

การแสดงผลภาพปิดแม่พิมพ์สำหรับข้อมูลเอกสารดิจิทัล



นายสันหทัย 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

**491740**

หัวข้อวิทยานิพนธ์

การแสดงผลภาพปิดแม่แบบสำหรับข้อมูลเอกสารดิจิทัล

โดย

นายสันหทัย นักรบ

สาขาวิชา


วิทยาศาสตร์คอมพิวเตอร์

อาจารย์ที่ปรึกษา

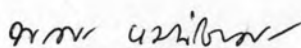
อาจารย์ ดร. โชติรัตน์ รัตนามัทธนะ


---


คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้หัวข้อวิทยานิพนธ์ฉบับนี้  
เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาโท

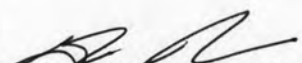
  
..... คณบดีคณะวิศวกรรมศาสตร์  
(ศาสตราจารย์ ดร.ดิเรก ลาวัณย์ศิริ)

คณะกรรมการสอบวิทยานิพนธ์

  
..... ประธานกรรมการ  
(รองศาสตราจารย์ ดร.พรศิริ หมั่นไชยศิริ)

  
..... อาจารย์ที่ปรึกษา  
(อาจารย์ ดร.โชติรัตน์ รัตนามัทธนะ)

  
..... กรรมการ  
(อาจารย์ ดร.อดิวงค์ สุชาโต)

  
..... กรรมการ  
(อาจารย์ ธงชัย โรจน์กั้งสดาล)

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

วิทยานิพนธ์นี้มีวัตถุประสงค์ในการแสดงผลภาพสำหรับข้อมูลเอกสารดิจิทัล โดยทำการแปลงข้อมูลในเอกสารจากตัวอักษรให้เป็นรูปภาพ เพื่อช่วยในการพิจารณาเปรียบเทียบความเหมือนและความแตกต่างของประเภทหรือหมวดหมู่เอกสาร ทำให้ผู้ใช้สามารถ จัดการ และจำแนกรูปแบบหรือประเภทของเอกสารได้ง่ายและรวดเร็วมากยิ่งขึ้น โดยไม่จำเป็นต้องเข้าไปพิจารณาเนื้อความในเอกสาร โดยการแสดงผลภาพมีแนวทางในการพัฒนาจากแนวคิดของทฤษฎีเคออสเกม ประยุกต์ร่วมกับการแสดงผลภาพบิตแม็บของข้อมูลอนุกรมเวลาโดยใช้วิธีการแบบแซ็ค

งานวิจัยนี้ได้ทำการวิเคราะห์รูปแบบและลักษณะต่างๆของเอกสาร โดยการปรับข้อมูลในเอกสาร และกำหนดพารามิเตอร์ที่สำคัญต่างๆ เพื่อให้การแสดงผลภาพบิตแม็บจากข้อมูลในเอกสารมีความชัดเจนและมีประสิทธิภาพ ซึ่งได้มาจากการทดลองด้วยข้อมูลจริง นอกจากนี้ยังได้ทำการทดสอบประสิทธิภาพของการแสดงผลภาพจากการพิจารณาเปรียบเทียบภาพบิตแม็บของข้อมูลเอกสาร ทั้งจากการสังเกตและใช้วิธีการจัดกลุ่มภาพบิตแม็บโดยอัตโนมัติ ซึ่งได้ผลสรุปจากการทดสอบว่า การแสดงผลภาพสำหรับข้อมูลเอกสารดิจิทัล สามารถช่วยในการพิจารณาเปรียบเทียบความเหมือนและความแตกต่างของประเภทหรือหมวดหมู่เอกสารดิจิทัลได้อย่างมีประสิทธิภาพ

ภาควิชา.....วิศวกรรมคอมพิวเตอร์..... ลายมือชื่อนิสิต..... *สันหทัย นักรบ*  
 สาขาวิชา.....วิทยาศาสตร์คอมพิวเตอร์..... ลายมือชื่ออาจารย์ที่ปรึกษา..... *Yltt Rm*  
 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.

Department..... Computer Engineering ..... Student's signature..... *สุนชัย นาคโรบ* .....  
 Field of study..... Computer Science ..... Advisor's signature..... *ชอติราต รตนamahatana* .....  
 Academic year ..... 2006 .....

## กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยความอนุเคราะห์ และความช่วยเหลืออย่างยิ่ง จาก อ.ดร.โชติรัตน์ รัตนามัทธนะ อาจารย์ที่ปรึกษา ซึ่งให้ข้อคิด แนวทาง และคำปรึกษา ตลอดจนเป็นผู้ตรวจทานแก้ไข ทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วง ขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณ รศ.ดร.พรศิริ หมั่นไชยศรี อ.ดร.อดิวงค์ สุขาโต และ อ.ธงชัย ไรจน์กั้งสดาล ประธานกรรมการและกรรมการสอบวิทยานิพนธ์ ที่กรุณาให้คำแนะนำในการแก้ไข วิทยานิพนธ์ให้มีคุณภาพยิ่งขึ้น ขอขอบพระคุณคณาจารย์ในภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัยทุกท่านที่ประสิทธิ์ประสาทความรู้อันมีค่ายิ่งแก่ผู้วิจัย

ที่สำคัญที่สุดขอขอบคุณ คุณพ่อ คุณแม่ และเพื่อนๆ ที่เป็นแรงผลักดัน เป็น กำลังใจ ที่สำคัญให้ตลอดการศึกษาค้นคว้าครั้งนี้

## 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: เครื่องมือการแสดงผลภาพสำหรับการ ทำงานกับฐานข้อมูลของข้อมูลอนุกรมเวลาขนาดใหญ่.....	11
2.5.2 สัญรูปอัจฉริยะ: การทำเหมืองข้อมูลขนาดย่อม และทำการแสดงผลภาพสู่ ระบบปฏิบัติการแบบส่วนต่อประสานด้วยภาพกับผู้ใช้.....	11
3 การออกแบบวิธีการแสดงผลภาพสำหรับข้อมูลเอกสารดิจิทัล .....	13
3.1 การแปลงข้อมูลจากเอกสารดิจิทัลไปเป็นข้อมูลอนุกรมเวลา .....	13



บทที่	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 การทดลองเพื่อสนับสนุนแนวทางการวิจัย .....	35
4.3.2 การทดลองเพื่อหาพารามิเตอร์และข้อมูลที่เหมาะสม.....	38



บทที่	หน้า
5	ผลภาพบิตแม็บและการวัดผลการแสดงผลภาพสำหรับข้อมูลเอกสารดิจิทัล ..... 44
5.1	ผลภาพบิตแม็บจากข้อมูลเอกสารประเภทเดียวกัน ..... 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 ระยะห่างระหว่างเอกสารจากการคำนวณแบบแมนฮัตตัน ..... 59
	ภาคผนวก ง ภาพบิตแม็บจากข้อมูลเอกสารที่นำมาทดลอง ..... 60
	ประวัติผู้เขียนวิทยานิพนธ์ ..... 62

## สารบัญตาราง

	หน้า
ตารางที่ 3.1 ตัวอย่างการเปรียบเทียบระหว่างเลขจำนวนจริงและตัวอักษรจากมาตรฐาน แอลกี .....	16
ตารางที่ 3.2 การเปรียบเทียบค่าสีอาร์จีบีกับค่าความถี่แบบบรทัดฐานมากที่สุดและ น้อยที่สุด .....	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

## 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	ตัวอย่างการกำจัดอักขระพิเศษ (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	การแปลงข้อมูลอนุกรมเวลาเป็นอักขระโดยกำหนดจุดชั้นจำนวน 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 และค่าความ ถี่ที่ได้มาจากการหาค่าสัดส่วนแบบปรับค่าบรรทัดฐานมากที่สุดและน้อยที่สุด .....	23
รูปที่ 3.13	ระดับแถบสีอาร์จีบีเปรียบเทียบกับค่าความถี่.....	24
รูปที่ 3.14	ตารางเมทริกซ์ที่กำหนดค่าความถี่ที่ได้มาจากค่าความถี่แบบบรรทัดฐานมาก ที่สุดและน้อยที่สุด กับรูปภาพบิตแม็บหลังจากทำการแปลงค่าความถี่เทียบกับ แถบสีอาร์จีบี .....	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 อัลกอริทึมของวิธีการการจัดกลุ่มแบบเคมีนที่ทำการปรับเพิ่มเติม .....	48
รูปที่ 5.4 ผลการจัดกลุ่มของภาพเอกสารบิตแม็บด้วยวิธีเคมีน .....	49