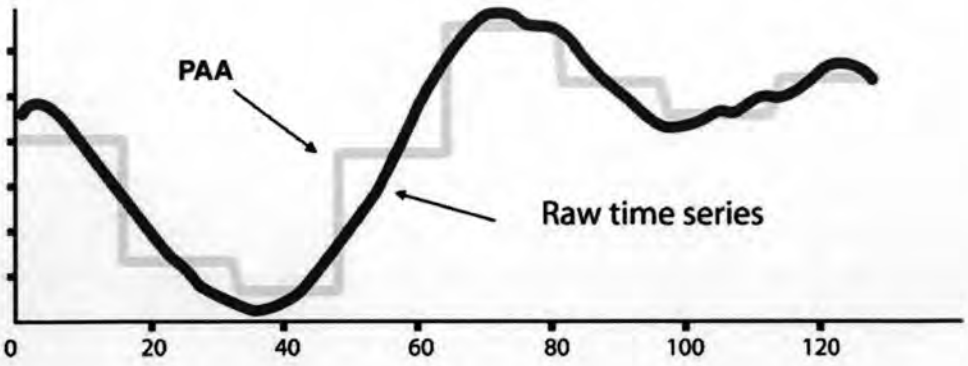
อย่างไรก็ตามการคำนวณค่าเฉลี่ยเคลื่อนที่แบบพื้นฐาน มีข้อจำกัดอยู่ที่ไม่สามารถคำนวนหาค่าเฉลี่ยของชุดข้อมูล n - 1 ตัวสุดท้ายได้ ทำให้ไม่สามารถทำการปรับเรียบข้อมูลช่วง n-1 สุดท้ายได้

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

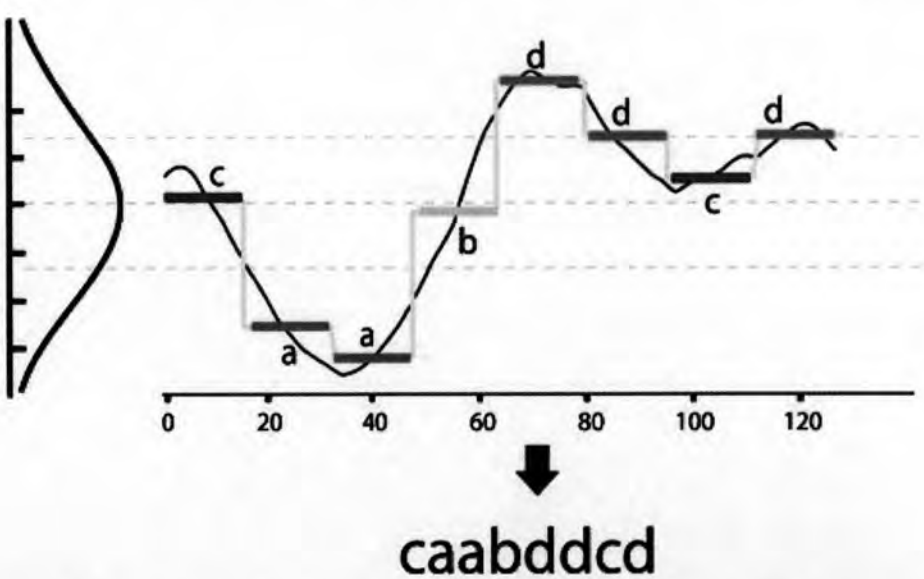
เนื่องจากข้อมูลอนุกรมเวลาที่ได้มาจากการแปลงข้อมูลเอกสาร ที่เป็นข้อมูลเลขจำนวนจริงและถูกทำข้อมูลให้เป็นบรรทัดฐาน ในหัวข้อ 3.2.2 ซึ่งจึงจัดว่าเป็นข้อมูลที่อยู่ในรูปแบบการกระจายแบบเกาส์เซียน ส่งผลให้ข้อมูลดังกล่าวมีลักษณะเป็นข้อมูลแบบไม่ต่อเนื่อง (Discrete Representation) ทำให้สามารถพิจารณาการแปลงข้อมูลเป็นอักขระโดยวิธีการของข้อมูลแบบไม่ต่อเนื่องได้ [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 มีผลให้จำเป็นต้องใช้ความยาวของข้อมูลเอกสารที่เพิ่มขึ้น เพื่อให้เพียงพอต่อความถี่ของแต่ละอักขระที่มีเพิ่มมากขึ้น 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 มาแทนค่าในแต่ละช่องของเมทริกซ์ ซึ่งมีค่าอยู่ระหว่าง 0 จนถึง 1 ที่มาจากการปรับค่าความถี่แบบค่าบรรทัดฐานมากที่สุดและน้อยที่สุด

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

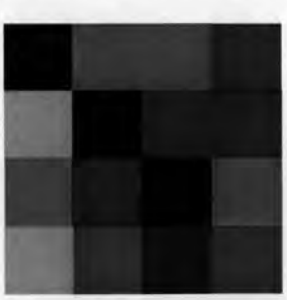
เข้มของสีปรากฏตั้งแต่ไม่มีค่าสีนี้อยู่เลย จนกระทั้งมีความเข้มของค่าสีนี้มากที่สุด สีอื่นๆที่เกิดขึ้นยกเว้นสีแดง สีเขียว และสีน้ำเงิน ซึ่งเกิดมาจากการผสมกันของค่าสี 3 สี ด้วยความเข้มของค่าสีที่แตกต่างกัน การเปรียบเทียบค่าสีอาร์จีบีกับค่าความถี่ของสายอักขระจะเป็นไปตาม ตารางที่ 3.2

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