NECTEC
Technical Journal
ISSN 1513-2145

วารสารวิชาการเนคเทค ปีที่ 3 ฉบับที่ 11 เดือน กรกฎาคม – ตุลาคม 2544 Vol. III, No. 11, July – October 2001

Short Paper

การบีบอัดภาพของกล้องจุลทรรศน์อิเล็กตรอน โดยใช้การแปลงเวฟเล็ตแบบดีสครีท

EM Image Compression Using the Discrete Wavelet Transform

ณัฐนันท์ ทัดพิทักษ์กุล[1] กิตติ อัตถกิจมงคล[2] สราวุฒิ สุจิตจร[3]
[1]นักศึกษาบัณฑิตศึกษา [2]อาจารย์ [3]รองศาสตราจารย์
สาขาวิชาวิศวกรรมไฟฟ้า สำนักวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

ABSTRACT – The image from the Electron Microscope (EM) is effectively used to analyze the fine details of the object's surface. When the discrete wavelet transform is applied to the image, the plain surface of the object will be in the low-frequency subband and the edge will be in the high-frequency subbands. Thus, compression algorithm for the EM image must take every subband of the wavelet coefficients into account. A powerful image compression algorithm we consider is the Set Partitioning in Hierarchical Tree (SPIHT). This coding scheme exploits the self-similarity of the wavelet coefficients across different scales and searches for the high magnitude coefficients in every subband. In this paper, we propose an improvement of the SPIHT algorithm. Several wavelets are then applied for comparison in order to find the best wavelet basis for the EM image with the improved algorithm.

KEY WORD – Image compression, SPIHT, Wavelet

บทคัดย่อ – ภาพของกล้องจุลทรรศน์อิเล็กตรอน (Electron Microscope, EM) เป็นภาพที่ใช้วิเคราะห์รายละเอียดของโครงสร้างระดับ ไมโครบนพื้นวัตถุ ทำให้ภาพ EM เป็นภาพที่มีรายละเอียดมาก เมื่อนำมาแปลงด้วยวิธีเวฟเล็ตแบบดีสครีท ข้อมูลที่เป็นพื้นผิววัตถุจะอยู่ที่สัมประสิทธิ์ที่แบนด์ย่อยความถี่ต่ำ และข้อมูลที่เป็นขอบหรือลายเส้นของวัตถุจะอยู่ที่สัมประสิทธิ์ที่แบนด์ย่อยความถี่สูง ดังนั้นการบีบอัดข้อมูลภาพ EM ด้วยการแปลงเวฟเล็ตจะต้องให้ความสำคัญกับสัมประสิทธิ์ทุกแบนด์ย่อย อัลกอริทึม Set Partitioning in Hierarchical Tree (SPIHT) เป็นอัลกอริทึมหนึ่งที่เหมาะสม เนื่องจากอัลกอริทึมนี้เข้ารหัสโดยให้ความสำคัญกับขนาดของสัมประสิทธิ์และไม่สนใจว่าสัมประสิทธิ์ตัวนั้นจะอยู่ในระดับแบนด์ย่อยใด ในบทความนี้เสนอวิธีการพัฒนาอัลกอริทึม SPIHT ให้สามารถบีบอัดข้อมูลได้เพิ่มขึ้น พร้อมทั้งหาเวฟเล็ตแม่ที่เหมาะสมกับอัลกอริทึมที่ทำการพัฒนา

คำสำคัญ – การบีบอัดข้อมูลภาพ, SPIHT, เวฟเล็ต

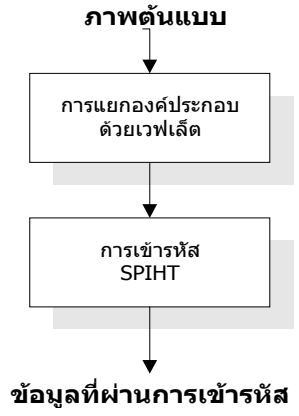
1. คำนำ

ปัจจุบันการศึกษาค้นคว้าสิ่งเร้นลับทางชีวภาพและกายภาพขนาดจิ๋ว ซึ่งไม่สามารถมองเห็นได้ด้วยตาเปล่าได้ เช่น การวิเคราะห์รายละเอียดของโครงสร้างระดับไมโครบนพื้นผิววัสดุ ไม่ว่าจะเป็นการกระจายของอนุภาค ความพรุน ความหนาแน่น หรือร่องรอยอันเกิดจากความผิดปกติของผิววัสดุ เป็นสิ่งจำเป็นอย่างยิ่งในการควบคุมคุณภาพของงานด้านโลหกรรมและวัสดุศาสตร์ และการผลิตประดิษฐกรรมทางไมโครอิเล็กทรอนิกส์ สิ่งหนึ่งที่ใช้ในการวิเคราะห์คือภาพไมโครกราฟที่ได้จากกล้อง EM จะมีอยู่ด้วยกัน 2 ประเภทคือ ภาพที่ได้จากกล้องโพลารอยด์ (Polaroid) ภาพที่ได้นั้นมักไม่สามารถแยกความเปรียบต่าง (contrast) ของภาพ รวมถึงความแตกต่างของรูปลักษณะตามบริเวณต่างๆ ได้ชัดเจน ทำให้เกิดความคลาดเคลื่อนในการวิเคราะห์สูง [12] และภาพที่ได้จากการเปลี่ยนสัญญาณภาพจากระบบสแกนความเร็วต่ำของ EM ที่อยู่ในรูปสัญญาณอนาล็อก จะถูกแปลงเป็นสัญญาณเชิงตัวเลข และบันทึกเป็นไฟล์ข้อมูล ซึ่งภาพจากไฟล์ข้อมูลยังสามารถนำไปผ่านกระบวนการปรับแต่งรายละเอียดภาพด้วยวิธีการกรองสิ่งรบกวน และผ่านกระบวนการชดเชยให้มีความคมชัด และความเปรียบต่างของภาพดีขึ้น ทำให้มีความเชื่อมั่นในผลการวิเคราะห์ได้สูงขึ้น [12]

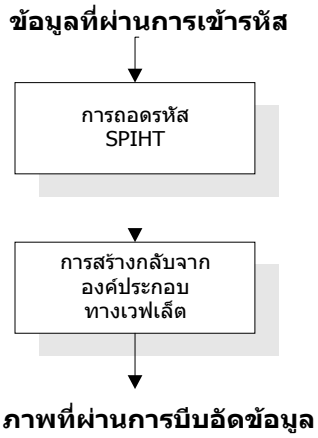
ภาพของ EM เป็นภาพที่มีรายละเอียดภาพสูง เมื่อนำมาแปลงเวฟเล็ตแบบดีสครีท ข้อมูลที่เป็นพื้นผิววัตถุจะอยู่ที่สัมประสิทธิ์ที่แบนด์ย่อย

ความถี่ต่ำ และข้อมูลที่เป็นขอบหรือลายเส้นของวัตถุจะอยู่ที่สัมประสิทธิ์ที่แบนด์ย่อยความถี่สูง ดังนั้นการบีบอัดข้อมูลภาพ EM ด้วยการแปลงเวฟเล็ตจะต้องให้ความสำคัญกับสัมประสิทธิ์ทุกแบนด์ย่อย อัลกอริทึม Embedded Zerotrees Wavelet (EZW) เป็นอัลกอริทึมหนึ่งที่เหมาะสม เนื่องจากอัลกอริทึมนี้เข้ารหัสโดยให้ความสำคัญกับขนาดของสัมประสิทธิ์และไม่สนใจว่าสัมประสิทธิ์ตัวนั้นจะอยู่ในระดับแบนด์ย่อยใด [8,9] และให้อัตราการบีบอัดข้อมูลดีกว่าการบีบอัดข้อมูลภาพด้วยวิธี Joint Photographic Group (JPEG) [9] และ EZW ยังไม่ทำให้เกิดบล็อกอาติแฟคต์ (Block Artifacts) เหมือน JPEG [10,11] ได้มีการพัฒนา EZW ให้มีประสิทธิภาพเพิ่มมากขึ้นหลายวิธี แต่ในบทความนี้เลือกใช้อัลกอริทึม Set Partitioning in Hierarchical Tree (SPIHT) ซึ่งให้อัตราการบีบอัดข้อมูลภาพที่ดีกว่า EZW [2,8] และเป็นอัลกอริทึมที่เร็วและมีประสิทธิภาพ [3] นอกจากนี้ในบทความนี้ยังทำการพัฒนาอัลกอริทึม SPIHT ให้มีอัตราการบีบอัดข้อมูลภาพที่ดีขึ้นด้วยการเพิ่ม List of Forbidden Coefficients (LFC) และเพิ่มเงื่อนไขการเข้ารหัสและการถอดรหัส พร้อมทั้งหาเวฟเล็ตแม่ที่เหมาะสมกับอัลกอริทึม SPIHT ที่ทำการพัฒนา

สำหรับการบีบอัดข้อมูลภาพในบทความนี้ ประกอบด้วยขั้นตอนการเข้ารหัส (Encode) และขั้นตอนการถอดรหัส (Decode) ซึ่งแต่ละขั้นตอนได้แสดงไว้ในรูปที่ 1 และรูปที่ 2 ตามลำดับ



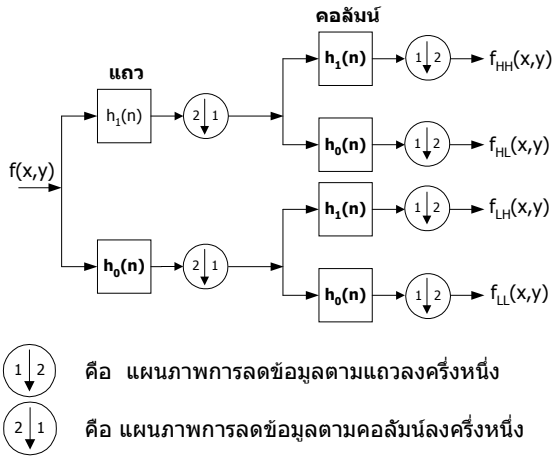
รูปที่ 1. แสดงการเข้ารหัสการบีบอัดข้อมูลภาพ



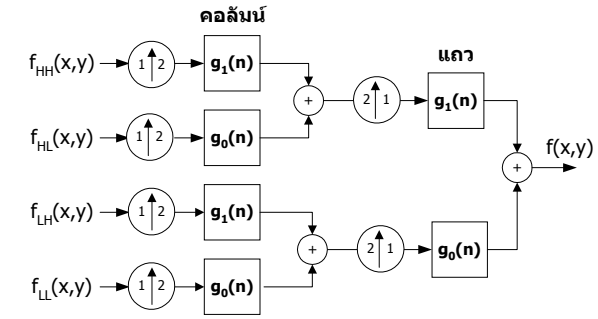
รูปที่ 2. แสดงการถอดรหัสการบีบอัดข้อมูลภาพ

2. การแปลงเวฟเล็ตแบบดีสครีท (Discrete Wavelet Transform)

การแปลงเวฟเล็ตของข้อมูลภาพ เป็นการแปลงเวฟเล็ตแบบดีสครีทแบบ 2 มิติ [1,5,7] ใช้หลักการแยกองค์ประกอบเป็นแบนด์ย่อย (Subband Decomposition) โดยมีวิธีแยกองค์ประกอบด้วยเวฟเล็ต (Wavelet Decomposition) เป็นดังรูปที่ 3 กำหนดให้ f(x,y) คือภาพต้นแบบ f_{LL} (x,y), f_{LH}(x,y), f_{HL}(x,y) และ f_{HH}(x,y) คือสัมประสิทธิ์เวฟเล็ต $h_0(n)$ และ $h_1(n)$ คือสัมประสิทธิ์ของเวฟเล็ตของการแยกองค์ประกอบ (ตัวฟิลเตอร์ที่กรองความถี่ต่ำและสูงตามลำดับ) และมีวิธีการสร้างกลับจากองค์ประกอบด้วยเวฟเล็ต (Wavelet Reconstruction) เป็นดังรูปที่ 4 กำหนดให้ $g_0(n)$ และ $g_1(n)$ คือสัมประสิทธิ์ของเวฟเล็ตของการสร้างกลับจากองค์ประกอบ (ตัวฟิลเตอร์ที่กรองความถี่ต่ำและสูงตามลำดับ)



รูปที่ 3 แสดงการแยกองค์ประกอบด้วยเวฟเล็ต



รูปที่. 4 แสดงการสร้างกลับจากองค์ประกอบด้วยเวฟเล็ต

3. การเข้ารหัส SPIHT [3,8]

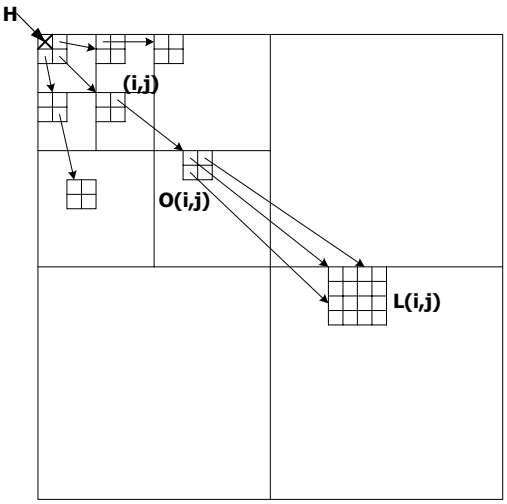
ตำแหน่งสัมประสิทธิ์เวฟเล็ตในการเข้ารหัสมีความสัมพันธ์แบบ Spatial Orientation Trees โดยมีความสัมพันธ์เป็นดังรูปที่ 5โดยปกติพลังงานของภาพ (Image Energy) ส่วนใหญ่จะอยู่ในส่วนที่มีความถี่ต่ำ ดังนั้นเราจะเคลื่อนที่จากแบนด์ย่อยที่สูงไปยังแบนด์ย่อยที่ต่ำกว่า นอกจากนี้จากการสังเกต spatial ของสัมประสิทธิ์มีความคล้ายกันระหว่างแบนด์ย่อยที่อยู่ในทิศทางเดียวกัน และคาดหวังว่าขนาดของสัมประสิทธิ์จะมีค่าน้อยลง โดยเริ่มจากแบนด์ย่อยสูงไปยังแบนด์ย่อยต่ำตามทิศทาง spatial ที่มีทิศทางเดียวกัน

การกำหนดตำแหน่ง (Coordinates) ของการเข้ารหัส

- (i,j) คือตำแหน่งของสัมประสิทธิ์เวฟเล็ตที่โหนด(node)
- O(i,j)คือกลุ่มสัมประสิทธิ์เวฟเล็ตที่ตำแหน่งของการสืบทอด (offspring) ของโหนด และ 1 โหนดมีการสืบทอด 4 ตำแหน่ง
- (k,l) คือตำแหน่งของสัมประสิทธิ์เวฟเล็ตที่สืบทอดจากโหนดนั้น คือ O(i,j) จะมี (k,l) อยู่ 4 ตำแหน่ง
- L(i,j)คือกลุ่มสัมประสิทธิ์เวฟเล็ตที่ตำแหน่งของการสืบทอดของ O (i,j) โดยมีการสืบทอดไปจนกว่าจะถึงแบนด์ย่อยต่ำสุด
- D(i,j) = O(i,j) + L(i,j) คือตำแหน่งของการสืบทอดทั้งหมดของโหนด
- H คือตำแหน่งทั้งหมดที่อยู่ในแบนด์ย่อยสูงสุดที่เป็นต้นกำเนิดโหนดและแต่ละตำแหน่งจะมีตำแหน่งของการสืบทอด 3 ตำแหน่ง

ในการเข้ารหัสและถอดรหัส จะมีตารางในการเข้ารหัสอยู่ด้วยกัน 3 ตารางคือ

1. List of Insignificant Set (LIS) คือตารางที่ใช้เก็บตำแหน่งของกลุ่มสัมประสิทธิ์เวฟเล็ตที่ยังไม่มีความสำคัญ ซึ่ง LIS นี้จะมีอยู่ด้วยกัน 2 ชนิดคือ ชนิด A และ ชนิด B ซึ่งค่าที่เก็บใน LIS ชนิด A คือ D(i,j) และค่าที่เก็บใน LIS ชนิด B คือ L(i,j)
2. List of Insignificant Pixel (LIP) คือตารางที่ใช้เก็บตำแหน่งของสัมประสิทธิ์เวฟเล็ตที่ยังไม่มีความสำคัญ
3. List of Significant Pixel (LSP) คือตารางที่ใช้เก็บตำแหน่งของสัมประสิทธิ์เวฟเล็ตที่มีความสำคัญ



รูปที่.5 แสดงความสัมพันธ์แบบ Spatial Orientation Trees

มีฟังก์ชันการเข้ารหัส $S_T(k)$ เป็นดังสมการ

$$S_T(k) = \begin{cases} 1, & \max_{(i,j)\in k}|c(i,j)| \geq T \\ 0, & \text{ortherwise} \end{cases} \quad (1)$$

$$n \leq \log_2(\max|c(i,j)|) \quad (2)$$

$$T = 2^n \quad (3)$$

เมื่อ k คือ สัมประสิทธิ์เวฟเล็ต ((i,j) หรือ (k,l)) หรือ
กลุ่มของสัมประสิทธิ์เวฟเล็ต (D(i,j) หรือ L(i,j))

3.1 อัลกอริทึมการเข้ารหัส

มีขั้นตอนดังนี้

1. กำหนดค่าเริ่มต้น
   - หาค่า n และ T จากสมการที่ 2 และ 3 ตามลำดับ
   - LSP กำหนดเป็นตารางว่าง
   - LIP กำหนดเป็นค่าสมาชิกของ H
   - LIS กำหนดเป็นค่าสมาชิกของ H และกำหนดให้เป็นชนิด A
2. การจัดลำดับ (Sorting)
   - สำหรับแต่ละ (i,j) ใน LIP
     - ส่งค่า $S_T((i,j))$ ออก

- ถ้าบิตเท่ากับ 1 ให้ส่งบิตเช็คเครื่องหมาย

- สำหรับแต่ละ (i,j) ใน LIS
  - ถ้าเป็น LIS ชนิด A
    * ส่งค่า $S_T(D(i,j))$ ออก
    * ถ้า $S_T(D(i,j))$ เท่ากับ 1

      สำหรับแต่ละ (k,l) ของ O(i,j)

      ~ ส่งค่า $S_T((k,l))$ ออก

      ~ ถ้าค่า $S_T((k,l))$ เท่ากับ 1 ให้ส่งบิตเช็คเครื่องหมาย และให้ส่ง (k,l) ไป LSP

      ~ ถ้าค่า $S_T((k,l))$ เท่ากับ 0 ให้ส่ง (k,l) ไป LIP

    * ถ้า L(i,j) มีสมาชิก ให้ส่ง (i,j) ไปท้าย LIS แล้วกำหนดให้เป็นชนิด B
  - ถ้าเป็น LIS ชนิด B
    * ส่งค่า $S_T(L(i,j))$ ออก

      ~ ถ้าค่า $S_T(L(i,j))$ เท่ากับ 1 ให้ส่งแต่ละสมาชิกของ O(i,j) ไปท้าย LIS แล้ว กำหนดให้เป็นชนิด A และย้าย (i,j) ออกจาก LIS

3. Quantization
   - n = n -1
   - สำหรับแต่ละ (i,j) ใน LSP
     - ส่งค่าบิตที่ n ของ (i,j) ออก

4. Update
   - ทำซ้ำขั้นตอนที่ 2 ถึงขั้นตอนที่ 3 และยุติเมื่อมีการส่งข้อมูลการบีบอัดครบตามต้องการ

3.2 การจัดเก็บเป็นไฟล์การบีบอัดข้อมูล

การจัดเก็บไฟล์แบ่งออกเป็น 2 ส่วน คือ ไฟล์ส่วนหัวและไฟล์ข้อมูลที่ได้จากการเข้ารหัส โดยไฟล์แต่ละส่วนจะมีการจัดเก็บค่าดังนี้

1. ไฟล์ส่วนหัวจะทำการจัดเก็บข้อมูลคือ
   - จำนวนข้อมูลทั้งหมดที่ทำการบีบอัด
   - ค่า n
   - จำนวนคอลัมน์ของภาพ
   - จำนวนแถวของภาพ
   - จำนวนระดับการแปลงเวฟเล็ต

2. ไฟล์ข้อมูลจะทำการจัดเก็บข้อมูลทั้งหมดที่ได้จากเข้ารหัสการบีบอัดข้อมูลด้วยวิธี SPIHT

3.3 อัลกอริทึมการถอดรหัส

ในการถอดรหัสจะใช้อัลกอริทึมเดียวกับการเข้ารหัส เพียงแต่เปลี่ยนจากการส่งค่าเป็นการรับค่า และนำค่านั้นมาพิจารณาเหมือนการเข้ารหัส

4. การพัฒนาอัลกอริทึม SPIHT

จากอัลกอริทึมที่กล่าวมามีจุดที่สามารถปรับปรุงพัฒนาได้ 2 จุดกล่าวคือ

1. กรณีที่ O(i,j) ไม่มีระดับความสำคัญแต่ L(i,j) มีระดับความสำคัญ จะมีการส่งบิตออกมา 2 บิต (ไม่นับรวมการเช็คแต่ละสมาชิกของ O(i,j)) ในกรณีนี้สามารถปรับอัลกอริทึมให้ส่งออกมาเพียง 1 บิตโดยการเพิ่มการเช็คการส่ง (k,l) เข้า LIP คือถ้ามีการส่ง (k,l) ทั้งหมดของ O(i,j) แสดงว่าได้เกิดกรณีนี้ขึ้น ดังนั้นสามารถทำการส่ง (k,l) เข้า LIS ให้เป็นชนิด A ได้เลย

2. กรณีที่ O(i,j) มีระดับความสำคัญแต่ L(i,j) ไม่มีระดับความสำคัญ จะมีการส่งบิตออกมาเหมือนกับข้อ 1 ในกรณีนี้สามารถปรับอัลกอริทึมให้ส่งข้อมูลให้น้อยลงได้ โดยการเพิ่ม List of Forbidden Coefficients (LFC) มีลักษณะการหา LFC ดังนี้

   - ทำการแบ่งสัมประสิทธิ์เวฟเล็ตแต่ละแบนด์ย่อยออกเป็น 4 ส่วน โดยเรียกแต่ละส่วนที่ทำการแบ่งว่า Label โดยแต่ละแบนด์ย่อยที่มีการแบ่งจะมีค่าตำแหน่ง Label ไม่ซ้ำกัน
   - หาค่าสัมประสิทธิ์สูงสุดของแต่ละ Label
   - หาค่าระดับของการควอนไทซ์ในแต่ละ Label และนำมาทำเป็น LFC โดยหาได้จากสมการดังนี้

$$LFC(w) \le \log_2(\max(c_{label}(w))) \quad (4)$$

เมื่อ W คือตำแหน่งของ Label

$C_{label}(W)$ คือสัมประสิทธิ์เวฟเล็ตทั้งหมดที่ W

LFC(W) คือระดับการควอนไทซ์ ที่ W

โดย LFC จะเป็นตัวบ่งบอกสถานะว่า O(i,j) และ L(i,j) ควรมีการส่งบิตออกมาหรือไม่ โดยจะดูว่า O(i,j) หรือ L(i,j) นี้อยู่ที่ Label ไหนและมีระดับการควอนไทซ์ที่เท่าใด ถ้าระดับของการควอนไทซ์ที่มีอยู่มีค่ามากกว่าระดับการควอนไทซ์ของ O(i,j) หรือ L(i,j) ก็ไม่ต้องทำการเช็ค O(i,j) หรือ L(i,j) สามารถเขียนเป็นฟังก์ชันได้ดังสมการที่ 3

$$F_n(p) = \begin{cases} 1, & \max_{w \in p} LFC(w) \le n \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

เมื่อ $F_n(p)$ คือฟังก์ชันที่ใช้เป็นเงื่อนไขในการเช็คค่าระดับการควอนไทซ์ของ O(i,j) หรือ L(i,j)

p คือค่า LFC ทั้งหมดที่เป็นของสมาชิกใน O(i,j) หรือ L(i,j)

4.1 อัลกอริทึมการเข้ารหัสที่มีการพัฒนา

มีการพัฒนาเฉพาะในส่วนการจัดลำดับในอัลกอริทึม SPIHT ส่วนอื่นจะเหมือนเดิม ดังนั้นจะแสดงเฉพาะในส่วนที่เป็นการจัดลำดับ และส่วนกำหนดค่าเริ่มต้น มีขั้นตอนดังนี้

1. กำหนดค่าเริ่มต้น
   - หาค่า n และ T จากสมการที่ 2 และ 3 ตามลำดับ
   - LSP กำหนดเป็นตารางว่าง
   - LIP กำหนดเป็นค่าสมาชิกของ H
   - LIS กำหนดเป็นค่าสมาชิกของ H และกำหนดให้เป็นชนิด A
   - LFC กำหนดค่าเป็นไปตามสมการที่ 4

2. การจัดลำดับ
   - สำหรับแต่ละ (i,j) ใน LIP
     - ส่งค่า $S_T((i,j))$ ออก
     - ถ้าบิตเท่ากับ 1 ให้ส่งบิตเช็คเครื่องหมาย
   - สำหรับแต่ละ (i,j) ใน LIS
     - เช็คค่า FO จาก $F_n(O(i,j))$
     - เช็คค่า FL จาก $F_n(L(i,j))$
     - ถ้าเป็น LIS ชนิด A และ FO หรือ FL เท่ากับ 1
       * ส่งค่า $S_T(D(i,j))$ ออก
       * ถ้า $S_T(D(i,j))$ เท่ากับ 1

         สำหรับแต่ละ (k,l) ของ O(i,j)
         ~ ส่งค่า $S_T((k,l))$ ออก
         ~ ถ้าค่า $S_T((k,l))$ เท่ากับ 1 ให้ส่งบิตเช็คเครื่องหมาย และให้ส่ง (k,l) ไป LSP
         ~ ถ้าค่า $S_T((k,l))$ เท่ากับ 0 ให้ส่ง (k,l) ไป LIP และเพิ่มค่าตัวเช็คการส่ง (k,l)
       * ถ้า L(i,j) มีสมาชิก
         ~ ถ้าสมาชิกทุกตัวของ O(i,j) ถูกส่งไป LIP ให้ทำการส่งสมาชิกทุกตัวของ O(i,j) ไปท้าย LIS และกำหนดให้เป็นชนิด A
         ~ ถ้าสมาชิกบางตัวของ O(i,j) ไม่ถูกส่งไป LIP ให้ทำการส่ง (i,j) ไปท้าย LIS แล้วกำหนดให้เป็นชนิด B
     - ถ้าเป็น LIS ชนิด B และ FL เท่ากับ 1
       * ส่งค่า $S_T(L(i,j))$ ออก
         ~ ถ้าค่า $S_T(L(i,j))$ เท่ากับ 1 ให้ส่งแต่ละสมาชิกของ O(i,j) ไปท้าย LISแล้ว กำหนดให้เป็นชนิด A และทำการย้าย (i,j) ออกจาก LIS

4.2 การจัดเก็บเป็นไฟล์การบีบอัดข้อมูล

การจัดเก็บไฟล์แบ่งออกเป็น 2 ส่วนคือ ไฟล์ส่วนหัวและไฟล์ข้อมูลที่ได้จากการเข้ารหัส โดยไฟล์แต่ละส่วนจะมีการจัดเก็บค่าดังนี้

1. ไฟล์ส่วนหัวจะทำการจัดเก็บข้อมูลคือ
   - จำนวนข้อมูลทั้งหมดที่ทำการบีบอัด
   - ค่า n
   - จำนวนคอลัมน์ของภาพ
   - จำนวนแถวของภาพ
   - จำนวนระดับการแปลงเวฟเล็ต
   - ค่า LFC

2. ไฟล์ข้อมูลจะทำการจัดเก็บข้อมูลทั้งหมดที่ได้จากเข้ารหัสการบีบอัดข้อมูลด้วยวิธี SPIHT

4.3 อัลกอริทึมการถอดรหัสที่มีการพัฒนา

ในการถอดรหัสจะใช้อัลกอริทึมเดียวกับการเข้ารหัส เพียงแต่เปลี่ยนจากการส่งค่าเป็นการรับค่า และนำค่านั้นมาพิจารณาเหมือนการเข้ารหัส

5. การทดสอบ

5.1 วิธีการทดสอบ

1. ทำการทดสอบหาตระกูลเวฟเล็ตแม่ ที่ให้ผลการบีบอัดข้อมูลภาพ EM ด้วยอัลกอริทึม SPIHT และอัลกอริทึม SPIHT ที่ผ่านการปรับปรุง ให้ได้ผลการบีบอัดข้อมูลภาพ EM ได้ดีที่สุด โดยใช้ข้อมูลของภาพต้นแบบจำนวน 1 ภาพคือ “Zucchini” (เป็นภาพตัวอย่างจากกล้อง EM รุ่น JEM 2010 ของบริษัท JEOL LTD) ซึ่งมีข้อมูลภาพขนาด512x512 พิกเซล แต่ละพิกเซลมี 8 บิต และตระกูลเวฟเล็ตแม่ที่ใช้ในการทดสอบมีอยู่ด้วยกัน 4 ตระกูลคือ bi9-7 [1] db4 [8] sym8 [5] และ coif5 [4] และทำการวัดประสิทธิภาพการบีบอัดข้อมูลภาพโดยใช้วิธี อัตราส่วนของสัญญาณต่อสัญญาณรบกวนสูงสุด (Peak signal-to-noise ratio, PSNR)

$$PSNR(dB) = 10\log\frac{255^2}{MSE} \tag{6}$$

$$MSE = \frac{1}{M*N}\sum_{x=1}^{M}\sum_{y=1}^{N}\left[f(x,y) - \hat{f}(x,y)\right]^2 \tag{7}$$

เมื่อ M คือจำนวนพิกเซลตามความกว้างภาพ
N คือจำนวนพิกเซลตามความสูงภาพ
$f(x,y)$ คือค่าของพิกเซลที่ตำแหน่ง (x,y) ของภาพต้นแบบ
$\hat{f}(x,y)$ คือค่าของพิกเซลที่ตำแหน่ง (x,y) ของภาพที่ผ่านการบีบอัดข้อมูล

ค่า PSNR จะสะท้อนถึงภาพที่ได้จากการบีบอัดว่ามีความผิดเพี้ยนจากภาพต้นแบบมากน้อยเพียงใด นั้นคือถ้าค่า PSNR มีค่ามากแสดงว่าภาพที่ได้จากการบีบอัดมีความผิดเพี้ยนน้อย และถ้าค่า PSNR มีค่าน้อยแสดงว่าภาพที่ได้จากการบีบอัดมีความผิดเพี้ยนมาก และวิธีอัตราบิต (Bit rate) คือค่าเฉลี่ยของจำนวนบิตต่อพิกเซล (Bits per pixel, bpp) ของภาพที่ผ่านการบีบอัดข้อมูล

$$\text{Bit rate(bpp)} = \frac{\text{bit compress}}{\text{pixel image}} \quad (8)$$

เมื่อ bit compress คือจำนวนบิตทั้งหมดของ ภาพที่ผ่านการบีบอัดข้อมูล

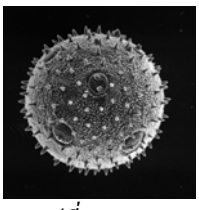
pixel image คือจำนวนพิกเซลทั้งหมดของภาพต้นแบบ

โดยอัตราบิต จะมีความสัมพันธ์กับขนาดของไฟล์ภาพที่ผ่านการบีบอัดข้อมูล เป็นดังตารางที่ 1 และทำการทดสอบที่อัตราบิตเดียวกันคือ 0.5, 1, 1.5 และ 2 bpp

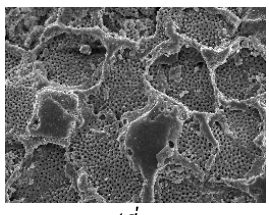
ตารางที่ 1. แสดงความสัมพันธ์ระหว่างอัตราบิต และขนาดของไฟล์ภาพที่ผ่านการบีบอัดข้อมูล

อัตราบิต (bpp)	ขนาดของไฟล์ภาพที่ผ่านการบีบข้อมูล
0.25	1/32 ของไฟล์ภาพต้นแบบ
0.5	1/16 ของไฟล์ภาพต้นแบบ
1	1/8 ของไฟล์ภาพต้นแบบ
2	1/4 ของไฟล์ภาพต้นแบบ

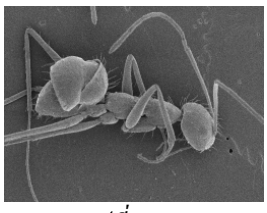
2. ทำการทดสอบอัลกอริทึม SPIHT และอัลกอริทึม SPIHT ที่ผ่านการปรับปรุง ที่อัตราบิตเดียวกันคือ 0.25, 0.5 และ 1 bpp โดยใช้ข้อมูลของภาพต้นแบบจำนวน 4 ภาพคือ “Zucchini” และ “sand” “ant” และ “LED” (เป็นภาพที่ถ่ายจากกล้อง EM รุ่น JSM-5800 LV ของบริษัท JEOL LTD) ซึ่งมีข้อมูลภาพขนาด 960x1280 พิกเซล แต่ละพิกเซลมี 8 บิต ซึ่งภาพที่นำมาทดสอบได้แสดงไว้ในรูปที่ 6 (a-d)



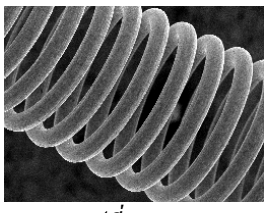
รูปที่ 6.a (Zucchini)



รูปที่ 6.b (sand)



รูปที่ 6.c (ant)



รูปที่ 6.d (LED)

รูปที่ 6 (a-d) แสดงรูปภาพที่ใช้ทดสอบการบีบอัดข้อมูลภาพ

5.2 ผลการทดสอบ

1. ผลการทดสอบหาตระกูลเวฟเล็ตแม่ที่ให้ผลการบีบอัดข้อมูลภาพ EM ได้ดีที่สุด พบว่าทั้งอัลกอริทึม SPIHT และอัลกอริทึม SPIHT ที่ผ่านการปรับปรุง ที่ใช้ bi9-7 เป็นเวฟเล็ตแม่ในการแปลงเวฟเล็ตให้ผลการบีบอัดข้อมูลภาพ EM ได้ดีที่สุด เป็นดัง ตารางที่ 2
2. ผลการทดสอบการเปรียบเทียบประสิทธิภาพการบีบอัดข้อมูลภาพ EM ด้วยอัลกอริทึม SPIHT และอัลกอริทึม SPIHT ที่ผ่านการปรับปรุง พบว่า อัลกอริทึม SPIHT ที่ผ่านการปรับปรุงให้ผลการบีบอัดข้อมูลภาพ EM ได้ดีกว่า เป็นดังตารางที่ 3 และมีตัวอย่างภาพที่ผ่านการบีบอัดข้อมูลด้วยอัลกอริทึม SPIHT ที่ผ่านการปรับปรุง เป็นดังรูปที่ 7
3. เวลาที่ใช้ในการบีบอัดข้อมูล ระหว่างอัลกอริทึม SPIHT และอัลกอริทึม SPIHT ที่ผ่านการปรับปรุง มีการใช้เวลาในการบีบอัดข้อมูลภาพที่ใกล้เคียงกันมาก เนื่องจากเงื่อนไขในการเข้ารหัสที่เพิ่มขึ้นในอัลกอริทึม SPIHT ที่ผ่านการปรับปรุง ไม่มีความซับซ้อน

ตารางที่ 2. แสดงผลการบีบอัดข้อมูลภาพ EM ที่ตระกูลเวฟเล็ตแม่ต่างๆ

ตระกูลเวฟเล็ตแม่	อัตราบิต (bpp)	PSNR (dB)	
		SPIHT	SPIHT*
bi9-7	0.5	25.9437	26.4209
	1	30.7854	31.4849
	1.5	35.0367	35.2730
	2	38.4622	38.7493
Db4	0.5	25.6627	25.7600
	1	30.0797	30.4873
	1.5	30.1865	34.7129
	2	37.3909	38.3268
Sym8	0.5	25.7788	25.8553
	1	30.1865	30.5499
	1.5	33.9292	34.6861
	2	37.3499	38.2538
Coif 5	0.5	25.5975	25.7016
	1	29.9516	30.2969
	1.5	33.6847	34.4694
	2	37.0617	38.0273

SPIHT * คือ SPIHT ที่ผ่านการปรับปรุง

ตารางที่ 3. แสดงผลการบีบอัดข้อมูลภาพ EM ด้วยอัลกอริทึม SPIHT

และอัลกอริทึม SPIHT ที่ผ่านการปรับปรุง

ภาพ	อัตราบิต (bpp)	PSNR (dB)	
		SPIHT	SPIHT*
Zucchini	0.25	22.9363	23.1409
	0.5	25.9437	26.4209
	1	30.7854	31.4849
Sand	0.25	23.3655	23.5302
	0.5	25.1323	25.3597
	1	27.7760	28.0055
Ant	0.25	34.5026	34.7356
	0.5	35.4126	35.4504
	1	37.1842	37.2121
LED	0.25	27.5024	27.6934
	0.5	28.8230	29.0162
	1	30.7168	30.9553

SPIHT * คือ SPIHT ที่ผ่านการปรับปรุง

6. สรุป

จากผลการวิจัยที่ได้นำเสนอ จากตารางที่ 1 พบว่าการใช้ bi9-7 เป็นเวฟเล็ตแม่ ให้ผลการบีบอัดข้อมูลภาพ EM ได้มีประสิทธิภาพมากที่สุด และจากตารางที่ 2 อัลกอริทึม SPIHT ที่ผ่านการปรับปรุง โดยการเพิ่ม LFC และเพิ่มเงื่อนไขการเข้ารหัสและการถอดรหัส ให้ผลการบีบอัดข้อมูลภาพที่ดีขึ้น

สำหรับงานวิจัยในอนาคต จะทำการพัฒนาการบีบอัดข้อมูลให้มีผลการบีบอัดข้อมูลที่ดีขึ้น ด้วยการนำมาเข้ารหัสเลขคณิต (Arithmetic Coding) และทำการหาข้อสรุปในการหาระดับอัตราบิตที่เหมาะสมของการบีบอัดข้อมูลภาพ EM จากการประเมินผลด้วยแบบสอบถาม

เอกสารอ้างอิง

[1] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies. “Image coding using wavelet transform” IEEE Transaction on Image Processing. Vol 1, April 1992. pp 205-220.

[2] A. Bovik, “Handbook of Image & Video Processing” New York: Academic Press. 2000

[3] B. A. Banister and T. R. Fischer. “Quadtree Classification and TCQ Image Coding.” IEEE Transactions on Circuit and Systems for Video Technology. Vol. 11, No 1, January 2001. pp 3-8.

[4] C. S. Burrus and J. E. Odegard. “Coiflet System and Zero Moments.” IEEE Transaction on Signal Processing. Vol 46, No 3, March 1998. pp. 761-766

[5] C. S. Burrus, R. A. Gopinath and H. Guo. “Introduction to Wavelets and Wavelet Transforms.” New York: Prentice-Hall International, Inc. 1998.

[6] M. J. Lai. “On the Digital-Associated with Daubechies’s Wavelets” IEEE Transaction on Signal Processing. VOL 43, September 1995. pp 2203-2205.

[7] S. G. Mallat. “A Theory for Multiresolution Signal Decomposition: The Wavelet Representation.” IEEE Transaction on Pattern and Machine Intelligence. Vol 11, July 1989. pp.674-693.

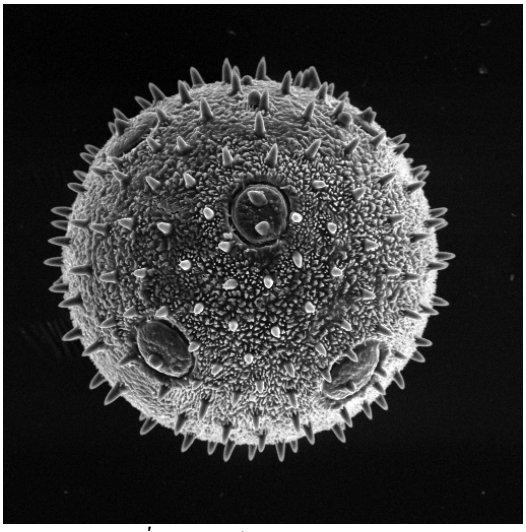
[8] A. Said and W. A. Pearlman., “A New Fast and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees.” IEEE Transaction on Circuit and Systems for Video Technology. Vol. 6, June 1996. pp 243-250.

[9] J. M. Shapiro. “Embedded image coding using zerotrees of wavelets coefficients.” IEEE Transaction on Signal Processing. Vol 41, Jan 1993. pp 3445 -3462.

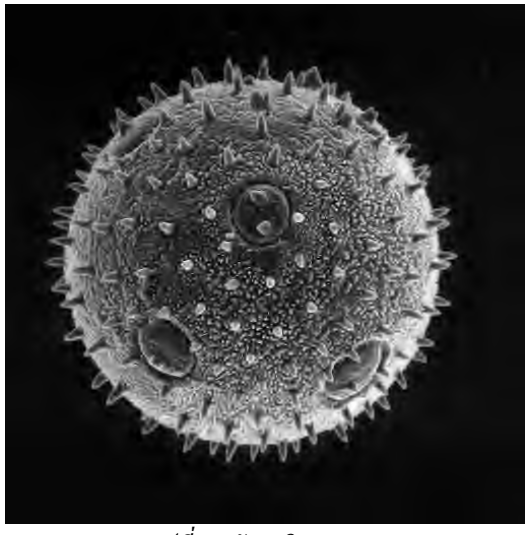
[10] ขวัญฤทัย ไพรีพ่ายฤทธิ์. ตระกูลเวฟเลตที่เหมาะสมสำหรับการเข้ารหัสแบบเอ็มเบตเด็ต ซีไร่ทรีเวฟเล็ตในการประยุกย์ใช้งานการแพทย์ทางไกล. วิทยา-นิพนธ์ปริญญาวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าธนบุรี. 2542.

[11] ศิริพร เดชะศิลารักษ์. การลดขอบบล็อคในภาพ JPEG ด้วยวิธีเวฟเล็ตเธรโชลดิ้งที่ปรับตัวเองได้. วิทยานิพนธ์ปริญญาวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าคุณทหารลาดกระบัง. 2543

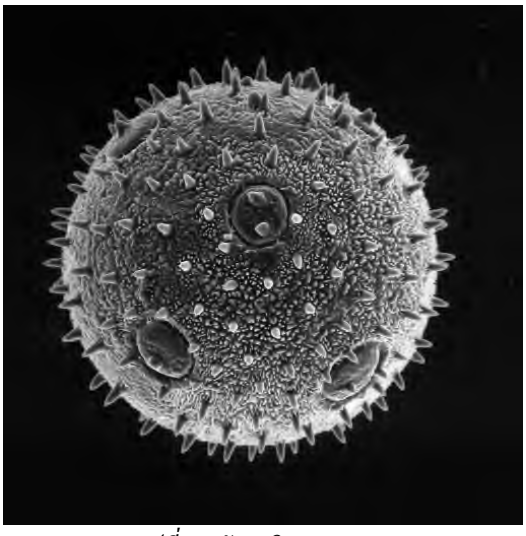
[12] เดโช ทองอร่าม และ สุวิทย์ ปุณณชัยยะ. กระบวนการทางภาพและการวิเคราะห์ภาพสำหรับกล้องจุลทรรศน์อิเล็กตรอนแบบสแกน. วารสารเครื่องมือวิจัยวิทยาศาสตร์และเทคโนโลยี. Vol. 5, 1995. หน้า 21 - 48.



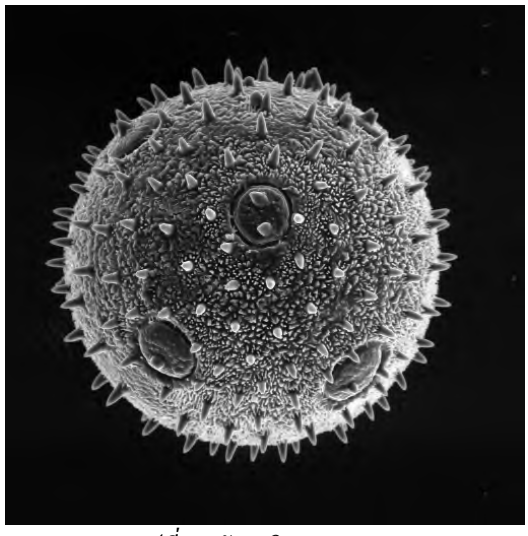
รูปที่ A1 ภาพต้นแบบ (Zucchini)
ขนาดไฟล์เท่ากับ 262,144 ไบต์



รูปที่ A2 อัตราบิต 0.25 bpp
ขนาดไฟล์เท่ากับ 8,192 ไบต์

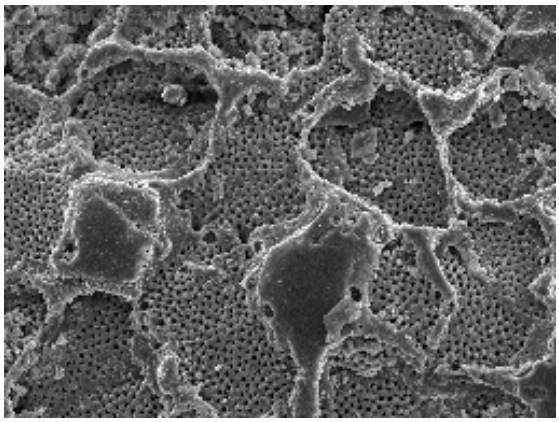


รูปที่ A3 อัตราบิต 0.5 bpp
ขนาดไฟล์เท่ากับ 16,384 ไบต์

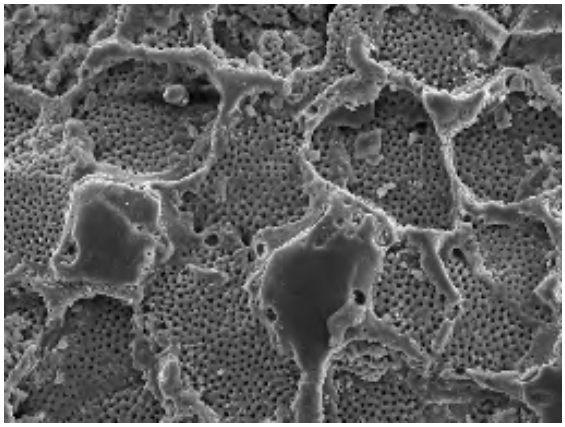


รูปที่ A4 อัตราบิต 1.00 bpp
ขนาดไฟล์เท่ากับ 32,768 ไบต์

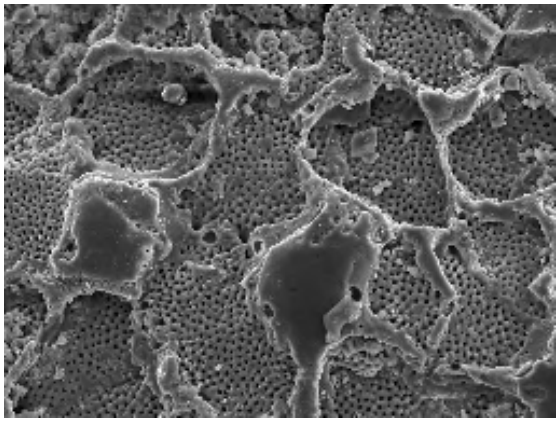
รูปที่ 7.1 แสดงรูปภาพต้นแบบ (A1) และภาพที่ผ่านการบีบอัดข้อมูลด้วยอัลกอริทึม SPIHT ที่ผ่านการปรับปรุง (A2-A4)



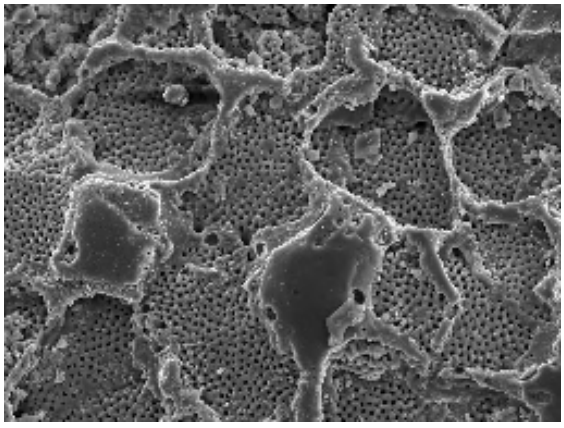
รูปที่ B1 รูปภาพต้นแบบ (sand)
ขนาดไฟล์เท่ากับ 1,228,800 ไบต์



รูปที่ B2 อัตราบิต 0.25 bpp
ขนาดไฟล์เท่ากับ 38,400 ไบต์



รูปที่ B3 อัตราบิต 0.50 bpp
ขนาดไฟล์เท่ากับ 76,800 ไบต์



รูปที่ B4 อัตราบิต 1.00 bpp
ขนาดไฟล์เท่ากับ 153,600 ไบต์

รูปที่ 7.2 แสดงรูปภาพต้นแบบ (B1) และภาพที่ผ่านการบีบอัดข้อมูลด้วยอัลกอริทึม SPIHT ที่ผ่านการปรับปรุง (B2-B4)

ณัฐนันท์ ทัดพิทักษ์กุล สำเร็จปริญญาตรีในสาขาวิชาวิศวกรรมโทรคมนาคม จากมหาวิทยาลัยเทคโนโลยีสุรนารี เมื่อ พ.ศ. 2543 ปัจจุบันกำลังศึกษาระดับปริญญาโทที่มหาวิทยาลัยเทคโนโลยีสุรนารี มีความสนใจงานวิจัยทางด้าน Digital Signal Processing, Image Processing และ Wavelet Transform

กิตติ อัตถกิจมงคล สำเร็จปริญญาเอกในสาขาวิศวกรรมไฟฟ้า จาก Vanderbilt University ประเทศสหรัฐอเมริกา เมื่อ พ.ศ. 2542 ปัจจุบันดำรงตำแหน่งเป็นอาจารย์ประจำสาขาวิชาวิศวกรรม ไฟฟ้า สำนักวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี มีความสนใจงานวิจัยทางด้าน Digital Signal Processing, Image Processing และ Wavelet Transform

สราวุฒิ สุจิตจร สำเร็จปริญญาเอกในสาขาวิศวกรรมไฟฟ้าจากมหาวิทยาลัยเบอร์มิงแฮม ประเทศอังกฤษ ปัจจุบันดำรงตำแหน่งรองศาสตราจารย์ และหัวหน้าสาขาวิศวกรรมไฟฟ้า สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี ดำเนินงานวิจัยทางด้านสัญญาณระบบและการควบคุม

การศึกษาการเกิด การเปลี่ยนระดับพลังงานในบ่อควอนตัมของสารกึ่งตัวนำ Er δ-doped InP โดยวิธีโฟโตเคอเรนต์สเปกโตรสโคปี

Study of Interband Transition in Single Quantum well of Er δ-doped InP by Photocurrent Spectroscopy

วิษณุ เพชรภา และ จิติ หนูแก้ว
ห้องปฏิบัติการวิจัยควอนตัมและสารกึ่งตัวนำทางแสง
ภาควิชาฟิสิกส์ประยุกต์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ABSTRACT - This work is the observation of the interband transition energies of Er δ-doped InP grown by Organometallic Vapor phase Epitaxy using Photocurrent Spectroscopy. The photocurrent spectra show the energy gap of ErP of 1.24 eV and three interband transition energies of InP/ErP single quantum well, 0.86 eV, 0.92 eV and 1.04 eV. The verification of these transition was made by developing a type-II single quantum well model to analyze the corresponding subband energy levels in conduction band and also the related transition energies. The results show good agreement to experimental data with acceptable band offset ratio.

บทคัดย่อ - งานวิจัยนี้เป็นการตรวจพบการเกิดการเปลี่ยนระดับพลังงานในบ่อควอนตัมของสารกึ่งตัวนำ Er δ-doped InP ที่ปลูกโดยกระบวนการ Organometallic Vapor Phase Epitaxy (OMVPE) สเปกตรัมของโฟโตเคอเรนต์ (Photocurrent) แสดงค่าพลังงานต้องห้ามของ ErP ที่ 1.24 eV นอกจากนั้นยังแสดงค่าพลังงานที่สอดคล้องกับการเปลี่ยนระดับพลังงานในบ่อควอนตัมระหว่าง InP และ ErP ซึ่งมีค่าดังนี้ 0.86 eV 0.92 eV และ 1.04 eV ในการตรวจสอบการเปลี่ยนระดับพลังงานดังกล่าว ได้สร้างแบบจำลองของบ่อศักย์ควอนตัมเดี่ยวแบบ Type-II ของ ErP/InP ขึ้น และคำนวณค่าระดับพลังงานย่อยในบ่อศักย์ของแถบนำ (Conduction Band) และสามารถนำไปคำนวณหาค่าพลังงานที่เกิดจากการเปลี่ยนระดับพลังงานของอิเล็กตรอนได้พบว่า แบบจำลองทางคณิตศาสตร์ที่สร้างขึ้นสอดคล้องกับผลการทดลองเป็นอย่างดี

คำสำคัญ - Photocurrent Spectroscopy, Er- doped InP, Type-II single quantum well

บทนำ

คุณสมบัติทางฟิสิกส์ของสารกึ่งตัวนำในกลุ่ม III-V ได้รับความสนใจเป็นอย่างมาก โดยเฉพาะอย่างยิ่งเมื่อสารนี้เจือด้วยธาตุหายาก (Rare Earth) เช่น Erbium (Er) นอกจากนั้นวิธีการปลูกผลึกสารกึ่งตัวนำที่พัฒนาขึ้น เช่นวิธีการ Organomatallic Vapor Phase Epitaxy (OMVPE) ทำให้ได้สารกึ่งตัวนำโครงสร้างซับซ้อน (Heterostructure) ที่มีคุณภาพสูง จากรายงาน [1] พบว่า เมื่อทำการเจือ Er^{3+} ลงบน InP แบบบางมาก (δ-doped) พบว่าเกิดสารประกอบ ErP (RE-V) ขึ้นที่มีโครงสร้างแบบ Rock Salt ที่มีค่า Lattice Constant a_0 = 0.5595 nm ค่าความหนาของชั้น ErP ที่ขึ้นอยู่กับเวลาในการเจือ Er (Exposure Time) สามารถวัดได้โดยกระบวนการ X-Ray Crystal Truncation Rod Scattering (X-Ray CTR) และ Rutherford Backscattering ในรายงานฉบับนี้จะกล่าวถึงการพบค่าพลังงานของการเปลี่ยนระดับพลังงานย่อย (Interband Transition Energy) ของ ErP ที่ได้จากการเจือ Er^{3+} แบบบาง (δ-doped) บน InP โดยวิธี OMVPE ซึ่งการตรวจพบทำได้โดยวิธีโฟโตเคอเรนต์สเปกโตรสโคปี (Photocurrent Spectroscopy) ในการคำนวณเพื่อหาค่าระดับพลังงานของชั้นย่อยในแถบการนำ (Conduction Band) และแถบวาเลนต์ (Valence Band) จะใช้แบบจำลองของบ่อ ควอนตัมแบบจำกัด Type-II (Type-II finite-quantum well) ระหว่าง InP และ ErP ที่มีค่า Energy Gap 1.35 eV และ 1.24 eV ตามลำดับ

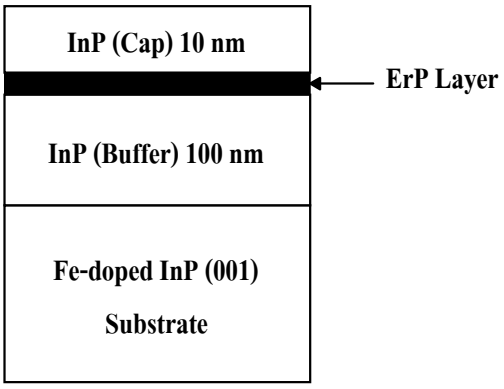
รายละเอียดการเตรียมสาร

ในการเตรียมสารกึ่งตัวนำ Er δ-doped InP โดยระบบ OMVPE แสดงโดยแผนผังดังรูปที่ 1 โดยมีรายละเอียดดังนี้ ลักษณะการไหลของก๊าซทำในลักษณะขนานกับผิวของแผ่นรอง โดยใช้ Quartz Reactor ที่ความดัน 0.1 atm ก๊าซที่ใช้ประกอบด้วย

TMIn (Trimethylindium)

TBP (Tertiarybutylphosphine) และ

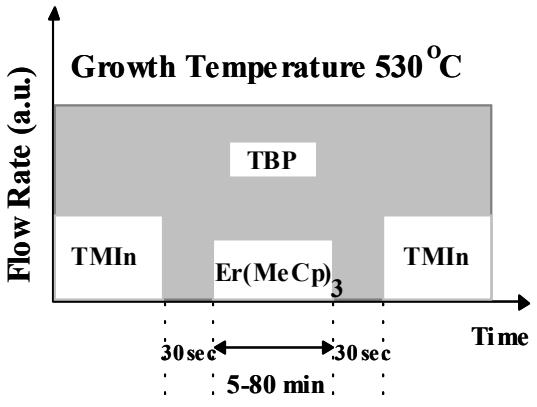
$Er(MeCp)_3$ (Tris(methylcyclopentadienyl)erbium) ที่เก็บไว้ที่ 100 $^{\circ}$C และใช้ H_2 เป็นตัวพาก๊าซทั้งสามเข้าสู่ Reactor ด้วยอัตรา 125 sccm อุณหภูมิของแผ่นรอง (Substrate) ควบคุมให้คงที่ที่ 530 $^{\circ}$C ซึ่งเป็นอุณหภูมิที่ทำให้เกิดการจับตัวและเรียงตัวของชั้น ErP ที่ดีที่สุด อุณหภูมิที่เปลี่ยนแปลงไปจากค่านี้ พบว่าจะทำให้จัดเรียงตัวของ ErP ไม่ดีพอ ชั้น InP ที่ไม่เจือ (Undoped InP Buffer Layer) หนา 100 nm เตรียมบน Fe-doped InP (001) โดยปล่อย TBP และ TMIn เข้ามาใน Reactor พร้อมกัน หลังจากนั้นจะหยุดการไหลของ TMIn ก่อนการไหลของ $Er(MeCp)_3$ ประมาณ 30 วินาที เพื่อเป็นการล้างผิวหน้า การเจือแบบ δ-doped โดยการไหล $Er(MeCp)_3$ ใช้ควบคุมความหนาของชั้น ErP ด้วยค่า Exposure time ตั้งแต่ 5-80 นาที หลังจากนั้นจะหยุดการไหลของ $Er(MeCp)_3$ ประมาณ 30 วินาที ก่อนจะปล่อย TMIn ไหลเข้าสู่ระบบเพื่อสร้างชั้นของ Undoped- InP หนา 10 nm โดยการเตรียมสารตามขั้นตอนดังกล่าว จะได้สารกึ่งตัวนำที่มีโครงสร้างดังรูปที่ 2



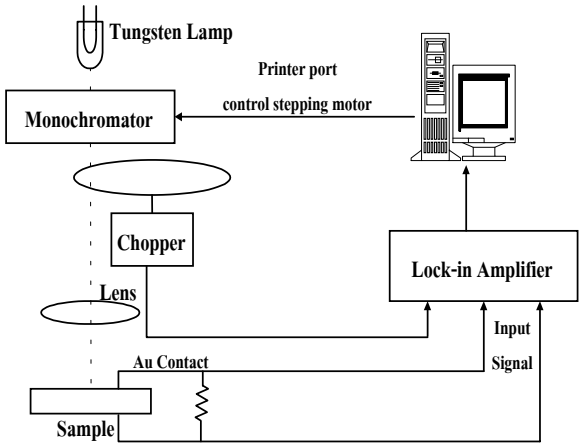
รูปที่ 2 แสดง โครงสร้างของชั้น Heterostructure ของ Er δ- doped InP

การทดลอง

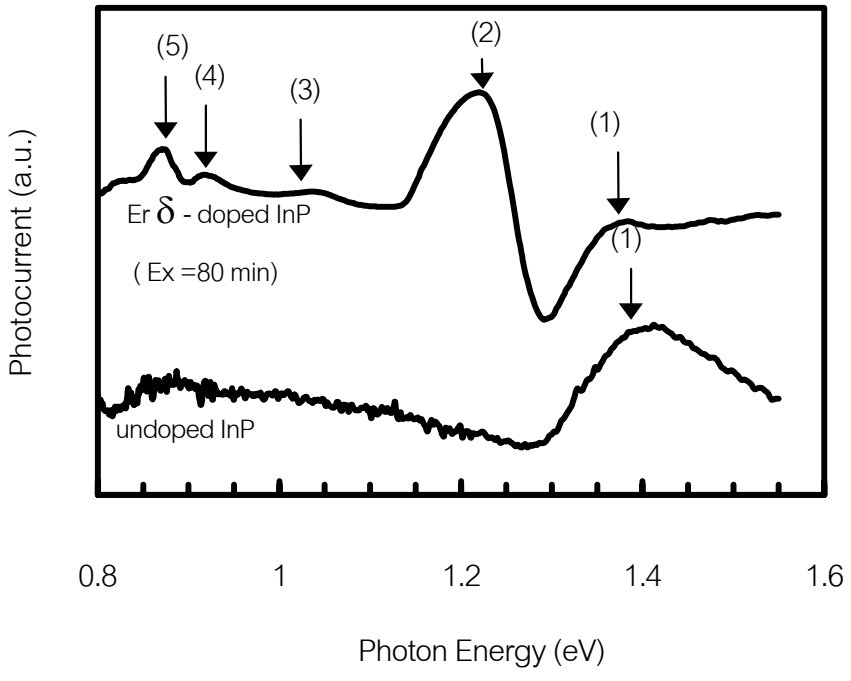
ระบบ Photocurrent Spectroscopy (PC) ที่ห้องปฏิบัติการวิจัยควอนตัมและสารกึ่งตัวนำทางแสงเป็นดังแผนผังในรูปที่ 3 แสงจากหลอดทังสเตนที่ให้แสงความยาวคลื่นตั้งแต่ 300-2200 nm จัดตั้งให้ผ่านโมโนโครมาเตอร์ (Monochromator) รุ่น SDMCI1-06 และตัวตัดแสง (Chopper) ที่หมุนด้วยความถี่ 450 Hz และทำหน้าที่มอดูเลตแสง แสงจากโมโนโครมาเตอร์ จะตกกระทบตัวอย่างสารกึ่งตัวนำ Er δ-doped InP 2 ตัวอย่างที่มีค่า Exposure Time ต่างกัน 2 ค่าคือ Exposure time 10 นาที และที่ Exposure time 80 นาที ซึ่งจะเปรียบสัญญาณนี้กับสารกึ่งตัวนำ InP ที่ไม่เจือ (Undoped InP) สัญญาณกระแสจากตัวอย่างนำเข้าสู่ Lock-in Amplifier รุ่น SR510 ที่จะทำหน้าที่ขยายสัญญาณ ข้อมูลที่ได้คือค่าของสัญญาณกระแสที่ความยาวคลื่นแสงต่าง ๆ จะควบคุมและเก็บไว้โดยคอมพิวเตอร์



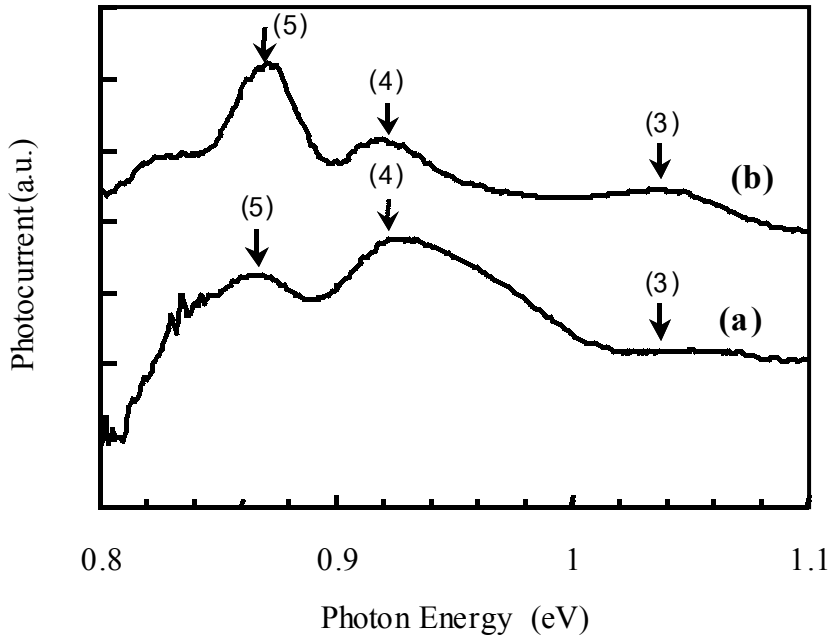
รูปที่ 1 แสดงแผนผังลำดับเวลาการไหลในการเตรียมสารกึ่งตัวนำ Er δ- doped InP โดย OMVPE



รูปที่ 3 แสดงระบบวัดโฟโตเคอเรนต์สเปกโตรสโคปี (Photocurrent Spectroscopy)



กราฟที่ 1 แสดงค่า Photocurrent ของ Undoped-InP และ Er δ-doped InP ที่ Exposure time 80 นาที



กราฟที่ 2 แสดงสัญญาณ Photocurrent ของ Er δ-doped InP ที่มี (a) Exposure time 10 นาที และ (b) 80 นาที

ผลการทดลองและการวิเคราะห์

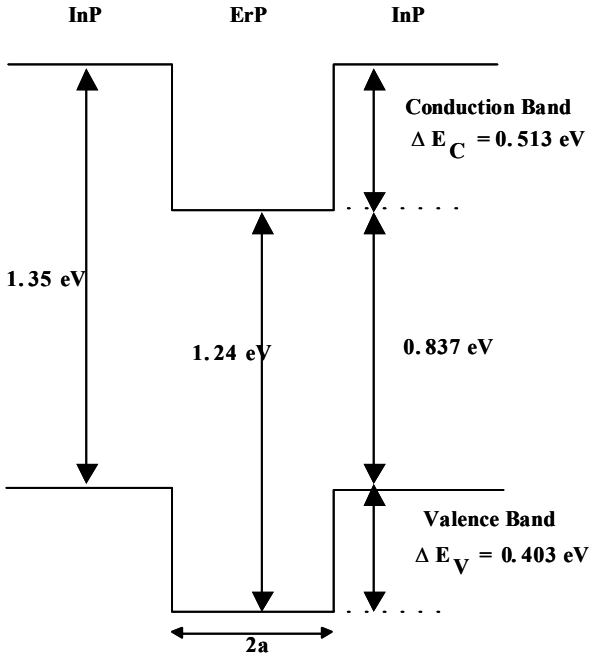
กราฟที่ 1 แสดงการเปรียบเทียบค่าของ Photocurrent ของ Undoped InP และ Er δ-doped InP ที่ Exposure time 80 นาที ค่าพลังงานโฟตอนมีค่าตั้งแต่ 0.8 eV ถึง 1.55 eV จากกราฟ สเปกตรัมที่ (1) ซึ่งตรวจพบในทั้งสองตัวอย่างมีค่าพลังงานโฟตอน 1.37 eV แสดงถึงค่าพลังงานต้องห้าม (Energy Gap: E_g) ของ InP สเปกตรัมที่ (2) ซึ่งตรงกับพลังงานโฟตอน 1.24 eV ควรเป็นสเปกตรัมที่แสดงถึงค่าพลังงานต้องห้ามของ ErP ใน Er δ-doped InP เนื่องจากค่าดังกล่าวไม่เกิดขึ้นในตัวอย่างของ Undoped-InP และค่าพลังงานต้องห้ามของ ErP นี้ยังสอดคล้องกับผลการคำนวณในรายงานอื่น [1] สเปกตรัมที่ (3) (1.08eV) สเปกตรัมที่ 4 (0.94 eV) และสเปกตรัมที่ 5 (0.87 eV) ซึ่งเกิดขึ้นในตัวอย่าง Er δ-doped InP เท่านั้น

ควรจะเกี่ยวข้องกับการเกิดการเปลี่ยนระดับพลังงานในบ่อควอนตัมของ ErP/InP ที่เป็นผลจากการเจือแบบบางด้วย Er

กราฟที่ 2 แสดงการเปรียบเทียบของสัญญาณ photocurrent ของ Er δ-doped InP ที่ Exposure time 10 นาที (กราฟ a) และ 80 นาที (กราฟ b) โดยที่ค่าพลังงานโฟตอนมีค่าตั้งแต่ 0.8 eV ถึง 1.1 eV พบว่าทั้งสองตัวอย่างให้ค่า สเปกตรัมที่ 3 (1.08 eV) สเปกตรัมที่ 4 (0.94 eV) และสเปกตรัมที่ 5 (0.87 eV) โดยที่ค่าลำดับสเปกตรัมเปรียบเทียบกับกราฟที่ 1 สเปกตรัมทั้งสามที่เกิดขึ้นในทั้งสองตัวอย่างแสดงถึงการเกิดการเปลี่ยนระดับพลังงาน (Interband Transition) ในทั้งสองตัวอย่าง ซึ่งมีค่าพลังงานโฟตอนนี้ใกล้เคียงกัน ในทางทฤษฎีทางควอนตัมพอสรุปได้ว่าค่าความกว้างของบ่อศักย์ที่หาได้จากความหนาของชั้น ErP นั้นมีค่าใกล้เคียงกันสำหรับสองตัวอย่างนี้ แสดงว่าค่าความหนาของชั้น ErP (วัดที่ FWHM) นั้นเปลี่ยนแปลงน้อยมากเมื่อเปลี่ยนค่า Exposure Time ซึ่งผลดังกล่าวสอดคล้องกับผลที่ได้โดยวิธีการ X-Ray CTR ในงานวิจัยอื่น [2,3] นอกจากนั้นสัญญาณ photocurrent ของตัวอย่างที่มีค่า Exposure 80 นาที มีความคมชัดของสัญญาณมากกว่าตัวอย่างที่ Exposure 10 นาที แสดงว่าเวลาในการเจือ Er โดยกระบวนการ OMVPE นั้นมีผลต่อการจับตัวและจัดเรียงตัวของ ErP กล่าวคือเมื่อเวลาในการเจือนานขึ้น จะทำให้การจับตัวระหว่าง Er กับ P และเกิดการจัดเรียงตัวของชั้น ErP ที่ดีขึ้น ซึ่งสอดคล้องกับผลการตรวจสอบ โดย X-Ray CTR [2,3] ที่พบว่า ชั้นของ ErP จะเริ่มเกิดขึ้นเมื่อเวลาในการเจือ Er นานกว่า 5 นาที การจัดเรียงตัวของชั้น ErP ที่ดีกว่านี้ ยังผลให้เกิดโครงสร้างบ่อควอนตัมที่ดีกว่า และทำให้ได้สัญญาณ Photocurrent ของการเปลี่ยนระดับพลังงานในบ่อควอนตัมที่คมชัดกว่า

แบบจำลองทางคณิตศาสตร์

แบบจำลองทางคณิตศาสตร์ที่สร้างขึ้นมาเป็นแบบจำลองอย่างง่ายเพื่อใช้ในการคำนวณหาค่าพลังงานที่เกิดจากการเปลี่ยนระดับพลังงานในบ่อควอนตัมของ ErP/InP ทั้งนี้ ค่าข้อมูลต่าง ๆ ที่ใช้ในแบบจำลองนี้ได้จากการทดลองและจากงานวิจัยอื่น [1-3] เนื่องจาก ErP ที่เจืออยู่บน InP มีค่า E_g ประมาณ 1.24 eV ขณะที่ InP มีค่า E_g เท่ากับ 1.35 eV ทำให้เกิดบ่อศักย์ควอนตัมที่จำกัด (Finite single quantum well) ขึ้น และเนื่องจากชั้นของ ErP เกิดจากการเจือแบบบาง (δ-doping) จึงมีผลทำให้บ่อศักย์ควอนตัมที่เกิดขึ้นควรมีลักษณะเป็นแบบ Type-II นอกจากนั้นจากผลการตรวจวัดค่าการกระจายของชั้น ErP โดยวิธีการ X-Ray CTR พบว่าลักษณะของการกระจายของชั้นErP นั้นมีลักษณะคล้ายรูปตัวยู (U-shaped) ดังนั้นลักษณะของบ่อศักย์ที่เกิดขึ้นสามารถประมาณให้เป็นรูปสี่เหลี่ยม (square well) ได้ ดังนั้นแบบจำลองบ่อศักย์ที่ใช้ในหาค่าพลังงานในการเปลี่ยนระดับพลังงานย่อย จะเป็นแบบ Type-II single finite square well ดังแสดงในรูปที่ 4



รูปที่ 4 แสดงแบบจำลองบ่อศักย์ควอนตัมของ InP/ErP/InP

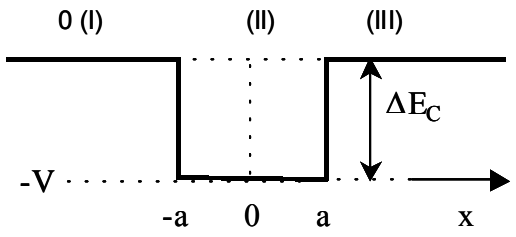
ค่าข้อมูลที่สำคัญที่ใช้ในการคำนวณแสดงในตารางที่ 1 ความกว้างของบ่อศักย์ของทั้งสองตัวอย่างนั้นได้จากการวิเคราะห์โดย X-ray CTR [2,3] มีค่าประมาณ 2.4 nm (8 ML) โดยวัดที่ Full width at Half Maximum ของสัญญาณที่ได้จาก X-ray CTR

ค่ามวลยังผล (Effective mass) ของ อิเล็กตรอน (m_e^*) และ โฮล (m_h^*) มีค่าเป็น $1.71m_0$ และ $0.14m_0$ ตามลำดับ [1] ส่วนค่าความลึกของบ่อศักย์สามารถคำนวณได้จากค่า Band offset ซึ่งโดยปกติสำหรับสารกึ่งตัวนำโครงสร้างบ่อควอนตัมโดยทั่วไปค่า Band offset จะอยู่ในช่วง 60:40 ในแบบจำลองนี้ค่า Band offset ที่ใช้มีค่าเป็น 56:44 ซึ่งเป็นค่าที่ยอมรับได้ ค่า Band offset ดังกล่าวจะให้ค่าความลึกบ่อศักย์ดังแสดงในรูปที่ 4 การคำนวณหาค่าระดับพลังงานย่อย (subband) ในบ่อควอนตัม ทำได้โดยการแก้สมการชโรดิงเจอร์ใน 1 มิติ (Schrödinger Equation)

ตารางที่ 1 แสดงค่าข้อมูลของ ErP ที่ใช้ในการคำนวณค่า ระดับพลังงานย่อยในแถบการนำ

m_e^*	$1.71m_0$
Energy gap of ErP	1.24 eV
Well Width (2a)	2.4 nm (8 ML)
Band offset	56:44

เมื่อพิจารณาบ่อศักย์ควอนตัมของแถบการนำ ดังรูป



ค่าของศักย์ที่บริเวณต่าง ๆ คือ

$V(x) = 0 \quad x < -a$ บริเวณ (I)

$= -V \quad -a < x < a$ บริเวณ (II)

$= 0 \quad a < x$ บริเวณ (III)

สมการชโรดิงเจอร์ใน 1 มิติ เขียนได้เป็น

$$\frac{d^2\psi}{dx^2} + \frac{2m_e^*}{\eta^2} E\psi = 0 \qquad \text{..... (1)}$$

สำหรับบริเวณ (I) and (III) และ

$$\frac{d^2\psi}{dx^2} + \frac{2m_e^*}{\eta^2} (E + V)\psi = 0 \qquad \text{..... (2)}$$

สำหรับบริเวณ (II)

กำหนดให้

$$k^2 = -\frac{2mE}{\eta^2} \quad \text{และ} \quad q^2 = \frac{2m(E+V)}{\eta^2}$$

คำตอบของสมการ (1) และ (2) ซึ่งจะนำไปหาค่าระดับพลังงานย่อยใน บ่อศักย์ อยู่ในรูปของ k และ q ที่สอดคล้องกับสมการ

$$\frac{k}{q} = \tan qa \qquad \text{สำหรับ even parity} \qquad \text{..... (3)}$$

$$-\frac{k}{q} = \cot qa \qquad \text{สำหรับ odd parity} \qquad \text{..... (4)}$$

โดยที่

$$\frac{k}{q} = \left(\frac{-E}{V+E}\right)^{\frac{1}{2}} \qquad \text{..... (5)}$$

และ

$$k^2 + q^2 = \frac{2m_e^*}{\eta^2} V \qquad \text{..... (6)}$$

ดังนั้นค่าของระดับพลังงานย่อยในบ่อศักย์สามารถได้จากจุดตัดระหว่างกราฟ tan qa, -cot qa และ k/q ที่ค่า qa ต่าง ๆ เมื่อทราบค่า k/q และ qa จากจุดตัดกราฟจะสามารถหาค่าระดับพลังงานได้จากสมการ

$$E = \frac{\eta^2}{2m_e^*}(qa)^2 - V \qquad \text{..... (7)}$$

จากรูปที่ 5 จะได้ค่า qa ที่จุดตัดของกราฟของสมการ (3) และ (4) ทั้งหมด 3 ค่าคือ 1.33, 2.66 และ 3.95 ตามลำดับ และค่า qa ทั้งสามค่านี้ให้ค่าระดับพลังงานย่อยที่สอดคล้องคือ E_{e1} = -0.486 eV, E_{e2} = -0.404 eV และ E_{e3} = -0.271 eV ตามลำดับ

สำหรับแถบวาเลนต์มีลักษณะของกำแพงศักย์แบบจำกัด 1 มิติ (One Dimensional Quantum Barrier) สำหรับโฮล จึงทำให้แถบพลังงานในแถบวาเลนต์มีลักษณะต่อเนื่องตลอดภายในแถบวาเลนต์ ดังนั้นเมื่อกระตุ้นด้วยแสง จะทำให้อิเล็กตรอนสามารถทะลุ (tunneling) จากบริเวณสูงสุดของแถบวาเลนต์ไปยังแถบพลังงานย่อยทั้งสามแถบในแถบการนำ ทำให้เกิดสเปกตรัม Photocurrent ที่สอดคล้องกับพลังงานที่เกิดจากการเปลี่ยนระดับพลังงาน interband Transition Energy (E_T)

ค่าพลังงานที่สอดคล้องกับระดับพลังงานย่อยต่างๆ มีค่าดังนี้

E_{T1} = 0.863 eV

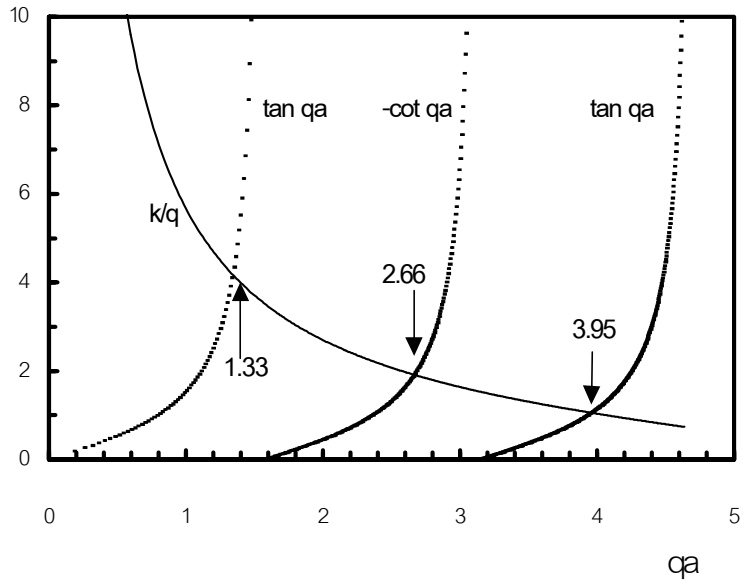
E_{T2} = 0.945 eV

E_{T3} = 1.079 eV

ค่าที่ได้จากแบบจำลองทางคณิตศาสตร์นี้สอดคล้องเป็นอย่างดีกับค่าสเปกตรัมที่ (5), (4) และ (3) ของสัญญาณ Photocurrent ตามลำดับ

สรุปผล

ค่าเปลี่ยนระดับพลังงานย่อยและค่า พลังงานต้องห้ามของ ErP ของสารกึ่งตัวนำ Er δ-doped InP ตรวจพบโดยวิธีการ Photocurrent Spectroscopy ในการคำนวณเพื่อจะหาค่าที่สอดคล้องกับค่าที่ได้จากการทดลอง ใช้แบบจำลองของ Type-II Finite square Quantum Well ระหว่าง InP และ ErP พบว่าค่าที่ได้จากแบบจำลองและการทดลองสอดคล้องกันที่ค่า Band offset หรืออัตราส่วนระหว่าง ΔE_C / ΔE_V = 56/44 ซึ่งเป็นช่วงที่ยอมรับได้



รูปที่ 5 แสดงกราฟของ tan qa, -cot qa และ k/q ที่ค่า qa ต่าง ๆ

กิตติกรรมประกาศ

ขอขอบคุณ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ที่ให้การสนับสนุนทุนวิจัย ขอบคุณ Prof. Dr. Y. Takada และนักวิจัยประจำห้องปฏิบัติการที่ภาควิชา Material Science and Engineering, Nagoya University ที่อนุเคราะห์ตัวอย่างสารกึ่งตัวนำที่ใช้ในการวิเคราะห์

เอกสารอ้างอิง

[1] L. Bolotov, T. Tsuchiya, A. Nakamura, T. Ito, T. Fujiwara, Y. Takada, Phys. Rev. B, Condensed Matter. Vol 59, No. 19, pp. 12236-9, (1999)

[2] Y. Takada, K. Fujita, M. Matsubara, N. Yamada, S. Ichiki, M. Tabuchi and Y. Fujiwara, J. Appl. Phys. 82(2),pp. 635-638, 15 July 1997

[3] K. Fujita, J. Tsuchiya, S. Ichiki, H. Hamamatsu, N. matsumoto, M. Tabushi, Y. Fujiwara, and Y. Takeda, Appl. Suf. Sci., 117/118,pp. 758-789, (1997)

[4] Doyeol, Phys. Rev. B, Condensed Matter. Vol 48, No. 11, pp. 781-4, (1993)

[5] A.G. Petukhov, W.R.L.Lambrecht, and B.Segall, Phys.Rev.B, 53, 4324 (1996)

[6] Y.K. Su, R.L.Wang, and H.H.Tsai, IEEE Transactionson Electron Devices, Vol. 40, NO. 12, 1993

วิษณุ เพชรภา จบการศึกษาระดับปริญญาตรี สาขาฟิสิกส์ (เกียรตินิยม) จากมหาวิทยาลัยเชียงใหม่ จบการศึกษาระดับปริญญาโท สาขาฟิสิกส์ จาก University of Central Florida ประเทศสหรัฐอเมริกา ปัจจุบันดำรงตำแหน่งอาจารย์ประจำภาควิชาฟิสิกส์ประยุกต์ และนักศึกษาระดับปริญญาเอกทางด้านฟิสิกส์ประยุกต์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

จิตติ หนูแก้ว จบการศึกษาระดับปริญญาตรีสาขาฟิสิกส์ จากมหาวิทยาลัยศรีนครินทรวิโรฒสงขลา พ.ศ. 2526 จบการศึกษาระดับปริญญาโทสาขาฟิสิกส์จากมหาวิทยาลัยเชียงใหม่ พ.ศ. 2532 และจบปริญญาเอกสาขา Material Sciences and Engineering จากมหาวิทยาลัยนาโกยา พ.ศ. 2541 ปัจจุบันดำรงตำแหน่งอาจารย์ประจำภาควิชาฟิสิกส์ประยุกต์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง หัวข้องานวิจัยที่สนใจได้แก่ Quantum Well Device, Semiconductor Physics และ Optical Instrumentation.

Probabilistic Loop Scheduling for Applications with Uncertain Execution Time

Sissades Tongsima, Edwin H.-M. Sha, Member, IEEE
Chantana Chantrapornchai, David R. Surma Member, IEEE and
Nelson Luiz Passos, Member, IEEE

ABSTRACT - One of the difficulties in high-level synthesis and compiler optimization is obtaining a good schedule without knowing the exact computation time of the tasks involved. The uncertain computation times of these tasks normally occur when conditional instructions are employed and/or inputs of the tasks influence the computation time. The relationship between these tasks can be represented as a data-flow graph where each node models the task associated with a probabilistic computation time. A set of edges represents the dependencies between tasks. In this research, we study scheduling and optimization algorithms taking into account the probabilistic execution times. Two novel algorithms, called *probabilistic retiming* and *probabilistic rotation scheduling*, are developed for solving the underlying non-resource and resource constrained scheduling problems respectively. Experimental results show that probabilistic retiming consistently produces a graph with a smaller longest path computation time for a given confidence level, as compared with the traditional retiming algorithm that assumes a fixed worst-case and average-case computation times. Furthermore when considering the resource constraints and probabilistic environments, probabilistic rotation scheduling gives a schedule whose length is guaranteed to satisfy a given probability requirement. This schedule is better than schedules produced by other algorithms that consider worst-case and average-case scenarios.

KEYWORDS - Scheduling, loop pipelining, probabilistic approach, retiming, rotation scheduling.

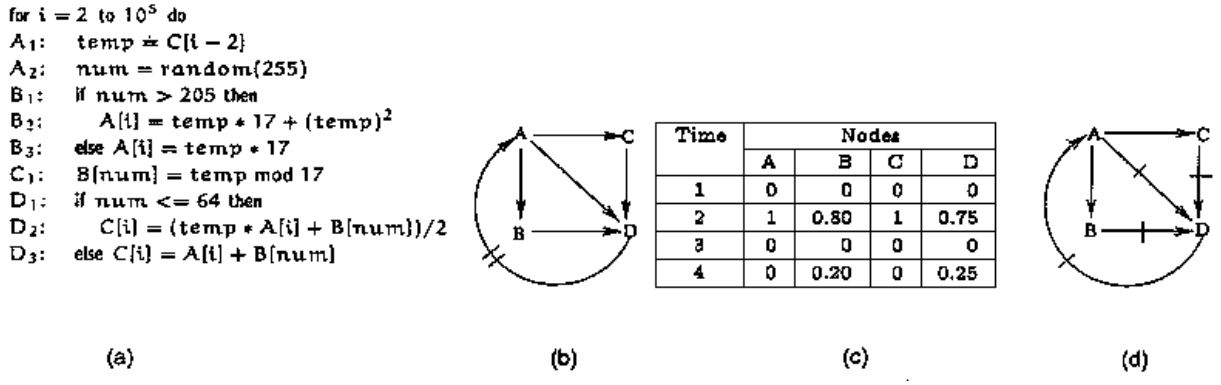
1. Introduction

In many practical applications such as interface systems, fuzzy systems, artificial intelligence systems, and others, the required tasks normally have uncertain computation times (called *uncertain* or *probabilistic* tasks for brevity). Such tasks normally contain conditional instructions and/or operations that could take different computation times for different inputs. A dynamic scheduling scheme may be considered to address the problem; however, the decision of the run-time scheduler which depends on the local on-line knowledge may not give a good overall schedule. Although many static scheduling techniques can thoroughly check for the best assignment for dependent tasks, existing methods are not able to deal with such uncertainty. Therefore, either worst-case or average-case computation times for these tasks are usually assumed. Such assumptions, however, may not be applicable for the real operating situation and may result in an inefficient schedule.

For iterative applications, statistics for the uncertain tasks are not difficult to collect. In this paper, two novel loop cheduling algorithms, *probabilistic retiming* (PR) and *probabilistic rotation scheduling* (PRS), are proposed to statically schedule these tasks for non-resource (assume *unlimited* number of target processors) and resource constrained (assume *limited* number of target processors) systems respectively. These algorithms expose the parallelism of the probabilistic tasks across iterations as well as take advantage of the inherent statistical data. For a system without resource constraints, PR can be applied to optimize the input graph (i.e., reduce the length of the longest path of the graph such that the probability of the longest path computation time being less than or equal to some given computation time, c, is greater than or equal to a given confidence probabilityθ). The resulting graph implies a schedule for the non-resource constrained system where the longest path computation time determines its *schedule length*. On the other hand, the PRS algorithm is used to schedule uncertain tasks to a fixed number of multiple processing elements. It produces a schedule length from the given graph and incrementally reduces the length so that the probability of it being less than the previous length is greater than or equal to the given confidence probability.

- *S. Tongsima is with the National Electronics and Computer Technology Center, National Science and Technology Development Agency, Bangkok 10400, Thailand. E-mail: stongsim@hpcc.nectec.or.th.*
- *E.H.-M. Sha is with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556. E-mail: esha@cse.nd.edu.*
- *C. Chantrapornchai is with the Faculty of Sience, Silpakorn University, Noakorn Pathom 73000, Thailand.*
- *D.R. Surma is with the Department of Mathematics and Computer Science, Valparaiso University, Valparaiso, IN 46383.*
- *N.L. Passos is with the Department of Computer Science, Midwestern State University, Wichita Falls, TX 76308.*

This paper is reprinted from a paper published in IEEE Transaction on Computer. Vol. 49, No. 1, January 2000.



Time	Nodes			
	A	B	C	D
1	0	0	0	0
2	1	0.80	1	0.75
3	0	0	0	0
4	0	0.20	0	0.25

Figure 1. Illustrates a sample code segment, the corresponding PG and its computation time, and the retimed graph. (a) Code segment. (b) PG. (c) Timing information. (d) Retimed PG.

In order to be compatible with the current high performance parallel processing technology, we assume that synchronization is required at the end of each iteration. Such a parallel computing style is also known as *synchronous parallelism* [19], [10]. Both PR and PRS take an input application which can be modeled as a probabilistic data-flow graph (PG), which is a generalized version of a data-flow graph (DFG) where a node corresponds to a task (a collection of statements), and a set of edges representing dependencies between these tasks and determine a schedule. The loop-carried dependences (dependency distances) between tasks in different iterations are represented by short bar lines on the corresponding edges. Since the computation times of the nodes can be either fixed or varied, a probability model is employed to represent the timing of the task.

Fig.1b shows an example of a PG consisting of four nodes. Note that such a graph models the code segment presented in Fig. 1a, where, for example, A in the PG corresponds to A_1 and A_2 of the code segment. Two bar lines on the edge between nodes D and A represent the dependency distances between these two nodes. The computation time of nodes A and C are known to be fixed (2 time units). In this code, the uncertainty occurs in the computation of nodes B and D. Assume that each arithmetic operation and the assignment operation (=) take 1 time unit. Furthermore, the computation time of the comparison and random number generating operations are assumed negligible. Hence, it may take either 4 or 2 time units to execute node B. Put another way, about 20 percent of the time (51 out of 256), statement B_2 will be executed and node B will take 4 time units; otherwise node B takes only 2 time units (B_3 has only one operation). Likewise, approximately 25 percent (64 out of 256), node D takes 4 time units, and about 75 percent, it will take 2 time units. Each entry in Fig. 1c shows a probability associated with each node's possible computation time (the probability distribution). By taking into account these varying timing characteristics, the proposed technique can be applied to a wide variety of applications in high-level synthesis and compiler optimization.

Considerable research has been conducted in the area of finding a schedule of a directed-acyclic graph (DAG) for multiple processing systems. (Note that DAGs are obtained from DFGs by ignoring edges of a DFG containing one or more dependency distances.) Many heuristics have been proposed to schedule DAGs, e.g., list scheduling, graph decomposition [13], [11], etc. These methods, however, consider neither exploring the parallelism across iterations nor addressing the problems of probabilistic tasks.

For *instruction level parallelism* (ILP) scheduling, trace scheduling [9] is used to globally schedule DAGs by rearraging some operations in the graphs. Percolation scheduling is used in a development environment [1] for microcode compaction, i.e., parallelism extraction of horizontal microcode. Nevertheless, the graph model used in these techniques does not reflect the uncertainty in node computation times. In the class of global cyclic scheduling, software pipelining [16] is used to overlap instructions, whereby the parallelism is exposed across iterations. This technique, however, expands the graph by unfolding or unrolling [22] it resulting in a larger code size. Loop transformations are also common techniques used to construct parallel compilers. They restructure loops from the repetitive code segment in order to reduce the total execution time of the schedule [2], [3], [20], [27], [28]. These techniques, however, do not consider that the target systems have limited number of processors or that task computation times are uncertain.

Modulo scheduling [24], [25], [26] is a popular technique in compiler design for exploiting ILP in loops which results in optimized codes. This framework specifies a lower bound, called initiation interval (II), to start with and strives to schedule nodes based on such knowledge. Much research was introduced to improve and/or expand the capability of modulo scheduling. For example, research was presented which improved modulo scheduling by producing schedules while considering limited number of registers [7], [8], [21]. In [17], a combination of modulo scheduling and loop unrolling was introduced and applied in the IMPACT compiler [4]. These ILP approaches, however, are limited to solving problems without considering uncertain computation times (probabilistic graph model).

Some research considers the uncertainty inherit in the computation time of nodes. Ku and De Micheli [14], [15] proposed a relative scheduling method which handles tasks

with unbounded delays. Nevertheless, their approach considers a DAG as an input and does not explore the parallelism across iterations. Furthermore, even if the statistics of the computation time of uncertain nodes is collected, their method will not exploit this information. A framework that is able to handle imprecise propagation delays is proposed by Karkowski and Otten [12]. In their approach, fuzzy set theory [29] was employed to model the imprecise computation times. Although their approach is equivalent to finding a schedule of imprecise tasks to a non-resource constrained system, their model is restricted to a simple triangular fuzzy distribution and does not consider probability values.

For scheduling under resource constraints, the *rotation scheduling* technique was presented by Chao et al. [5], [6] and was extended to handle multi-dimensional applications by Passos et al. [23]. Rotation scheduling attempts to pipeline a loop by assigning nodes from the loop to the system with a limited number of processing elements. It implicitly uses traditional retiming [18] in order to reduce the total computation time of the nodes along the longest paths (also called the critical paths), in the DFG. In other words, the graph is transformed in such a way that the parallelism is exposed but the behavior of the graph is preserved. In this paper, the rotation scheduling technique is extended so that it can deal with uncertain tasks.

Since the computation time of a node in a PG is a random variable, the total computation time of this graph is also a random variable. The concept of a *control step* (the synchronization of the tasks "within" each iteration) is no longer applicable. A schedule conveys only the execution *order* or *pattern* of the tasks being executed in a functional unit and/or between different units. In order to compute the total computation time of this ordering, a *probabilistic task-assignment graph* (PTG) is constructed. A PTG is obtained from a PG in which non-zero dependency distance edges are ignored and each node is assigned to a specific functional unit in the system. The PTG also contains additional edges, called *flow-control* edges where a connection from *u* to *v* means that *u* is executed immediately before *v* using the same functional unit. Note that in the non-resource constrained scenario, the PTG will be the DAG portion of the PG (a subgraph that contains only no dependency distance edges).

Let us use the example in Fig. 1b. Assume that the term longest path computation time entails finding the maximum of the summation of computation times of nodes along paths which contain no dependency distances. After examining all possible longest paths of this graph, it is likely (60 percent) that its longest path computation time is less than or equal to 8. The details of how this value is determined is given in Section 3. Note that if all nodes in this graph are assigned their worst-case values, the longest path computation time (or schedule length for non-resource constrained systems) of this graph will be 10. One might wish to reduce the longest path of this graph in nearly all cases, for example reducing the chance of the clock period being greater than 6. By applying probabilistic retiming, the longest path computation time of the graph may be improved with respect to the given constraint. The modified graph after retiming is shown in Fig. 1d. The longest path computation time of this graph is less than than or equal to 6 with 20 percent chance.

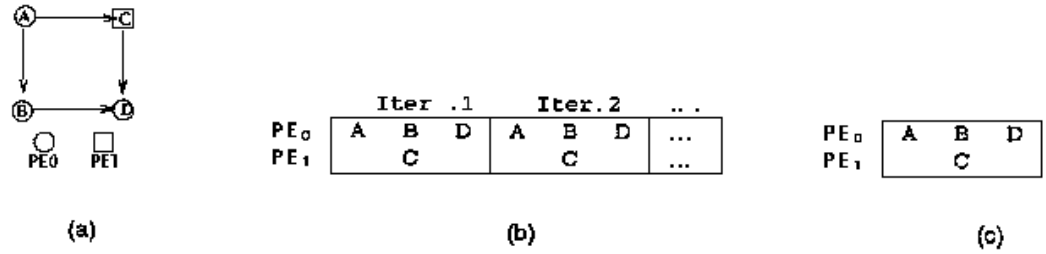


Figure 2. Illustrates an example of PTG, its corresponding repeated pattern, and the static execution order. (a) The PTG. (b) Initial execution pattern. (c) Schedule.

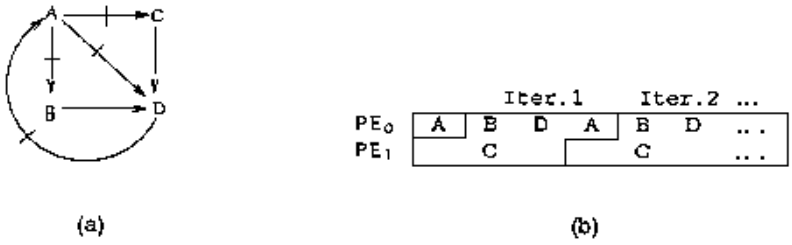


Figure 3. Illustrates the corresponding retimed PG and the repeated pattern after changing iteration window. (a) Rotate A. (b) Reshaping iteration window.

If we need to schedule nodes from the PG to two homogeneous functional units, a possible PTG can be constructed as shown in Fig. 2a. Since the input graph is cyclic, an *execution pattern* of this PTG is repeated and the synchronization is applied at the end of each iteration, as shown in Fig. 2a. The solid edges in this PTG represent those zero dependency distance edges, called dependency edges, from the input graph (see Fig. 1b). In this figure, nodes *A, B* and *D* are assigned to PE_0 and node *C* is bound to PE_1. Note that *D* is implicitly executed after *A*; therefore, the direct edge from *A* to *D* from the original input graph can be omitted. A corresponding static schedule which shows only one iteration from the execution pattern is shown in Fig. 2c.

The resulting longest path computation time of the PTG is less than 9 units with 90 percent certainty. This longest path timing and its probability are also known as a schedule length for resource constrained systems. We can improve the resulting schedule length by applying our probabilistic rotation scheduling algorithm to the PG and its PTG. In this case the algorithm first selects the root node *A* to be rescheduled. Then one dependency distance from the incoming edges of node *A* is moved to all its outgoing edges. Fig.3a. shows the resulting transformation graph of the PG. This new graph will be used as a reference to later update the PTG. The new execution pattern is equivalent to reshaping the iteration window as presented in Fig. 3b.

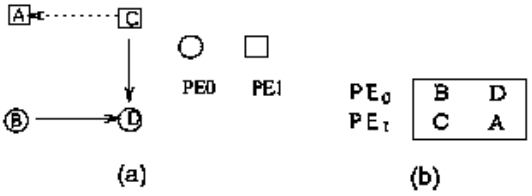


Figure 4. Illustrates the resulting PTG and its execution order after rescheduling A. (a) PTG. (b) Static execution order.

By applying the PRS algorithm, node *A* from the next iteration (see Fig. 3b.) is introduced to the static execution pattern. Note that node *A* has no inter-iteration dependencies associated with it. Therefore, *A* can be rescheduled to any available functional unit. One possible schedule is to assign node *A* immediately after node *C* in PE_1. The resulting PTG and the new execution order are shown in Fig. 4a and 4b, respectively. The dotted arrow from *C* to *A* in this new PTG represents the flow-control edge. For this PTG, the resulting schedule length will be less than seven with higher than 90 percent confidence.

The remainder of this paper is organized as follows. Section 2 presents the graph model used in this work. Required terminology and fundamental concepts are also presented. Section 3 discusses probabilistic retiming and the algorithm for computing a total computation time of a probabilistic graph. The probabilistic rotation scheduling algorithm and the supported routines will be discussed in Section 4. Experimental results are discussed in Section 5. Finally, Section 6 draws conclusions of this research.

2. Preliminaries

In this section, the graph model which is used to represent tasks with uncertain computation times is introduced. Terminology and notations relevant to this work are also discussed. We begin by examining a DFG that contains tasks with uncertain computation time which can be modeled as a *probabilistic graph* (PG). The following gives the formal definition for such a graph.

Definition 2.1. *A probabilistic graph (PG) is a vertex-weighted, edge-weighted, directed graph* $G = \langle V,E,d,T \rangle$, *where V is the set of vertices representing tasks, E is the set of edges representing the data dependencies between vertices, d is a function from E to the set of non-negative integers, representing the number of dependency distance on an edge, and* T_v *is a random variable representing the computation time of a node* $v \in V$.

Note that traditional DFGs are a special case of PGs where all probabilities equal one. Each vertex $v \in V$ is weighted with a *probability distribution* of the computation time, given by T_v, where T_v is a discrete random variable corresponding to the computation time of *v* such that $\Sigma_{\forall x} \Pr(T_v = x) = 1$. The notation $\Pr(T_v = x)$ is read "the probability that random variable *T* assumes value *x*". The *probability distribution* of *T* is assumed to be discrete in this paper. The granularity of the resulting probability distribution, if necessary, depends on the needed degree of accuracy.

An edge $e \in E$ from *u* to *v*, $u, v \in V$, is denoted by $u \xrightarrow{e} v$ and a path *p* starting from *u* and ending at *v* is indicated by the notation $u \overset{p}{\rightsquigarrow} v$. The number of dependency distances of path $p(d(p))$, $p = v_0 \xrightarrow{e_0} v_1 \xrightarrow{e_1} \cdots \xrightarrow{e_{k-1}} v_k$ is $d(p) = \sum_{i=0}^{k-1} d(e_i)$. As an example, Fig. 1b has the set of edges $E = \{A \xrightarrow{e_1} B,\ A \xrightarrow{e_2} C,\ A \xrightarrow{e_3} D,\ B \xrightarrow{e_4} D,\ C \xrightarrow{e_5} D,\ D \xrightarrow{e_7} A\}$. The number of dependency distances on each edge $e \in E$ is given by $d(e)$, where, for $i = 1,\ldots, 6$, $d(e_i) = 0$ and $d(e_7) = 2$.

The *execution order* or *execution pattern* of a PG are determined by the precedence relations in the graph. During one iteration of the graph each vertex in the execution order is computed exactly one time. Multiple iterations are identified by index *i*, starting from 0. Inter-iteration dependencies are represented by weighted edges or dependency distances. For any iteration *j*, an edge *e* from *u* to *v* with dependency distance $d(e)$ conveys that the computation of node *v* at iteration *j* depends on the execution of node *u* at iteration $j - d(e)$. An edge with no dependency distances represents a data dependency within the same iteration. A legal data flow graph must have strictly positive dependency distance cycles, i.e., the summation of the $d(e)$ along any cycle cannot be less than or equal to zero.

2.1 Retiming Overview

Retiming operations rearrange registers in a circuit or dependency distances in a data-flow graph in such a way that the behavior of the circuit is preserved while achieving a faster circuit. Traditionally, retiming [18] optimizes a synchronous circuit (graph) $G = \langle V, E, d, t\rangle$ which has non-probabilistic functional elements, i.e., each of the vertices $v \in V$ is associated with a fixed numerical timing value. The optimization goal is normally to reduce the *clock period* or *cycle period* $\Phi(G)$ (also known as longest path computation time). The cycle period represents the execution time of the longest path (referred to as the critical path) that has all zero dependency distance edges. It is defined by the equations

$\Phi(G) = \max\{t(p) : d(p) = 0\}$, where

$p = v_0 \xrightarrow{e_0} v_1 \xrightarrow{e_1} \cdots \xrightarrow{e_{k-1}} v_k$, $t(p) = \sum_{i=0}^{k} t(v_i)$, and $d(p) = \sum_{i=0}^{k-1} d(e_i)$.

Retiming of a graph $G = \langle V, E, d, t\rangle$ is a transformation function from vertices to the set of integers, $r : V \alpha Z$. The retiming function describes the movement of dependency distances with respect to the vertices so as to transform G into a new graph $G_r = \langle V, E, d_r, t\rangle$, where d_r represents the number of dependency distances on the edges of G_r. The positive (or negative) value of the retiming function determines the movement of the dependency distances. During retiming the same number of dependency distances is pushed from all incoming (outgoing) edges of a node to all outgoing (incoming) edges. If a single dependency distance is pushed from all incoming edges of node $u \in V$ to all outgoing edges of node u, then $r(u) = 1$. Conversely, if one dependency distance is pushed from all outgoing to all incoming edges of u, then $r(u) = -1$. The absolute value of the retiming function conveys the number of dependency distances that are pushed. An algorithm to find a set of retiming functions to minimize the clock period of the graph presented in [18] is a polynomial time algorithm which has the time complexity of $O(|V| |E| \log |V|)$.

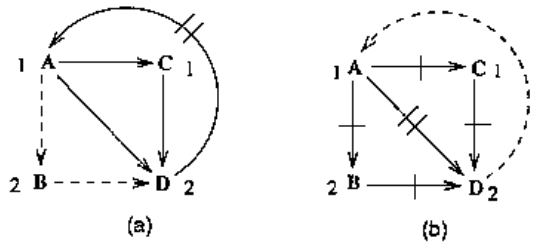


Figure 5. Illustrates retiming transformations ((a) before and (b) after retiming) where dotted edges represent the critical path.

Consider Fig. 5a, which illustrates a simple graph with four vertices, A, B, C and D. The numbers next to the vertices in the figure represent the required computation times. Fig. 5b represents a retimed version of Fig. 5a where $r(B) = r(C) = 1$, $r(A) = 2$, and $r(D) = 0$. In this case, the movement of dependency distances is as follows: $r(A) = 2$ is equivalent to removing two dependency distances from the incoming edge of vertex A, $D \xrightarrow{e} A$ and adding them onto edges $A \xrightarrow{e} B$, $A \xrightarrow{e} C$, and $A \xrightarrow{e} D$. The retiming functions for nodes C and B are $r(B) = r(C) = 1$. This means that one dependency distance from $A \xrightarrow{e} B$ is pushed through vertex B to edge $B \xrightarrow{e} D$. Similarly, one dependency distance from edge $A \xrightarrow{e} C$ is pushed through vertex C to $C \xrightarrow{e} D$. An equivalent set of retimings in Fig. 5b is $r(B) = r(C) = -1$, $r(D) = -2$, and $r(A) = 0$. This equivalent set of retimings produces the same graph by pushing the dependency distances backward through nodes D, B and C, instead of forward through nodes A, B and C. The dotted lines in Fig. 5a represent the critical path of the graph, for which $\Phi(G) = 5$. After retiming, the critical path becomes $\Phi(G) = 3$, as illustrated by the dotted line in Fig. 5b.

The following summarizes some essential properties of the retiming transformation:

1. r is a legal retiming if $d_r(e) \geq 0$, $\forall e \in E$.
2. For an edge $u \xrightarrow{e} v$, where $u, v \in V$, $d_r(e) = d_r(e) + r(u) - r(v)$.
3. For a path $u \overset{p}{\rightsquigarrow} v$, where $u, v \in V$, $d_r(p) = d_r(p) + r(u) - r(v)$.
4. In any directed cycle (l) of G and G_r, $d_r(l) = d(l) > 0$.

Property 1 guarantees that the retimed graph will not have any edge containing a negative number of dependency distances. Properties 2 and 3 explain the movement of such distances. If $r(v)$, $v \in V$, has a positive value, the distances will be deleted from the incoming edge(s) of v and inserted onto the outgoing edge(s), and vice versa if $r(v)$ has the negative value. Finally, Property 4 ensures that the number of dependency distances in any loop of the graph remains constant. That requires that all cycles have at least one dependency distance. Since retiming is an optimization technique which is subject to unlimited number of target resources, the resulting longest path computation time after the transformation is the underlying schedule length. Consider only a DAG part of the retimed graph where edges with nonzero dependency distances in the retimed graph are ignored. The iteration boundaries of this schedule will be at the root nodes (beginning of the iteration) and at the leaf nodes (end of the iteration).

2.2 Rotation Scheduling

In [5], Chao et al. proposed an algorithm, called *rotation scheduling*, which uses the retiming algorithm to deal with scheduling a cyclic DFG under resource constraints. The input to the rotation scheduling algorithm is a DFG and its corresponding static schedule, i.e., a synchronized order of the nodes in the DFG. Rotation scheduling reduces the schedule length (the number of control steps needed to execute one iteration of the schedule) by exploiting the parallelism between iterations. This is accomplished by shifting the scope of a static schedule in one iteration, called

the iteration window, down by one control step. Looking at a static iteration, rotation scheduling analogously rotates tasks from the top of the schedule of each iteration down to the end. This process is equivalent to retiming those tasks (nodes in the DFG) in which one dependency distance will be deleted from all their incoming edges and added to all their outgoing edges resulting in an intermediate retimed graph. Once the parallelism is extracted, the algorithm reassigns the rotated nodes to new positions so that the schedule length is shorter.

As an example, the cyclic DFG in Fig. 6a is to be scheduled using two processing elements. Fig. 6b presents one possible static schedule for such a graph. By using rotation scheduling, this schedule can be optimized. First, the algorithm uses node A from the next iteration. The original graph is retimed by $r(A) = 1$, i.e., one dependency distance from $E \xrightarrow{e} A$ is moved to all outgoing edges of A (see Fig. 6c). By doing so, node A now can be executed at any control step in this new iteration window. Assume that rotation scheduling uses a remapping strategy that places node A immediately after node C in PE_1. The resulting static schedule length is then reduced by one control step as shown in Fig. 6d. In Section 4, the concept of the schedule length and the remapping strategy will be extended to handle probabilistic inputs.

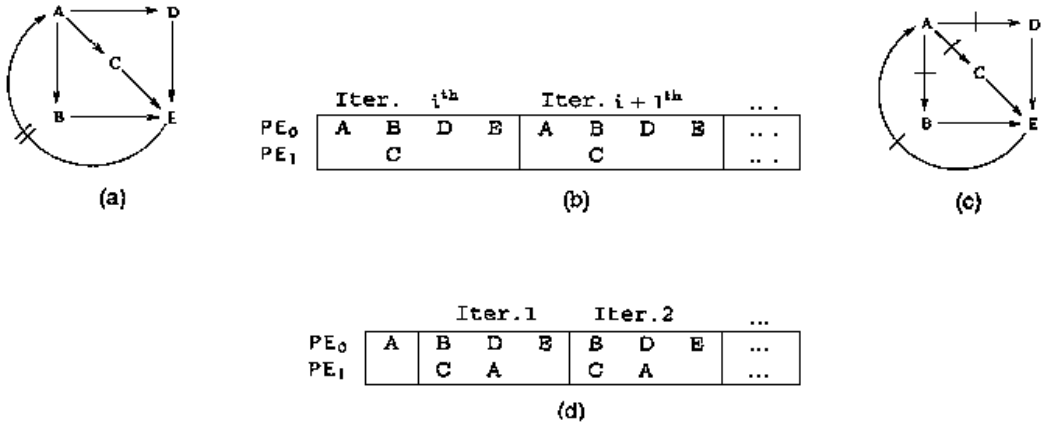


Figure 6. Illustrates an example to represent how rotation scheduling optimizes the underlying schedule length. (a) Cyclic DFG. (b) Static schedule. (c) Retimed. (d) Resulting schedule.

3. Nonresource Constrained Scheduling

Assuming there are infinite available resources, one can optimize a PG with respect to a desired longest path computation time and confidence level, i.e., attempt to reduce the longest path computation time of the graph. The distribution of dependency distances in the PG is done according to a probabilistic timing constraint where the *probability* of obtaining the timing result (longest path computation time) being less than or equal to a given value c is greater than some confidence probability value θ. This resulting timing information is essentially the schedule length of the nonresource constrained problem. This section presents an efficient algorithm for optimizing a probabilistic graph with respect to a desired computation time (c) and its corresponding confidence probability (θ). In order to evaluate the modified graph, we need to know the probability distribution associated with its computation time. The remaining subsections discuss these issues.

3.1 Computing Maximum Reaching Time

Let G_{dag} be the DAG portion (the subgraph that has only edges with no dependency distances) of a probabilistic graph G. Assume that two dummy nodes, v_s and v_d, are added to G_{dag}, where v_s connects to all source nodes (roots) and v_d is connected by all sink nodes (leaves). Traditionally, the longest path computation time of a graph is computed by maximizing the summation of computation times of nodes along the critical (longest) paths between these dummy nodes. Likewise, for a probabilistic graph, we can compute the summation of the computation time for each path from v_s to v_d in the graph. In this case the largest summation value is called the maximum reaching time or mrt of the graph. The mrt of a PG exhibits a possible longest path computation time of the graph and its associated probability. Therefore, unlike the traditional approach, the summation and maximum functions of computation time along the paths in a PG become functions of multiple random variables.

To compute an mrt of a PG, we need to modify the graph so that v_s and v_d are connected to the DAG portion of the original graph. Formally, a set of zero dependency distance edges is used to connect vertex v_s to all roots, and to connect all leaves to vertex v_d. Since it is nontrivial to efficiently compute a function of dependent random variables, Algorithm 1 computes the mrt(G) assuming that the random variables are independent. This algorithm traverses the input graph in the breadth-first fashion starting from v_s and ending at v_d. In general, the algorithm accumulates the probabilistic computation times along each traversed path. When it reaches a node that has more than one parent, all the values associated with its parents are maximized.

Algorithm 1 Calculate maximum reaching time of graph G
Require probabilistic graph PG

Ensure mrt(G) = temp$_{mrt}$ (v_s, v_d)

1. $G_0 = \langle V_0, E_0, d, T\rangle$ such that $V_0 = V + \{v_s, v_d\}$
2. $E_0 = E - \{e \in E : d(e) \neq 0\} + \{v_s \xrightarrow{e} v \in V_r, \quad u \in V_l \xrightarrow{e} v_d\}$
3. $\forall\ u \in V_0$, temp$_{mrt}(v_s, u) = 0$, $T_{v_s} = T_{v_d} = 0$, $Queue = v_s$
4. **while** $Queue \neq 0$ **do**
5. get node u from top of the $Queue$
6. temp$_{mrt}(v_s, u)$ = temp$_{mrt}(v_s, u) + T_u$
7. **for all** $u \xrightarrow{e} v$ **do**
8. decrement the incoming degree of node v by one
9. temp$_{mrt}(v_s, v)$ = max(temp$_{mrt}(v_s, u)$, temp$_{mrt}(v_s, v)$)
10. **if** incoming degree of node v becomes 0 **then**
11. insert node v into the $Queue$
12. **end if**
13. **end for**
14. **end while**

Lines 1 and 2 produce DAG G_0 from G containing only edges $e \in E$, with $d(e) = 0$, and the additional zero dependency distance edges connecting v_s to every root node $v \in V_r$ of G and connecting every leaf node $u \in V_l$ of G to v_d. Line 3 initializes the temp$_{mrt}(v_s, u)$ value for each vertex u in the new graph and sets the computation time of T_{v_s} and T_{v_d} to zero. Lines 4-12 traverse the graph in topological order and compute the mrt of each v with respect to v_s (temp$_{mrt}(v_s, v_d)$). Note that the temp$_{mrt}$ for node v with respect to v_s is originally set to zero. It stores the current maximum computation time of all node v's visited parents. When the first parent of v is dequeued, v has its indegree reduced by one (Line 8) and temp$_{mrt}$ mrt is updated (Line 9). Vertex v's other parents are in turn dequeued, and the process is repeated. Eventually, the last parent of node v will be dequeued and maximized. At this point, node v will be inserted into the queue since all parents have been considered, i.e., indegree of v equals zero (Line 10). Node v will be eventually dequeued by Line 5. Line 6 will then add T_v to the temp$_{mrt}$ of node v producing the final mrt with respect to all paths reaching node v.

Note that the initial computation times are integers and the probabilities associated with these times being greater than the given value c are accumulated as one value in the algorithm. Only O(c+1) values need to be stored for each vertex. Therefore, the time complexity for calculating the summation (Line 6), or the maximum (Line 9) of two vertices is O(c^2). Since the algorithm computes the result in a breadth first fashion, the running time of Algorithm 1 is O ($c^2|V|\ |E|$), while the space complexity is bounded by O(c$|V|$).

3.2 Probabilistic Retiming

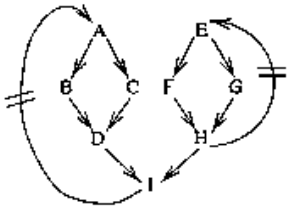
Using the concept of mrt, Algorithm 2 presents the probabilistic retiming algorithm which reduces the longest path computation time of the given PG to meet a timing constraint. Such a constraint is that Pr(mrt(v_s, v_d) < c) > θ where c is the desired longest path computation time of the graph and θ is the confidence probability. This requirement can be rewritten as Pr(mrt(v_s, v_d) < c) ≤ δ, where δ is 1 - θ. The algorithm retimes vertices whose probability of computation time being greater than c is larger than the acceptable probability value. Initially, the retiming value for each node is set to zero and nonzero dependency distance edges are eliminated. Then, v_s is connected to the root-vertices of the resulting DAG and v_d is connected by the leaf-vertices of the DAG. Lines7-17 traverse the DAG in a breath-first search manner and update the temp$_{mrt}$ for each node as in Algorithm 1. After updating a vertex, the resulting temp$_{mrt}$ is tested to see if the requirement, Pr(temp$_{mrt}(G) > c$) ≤ δ, is met. Line 19, then decreases the retiming value of any vertex v that violates the requirement unless the vertex has previously been retimed in a current iteration. The algorithm then repeats the above process using the retimed graph obtained from the previous iteration. If the algorithm finds the solution for a given clock period, the final retimed graph implies the number of required resources to achieve such a schedule length.

Algorithm 2 Probabilistic retiming
Require probabilistic graph, a requirement Pr(temp$_{mrt}(G) > c$) ≤ δ
Ensure retiming function r for each node to meet the requirement

1. $\forall$ node $v \in V$ initialize retiming function $r(v)$ to 0
2. **for** i = 1 to $|V|$ **do**
3. retime graph G_r with the retiming function $r(v)$
4. G_0 = directed acyclic portion (DAG) of G_r
5. prepend dummy node v_s to G_0 {connects to all root nodes}
6. append dummy node v_d to G_0 {connected by all leaf nodes}
7. **for all** nodes in G_0 **do**
8. temp$_{mrt}(v_s, u)$ = 0
9. insert v_s into $Queue$
10. $T_{v_s} = T_{v_d} = 0$ {set timing of two dummies to zero}
11. **end for**
12. **while** $Queue \neq 0$
13. get node u from the $Queue$
14. temp$_{mrt}(v_s, u)$ = temp$_{mrt}(v_s, u) + T_u$ {adding two random variables}
15. **for all** $u \xrightarrow{e} v \in G_0$ **do**
16. decrement number of incoming degrees of node v by one
17. temp$_{mrt}(v_s, v)$ = max(temp$_{mrt}(v_s, u)$, temp$_{mrt}(v_s, v)$) {maximizing two random variables}
18. **if** Pr(temp$_{mrt}(v_s, v) > c) > \delta$ **and** u has not been retimed **then**
19. $r(u) = r(u)$ -1 {move one dependency distances from all outgoing edges to all incoming edges}
20. **end if**
21. **if** number of incoming edges of node v is 0 **then**
22. insert node v into a ready Queue
23. **end if**
24. **end for**
25. **end while**
26. **end for**

Line 19 pushes a dependency distance onto all incoming edges of a node that violates the timing constraint. Since all descendents of this node will also be retimed, Line 19 in essence moves a dependency distance from below v_d to above this node. In other words, $r(u) = r(u) - 1$ for all nodes from u to v_d. Hence only the incoming edges of vertex u will have an additional dependency distance. Once all nodes are not retimed in the current iteration, the requirement $\Pr(\text{mrt}(v_s, v_d) > c) \leq \delta$ is met. Then the algorithm stops and reports the resulting retiming functions associated with nodes in the graph. Otherwise, the algorithm repeats at most $|V|$ times. Since the computation of the maximum reaching time is performed in every iteration, the time complexity of this algorithm is $O(c^2|V|^2|E|)$ while the space complexity remains the same as in the maximum reaching time algorithm. The resulting retiming function returned by Algorithm 2 guarantees (necessary condition) the following:

Theorem 3.1. *Given $G = \langle V, E, d, T\rangle$, a desired cycle period, c, and a confidence probability, $\theta = 1 - \delta$, if Algorithm 2 (probabilistic retiming algorithm) finds a solution, then the resulting retimed graph G_r satisfies the requirement* $\Pr(\text{mrt}(G) \leq c) \geq \theta$.

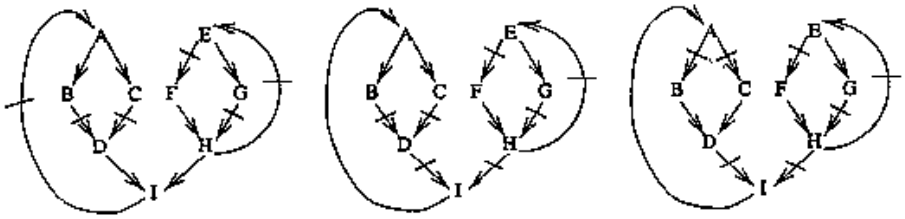


(a)

Time	Nodes								
	A	B	C	D	E	F	G	H	I
1	0.3	0	0	0.9	0	0.5	0.9	0.5	0
2	0.7	0.8	0	0.1	0.5	0	0	0	0.5
3	0	0	0.5	0	0	0.5	0	0.5	0.5
4	0	0	0.5	0	0.5	0	0.1	0	0
5	0	0.2	0	0	0	0	0	0	0

(b)

Figure 7. Illustrates an example of a 9-node graph and its corresponding probabilistic timing information.



(a) (b) (c)

Figure 8. Illustrates retimed graph corresponding to Table 1, 2, and 3.

Table 1. Shows first iteration showing probability distribution of mrt(v_s, v), $v \in V$

$v \in V$	Pr(mrt(v_s, v)) for different time							$r(v)$
	1	2	3	4	5	6	> c	
A	0.3	0.7	0	0	0	0	0	0
B	0	0	0.24	0.56	0	0.06	0.14	0
C	0	0	0	0.15	0.5	0.35	0	0
D	0	0	0	0	0.108	0.372	0.520	-1
E	0	0.5	0	0.5	0	0	0	0
F	0	0	0.250	0	0.5	0	0.25	-1
G	0	0	0.45	0	0.45	0.05	0.050	0
H	0	0	0	0.056	0	0.338	0.606	-1
I	0	0	0	0	0	0	1	-1
v_d	0	0	0	0	0	0	1	0

Table 2. Shows second iteration showing probability distribution of mrt(v_s, v), $v \in V$

$v \in V$	Pr(mrt(v_s, v)) for different time							$r(v)$
	1	2	3	4	5	6	> c	
A	0.3	0.7	0	0	0	0	0	0
B	0	0	0.24	0.56	0	0.06	0.14	0
C	0	0	0	0.15	0.5	0.35	0	0
D	0.9	0.1	0	0	0	0	0	-1
E	0	0.5	0	0.5	0	0	0	0
F	0.5	0	0.5	0	0	0	0	-1
G	0.9	0	0	0.1	0	0	0	0
H	0	0.225	0	0.45	0.05	0.225	0.5	-1
I	0	0	0	0.112	0.112	0.225	0.550	-2
v_d	0	0	0	0.013	0.103	0.27	0.613	0

Table 3. Shows third iteration showing probability distribution of mrt(v_s, v), $v \in V$

$v \in V$	Pr(mrt(v_s, v)) for different time							$r(v)$
	1	2	3	4	5	6	> c	
A	0	0	0.15	0.5	0.35	0	0	0
B	0	0	0	0	0.12	0.4	0.48	-1
C	0	0	0	0	0	0.075	0.925	-1
D	0.9	0.1	0	0	0	0	0	-1
E	0	0.5	0	0.5	0	0	0	0
F	0.5	0	0.5	0	0	0	0	-1
G	0.9	0	0	0.1	0	0	0	0
H	0	0.225	0	0.45	0.05	0.225	0.05	-1
I	0	0.5	0.5	0	0	0	0	-2
v_d	0	0	0	0	0	0.037	0.963	0

3.3 Example

Consider the PG and the probability distribution associated with nodes in the graph in Fig.7. For this experiment, let c = 6 be the desired longest path computation time and $\delta = 0.2$ be the acceptable probability. Algorithm 2 works by first checking and computing mrt from v_s to A and E. Then, it topologically calculates the mrt of the adjacent nodes of A and E. After it computes the mrt of node I, mrt(v_s, v_d) is obtained.

Three iterations of Algorithm 2 computing the results of the maximum reaching time from v_s to v including v_d are tabulated in Tables 1, 2 and 3. After the first iteration, the retiming value associated with nodes D, F, H, and I are shown in Column $r(v)$ of Table 1. The values in Columns 2-8 show the probability that the mrt(v_s,v), $\forall v \in V$, ranges from 1 to 6 and greater than 6 (>6), respectively. The retimed graph associated with the retiming value in Table 1 after the first iteration is presented in Fig. 8a. Table 2 presents the maximum reaching time from the dummy node v_s to each node $v \in V$ as well as the retiming function for each vertex after the second iteration. Fig. 8b presents the retimed graph corresponding to the retiming function presented in Table 2. By computing the mrt(v_s,v) of the retimed graph in Fig. 8b, it becomes apparent that nodes B and C need to be retimed. Fig. 8c illustrates the final retimed graph in accordance with the retiming function presented in Table 3. Note that Table 3 also presents the final maximum reaching time and retiming value for each vertex which satisfies the required configuration. From this final retimed graph, one could, therefore, allocate a minimum of five processing elements in order to compute the graph in six time units with 80 percent confidence.

4. Resource-Constrained Scheduling

In this section, we present a probabilistic scheduling algorithm which considers the limited number of resources. The traditional rotation scheduling framework is extended to handle the probabilistic environment. We call this algorithm *probabilistic rotation scheduling* (PRS). Given a PG, the algorithm iteratively optimizes the PG with respect to the confidence probability and the number of resources.

Before presenting this algorithm, we first discuss two important concepts that make scheduling under the probabilistic environment different from traditional scheduling problems. First, in the probabilistic model, a synchronization control step is not available. A node can begin its execution if all of its parents have already been executed. This is similar to the asynchronous model where data request and handshaking signals are used to communicate between nodes. The schedule can be viewed as a directed graph where edges show either the data requirement to execute a node or the order that a node can be executed in a particular functional unit. Note that a synchronization will be applied at the end of each iteration. Second, the task remapping strategy for PRS should take the probabilistic nature of the problem into account. The following subsection discuss these concepts in more details.

4.1 Schedule Length Subject to the Confidence

The concept of mrt can be used to compute the underlying schedule length. Hence, the conventional way of calculating schedule length has to be redefined to include the mrt notion. In order to do so, we update the probabilistic data flow graph by adding the resource information and extra edges between two nodes executed consecutively in the same functional unit and have no data dependencies between them. This graph, called the *probabilistic task-assignment graph* (PTG), represents a schedule under the probabilistic model.

Definition 4.1. *A probabilistic task-assignment graph* (*PTG*) $G = \langle V, E, w, T, b \rangle$, *is a vertex-weighted, edge-weighted, directed acyclic graph, where V is the set of vertices representing tasks, E is the set of edges representing the data dependencies between vertices, w is a edge-type function from* $e \in E$ *to* $\{0,1\}$, *where* 0 *represents the type of dependency edge and* 1 *represents the type of flow-control edge,* T_v *is a random variable representing the computation time of a node* $v \in V$, *and b is a processor binding function from* $v \in V$ *to* $\{PE_i, 1 \le i \le n\}$, *where* PE_i *is processing element i and n is the total number of processing elements*.

Figure 9. Illustrates an example of a probabilistic task-assignment graph (PTG) where the nodes are assigned to PE0 and PE_1.

As an example, Fig. 9 shows an example of the PTG with two functional units PE_0 and PE_1. Nodes B and D are assigned to PE_0. That is, $b(B_0 = b(D) = PE_0$. Meanwhile, $b(C) = b(A) = PE_1$. Edges consists of $C \xrightarrow{e_1} A$, $C \xrightarrow{e_2} D$, $B \xrightarrow{e_3} D$, where $w(e_1) = 1$ and $w(e_2) = w(e_3) = 0$. Note here that if there exist edges $A \rightarrow B$ and $B \rightarrow D$ and all of the nodes are scheduled to the same processor, edge $A \rightarrow D$, which was a true dependence edge, can be ignored. Note also that removing redundancy edges is simple and should be utilized to speed up the calculation of mrt. In Fig. 9, edge $C \xrightarrow{e_1} A$ is control-typed since A now has no dependency to C but has to execute after C due to resource constraints. Other edges represent data dependencies. Applying the mrt algorithm to the PTG, we can define the *probabilistic schedule length*. This length is expressed in terms of confidence probabilistic as follows.

Definition 4.2. *A probabilistic schedule length of PTG* $G = \langle V, E, w, T, b \rangle$ *with respect to a confidence level* θ, $\text{psl}(G, \theta)$, *is the smallest computation time c such that* $\Pr(\text{mrt}(G) > c) < 1 - \theta$.

For example, consider the probability distribution of the mrt (G) shown in table 4. Given a confidence probability $\theta = 0.8$, the probabilistic schedule length $\text{psl}(G, 0.8)$ is 14. This because the smallest possible computation time is 14, where $\Pr(\text{mrt}(G) > 14) < 0.2$, i.e., $0.04365 + 0.02293 + 0.00875 = 0.07818 < 0.2$. Therefore, with 80 percent confidence, the computation time of G is less than 14.

4.2 Task Remapping Heuristic: Template Scheduling

In this section, we propose a heuristic, called *template scheduling* (TS), to search for a place to reschedule a task. This remapping phase plays an important role in reducing the probabilistic schedule length in PRS. Since the computation time is a random variable, there is no fixed control step within an iteration. As long as a node is placed after its parents, any scheduling location is legal.

In template scheduling, a *schedule template* is computed using the expectation of the computation time of each node. This template implies not only the execution order, but also the expected control step that a node can start execution. In order to determine an expected control step, each node in a PTG is visited in the topological order and the following is computed:

Definition 4.3. *The expected control step of node v of PTG* $G = \langle V, E, w, T, b \rangle$, $\text{Ecs}(v)$, *is computed by*

$$\text{Ecs}(v) = \max(\text{Ecs}(u_i) + ET_{u_i}),$$

where $u_i \xrightarrow{e} v \in E$, ET_u *represents the expected computation time of node u and* $\text{Ecs}(v_i) = 0$ *for all root nodes* $v_i \in V$.

This definition assumes node v can start its execution right after all parents finish their execution. By observing this template, one can ascertain how long (the number of control steps) each processing element would be idle. The template scheduling decides where to reschedule a node using "their degree of flexibility."

Table 4. Shows possible computation time of the mrt(G)

	8	9	10	11	12	13	14	15	16
Prob	0.00197	0.04373	0.20902	0.25140	0.23661	0.18194	0.04365	0.02293	0.00875

Definition 4. *Given a PTG $G = \langle V, E, w, T, b\rangle$, a degree of flexibility of node u with respect to the processing element PE_i,* dflex(*u, i*), *is computed by*:

$$\text{dflex}(u, i) = \text{Ecs}(v) - \text{Ecs}(u) - ET_u,$$

where $u \xrightarrow{e} v \in E$ *and u and v are assigned to* PE_i.

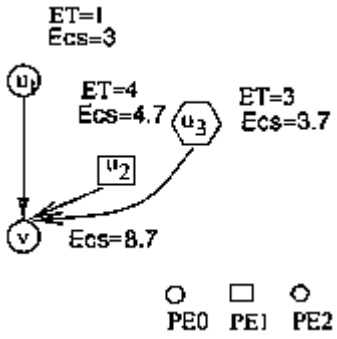


Figure 10. Illustrates an example of how to obtain the expected control step.

The degree of flexibility conveys the expected size of available time slot within PE_i. Fig. 10 shows a typical case where node *v* has more than one parent. u_1, u_2 and u_3 are parents of node *v* and each of these parents has the expected computation time 1, 4, and 3, respectively. In the same order, the expected control steps of these nodes are 3, 4.7, and 3.7, respectively. Therefore, the expected control step Ecs(*v*) = 8.7. According to Definition 4.3, the degree of flexibility of *u* with respect to PE_0, is 8.7 – 3 – 1 = 4.7. This value conveys how long PE_0 has to wait before *v* can be executed. Note that the degree of flexibility of a node, which is executed at last in any *PE*, is undefined. The following steps compute the new *G* after rescheduling node *v*.

Algorithm 3 Rescheduling rotated nodes using the template scheduling heuristic
Require: PTG, rescheduled node *v* and confidence probability θ
Ensure: Rescheduling new PTG with shortest psl

1. Assume that all nodes in the PTG have their expected computation times precomputed
2. ∀ node $u \in V$ compute Ecs(*u*) and dflex(*u*)
3. **for** each of target processors (PE_i) **do**
4. Using the maximum dflex to select node *x* which is scheduled to PE_i
5. schedule *v* after *x*
6. reconstruct a new PTG (assigned with PE_i) with this assignment
7. compare it with others PTG and get the one that has the best psl
8. **end for**

This rescheduling policy hopes that placing a node in the processor with the expected biggest idle time slot results in the least potential of increasing the total execution time. If a computation time of the node is much smaller than the expected time slot, this approach may allow the next rescheduled node to be placed here also. This is similar to worst-fit policy, where the scheduler strives to schedule a node to biggest slot. In section 5, we demonstrate the effectiveness of this heuristic over the method that exhaustively finds the best place for a node. Note that this exhaustive search is not performed globally, rather the search is done locally in each remapping iteration. We call this heuristic a *local search* (LS).

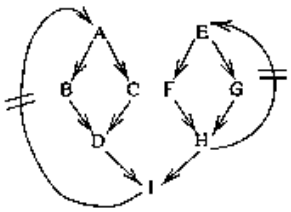
4.3 Rotation Phase

Having discussed the rescheduling heuristic, the following presents the probabilistic rotation scheduling (PRS). Note that the previous heuristic or any rescheduling heuristic can be used as rescheduling part of this PRS algorithm. The experiments in Section 5 show the efficacy of the PRS framework with different rescheduling heuristics.

Algorithm 4 Probabilistic Rotation Scheduling
Require: PG and designer's confidence probability θ
Ensure: PTG with a shortest psl

1. $\forall u \in V$ compute ET_u
2. $G_s \Leftarrow$ find_initial_schedule {finding an initial schedule for PG and keep it in G_s}
3. **for** $i = 1$ to $2|V|$ **do**
4. $R \Leftarrow$ all roots of a DAG portion of G_s {these are nodes to be rotated}
5. retime each of the nodes in *R*
6. reschedule these nodes one by one using the heuristic previously presented
7. compute psl of the new graph with respect to θ
8. **if** psl(G_s, θ) ≤ psl (G_{best}, θ) **then**
9. $G_{best} \Leftarrow G_s$ {considering that G_{best} is initialized to G_s first}
10. **end if**
11. **end for**

In order to use template scheduling, an expected computation time of each task will be precomputed. After that, an initial schedule is constructed by find_initial_schedule. Note that the algorithm for creating the initial schedule can be any DAG scheduling, e.g., probabilistic list scheduling discussed previously. Rotation scheduling loops for $2|V|$ times to reschedule all nodes in the graph at least once. Like traditional rotation scheduling, only nodes that have all their incoming edges with nonzero dependency distances will be drawn from each of these edges and placed on their outgoing edges. Then, these rotated nodes will be rescheduled one by one using the template scheduling technique. After all rotated nodes are scheduled, if the resulting PTG is better than the current one, Algorithm 4 will save the better PTG.

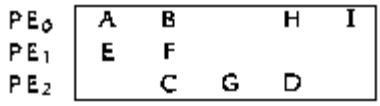


(a)

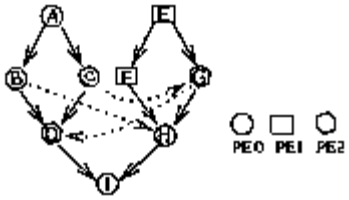
Time	Nodes								
	A	B	C	D	E	F	G	H	I
1	0.3	0	0	0.9	0	0.5	0.9	0.5	0
2	0.7	0.8	0	0.1	0.5	0	0	0	0.5
3	0	0	0.5	0	0	0.5	0	0.5	0.5
4	0	0	0.5	0	0.5	0	0.1	0	0
5	0	0.2	0	0	0	0	0	0	0

(b)

Figure 11. Illustrates an example of the computation time of graph in Figure 1b.

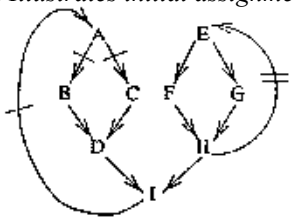


(a)



(b)

Figure 12. Illustrates initial assignment and the corresponding execution order. (a) Static execution order. (b) PTG.



(a)

	A	B	C	D	E	F	G	H	I
ET_v	1.7	2.6	3.5	1.1	3.0	2.0	1.3	2.0	2.5
Ecs	-	1.7	1.7	6.5	0.0	3.0	5.2	6.5	8.5
dflex	0.0	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0

(b)

Figure 13. Illustrates the probabilistic graph after A is rotated and the template values. (a) New PTG. (b) Ecs and dflex.

Table 5. Shows possible computation time of the mrt of a PTG

	8	9	10	11	12	13	14	15	16
Prob	0.00197	0.04373	0.20902	0.25140	0.23661	0.18194	0.04365	0.02293	0.00875

4.4 Example

Let us revisited the PG example in Section 3.3 as shown in Fig. 11a and the corresponding computation time in Fig. 11b. The confidence probability is given as θ = 0.8. After list scheduling is applied, the initial execution order is constructed as shown in Fig. 12a. The corresponding PTG is presented in Fig. 12b. Nodes *A*, *B*, *H* and *I* are assigned to PE_0, nodes *E* and *F* are scheduled PE_1 and nodes *C*, *G* and *D* are assigned to PE_2. Edges $B \xrightarrow{e} H$ and $C \xrightarrow{e} G$ and $G \xrightarrow{e} D$ are flow-control edges.

For this assignment, the mrt of such a PTG is computed as shown in Table 5. Therefore, with higher than 80 percent confidence probability, psl(*G*, 0.8) = 14.

According to the structure of the PTG, either *A* or *E* can be rescheduled. In the first rotation, PRS selects *A* to be rescheduled. One dependency distance is moved from all

incoming edges of *A* and pushed to all outgoing edges of *A*. The resulting retimed graph PG is shown in Fig. 13a. In this graph, node *A* requires no direct data dependency from any node. Therefore, *A* can be placed at any position in the schedule. Fig. 13b shows the expected computation time, the expected control step, and the degree of flexibility of each node in this PTG.

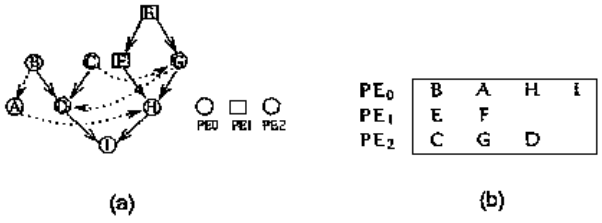


Figure 14. Illustrates the PTG, execution order and its mrt after the first rotation where psl (G, 0.8) = 12. (a) PTG. (b) Execution order.

5. Experimental Results

In this section we perform experiments both using nonresource and resource constrained scheduling on two general classes of problems. The first class are real applications which may have a combination of nodes with probabilistic computation times and with fixed computation times. The second are well-known DSP filter benchmarks. Since these benchmarks contain two uniform types of nodes, namely multiplication and addition, the basic timing information consisting of three probability distribution are assigned to each benchmark graphs. In order to show the usability of the proposed algorithm, three applications are profiled to get their probabilistic timing information. The profiler reports the processing time requirement in these applications and the corresponding frequency of this time value. The frequency of timing occurences is used to obtain a node probability distributions. A node in these graphs may represent a large number of operations which cause the uncertain computation time as well as operations which have fixed timing information. Each timing information is discretized to a smaller unit such as nanoseconds.

Based on the values in the table from Fig. 13b, it is obvious that an expected waiting time between *B* and *H* in PE_0 where psl can be reduced. The resulting PTG and its execution order are shown in Fig. 14, where psl(*G*, 0.8) = 12. After running PRS for 18 iterations, the shortest possible schedule length was found in the 15th iteration. In Fig. 15, we present the resulting schedule length of the this trial which is less than 9 with probability greater than 80 percent (psl(*G*, 0.8) = 9).

The DSP filter benchmarks used in these experiments include a Biquadratic IIR filter, a 3-stage IIR filter, a fourth-order Jaunmann wave digital filter, a fifth-order elliptic filter, an unfolded fifth-order elliptic filter with an unfolding factor equal to 4 (uf = 4), an all-pole lattice filter, an unfolded all-pole lattice filter (uf = 2), an unfolded all-pole lattice filter (uf = 6), a differential equation solver and a Volterra filter. The rest of the benchmarks are the application for image processing (Floyd-Steinberg), the application to search for a solution which maximize some unknown function by using genetic algorithm, and the famous example of the application in the fuzzy logic area, the inverted pendulum problem. All of the experiments were performed using SUN UltraSparc.

Table 6. Shows probabilistic retiming versus worst case traditional retimimg

Benchmark	#nodes	c worst	θ = 0.9		θ = 0.8	
			c	%	c	%
Biquad IIR	8	78	60	23	57	26
Diff. Equation	11	118	81	31	77	35
3-stage direct IIR	12	54	44	19	41	24
All-poll Lattice	15	157	120	24	117	25
4^{th} order WDF	17	156	116	26	112	28
Volterra	27	276	216	22	212	23
5^{th} Elliptic	34	330	240	28	236	29
All-poll Lattice (uf=2)	45	468	350	25	346	25
All-poll Lattice (uf=6)	105	1092	811	26	806	26
5^{th} Elliptic (uf=4)	170	1633	1185	27	1174	28
Floyd-Steinberg	13	30	23	23	21	30
Genetic application	18	202	180	11	127	37
Fuzzy application	24	19	17	11	17	11

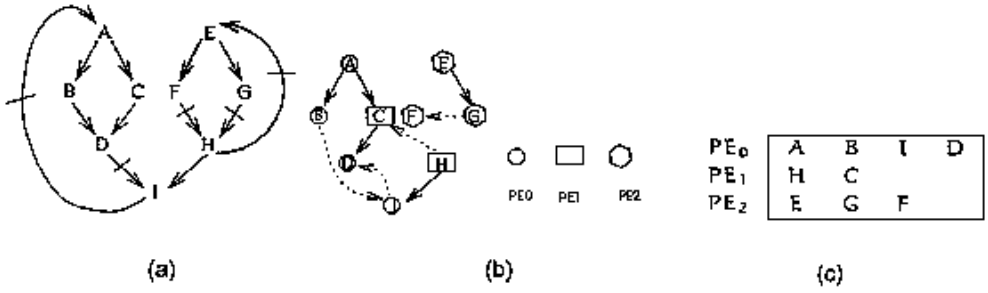


Figure 15. Illustrates the final PG, PTG, execution order where psl(G, 0.8) = 9. (a) PG. (b) PTG. (c) Final execution order.

Table 7. Shows probabilistic retiming versus average-case analysis

Benchmark	Traditional	Algorithm 2	
	mrt(G^r_{avg})	θ = 0.9	θ = 0.8
Biquad IIR	70.40	52.64	52.30
Diff. Equation	76.05	73.07	72.50
3-stage direct IIR	41.90	37.70	38.36
All-poll Lattice	114.45	111.77	111.40
4th order WDF	106.73	106.44	105.98
Volterra	204.00	202.44	202.00
5th Elliptic	233.30	228.41	227.59
All-poll Lattice (uf=2)	342.17	338.11	337.62
All-poll Lattice (uf=6)	800.51	794.02	793.39
5th Elliptic (uf=4)	800.51	794.02	793.39
Floyd-Steinberg	18.47	18.45	18.17
Genetic application	150.89	144.01	112.46
Fuzzy application	18.03	16.08	16.08

5.1 Nonresource Constrained Experiments

In each experiment, for a given confidence level θ = 1 - δ, Algorithm 2 is used to search for the best longest path computation time. In order to do this, the current desired longest path computation time (*c*) is varied based on whether or not a feasible solution is found. For instance, if *c* is too small, the algorithm will report that no feasible solution exists. In this case, *c* is increased and Algorithm 2 is reapplied. This process will repeated, until the smallest feasible *c* is found.

Table 6 shows the results for traditional retiming using worst-case computation time assumptions (column *c* worst) and the probabilistic model with two high confidence probabilities (θ = 0.9 and 0.8). The average running time for these experiments was determined to be below 60 seconds including the input/output interfaces. The algorithms are implemented in a straightforward way where array is used to store probability distributions. Column 3 in the table presents the optimized longest path computation times obtained from applying traditional retiming using the worst-case computation time for each node in the graph benchmarks. For columns where θ = 0.9 and θ = 0.8, the probabilistic retiming algorithm is applied to the benchmarks (*G*) while each of these confidence probabilities is used as its input. The numbers show in both columns are the given *c* where Pr(mrt (*G*) ≤ *c*) ≥ θ. The value *c* from this requirement is the smallest input value which Algorithm 2 can find a solution to satisfy such a requirement. Notice that for all benchmarks the longest path computation time with θ = 0.9 are still smaller than the computation time in Column 3. In order to quantify the improvement of the probabilistic retiming algorithm, the "%" columns list the percent of computation time reductions with respect to the value from Column 3.

Table 7 compares the probabilistic retiming algorithm to the traditional retiming algorithm with average computation times used for each node in the graphs. First, the probabilistic nodes of each input graph are converted to fixed time nodes resulting in G_{avg}, i.e., each node assumes its average computation time rather than probabilistic computation time. Traditional retiming is then applied to the resulting graph, resulting in graph G^r_{avg}. The purpose of this table is to compare G^r_{avg} (obtained from running traditional retiming on G^r_{avg}) with retimed PGs. In order to compare with the results produced by the proposed algorithm, the placement of dependency distance in each G^r_{avg} is preserved while the original probabilistic computation times are replaced with the average computation times. Put another way, we transformed each G^r_{avg} back to a probabilistic graph. Algorithm 1 is then used to evaluate these graphs while only the *expected values* of each result are shown in the table. Columns 4 and 5 present the expected values of the results obtained from running probabilistic retiming on each PG where the confidence probability of 0.9 and 0.8 are considered. Note that these results are consistently better (smaller value) than the results obtained from running traditional retiming on each of G^r_{avg}. Hence, the approach of using the expected values

for each node is neither a good heuristic in the initial design phase nor does it give any quantitative confidence to the resulting graphs.

5.2 Resource-Constrained Experiments

We tested the probabilistic rotation scheduling (PRS) algorithm on the selected filter and application benchmarks: the fifth elliptic filter, 3 stage-IIR filter, Volterra filter, Lattice filter, and Floyd-Steinberg, Genetic algorithm, Fuzzy logic applications. Table 8 demonstrates the effectiveness of our approach on both 2-adder, 1-multiplier and 2-adder, 2-multiplier systems for those filter benchmarks. The specification of 3 and 4 general purpose processors (PEs) are adopted for the other three application benchmarks. The performance of PRS is evaluated by comparing the resulting schedule length with the schedule length obtained from the modified list scheduling technique (capable of handling the probabilistic graphs). We also show the effectiveness of template scheduling (TS) by comparing its results with other heuristics, namely, local search (LS), and as-late-as possible scheduling (AL). The average execution times of AL and TS are very comparable (about 12 seconds running on UltraSparc) while LS takes much longer time and does not give the outstanding results comparing with those from TS.

In each rescheduling phase of PRS, the LS approach strives to reschedule a node to all possible legal location (local search) and returns the assignment which yields the minimum psl(G, θ). This method is simple and gives a good schedule; however, it is time consuming and not practical to try all possible scheduling places in every iteration of PRS. Furthermore, a PTG needs to be temporarily updated in every trial in order to compute the possible schedule length. On the contrary, the AL method reduces the number of trials by attempting to schedule a task only once at the farthest legal position in each functional unit or processor while the TS heuristic re-maps the scheduled node after the node with the highest degree of flexibility in each functional unit.

Table 8. Shows comparison of the results obtained from applying benchmarks modified list and probabilistic rotation scheduling (using different remapping heuristics)

Spec.	Benchmarks	#nodes	θ = 0.9				θ = 0.8			
			PL	PRS			PL	PRS		
				AL	LS	TS		AL	LS	TS
2 Adds. 1 Mul.	Diff. Equation	11	169	152	133	**133**	165	147	131	**131**
	3-stage direct IIR	12	188	184	151	**151**	184	179	147	**147**
	All-poll Lattice	15	229	225	142	**141**	225	220	138	**138**
	Volterra	27	526	468	361	**361**	519	461	354	**354**
	5th Elliptic	34	318	298	293	**293**	314	294	289	**289**
3 PEs	Floyd-Steinberg	13	42	32	**27**	30	38	28	28	**27**
	Genetic application	18	434	295	**216**	275	438	180	150	**150**
	Fuzzy application	24	52	46	45	**45**	52	45	43	**43**
2 Adds. 2 Mul.	Diff. Equation	11	120	103	**83**	90	117	100	**83**	91
	3-stage direct IIR	12	124	120	87	**87**	120	110	83	**82**
	All-poll Lattice	15	229	225	140	**139**	225	220	136	**136**
	Volterra	27	359	270	**237**	259	353	265	**221**	256
	5th Elliptic	34	288	288	274	**271**	284	274	270	**267**
4 PEs	Floyd-Steinberg	13	42	30	**26**	28	38	24	24	**24**
	Genetic application	18	434	291	**205**	275	337	180	180	**180**
	Fuzzy application	24	49	41	**38**	40	47	38	**35**	36

Table 9. Shows comparing probabilistic rotation with traditional rotation running on graphs with average computation times

Spec.	Benchmarks	#nodes	worst case		θ = 0.9			θ = 0.8		
			L	R	PL	PRS	AVG	PL	PRS	AVG
2 Adds. 1 Mul.	Diff. Equation	11	228	180	169	**133**	136	165	**131**	131
	3-stage direct IIR	12	252	204	188	**151**	163	184	**147**	179
	All-poll Lattice	15	312	204	229	**141**	153	225	**138**	149
	Volterra	27	750	510	526	**361**	526	519	**354**	519

[6] L. Chao and E. Sha, "Static Scheduling for Synthesis of DSP Algorithms on Various Models," *J. VLSI Signal Processing*, pp. 207-223, Oct. 1995.

[7] A.E. Eichenberger and E.S. Davidson, "Stage Scheduling: A Technique to Reduce the Register Requirements of a Modulo Schedule," *Proc. 28th Int'l Symp. Microarchitecture*, pp. 338-349, Nov. 1995.

[8] A.E. Eichenberger, E.S. Davidson, and S.G. Abraham, "Minimum Register Requirements for a Modulo Schedule," *Proc. 27th Int'l Symp. Microarchitecture*, pp. 75-84, Nov. 1994.

[9] J.A. Fisher, "Trace Scheduling: A Technique for Global Microcode Compaction," *IEEE Trans. Computers*, vol. 30, no. 7, pp. 478-490, July 1981.

[10] I. Foster, *Designing and Building Parallel Program: Concepts and Tools for Parallel Software Engineering*. Addison-Wesley, 1994.

[11] R.A. Kamin, G.B. Adams, and P.K. Dubey, "Dynamic List-Scheduling with Finite Resources," Proc. *1994 Int'l Conf. Computer Design*, pp. 140-144, Oct. 1994.

[12] I. Karkowski and R.H.J.M. Otten, "Retiming Synchronous Circuitry with Imprecise Delays," *Proc. 32nd Design Automation Conf.*, pp. 322-326, 1995.

[13] A.A. Khan, C.L. McCreary, and M.S. Jones, "A Comparison of Multiprocessor Scheduling Heuristic," *Proc. 1994 Int'l Conf. Parallel Processing*, vol. II, pp. 243-250, 1994.

[14] D. Ku and G. De Micheli, *High-Level Synthesis of ASICS under Timing and Synchronization Constraints*. Kluwer Academic, 1992.

[15] D. Ku and G. De Micheli, "Relative Scheduling under Timing Constraints: Algorithm for High-Level Synthesis," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, pp. 697-718, June 1992.

[16] M. Lam, "Software Pipelining," *Proc. ACM SIGPLAN '88 Conf. Programming Language Design and Implementation*, pp. 318-328, June 1988.

[17] D.M. Lavery and W.W. Hwu, "Unrolling-Based Optimization for Modulo Scheduling," *Proc. 28th Int'l Symp. Microarchitecture*, pp. 327-337, Nov. 1995.

[18] C.E. Leiserson and J.B. Saxe, "Retiming Synchronous Circuitry," *Algorithmica*, vol. 6, pp. 5-35, 1991.

[19] B.P. Lester, *The Art of Parallel Programming*. Englewood Cliffs, N.J.: prentice Hall, 1993.

[20] W. Li and K. Pingali, "A Singular Loop Transformation Framework Based on Non-Singular Matrices," Technical report TR 92-1294, Cornell Univ., Ithaca, N.Y., July 1992.

[21] J. Llosa, M. Valero, and E. Ayguadé, "Heuristics for Register-Constrained Software Pipelining," *Proc. 29th Int'l Symp. Microarchitecture*, pp. 250-261, Dec. 1996.

[22] K.K. Parhi and D.G. Messerschmit, "Static Rate-Optimal Scheduling of Iterative Data-Flow Program via Optimum Unfolding," *IEEE Trans. Computers*, vol. 40, no. 2, pp. 178-195, Feb. 1991.

[23] N.L. Passos, E. Sha, and S.C. Bass, "Loop Pipelining for Scheduling Multi-Dimensional Systems via Rotation," *Proc. 31st Design Automation Conf.*, pp. 485-490, June 1994.

[24] B.R. Rau and J.A. Fisher, "Instruction-Level Parallel Processing," *J. Supercomputing*, vol. 7, pp. 9-50, July 1993.

[25] B.R. Rau and C.D. Glaeser, "Some Scheduling Techniques and a Easily Schedulable Horizontal Architecture for High Performance Scientific Computing," *Proc. 14th Ann. Workshop Microprogramming*, pp. 183-198, Oct. 1981.

[26] B. Ramakrishna Rau, "Iterative Modulo Scheduling: An Algorithm for Software Pipelining Loops," *Proc. 27th Ann. Int'l Symp. Microarchitecture*, pp. 63-74, Nov. 1994.

[27] M.E. Wolfe, *High Performance Compilers for Parallel Computing*, chapter 9, Redwood City, Calif.: Addison-Wesley, 1996.

[28] M.E. Wolfe and M.S. Lam, "A Loop Transformation Theory and Algorithm to Maximize Parallelism," *IEEE Trans. Parallel and Distributed Systems*, vol. 2, no. 4, pp. 452-471, Oct. 1991.

[29] L.A. Zadeh, "Fuzzy Sets as a Basis for a Theory of Possibility," *Fuzzy Seta and Systems*, vol. 1, pp. 3-28, 1978.

Sissades Tongsima received the BEng degree in industrial instrumental engineering in 1991 from Kink Mongkut Institute of Technology, Ladkrabang, Thailand. In 1992, he was awarded the Royal Thai Government Scholarship to pursue his studies in computer science, majoring in the parallel and distributed computing area. He obtained his MS and PhD degrees from the Department of Computer Science and Engineering, University of Notre Dame, in 1995 and 1999, respectively. Dr. Tongsima's research interests while studying at Notre Dame include data scheduling and partitioning on parallel and distributed systems, high-level synthesis, loop transformations, rapid prototyping, and fuzzy systems. Upon completing of his studies, he returned to Thailand, where, in June 1999, he joined the National Electronics and Computer Technology Center (NECTEC). He is currently a researcher in the High Performance Computing Division at NECTEC.

Edwin H.-M. Sha (S'88-M'92) received the MA and PhD degrees from the Department of Computer Science, Princeton University, Princeton, New Jersey, in 1991 and 1992, respectively. Since August 1992, he has been with the University of Notre Dame, Notre Dame, Indiana. He is now an associate professor and the associate chairman of the Department of Computer Science and Engineering. His research areas include multimedia, memory systems, parallel and pipelined architectures, loop transformations and parallelizations, software tools for parallel and distributed systems, high-level synthesis in VLSI, fault-tolerant computing, VLSI processor arrays, and hardware and software co-design.

He had published more than 100 research papers in referred conferences and journals during the past seven years. In 1994, he was the program committee chair for the Fourth IEEE Great lakes Symposium on VLSI and he served as the program committee cochair for the 2000 ISCA 13th International Conference on Parallel and Distributed Computing Systems. He has served on program committees for many international conferences. He has served as an associate editor for several journals. He received the Oak Ridge Association Foundation CAREER Award in 1995. He was also invited to be a guest editor for special issue on Low Power Design of the *IEEE Transactions on VLSI Systems* and is now serving as an editor for the *IEEE Transactions on Signal Processing*. He received the CSE Undergraduate Teaching award in 1998. He is a member of the IEEE.

Chantana Chantrapornchai received her PhD degree in computer science and engineering from the University of Notre Dame in 1999. She received her bachelor's degree in computer science from Thammasart University, Bangkok, Thailand, in 1992 and her MS degree in computer science from Northeastern University, Boston, in 1994. Currently Dr. Chantrapornchai is a faculty member in the Department of Mathematics, Faculty of Science, Silpakorn University, Thailand. Her research interests include high-level synthesis, rapid prototyping, and fuzzy systems.

David R. Surma received the BS degree in electrical engineering from Valparaiso University, Valparaiso, Indiana, where he was a Presidential Scholar, in 1985. He received MS degree in electrical engineering with a specialty in computer engineering from the University of Arizona, Tucson, in 1989. In 1998, he was awarded the PhD degree in computer science and engineering from the University of Notre Dame, Notre Dame, Indiana. From 1990-1996, he taught electrical and computer engineering at Valparaiso University. Since then he has taught computer science at Indiana University and has held a visiting assistant professorship at Notre Dame. Currently, He is an assistant professor of computer science at Valparaiso University. His research interests include parallel and distributed systems, communication and data scheduling, computer networks and real-time systems. He is a member of the IEEE.

Nelson Luiz Passos received the BS degree in electrical engineering, with specialization in telecommunications, from the University of Sao Paulo, Brazil, in 1974. He received the Ms degree in computer science from the University of North Dagota, Grand Forks, in 1992, and the PhD degree in computer science and engineering from the University of Notre Dame, Indiana, in 1996.

From 1974 to 1990, he worked at Control Data Corporation. Since1996, he has been at Midwestern State University, Wichita Falls, Texas, where he is currently an associate professor with the Department of Computer Science. His research interests include parallel processing, VLSI design, loop transformations.

Dr. Passos was awarded the Professional Excellence and Bill Norris Shark Club Awards at Control Data Corporation. While with the Fellowship. He received a U.S. National Science Foundation grant award in 1997 to support to new research om multidimensional retiming. He is a member of the IEEE.

	5th Elliptic	34	438	396	318	**293**	299	314	**289**	294
3 PEs	Floyd-Steinberg	13	59	38	42	**30**	40	38	**27**	31
	Genetic application	18	500	400	434	**275**	416	438	**180**	410
	Fuzzy application	24	69	55	52	**45**	66	52	**43**	63

Columns θ = 0.8 and θ = 0.9 show the results when considering the probabilistic situations with confidence probabilities 0.8 and 0.9. Column "PL" presents the probabilistic schedule length (psl) after modified list scheduling is applied to the benchmarks. Columns "LS", "AL", and "TS" show the resulting psl, after running PRS against those benchmarks using the remapping heuristics LS, AL and TS respectively. Among these three heuristics, the TS scheme produces better results than AL which uses the simplest criteria. Further, it yields as good as or sometimes even better results than given by the LS approach, while TS taking less time to select a re-scheduled position for a node. This is because in each iteration the LS method finds the local optimal place. However, scheduling nodes to these positions does not always result in the global optimal schedule length.

In Table 9, based on the system that has 2 adders and 1 multiplier (for filter benchmarks) and 3PEs (for application benchmarks), we present the comparison results obtained from applying the following techniques to the benchmarks: modified list scheduling, traditional rotation scheduling, probabilistic rotation scheduling using template scheduling heuristic, and traditional rotation scheduling considering average computation times. Columns "L" and "R" show the schedule length obtained from applying modified list scheduling and traditional rotation scheduling respectively to the benchmarks where all probabilistic computation times are converted into their worst-case computation times. Obviously, considering the probabilistic case gives the significant improvement of the schedule length over the worst case scenario.

Column "PL" presents the initial schedule lengths obtained from using the modified list scheduling approach. The results in column "PRS" are obtained from Table 8 (PRS using template scheduling heuristic). In column "AVG", the psls are computed by using the graphs (PTGs) retrieved from running traditional rotation to the benchmarks where the average computation time is assigned to each node. These results demonstrate that considering the probabilistic situation while performing rotation scheduling can consistently give better schedules than considering only worst-case or average-case computation times.

6. Conclusion

We have presented scheduling and optimization algorithms which operate in probabilistic environment. A probabilistic data-flow graph is used to model an application which takes this probabilistic nature into account. The probabilistic retiming algorithm is used to optimize the given application when nonresource constrained environments are assumed. Given an acceptable probability and a desired longest path computation time, the algorithm reduces the computation time of the given probabilistic graph to the desired value. The concept of maximum reaching time is used to calculate timing values of the probabilistic graph. When a limited number of processing elements is considered, the probabilistic rotation scheduling algorithm (where the probabilistic concept and loop pipelining are integrated to optimize a task schedule) is proposed. Based on the maximum reaching time notion, the probabilistic schedule length is used to measure the total computation time of these tasks being scheduled in one iteration. Given a probabilistic graph, the schedule is constructed by using the task-assignment probabilistic graph and the probabilistic schedule length is computed with respect to a given confidence probability θ. Probabilistic rotation scheduling is applied to the initial schedule in order to optimize the schedule. It produces the best optimized schedule with respect to the confidence probability. The remapping heuristic, template scheduling, is incorporated in the algorithm in order to find the scheduling position for each node.

Acknowledgment

The research was partially supported by U.S. National Science Foundation Career Grant MIP 95-01006.

References

[1] A. Aiken and A. Nicolau, "Development Environment for Horizontal Micronode," *IEEE Trans. Software Eng.*, vol. 14, Feb. 1987.

[2] A. Aiken and A. Nicolau, "Optimal Loop Parallelization," *Proc. ACM SIGPLAN Conf. Programming Language Design and Implementation*, pp. 308-317, June 1988.

[3] U. Banerjee, "Unimodular Transformations of Double Loops," *Proc. Workshop Advances in Languages and Compilers for Parallel Processing*, pp. 192-219, Aug 1990.

[4] P.P. Chang et al., "Impact: An Architectural Framework for Multiple Instruction Issue Processor," *Proc. 18th Int'l Symp. Computer Architecture*, pp. 266-275, 1991.

[5] L. Chao, A. LaPaugh, and E. Sha, "Rotation Scheduling: A Loop Pipelining Algorithm," *Proc. 30th Design Automation Conf.*, pp. 566-572, June 1993.

[6] L. Chao and E. Sha, "Static Scheduling for Synthesis of DSP Algorithms on Various Models," *J. VLSI Signal Processing*, pp. 207-223, Oct. 1995.

[7] A.E. Eichenberger and E.S. Davidson, "Stage Scheduling: A Technique to Reduce the Register Requirements of a Modulo Schedule," *Proc. 28th Int'l Symp. Microarchitecture*, pp. 338-349, Nov. 1995.

[8] A.E. Eichenberger, E.S. Davidson, and S.G. Abraham, "Minimum Register Requirements for a Modulo Schedule," *Proc. 27th Int'l Symp. Microarchitecture*, pp. 75-84, Nov. 1994.

[9] J.A. Fisher, "Trace Scheduling: A Technique for Global Microcode Compaction," *IEEE Trans. Computers*, vol. 30, no. 7, pp. 478-490, July 1981.

[10] I. Foster, *Designing and Building Parallel Program: Concepts and Tools for Parallel Software Engineering*. Addison-Wesley, 1994.

[11] R.A. Kamin, G.B. Adams, and P.K. Dubey, "Dynamic List-Scheduling with Finite Resources," Proc. *1994 Int'l Conf. Computer Design*, pp. 140-144, Oct. 1994.

[12] I. Karkowski and R.H.J.M. Otten, "Retiming Synchronous Circuitry with Imprecise Delays," *Proc. 32nd Design Automation Conf.*, pp. 322-326, 1995.

[13] A.A. Khan, C.L. McCreary, and M.S. Jones, "A Comparison of Multiprocessor Scheduling Heuristic," *Proc. 1994 Int'l Conf. Parallel Processing*, vol. II, pp. 243-250, 1994.

[14] D. Ku and G. De Micheli, *High-Level Synthesis of ASICS under Timing and Synchronization Constraints*. Kluwer Academic, 1992.

[15] D. Ku and G. De Micheli, "Relative Scheduling under Timing Constraints: Algorithm for High-Level Synthesis," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, pp. 697-718, June 1992.

[16] M. Lam, "Software Pipelining," *Proc. ACM SIGPLAN '88 Conf. Programming Language Design and Implementation*, pp. 318-328, June 1988.

[17] D.M. Lavery and W.W. Hwu, "Unrolling-Based Optimization for Modulo Scheduling," *Proc. 28th Int'l Symp. Microarchitecture*, pp. 327-337, Nov. 1995.

[18] C.E. Leiserson and J.B. Saxe, "Retiming Synchronous Circuitry," *Algorithmica*, vol. 6, pp. 5-35, 1991.

[19] B.P. Lester, *The Art of Parallel Programming*. Englewood Cliffs, N.J.: prentice Hall, 1993.

[20] W. Li and K. Pingali, "A Singular Loop Transformation Framework Based on Non-Singular Matrices," Technical report TR 92-1294, Cornell Univ., Ithaca, N.Y., July 1992.

[21] J. Llosa, M. Valero, and E. Ayguadé, "Heuristics for Register-Constrained Software Pipelining," *Proc. 29th Int'l Symp. Microarchitecture*, pp. 250-261, Dec. 1996.

[22] K.K. Parhi and D.G. Messerschmit, "Static Rate-Optimal Scheduling of Iterative Data-Flow Program via Optimum Unfolding," *IEEE Trans. Computers*, vol. 40, no. 2, pp. 178-195, Feb. 1991.

[23] N.L. Passos, E. Sha, and S.C. Bass, "Loop Pipelining for Scheduling Multi-Dimensional Systems via Rotation," *Proc. 31st Design Automation Conf.*, pp. 485-490, June 1994.

[24] B.R. Rau and J.A. Fisher, "Instruction-Level Parallel Processing," *J. Supercomputing*, vol. 7, pp. 9-50, July 1993.

[25] B.R. Rau and C.D. Glaeser, "Some Scheduling Techniques and a Easily Schedulable Horizontal Architecture for High Performance Scientific Computing," *Proc. 14th Ann. Workshop Microprogramming*, pp. 183-198, Oct. 1981.

[26] B. Ramakrishna Rau, "Iterative Modulo Scheduling: An Algorithm for Software Pipelining Loops," *Proc. 27th Ann. Int'l Symp. Microarchitecture*, pp. 63-74, Nov. 1994.

[27] M.E. Wolfe, *High Performance Compilers for Parallel Computing*, chapter 9, Redwood City, Calif.: Addison-Wesley, 1996.

[28] M.E. Wolfe and M.S. Lam, "A Loop Transformation Theory and Algorithm to Maximize Parallelism," *IEEE Trans. Parallel and Distributed Systems*, vol. 2, no. 4, pp. 452-471, Oct. 1991.

[29] L.A. Zadeh, "Fuzzy Sets as a Basis for a Theory of Possibility," *Fuzzy Seta and Systems*, vol. 1, pp. 3-28, 1978.

Sissades Tongsima received the BEng degree in industrial instrumental engineering in 1991 from Kink Mongkut Institute of Technology, Ladkrabang, Thailand. In 1992, he was awarded the Royal Thai Government Scholarship to pursue his studies in computer science, majoring in the parallel and distributed computing area. He obtained his MS and PhD degrees from the Department of Computer Science and Engineering, University of Notre Dame, in 1995 and 1999, respectively. Dr. Tongsima's research interests while studying at Notre Dame include data scheduling and partitioning on parallel and distributed systems, high-level synthesis, loop transformations, rapid prototyping, and fuzzy systems. Upon completing of his studies, he returned to Thailand, where, in June 1999, he joined the National Electronics and Computer Technology Center (NECTEC). He is currently a researcher in the High Performance Computing Division at NECTEC.

Edwin H.-M. Sha (S'88-M'92) received the MA and PhD degrees from the Department of Computer Science, Princeton University, Princeton, New Jersey, in 1991 and 1992, respectively. Since August 1992, he has been with the University of Notre Dame, Notre Dame, Indiana. He is now an associate professor and the associate chairman of the Department of Computer Science and Engineering. His research areas include multimedia, memory systems, parallel and pipelined architectures, loop transformations and parallelizations, software tools for parallel and distributed systems, high-level synthesis in VLSI, fault-tolerant computing, VLSI processor arrays, and hardware and software co-design.

He had published more than 100 research papers in referred conferences and journals during the past seven years. In 1994, he was the program committee chair for the Fourth IEEE Great lakes Symposium on VLSI and he served as the program committee cochair for the 2000 ISCA 13th International Conference on Parallel and Distributed Computing Systems. He has served on program committees for many international conferences. He has served as an associate editor for several journals. He received the Oak Ridge Association Foundation CAREER Award in 1995. He was also invited to be a guest editor for special issue on Low Power Design of the *IEEE Transactions on VLSI Systems* and is now serving as an editor for the *IEEE Transactions on Signal Processing*. He received the CSE Undergraduate Teaching award in 1998. He is a member of the IEEE.

Chantana Chantrapornchai received her PhD degree in computer science and engineering from the University of Notre Dame in 1999. She received her bachelor's degree in computer science from Thammasart University, Bangkok, Thailand, in 1992 and her MS degree in computer science from Northeastern University, Boston, in 1994. Currently Dr. Chantrapornchai is a faculty member in the Department of Mathematics, Faculty of Science, Silpakorn University, Thailand. Her research interests include high-level synthesis, rapid prototyping, and fuzzy systems.

David R. Surma received the BS degree in electrical engineering from Valparaiso University, Valparaiso, Indiana, where he was a Presidential Scholar, in 1985. He received MS degree in electrical engineering with a specialty in computer engineering from the University of Arizona, Tucson, in 1989. In 1998, he was awarded the PhD degree in computer science and engineering from the University of Notre Dame, Notre Dame, Indiana. From 1990-1996, he taught electrical and computer engineering at Valparaiso University. Since then he has taught computer science at Indiana University and has held a visiting assistant professorship at Notre Dame. Currently, He is an assistant professor of computer science at Valparaiso University. His research interests include parallel and distributed systems, communication and data scheduling, computer networks and real-time systems. He is a member of the IEEE.

Nelson Luiz Passos received the BS degree in electrical engineering, with specialization in telecommunications, from the University of Sao Paulo, Brazil, in 1974. He received the Ms degree in computer science from the University of North Dagota, Grand Forks, in 1992, and the PhD degree in computer science and engineering from the University of Notre Dame, Indiana, in 1996.

From 1974 to 1990, he worked at Control Data Corporation. Since1996, he has been at Midwestern State University, Wichita Falls, Texas, where he is currently an associate professor with the Department of Computer Science. His research interests include parallel processing, VLSI design, loop transformations.

Dr. Passos was awarded the Professional Excellence and Bill Norris Shark Club Awards at Control Data Corporation. While with the Fellowship. He received a U.S. National Science Foundation grant award in 1997 to support to new research om multidimensional retiming. He is a member of the IEEE.

ระบบสารสนเทศภูมิศาสตร์กับการประยุกต์

ผกาสิน พูนพิพัฒน์[1] และ ภัทรชัย ลลิตโรจน์วงศ์[2]
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ABSTRACT – A Geographic Information System (GIS) is not limited to be used in cartography only, but it has grown enormously across many disciplines and in numerous business sectors involved in map analysis. This article describes the uses of GIS in a variety of ways, such as agricultural production improvement, earthquake analysis, public transport management and urban transport planning. In Thailand, GIS also has been applied for decision supports in the fields of agriculture, route analysis, tax collection, and real estate affairs.

KEY WORDS – GIS, Geographic Information System, Precision Farming, Geo-Informatic, ROMANSE, Road Management System, Urban GIS, Tax Map, Property Information System

บทคัดย่อ – ปัจจุบันระบบเทคโนโลยีสารสนเทศภูมิศาสตร์ (Geographic Information System, GIS) ไม่ได้ถูกจำกัดการใช้งานเฉพาะในกิจกรรมที่มีการใช้แผนที่โดยตรงเท่านั้น หากแต่มีการนำไปประยุกต์ใช้ในกิจกรรมอื่นๆ ที่ต้องอาศัยแผนที่ในการวิเคราะห์ ดังนั้นบทความนี้จะขอนำเสนอตัวอย่างการประยุกต์ใช้ GIS กับงานด้านต่างๆ ได้แก่ การเพิ่มผลผลิตทางการเกษตร การวิเคราะห์และพยากรณ์ปรากฏการณ์แผ่นดินไหว การจัดการระบบขนส่งมวลชนและการวางการผังจราจร เป็นต้น สำหรับประเทศไทยเองก็มีการนำเทคโนโลยี GIS มาใช้เป็นเครื่องมือช่วยในการตัดสินใจในการปลูกข้าว วางแผนการเดินทาง ตลอดจนนำมาใช้กับงานด้านการจัดเก็บภาษี และงานด้านอสังหาริมทรัพย์ เป็นต้น

คำสำคัญ – ระบบสารสนเทศภูมิศาสตร์, เกษตรกรรมแบบพอเหมาะ, ข้อมูลทางธรณีวิทยา, ระบบบริหารการเดินรถ, การวางผังการจราจร, แผนที่ภาษี, สารสนเทศทางอสังหาริมทรัพย์

1. บทนำ

ระบบสารสนเทศภูมิศาสตร์หรือที่เรียกว่า Geographic Information System (GIS) เป็นระบบสารสนเทศที่ใช้ในการจัดเก็บ จัดการ ค้นคืน ค้นหา วิเคราะห์และแสดงผลข้อมูลเชิงพื้นที่ (Spatial Data) ร่วมกับข้อมูลอรรถาธิบาย (Non-Spatial Data)

โดยเราสามารถแบ่งข้อมูลเชิงพื้นที่ที่ใช้สำหรับพัฒนา GIS ได้เป็น 3 ประเภท [1] คือ

1. ข้อมูลแผนที่มูลฐาน คือ ข้อมูลที่แสดงถึงสภาพภูมิประเทศและสิ่งต่างๆ ที่ปรากฏอยู่บนภูมิประเทศ เช่น เส้นชั้นความสูง ถนน เส้นทางน้ำ เป็นต้น

2. ข้อมูลเฉพาะด้าน คือ ข้อมูลที่ใช้ร่วมกันภายในกลุ่มธุรกิจ กลุ่มวิชาชีพ หรือหน่วยงานที่มีลักษณะคล้ายคลึงกัน ตัวอย่างของข้อมูลประเภทนี้ได้แก่ ชนิดของดิน แปลงที่ดิน ตำแหน่งสถานพยาบาล ร้านขายยา สถานีบริการน้ำมัน เป็นต้น ซึ่งความสำคัญของข้อมูลประเภทนี้จะลดน้อยลงเมื่อถูกนำไปใช้ข้ามกลุ่มธุรกิจ ยกตัวอย่างเช่น

[1] นักศึกษาคณะเทคโนโลยีสารสนเทศ สจล.

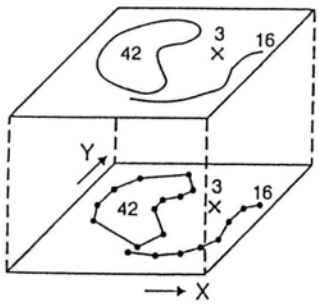
[2] อาจารย์ประจำคณะเทคโนโลยีสารสนเทศ สจล.

ตำแหน่งสถานพยาบาลและร้านขายยาเป็นข้อมูลที่มีความสำคัญมากต่อบริษัทผู้ผลิตยา แต่จะไม่มีความสำคัญใดๆ ต่อเกษตรกรเพื่อทำการเพาะปลูก

3. ข้อมูลเฉพาะองค์กร คือ ข้อมูลที่แต่ละองค์กรได้พัฒนาขึ้นใช้เอง โดยส่วนมากแล้วมักจะเป็นข้อมูลที่เป็นความลับของแต่ละองค์กร ตัวอย่างข้อมูลประเภทนี้ได้แก่ ตำแหน่งที่อยู่ของลูกค้าขององค์กร เป็นต้น

ซึ่งเราสามารถนำข้อมูลเชิงพื้นที่ดังกล่าวเข้าสู่ GIS ได้หลายวิธี เช่น โดยการกราดตรวจ (Scanning), การแปลงข้อมูลเวกเตอร์ (Vector Convertion), การแปลงข้อมูลภาพ (Imagery Data), การทำดิจิไทซ์ (Digitizing), หรือนำเข้าข้อมูลจากข้อมูลดิจิทัลโดยตรง เช่น ข้อมูลที่มาจากแผนที่เชิงเลข (Digital Mapping) หรือ ข้อมูลที่ได้จากการสำรวจจากระยะไกล (Remote Sensing Data) เป็นต้น โดยข้อมูลเหล่านี้จะถูกเก็บอยู่ในรูปของข้อมูลเชิงเลข (Digital Data) [2] อันได้แก่

- ข้อมูลเวกเตอร์ (Vector) เป็นการจัดเก็บข้อมูลในรูปของคู่ลำดับทางภูมิศาสตร์ เช่น อยู่ในรูปของข้อมูลตัวเลขแสดงรายละเอียดของละติจูด และลองจิจูด (x, y) หรือ North-East ดังแสดงในรูปที่ 1 และตารางที่ 1

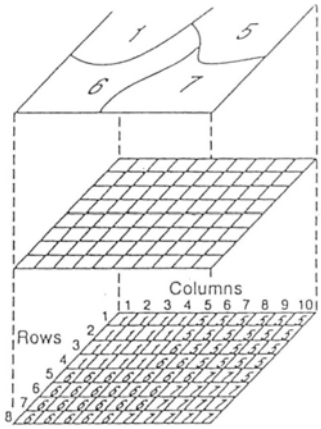


รูปที่ 1. แสดงรูปแบบข้อมูลเวกเตอร์ (Vector) ในระบบสารสนเทศภูมิศาสตร์ [2, หน้า 160]

ตารางที่ 1. โครงสร้างข้อมูลเวกเตอร์ (Vector) [2, หน้า 160]

ลักษณะ	หมายเลข (Class หรือ ID)	ตำแหน่ง
จุด	3	(X, Y) จุดเดี่ยว
เส้น	16	$X_1Y_1, X_2Y_2, \ldots, X_nY_n$
เส้นรอบพื้นที่ (รูปปิด)	42	$X_1Y_1, X_2Y_2, \ldots, X_1Y_1$

- ข้อมูลราสเตอร์ (Raster) เป็นการจัดเก็บข้อมูลในรูปของจุดภาพ (Pixel) โดยมีค่าพิกัดเป็นพิกัดจุดภาพ (Pixel Coordinate) ตามความสัมพันธ์เชิงพิกัด (Row, Column) ดังแสดงในรูปที่ 2 และตารางที่ 2

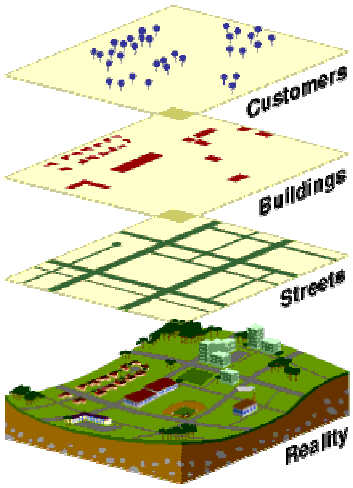


รูปที่ 2. แสดงข้อมูลเชิงพื้นที่ในรูปแบบราสเตอร์ (Raster) ในระบบสารสนเทศภูมิศาสตร์ [2, หน้า 160]

ตารางที่ 2. โครงสร้างข้อมูลเชิงพื้นที่ (Raster) [2, หน้า 160]

Row	1	1	1	1	1	1	1
Column	1	2	3	4	5	6	7
ค่า (Value)	1	1	1	1	5	5	5

โดยข้อมูลเชิงเลขเหล่านี้จะถูกจัดเก็บในลักษณะที่เป็นแผ่นข้อมูล (Layer) วางซ้อนกันตามลักษณะและประเภทของข้อมูล (ดังรูปที่ 3) ซึ่งข้อมูลทั้งหมดจะสามารถเชื่อมโยงถึงกันได้โดยผ่านทางคู่ลำดับทางภูมิศาสตร์



รูปที่ 3. แสดงตัวอย่างการวางซ้อนแผ่นข้อมูล [3]

เมื่อทำการจัดเก็บข้อมูลทั้งหมดสมบูรณ์แล้ว จะสามารถนำข้อมูลที่อยู่ในรูปเชิงแผนที่มาทำการสอบถาม (Query) หรือวิเคราะห์ (Analyse)

ได้ และถ้าหากข้อมูลที่นำมาวิเคราะห์นั้นอยู่คนละแผ่นข้อมูลกัน GIS จะนำชั้นข้อมูลที่ต้องการในส่วนนั้นมาวางซ้อนกัน ทำให้เกิดแผ่นข้อมูลใหม่อีก 1 ชั้น ซึ่งจะมีข้อมูลในส่วนที่นำมาวางซ้อนกันผสมกันอยู่ด้วย วิธีการดังกล่าวนี้เรียกว่า การวางซ้อน (Overlay) ยกตัวอย่างเช่น บริษัทสามารถหาการกระจุกตัวของผู้ใช้บริการได้จากการนำข้อมูลผู้ใช้บริการ อาคาร ถนน ฯลฯ มาวางซ้อนกัน (ดังรูปที่ 3) นอกจากนี้เรายังสามารถทำให้ GIS แสดงผลออกมาในรูปของแผนที่ หรือกราฟ ซึ่งแม้แต่ผู้ที่ไม่มีความรู้เรื่อง GIS ก็สามารถทำความเข้าใจได้โดยง่าย และสามารถนำมาใช้ช่วยในการตัดสินใจได้อย่างรวดเร็ว

2. GIS กับการประยุกต์ใช้

ด้วยประสิทธิภาพในการใช้งานของ GIS ดังที่กล่าวมา ทำให้มีการนำ GIS ไปประยุกต์ใช้ในงานด้านต่างๆ อย่างมากมาย สำหรับบทความนี้จะขอยกตัวอย่างการนำไปประยุกต์ใช้กับงานด้านเกษตรกรรม การวิเคราะห์และพยากรณ์ปรากฏการณ์แผ่นดินไหว การจัดการระบบขนส่งมวลชนและการวางผังการจราจร ดังนี้

2.1 การประยุกต์ใช้ GIS กับงานด้านเกษตรกรรม

เราสามารถนำ GIS มาประยุกต์ใช้ในการช่วยบริหารการทำเกษตรกรรมเพื่อเพิ่มผลกำไรและลดต้นทุน แต่ยังคงรักษาสภาพแวดล้อมไว้ โดยการประยุกต์ใช้ระบบข้อมูลทางการเกษตร ระบบจะให้คำแนะนำเพื่อช่วยในการตัดสินใจ คำนวณ สร้างแบบจำลอง วิเคราะห์ และทำนายปริมาณผลผลิตกับต้นทุน ทำให้เกษตรกรมองเห็นภาพที่เป็นรูปธรรม (Visualization) มากยิ่งขึ้น ตัวอย่างเช่น การจัดการในการให้สารเคมีกำจัดศัตรูพืชในปริมาณที่จำเป็น การจัดการในการให้ปุ๋ยแก่พืชตามความเหมาะสม การปลูกพืชที่เหมาะกับประเภทสารอาหารในดิน การปรับเปลี่ยนการเพาะปลูกเพื่อช่วยปรับสภาพสารอาหารในดิน ฯลฯ ซึ่งสิ่งเหล่านี้จะช่วยลดต้นทุนการผลิตและลดผลกระทบต่อสภาพแวดล้อม

โดยเราสามารถแบ่งองค์ประกอบของ GIS เพื่อช่วยในการทำเกษตรกรรมได้ 2 องค์ประกอบ [4] คือ

1) Map Generator เป็นองค์ประกอบที่ใช้สร้างแผนที่ในการทำเกษตรกรรม โดยอาศัยข้อมูลทางการเกษตรมาวิเคราะห์และสร้างแบบจำลองเพื่อช่วยในการทำเกษตรกรรม ตัวอย่างข้อมูลทางการเกษตรได้แก่ ความเข้มข้นของแร่ธาตุ NPK (Nitrogen, Phosphorous, Potassium) ในดิน ความลาดของภูมิประเทศ ปริมาณน้ำฝนโดยเฉลี่ยในแต่ละปี เป็นต้น

2) Task Controller เป็นองค์ประกอบที่ติดตั้งอยู่ที่แหล่งเพาะปลูกเพื่อทำหน้าที่ควบคุมอุปกรณ์ทางการเกษตร โดยรับคำสั่งมาจาก Map Generator ผ่านระบบกำหนดตำแหน่งบนโลก (Global Position System, GPS) นำมาสั่งงานให้อุปกรณ์ทางการเกษตรทำงาน ตัวอย่างเช่น ควบคุมอุปกรณ์ให้ปุ๋ย ควบคุมอุปกรณ์ให้สารเคมีกำจัดศัตรูพืช หรือควบคุมอุปกรณ์หว่านเมล็ดพันธุ์พืช

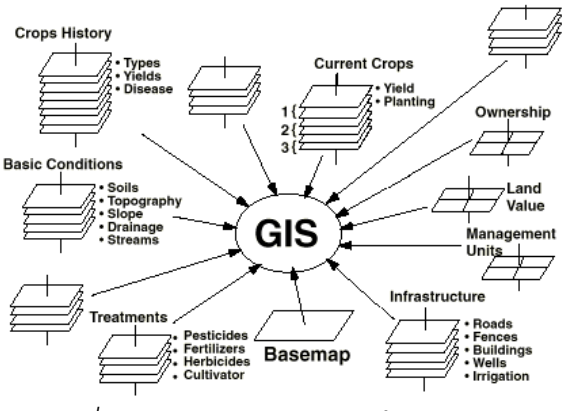
การสื่อสารระหว่าง Map Generator และ Task Controller ผ่าน GPS จะเป็นไปแบบ 2 ทิศทาง คือ คำสั่งจาก Map Generator จะส่งไปยัง Task Controller เพื่อควบคุมอุปกรณ์ทางการเกษตรให้ทำงาน แต่ในขณะเดียวกัน ข้อมูลการเพาะปลูกซึ่ง Task Controller ได้จัดเก็บไว้จากการควบคุมอุปกรณ์ทางการเกษตร ณ บริเวณพื้นที่เพาะปลูก ตัวอย่างเช่น ปริมาณสารเคมีกำจัดศัตรูพืชที่ได้ให้ไว้แล้ว จะถูกส่งกลับไปที่ Map Generator เพื่อบันทึกและปรับปรุงประวัติการเพาะปลูกอีกครั้งหนึ่ง ข้อมูลประวัติที่ส่งกลับมานี้จะเป็นประโยชน์อย่างมากสำหรับช่วยวางแผนการเพาะปลูกในอนาคต

การพัฒนา GIS เพื่อช่วยในการทำเกษตรกรรมจะเริ่มต้นจากการจัดเตรียมแผนที่มูลฐาน (Basemap) ของบริเวณที่จะทำการเพาะปลูก แล้วทำการเก็บข้อมูลทางการเกษตรเข้าไปในระบบแผนที่ก่อน ตัวอย่างข้อมูลทางการเกษตรซึ่งแสดงในรูปที่ 4 ได้แก่

- ข้อมูลคุณสมบัติของดิน ซึ่งได้จากการสุ่มตัวอย่างดินมาวิเคราะห์หาระดับความเข้มข้นของธาตุ NPK และทำการบันทึกข้อมูลในลักษณะดังนี้ คือ รหัสพิกัด ความเข้มข้น N ความเข้มข้น P และความเข้มข้น K

- Crop Type เป็นข้อมูลจำเพาะของพันธุ์พืชแต่ละชนิด ประกอบไปด้วย รหัสพืช ชื่อพืช และคุณสมบัติเฉพาะของพืชแต่ละชนิด

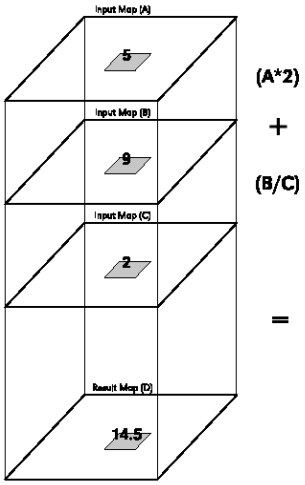
- Slope Map เป็นแผนที่แสดงความลาดของภูมิประเทศ เนื่องจากว่า พื้นที่ซึ่งมีความลาดเอียงมากอาจไม่เหมาะในการเพาะปลูกพืชบางชนิด



รูปที่ 4. แสดงข้อมูลทางการเกษตรในแบบจำลองเพื่อช่วยในการทำเกษตรกรรม [4, หน้า 7]

ข้อมูลทางการเกษตรเหล่านี้จะถูกนำมาใส่เข้าไปในแบบจำลอง (Model) ซึ่งได้สร้างความสัมพันธ์ในเชิงพีชคณิต (Algebra) ระหว่างข้อมูลทางการเกษตรไว้แล้ว แบบจำลองจะให้คำแนะนำการเพาะปลูกแสดงพร้อมกับแผนที่

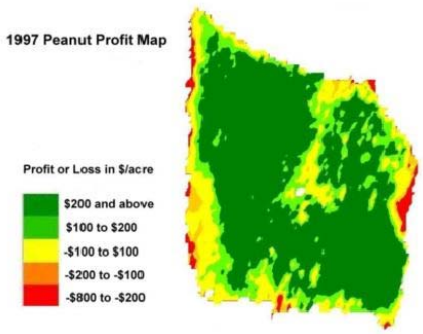
ตัวอย่างเช่น จากรูปที่ 5 แสดงการสร้าง Map D โดยการสร้างความสัมพันธ์ของ Map A, Map B และ Map C ตามสมการ 2A+(B/C) ได้ Map D ซึ่งเป็นแผ่นข้อมูลใหม่



รูปที่ 5. แสดงตัวอย่างการสร้าง Map D ด้วยหลักการทางพีชคณิต [4, หน้า 12]

ในรูปที่ 6 แสดงตัวอย่างแผนที่พื้นที่เพาะปลูกกับผลกำไรด้วยการสร้างแผ่นข้อมูล ผลกำไร (*P*) จากแผ่นข้อมูล ข้อมูลที่ได้จัดทำขึ้นอยู่ใน GIS คือ รายได้ (*GI*) และ รายจ่าย (*I*) โดยค่าทั้งหมดมีหน่วยเป็น ดอลลาร์ต่อเอเคอร์ ($/acre)

$$P = GI - I$$



รูปที่ 6. แสดงแผนที่เพาะปลูกกับผลกำไร [5, หน้า 5]

ด้วยแผนที่ดังกล่าวนี้เอง เกษตรกรสามารถเลือกพื้นที่ทำการเพาะปลูกในบริเวณที่คาดว่าน่าจะให้ผลตอบแทนสูง และหลีกเลี่ยงการเพาะปลูกในบริเวณที่น่าจะให้ผลตอบแทนต่ำหรือขาดทุน

แบบจำลองทางการทำการเกษตรนี้จะใช้ช่วยในการตัดสินใจในการทำเกษตรกรรม โดยอาศัยข้อมูลทางการเกษตรมาใช้คำนวณตาม สมการที่ได้ใส่ไว้ในแบบจำลองตั้งแต่ต้น ตัวอย่างเช่น ปริมาณการให้น้ำ ปริมาณการให้ปุ๋ย และปริมาณการใช้สารเคมีกำจัดศัตรูพืช ในกรณีที่มีการเพาะปลูกพืชหลายชนิดในแหล่งเพาะปลูกเดียวกัน เกษตรกรสามารถประยุกต์ใช้แบบจำลองเดิมโดยเพียงแต่แก้ไขค่าปัจจัยเฉพาะ (Factor) ซึ่งมีค่าแตกต่างกันในพืชแต่ละชนิดเข้าไปในแบบจำลองมาตรฐาน โดยไม่จำเป็นต้องเริ่มต้นพัฒนาระบบตั้งแต่แรก

ในกรณีที่คำแนะนำในการทำเกษตรกรรม ซึ่งได้มาจากหลักการทางพีชคณิตนั้น มีความไม่ถูกต้องและไม่เหมาะสม เกษตรกรผู้ใช้สามารถเข้าไปแก้ไขปรับปรุงคำแนะนำนั้น โดยอาศัยประสบการณ์ความชำนาญของเกษตรกรเองได้

2.2 การประยุกต์ใช้ GIS กับการวิเคราะห์และพยากรณ์ปรากฏการณ์แผ่นดินไหว

แผ่นดินไหว คือ มหันตภัยธรรมชาติซึ่งเกิดจากการเปลี่ยนแปลงอันเนื่องมาจากแรงดันที่ซ่อนอยู่ใต้พื้นพิภพ ทำให้เกิดการเคลื่อนตัวของเปลือกโลก และทำให้เกิดเหตุการณ์แผ่นดินไหว ซึ่งปรากฏการณ์แผ่นดินไหวนี้ได้สร้างความสูญเสียอย่างร้ายแรงต่อชีวิตและทรัพย์สิน โดยที่มนุษย์มิอาจจะควบคุมหรือระงับการเกิดปรากฏการณ์ได้ แต่มนุษย์สามารถทำนายพยากรณ์ล่วงหน้าได้ว่า จะเกิดเหตุการณ์แผ่นดินไหวขึ้นที่ใดและเมื่อใด โดยใช้หลักการทางวิทยาศาสตร์ ทำให้มนุษย์สามารถหาทางป้องกันได้ทันเวลาก่อนที่เหตุการณ์จะเกิดขึ้น และช่วยลดความสูญเสียที่อาจจะเกิดขึ้นได้ ตัวอย่างเช่น อาจทำการเคลื่อนย้ายสิ่งมีชีวิตออกจากพื้นที่ก่อนที่จะเกิดแผ่นดินไหว เป็นต้น

นักวิทยาศาสตร์ได้ศึกษาและตั้งสมมติฐานเพื่ออธิบายการเคลื่อนตัวของเปลือกโลก (Plate Tectonic Theory) โดยอ้างอิงถึงสถิติข้อมูลของการเกิดแผ่นดินไหว ศึกษาหลักทางธรณีวิทยา เรียนรู้ความสัมพันธ์ของการเปลี่ยนแปลงบรรยากาศกับการเกิดแผ่นดินไหว ศึกษาคุณสมบัติของชั้นหิน เป็นต้น สิ่งต่างๆ เหล่านี้จะถูกนำมาประมวลความสัมพันธ์ สร้างเป็นทฤษฎีและแบบจำลองต้นฉบับ (Prototype) เพื่อใช้อธิบายเหตุการณ์ที่เกิดขึ้น ทำให้สามารถทำนายพยากรณ์เหตุการณ์การเกิดแผ่นดินไหวได้ล่วงหน้า

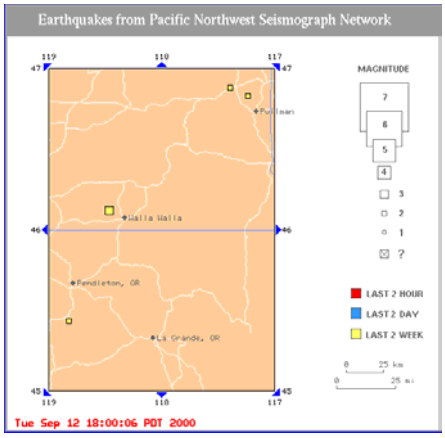
ปัจจุบัน GIS มีบทบาทสำคัญอย่างยิ่งในการสร้างแบบจำลองเพื่อใช้อธิบายปรากฏการณ์การเกิดแผ่นดินไหว และนำมาใช้ทดสอบสมมติฐานต่างๆ ที่นักวิทยาศาสตร์ได้สร้างขึ้นมา ทำให้สะดวกและรวดเร็ว มีความถูกต้องแม่นยำมากยิ่งขึ้น

บทบาทที่สำคัญอีกอย่างหนึ่งของ GIS คือ การใช้ระบบเป็นเครื่องมือในการจัดการและจัดเก็บสถิติข้อมูลของการเกิดแผ่นดินไหว โดยความสั่นสะเทือนจากเหตุการณ์แผ่นดินไหวจะถูกวัดและจัดเก็บไว้เป็นแฟ้มประวัติ ซึ่งผู้ที่สนใจสามารถเข้าไปศึกษา ค้นคืน หรือค้นคว้าข้อมูลการเปลี่ยนแปลงของเปลือกโลกในลักษณะ Online ผ่านทาง Website ตัวอย่างเช่น http://www.geophys.washington.edu [6]

ในรูปที่ 7 แสดงเหตุการณ์ล่าสุดของการเกิดแผ่นดินไหวบนแผนที่ในบริเวณ Walla Walla ในมลรัฐวอชิงตัน สหรัฐอเมริกา ผ่านทาง

http://spike.geophys.washington.edu/recenteqs/ ซึ่งมีข้อมูลประกอบดังนี้

Magnitude: 3.0 – duration magnitude (Md)
Time: September 5, 2000 at 11:54:18 PM (PDT)
Distance from: Walla Walla, WA - 11 km
Pasco, WA - 51 km
Kenewikck, WA - 52 km
Yakima, WA - 166 km
Coordinate: 46 deg. 7.2 min. N (46.120 N),
118 deg. 27.6 min. W (118.459W)



รูปที่ 7. แสดงตัวอย่างข้อมูลการเกิดแผ่นดินไหวผ่านทาง Internet [6]

Global Earthquake Mapping and Information Package (GEMIP) [7] เป็นระบบฐานข้อมูลการเกิดแผ่นดินไหวอีกระบบหนึ่งที่ได้รับความนิยมและมีความน่าเชื่อถือสูง มีจุดประสงค์ในการให้บริการข้อมูลแก่นักวิทยาศาสตร์ที่ศึกษาการเปลี่ยนแปลงของเปลือกโลกหรือเหตุการณ์แผ่นดินไหว ข้อมูลจะได้รับการปรับปรุงอยู่ตลอดเวลาและต่อเนื่อง ดังนั้นจึงทำให้แบบจำลองที่พัฒนาขึ้นมาใหม่ สามารถใช้อธิบายและทำนายเหตุการณ์แผ่นดินไหวได้อย่างน่าเชื่อถือและถูกต้องแม่นยำมากยิ่งขึ้น

นักวิทยาศาสตร์ได้จัดแบ่งลักษณะภูมิศาสตร์การเกิดแผ่นดินไหวออกเป็นโซน (Zone) ตามคุณสมบัติทางธรณีวิทยาของชั้นเปลือกโลก และกำหนดค่าคุณสมบัติเฉพาะของการเกิดแผ่นดินไหวไว้ ซึ่งค่าดังกล่าวนี้จะได้รับการปรับปรุงแก้ไขเป็นประจำ

การให้คะแนนจะจัดแบ่งออกเป็น 4 ช่วงคะแนนใหญ่ๆ คือ ช่วงที่ 1: 0-3 คะแนน เป็นช่วงที่ไม่สามารถพยากรณ์ได้ (Unpredictable), ช่วงที่ 2: 4-7 คะแนน เป็นช่วงที่ยากต่อการพยากรณ์ (Poor), ช่วงที่ 3: 8-11 คะแนน เป็นช่วงระดับปานกลางที่สามารถพยากรณ์ได้ (Medium), ช่วงที่ 4: 12-13 คะแนน เป็นช่วงที่พยากรณ์ได้แน่นอน (Strong)

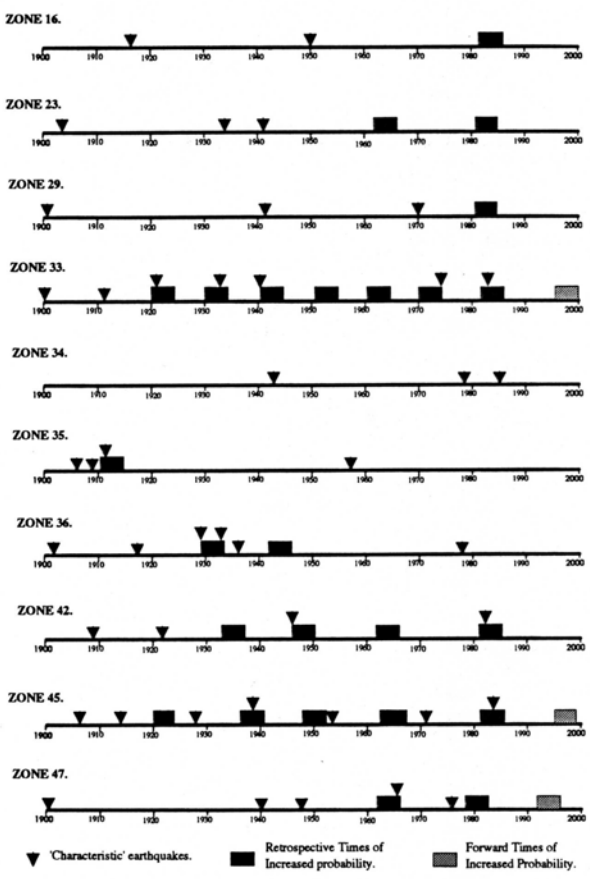
ในเขตอเมริกาใต้สามารถแบ่งโซนการเกิดแผ่นดินไหวออกได้เป็น 46 โซน ซึ่งจะประกอบไปด้วยโซนที่อยู่ในช่วงที่ไม่สามารถพยากรณ์ได้ 25โซน และเป็นช่วงที่ยากต่อการพยากรณ์อีก 11 โซน ดังนั้นจะเหลือเพียง 10 โซน ซึ่งนักวิทยาศาสตร์ให้ความสนใจในการพยากรณ์ ดังแสดงในรูปที่ 8

โดยสรุปแล้ว GIS ได้กลายเป็นเครื่องมือที่ใช้ในการจัดเก็บข้อมูลเหตุการณ์แผ่นดินไหว ใช้แสดงภาพเหตุการณ์จริงหรือเหตุการณ์จำลองใช้ในการทดสอบสมมติฐานการเกิดแผ่นดินไหว ทำให้นักวิทยาศาสตร์สามารถวิเคราะห์พยากรณ์เหตุการณ์แผ่นดินไหวได้อย่างง่ายดาย รวดเร็ว และถูกต้องมากยิ่งขึ้น ช่วยให้เราสามารถป้องกันการสูญเสียที่อาจจะเกิดขึ้นได้ล่วงหน้า ส่วนผู้ที่อาศัยอยู่ในบริเวณที่มีความเสี่ยงต่อการเกิดแผ่นดินไหวสูงนั้น ก็สามารถมั่นใจได้ว่าเขาจะรู้ตัวก่อนเกิดแผ่นดินไหวและสามารถหลีกเลี่ยงเหตุการณ์นั้นได้ทัน ดังนั้นมนุษย์จึงไม่จำเป็นต้องกลัวปรากฏการณ์แผ่นดินไหวอีกต่อไป

2.3 การประยุกต์ใช้ GIS กับการจัดการระบบขนส่งมวลชนและการวางผังการจราจร

GIS ได้ถูกนำมาพัฒนาเพื่อนำมาประยุกต์ใช้ช่วยในการจัดการระบบขนส่งมวลชน (Public Transport Management) ให้มีประสิทธิภาพมากยิ่งขึ้น ด้วยระบบนี้จะทำให้การให้บริการมีประสิทธิภาพสูงขึ้น ผู้ใช้บริการจะได้รับความสะดวกรวดเร็วในการเดินทางมากขึ้น ในขณะที่ระบบขนส่งมวลชนก็สามารถลดค่าใช้จ่ายลงไป

การที่เทคโนโลยี GIS จะสามารถช่วยในการจัดการระบบขนส่งมวลชนได้นั้น จำเป็นต้องอาศัยอุปกรณ์การสื่อสารระหว่างศูนย์ควบคุมกลางกับรถโดยสารในลักษณะแบบทันที (Real Time) โดยข้อมูลที่รับส่งนี้จะถูกใช้ในการปรับปรุงเพื่อให้การบริการมีประสิทธิภาพอยู่ตลอดเวลาและต่อเนื่อง โดยข้อมูลการจราจรทั้งหมดจะถูกจัดเก็บไว้เพื่อใช้ในการวางแผนเพื่อจัดเตรียมปริมาณรถโดยสารหรือเปลี่ยนเส้นทางการจราจรให้สอดคล้องกับอุปสงค์และอุปทาน (Demand and Supply Balance) ที่จะเกิดขึ้นในอนาคต ทำให้ระบบขนส่งมวลชนรักษาประสิทธิภาพในการให้บริการอยู่เสมอ



รูปที่ 8. แสดงโซนการเกิดแผ่นดินไหวซึ่งสามารถพยากรณ์ได้ในเขตอเมริกาใต้ [7, หน้า 242]

จุดประสงค์สำคัญของการพัฒนาระบบขนส่งมวลชนแบบใหม่นี้ก็เพื่อที่จะส่งเสริมให้ประชาชนหันกลับมาใช้บริการระบบขนส่งมวลชนแทนการใช้รถยนต์ส่วนบุคคล ทำให้เกิดการอนุรักษ์พลังงานเชื้อเพลิงและเป็นการลดมลพิษจากการเผาไหม้ที่เกิดขึ้นได้

สาเหตุที่ทำให้ระบบขนส่งมวลชนไม่ประสบความสำเร็จและไม่เป็นที่นิยมในอดีตนั้น เนื่องมาจากการให้บริการระบบขนส่งมวลมีความล่าช้า ขาดการบริหารตารางเวลาเดินรถที่มีประสิทธิภาพ เส้นทางเดินรถไม่ตรงและไม่เพียงพอกับความต้องการ ขาดความสะดวกสบาย มีความแออัด ยัดเยียด เป็นต้น ด้วยการนำ GIS มาประยุกต์ใช้จะทำให้การจัดการระบบขนส่งมวลชนเป็นเรื่องที่ง่าย และจะทำให้ระบบขนส่งมวลชนมีประสิทธิภาพ ช่วยแก้ไขปัญหาต่างๆ ดังที่ได้กล่าวมาแล้ว

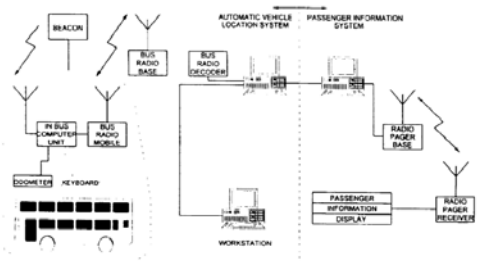
องค์ประกอบที่สำคัญในการพัฒนาระบบขนส่งมวลชน มีดังนี้คือ

1) ข้อมูลสารสนเทศต้องมีความน่าเชื่อถือและถูกต้องแม่นยำ
2) ข้อมูลสารสนเทศต้องเป็นแบบทันที เพื่อที่จะนำข้อมูลมาปรับปรุงการให้บริการได้อย่างตลอดเวลาและทันเวลา
3) ต้องมีระบบสื่อสารระหว่างรถโดยสารและศูนย์ควบคุม
4) มีการให้บริการคำแนะนำข้อมูลการจราจร เช่น แนะนำเส้นทางจราจรที่สะดวกในการเดินทาง
5) มีระบบให้คำปรึกษาซึ่งสามารถโต้ตอบกับผู้ใช้บริการในเชิงโต้ตอบ (Interactive Technology)

ในสหราชอาณาจักรมีการพัฒนาโครงการทดลองที่เรียกว่า ROMANSE (ROad MANagement System for Europe) [8] เพื่อพัฒนาระบบขนส่งมวลชนให้มีประสิทธิภาพสูง ด้วยการประยุกต์ใช้ GIS โดยโครงการนี้มีวัตถุประสงค์หลัก 2 ประการคือ จูงใจให้ประชาชนหันมาใช้บริการระบบขนส่งมวลชนแทนการใช้รถยนต์ส่วนบุคคล และพัฒนาระบบให้บริการข้อมูลเพื่อแนะนำเส้นทางและวิธีการเดินทางทั้ง**ก่อนและ**ระหว่างการเดินทาง โดยทำการติดตั้งคอมพิวเตอร์เพื่อแนะนำข้อมูลการเดินทางตามสถานที่ต่างๆ

ROMANSE เป็นการนำเทคโนโลยีระบบติดตามและบอกตำแหน่งยานพาหนะอัตโนมัติ หรือที่เรียกว่า Automatic Vehicle Location (AVL) โดยอาศัยองค์ประกอบ 3 องค์ประกอบ ดังแสดงในรูปที่ 9 คือ

1) Bus Radio Mobile เป็นเครื่องมือที่ติดตั้งไว้ที่รถโดยสารเพื่อใช้ในการส่งสัญญาณบอกตำแหน่งและข้อมูลการเดินรถต่างๆ (เช่น เลขกิโลเมตร ณ ขณะนั้น) ไปยัง Intelligent Bus Stop
2) Intelligent Bus Stop (IBS) เป็นเครื่องมือที่ติดตั้งไว้ที่ป้ายจอดรถโดยสาร เพื่อทำหน้าที่รับข้อมูลซึ่งส่งมาจากรถโดยสารขณะที่แล่นเข้ามาใกล้ และในขณะเดียวกัน ก็ทำหน้าที่ในการติดต่อกับศูนย์ควบคุมระบบกลาง หรือที่เรียกว่า Traffic and Traveller Information Centre
3) Traffic and Traveller Information Centre (TTIC) เป็นศูนย์ควบคุมระบบกลาง ซึ่งทำหน้าที่ในการบังคับและควบคุมระบบ ROMANSE ทั้งระบบ



รูปที่ 9. แสดงการทำงานของระบบ ROMANSE [8, หน้า 9/4]

จากรูปที่ 9 ขณะที่รถโดยสารแล่นเข้าใกล้ป้ายจอดรถ ข้อมูลตำแหน่งรถและข้อมูลการเดินรถต่างๆ (จาก Odometer) จะถูกส่งออกจาก Bus Radio Mobile ไปยังแผงรับสัญญาณ (Beacon) ซึ่งติดตั้งอยู่ที่ IBS และจากนั้นข้อมูลก็จะถูกส่งต่อไปยังศูนย์ควบคุม TTIC ดังนั้นระบบ ROMANSE จะทราบตำแหน่งและสถานภาพของรถทุกคันได้ตลอดเวลา เพื่อที่จะนำมาใช้ควบคุมระบบ AVL ให้ปฏิบัติงานได้อย่างมีประสิทธิภาพ

นอกจากนี้แล้ว ข้อมูลการจราจรทั้งหมดที่ส่งเข้ามาจะถูกจัดเก็บและนำมาใช้วิเคราะห์เพื่อปรับปรุงประสิทธิภาพ ใช้ในการวางแผนพัฒนาเส้นทางที่สะดวกรวดเร็ว หรือใช้ในการวางแผนจัดเตรียมจำนวนรถโดยสารให้พอเหมาะกับสถานการณ์ในอนาคต

ในระหว่างการเดินรถ หากเกิดความผิดปกติกับรถโดยสารหรือสัญญาณของรถโดยสารได้หายไปจากระบบ พนักงานที่ศูนย์ควบคุม TTIC สามารถเข้าตรวจสอบความผิดปกตินี้ และทำการติดต่อพูดคุยกับพนักงานขับผ่านระบบวิทยุสื่อสารได้โดยตรง เพื่อให้คำแนะนำกับพนักงานขับรถได้อย่างทันท่วงที หรืออาจทำการจัดรถเข้าเสริมให้บริการแทน ทำให้ลดผลกระทบต่อประสิทธิภาพการให้บริการของระบบขนส่งมวลชน และในทางกลับกัน เมื่อพนักงานขับรถโดยสารมีความประสงค์ที่จะติดต่อกับศูนย์ควบคุม TTIC ก็สามารถส่งสัญญาณเรียกเตือน เพื่อขอคำแนะนำจากศูนย์ควบคุมได้ทันทีเช่นกัน

นอกเหนือจากการควบคุมระบบการให้บริการรถโดยสารที่มีประสิทธิภาพแล้วนั้น ศูนย์ควบคุม TTIC ยังมีหน้าที่อีกอย่าง คือ ให้คำแนะนำไปยังป้ายจอดรถ IBS เพื่อให้คำแนะนำกับผู้โดยสารที่รออยู่ที่ป้ายจอดรถ ในการเลือกเส้นทางและวิธีการเดินทางที่เหมาะสมที่สุดในขณะนั้น โดยให้คำแนะนำดังนี้ (ดังแสดงในรูปที่ 10)

1) หมายเลขของสายเดินรถโดยสารที่จะผ่านป้ายจอดรถโดยสาร (Route Number) นั้นๆ
2) จุดหมายปลายทางที่รถโดยสารนั้นจะไป
3) เวลาโดยประมาณที่รถโดยสารคันถัดไปจะมาถึงป้ายจอดรถโดยสารที่นักเดินทางรออยู่

รูปที่ 10. แสดงรูปป้ายรถโดยสารพร้อมบอร์ดที่สหราชอาณาจักร [9]

นอกเหนือจากบอร์ดแสดงคำแนะนำในการเดินทางแล้ว ได้มีการพัฒนาระบบการให้คำแนะนำด้วยระบบเสียง หรือที่เรียกกันว่า Voice Synthesizer เพื่อให้บริการแก่บุคคลซึ่งไม่สะดวกที่จะอ่านคำแนะนำบนบอร์ด (เช่น ผู้พิการด้านสายตา) โดยผู้ใช้บริการจะต้องพกป้าย (Tag) ซึ่งส่งสัญญาณกระตุ้นไปยัง IBS ในการให้คำแนะนำด้วยเสียง

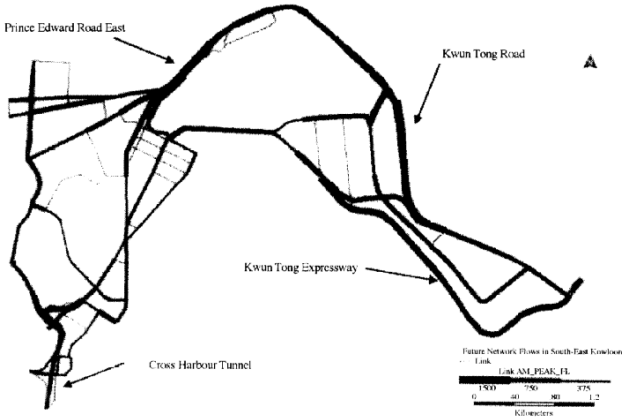
ดังนั้นการนำ GIS มาใช้จะช่วยเพิ่มประสิทธิภาพในการให้บริการขนส่งมวลชน ให้มีความสะดวกรวดเร็ว ลดเวลาที่ใช้ในการเดินทาง เพิ่มความแน่นอน ลดเวลาระหว่างรอรถโดยสาร และช่วยเพิ่มสมรรถนะในการควบคุมและสั่งการ ทำให้สามารถตรวจสอบสถานภาพของรถโดยสารที่อยู่ในระหว่างปฏิบัติงาน ณ เวลาปัจจุบันได้ ซึ่งสิ่งเหล่านี้จะช่วยจูงใจให้ประชาชนหันมาใช้บริการระบบขนส่งมวลชนแทนการใช้รถยนต์ส่วนบุคคลมากยิ่งขึ้น เป็นการอนุรักษ์ทรัพยากรเชื้อเพลิงและยังเป็นการช่วยลดปัญหามลภาวะ

นอกจากนี้เรายังสามารถนำ GIS มาประยุกต์ใช้กับการวางผังการจราจร หรือ Urban Transport Planning System (UTPS) [10] ซึ่งเป็นระบบการวางแผนพัฒนาเส้นทางการจราจร เพื่อให้สามารถรองรับปริมาณความต้องการ และพร้อมสำหรับการเปลี่ยนแปลงที่กำลังจะเกิดขึ้นในอนาคต โดยนำ GIS มาใช้เป็นเครื่องมือช่วยในการวางแผนปรับปรุงและพัฒนาเส้นทางให้มีขนาดช่องทางเดินรถและจำนวนเส้นทาง ในปริมาณที่เหมาะสมกับความต้องการและสามารถรองรับกับการเปลี่ยนแปลงที่กำลังจะเกิดขึ้น ตัวอย่างเช่น ขนาดหรือลักษณะของชุมชนที่เปลี่ยนแปลงไป ระบบสาธารณูปโภคพื้นฐานที่กำลังจะเกิดขึ้นใหม่ ประเภทของกิจกรรมการใช้พื้นที่ที่เปลี่ยนแปลงไป เช่น เปลี่ยนจากพื้นที่เกษตรกรรมเป็นพื้นที่อุตสาหกรรม เป็นต้น โดยมีขั้นตอนที่เกี่ยวข้อง ดังต่อไปนี้

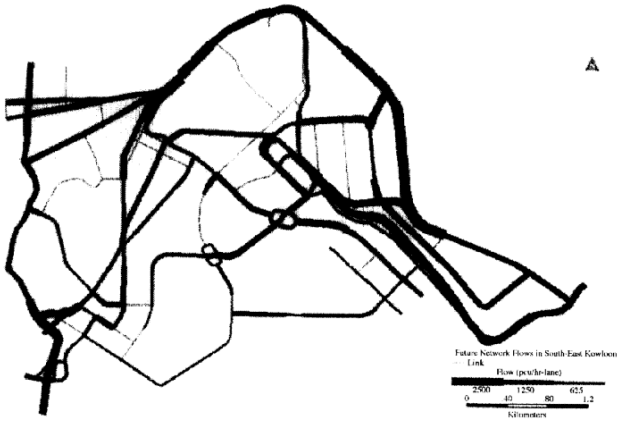
1) Trip Generation เพื่อพยากรณ์จำนวนการเดินทาง (Trip) โดยคำนึงถึงขนาดและจำนวนครัวเรือน รายได้ในแต่ละพื้นที่ และประเภทของกิจกรรมการใช้พื้นที่ เช่น เป็นแหล่งที่อยู่อาศัย หรือแหล่งอุตสาหกรรม
2) Trip Distribution เพื่อใช้หาจุดเริ่มต้นการเดินทางและจุดหมายปลายทาง (Original-Destination Matrix, O-D Matrix)
3) Modal Split เพื่อทำหน้าที่กระจาย O-D Matrix แต่ละตัวด้วยวิธีการขนส่งต่างๆ
4) Network Assignment เพื่อใช้ในการกำหนดเส้นทาง และวิธีการเดินทาง โดยคำนวณจากระยะทาง ความเร็วที่รถสามารถวิ่งได้ ความจุของรถยนต์ และขนาดของถนน

ดังนั้นนักวางแผนพัฒนาเส้นทางจึงสามารถนำ GIS มาเป็นเครื่องมือช่วยในการสร้างแบบจำลองสถานการณ์ เพื่อแสดงการเปลี่ยนแปลงของเส้นทางการจราจรตามสถานการณ์ที่ได้ตั้งสมมติฐานและกำหนดไว้ ทำให้สามารถนำมาใช้เปรียบเทียบกับสภาพเส้นทางที่มีอยู่ก่อนการเปลี่ยนแปลงได้อย่างรวดเร็ว และสามารถใช้ผลจากคำแนะนำมาใช้ช่วยในการทำแผนปรับปรุงพัฒนาเส้นทางการจราจร เพื่อรองรับการใช้งานในอนาคต

ดังเช่นที่เขตบริหารพิเศษฮ่องกงได้นำ GIS มาช่วยในการวางผังการจราจรสำหรับการเตรียมการย้ายสนามบินแห่งชาติ จาก Kai Tak ในเขตเกาลูน ซึ่งมีความแออัด มาเป็น Chek Lap Kok ในฝั่ง Lantau โดยพื้นที่เดิมจะเปลี่ยนเป็นพื้นที่พัฒนาอุตสาหกรรมใหม่



รูปที่ 11. แสดงถนนในเขตเกาลูนก่อนการเปลี่ยนแปลง [10, หน้า 440]



รูปที่ 12. แสดงถนนในเขตเกาลูนหลังการเปลี่ยนแปลง [10, หน้า 440]

หลังจากที่ได้มีการบันทึกข้อมูลจำนวนประชากรและขนาดครอบครัวตามข้อมูลการวางแผนสำมะโนครัวทั้งก่อนและหลังการวางแผนการพัฒนาแหล่งอุตสาหกรรมแห่งใหม่ โดยพิจารณาผลกระทบของการย้ายสนามบินไปอยู่ที่เกาะ Lantau โดยทำการจำลองสถานการณ์ทั้งก่อนและหลัง เพื่อคำนวณเส้นทางการจราจรที่เหมาะสม พบว่าสภาพถนนในเขตเกาลูนนี้ จะไม่สามารถรองรับการเปลี่ยนแปลงที่เกิดขึ้นใหม่ได้อีกต่อไป ดังนั้น จึงจำเป็นที่จะต้องทำการพัฒนาระบบเส้นทางจราจรในเขตเกาลูนจากสภาพการจราจรในรูปที่ 11 ให้เป็นระบบเส้นทางการจราจรที่อยู่ในรูปที่ 12 โดยจะต้องทำการตัดถนนเล็กๆ เพื่อเชื่อมต่อถนนที่มีอยู่เดิม มีการขยายช่องทางจราจรให้กว้างขึ้น สังเกตได้จากขนาดของเส้นทางที่มีความหนามากขึ้น นอกจากนี้จะมีการสร้างถนนใหม่ในเขตอุตสาหกรรม Kwun Tong และพัฒนาถนนที่เชื่อมจาก Tsim Sha Tsui ไปยังสนามบินแห่งชาติแห่งใหม่ เพื่อให้สามารถรองรับกับปริมาณการใช้งานที่เพิ่มมากขึ้น

โดยสรุปแล้ว GIS มีบทบาทสำคัญในการช่วยวางผังการจราจร เพราะจะช่วยในการสร้างแบบจำลองเพื่อพัฒนาเส้นทางการจราจรใหม่ ให้รองรับกับการเปลี่ยนแปลงได้อย่างรวดเร็ว และให้ความสะดวกในการวิเคราะห์เปรียบเทียบความแตกต่างที่เกิดขึ้นระหว่างเส้นทางการจราจรปัจจุบันและอนาคต เพื่อจะได้นำมาใช้ในการวางผังการจราจรสำหรับอนาคต

3. GIS ในประเทศไทย

ประเทศไทยเราได้มีการนำ GIS มาประยุกต์ใช้กับกิจกรรมด้านต่างๆ มาเป็นระยะเวลานานแล้ว โดยจะขอกล่าวถึงตัวอย่างการประยุกต์ใช้ GIS ที่โดดเด่น 4 ตัวอย่าง ได้แก่ การนำ GIS มาใช้เป็นเครื่องมือช่วยในการตัดสินใจผลิตพืชข้าว เป็นเครื่องมือช่วยในการวางแผนการเดินทาง นำมาใช้กับงานด้านการจัดเก็บภาษี และงานด้านอสังหาริมทรัพย์

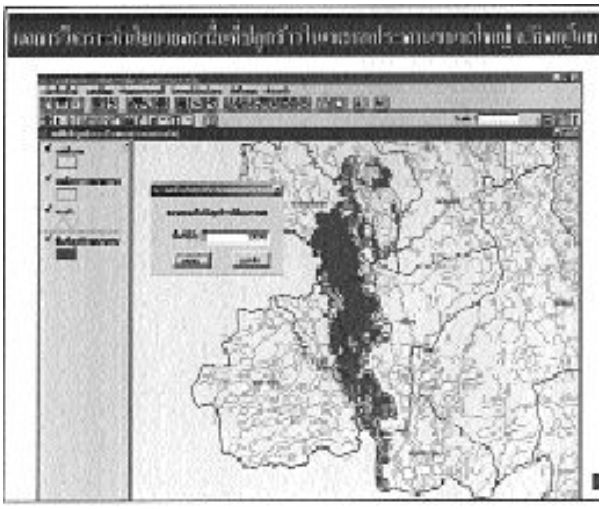
3.1 การประยุกต์ใช้ GIS เป็นเครื่องมือช่วยในการตัดสินใจผลิตพืชข้าว

โปรแกรม “โพสพ 1.0” [11] ดังแสดงในรูปที่ 13 ได้ถูกพัฒนาขึ้นโดยทีมงานวิจัยโครงการระบบสนับสนุนการตัดสินใจผลิตพืชข้าวในภาคเหนือ ของมหาวิทยาลัยเชียงใหม่ โดยการวิจัยนี้ได้แบ่งขั้นตอนการวิจัยออกเป็น 3 ส่วน ดังนี้

ขั้นตอนที่ 1: ออกแบบและสร้างสมมติฐานข้อมูลเชิงพื้นที่ โดยหาข้อมูลมาวิเคราะห์ เช่น แหล่งปลูกข้าว ชนิดของดิน ภูมิอากาศ สภาพความสูงต่ำของภูมิอากาศ พื้นที่รับน้ำชลประทาน และพื้นที่เขตน้ำท่วม นอกจากนี้ยังมีข้อมูลโครงสร้างพื้นฐานที่สนับสนุนการขนส่งและการตลาดข้าว อันได้แก่ เขตการปกครอง ถนน และโรงสีข้าว

ขั้นตอนที่ 2: การทดสอบ โดยจำลองพื้นที่ในแปลงทดสอบก่อน แล้วนำพันธุ์ข้าวที่นิยมปลูกในภาคเหนือ เช่น สันป่าตอง ดอกมะลิ มาทำการเพาะปลูก ซึ่งผลการทดสอบในแปลงนาพบว่า แบบจำลองสามารถจำลองการเจริญเติบโต และได้ผลผลิตข้าวพันธุ์ดังกล่าวได้เป็นอย่างดี และโครงการวิจัยนี้สามารถนำมาใช้ในการประมาณการณ์เพิ่มผลผลิตข้าวในภาคเหนือ

ขั้นตอนที่ 3: เป็นงานส่วนสุดท้ายและค่อนข้างสำคัญ เพราะเป็นการนำเอาข้อมูลจากฐานข้อมูล และผลของแปลงจำลองการปลูกข้าวทั้ง 2 ส่วนมาประกอบเข้าภายใต้โปรแกรมโพสพ 1.0 ซึ่งเป็นโปรแกรม GIS เพื่อใช้เป็นเครื่องมือในการตัดสินใจ ทำให้เกษตรกรสามารถเลือกพื้นที่นาที่จะเพิ่มผลผลิตได้ หลังจากนั้นเกษตรกรก็สามารถเรียกดูข้อมูลปัจจัยทางกายภาพที่มีผลต่อผลผลิต โดยเกษตรกรสามารถกำหนดพันธุ์ข้าวที่จะปลูก วันที่ปลูก ปริมาณข้าว วันที่ใส่ปุ๋ย และจำนวนครั้งของการใส่ปุ๋ย



รูปที่ 13. แสดงหน้าจอจากโปรแกรม "โพสพ 1.0" [11, หน้า 10]

รูปที่ 14. แสดงตัวอย่างหน้าจอการค้นหาตำแหน่งสถานที่ทาง Internet [12]

โครงการ "โพสพ" นี้จัดได้ว่าเป็นโครงการนำร่องในการนำ GIS มาประยุกต์ใช้ในการทำการเกษตรกรรม ซึ่งโครงการในลักษณะนี้น่าจะเป็นโครงการที่ได้รับการส่งเสริมมากขึ้น เพราะว่าประเทศไทยเราเป็นประเทศเกษตรกรรม ยิ่งไปกว่านั้น การประยุกต์ใช้ GIS จะช่วยให้เกษตรกรสามารถลดต้นทุนและเพิ่มผลผลิตทางการเกษตร เป็นการสร้างรายได้ และทำให้ชีวิตประชากรส่วนใหญ่ของประเทศมีความเป็นอยู่ที่ดีขึ้น ยังให้เกิดประโยชน์ในภาพรวมของประเทศต่อไป

3.2 การประยุกต์ใช้ GIS เป็นเครื่องมือช่วยตัดสินใจวางแผนการเดินทาง

เส้นทางการจราจรในกรุงเทพมหานครมีความสลับซับซ้อน และสภาพการจราจรก็มีความแออัดสูง นักเดินทางอาจต้องใช้เวลาในการเดินทางโดยเฉพาะช่วงเวลาเร่งด่วนนานมาก ดังนั้นการเลือกเส้นทางที่เหมาะสมก่อนออกเดินทาง รวมไปถึงการทราบข้อมูลความหนาแน่นของเส้นทางจราจรที่จะใช้เดินทาง จึงเป็นวิธีการเตรียมตัววิธีหนึ่งที่จะช่วยให้นักเดินทางประหยัดเวลาในการเดินทางได้

ด้วยเหตุนี้บริษัทแห่งหนึ่งจึงได้ออกแบบเว็บไซต์ (website) ที่ให้ข้อมูลเกี่ยวกับสถานที่และการเดินทาง โดยนักเดินทางหรือผู้ขับขี่สามารถเข้าไปค้นหาตำแหน่งที่ตั้งสถานที่ที่ต้องการ ผ่านทางระบบ อินเทอร์เน็ตได้ โดยเข้าไปที่ www.thaimapguide.com [12] ดังแสดงในรูปที่ 14

นอกจากความสามารถในการสืบค้นตำแหน่งสถานที่แล้ว เว็บไซต์ดังกล่าวยังมีการเชื่อมต่อเข้ากับ GIS โดยให้ข้อมูลความหนาแน่นของการจราจรในแต่ละเส้นทาง ด้วยการแสดงสีที่แตกต่างกัน (ดังรูปที่ 15) และแสดงตำแหน่งที่มีการเกิดอุบัติเหตุหรือไฟไหม้ เพื่อบ่งชี้ให้นักเดินทางหรือผู้ขับขี่สามารถเลือกใช้เส้นทางการจราจรที่มีความคล่องตัวสูงไม่แออัด และคาดว่าจะเป็นเส้นทางที่ใช้เวลาในการเดินทางน้อยที่สุด นับเป็นเทคโนโลยีที่จะช่วยพัฒนาการจราจร โดยเฉพาะอย่างยิ่งการจราจรในเมืองใหญ่ ดังเช่น กรุงเทพมหานคร หรือ เมืองเชียงใหม่

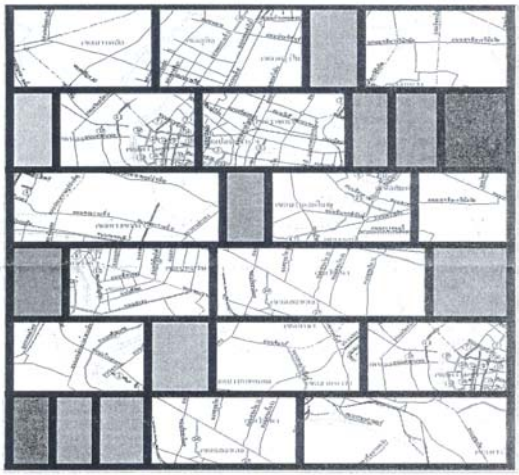
รูปที่ 15. แสดงหน้าจอแสดงความหนาแน่นการจราจรผ่านระบบ Online [12]

3.3 การประยุกต์ใช้ GIS กับการจัดเก็บภาษี

GIS ได้ถูกนำมาใช้เพิ่มประสิทธิภาพการจัดเก็บภาษีในประเทศไทยตั้งแต่ปี 2541 โดยเริ่มโครงการทดลองในเขตคลองเตย กรุงเทพมหานคร [13] พบว่า ในอดีต ร้อยละ 30 ของผู้เสียภาษีไม่มีชื่ออยู่ในระบบการจัดเก็บภาษีแบบเดิม ส่งผลให้รัฐบาลไม่สามารถจัดเก็บภาษีได้อย่างเต็มเม็ดเต็มหน่วย ทั้งนี้ ด้วยวิธีการจัดเก็บภาษีแบบใหม่นี้ จะช่วยแก้ปัญหาดังกล่าวให้หมดไป ช่วยเพิ่มรายได้การเก็บภาษีในเขตคลองเตยประมาณ ร้อยละ 30-50 หรือเมื่อเทียบเป็นเงินประมาณ 40 ล้านบาท ภายในระยะ 2 ปีนับแต่เริ่มต้นโครงการทดลองในเขตคลองเตย

นอกจากนี้ ระบบการจัดเก็บภาษีแบบใหม่ ยังช่วยให้การทำงานของเจ้าหน้าที่สะดวกขึ้น เพราะว่าเจ้าหน้าที่สามารถตรวจสอบสถานะการเสียภาษีจากแผนที่ภาษีบนหน้าจอ ซึ่งจะแสดงสถานะการชำระภาษีในแต่ละบริเวณหรือในแต่ละรายที่สนใจด้วยความแตกต่างของสี (ดูรูปที่ 16) กล่าวคือ สีแดง หมายถึง ยังมิได้ชำระภาษี และสีเขียว หมายถึง ได้ชำระภาษีไปแล้ว สำหรับผู้ที่ยังมิได้ชำระภาษีนั้น เจ้าหน้าที่ก็จะพิมพ์และส่งจดหมายเตือนไปยังผู้นั้นได้ทันที และยิ่งไปกว่านั้น สำหรับเจ้าหน้าที่ภาษีระดับสูงจะสามารถนำข้อมูลของการจัดเก็บภาษีทั้งหมดจากระบบมาใช้วิเคราะห์และจัดทำนโยบายการจัดเก็บภาษีในอนาคตได้โดยง่าย

จะเห็นว่า GIS มีส่วนช่วยในการปรับปรุงและเพิ่มประสิทธิภาพการจัดเก็บภาษีของรัฐบาลไทย ช่วยให้เจ้าหน้าที่ทำงานได้สะดวกและรวดเร็วมากขึ้นกว่าเดิม ลดปัญหาการจัดเก็บภาษีไม่เต็มเม็ดเต็มหน่วยเช่นในอดีตที่ผ่านมาได้เป็นอย่างดี นับได้ว่าเป็นโครงการที่ประสบความสำเร็จอย่างสูง ส่งผลให้ได้รับการสนับสนุนและขยายผลการใช้งานไปในวงกว้าง



รูปที่ 16. แสดงสถานะการจัดเก็บภาษีด้วย GIS [13]

จะเห็นว่า GIS มีส่วนช่วยในการปรับปรุงและเพิ่มประสิทธิภาพการจัดเก็บภาษีของรัฐบาลไทย ช่วยให้เจ้าหน้าที่ทำงานได้สะดวกและรวดเร็วมากขึ้นกว่าเดิม ลดปัญหาการจัดเก็บภาษีไม่เต็มเม็ดเต็มหน่วยเช่นในอดีตที่ผ่านมาได้เป็นอย่างดี นับได้ว่าเป็นโครงการที่ประสบความสำเร็จอย่างสูง ส่งผลให้ได้รับการสนับสนุนและขยายผลการใช้งานไปในวงกว้าง

3.4 การประยุกต์ใช้ GIS กับงานด้านอสังหาริมทรัพย์

สถาบันการเงินบางแห่งในประเทศไทย มีการนำ GIS มาประยุกต์ใช้กับงานด้านอสังหาริมทรัพย์ โดยเฉพาะอย่างยิ่งในเขตกรุงเทพมหานครและปริมณฑล [14] เพื่อใช้ในการวิเคราะห์สินเชื่อโครงการ ประเมินราคาหลักประกันเบื้องต้น และความเสี่ยงของการปล่อยสินเชื่อ ซึ่งระบบดังกล่าวนี้ประกอบไปด้วยข้อมูลแผนที่ ข้อมูลโครงการบ้านจัดสรร ข้อมูลราคาที่ดิน ข้อมูลถนนตัดใหม่ ข้อมูลแนวเวนคืน ข้อมูลสีผังเมือง และความหนาแน่นของประชากร ฯลฯ โดยที่สถาบันการเงินสามารถสร้างชั้นข้อมูลของตนเอง และทำการเชื่อมโยงโดยอาศัยตำแหน่งพิกัดบนแผนที่ได้ และนอกจากนี้ยังสามารถเชื่อมโยงเข้ากับแฟ้มข้อมูลต่างๆ ไม่ว่าจะเป็นแผนผัง รูปโครงการ และแผนภูมิเข้าไว้ด้วยกัน ทั้งนี้ เพื่อความสะดวกในการประเมินราคาหลักประกันเบื้องต้น โดยสามารถเปรียบเทียบหลักประกันของลูกค้ากับหลักประกันในบริเวณใกล้เคียงที่เคยประเมินเอาไว้แล้ว ทั้งยังสามารถเปรียบเทียบราคากับโครงการจัดสรรในบริเวณเดียวกันจากฐานข้อมูลของบริษัทได้อีกด้วย

ในด้านการค้นหาข้อมูลสามารถทำได้โดยการระบุตำแหน่งที่ต้องการ ระบบก็จะทำการค้นหาและแสดงให้เห็นอย่างชัดเจนบนแผนที่ ซึ่งผู้ใช้สามารถเลือกดูข้อมูลที่มีอยู่หลายประเภทได้ตามความต้องการ ทำให้มองเห็นความแตกต่างและเข้าใจมูลค่าของหลักทรัพย์ได้ดีขึ้น ตัวอย่าง เช่น หลักทรัพย์ 2 แห่งหรือมากกว่า 2 แห่ง ที่อยู่ต่างตำบล/อำเภอ อาจทำให้ผู้อ่านคิดว่าหลักประกันอยู่ไกลกัน แต่เมื่อมองจากแผนที่กลับพบว่าอยู่ใกล้กัน หรือห่างกันเพียงแค่ข้ามถนนเท่านั้น ดังนั้นราคาหลักทรัพย์ทั้ง 2 แห่ง จึงควรมีค่าใกล้เคียงกัน หรือ ในกรณีที่หลักทรัพย์ 2 แห่ง หรือมากกว่า 2 แห่งขึ้นไป ตั้งอยู่ในอำเภอ/ตำบลเดียวกัน แต่มีราคาประเมินแตกต่างกันมาก ทั้งนี้เนื่องจากสีผังเมืองนั่นเองที่เป็นตัวทำให้ราคาแตกต่างกัน ทำให้สามารถทราบมูลค่าหลักทรัพย์ในเบื้องต้นจากการดูตำแหน่งบนแผนที่ ประกอบกับมูลค่าของหลักทรัพย์ในบริเวณใกล้เคียงที่สถาบันการเงินเคยประเมินไว้ ผนวกกับราคาโครงการบ้านจัดสรรที่ยังมีการซื้อขายกันในปัจจุบันที่ได้ทำการสำรวจไว้ จะทำให้ท่านทราบศักยภาพและมูลค่าคร่าว ๆ ในเบื้องต้นได้ (ดังรูปที่ 17) นอกจากนี้ ระบบยังสามารถค้นหาข้อมูลแบบตั้งเงื่อนไขได้ เช่น เมื่อต้องการเรียกหลักประกันที่มีมูลค่าประเมินมากกว่า 25 ล้านบาท และประเมินครั้งล่าสุดเมื่อกลางปี 2542 เพื่อนำข้อมูลมาประเมินใหม่ โดยโปรแกรมจะทำการประมวลผลแล้วแสดงผลลัพธ์บนจอทันที

รูปที่ 17.แสดง หน้าจอจากโปรแกรมสารสนเทศทางอสังหาริมทรัพย์ [14]

การนำ GIS มาใช้จึงช่วยเพิ่มประสิทธิภาพการวิเคราะห์หลักทรัพย์ โดยพนักงานวิเคราะห์หลักทรัพย์สามารถตรวจสอบราคาหลักทรัพย์คร่าวๆ ก่อนการคำนวณมูลค่าหลักทรัพย์จริง โดยการเปิดดูข้อมูลที่เคยให้ราคาไว้ทุกหลักทรัพย์ในบริเวณนั้น นอกจากนี้การเชื่อมโยงข้อมูลระหว่างฝ่าย จะช่วยให้ฝ่ายวิเคราะห์สามารถพิจารณาราคาหลักทรัพย์กับลูกค้าได้ทันทีในเบื้องต้น เป็นการลดเวลาพิจารณาหลักทรัพย์ แก้ปัญหาการคาดเดามูลค่าทรัพย์สิน ซึ่งวิธีนี้ทำให้ประหยัดเวลาและค่าใช้จ่ายที่อาจเกิดขึ้นโดยไม่จำเป็น นอกจากนั้น การมองเห็นภาพรวมของหลักทรัพย์ทั้งหมดขององค์กร จะช่วยให้การตัดสินใจปล่อยสินเชื่อ ซื้อ-ขายหลักทรัพย์ หรือบริหารทรัพย์สิน เป็นไปอย่างมีประสิทธิภาพมากยิ่งขึ้น

ในแง่ของผู้บริหารเอง ผู้บริหารสามารถมองเห็นภาพรวมของหลักทรัพย์ทั้งหมด เนื่องจากผู้บริหารสามารถเข้าไปประมวลข้อมูลจากทุกฝ่ายได้ ทำให้ผู้บริหารมองเห็นภาพการกระจายความเสี่ยงของหลักทรัพย์ทั้งหมดจากการจัดชั้นความเสี่ยงของหลักทรัพย์ และเทียบเคียงศักยภาพของหลักทรัพย์ในแต่ละทำเลด้วยข้อมูลปัจจัยภายในและปัจจัยภายนอก ซึ่งทำให้ผู้บริหารสามารถบริหารหลักทรัพย์ได้ดีและไม่ผิดพลาด

4. บทสรุป

GIS เป็นระบบสารสนเทศที่มีประสิทธิภาพในการจัดเก็บ จัดการ ค้นคืน ค้นหา วิเคราะห์ และแสดงผลข้อมูลเชิงพื้นที่ร่วมกับข้อมูลอรรถาธิบาย ทำให้การประมวลและการวิเคราะห์ข้อมูลมีประสิทธิภาพมากยิ่งขึ้น เนื่องจากระบบสามารถทำการซ้อนทับชั้นข้อมูลที่ผู้ใช้ต้องการเข้าด้วยกันตามความสัมพันธ์หรือเงื่อนไขที่ผู้ใช้กำหนด อีกทั้งผู้ใช้ยังสามารถสร้างแบบจำลองเพื่อใช้ในการทดลอง ทดสอบ และเปรียบเทียบทางเลือก เพื่อช่วยในการตัดสินใจก่อนที่จะนำไปปฏิบัติจริงได้ โดยผู้ใช้สามารถกำหนดให้ระบบแสดงผลออกมาในรูปของแผนที่ หรือกราฟที่ง่ายต่อการเข้าใจและช่วยให้มองเห็นภาพที่ชัดเจนมากยิ่งขึ้น

ด้วยเหตุนี้จึงทำให้ GIS ได้รับความนิยม และมีการนำมาประยุกต์ใช้กับงานทุกประเภท ทุกวงการ นอกเหนือจากตัวอย่างที่กล่าวมา โดยเราสามารถนำมาประยุกต์ใช้กับงานด้านการอนุรักษ์และจัดการสิ่งแวดล้อม การจัดการอุทยานแห่งชาติ การควบคุมและติดตามมลภาวะ การอนุรักษ์และการจัดการด้านทรัพยากร การจัดการระบบชลประทาน การจัดการด้านสาธารณะภัย การโทรคมนาคม การบริการสาธารณูปโภค การให้บริการห้องสมุด การทหาร การปราบปรามอาชญากรรม การท่องเที่ยว ตลอดจนการวิเคราะห์ด้านการตลาด เป็นต้น

ทั้งนี้การนำ GIS มาประยุกต์ใช้จะมีประสิทธิภาพและประสบความสำเร็จได้ จำเป็นต้องอาศัยองค์ประกอบหลายด้านด้วยกัน ได้แก่

1) ระบบฮาร์ดแวร์ (Hardware) ที่มีประสิทธิภาพ
2) ระบบซอฟต์แวร์ (Software) ที่เหมาะสมกับลักษณะการใช้งาน
3) ระบบข้อมูล (Data) ที่มีการนำเข้าและจัดเก็บอย่างมีประสิทธิภาพและเหมาะสมกับการใช้งาน
4) บุคลากร (Peopleware) ที่มีประสบการณ์ ความรู้ และความชำนาญทั้งทางด้านคอมพิวเตอร์และภูมิศาสตร์
5) วิธีการ และการกำหนดขั้นตอนการปฏิบัติงานที่เหมาะสมกับลักษณะการทำงาน

หากขาดสิ่งหนึ่งสิ่งใดไป จะทำให้ GIS ด้อยประสิทธิภาพลงและทำให้ไม่สามารถสร้างข้อมูลที่มีประโยชน์ต่อการตัดสินใจได้ดีเท่าที่ควรจะเป็น

เอกสารอ้างอิง

[1] อิทธิ ตริสิริสัตยวงศ์. 2540. “โครงการวิจัยเพื่อพัฒนาระบบสารสนเทศปริภูมิเพื่อการค้าการขนส่ง.” รายงานผลงานวิจัย, สถาบันพาณิชยนาวี, จุฬาลงกรณ์มหาวิทยาลัย.

[2] นิภา เจียรภัทรานนท์. 2539. “การประเมินทางธรณีวิทยาสิ่งแวดล้อมโดยใช้ระบบสารสนเทศภูมิศาสตร์ เพื่อการวางแผนการใช้ที่ดินในพื้นที่ จ.สระบุรี ประเทศไทย.” วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต, สหสาขาวิชาวิทยาศาสตร์สภาวะแวดล้อม, จุฬาลงกรณ์มหาวิทยาลัย.

[3] ESRI. 2001. **What is GIS?.** [Online]. Available : http://www.gis.com/whatisgis/index.html.

[4] ESRI. 1997. “GIS Solutions for Production Agriculture.” White Paper. Environmental Systems Research Institute, ESRI.

[5] University of Georgia. 2000. "Precision Farming: An Introduction." Cooperative Extension Service, College of Agricultural and Environmental Sciences, University of Georgia.

[6] University of Washington. 2000. **Pacific Northwest Seismograph Network Information–from the PNSN.** [Online]. Available : http://spike.geophys.washington.edu/recenteqs/.

[7] Fisher, P. 1995. **Innovation in GIS: 2 Selected Papers from the Second National Conference on GIS Research**. London : Taylor & Francis.

[8] Powell, R. et. al. 1993. "Real time passenger information system for the ROMANSE project." pp. 9/1-9/4. in **IEE Colloquium on Public Transport Information and Management Systems.**

[9] Transport Research Laboratory. 2001. **Bus Priority in SCOOT.** [Online]. Available : http://www.scoot-utc.com/frame2/ busprior.htm.

[10] Thong, C.M. and Wong, W.G. 1998. "Using GIS to design a traffic information database for urban transport planning." 21(6) : 423-425. in **Computer, Environment and Urban Systems**.

[11] นิรนาม. 2543, 1 มิถุนายน. "โปรแกรม 'โพสพ' ทางเลือกคนไทยเพิ่มผลผลิตข้าว." **กรุงเทพธุรกิจ**. หน้า 10-11.

[12] Thaimapguide. 2001. **Thaimapguide**. [Online]. Available : http://www.thaimapguide.com.

[13] Sutharoj, P. 2000, 11 April. "BMA improving tax collection with GIS," **The Nation.**

[14] Agency for Real Estate Affairs Company Limited. 2002. **Detail of GeoSearch Program.** [Online]. Available : http://www.area.co.th/bar21e.html.

ผกาสิน พูนพิพัฒน์ เกิดวันที่ 5 ตุลาคม 2513 จบมัธยมศึกษาจากโรงเรียนสตรีวิทยา และศึกษาศาสตร์บัณฑิต (ภาษาอังกฤษ) จากมหาวิทยาลัยศิลปากร ปัจจุบันกำลังศึกษาปริญญาโท สาขาการจัดการเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง และทำงานเป็นผู้ช่วยนักการทูต แผนกเศรษฐกิจ สถานเอกอัครราชทูตญี่ปุ่น ประจำประเทศไทย โดยมีความสนใจเป็นพิเศษเกี่ยวกับ GIS แล E-Learning
เบอร์โทรศัพท์ที่ติดต่อ 0-2252-6151 ต่อ 239
Emial : pakasin@yahoo.com

ภัทรชัย ลลิตโรจน์วงศ์ ปัจจุบันเป็นอาจารย์ประจำคณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง โดยมีความสนใจเป็นพิเศษเกี่ยวกับ Artificial Intelligence, Cognitive Modeling, Automated Reasoning, Neural Network Application, Intelligent Agents/ Intelligent Systems/ Expert Systems, Object-Oriented Methodology , และ Database, Data Warehouse, Data Mining
เบอร์โทรศัพท์ที่ติดต่อ 0-2737-2551-4 ต่อ 534
E-mail: pattarachai@it.kmitl.ac.th, pattarachai_l@hotmail.com

การใช้เทคนิคดาต้าไมน์นิงเพื่อพัฒนาคุณภาพการศึกษาคณะวิศวกรรมศาสตร์

กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ และธนาวินท์ รักธรรมานนท์
อาจารย์ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์
และนิสิตปริญญาโทวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

ABSTRACT- The aim of our work is to contribute to an improved quality of engineering graduates by proposing a data mining system that assists students in selecting the appropriate major according to their profile and in their course registering.

KEY WORDS – data mining, knowledge

บทคัดย่อ - บทความนี้เป็นการศึกษาและวิเคราะห์ระบบฐานข้อมูลนิสิต โดยนำความรู้ทางด้านดาต้าไมน์นิงมาประยุกต์ใช้กับข้อมูลนิสิตคณะวิศวกรรมศาสตร์ เพื่อเป็นแนวทางในการแก้ไขปัญหาต่างๆ อาทิเช่น ปัญหาการเลือกสาขาวิชาไม่ตรงกับความสามารถที่แท้จริง ปัญหาผลการเรียนของนิสิตตกต่ำจนต้องออกจากสถาบันการศึกษา อันเป็นผลทำให้ไม่ได้มาซึ่งบุคลากรที่มีความสามารถสูงสุด

คำสำคัญ – ดาต้าไมน์นิง, ความรู้

1. บทนำ

การสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจบนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Databases: KDD) หรือที่เรียกกันว่า ดาต้าไมน์นิง (Data Mining) เป็นสาขาหนึ่งในวิทยาศาสตร์คอมพิวเตอร์ที่กำลังได้รับความสนใจอย่างสูงในปัจจุบัน [2, 5, 8] เมื่อใช้เทคนิคดาต้าไมน์นิง ข้อมูลขนาดใหญ่จะถูกวิเคราะห์และสืบค้นความรู้หรือสิ่งที่สำคัญออกมา จากนั้นจะรวบรวมความรู้ที่ได้ให้อยู่ในรูปฐานความรู้ (Knowledge Base) เพื่อนำไปใช้ประโยชน์ต่อไป โดยในปัจจุบันได้มีการนำเทคนิคดาต้าไมน์นิงไปประยุกต์ใช้ในงานด้านต่างๆ มากขึ้น ทั้งในด้านการส่งเสริมการขายสินค้าในห้างสรรพสินค้า, ด้านการวิเคราะห์เครดิตลูกค้าในธนาคาร และในด้านอื่นๆ อีกมาก แต่ไม่มีการนำมาประยุกต์กับด้านการศึกษาอย่างจริงจัง ทั้งที่ในปัจจุบันตามสถาบันการศึกษาส่วนใหญ่มีข้อมูลนิสิตที่ได้จัดเก็บไว้เป็นเวลานาน แต่มิได้ถูกนำมาใช้ให้เกิดประโยชน์เท่าที่ควร โดยในบทความนี้ได้นำเสนอเทคนิคต่างๆ ที่สำคัญของดาต้าไมน์นิงมาประยุกต์ใช้ในการสืบค้นสิ่งที่น่าสนใจออกมาจากข้อมูลนิสิต

ข้อมูลนิสิตที่นำมาวิจัยนี้เป็นข้อมูลของนิสิตคณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ โดยประกอบไปด้วยข้อมูล 2 ส่วน ส่วนแรกคือ ฐานข้อมูลการลงทะเบียนเรียนของนิสิต ที่แต่และแถวแสดงถึงวิชาที่นิสิตได้ลงทะเบียนเรียนและผลการเรียนในวิชาต่างๆ และส่วนที่ 2 คือ ฐานข้อมูลประวัติส่วนตัวของนิสิต เช่น อายุ เพศ ที่อยู่ ประวัติการศึกษาก่อนเข้ามาในมหาวิทยาลัย เกรดเฉลี่ยสะสม เป็นต้น

ในบทความนี้ เริ่มต้นด้วยบทนำที่กล่าวถึงความสำคัญและแนวทางในการพัฒนาคุณภาพการศึกษาโดยใช้เทคนิคดาตาไมน์นิงในส่วนแรก ในส่วนที่ 2 กล่าวถึงรายละเอียดของแนวคิดและวิธีการที่นำมาประยุกต์ใช้ในงานนี้ ได้แก่ การสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจบนฐานข้อมูลขนาดใหญ่ ตามด้วยเทคนิคดาต้าไมน์นิงที่สำคัญ 3 ประการคือ การค้นหากฎความสัมพันธ์ (association rule discovery) , การจำแนกข้อมูล (data classification) และ การพยากรณ์ข้อมูล (data prediction) ในส่วนที่ 3 กล่าวถึงการนำเทคนิคดาต้าไมน์นิงมาประยุกต์ใช้ในการช่วยในการช่วยนิสิตเลือกสาขาวิชาที่เหมาะสม ส่วนที่ 4 กล่าวถึงการนำเทคนิคดาต้าไมน์นิงมาประยุกต์ใช้ในการทำนายเกรดแต่ละรายวิชาในภาคการศึกษาต่อไป ส่วนที่ 5 เป็นการสรุปงานวิจัยนี้และกล่าวถึงการปรับปรุงงานต่อไปในอนาคต

2. แนวคิดและวิธีการเบื้องต้น

การสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจบนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Databases: KDD) หรือที่เรียกกันว่าดาต้าไมน์นิงเป็นเทคนิคที่ใช้จัดการกับข้อมูลขนาดใหญ่ โดยจะนำข้อมูลที่มีอยู่มาวิเคราะห์แล้วดึงความรู้ หรือสิ่งที่สำคัญออกมา เพื่อใช้ใน

การวิเคราะห์ หรือทำนายสิ่งต่างๆ ที่จะเกิดขึ้น กระบวนการ KDD ประกอบไปด้วย 3 ส่วนหลักๆ คือ

1) Pre-processing คือ ขั้นตอนการเตรียมข้อมูลให้เหมาะสม และให้อยู่ในรูปแบบที่สามารถนำไปใช้งานได้

2) Data Mining คือ ขั้นตอนการเลือกเทคนิคที่เหมาะสม สำหรับงานที่ต้องการ โดยสามารถรวมเทคนิคได้มากกว่าหนึ่งเทคนิคมา ประมวลผลเพื่อดึงความรู้หรือสิ่งที่น่าสนใจจากข้อมูลที่ผ่านขั้นตอน Pre-precessing แล้ว โดยผลลัพธ์ที่ได้จากขั้นตอนนี้คือฐานความรู้

3) Post-processing คือขั้นตอนการนำฐานความรู้ที่ได้จากขั้นตอนดาต้าไมน์นิงมาทดสอบและพิจารณาว่าถูกต้องตามความต้องการหรือไม่ ซึ่งบางครั้งอาจต้องปรับแก้ค่าและนำเข้าสู่ขั้นตอนดาต้าไมน์นิงใหม่อีกครั้ง จนกว่าจะได้ความรู้หรือสิ่งที่น่าสนใจตามที่ต้องการออกมา

จะเห็นได้ว่าดาต้าไมน์นิงเป็นองค์ประกอบหลักที่สำคัญในกระบวนการ KDD ในลำดับต่อไป เราจะกล่าวถึงเทคนิคดาต้าไมน์นิงที่สำคัญ 3 เทคนิค ที่ได้นำมาประยุกต์ใช้ คือ การค้นหากฎความสัมพันธ์ (Association rule discovery) , การจำแนกประเภทข้อมูล (data classification) และ การพยากรณ์ข้อมูล (data prediction)

2.1 การค้นหากฎความสัมพันธ์ (Association rule discovery)

การค้นหากฎความสัมพันธ์ [7] เป็นการค้นหาความสัมพันธ์ของข้อมูลจากฐานข้อมูลขนาดใหญ่ที่มีอยู่ เพื่อนำไปใช้ในการวิเคราะห์ หรือทำนายปรากฏการณ์ต่างๆ โดยเทคนิคนี้ใช้กันอย่างแพร่หลายในการขายสินค้า หรือการวิเคราะห์ข้อมูลที่เป็นทรานแซกชัน เราขอยกตัวอย่างการนำเทคนิคการค้นหากฎความสัมพันธ์มาประยุกต์ใช้ในข้อมูลการขายสินค้าดังตารางที่ 1

ตารางที่ 1 แสดงตัวอย่างข้อมูลการขายสินค้า

หมายเลขทรานแซกชัน	สินค้าที่ซื้อ
1	น้ำตาล, ขนมปัง
2	นมสด, น้ำตาล, ขนมปัง
3	ขนมปัง
4	นมสด, น้ำตาล

จากตารางที่ 1 สามารถบอกได้ว่าน้ำตาลและขนมปังจะถูกซื้อด้วยกันในทรานแซกชันที่ 1 หลังจากที่นำข้อมูลไปผ่านกระบวนการดาต้าไมน์นิงแล้ว จะได้ความสัมพันธ์อยู่ในรูป X -> Y หมายความว่า เมื่อซื้อ X แล้วจะซื้อY ด้วย ยกตัวอย่างเช่น นมสด -> น้ำตาล หมายความว่า เมื่อลูกค้าซื้อนมสดแล้วจะซื้อน้ำตาลด้วย

กฎหรือความสัมพันธ์ต่างๆ ที่ได้มาจากกระบวนการดาต้าไมน์นิงนั้นจะมีความน่าสนใจต่างกัน จึงต้องมีเกณฑ์ในการวัดความน่าสนใจของกฎ การค้นหากฎความสัมพันธ์นี้ มีเกณฑ์ในการวัดความน่าสนใจ 2 แบบ ได้แก่ ค่าสนับสนุน (support) และค่าความมั่นใจ (confidence)

- ค่าสนับสนุน (support) แสดงถึงเปอร์เซ็นต์ของข้อมูลที่เป็นไปตามกฎจากข้อมูลทั้งหมด จากกฎ นมสด -> น้ำตาล เมื่อพิจารณาจากข้อมูลในตารางที่ 1 จะเห็นว่าข้อมูลในทรานแซกชันที่ 2 และ 4 ที่เป็นไปตามกฎ จากข้อมูลทั้งหมด 4 ทรานแซกชัน ดังนั้น จะได้ว่าค่าสนับสนุนจากกฎนี้คือ 2/4 คือ 50 %
- ค่าความมั่นใจ (confidence) แสดงถึงความน่าเชื่อถือของกฎ จากกฎ นมสด -> น้ำตาล จะพบว่า มีทรานแซกชันที่ซื้อนมสด 2 ทรานแซกชัน และมีทรานแซกชันที่ซื้อทั้งนมสดและน้ำตาลพร้อมกัน 2 ทรานแซกชัน ดังนั้น จากกฎนี้ มีค่าความมั่นใจ เท่ากับ 100 %

เราจะพิจารณาเฉพาะความสัมพันธ์ที่มีค่าสนับสนุนและค่าความมั่นใจสูงกว่าค่าสนับสนุนต่ำสุด (minimum support) และ ค่าความมั่นใจต่ำสุด (minimum confidence) ตามลำดับ จากตารางที่ 1 สมมติกำหนดค่าสนับสนุนต่ำสุดเท่ากับ 50 % และความมั่นใจต่ำสุดเท่ากับ 70 % จะได้ว่ามีเพียงกฎ นมสด -> น้ำตาลเท่านั้นที่ผ่านเกณฑ์ตามที่กำหนดไว้

2.2 การจำแนกประเภทข้อมูล (Data Classification)

การจำแนกประเภทข้อมูล [3] เป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ โดยจะนำข้อมูลส่วนหนึ่งมาสอนให้ระบบเรียนรู้ (training data) เพื่อจำแนกข้อมูลออกเป็นกลุ่มตามที่ได้กำหนดไว้ ผลลัพธ์ที่ได้จากการเรียนรู้คือ โมเดลจำแนกประเภทข้อมูล (classifier model) และจะนำข้อมูลส่วนที่เหลือจากข้อมูลสอนระบบเป็นข้อมูลที่ใช้ทดสอบ (testing data) ซึ่งกลุ่มที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับกลุ่มที่หามาได้จากโมเดลเพื่อทดสอบความถูกต้องและปรับปรุงโมเดลจนกว่าจะได้ค่าความถูกต้องในระดับที่น่าพอใจ หลังจากนั้น เมื่อมีข้อมูลใหม่เข้ามา เราจะนำข้อมูลมาผ่านโมเดล โดยโมเดลจะสามารถทำนายกลุ่มของข้อมูลนี้ได้

ต้นไม้ช่วยการตัดสินใจ (Decision tree) [4] เป็นวิธีหนึ่งที่สำคัญในการจำแนกประเภทข้อมูล โดยต้นไม้ช่วยการตัดสินใจจะมีลักษณะคล้ายโครงสร้างต้นไม้ที่แต่ละโหนดแสดงคุณลักษณะ (attribute), แต่ละกิ่งแสดงผลในการทดสอบ และลีฟโหนด (leaf node) แสดงกลุ่มที่กำหนดไว้ ซึ่งต้นไม้ช่วยการตัดสินใจนี้ง่ายต่อการปรับเปลี่ยนเป็นกฎการจำแนกประเภทข้อมูล (classification rule)

2.3 การพยากรณ์ข้อมูล (Data Prediction)

การพยากรณ์ข้อมูล [3] เป็นกระบวนการสร้างโมเดลเพื่อทำนายหาค่าที่

ต้องการจากข้อมูลที่มีอยู่ โดยมีกระบวนการสร้างโมเดลคล้ายกับการจำแนกประเภทข้อมูลดังที่ได้กล่าวมาข้างต้น ต่างกันตรงที่การพยากรณ์ข้อมูลไม่มีการจัดข้อมูลเข้ากลุ่มตามที่ได้กำหนด แต่การพยากรณ์ข้อมูลนี้เป็นการพยากรณ์หาค่าที่ต้องการออกมาเป็นตัวเลข ตัวอย่างเช่น หายอดขายของเดือนถัดไปจากข้อมูลการขายทั้งหมดที่ผ่านมา หรือทำนายเกรดเฉลี่ยของนักเรียนในปีการศึกษาหน้า จากข้อมูลการลงทะเบียนของนิสิตทั้งหมด เป็นต้น

3. การนำเทคนิคดาต้าไมน์นิงมาประยุกต์ใช้ในการช่วยนิสิตเลือกสาขาวิชาที่เหมาะสม

การเลือกสาขาวิชาที่เหมาะสมของนิสิตนั้น เนื่องด้วยนิสิตยังขาดประสบการณ์ และไม่รู้จักแต่ละสาขาวิชามากพอ นิสิตส่วนใหญ่จึงใช้ความรู้สึกความชอบ หรือสภาพแวดล้อมทั้งเพื่อน หรือผู้ปกครองเป็นหลักใหญ่ โดยอาจไม่ทราบถึงสาขาวิชาที่เหมาะสมกับความสามารถและลักษณะเฉพาะของตัวเอง จึงอาจทำให้เมื่อเข้าไปเรียนจริงๆ แล้วเพิ่งค้นพบตัวเองในภายหลังว่าไม่เหมาะสมกับสาขาวิชานี้ จนอาจทำให้เกิดผลกระทบต่างๆ ตามมา โครงงานนี้ได้นำเทคนิคใหม่ทางวิทยาการคอมพิวเตอร์ คือ เทคนิคดาต้าไมน์นิงมาประยุกต์ใช้กับข้อมูลนิสิต เพื่อช่วยในการชี้แนะแนวทางการเลือกสาขาวิชาที่เหมาะสมกับนิสิตแต่ละคนให้มากที่สุด

ในการที่จะบรรลุให้ได้ตามวัตถุประสงค์ คือ การใช้เทคนิคดาต้าไมน์นิงเพื่อเลือกสาขาวิชาที่เหมาะสมให้กับนิสิตนั้น สามารถทำได้หลายแนวทาง ในบทความนี้ขอเสนอ 2 แนวทางที่มีผลการทดสอบความถูกต้องค่อนข้างสูง คือ การทำเทคนิคการจำแนกประเภทข้อมูล และเทคนิคการพยากรณ์ข้อมูลมาสร้างโมเดล

การสร้างโมเดลสำหรับการเลือกสาขาวิชานั้นสามารถทำได้หลายแนวทาง แนวทางหนึ่งคือการนำข้อมูลของนิสิตที่เรียนดีในทุกวิชามา สร้างโมเดลกลางการจำแนกประเภทข้อมูล [6] โดยแต่ละโหนดภายในต้นไม้ช่วยการตัดสินใจ (decision tree) บ่งบอกถึงลักษณะและผลการเรียนในรายวิชาต่างๆ ของนิสิตและคลาสปลายทางแทนสาขาวิชาต่างๆ เพื่อที่จะทำนายว่าลักษณะของนิสิตแต่ละคนนั้นคล้ายคลึงกับลักษณะของนิสิตที่เรียนดีในสาขาวิชาใดมากที่สุด วิธีดังกล่าวมีข้อดีคือสามารถทำนายสาขาวิชาที่เหมาะสมที่สุดให้กับนิสิตได้ แต่วิธีการดังกล่าวนั้นมีข้อเสียบางประการ เช่น โมเดลจะทำนายแนวโน้มโอนเอียงไปทางสาขาวิชาที่มีจำนวนนิสิตมากเป็นผลทำให้ความถูกต้องของโมเดลที่ได้ค่อนข้างต่ำ คือประมาณ 50 % ดังนั้นเราจึงได้คิดวิธีการสร้างโมเดลแบบอื่นเพื่อปรับปรุงประสิทธิภาพให้มากขึ้นกว่าเดิม

ในลำดับต่อมา เราได้สร้างโมเดลจำแนกประเภทข้อมูลสำหรับแต่ละสาขาวิชาโดยพิจารณาว่านิสิตเหมาะสมกับสาขาวิชานั้นๆ หรือไม่ และโมเดลการพยากรณ์ข้อมูลในแต่ละสาขาวิชาเพิ่มเติมเพื่อลดข้อผิดพลาดที่เกิดขึ้นจากโมเดลกลางการจำแนกประเภทข้อมูล ในส่วนที่ 3.1 กล่าวถึงโมเดลจำแนกประเภทข้อมูลสำหรับแต่ละสาขาวิชาโดยพิจารณาว่านิสิตเหมาะสมกับสาขาวิชานั้นๆ หรือไม่โดยการใช้เทคนิคการจำแนกประเภทข้อมูล (Classification) และส่วนที่ 3.2 จะนำเสนอโมเดลการพยากรณ์ข้อมูลในแต่ละสาขาวิชาโดยการใช้เทคนิคการพยากรณ์ข้อมูล (Prediction) การนำเสนอในแต่ละส่วนจะกล่าวถึงการเตรียมข้อมูล การสร้างโมเดล การแปลความหมายจากโมเดล การนำความรู้ที่ได้มาใช้งานจริง และการเปรียบเทียบประสิทธิภาพของแต่ละโมเดล ในรายละเอียดต่อไป

3.1 โมเดลการจำแนกประเภทข้อมูลสำหรับแต่ละสาขาวิชาโดยพิจารณาว่านิสิตเหมาะสมกับสาขาวิชานั้นๆ หรือไม่ (Classification data model for each major)

3.1.1 การเตรียมข้อมูล

ในงานวิจัยนี้ใช้ข้อมูลการลงทะเบียนเรียนของนิสิตคณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ตั้งแต่ปี 2535-2542 รวมทั้งสิ้น 476,085 แถว จากนิสิตกว่า 10,000 คน โดยข้อมูลชุดนี้ ประกอบไปด้วยข้อมูล 2 ส่วน ส่วนแรกคือ ข้อมูลประวัติส่วนตัวของนิสิต เช่น ชื่อ ที่อยู่ ภูมิลำเนา อายุ ฯลฯ ดังตารางที่ 2

ตารางที่ 2 แสดงตัวอย่างข้อมูลประวัติส่วนตัวนิสิต

Stu_code	Sex	Address	SchoolGPA	…	GPA
37058063	male	Bangkok	2.5	….	2.3
37058167	male	Songkla	3.4	….	3.2
…….	….	…….	……	….	….

ข้อมูลอีกส่วนหนึ่งคือ ข้อมูลการลงทะเบียนของนิสิต เนื่องมาจาก คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์นี้จะมีการเลือกสาขาวิชาขึ้นในปี 2 ดังนั้น ในการวิจัยนี้จึงนำเฉพาะวิชาในปีที่ 1 ของการศึกษา (9 วิชา) มาเป็นตัวพิจารณา ในแต่ละแถวประกอบไปด้วยวิชาและผลการเรียนของนิสิตดังตารางที่ 3

ตารางที่ 3 แสดงตัวอย่างข้อมูลการลงทะเบียนเรียนของนิสิต

Stu_code	Sub_code	Section	Term	Year	Grade
37058063	204111	2	1	2537	C+
37058063	403111	6	1	2537	D
37058063	208111	1	1	2537	B+
……	……..	……	…….	……	…

จากตารางที่ 3 ข้อมูลอยู่ในระดับรายวิชาจำเป็นต้องแปลงข้อมูลให้อยู่ในระดับของนิสิต เพื่อให้ได้ลักษณะโครงสร้างข้อมูลตรงตามเป้าหมายที่ต้องการคือศึกษาพฤติกรรมและลักษณะของนิสิตแต่ละคน โดยใช้วิธีแบ่งกลุ่มของวิชาต่างๆ ที่ลงทะเบียนตามรหัสนิสิต จากนั้น จึงแปลงให้แต่ละแถวแทนนิสิตแต่ละคน และคอลัมน์แทนรายชื่อวิชาต่างๆ นอกจากนี้เพื่อต้องการลดการกระจายของข้อมูลเกรดของนิสิต จึงจัดกลุ่มเกรดของนิสิตเป็น 3 กลุ่ม ตามวิธีที่ได้เคยกล่าวมาในส่วน 3.1 แล้ว ผลของตารางแสดงได้ดังในตารางที่ 4

ตารางที่ 4 แสดงตัวอย่างข้อมูลที่จัดในระดับนิสิต

Stu_code	Sex	204111	403111	…	GPA
37058063	male	Medium	Low	…	2.3
37058167	male	High	High	…	3.2
……	….	…..	…..	…	…..

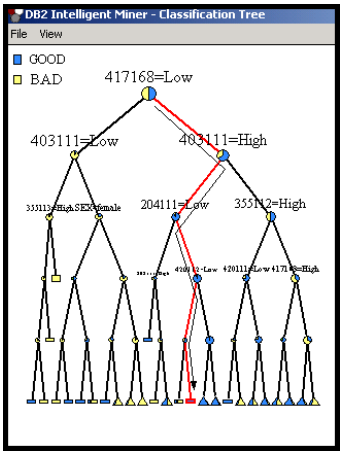
ในการสร้างแต่ละโมเดล แบ่งข้อมูล 70 % จากข้อมูลทั้งหมดเป็นข้อมูลสอนระบบ (training data) เพื่อสร้างโมเดล และข้อมูล 30 % ที่เหลือเป็นข้อมูลที่ใช้ทดสอบ (testing data) เพื่อทดสอบความถูกต้องของโมเดล

3.1.2 การสร้างโมเดล

ในโมเดลการจำแนกประเภทข้อมูลนี้ แต่ละกิ่งแทนลักษณะต่างๆ ของนิสิตดังที่ได้กล่าวมา และลีฟโหนด (คลาสปลายทาง) มี 2 คลาส คือ GOOD และ BAD โดยมีเกณฑ์ว่า GOOD คือนิสิตที่มีเกรดเฉลี่ยสะสมอยู่ในอันดับ 40% แรกของแต่ละสาขาวิชา และ BAD คือ นิสิตที่มีเกรดเฉลี่ยสะสมอยู่ในอันดับ 40% สุดท้ายในแต่ละสาขาวิชา โดยเปอร์เซ็นต์นี้สามารถปรับเปลี่ยนให้เหมาะสมได้ตามลักษณะของข้อมูล ในการวิจัยนี้พบว่า ที่ 40% เป็นอัตราส่วนที่เหมาะสมที่สุด เพราะจะได้ข้อมูลที่นำมาวิจัยมาก และมีการเว้นช่วงผิดพลาดที่สามารถเกิดได้ในช่วงกลางของนิสิต (ช่วงระหว่างเกรด GOOD และ BAD) ดังนั้น ในการวิจัยนี้ข้อมูลสอนระบบคือข้อมูลนิสิตที่มีลักษณะทั้ง GOOD และ BAD
จากวิธีการที่ได้นำเสนอข้างต้น ต้องสร้างโมเดลดังกล่าวกับทุก สาขาวิชา ในบทความนี้ขอยกตัวอย่างกับการนำข้อมูลนิสิตภาคไฟฟ้ามาสร้างโมเดล ผลที่ได้จากข้อมูลและวิธีข้างต้นแสดงดังรูปที่ 1

3.1.3 การแปลความหมายข้อมูลและการนำความรู้ที่ได้มาใช้

เมื่อต้องการทราบสาขาวิชาที่เหมาะสมของนิสิตคนหนึ่ง จะพิจารณาลักษณะต่างๆ ของนิสิตที่แสดงเงื่อนไขกำกับไว้ในโหนดต่างๆ ในต้นไม้ช่วยการตัดสินใจ (decision tree) โดยพิจารณาลงมาตามทางของต้นไม้ที่ละโหนดที่ตรงกับลักษณะของนิสิต จนกระทั่งถึงคลาสปลายทาง หากคลาสปลายทางที่ได้ประกอบด้วยสัดส่วน GOOD มากกว่า BAD มากๆ ย่อมแสดงว่า นิสิตที่มีลักษณะเดียวกับนิสิตคนนี้เข้าภาคไฟฟ้าแล้วจะมีผลการเรียนดี (GOOD) มากกว่าผลการเรียนไม่ดี (BAD) ดังนั้น สาขาวิชาไฟฟ้านี้ เป็นสาขาวิชาหนึ่งที่นิสิตคนนี้ควรพิจารณา โดยจะทดสอบข้อมูลนิสิตคนนี้กับโมเดลของทุกสาขาวิชา และเลือกเฉพาะสาขาวิชาที่คลาสปลายทางมีสัดส่วนของ GOOD มากกว่า BAD ถ้าผลออกมาว่ามีหลายสาขาวิชาที่เหมาะสม สามารถเลือกสาขาวิชาที่ดีที่สุดได้โดยพิจารณาจากสัดส่วนของ GOOD ในโมเดลที่มากกว่าเป็นหลัก หลังการทดสอบ ได้ผลลัพธ์เป็นที่น่าพอใจ คือ ได้ผลการทดสอบถูกต้องเฉลี่ย 84.58 % ในทุกโมเดล



รูปที่ 1 แสดง Decision tree การเลือกสาขาวิชาในโมเดลการจำแนกประเภทข้อมูลของสาขาวิศวกรรมไฟฟ้า

3.2 โมเดลการพยากรณ์ข้อมูลในแต่ละสาขาวิชา (Prediction data model for each major)

3.2.1 การเตรียมข้อมูล

ปรับเปลี่ยนข้อมูลจากการเตรียมข้อมูลที่นำไปสร้างโมเดลการจำแนกประเภทข้อมูลเล็กน้อย คือ แทนผลการเรียนในวิชาต่างๆ ของโมเดลการพยากรณ์ข้อมูลนั้นด้วยตัวเลขที่แทนเกรดจริงๆ ที่นิสิตได้โดยไม่ต้องจัดกลุ่มเพื่อลดการกระจายของเกรด และประวัติส่วนตัวของนิสิตแทนด้วยตัวเลข 1 หรือ 2 นอกจากนี้คอลัมน์ที่ทำนายโดยโมเดลการจำแนกประเภทข้อมูลนั้นเป็นกลุ่มของข้อมูล ดังตารางที่ 4 แต่สำหรับตารางข้อมูลในโมเดลการพยากรณ์ข้อมูลนี้ คอลัมน์การทำนายคือ เกรดเฉลี่ยที่เป็นตัวเลข ดังตารางที่ 5

ตารางที่ 5 แสดงตัวอย่างข้อมูลสำหรับโมเดล Prediction ในแต่ละสาขาวิชา

Stu_code	Sex	Address	204111	…	GPA
37058063	1	1	1	…	2.3

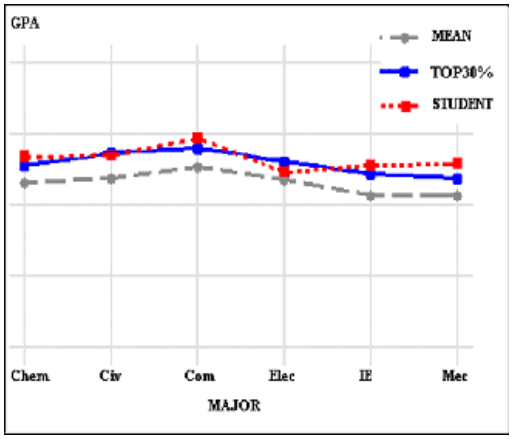
37058167	1	1	3.5	…	3.2
……	….	…..	…..	…	….

3.2.2 การสร้างโมเดล แปลความหมายข้อมูลและการนำความรู้ที่ได้มาใช้

สร้างโมเดลโดยนำผลการเรียนและลักษณะต่างๆ ของนิสิตแต่ละสาขาวิชามาเป็นตัวพิจารณา และสร้างโมเดลเพื่อทำนายเกรดเฉลี่ยตามลักษณะในแต่ละสาขาวิชา หลังจากสร้างโมเดลของทุกสาขาวิชาแล้ว เมื่อมีนิสิตคนหนึ่งต้องการทราบว่าตนเหมาะสมกับสาขาวิชาใด จะนำข้อมูลต่างๆ ของนิสิตคนนั้น ทั้งข้อมูลการศึกษา และข้อมูลประวัติส่วนตัวมาเป็นปัจจัยในแต่ละโมเดลเพื่อที่จะทำนายค่าเกรดเฉลี่ยที่ต้องการออกมา เมื่อนำข้อมูลนิสิตมาผ่านทุกโมเดลแล้ว จะได้เกรดเฉลี่ยของนิสิตคนนั้นตามโมเดลของแต่ละสาขาวิชา นั่นก็คือ เป็นการทำนายผลการเรียนของนิสิตในทุกๆ สาขาวิชา

จากนั้นจะหาเกณฑ์ผลการเรียนในทุกสาขาวิชาเพื่อที่จะทดสอบความถูกต้องของการทำนายเกรดเฉลี่ยในแต่ละโมเดล 1) ค่า MEAN คือ ค่าที่บอกผลการเรียนโดยเฉลี่ยของแต่ละสาขาวิชา 2) ค่า TOP30% คือ ค่าที่บอกถึงผลการเรียนของนิสิตที่ถือได้ว่าเรียนได้เป็นอันดับ 30 % แรกของแต่ละสาขาวิชา โดยถ้าโมเดลทำนายเกรดให้ไปอยู่ในเกณฑ์หนึ่ง แต่เกรดจริงๆ แล้วไปอยู่อีกในเกณฑ์หนึ่งนั้น แสดงว่าโมเดลทำนายผิดพลาด ผลการทดสอบความถูกต้องได้ค่อนข้างสูงมาก ข้อดีของโมเดลการพยากรณ์ข้อมูล

การนำเสนอสาขาวิชาให้กับนิสิตโดยใช้โมเดลการพยากรณ์ข้อมูล เป็นดังรูปที่ 2



รูปที่ 2 แสดงตัวอย่างการนำเสนอสาขาวิชาให้กับนิสิตโดยโมเดลการพยากรณ์ข้อมูลในแต่ละสาขาวิชา

จากกราฟในรูปที่ 2 รายละเอียดต่างๆ ของกราฟเป็นดังนี้

- กราฟเส้นประห่าง แสดงเกรดเฉลี่ยสะสมของนิสิตทุกคนในแต่ละสาขาวิชา
- กราฟเส้นทึบ แสดงเกรดเฉลี่ยสะสมสูงสุด 30 % แรกของนิสิตในแต่ละสาขาวิชา ซึ่งแสดงให้เห็นถึงผลการเรียนของนิสิตที่เรียนดีในแต่ละสาขาวิชา
- กราฟเส้นประถี่ แสดงเกรดเฉลี่ยเมื่อจบการศึกษาที่ทำนายได้ของนิสิตในแต่ละสาขาวิชา

จากรายละเอียดดังกล่าว จะเห็นได้ว่า เส้นกราฟเส้นประห่างและเส้นทึบของนิสิตทุกคนจะมีลักษณะเหมือนกัน เพราะเป็นกราฟที่แสดงผลการเรียนของนิสิตในสาขาวิชาต่างๆ ที่คงที่ ส่วนกราฟเส้นประถี่ของนิสิตแต่ละคนนั้นมีลักษณะแตกต่างกันไปตามค่าเกรดเฉลี่ยที่ทำนายได้ของนิสิตแต่ละคนในแต่ละสาขาวิชา การนำเสนอกราฟเส้นประห่างและเส้นทึบให้กับนิสิตนั้นเพื่อที่นิสิตจะได้ทราบว่าผลการเรียนของนิสิตเป็นอย่างไรเมื่อเปรียบเทียบกับนิสิตส่วนใหญ่ในแต่ละสาขาวิชา จากรูปที่ 2 แสดงให้เห็นว่าถ้านิสิตเรียนในสาขาวิชาวิศวกรรมคอมพิวเตอร์แล้วจะมีแนวโน้มผลการเรียนตอนจบการศึกษาสูงกว่าสาขาวิชาอื่นๆ แต่สาขาวิชาที่นิสิตคนนี้เรียนแล้วมีแนวโน้มผลการเรียนสูงกว่านิสิตส่วนใหญ่ และสูงกว่านิสิตที่เรียนดีมากที่สุด คือ สาขาวิชาวิศวกรรมเครื่องกล เพราะจากกราฟจะเห็นได้ว่าระยะระหว่างกราฟเส้นประถี่กับกราฟเส้นทึบของสาขาวิชาวิศวกรรมเครื่องกลนี้จะมากกว่าของสาขาวิชาอื่นๆ ซึ่งการนำเสนอแบบนี้จะไม่บอกนิสิตว่าสาขาวิชาที่เหมาะสมที่สุด แต่ระบบจะนำเสนอผลที่ได้ในทุกสาขาวิชาให้กับนิสิตเพื่อที่นิสิตจะได้สามารถนำไปพิจารณาประกอบกับความต้องการของนิสิตอย่างเหมาะสม

3.3 การเปรียบเทียบโมเดล

เปรียบเทียบโมเดลที่ได้นำเสนอมาทั้ง 3 โมเดลดังตารางที่ 6

ตารางที่ 6 แสดงการเปรียบเทียบโมเดลทั้งสามโมเดล

โมเดลกลางการจำแนกประเภทข้อมูล	โมเดลการจำแนกประเภทข้อมูลในแต่ละสาขาวิชา	โมเดลการพยากรณ์ข้อมูลในแต่ละสาขาวิชา
1. ความน่าเชื่อถือน้อย ประมาณ 50% เนื่องจากกลุ่มเป้าหมายมาก	1. ผลการทดสอบที่ได้มีความถูกต้องสูง 84.58 %	1. ผลการทดสอบที่ได้มีความถูกต้องสูง 96.84 %
2. ข้อมูลนิสิตในสาขาวิชาต่างๆ มีจำนวนแตกต่างกันมาก ทำให้โมเดลการทำนายโอนเอียงไปทางสาขาวิชาที่มีนิสิตมาก	2. ข้อมูลแต่ละสาขาวิชาไม่ส่งผลกระทบต่อกัน เนื่องมาจากวิธีนี้ได้สร้างโมเดลแยกกันในแต่ละสาขาวิชา	2. ข้อมูลแต่ละสาขาวิชาไม่ส่งผลกระทบต่อกัน เนื่องมาจากวิธีนี้ได้สร้างโมเดลแยกกันในแต่ละสาขาวิชา

3. ต้องมีการจัดกลุ่มผลการเรียนในแต่ละรายวิชา (High, Medium, Low) เพื่อลดการกระจายตัวของข้อมูล ทั้งนี้ ถ้าไม่มีการจัดกลุ่มข้อมูล จะทำให้โมเดลที่ได้กระจายตัว ข้อมูลในแต่ละเส้นทางของโมเดลมีจำนวนน้อย เป็นผลทำให้ความถูกต้องของโมเดลลดลงอย่างมาก	3. ต้องมีการจัดกลุ่มผลการเรียนในแต่ละรายวิชา (High, Medium, Low) เพื่อลดการกระจายตัวของข้อมูล ทั้งนี้ ถ้าไม่มีการจัดกลุ่มข้อมูล จะทำให้โมเดลที่ได้กระจายตัว ข้อมูลในแต่ละเส้นทางของโมเดลมีจำนวนน้อย เป็นผลทำให้ความถูกต้องของโมเดลลดลงอย่างมาก	3. ข้อมูลผลการเรียนในแต่ละรายวิชาเป็นข้อมูลผลการเรียนจริง (A, B+, B, C+, C, D+, D, F) ที่มิได้มีการจัดกลุ่ม ทำให้ข้อมูลที่นำมาสร้างโมเดล Prediction นั้นมีความละเอียดและแม่นยำมากกว่าการจัดกลุ่มดังเช่นโมเดลการจำแนกประเภทข้อมูล
4. โมเดลนำเสนอเพียงสาขาวิชาเดียวที่เหมาะสม ซึ่งส่งผลกระทบโดยตรงกับการตัดสินใจของนิสิต	4. โมเดลนำเสนอเฉพาะสาขาวิชาที่เหมาะสมให้กับนิสิตเท่านั้น สำหรับนิสิตบางส่วนที่มีผลการเรียนดี โมเดลจะเสนอทุกสาขาวิชาให้กับนิสิตเป็นสาขาวิชาที่เหมาะสม และสำหรับนิสิตบางส่วนที่มีผลการเรียนไม่ดี โมเดลจะไม่นำเสนอสาขาวิชาใดๆ ที่เหมาะสมให้กับนิสิตเลย ทำให้การตัดสินใจทั้งหมดไปตกอยู่กับนิสิต โดยที่โมเดลมิได้ช่วยนิสิตในกลุ่มเหล่านี้เลย	4. โมเดลนำเสนอแนวโน้มเกรดเฉลี่ยสะสมเมื่อจบการศึกษาของนิสิตในทุกสาขาวิชา ทำให้นิสิตได้เห็นแนวโน้ม และเห็นความแตกต่างของผลการเรียนของตน เมื่อเข้าไปศึกษาในสาขาวิชาที่แตกต่างกัน นอกจากนี้การนำเสนอได้เพิ่มในส่วนของ MEAN และ TOP30% ทั้งที่ได้เคยกล่าวไป ทำให้ช่วยให้นิสิตได้เห็นความแตกต่างในการเรียนในแต่ละสาขาวิชามากยิ่งขึ้น

4. การนำเทคนิคดาตาไมน์นิงมาประยุกต์ใช้ในการช่วยทำนายแนวโน้มเกรดรายวิชาต่างๆ ในภาคเรียนต่อไป

ปัญหาผลการเรียนตกต่ำของนิสิตนั้น มีสาเหตุหลายประการ เช่น นิสิตไม่มีความตั้งใจในการเรียน , การเตรียมตัวสอบไม่ดีพอ , นิสิตลงวิชาที่ไม่เหมาะกับความสามารถของตน หรือวิชาพื้นฐานที่จำเป็นไม่ดีพอ ในโครงงานนี้ได้เสนอแนวทางการแก้ปัญหาเหล่านี้บางส่วน คือ การทำนายแนวโน้มเกรดแต่ละวิชาในภาคเรียนต่อไป เพื่อเป็นแนวทางให้กับนิสิตในการเลือกลงรายวิชา และสามารถปฏิบัติตนในการเรียนในแต่ละวิชาได้อย่างเหมาะสม ซึ่งโมเดลที่จัดทำขึ้นนี้เป็นสิ่งที่ช่วยแนะแนวในการปฏิบัติให้กับนิสิตเท่านั้น การแก้ปัญหาหลักนั้นขึ้นอยู่กับตัวของนิสิต

งานวิจัยได้นำเทคนิคดาตาไมน์นิงหลายเทคนิคมาประยุกต์ใช้ในการทำนายแนวโน้มเกรด ในบทความนี้ขอนำเสนอการนำเทคนิคการค้นหากฎความสัมพันธ์ (Association rule discovery) มาประยุกต์ใช้ เพราะจากการทดสอบแล้วพบว่าโมเดลที่ได้ให้ผลการทดสอบที่น่าเชื่อถือมากที่สุด ในลำดับต่อไปนี้จะได้กล่าวถึงรายละเอียดต่างๆ ของการใช้เทคนิคการค้นหาความสัมพันธ์มาช่วยทำนายแนวโน้มเกรด อันได้แก่ การเตรียมข้อมูล การสร้างโมเดล การแปลความหมายข้อมูล และการนำความรู้ที่ได้มาใช้

4.1 การใช้เทคนิคการค้นหากฎความสัมพันธ์มาช่วยทำนายแนวโน้มเกรด

เทคนิคการสืบค้นกฎความสัมพันธ์ (Association rule discovery) เป็นการหาความสัมพันธ์ของข้อมูลที่มีอยู่ ในงานวิจัยนี้ได้ประยุกต์ใช้เทคนิคการสืบค้นกฎความสัมพันธ์กับข้อมูลผลการเรียนนิสิต โดยหาความสัมพันธ์ของผลการเรียนในแต่ละวิชาที่ส่งผลต่อกัน ซึ่งจะทำให้ได้ว่าวิชาใดบ้างที่มีผลต่อวิชาที่ต้องการจะทำนายเกรดล่วงหน้า

4.1.1 การเตรียมข้อมูล

งานวิจัยส่วนนี้ได้นำข้อมูลการลงทะเบียนเรียนของนิสิตคณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ตั้งแต่ปี 2535-2542 รวมทั้งสิ้น 476,085 แถว จากนิสิตกว่า 10,000 คน เช่นเดียวกับในส่วนที่ 3 มาใช้ โดยการเตรียมข้อมูลนั้นคล้ายกับการเตรียมข้อมูลในตารางที่ 4 แต่ในการเตรียมข้อมูลครั้งนี้จะคัดเลือกเฉพาะคอลัมน์ผลการเรียนของนิสิต โดยจะไม่นำคอลัมน์ประวัติส่วนตัวของนิสิตมาพิจารณา ทั้งนี้เพราะความสัมพันธ์ของการเรียนในแต่ละวิชานั้นส่งผลต่อผลการเรียนในอีกวิชาหนึ่งมากกว่าประวัติส่วนตัวของนิสิต นอกจากนี้ในการเตรียมข้อมูลครั้งนี้จะนำผลการเรียนในทุกวิชา ของทุกภาคการศึกษามาเป็นตัวพิจารณาด้วย เพราะเราต้องการหาความสัมพันธ์ของทุกวิชาในทุกปีการศึกษา รูปแบบตารางที่อยู่ในรูปแบบที่เหมาะสมกับเทคนิคการสืบค้นกฎความสัมพันธ์เป็นดังตารางที่ 7 และเนื่องมาจากในสาขาวิชาต่างๆ มีรายวิชาแตกต่างกันออกไป ดังนั้นจึงต้องแบ่งข้อมูลออกตามแต่ละสาขาวิชา แล้วจึงสร้างโมเดลการสืบค้นกฎความสัมพันธ์ สำหรับแต่ละสาขาวิชานั้นๆ

ตาราง 7 แสดงตัวอย่างข้อมูลสำหรับการใช้เทคนิคการค้นหากฎความสัมพันธ์ในการทำนายเกรด

Stu_code	Subject1	Subject2	…
37058063	204111Medium	403111Low	…
37058167	204111High	403111High	…
……	…..	…..	…

4.1.2 การสร้างโมเดล

จากตารางที่ 7 แสดงผลการเรียนในรายวิชาต่างๆ ของนิสิตแต่ละคน จากตารางที่ 7 ได้ตัวอย่างความสัมพันธ์หนึ่งเป็น

[204111Medium] + [403111Low] -> [417167Low]

จากความสัมพันธ์นี้ สามารถอธิบายได้ว่าเมื่อเรียนวิชา 204111 ได้เกรดอยู่ในช่วง C+, C และเรียนวิชา 403111 ได้เกรดอยู่ในช่วง D+, D, F แล้วจะได้ผลการเรียนวิชา 417167 อยู่ในช่วง D+, D, F ซึ่งจากความสัมพันธ์นี้เป็นเพียงตัวอย่างหนึ่งของนิสิตคนหนึ่งเท่านั้น แต่ต้องหาความสัมพันธ์ทั้งหมดของทุกรายวิชาและของนิสิตทุกคน ซึ่งจะได้ความสัมพันธ์ที่มีค่าสนับสนุน (support) และค่าความมั่นใจ (confidence) แตกต่างกันออกไป

ผลที่ได้จากการใช้เทคนิคการค้นหากฎความสัมพันธ์จะอยู่ในรูปความสัมพันธ์ของวิชาต่างๆ มากมาย ดังรูปที่ 3

Support(%	Confidence(%	Type	Lift	Rule Body		Rule Head
3.1769	84.6200	.	4.9...	[417167High]+[204111High]+[417168High]+[355113High]	=..	[403111High]
4.1155	77.0300	+	4.4...	[417167High]+[417168High]+[355113High]	=..	[403111High]
3.7545	74.2900	+	4.3...	[417167High]+[204111High]+[417168High]	=..	[403111High]
3.4657	73.8500	+	4.2...	[204111High]+[417168High]+[355113High]	=..	[403111High]
3.5379	70.0000	+	4.2...	[417167High]+[420112High]	=..	[417168High]
3.4657	67.6100	+	4.1...	[417167High]+[420111High]	=..	[417168High]
3.8989	69.2300	.	4.0...	[417167High]+[204111High]+[355112High]+[355113High]	=..	[403111High]
5.0542	67.9600	+	3.9...	[417167High]+[204111High]+[355113High]	=..	[403111High]
3.5379	66.2200	.	3.8...	[417167High]+[208111High]+[355113High]	=..	[403111High]
3.1769	62.8600	.	3.8...	[417167High]+[204111High]+[355113High]+[403111High]	=..	[417168High]
3.3213	65.7100	+	3.8...	[417167High]+[420112High]	=..	[403111High]
3.1769	64.7100	.	3.7...	[417168High]+[355112High]+[355113High]	=..	[403111High]
3.7545	59.7700	.	3.6...	[417167High]+[204111High]+[403111High]	=..	[417168High]
3.0325	62.6900	.	3.6...	[420111Medium]+[417167High]+[204111High]+[355113High]	=..	[403111High]
3.0325	62.6900	+	3.6...	[420112High]+[355113High]	=..	[403111High]
4.2599	62.1100	+	3.6...	[204111High]+[417168High]	=..	[403111High]
3.4657	96.0000	+	3.5...	[417168High]+[420111High]	=..	[417167High]
4.6209	60.9500	+	3.5...	[417168High]+[355113High]	=..	[403111High]
3.0325	60.0000	.	3.4...	[208111Medium]+[417167High]+[355112High]+[355113High]	=..	[403111High]
3.6823	60.0000	.	3.4...	[420111Medium]+[417167High]+[355112High]+[355113High]	=..	[403111High]
4.1877	59.7900	+	3.4...	[417167High]+[204111High]+[355112High]	=..	[403111High]
3.6101	56.8200	+	3.4...	[420111High]	=..	[417168High]
5.7040	58.5200	.	3.3...	[417167High]+[355112High]+[355113High]	=..	[403111High]

รูปที่ 3 แสดงตัวอย่างความสัมพันธ์ของผลการเรียนในรายวิชาต่างๆ

ความสัมพันธ์ที่ได้ในแต่ละสาขาวิชามีมากกว่า 200,000 กฎ ซึ่งต้องตัดความสัมพันธ์บางส่วนโดย

1) กำหนดค่าสนับสนุนต่ำสุด (minimum support) และค่าความมั่นใจต่ำสุด (minimum confidence) ไว้เพื่อเลือกเฉพาะความสัมพันธ์ที่มีจำนวนนิสิตมาก และน่าเชื่อถือ มานำเสนอให้กับนิสิต

2) กำจัดความสัมพันธ์ที่ลำดับของวิชาไม่เป็นไปตามข้อกำหนดหลักสูตร และคัดเลือกเฉพาะความสัมพันธ์ที่ทางด้านซ้ายมือของกฎเป็นวิชาที่นิสิตเคยเรียน และทราบผลการเรียนแล้ว

4.1.3 การแปลความหมายข้อมูลและการนำความรู้ที่ได้มาใช้

เมื่อนิสิตต้องการทำนายผลการเรียนในรายวิชาหนึ่ง โมเดลจะทำนายผลการเรียนให้กับนิสิตโดยพิจารณาวิชาที่นิสิตได้เคยลงทะเบียนเรียนที่ทราบผลการเรียนมาแล้วทุกวิชา กับกฎความสัมพันธ์ต่างๆ ที่หาออกมาได้ ซึ่งจะเห็นได้ว่าสามารถหากฎความสัมพันธ์ที่ตรงกับเงื่อนไขความต้องการ (คือกฎความสัมพันธ์ที่ด้านขวาของกฎเป็นวิชาที่เราต้องการทำนาย และด้านซ้ายของกฎเป็นวิชาและผลการเรียนในวิชาต่างๆ ของนิสิตที่เคยเรียนมา) ออกมาได้หลายกฎความสัมพันธ์

ตัวอย่างเช่น จากตารางที่ 7 ถ้าต้องการทำนายเกรดในวิชา 417168 ของนิสิตรหัส 37058063 เมื่อพิจารณากฎตามเงื่อนไขความต้องการข้างต้นแล้วได้ความสัมพันธ์ดังนี้

(1) [204111Medium]+[403111High]+[417167Medium] -> [417168Medium] confidence = 86.5

(2) [403111Low] + [417167Low] -> [417168Low] confidence = 80.3

(3) [403111Medium] + [204111Medium] -> [417168Medium] confidence = 84.2

จากตัวอย่างความสัมพันธ์ จะเห็นได้ว่า กฎที่ได้มาทั้งหมดนั้นสามารถทำนายผลการเรียนในวิชาที่ต้องการทำนาย (417168) ได้ออกมาหลายแบบด้วยกัน ในกรณีนี้ทำนายได้เป็น Medium และ Low ซึ่งควรนำเสนอผลการเรียนเพียงช่วงเดียวให้กับนิสิต ดังนั้นต้องกำหนดหลักเกณฑ์ในการตัดสินว่าจะเลือกความสัมพันธ์หรือนำเสนอผลการเรียนใดให้กับนิสิตจึงจะถูกต้องมากที่สุด การลำดับความสำคัญของเกณฑ์การเลือกความสัมพันธ์เป็นดังนี้ [1]

1) เลือกความสัมพันธ์ที่ทางด้านซ้ายมือของกฎมีวิชาและผลการเรียนตรงกับนิสิตคนนั้นมากที่สุดเป็นอันดับแรก
2) เลือกความสัมพันธ์ที่มีค่าความมั่นใจ (confidence) สูงสุดเมื่อเกณฑ์ในข้อ 1) เท่ากัน
3) เลือกความสัมพันธ์ที่มีค่าสนับสนุน (support) สูงสุดเมื่อเกณฑ์ในข้อ 2) เท่ากัน

จากตัวอย่าง เมื่อพิจารณาตามเกณฑ์ข้อ 1) เป็นอันดับแรก จะเห็นได้ว่าความสัมพันธ์ (1) มีวิชาและผลการเรียนที่ตรงตามเกณฑ์ที่ 1) 1 วิชาคือ [204111Medium] ส่วนความสัมพันธ์ (2) และ (3) มีวิชาและผลการเรียนที่ตรงตามเกณฑ์เท่ากับ 2 และ 1 ตามลำดับ ดังนั้น จากความสัมพันธ์นี้ จะสรุปได้ว่า ความสัมพันธ์ที่ (2) นั้นตรงตามหลักเกณฑ์มากที่สุด และทำนายได้ว่าเกรดในวิชา 417168 ของนิสิตคนนี้ในภาคการศึกษาต่อไปจะอยู่ในช่วง Low (D+, D, F) นอกจากนี้ ระบบได้นำเสนอเปอร์เซ็นต์ความเป็นไปได้ที่นิสิตคนนั้นจะได้เกรดอยู่ในช่วงที่เราทำนายโดยนำมาจากค่าความมั่นใจ (confidence) ให้กับนิสิตด้วย ทั้งนี้เพราะโมเดลทำนายโดยอ้างอิงมาจากข้อมูลเดิมของนิสิตที่เคยเรียนมาและได้ผลการเรียนเช่นเดียวกันกับนิสิตคนนั้น เช่นถ้ามีความสัมพันธ์ที่มีค่าความมั่นใจเท่ากับ 100 % นั้น มิได้หมายความว่านิสิตคนนั้นจะได้เกรดตามที่ทำนาย 100 % แต่หมายความว่า ข้อมูลของนิสิตทั้งหมดที่ได้นำมาสร้างโมเดลที่มีผลการ

เรียนในรูปแบบเดียวกันกับนิสิตคนนี้ได้ผลการเรียนในวิชาดังกล่าวเป็นแบบเดียวกัน 100 %

โมเดลที่ได้นี้เป็นเพียงสิ่งที่ช่วยแนะแนวทางให้กับนิสิต ในกรณีที่นิสิตสามารถเลือกวิชาเรียนได้ เช่น วิชาเลือก โมเดลจะเป็นสิ่งที่ช่วยนิสิตตัดสินใจในการเลือกลงทะเบียนเรียนในวิชาที่เหมาะสมกับความสามารถของตน ในกรณีที่เป็นวิชาบังคับและนิสิตต้องลงทะเบียนเรียน โมเดลนี้จะเป็นสิ่งที่ช่วยชี้แนะแนวทางในการปฏิบัติตนที่ดีให้กับนิสิตต่อวิชานั้นๆ ถ้าโมเดลทำนายออกมาว่าผลการเรียนจะออกมาไม่ดี (D+, D, F) นั้นหมายความว่า นิสิตควรจะใส่ใจ และตั้งใจในการเรียนวิชานี้เป็นพิเศษ ถ้าโมเดลทำนายผลการเรียนได้ว่าค่อนข้างดี (A, B+, B) ก็มิได้หมายความว่าจะให้นิสิตละเลย หรือไม่ต้องใส่ใจในวิชานี้ แต่สิ่งที่ได้จากโมเดลหมายความว่า นิสิตคนนั้นมีโอกาสที่จะเรียนในวิชานั้นได้ดี เนื่องมาจากผลการเรียนในบางวิชาที่มีผลต่อวิชานั้นของนิสิตคนนี้ค่อนข้างดี แสดงว่านิสิตมีพื้นฐานได้เปรียบกว่านิสิตคนอื่น ดังนั้นนิสิตควรที่จะตั้งใจเรียนเพื่อให้ได้ผลการเรียนดีตามที่โมเดลได้ทำนายไว้

ตัวอย่างการนำเสนอการทำนายแนวโน้มผลการเรียนให้กับนิสิตแสดงได้ดังรูปที่ 4

จากรูปที่ 4 แสดงแนวโน้มผลการเรียนในรายวิชาต่างๆ ที่นิสิตต้องการทราบ โดยแสดงชื่อรายวิชา ช่วงเกรดที่ทำนายได้ เปอร์เซ็นต์ความเป็นไปได้ในการทำนายในรายวิชานั้นๆ จำนวนนิสิตที่อ้างอิงโดยมีลักษณะการเรียนคล้ายคลึงกับนิสิตในรายวิชานั้นๆ และจำนวนนิสิตทั้งหมดในสาขาวิชานั้นที่นำมาสร้างโมเดล

ผลการทำนายเกรดรายวิชาต่าง ๆ จากจำนวนนิสิต 437 คน

รหัสวิชา	วิชา	เกรดที่ทำนาย	% ความเป็นไปได้	จำนวนนิสิต
417168	คณิตศาสตร์วิศวกรรม II	C+ หรือ C	68.7000	78
417267	คณิตศาสตร์วิศวกรรม III	A หรือ B+ หรือ B	95.5600	39
417268	คณิตศาสตร์วิศวกรรม IV	C+ หรือ C	77.7800	13
208111	การเขียนแบบวิศวกรรม	A หรือ B+ หรือ B	60.8700	69
208221	กลศาสตร์วิศวกรรม I	A หรือ B+ หรือ B	88.8900	39
208222	กลศาสตร์วิศวกรรม II	A หรือ B+ หรือ B	71.8800	21
208281	การฝึกงานโรงงาน	A หรือ B+ หรือ B	100.0000	148
204331	การโปรแกรมระ บบ	A หรือ B+ หรือ B	90.9100	48
204341	วิศวกรรมซอฟต์แวร์	A หรือ B+ หรือ B	82.3500	13
204351	ฐานข้อมูลและการสืบค้นสารสนเทศ	A หรือ B+ หรือ B	95.3800	61
204371	เทคนิคการแปลงและการวิเคราะห์สัญญาณ	A หรือ B+ หรือ B	87.5000	26
204497	สัมมนา	A หรือ B+ หรือ B	96.0000	43
204499	โครงงานวิศวกรรมคอมพิวเตอร์	A หรือ B+ หรือ B	94.4400	13
205211	การวิเคราะห์วงจรไฟฟ้า I	D+ หรือ D หรือ F	78.0000	34
205213	ปฏิบัติการวงจรไฟฟ้า	A หรือ B+ หรือ B	100.0000	13
205251	วงจรและระ บบอิเล็กทรอนิกส์ I	D+ หรือ D หรือ F	82.3500	13

รูปที่ 4 แสดงตัวอย่างการนำเสนอการทำนายแนวโน้มผลการเรียน

สาเหตุที่นำเสนอเกรดให้นิสิตเป็นช่วงนั้น เนื่องมาจาก วัตถุประสงค์ของงานวิจัยนี้ที่ต้องการชี้แนะแนวทางในการปฏิบัติตนในการเรียนในวิชานั้นๆ อย่างเหมาะสม ซึ่งการที่นิสิตทราบว่า ผลการเรียนมีแนวโน้มจะออกมาดี ปานกลาง หรือไม่ดีนั้น เป็นการนำเสนอที่มีผลต่อดีต่อการปฏิบัติตนของนิสิต โดยถ้านำเสนอเกรดให้กับนิสิตแบบเจาะจงว่านิสิตจะได้ A นั้นอาจส่งผลเสียทำให้นิสิตเกิดความลำพองใจ ไม่ตั้งใจเรียน ทำให้ผลการเรียนตกต่ำกว่าที่ควรจะเป็น หรือได้ทำนายว่านิสิตจะได้ F อาจเป็นการบั่นทอนกำลังใจของนิสิตลงไปมาก ทำให้นิสิตเกิดความท้อแท้ในการเรียนรายวิชานั้นๆ ได้

บทสรุป

การพัฒนาคุณภาพการศึกษาสามารถทำได้ในหลายรูปแบบ ไม่ใช่ว่าทำเพียงแบบใดแบบหนึ่งแล้วจะสำเร็จได้ แต่ต้องอาศัยวิธีการต่างๆ มาประกอบกันเพื่อให้ได้ผลที่ดีที่สุด อีกทั้งต้องคำนึงถึงปัจจัยต่างๆ หลายประการ ที่จะส่งผลกระทบต่อกัน ซึ่งในบางครั้งปัจจัยเหล่านี้มีมากเกินกว่าที่จะพิจารณาได้ด้วยตาเปล่า

ในบทความนี้ได้นำเสนอการนำเทคนิคดาต้าไมน์นิงมาประยุกต์ใช้เพื่อช่วยพิจารณาหารูปแบบความสัมพันธ์ของข้อมูลในมุมมองต่างๆ เพื่อที่จะได้นำสิ่งที่เป็นประโยชน์ที่ได้จากดาต้าไมน์นิงไปเป็นส่วนหนึ่งในการพัฒนาคุณภาพการศึกษาต่อไป ซึ่งผลลัพธ์ที่ได้จากงานวิจัยนี้ค่อนข้างเป็นที่น่าพอใจ โดยมีเปอร์เซ็นต์ความถูกต้องค่อนข้างสูง แต่มีปัญหาบางประการ ได้แก่ จำนวนข้อมูลในบางสาขาวิชามีปริมาณค่อนข้างน้อยทำให้โมเดลที่ได้ไม่แม่นยำเท่าที่ควร หากต้องการกำจัดความผิดพลาดที่เกิดจากปริมาณข้อมูลน้อยเกินไปจำเป็นต้องใช้ข้อมูลอย่างน้อยพันคนในแต่ละสาขาวิชาที่ต้องการทำนาย, วิธีที่นำเสนอไปนั้นอาจยังไม่ใช่วิธีที่ดีที่สุดในการวิเคราะห์แนวโน้ม, หน่วยงานการศึกษาที่ต้องการนำโมเดลนี้ไปใช้งาน จะต้องเตรียมข้อมูลให้ตรงตามโครงสร้างที่โมเดลนี้ได้ออกแบบไว้ แต่สามารถปรับปรุงโมเดลให้เหมาะสมกับโครงสร้างข้อมูลในแต่ละหน่วยงานได้

เอกสารอ้างอิง

[1] B.Liu and W.Hsu and Y.Ma. “Integrating Classification and Association Rule Mining” , .In Proceeding of International Conference KDD’98, 1998.

[2] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. 1996. “Advances in knowledge discovery and data mining”, AAAI/MIT Press.

[3] Han, J., Kamber, M. 2000., “Data Mining Concepts and Techniques”, Morgan Kaufmann.

[4] J.Gehtke, R.Ramakrishnan, and V.Ganti. Rainforest., "A framework for fast decision tree construction of large datasets", In Proceeding of International Conference Very Large Database, p.416-427, 1998.

[5] Kitsana Waiyamai and Lotfi Lotfi Lakhal. 2000., "Knowledge Discovery from Very Large Databases Using Frequent Concept Lattices", p.437-445, 2000.

[6] Kitsana Waiyamai, Chidchanok Songsiri and Thanawin Rakthanmanon., "A Data Mining based Approach for Improving Quality of Engineering Graduates". In Proceeding of 4th International Conference UNESCO International center for Engineering Education (UICEE ' 2001), p.84-88, 2001.

[7] R.Agrawal, H. Mannila, R. Srikant and A. I. Verkamo., "Fast Discovery of Association Rules". In U.M. Fayyad, G.Piatetsky-Shapiro, P.Smyth, and R.Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, AASAi/MIT Press, p. 307-328, 1996.

[8] กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ และ ธนาวินท์ รักธรรมานนท์, มารู้จักกับ Data Mining, "ไมโครคอมพิวเตอร์" (MICRO COMPUTER), Vol. 18, No. 187, p. 179-181, December 2000.

กฤษณะ ไวยมัย จบปริญญาตรีและปริญญาโททางวิทยาศาสตร์คอมพิวเตอร์ จาก University of Picardie ประเทศฝรั่งเศส จบปริญญาเอกทางด้านวิทยาการคอมพิวเตอร์จาก University of Clermont ประเทศฝรั่งเศส งานวิจัยหลักประกอบด้วยสามส่วนสำคัญ ส่วนแรกคือ ระบบสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Databases: Data Mining) ส่วนที่สองคือ ระบบสารสนเทศ (Information System) และ ส่วนที่สามคือ ระบบฐานข้อมูล (Database Management System) ปัจจุบัน ดร.กฤษณะ ไวยมัย ดำรงตำแหน่งอาจารย์ประจำภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

ชิดชนก ส่งศิริ จบปริญญาตรีทางวิศวกรรมคอมพิวเตอร์จากมหาวิทยาลัยเกษตรศาสตร์ งานวิจัยหลักประกอบด้วยการวิจัยและการพัฒนาระบบสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ (knowledge Discovery from very large Databases: Data Mining) , ระบบคลังข้อมูล (Data Warehousing) และระบบจัดการฐานข้อมูล (Database management) ปัจจุบันศึกษาปริญญาโทวิศวกรรมคอมพิวเตอร์จากมหาวิทยาลัยเกษตรศาสตร์

ธนาวินท์ รักธรรมานนท์ จบปริญญาตรีทางวิศวกรรมคอมพิวเตอร์จากมหาวิทยาลัยเกษตรศาสตร์ งานวิจัยหลักประกอบด้วยการวิจัยและการพัฒนาระบบสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ (knowledge Discovery from very large Databases: Data Mining) , ระบบคลังข้อมูล (Data Warehousing) และระบบจัดการฐานข้อมูล (Database management) ปัจจุบันศึกษาปริญญาโทวิศวกรรมคอมพิวเตอร์จากมหาวิทยาลัยเกษตรศาสตร์

การใช้เทคนิค Association Rule Discovery เพื่อการจัดสรรกฎหมายในการพิจารณาคดีความ

กฤษณะ ไวยมัย และ ธีระวัฒน์ พงษ์ศิริปรีดา
อาจารย์ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์
และนิสิตปริญญาโทวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

ABSTRACT - Data Classification and Association Rule Discovery are two important data mining techniques. In this paper, we apply the two techniques to find appropriate laws and articles for lawsuits. We propose a data classification method using a classifier generated from association rule discovery technique. We utilize the classifier to help selecting laws and articles to be tried for a lawsuit. Our experiment shows that the classifier generated from a proposed method yields better accuracy than one generated from a general data classification method. In addition, our method increases efficiency in multi-group and multi-level classification.

KEY WORDS – data mining, knowledge discovery from large database, lawsuit

บทคัดย่อ – เทคนิค Data classification และเทคนิค Association rule discovery เป็นเทคนิคที่สำคัญของดาต้าไมน์นิง ในงานวิจัยฉบับนี้ เราได้นำเทคนิคทั้งสองมาประยุกต์ใช้ในการจัดสรรกฎหมายที่เหมาะสมกับคดีความ โดยนำเทคนิค Data classification มาสร้างตัวจำแนกข้อมูลจากกฎเกณฑ์ที่ได้จากเทคนิค Association rule discovery อีกทีหนึ่ง ซึ่งตัวจำแนกข้อมูลที่ได้นี้จะสามารถนำไปใช้ทำนายคดีความแต่ละคดีว่าควรใช้กฎหมายฉบับใดในการพิจารณา ผลการวิจัยที่ได้ แสดงให้เห็นว่าการสร้างตัวจำแนกตามวิธีที่เสนอนี้ได้ประสิทธิภาพดีกว่าการสร้างตัวจำแนกตามเทคนิค Data classification แบบปกติ นอกจากนี้ วิธีดังกล่าวยังช่วยแก้ปัญหาต่างๆ ที่เกิดขึ้นจากการใช้เทคนิค Data classification โดยทั่วไป

คำสำคัญ – ดาต้าไมน์นิง, การสืบค้นความรู้จากฐานข้อมูลขนาดใหญ่, การจัดสรรกฎหมาย

1. บทนำ

ในการพิจารณา หรือตัดสินคดีความใดๆ นักกฎหมายจำเป็นต้องอ้างอิงตัวบทกฎหมายเพื่อประกอบการพิจารณากฎหมายไทยที่ใช้อยู่ในปัจจุบันสามารถจำแนกออกได้เป็น พระราชบัญญัติ และพระราชกำหนด [1] ซึ่งมีจำนวนรวมกันประมาณ 260 ฉบับ ในแต่ละฉบับจะประกอบด้วยมาตราย่อยต่างๆ มากน้อยแตกต่างกันไป บางฉบับอาจมีจำนวนมาตราสูงถึงหลายพันมาตรา จำนวนกฎหมายที่มีปริมาณมากเช่นนี้ ทำให้ยากต่อการที่จะศึกษาได้อย่างครบถ้วน เมื่อนักกฎหมายจำเป็นต้องพิจารณาคดีความใดๆ จึงเกิดปัญหาในการเลือกกฎหมายฉบับที่เหมาะสม มาใช้ประกอบการพิจารณาคดีความ ในปัจจุบันนักกฎหมายโดยทั่วไปมักจะใช้ทักษะความชำนาญในสาขาวิชาชีพที่มีอยู่ ประกอบกับการสืบค้นคดีความเก่าๆ ขึ้นมาพิจารณา เพื่อดูว่าคดีที่เคยตัดสินไปแล้วนั้น อ้างถึงกฎหมายฉบับใดบ้าง เพื่อใช้เป็นแนวทางในการพิจารณาคดีที่กำลังสนใจอยู่ต่อไป

ที่ผ่านมาได้มีการใช้การค้นหาคำ หรือวลีแบบ Full-Text-Search [2] เข้ามาช่วย โดยการค้นหาคำหรือวลีจากคดีความเก่าๆ ที่เก็บไว้ เพื่อนำมาพิจารณาว่าสอดคล้องกับคดีปัจจุบันที่กำลังพิจารณาอยู่หรือไม่ ซึ่งผลลัพธ์ที่ได้ยังไม่เป็นที่น่าพอใจ เนื่องจากผู้ใช้ไม่สามารถตัดสินใจได้ว่าควรเลือกใช้ข้อความส่วนใดบ้างในการค้นหา เพราะถ้าระบุคำค้นมากอาจทำให้ค้นหาคดีความเก่าไม่พบ หรือถ้าระบุคำค้นน้อยเกินไป ก็จะทำให้ได้ผลลัพธ์มากเกินจนไม่มีความหมาย นอกจากนี้ผู้ใช้ยังต้องเสียเวลาในการอ่านคดีความเก่าๆที่ค้นมาได้ เพื่อพิจารณาว่าใจความของคดีเก่านั้น มีความหมายตรงตามที่ต้องการหรือไม่

งานวิจัยฉบับนี้ได้มุ่งเน้นที่จะศึกษาและคิดค้นเพื่อสร้างระบบที่สามารถทำนายได้ว่าแต่ละคดีมีความน่าจะเป็นที่จะใช้กฎหมายฉบับไหน และมาตราใดประกอบการพิจารณา เพื่อเป็นแนวทางในการศึกษา หรือทำงานกับคดีความนั้นๆ ต่อไป ระบบนี้นอกจากจะสามารถช่วยนักกฎหมายในการทำงานให้สะดวกรวดเร็วขึ้นแล้ว ยังสามารถเอื้อประโยชน์อย่างมากต่อประชาชนทั่วไป ที่ไม่มีความรู้ ความชำนาญทาง

ด้านกฎหมาย ในการที่จะหากฎหมายที่เหมาะสมมาศึกษา เพื่อใช้ประกอบคดีความที่กำลังสนใจอยู่ โดยในการสร้างระบบดังกล่าว เราเลือกใช้เทคนิคการสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจจากฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Database: KDD) หรือที่เรียกกันว่า Data Mining [3,4,5,6,7]

Data Classification [8,9] และ Association Rule Discovery [10,11,12] เป็นเทคนิคที่สำคัญของ KDD เทคนิค Data Classification เป็นเทคนิคที่ใช้หากฎเกณฑ์ความสัมพันธ์ของข้อมูล เพื่อนำมาสร้างตัวจำแนกประเภทของข้อมูล ส่วนเทคนิค Association Rule Discovery เป็นเทคนิคที่ใช้หากฎเกณฑ์ความสัมพันธ์ทั้งหมดของข้อมูล ที่มีคุณสมบัติสอดคล้องตามค่า Minimum Support และ Minimum Confidence ข้อแตกต่างที่ชัดเจนของเทคนิคทั้งสอง คือ Data Classification เป็นเทคนิคที่กำหนดผลลัพธ์แน่นอน ซึ่งก็คือจำนวนประเภทของข้อมูลที่ต้องการจำแนก ส่วน Association Rule Discovery เป็นเทคนิคที่ไม่มีการกำหนดขอบเขตของผลลัพธ์ล่วงหน้า

ในงานวิจัย *Using Data Mining technique to select the laws and articles for lawsuits* [13] ได้มีการนำเทคนิค Data Classification มาใช้สร้างระบบจัดสรรกฎหมายที่เหมาะสมให้กับคดีความแบบอัตโนมัติ ผลการทดลองที่ได้ยังไม่เป็นที่น่าพอใจ ซึ่งสามารถสรุปได้ 3 ประเด็นคือ

ประเด็นที่ 1 การตัดคำในขั้นตอนการเตรียมข้อมูลยังขาดประสิทธิภาพ ทำให้ผลการทำนายของระบบมีความถูกต้องต่ำ

ประเด็นที่ 2 ระบบไม่สามารถทำนายกฎหมายให้กับคดีความได้ครั้งละมากกว่าหนึ่งกฎหมาย

ประเด็นที่ 3 ระบบไม่สามารถทำนายกฎหมายแบบมีลำดับชั้นในเวลาเดียวกันได้

จากงานวิจัย *Integrating Classification and Association Rule Mining* [14] ได้แสดงให้เห็นว่าเทคนิค Association Rule Discovery สามารถนำมาประยุกต์ร่วมกับเทคนิค Data Classification เพื่อใช้ในการสร้างตัวจำแนกข้อมูล โดยการใช้เทคนิค Association Rule Discovery หากฎเกณฑ์ความสัมพันธ์ของข้อมูลที่มีต่อประเภทของข้อมูลนั้นๆ กฎเกณฑ์ที่ได้จากขั้นตอนนี้เรียกว่า Class Association Rules (CARs) ซึ่งจะถูกนำไปใช้ในการสร้างตัวจำแนกข้อมูลต่อไป

เราได้ประยุกต์เทคนิคดังกล่าว มาใช้ในการสร้างตัวทำนายกฎหมายของระบบ ซึ่งผลการทดลองที่ได้ แสดงให้เห็นว่าการสร้างตัวทำนายด้วยวิธีดังกล่าว ให้เปอร์เซ็นต์ความถูกต้องสูงกว่าการสร้างตัวทำนายด้วยวิธี Data Classification แบบทั่วไป นอกจากนี้ วิธีดังกล่าวยังช่วยแก้ปัญหาทั้ง 3 ประการของระบบเดิมดังที่ได้กล่าวมาแล้ว

ในงานวิจัยฉบับนี้ เราได้อธิบายถึงเทคนิค และรายละเอียดการทดลองซึ่งแบ่งออกเป็น 3 ขั้นตอน คือ

1. ขั้นตอนการเตรียมข้อมูล
2. ขั้นตอนการหา Class Association Rules
3. ขั้นตอนการสร้างตัวจำแนกข้อมูลจาก Class Association Rules

ในตอนท้าย เราได้สรุปผลการทดลองโดยการแสดงกราฟเปรียบเทียบประสิทธิภาพของระบบใหม่ กับระบบเก่า พร้อมทั้งได้เสนอแนะแนวทางในการพัฒนาระบบต่อไป

2. ขั้นตอนการเตรียมข้อมูล (Pre-processing)

2.1 ข้อมูลคดีความ

ข้อมูลที่ใช้ทดสอบระบบ คือข้อมูลคดีความของศาลฎีกา ซึ่งเป็นศาลสูงสุด คดีความที่เข้าสู่ขั้นตอนการพิจารณาของศาลฎีกาเหล่านี้ เป็นคดีความที่ผ่านขั้นตอนการพิจารณาของศาลชั้นต้น และศาลอุทธรณ์มาแล้ว ซึ่งประกอบด้วยคดีอาญา และคดีแพ่ง ข้อมูลคดีความของศาลฎีกาที่นำมาทดสอบนี้ ประกอบด้วยรายละเอียดของคดีความ ได้แก่ ปีที่มีการวินิจฉัย ชื่อโจทย์ และจำเลย ชื่อกฎหมายที่ใช้ประกอบการพิจารณาตัดสินคดี และตัวเนื้อคดีความ ในตัวคดีความจะเป็นส่วนการบรรยายรูปคดีที่ผ่านขั้นตอนการพิจารณาจากศาลชั้นต้น และศาลอุทธรณ์มาแล้ว ข้อมูลทั้งหมดสามารถแบ่งได้เป็น 30,000 คดี ตั้งแต่ปี พ.ศ. 2489 ถึงปี พ.ศ. 2541 ซึ่งโดยเฉลี่ยแล้ว แต่ละคดีความจะใช้กฎหมาย 1-2 ฉบับ ฉบับละ 3-4 มาตราในการพิจารณาจากข้อมูลทั้งหมด เราสามารถแบ่งกลุ่มกฎหมายที่ใช้พิจารณาคดีความออกได้เป็น 20 กลุ่ม และกลุ่มมาตราได้ 2200 กลุ่ม

2.2 การตัดคำ

คดีความแต่ละคดีจะถูกตัดคำด้วยพจนานุกรมภาษาไทย เพื่อแบ่งคดีความออกเป็นคำสั้นๆ สำหรับใช้ประมวลผลในขั้นตอนต่อไป พจนานุกรมภาษาไทยที่ใช้ในระบบนี้ เราสร้างขึ้นมาจากการหาวลีจากคดีความทั้งหมด โดยการใช้เทคนิค Suffix array [15]

เทคนิค Suffix array เป็นเทคนิคที่ใช้กันอย่างแพร่หลายเพื่อพิจารณาความสัมพันธ์ของชุดข้อมูล เราใช้เทคนิคนี้ในการหา กลุ่มของตัวอักษรที่ยาวที่สุดที่ปรากฏซ้ำๆอยู่ในชุดข้อมูล ซึ่งก็คือวลีที่เราต้องการนั้นเอง

ให้ A แทน String ใดๆ (ซึ่งถ้ามองอีกรูปหนึ่ง A ก็คือ array of characters นั้นเอง) เราจะได้ S แทน Suffix array ของ A โดยที่ S คือ Array of pointer ที่ชี้ไปที่ตำแหน่งในหน่วยความจำของสมาชิกใน A แต่ละตัว

ฉะนั้นสมาชิกของ S แต่ละตัวก็คือ sub string ที่เริ่มต้นด้วยตัวอักษรตำแหน่งต่างๆ ของ A

ยกตัวอย่างเช่น A = banana
เราจะได้ S[0] = banana
S[1] = anana
S[2] = nana
S[3] = ana
S[4] = na
S[5] = a

จาก S ที่ได้ เรานำสมาชิกทั้งหมดของ S มาจัดเรียงลำดับจากน้อยไปมาก ซึ่งจะได้ผลดังต่อไปนี้

S[5] = a
S[3] = ana
S[1] = anana
S[0] = banana
S[4] = na
S[2] = nana

จาก S ที่ได้ผ่านการเรียงลำดับแล้ว จะทำให้เราสามารถหากลุ่มของตัวอักษรที่ยาวที่สุดที่ซ้ำกันในชุดของข้อมูลที่ติดกัน ซึ่งก็คือ ana ที่ปรากฏอยู่ที่ S[3] และ S[1] นั่นเอง

รูปที่ 1 แสดงอัลกอริทึมของโปรแกรมการหาวลีจากชุด string ที่ผ่านการเรียงข้อมูลแล้ว

```
1  FOR  m = 1  TO  Count(S) - FNUM  DO
2      sl = 999;
3      FOR  i = m  TO  m + FNUM  DO
4          FOR  c = 1  TO  Length(Si)  DO
5              IF  Si [c] < > Si+1 [c]  THEN
6                  IF  c < sl  THEN  sl = c
7                  break;
8              END
9          END
10     END
11     IF  sl > LNUM  THEN  ReturnPhase ( S 1..sl )
12 END
```

รูปที่ 1. แสดงอัลกอริทึมของโปรแกรมหาวลี

S คือ จำนวนชุดตัวอักษรย่อยที่ผ่านการจัดเรียงลำดับแล้ว

LNUM คือ จำนวนตัวอักษรน้อยที่สุดของชุดตัวอักษร ที่จะถูกเลือกเป็นวลี

FNUM คือ จำนวนความถี่น้อยที่สุดในการเกิดชุดตัวอักษรซ้ำๆกัน ที่จะถูกเลือกเป็นวลี

sl คือ ตัวแปรสำหรับนับจำนวนตัวอักษรที่เกิดซ้ำกันในชุดข้อมูลย่อยทั้ง 2 ชุด

โปรแกรมจะวนรอบเป็นจำนวนครั้ง เท่ากับ จำนวนชุดตัวอักษรย่อย Count (S) ลบด้วยค่า FNUM ในแต่ละรอบ โปรแกรมจะเริ่มด้วยการให้ค่าเริ่มต้นตัวแปร sl มีค่าสูงสุดซึ่งคือ 999 (บรรทัดที่ 2) ถัดมาจะวนรอบเพื่อเปรียบเทียบชุดตัวอักษรเป็นจำนวนครั้ง เท่ากับค่า FNUM (บรรทัดที่ 3) ในรอบเปรียบเทียบชุดตัวอักษรนี้ โปรแกรมจะเปรียบเทียบตัวอักษรทีละตัวเป็นจำนวนครั้งเท่ากับจำนวนตัวอักษรในชุดตัวอักษรนั้น (บรรทัดที่ 4) หรือเมื่อพบตัวอักษรที่ต่างกันของ 2 ชุดตัวอักษร (บรรทัดที่ 5 ถึง 8) ทุกครั้งที่พบตัวอักษรที่ต่างกัน โปรแกรมจะตรวจสอบค่า sl กับ ตำแหน่งตัวอักษรที่เริ่มต่างกันนี้ (บรรทัดที่ 6) ถ้าหากว่าตำแหน่งตัวอักษรนี้น้อยกว่าค่า sl ก็จะ กำหนดค่า sl ให้เท่ากับตำแหน่งตัวอักษรนั้น ด้วยวิธีนี้เมื่อโปรแกรมทำงานผ่านไปครบจำนวน FNUM รอบ จะได้ค่า sl เป็นค่าของจำนวนตัวอักษรของชุดตัวอักษรที่เกิดซ้ำกันในชุดตัวอักษรทั้ง FNUM ชุด ซึ่งถ้าหากว่าค่า sl นี้ไม่น้อยกว่าค่า LNUM ที่กำหนดไว้ โปรแกรมก็จะเลือกให้ตัวอักษรตัวที่ 1 ถึง sl เป็นวลี (บรรทัดที่ 11)

เรานำเทคนิคดังกล่าว มาใช้กับคดีความที่มีอยู่ โดยแบ่งกลุ่มคดีความเป็นกลุ่มๆ 70 กลุ่ม ตามกฎหมายที่ใช้พิจารณาระดับมาตรา แต่ละกลุ่มมี 100 คดีความ หลังจากนั้น เราจะโหลดคดีความครั้งละกลุ่มเข้าสู่หน่วยความจำเพื่อสร้างเป็น Array of characters (A) ในขั้นตอนนี้ เราจะได้ A คือ ชุดของตัวอักษรในคดีความเรียงต่อเนื่องกัน ซึ่งมีขนาดประมาณ 200 Kbytes

เราให้ S เป็น Array of pointer ที่ชี้ไปที่ตำแหน่งสมาชิกแต่ละตัวของ A เมื่อเราจัดเรียง S ด้วยเทคนิค quick sort ก็จะได้ผลลัพธ์คือ sub string ย่อยๆ จำนวน 200 K strings ที่เรียงลำดับจากน้อยไปมาก หลังจากนั้นเราสามารถหากลุ่มของตัวอักษรยาวที่สุดที่ซ้ำๆกัน ซึ่งก็คือวลีต่างๆ ที่เราต้องการนั้นเอง

จากวิธีการดังกล่าว ขั้นตอนที่ใช้เวลานานที่สุดคือขั้นตอนการเปรียบเทียบชุดตัวอักษรเพื่อหาวลี ซึ่งเราได้ปรับปรุงวิธีการให้เร็วขึ้น โดยการตัดคำชุดของตัวอักษร A ด้วยพจนานุกรมคำเดี่ยวเสียก่อน หลังจากนั้นให้ S แต่ละตัว ชี้ไปที่ตำแหน่งหน่วยความจำของตัวอักษรใน A ที่เป็นตัว

เริ่มต้นคำใหม่เท่านั้น ด้วยวิธีการนี้จะทำให้ S มีขนาดลดลง 70% ซึ่งทำให้ขั้นตอนการหาวลีจากชุดตัวอักษรมีความเร็วขึ้นมาก

เมื่อเราทำขั้นตอนดังกล่าวกับข้อมูลครบทั้ง 70 ชุด เราจะได้ผลลัพธ์คือ วลี จำนวน 23,722 วลี ซึ่งจะถูกนำไปสร้างเป็นพจนานุกรมเพื่อใช้ตัดคำต่อไป

2.3 การแปลงข้อมูล

พจนานุกรมที่ได้จากขั้นตอน 2.2 จะถูกจัดเก็บลงฐานข้อมูล โดยแต่ละวลีจะมีค่าตัวเลขกำกับอยู่เพื่อใช้ในการแปลงวลีให้อยู่ในรูปตัวเลข เพื่อความสะดวกในการทำงานขั้นตอนต่อไป สำหรับข้อมูลคดีความที่มีอยู่เราได้แปลงเข้าสู่ฐานข้อมูลเช่นกัน โดยเนื้อความของ แต่ละคดี จะถูกตัดคำด้วยพจนานุกรมดังกล่าว ซึ่งจะได้ผลลัพธ์เป็นตัวเลขแทนวลีที่ปรากฏอยู่ในคดีความนั้นๆ ในแต่ละคดีความมีรหัสกฎหมายที่ใช้ในการพิจารณา ระบุอยู่ด้วย ซึ่งแบ่งเป็น 2 ระดับคือระดับกฎหมาย และระดับมาตรา รูปแบบของฐานข้อมูลแสดงได้ดังรูปที่ 2 ซึ่งประกอบไปด้วย รหัสกฎหมาย และ มาตรา (LAW_CHAPTER) รหัสคดีความ (CASE) และวลีที่ปรากฏอยู่ในแต่ละคดีความ (WORDID)

LAW_CHAPTER	CASE	WORDID
01_249	2489_0002	17566
01_249	2489_0002	14690
01_249	2489_0002	14690
01_249	2489_0002	9830
01_249	2489_0002	20880
01_249	2489_0002	5189
01_249	2489_0002	9231
01_249	2489_0002	21113
01_249	2489_0002	682
01_249	2489_0002	5138
01_249	2489_0002	11008
01_249	2489_0002	5189
01_249	2489_0009	15651
01_249	2489_0009	21122
01_249	2489_0009	5189
01_249	2489_0009	1532
01_249	2489_0009	21549
01_249	2489_0009	1913
01_249	2489_0009	284
01_249	2489_0009	23144
01_249	2489_0009	7889
01_249	2489_0009	22295
01_249	2489_0009	2846

รูปที่ 2. แสดงโครงสร้างข้อมูลที่ผ่านการเตรียม

3. การหา Class Association Rules (CARs)

ด้วยเทคนิคของ Association Rule Discovery เราสามารถหากฎเกณฑ์ ที่บอกความสัมพันธ์ของข้อมูล โดยกฎเกณฑ์ที่ได้จะต้องมีคุณสมบัติสอดคล้องตาม ค่า MinSupt และ MinConfd ที่กำหนดโดยผู้ใช้ กฎเกณฑ์ที่ได้นี้เรียกว่า Association Rules ซึ่งจะมีรูปแบบดังต่อไปนี้

{item1, item2} → item3 Supt% Confd%

- *Supt* คือ เปอร์เซ็นต์ของจำนวนข้อมูลที่มีสมาชิกสอดคล้องตามกฎ ต่อจำนวนข้อมูลทั้งหมด
- *Confd* คือ เปอร์เซ็นต์ของจำนวนข้อมูลที่สอดคล้องตามกฎ ต่อจำนวนข้อมูลทั้งหมดที่มีสมาชิกตามกฎฝั่งซ้ายมือ
- *MinSupt* คือ ค่า Supt ต่ำสุดที่ทำให้เกิดกฎเกณฑ์
- *MinConfd* คือ ค่า Confd ต่ำสุดที่ทำให้เกิดกฎเกณฑ์

จากตัวอย่างกฎข้างบน มีความหมายว่า เมื่อมี item1 และ item2 ปรากฏอยู่ในข้อมูล จะมี item3 ปรากฏอยู่ในข้อมูลชุดนั้นด้วยเสมอ ตามเปอร์เซ็นต์ความเชื่อมั่น (Confd) และตามเปอร์เซ็นต์ปริมาณข้อมูลที่สอดคล้องตามกฎ (Supt)

Class Association Rules (CARs) คือกฎเกณฑ์ที่ได้จากกระบวนการ Association Rule Discovery เช่นกัน แต่เป็นกฎเกณฑ์ที่บอกความสัมพันธ์ของข้อมูลที่มีต่อประเภทข้อมูลนั้นๆ (Class) กฎเกณฑ์ที่ได้นี้จะมีลักษณะดังต่อไปนี้

{item1, item2} → Class Supt% Confd%

เราได้นำเทคนิคดังกล่าวมาใช้กับงานกฎหมาย โดยอาศัยข้อมูลที่ได้ผ่านขั้นตอนการเตรียมข้อมูลมาแล้วส่วนหนึ่งคือ 70% เพื่อสอนให้ระบบเรียนรู้การจำแนกประเภทข้อมูล เป็นการเรียนรู้เพื่อหากฎเกณฑ์ความสัมพันธ์ของวลีในคดีความ ที่มีต่อกฎหมายที่ใช้พิจารณาคดีความนั้นๆ ในกรณีนี้ประเภทของข้อมูลก็คือกฎหมายฉบับที่ใช้พิจารณาคดีความนั้นเอง กฎเกณฑ์ที่ได้นี้จะอยู่ในรูปต่อไปนี้

วลี → กฎหมาย Supt% Confd%

จากรูปแบบ CARs ที่แสดงข้างบน มีความหมายว่า ถ้าปรากฏวลีหนึ่งในคดีความ มีความน่าจะเป็นที่คดีความนั้นจะใช้กฎหมาย ที่ระบุอยู่ทางฝั่งขวามือของกฎเป็นตัวพิจารณาคดี ตามค่าความเชื่อมั่น Confd% และมีความถี่ของการเกิดเหตุการณ์เช่นนี้ Supt%

ซ้ายมือของกฎก็คือวลีต่างๆ ที่หาได้จากคดีความ เรากำหนดให้ฝั่งซ้ายมือของกฎเกณฑ์มีค่าเป็นวลี 1 วลี ไม่ใช่เซตของวลี เนื่องจากว่าจำนวน

วลีในระบบมีมากถึง 23,722 วลี ทำให้ไม่สามารถหากฎเกณฑ์ของเซตวลีได้ และเนื่องจากว่าวลีที่ใช้ในระบบนี้คือวลีที่มีความยาวพอสมควร (ผ่านการตัดคำแบบวลีในขั้นตอนการเตรียมข้อมูล) ซึ่งในตัววลีแต่ละตัวจะประกอบไปด้วยคำสั้นๆ รวมกันอยู่แล้ว ทำให้ไม่จำเป็นต้องหากฎเกณฑ์ของเซตวลีอีกต่อไป

ฝั่งขวามือของกฎคือ กฎหมาย ซึ่งอาจหมายถึงกฎหมายฉบับที่ใช้ หรือมาตราที่ใช้ประกอบการพิจารณาตัดสินคดีความนั้นๆ

สำหรับค่า MinSupt และค่า MinConfd เรากำหนดให้มีค่า 0 เปอร์เซ็นต์ในระหว่างทำการทดลอง เนื่องจากว่าข้อมูลในระบบนี้ที่ทำการทดลองมีการกระจายตัวสูง ทำให้ค่า Supt และค่า Confd ของ CARs ที่ได้มีค่าต่ำมาก อย่างไรก็ตาม CARs ต่างๆที่หาได้จากขั้นตอนนี้ จะถูกคัดเลือกเพื่อนำไปใช้สร้างตัวจำแนกข้อมูลอีกครั้ง ในขั้นตอนต่อไป ฉะนั้นค่า Supt และค่า Confd จึงไม่มีความหมายต่อการหา CARs ในขั้นตอนนี้

ยกตัวอย่างข้อมูลที่นำมาสร้าง CARs ดังแสดงในตารางที่ 1

ตารางที่ 1. แสดงตัวอย่างข้อมูลที่นำมาสร้าง CARs

กฎหมาย และมาตรา	วลี
กฎหมายลักษณะอาญา มาตรา 1	ทำร้ายร่างกาย
กฎหมายลักษณะอาญา มาตรา 2	ทำร้ายร่างกาย
กฎหมายวิธีพิจารณาความอาญา มาตรา 3	ทำร้ายร่างกาย
กฎหมายแพ่งและพาณิชย์ มาตรา 4	ไม่ชำระเงินตามสัญญา
กฎหมายแพ่งและพาณิชย์ มาตรา 4	ไม่ชำระเงินตามสัญญา

เนื่องจากประเภทของข้อมูล ที่ต้องการจำแนก มีอยู่ 2 ประเภท คือ รหัสกฎหมาย และมาตราของกฎหมาย ทำให้การหากฎเกณฑ์ต้องแบ่งออกเป็น 2 ชุด ซึ่งจะมีลักษณะดังต่อไปนี้

ตารางที่ 2. แสดง CARs ระดับกฎหมาย

Rules	Supt	Confd
ทำร้ายร่างกาย → กฎหมายลักษณะอาญา	40.00%	66.67%
ทำร้ายร่างกาย → กฎหมายวิธีพิจารณาความอาญา	20.00%	33.33%
ไม่ชำระเงินตามสัญญา → กฎหมายแพ่งและพาณิชย์	40.00%	100.00%

จากอัลกอริทึมรูปที่ 3 โปรแกรมจะเริ่มต้นด้วยการกำหนดค่าเริ่มต้นให้กับตัวแปร f มีค่าเป็น 0 (บรรทัดที่ 1) หลังจากนั้นจะเริ่มวนรอบทำงานกับข้อมูลของวลีทีละตัว เป็นจำนวน Count(Word) รอบ ซึ่งเท่ากับจำนวนวลีที่มีอยู่ในพจนานุกรม โดยในแต่ละรอบจะเริ่มต้นด้วยการดึงข้อมูลทั้งหมดของวลีนั้นขึ้นมาก่อน ด้วยฟังก์ชัน GetDataOfWord (บรรทัดที่ 3) ตัวแปรที่ผ่านเข้าสู่ฟังก์ชันคือค่าตัวแปร w ซึ่งก็คือรหัสของวลีในพจนานุกรมนั่นเอง ข้อมูลที่ดึงออกมานี้คือข้อมูลที่ได้จากขั้นตอนการเตรียมข้อมูล ซึ่งประกอบด้วย รหัสวลี, รหัสกฎหมาย และมาตรา (ในกรณีนี้เราจะไม่นำรหัสคดีความมาพิจารณาด้วย) ข้อมูลที่ได้ออกมาจากฟังก์ชันจะถูกจัดเรียงลำดับตามรหัสกฎหมาย และมาตรา (Class)

ตารางที่ 3. แสดง CARs ระดับมาตรา

Rules	Supt	Confd
ทำร้ายร่างกาย → กฎหมายลักษณะอาญา มาตรา 1	20.00%	33.33%
ทำร้ายร่างกาย → กฎหมายลักษณะอาญา มาตรา 2	20.00%	33.33%
ทำร้ายร่างกาย → กฎหมายวิธีพิจารณาความอาญา มาตรา 3	20.00%	33.33%
ไม่ชำระเงินตามสัญญา → กฎหมายแพ่งและพาณิชย์ มาตรา 4	40.00%	100.00%

สำหรับอัลกอริทึมในการหา CARs สามารถแสดงได้ดังรูปที่ 3 โดยการหา CARs ระดับกฎหมาย และ CARs ระดับมาตราไม่มีความแตกต่างกัน

```
1   f = 0
2   FOR  w = 1  TO  Count(Word) DO
3       GetDataOfWord(w)
4       FOR  i = 1  TO  Count(Data)  DO
5           IF  (i > 1) and (Class_i <> Class_i-1) THEN
6               supt = f / Count(All)*100
7               conf = f / Count(Data) * 100
8               ReturnRule( W → Class , supt, conf)
9               f = 0
10          ELSE
11              f = f+1
12          END
13      END
14  END
```

รูปที่ 3. แสดงอัลกอริทึมของโปรแกรมสร้าง CARs

หลังจากนั้นโปรแกรมจะวนรอบเป็นจำนวน Count(Word) ครั้ง ซึ่งก็คือจำนวนรายการข้อมูลที่ได้จากฟังก์ชัน GetDataOfWord เพื่อหา CARs ของวลีนั้นๆออกมา (บรรทัดที่ 4 ถึง 13) สำหรับการหา CARs ระดับกฎหมายนั้น ให้พิจารณาค่าตัวแปร Class ที่ปรากฏอยู่ในบรรทัด ที่ 5 และ บรรทัดที่ 8 เป็นรหัสกฎหมาย โปรแกรมจะวนรอบนับจำนวนรายการแล้วเก็บไว้ที่ตัวแปร f เรื่อยๆ จนกระทั่งเมื่อ Class หรือรหัส

กฎหมายของรายการข้อมูลเปลี่ยนไปจากรายการข้อมูลชุดก่อนหน้า (บรรทัดที่ 5) ซึ่งหมายความว่ารายการข้อมูลของรหัสกฎหมายนั้นๆ ได้หมดไปแล้ว และกำลังเริ่มเข้าสู่รายการของรหัสกฎหมายตัวใหม่เนื่องจากว่าข้อมูลมีการเรียงลำดับไว้แล้วตามรหัสกฎหมาย ถึงขั้นตอนนี้โปรแกรมจะสร้างกฎขึ้นมา 1 กฎ (บรรทัดที่ 8) และกำหนดค่าตัวแปร f ให้กลับเป็นค่า 0 อีกครั้งหนึ่ง สำหรับกฎที่โปรแกรมสร้างขึ้นนี้ จะมีค่าฝั่งซ้ายมือเป็น w ซึ่งก็คือรหัสวลีนั่นเอง ส่วนทางฝั่งขวามือของกฎจะมีค่า Class ซึ่งก็คือรหัสกฎหมายนั่นเอง ส่วนค่า Supt สามารถหาได้จากจำนวนรายการที่นับไว้ที่ตัวแปร f หารด้วย จำนวนรายการข้อมูลทั้งหมดของระบบ คูณด้วย 100 (บรรทัดที่ 6) และค่า Confd สามารถหาได้จากจำนวนรายการที่นับไว้ที่ตัวแปร f หารด้วย จำนวนรายการของวลีนั้นๆ คูณด้วย 100 (บรรทัดที่ 7)

สำหรับการหา CARs ระดับมาตรา ก็ทำตามอัลกอริทึมที่แสดงไว้เช่นกัน แต่ตัวแปร Class ที่ปรากฏอยู่ในบรรทัด ที่ 5 และ บรรทัดที่ 8 ให้พิจารณาเป็นรหัสมาตราของกฎหมายแทน

4. การสร้างตัวจำแนกจาก Class Association Rules

ขั้นตอนนี้เป็นขั้นตอนการสร้างตัวจำแนก โดยการนำ CARs ของแต่ละวลีมาจัดเรียงตามค่าความสำคัญ (Precedence) จากมากไปน้อย ซึ่งค่าความสำคัญของกฎจะพิจารณาจากค่า Confd และค่า Supt ซึ่งมีรายละเอียดดังนี้

ถ้าค่า Confd ของกฎที่ 1 มากกว่าค่า Confd ของกฎที่ 2 ให้กฎที่ 1 มีค่าความสำคัญมากกว่ากฎที่ 2

ถ้าค่า Confd ของกฎที่ 1 เท่ากับค่า Confd ของกฎที่ 2 ให้พิจารณาค่า Supt ถ้าค่า Supt ของกฎที่ 1 มากกว่าค่า Supt ของกฎที่ 2 ให้กฎที่ 1 มีค่าความสำคัญมากกว่ากฎที่ 2

ถ้าค่า Confd และค่า Supt ของกฎที่ 1 และกฎที่ 2 มีค่าเท่ากัน ให้กฎที่มาก่อนมีค่าความสำคัญมากกว่า

เนื่องจากเราได้จัดเก็บ CARs ไว้ในฐานข้อมูล ทำให้การเรียงลำดับค่าความสำคัญของกฎสามารถทำได้ด้วยภาษา SQL คือ *"SELECT Rules FROM CARs ORDER BY Confd DESC, Supt DESC"*
กฎเกณฑ์ที่ได้ผ่านการจัดเรียง ค่าความสำคัญแล้ว ก็คือตัวจำแนกข้อมูลนั้นเอง ซึ่งใช้สัญลักษณ์ว่า CARsp (Class Association Rules sorted with Precedence)

เมื่อมีคดีความใหม่ ที่ต้องการทำนายว่าใช้กฎหมายฉบับไหนในการพิจารณาเข้าสู่ระบบ ระบบจะทำงาน โดยแบ่งเป็น 2 ขั้นตอนดังต่อไปนี้

ขั้นตอนที่ 1 ตัดคำคดีความใหม่ที่เข้ามา ด้วยพจนานุกรมวลี เช่นเดียวกับการตัดคำในขั้นตอนการเตรียมข้อมูล ผลลัพธ์ที่ได้จะถูกแปลงเป็นตัวเลขซึ่งแทนรหัสของวลีต่างๆในคดีความนั้น

ขั้นตอนที่ 2 เลือกกฎเกณฑ์ของวลีแต่ละวลีในคดีความ จาก CARsp ออกมา กฎเกณฑ์แต่ละชุดที่ได้นี้จะถูกจัดเรียงลำดับตามค่าความสำคัญ (Precedence) อีกครั้งหนึ่ง เนื่องจากเราได้จัดเก็บ CARs ไว้ในฐานข้อมูล ทำให้การเรียงลำดับค่าความสำคัญของกฎสามารถทำได้ด้วยภาษา SQL เช่นเดิม คือ *"SELECT Rules FROM CARs WHERE Word in (w1,w2,w3,...) ORDER BY Confd DESC, Supt DESC"*

กฎเกณฑ์ที่มีค่าความสำคัญสูงที่สุด (รายการแรกของผลการทำงานตามคำสั่งภาษา SQL) ก็คือผลการทำนายประเภทข้อมูลที่ได้ ซึ่งจะทำให้เรารู้ว่าคดีความนั้นๆ ใช้กฎหมายฉบับไหนในการพิจารณา

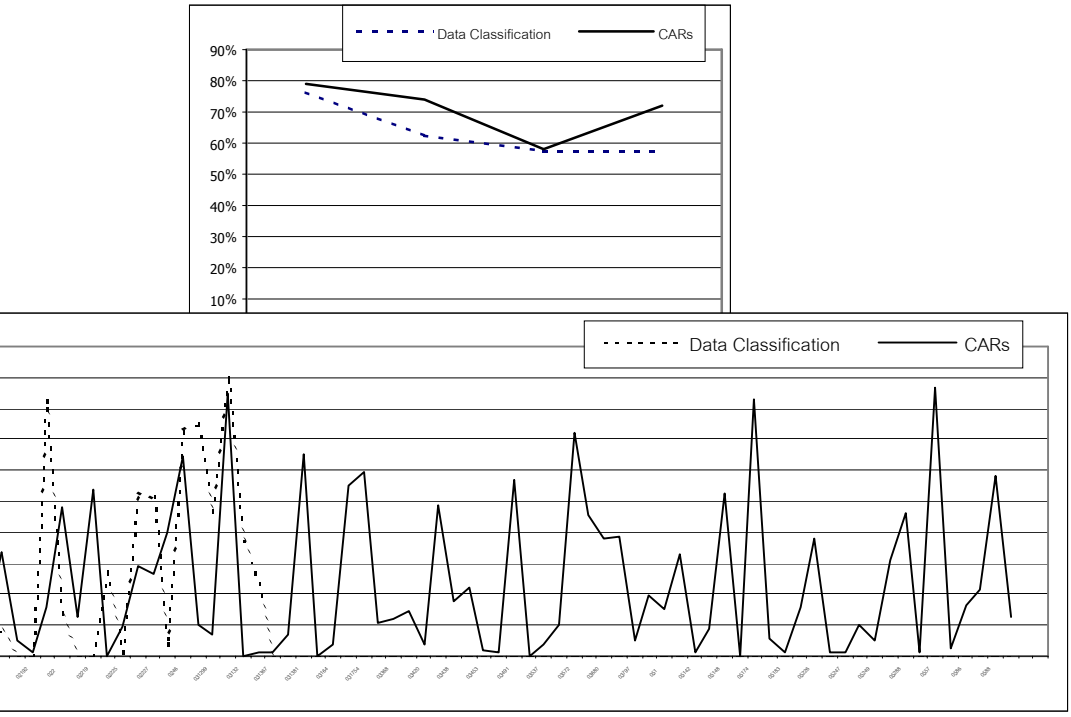
5. ผลการทดลอง

จากข้อมูลที่ผ่านขั้นตอนการเตรียมข้อมูลแล้ว เรานำข้อมูลที่เหลือ 30% มาทดสอบกับระบบใหม่ที่ได้ เพื่อเปรียบเทียบเปอร์เซ็นต์ความถูกต้องกับผลลัพธ์ที่ได้จากระบบเดิมที่มีการสร้างตัวจำแนกด้วยเทคนิค Data Classification แบบทั่วไป ซึ่งก็คือเทคนิค Decision Tree [9] โดยเราแบ่งการทดสอบออกเป็น 2 ชุด คือชุดทดสอบระดับกฎหมาย ซึ่งประกอบด้วยประเภทข้อมูลที่ต้องการจำแนก 4 ประเภท และชุดทดสอบระดับมาตรา ซึ่งประกอบด้วยประเภทข้อมูลที่ต้องการจำแนก 70 ประเภท ผลการเปรียบเทียบแสดงได้ดังกราฟรูปที่ 4.1 และ 4.2 โดยแกนตั้งของกราฟแสดงเปอร์เซ็นต์ความถูกต้องของผลการจำแนกข้อมูล และแกนนอนแสดงรหัสกฎหมาย และมาตราของกฎหมายตามลำดับ

จากกราฟ 4.1 และ 4.2 แสดงให้เห็นว่าการสร้างตัวจำแนก ด้วยเทคนิคใหม่ที่เสนอได้ผลลัพธ์ในการทำนายดีกว่าตัวจำแนกที่สร้างขึ้นด้วยเทคนิค Data Classification แบบทั่วไป โดยการทดสอบในระดับกฎหมาย พบว่าระบบใหม่ที่เสนอ ให้เปอร์เซ็นต์ความถูกต้องสูงกว่าในกฎหมายทุกฉบับ ส่วนในการทดสอบระดับมาตรานั้น ระบบใหม่ที่เสนอได้เปอร์เซ็นต์ความถูกต้องใกล้เคียงกับระบบเดิมเมื่อเทียบมาตราต่อมาตรา แต่ระบบใหม่สามารถจำแนกประเภทข้อมูลได้จำนวนกลุ่มมากกว่า คือสามารถทำนายผลลัพธ์ได้โดยที่เปอร์เซ็นต์ความถูกต้องไม่เป็น 0 เกือบทุกมาตรา ต่างจากระบบเดิมที่สามารถทำนายได้เพียง 20 มาตราจากทั้งสิ้น 70 มาตรา

จากกราฟรูปที่ 4.2 เมื่อเปรียบเทียบเปอร์เซ็นต์ความถูกต้องของทั้ง 2 ระบบโดยพิจารณามาตราต่อมาตรา จะพบว่าทั้ง 2 ระบบมีเปอร์เซ็นต์ความถูกต้องสอดคล้องกัน โดยในมาตราที่เปอร์เซ็นต์ความถูกต้องต่ำ ก็จะได้เปอร์เซ็นต์ความถูกต้องต่ำทั้ง 2 ระบบ และในมาตราที่เปอร์เซ็นต์

ความถูกต้องสูง ก็จะสูงตามกันทั้ง 2 ระบบเช่นกัน แสดงให้เห็นว่าการทำงานของระบบทั้ง 2 นั้นให้ผลลัพธ์ที่สอดคล้องตรงกัน ในมาตราที่ได้มาตราเปอร์เซ็นต์ความถูกต้องสูง มีสาเหตุจากการที่มาตรานั้นๆ มีวลีที่เด่นชัดแตกต่างจากคำในมาตราอื่นๆ ส่วนในมาตราที่ได้เปอร์เซ็นต์ความถูกต้องต่ำ ก็เนื่องมาจากนั้นๆ ไม่มีวลีเด่น ที่สามารถแยกความแตกต่างจากวลีในมาตราอื่นๆได้เลย



รูปที่ 4.1 แสดงกราฟเปรียบเทียบเปอร์เซ็นต์ความถูกต้องของตัวจำแนกที่สร้างขึ้นด้วยเทคนิค Data Classification ทั่วไป กับตัวจำแนกที่สร้างจาก CARs ระดับกฎหมาย

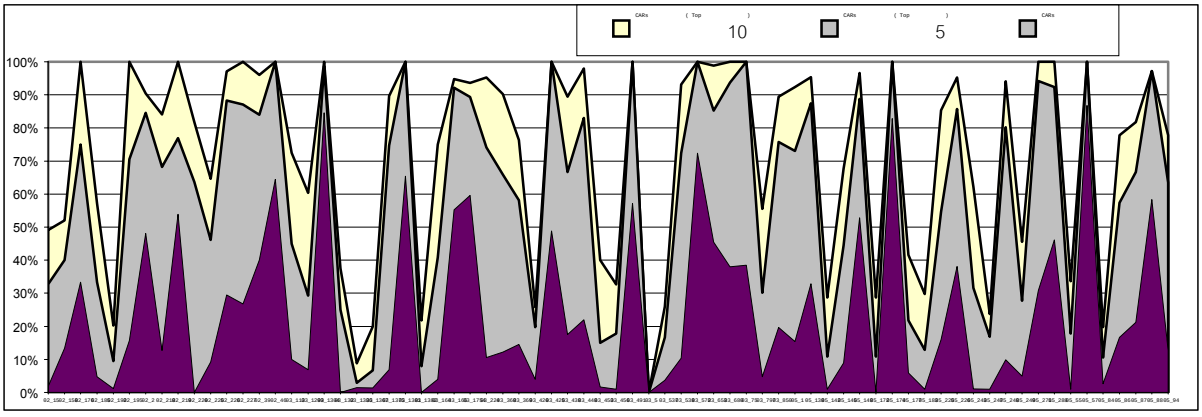
รูปที่ 4.2 แสดงกราฟเปรียบเทียบเปอร์เซ็นต์ความถูกต้องของตัวจำแนกที่สร้างขึ้นด้วยเทคนิค Data Classification ทั่วไป กับตัวจำแนกที่สร้างจาก CARs ระดับมาตรา

รูปที่ 5. แสดงกราฟเปรียบเทียบเปอร์เซ็นต์ความถูกต้องของตัวจำแนกข้อมูลที่สร้างขึ้นด้วยการตัดคำแบบเดิม กับการตัดคำแบบวลี

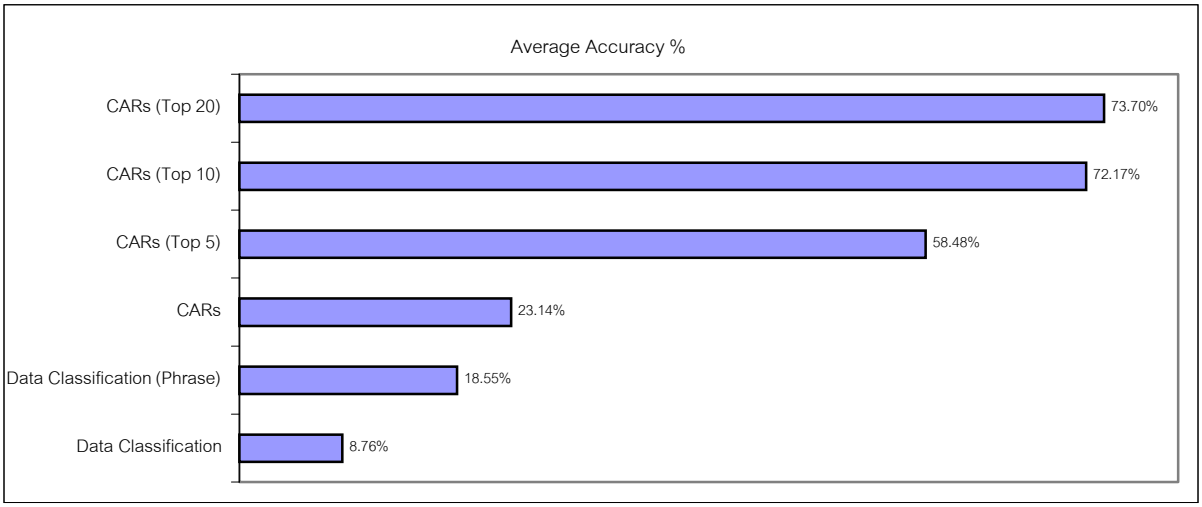
สาเหตุที่ทำให้ระบบใหม่ที่เสนอ มีเปอร์เซ็นต์ความถูกต้องของการทำนายสูงขึ้น เกิดจากปัจจัย 2 ประการ ปัจจัยแรกคือการใช้เทคนิค Association Rule Discovery มาช่วยในการสร้างตัวจำแนก และอีกปัจจัยคือการเปลี่ยนรูปแบบการตัดคำเป็นการตัดวลี เพื่อสนับสนุนสมมติฐานดังกล่าว เราได้ทำการทดลองกับระบบเดิมที่มีการสร้างตัวจำแนกข้อมูลด้วยเทคนิค Data Classification แบบทั่วไปอีกครั้ง แต่ปรับเปลี่ยนฟังก์ชันในการตัดคำ เป็นการตัดวลีแทน ผลลัพธ์ที่ได้ เป็นไปตามกราฟรูปที่ 5 โดยแกนตั้งของกราฟแสดงเปอร์เซ็นต์ความถูกต้องของผลการจำแนกข้อมูล และแกนนอนแสดงกฎหมายมาตราต่างๆ

จากกราฟรูปที่ 5 แสดงให้เห็นว่าการทดลองใหม่ คือการตัดคำแบบวลี ให้เปอร์เซ็นต์ความถูกต้องใกล้เคียงกับการทดลองเก่า เมื่อเปรียบเทียบมาตรา ต่อมาตรา แต่พบว่าในการทดลองใหม่ ตัวจำแนกสามารถจำแนกข้อมูลได้มากประเภทกว่า

เนื่องจากว่ารูปแบบของข้อมูลทางกฎหมาย ที่ต้องการหากฎหมายที่เหมาะสมสำหรับการพิจารณาคดีความ โดยปกติแล้วในแต่ละคดีความจะใช้กฎหมายมากกว่า 1 ฉบับในการพิจารณา ซึ่งไม่สามารถทำได้ด้วยเทคนิค Data Classification แบบทั่วไปดังที่ได้กล่าวมาแล้ว เมื่อเราปรับเปลี่ยนขั้นตอนการสร้างตัวจำแนก มาเป็นการสร้างตัวจำแนกจาก CARs ทำให้สามารถแก้ปัญหาดังกล่าว คือแทนที่จะเลือกกฎเกณฑ์จาก CARsp 1 กฎที่มีค่าความสำคัญสูงสุด ก็เปลี่ยนมาเป็นเลือกกฎเกณฑ์ ครั้งละ 5 หรือ 10 กฎแรกที่มีค่าความสำคัญสูงสุดแทน วิธีนี้ทำให้ระบบสามารถทำนายผลการจำแนกได้ครั้งละมากกว่า 1 ประเภท ซึ่งหมายความว่าระบบได้เสนอกฎหมายมาตราต่างๆ ครั้งละ 5 หรือ 10 มาตรา ที่คิดว่าเหมาะสมสำหรับคดีความนั้นๆ เพื่อให้ผู้ใช้นำไปศึกษาอันทำให้ก่อประโยชน์ต่อไป การทำนายผลเช่นนี้ จะทำให้ผลการทำนายมีเปอร์เซ็นต์ความถูกต้องสูงขึ้น ซึ่งสามารถดูค่าเปรียบเทียบได้จากกราฟ รูปที่ 6



รูปที่ 6. แสดงกราฟเปรียบเทียบเปอร์เซ็นต์ความถูกต้องของตัวจำแนก เมื่อทำนายผลครั้งละ 1, 5 และ 10 มาตรา



รูปที่ 7. แสดงกราฟเปรียบเทียบเปอร์เซ็นต์ความถูกต้องของการทดลองแบบต่างๆ

อีกปัญหาหนึ่งที่ได้กล่าวไว้แล้วในตอนต้น คือเรื่องการแบ่งกลุ่มข้อมูลแบบมีลำดับชั้น ซึ่งไม่สามารถทำได้ในระบบเดิม ทำให้ต้องแบ่งตัวจำแนกข้อมูลเป็น 2 ชุด โดยมีการใช้งานแยกจากกันโดยเด็ดขาด ผู้ใช้ต้องกำหนดว่าจะเลือกใช้ตัวจำแนกชุดไหน สำหรับระบบใหม่ที่เสนอนี้ ถึงแม้ว่าตัวจำแนกข้อมูลจะยังคงต้องแบ่งเป็น 2 ชุดเช่นเดิม แต่ด้วยรูปแบบของการจำแนกที่ใช้การเลือกกฎเกณฑ์ที่มีค่าความสำคัญสูงสุด ขึ้นมาทำนายประเภทข้อมูล ทำให้ตัวจำแนกใหม่นี้ สามารถเลือกกฎเกณฑ์จากตัวจำแนกทั้ง 2 ชุดได้ในขณะเดียวกัน โดยสามารถกำหนดเงื่อนไขได้ว่า ถ้าหากค่าความสำคัญของกฎเกณฑ์ระดับมาตรา มีค่าน้อยเกินไป ก็สามารถเลือกทำนายผลเฉพาะระดับกฎหมายได้โดยอัตโนมัติ

เมื่อเปรียบเปอร์เซ็นต์ความถูกต้องโดยเฉลี่ย ของตัวจำแนกมาตราแบบต่างๆ ให้ผลออกมาตาม กราฟรูปที่ 7 เมื่อแกนนอนคือเปอร์เซ็นต์ความถูกต้องโดยเฉลี่ยของการจำแนก และแกนตั้งคือตัวจำแนกแบบต่างๆ เริ่มต้นจากตัวจำแนกระบบเดิม ที่ถูกสร้างด้วยเทคนิค Data Classification แบบทั่วไป ถัดมาคือตัวจำแนก ที่ได้เปลี่ยนฟังก์ชันการตัดคำเป็นการตัดวลี ถัดมาคือตัวจำแนกที่สร้างขึ้นจาก CARs และที่เหลือคือตัวจำแนกที่สร้างขึ้นจาก CARs เช่นกันแต่มีการทำนายผลครั้งละ 5, 10 และ 20 มาตราตามลำดับ

6. สรุป

จากกราฟรูปที่ 7 แสดงให้เห็นอย่างชัดเจนว่า งานวิจัยนี้สามารถปรับปรุงเทคนิคในการสร้างตัวจำแนก ให้มีประสิทธิภาพ ทางด้านความถูกต้องในการจำแนกเพิ่มมากขึ้น นอกจากนี้ระบบใหม่ยังช่วยแก้ปัญหาต่างๆ ที่เกิดขึ้นจากระบบเดิม ซึ่งประกอบด้วย

1. ปัญหาเรื่องข้อจำกัดทางด้านจำนวนประเภทข้อมูลที่ระบบสามารถจำแนกได้ ซึ่งก็คือจำนวนมาตราที่ระบบสามารถทำนายได้นั่นเอง ในระบบเดิม ตัวจำแนกสามารถทำนายผลได้เพียง 20 มาตราเท่านั้น ซึ่งจากผลการทดลองระบบใหม่ ได้แสดงให้เห็นแล้วว่าระบบใหม่ที่เสนอ สามารถทำนายผลได้ครบทุกมาตรา

2. ปัญหาเรื่องการทำนายผลให้กับคดีความครั้งละมากกว่า 1 มาตรา ซึ่งระบบเดิมไม่สามารถทำได้ จากการทดลองแสดงให้เห็นแล้วว่าระบบใหม่ที่มีการสร้างตัวจำแนกจาก CARs สามารถทำนายผลลักษณะเช่นนี้ได้ โดยสามารถระบุจำนวนมาตราที่ต้องการให้ระบบทำนายได้ นอกจากนี้ด้วยวิธีการดังกล่าว ยังเป็นการเพิ่มเปอร์เซ็นต์ความถูกต้องเฉลี่ยของการทำนายด้วย

3. ปัญหาเรื่องการทำนายผลแบบมีระดับชั้น คือระดับกฎหมาย และระดับมาตรา ระบบใหม่ที่เสนอนี้ ถึงแม้ว่าตัวกฎเกณฑ์หรือ CARs จะถูกจัดเก็บแยกกันเป็น 2 ชุด เช่นเดียวกับระบบเดิม แต่ในขั้นตอนการทำนายผล ระบบสามารถทำนายผลระดับกฎหมาย และระดับมาตราได้ในครั้งเดียว โดยระบบสามารถตรวจสอบได้ว่าถ้าเปอร์เซ็นต์ความถูกต้อง ของผลการทำนายระดับมาตรา มีแนวโน้มที่ต่ำก็จะยกเลิกการทำนายระดับมาตรานั้นแล้วทำนายเฉพาะระดับกฎหมาย

ระบบที่เสนอในงานวิจัยนี้ มุ่งเน้นที่จะแก้ไขปัญหาเฉพาะทาง โดยมุ่งเน้นที่จะพัฒนาระบบที่สามารถทำนายกฎหมายที่ใช้สำหรับพิจารณาคดีความ โดยอัตโนมัติ ซึ่งทำให้การออกแบบ และเทคนิคที่เสนอมีความเฉพาะเจาะจงสำหรับรูปแบบข้อมูล และวิธีการในการทำงานเช่นนี้ ฉะนั้นการนำเทคนิคที่เสนอในงานวิจัยนี้ไปประยุกต์ใช้กับงานลักษณะอื่นๆ ก็จำเป็นต้องปรับปรุงเทคนิคและวิธีการหลายจุด เพื่อให้ระบบสามารถตอบสนองต่อความต้องการของระบบนั้นๆ ได้

เปอร์เซ็นต์ความถูกต้องที่ได้จากระบบที่เสนอในงานวิจัยฉบับนี้ยังไม่สูงนัก ซึ่งมีแนวโน้มว่าจะสามารถปรับปรุงให้ระบบมีความสามารถเพิ่มมากขึ้นได้ ไม่ว่าจะเป็นการเสนอเทคนิคในการสร้างตัวจำแนกด้วยวิธีอื่นๆ หรือการปรับปรุงขั้นตอนการเตรียมข้อมูลให้มีประสิทธิภาพเพิ่มมากขึ้น

เอกสารอ้างอิง

[1] คณาจารย์คณะนิติศาสตร์มหาวิทยาลัยกรุงเทพ. ความรู้เบื้องต้นเกี่ยวกับกฎหมาย. แผนกตำราและคำสอน มหาวิทยาลัยกรุงเทพ. หน้า 5, 2540.

[2] Witten, I.H., Moffat, A. and Bell, T.C. Managing Gigabytes: Compressing and indexing Documents and Images. Morgan Kaufmann Publishers, page 4, 1999.

[3] K. Waiyamai and L. Lakhal 2000. Knowledge Discovery Very Large Database using Frequent Concept Lattices. Springer-Verlag Lecture Notes in Artificial Intelligent, pages 437-445, June 1810.

[4] M. S. Chen, J. Han, and P. S. Yu. Data Mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.

[5] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of ACM, 39:58-64, 1996.

[6] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advance in knowledge Discovery and Data Mining. AAAi/MIT Press, 1996.

[7] M. Berry and G. Linoff. Data Mining Techniques: For Marketing Sales and Customer Support, John Wiley & Sons, 1997.

[8] P.K.Chan and S.J.Stolfo. Learning arbiter and combiner trees from partitioned data for scaling machine learning. In Proc. 1st.Int.Conf. Knowledge Discovery and Data Mining (KDD'95), pages 39-44, Montreal, Canada, August 1995.

[9] J.Gehrke, R.Ramakrishnan, and V.Ganti. Rainforest: A framework for fast decision tree construction of large datasets. In

Proc. 1998 Int. Conf. Very Large Datacase, pages 416-427, New York, NY, August 1998.

[10] K. Waiyamai 1998. A novel Approach for Association Rule Discovery. The National Computer Science and Engineering Conferrence (NCSEC'98). Bangkok, Thailand, 19-21 October.

[11] R. Agrawal, T. Imielinski, and A. Sawami. Mining Association Rules between sets of items in large database. SIGMOD'93, pages 207-216, Washington,D.C.

[12] R. Feldman & I. Dagan, "Knowledge Discovery in Textual Database." In Proceedings of the International Conference on Knowledge Discovery and Data Mining, 1995.

[13] T. Pongsiripreeda and K. Waiyamai. Using Data Mining technique to select the laws and articles for lawsuits (in thai language). The National Computer Science and Engineering Conferrence (NCSEC'2000). Bangkok, Thailand, November.

[14] B.Liu and W.Hsu and Y.Ma. Integrating Classification and Association Rule Mining. In Proc. 1998 Int. Conf. KDD'98, 1998.

[15] J.L. Bentley. Programming Pearls. Addision-Wesley, pages 164-167, May 1989.

กฤษณะ ไวยมัย จบปริญญาตรีและปริญญาโททางวิทยาศาสตร์คอมพิวเตอร์ จาก University of Picardie ประเทศฝรั่งเศส จบปริญญาเอกทางด้านวิทยาการคอมพิวเตอร์จาก University of Clermont ประเทศฝรั่งเศส งานวิจัยหลักประกอบด้วยสามส่วนสำคัญ ส่วนแรกคือ ระบบสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Databases: Data Mining) ส่วนที่สองคือ ระบบสารสนเทศ (Information System) และ ส่วนที่สามคือ ระบบฐานข้อมูล (Database Management System) ปัจจุบัน ดร.กฤษณะ ไวยมัย ดำรงตำแหน่งอาจารย์ประจำภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

ธีระวัฒน์ พงษ์ศิริปรีดา จบปริญญาตรีทางวิทยาศาสตร์คอมพิวเตอร์ จากมหาวิทยาลัยธรรมศาสตร์ จบปริญญาโททางด้านวิศวกรรมคอมพิวเตอร์จากมหาวิทยาลัยเกษตรศาสตร์ งานวิจัยหลักคือระบบสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Databases: Data Mining) ปัจจุบันเป็นพนักงานบริษัท เอเทรียม เทคโนโลยี จำกัด ในตำแหน่ง System Analyst

Information Technology Strategic Planning Process for Institutions of Higher Education in Thailand*

Wanwipa Titthasiri
School of Information Technology, Rangsit University
Pathumthani, Thailand

ABTRACT - This study presents a survey on the status of Information Technology (IT) and IT strategic planning in both public and private higher educational institutions in Thailand. Questionnaires were sent to selected personnel from 18 public institutions and 16 private institutions. A total of 112 persons constituting 82.4 percent of the population responded. The findings revealed that there was great interest in using IT in both academic and administrative areas. One of the problems of IT management in Thai higher educational institutions was determined to be the lack of planning. Some important obstacles encountered when developing plans were financial and IT manpower problems. Nevertheless, the findings of the study identified a major factor in that Thai institutions lack understanding of how to develop an IT strategic plan. The data indicated that only half of the representatives of Thai institutions made use of an IT strategic plan but it seems that even then components and processes of many were incomplete. To support the development of IT strategic planning in Thai institutions, an IT strategic planning process for Thai higher educational institutions was proposed in this study. This process was developed, based on the perceptions of the administrative and the IT professional groups, as well as a review of the literature. It is hoped that this process would be used as a guideline to enable planners to develop IT strategic plans in Thai higher educational institutions.

General Terms - Higher Education, Institutions, Planning, Process, Status, Thailand.
Additional Key Words and Phrases - Information Technology (IT), Development, Private Institutions, Public Institutions, Questionnaires, Strategic Planning.

1. INTRODUCTION

Necessary to the success of the information age, society must constantly evolve to accommodate and keep pace with new technological advancements—advancements that can change the way people live, work, and learn. Today Information Technology (IT) is used in almost all facets of society, and as IT undergoes rapid changes, Rowley, Lujan, and Dolence [6] pointed out that focused attention must be applied to stimulating innovations in pedagogy, research, and management through computer usage. Realizing this, universities have already begun to adopt IT. But to accomplish what Rowley, Lujan, and Dolence have stated, not only do students need to become more highly computer literate, classrooms and laboratories must be designed to accommodate and provide new types of technology; also, the university's administrative system must be modernized and integrated into a single platform, and a network designed to create an interface between the campus and the worldwide community.

IT's role in facilitating university activities focused on both administrative and academic areas, where IT is being used in university management, administrative processes, improvement of research, and the teaching-learning process. Because IT is rapidly changing with no fixed limits, universities need a strategic IT plan to guide their future development. Such a plan will help universities address the challenges of budget constraints for new or increased investments in IT, better respond to the rapidly changing IT environment, provide technical support for IT, and develop software and tools not only for research but also for the teaching-learning and administrative processes.

If a university has a good strategic plan, then the risk involved in IT decision making can be reduced. However, many universities are unable to create this important strategic plan because they do not have the proper information and experience to strategically plan and utilize IT. Therefore, IT strategic plans must be carefully developed. Thai universities, in particular, are in need of the IT strategic planning process.

* This full paper is as a dissertation, licensed by The University of Pittsburgh, PA, U.S.A., Year 2000.

The objectives of this study were to describe the current and preferred status of IT and IT strategic planning components and processes in Thai higher education and to develop a proposed IT strategic planning process for Thai higher educational institutions in Thailand.

The expected outcomes of this study were:

(1) to assess the perceptions of administrator and IT professional in Thai higher educational institutions as to the current and preferred status of IT and IT strategic planning; and

(2) to develop an IT strategic planning process for institutions of higher education in Thailand. This process was intended to provide guidance to Thai higher educational institutions in developing IT strategic plans.

2. RESEARCH DESIGN

This study employed a field survey research methodology. The survey instruments measured the current and preferred status of IT and IT strategic planning perceptions of administrators and IT professionals. Based on the survey information, the data were analyzed and the results used to develop an IT strategic planning process for higher educational institutions in Thailand.

2.1 Research Population and Sample

The research population selected for this study was a finite population that included:

(1) Administrators of Thai higher educational institutions, including (a) the vice-presidents for administration, (b) the vice-presidents for academic affairs, and (c) IT department chairpersons (three persons per institutions).

(2) IT professionals in Thai higher educational institutions, including the computing center directors or Chief Information Officer (one person per institution).

The population frame included both public and private higher educational institutions in Thailand, which were founded before 1995, and are still in existence. Further, they were universities, with a student body of at least 4,000 students. The population was comprised of 34 institutions, including 16 private institutions and 18 public institutions. Thus, the total population was 136 persons (102 administrators and 34 IT professionals). Data obtained from all the respondents were used in this study.

2.2 Instrumentation

Two survey instruments, the administrator's questionnaire and the IT professionals' questionnaire were based on a literature review of institutional strategic planning concepts and components of IT systems. The administrator's questionnaire was developed to obtain the perceptions of administrators regarding the current and preferred status of IT and of IT strategic planning. The IT professionals' questionnaire was developed to obtain background data; it also included all questions presented in the administrators' questionnaire. The questions focused on: (a) the background of the administrators and IT professionals, (b) the background of IT systems and IT strategic planning in higher educational institutions in Thailand as reported by IT professionals, and (c) the perceptions of administrators and IT professionals in Thai higher educational institutions on the current and preferred status of IT and IT strategic planning.

2.3 Research Implementation and Data Collection

The participants selected to answer the questionnaires were administrators and IT professionals in Thai higher educational institutions. All 102 administrators and 34 IT professionals from 34 institutions were requested to complete the questionnaires specified by the researcher. First, the questionnaires were mailed to these two groups. As a follow-up, each respondent was contacted by telephone to make sure the questionnaire was received, fill out, and returned. A total of 112 persons constituting 82.4% of the population responded to the researcher's request. Among them, 83 responses (81.4%) were from the administrative group and 29 responses (85.3%) were from the IT professional group. It was believed that this response rate was sufficient for the study and analysis of the current status of IT and IT strategic planning in Thai higher education in general and the development of an IT strategic planning process.

2.4 Data Analysis

The data were analyzed by SPSS PC program. The current and preferred status were aggregated for each institution to determine a mean and a percentage value for each of sample group and for all the institutions. The independent samples t-test was also used to analyze the significantly different means in some part of data.

3. RESULTS AND DISCUSSIONS

3.1 Demographic Data

The returned responses indicated the work experience and the years served in their current position by respondents. The data gave a clear indication of the respondents regarding their work experience in Thai higher education and years served in their current positions. Most administrators and IT professionals have had at least ten years experience in Thai higher education, and have served in their current position between one and three years.

3.2 Current Status of IT

The organizational structure of IT in Thai higher educational institutions was categorized as centralized, centralized (dual), coordinated, and decentralized. This study found that most Thai higher educational institutions (37.9%) were organized

IT administrative area function, 96.6% of Thai institutions implemented IT in their registration systems. Additionally, 93.1%, 75.9%, and 75.9%, respectively of Thai institutions implemented IT in their accounting, personnel, and entrance systems. Second, in the IT academic area function, computer laboratory rooms existed in all Thai institutions. About 97% of Thai institutions had classrooms equipped with IT capabilities, and also had implemented a digital library. Only 27.6% of Thai institutions had equipped their research centers with IT capabilities. Moreover, IT systems were up-to-date and were developed in-house. Changes in user requirements were confronted, as well as the user computer competence during IT system implementation. The availability and capability of IT staff and user competence were important issues in IT functional implementation. As would be expected, the level of staff expertise was one of the critical needs for successful implementation. The study revealed that a computing center existed in all Thai institutions indicating that IT is an important and rapidly expanding tool in Thai institutions. (For more detailed data tables; see [8])

The study indicates that administrators, faculty members, staff, and students in Thai institutions were perceived as enthusiastic about IT utilization. This means that people in Thai institutions were interested and aware of the potential benefits of IT utilities. Software investment was lower than hardware investment in both administrative and academic functions reflecting the ability of institutions to develop some of their own software, while depending on outside vendors for equipment. The use of campus network infrastructures and access, provision of IT training and workshops, and IT support and services were viewed as moderately satisfactory. The availability and capacity of IT manpower and finance are important problems that affect the progress of IT and its implementation in Thai higher educational institutions. Regarding administrators and IT professionals perception of the current status of IT, it was found that there was no significant difference between the means of the administrative and IT professional groups, at a confidence interval of 95%.

All Thai institutions have attempted to update IT, including both hardware and software. Reflecting campuses were under increasing pressure from their constituents, faculty, staff, and students to replace and upgrade out-dated equipment and software. Therefore, because of funding considerations as well as manpower availability, a balance between the economical management and retention of sufficient operating flexibility to deal with innovation and change must be addressed. Not only hardware and software are involved with the current operating IT, but also people. IT training and support is essential for both IT staff and users and must be provided. This will require an expanded institutional-wide training effort to achieve. While currently there is a modest level of training offered, much more must be done to meet the requirements of a rapidly-changing technology. It was believed by administrators and IT professionals that an IT strategic plan was required in Thai institutions for effective IT management.

3.2 The IT Strategic Planning Components

Most current IT strategic planning components in Thai institutions included background, organizational development, IT functional areas, and sub-organizational plans. Only half of the current IT strategic planning included environmental assessment as a component. The study found that changes external to the institution were not an important input to the current strategic planning situation. This was an unexpected finding since environmental scanning is considered to be a basic activity in the development of a strategic plan. The absence of this component could easily affect the success of the final plan since external changes could be occurring that may result in a total redirection of software or hardware utilization. The only reason this component may not be so critical is that many of the plans were relatively short-term covering only a few years.

However, background, environmental assessment, organizational development, IT functional areas, and sub-organizational area plans were preferred as components of the IT strategic planning of Thai institutions. Nevertheless, it was found that the components of background and environmental assessment were preferred at a lower level than other components. It might be because of the lack of understanding by people involved in IT strategic planning. Most IT personnel are not formally trained in management and have to learn on the job. If IT planning is to be strategic, then it must address the external environment, particularly if it is a multi-year plan. In addition, it was found that the preferred level of the IT professionals and the administrators had significantly different means in the components of organizational development, proved by the independent samples t-test. Moreover, the study reflected the fact that organizational development was considered as an important component in both actual and preferred strategic planning.

3.4 The IT Strategic Planning Process

Among the existing IT strategic planning processes in Thai institutions, the responsibility for planning was under: (a) an IT strategic planning committee; and (b) a computer or IT-related center. The first approach, an IT strategic planning committee, had members who typically included the computer or IT-related center director, the vice-president for development and planning, and the dean of the engineering or IT school, with an average membership of 14. In the second approach, a computer or IT-related center approach, the president, the vice-presidents for development and planning, and the IT staff played important roles in the IT strategic planning, apart from the director of computer or IT-related center itself. Another important aspect was that the president, the vice-president for development and planning, the IT staff, the director of the computer or IT-related center, and the computer science or IT-related department chairperson, have been involved in the process at a high percentage, but that students have not been involved. Additionally, most of the steps in the process agreed with the IT and strategic planning literature [1][5] but a lower percentage of respondents agreed with environmental assessment, functional strategic formulation, IT strategies and action plan dissemination, and IT strategic planning process evaluation. This means that in many cases existing IT strategic

planning is not an ongoing process. Therefore, the IT plan will not represent the current situation in planning. A university-wide IT advisory committee is usually responsible for the IT strategic plan review and approval. A formal written IT strategic plan is required by most institutions covering a one-to-three year time frame. This seems to be too little time for a strategic plan and represents more of an operational type of plan.

The preferred IT strategic planning process was in agreement with IT literature [1][6] which would indicate an increasingly better understanding by people in IT strategic planning. The preferred IT strategic planning process for Thai higher educational institutions specifies that an IT strategic planning committee be responsible for IT strategic planning. This committee's members should include the director of computer or IT-related center, the vice-president for development and planning, the IT staff, the vice-president for academic affairs, the president, the dean of the engineering or IT school, the vice-president for administration, with an average of seven or eight members in all. Another important committee should be the university-wide IT advisory committee that is responsible for the review and approval of the IT strategic plan. A formal written plan should be prepared by the IT strategic planning committee. The directors of computer or IT-related centers, the IT staff, the vice-presidents for development and planning, and the computer science or IT-related department chairperson need to get involved in the process. Additionally, the planning process steps which were included in the existing IT strategic planning process at low percentages—environmental assessment, functional strategic formulation, IT strategies and action plan dissemination, and IT strategic planning process evaluation—were in need of strengthening.

However, the researcher agrees with the preferred IT strategic planning process that it would be appropriate to improve the planning process for Thai institutions. IT strategic planning should be under an IT strategic planning committee instead of the computer or IT-related center director as is done at present. The university-wide IT advisory committee should be responsible for the review and approval of the IT strategic plan. A formal written plan should be an output of the planning process. The process should start with an initiation and agreement on IT strategic planning and development of a project plan for implementing the IT plan. Afterwards, the environmental assessment should be implemented to define vision, mission, goals, and objectives and to identify the IT strategic issues in specific functional areas, and then a prioritization of strategies completed according to costs, benefits, and needs. Later, functional strategies should be formulated and reviewed by key stakeholders. After review, the IT strategic plan should be announced and disseminated, and then an action plan should be developed and discussed with key people before dissemination. The final phase should be to execute program implementation and then to evaluate the IT strategic planning process. The last thing is that planning has to provide for the continual implementation of the plan. However, significantly different means were found among the IT professionals and administrators, at a confidence interval of 95%, in the steps of defining vision, mission, goals, and objectives, identifying the IT strategic issues in specific functional areas, inviting key stakeholder group to review IT strategic plan, developing operational strategy, discussing the action plan with key people, executive program implementation, and evaluating IT strategic planning process.

Regarding the obstacles to implementing IT strategic planning in Thai institutions, lack of understanding by people involved in IT strategic planning, IT manpower problems, and financial problems were predicted by most administrators. In general, the IT strategic planning should be developed under institutional strategic planning, but more Thai institutions had IT than institutional strategic plans, as shown by the study. Therefore, knowledge or concepts of strategic planning should be introduced in Thai higher educational institutions. In addition, planning has to be considered, based upon the availability and capability of IT staff and funding.

However, it was agreed that IT is still rapidly expanding and is offering an increasing number of new opportunities at a lower cost. Selection of alternative strategies was a complex and difficult task because of unknowns, uncertainties, and constraints. Therefore, Thai higher education is now in need of more and better strategic planning in order to make correct decisions.

1. DEVELOPMENT OF AN INFORMATION TECHNOLOGY STRATEGIC PLANNING PROCESS

According to this study, evidence was cited that IT strategic plans are now required for IT functions in Thai higher educational institutions. Almost 60% of Thai institutions had IT strategic plans covering a one-to-three-year time period and another 30% were developing IT strategic plans. However, it was found that existing IT strategic planning in Thai higher educational institutions had several weaknesses in terms of both components and processes. This might result from a lack of understanding of the strategic planning process. Therefore, based on these research findings, need exists for additional support to develop an effective IT strategic planning process in Thai institutions.

This section presents the concepts of an IT strategic planning process designed to assist planners in developing an IT strategic plan for Thai higher educational institutions. The proposed process is drawn from the results of this study (based on the preferred status), the IT literature[5], and the strategic planning literature[1]. This study reflected the fact that most Thai institutions preferred producing a formal written plan as an output of the IT strategic planning process. This planning document would enable people to understand where they are now (i.e., what exists), to imagine where they want to be, and to understand how the IT functions have been successfully implemented. The results of the study indicated that strategic planning produces not only a document, but also provides for the continual implementation of the plan. IT strategic planning is long-term and continuous. The study further indicated that all stakeholders, including administrators, faculty, staff members, students, and others who will benefit from the implementation of the plan, should be included in the planning process.

The proposed IT strategic planning process is composed of four major phases. From the literature and the results of study, the process includes: (a) organizing a planning team, (b) fact-finding and trend assessment, (c) determination and dissemination of IT strategies; and (d) implementation and revision of programs (see Figure 1). It is assumed that a larger institutional strategic planning framework exists.

Phase 1: Organizing a Planning Team

Based on the literature [6], IT strategic planning, which is a subset of institutional strategic planning, begins with a review of institutional goals and objectives before initiating any IT strategic planning and an agreement to proceed. This means that after discussions within an institutional policy committee and among the executive officers of a university, agreement should be reached that the interest of the university would be served by having an IT plan. Once that is affirmed, then IT strategic planning activities can be initiated.

First, a university-wide IT advisory committee should be established. Study findings revealed that the president was most frequently the chairman of this committee. This may be because in Thai culture, presidents play roles at the broad institutional level. The task of this committee would be to develop and recommend to the administration approaches to institutional IT problems. It is therefore recommended from the findings that this committee should be composed of the president as chairman, vice-president for administration, vice-president for academic affairs, vice-president for development and planning, senior faculty and staff representatives, and external consultants. The aim being to provide a forum for discussion of IT problems, needs, future planning, and review of an IT strategic plan.

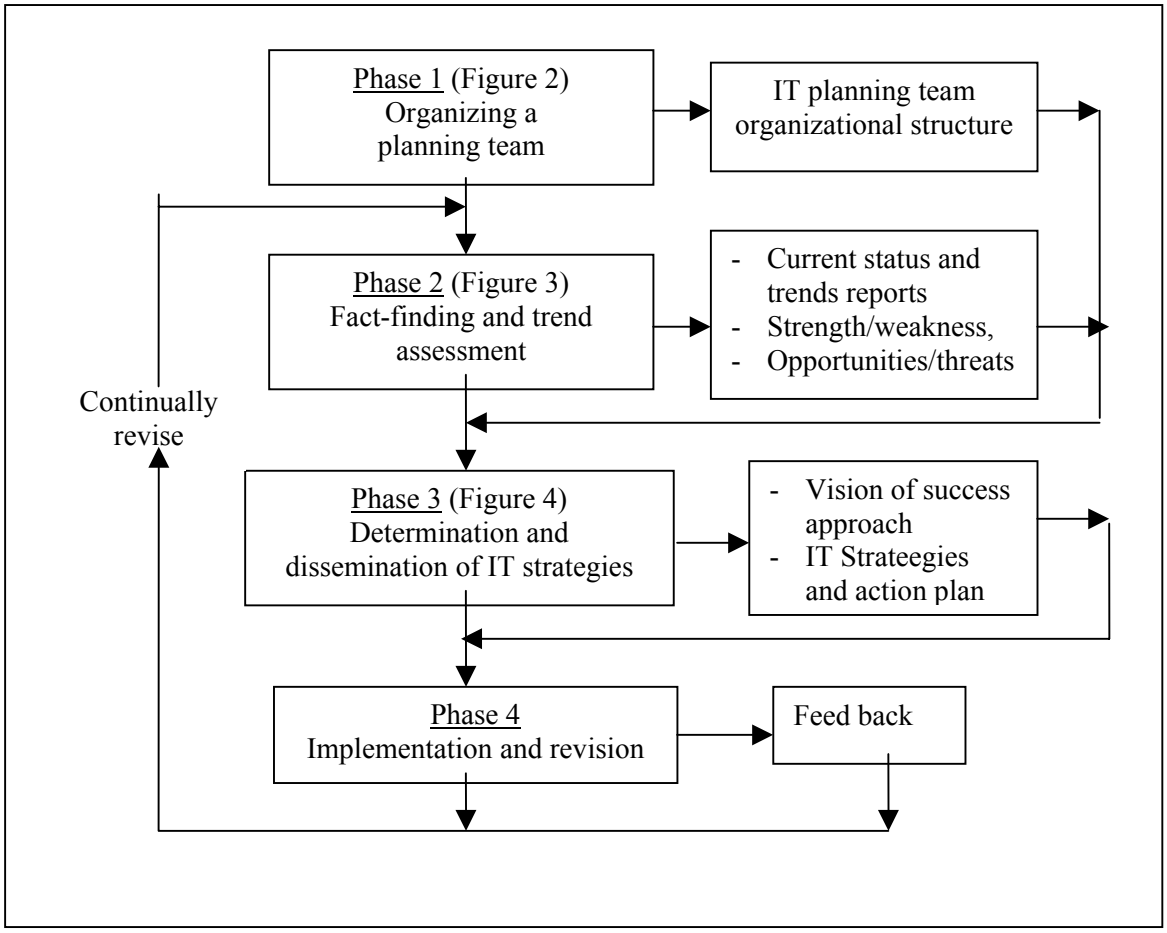


Figure 1. Illustrates A Proposed IT Strategic Planning Process for Institutions of Higher Education in Thailand

Next, an IT strategic planning committee should be established as an IT task group responsible for planning. Based on the study of the preferred IT strategic planning process, the committee should be responsible for IT planning instead of a computing center, as is presently done. This may be because most IT personnel in Thai institutions are not formally trained in managing of IT strategic planning. The proposed committee would include many perspectives and areas of expertise in higher education. Based on the literature [10], it was found that IT functions in higher education could be classified into three major areas: (a) administrative; (b) academic; and (c) infrastructure and access areas. The infrastructure and access area was considered as resource sharing and was thought to be most appropriately developed under the administrative area function since in Thai higher education, the administrative IT function was developed and considered as a core part of the IT systems. Therefore, as shown in Figure 2, this IT task group should have two subgroups: (a) an IT administrative planning group and (b) an IT academic planning group.

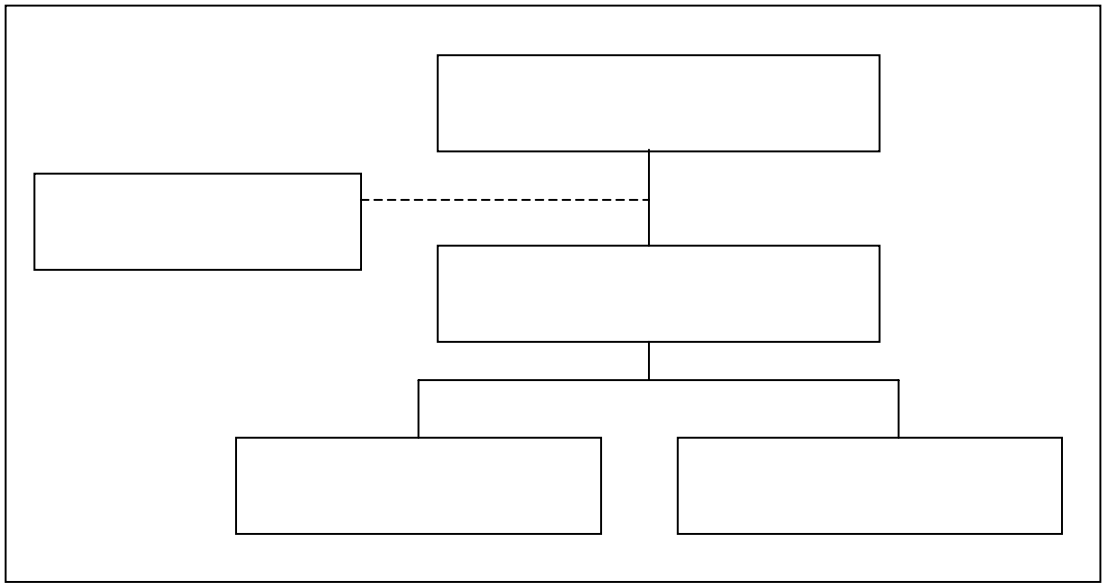


Figure 2. Illustrates IT Planning Team Organizational Structure

Research findings support the preference for an IT strategic planning committee composed of the director of the computing or IT-related center or CIO[1], vice-president for development and planning, dean of the engineering or IT school, vice-president for academic affairs, vice-president for administration, IT staff, external consultants, and the president. Although some institutions include the president in the IT task group, this researcher recommends that the president should not be included in the IT task group since the president is usually chairman of a university-wide IT advisory committee responsible for reviewing and approving an IT strategic plan developed by this IT strategic planning committee. For the same reason, the director of department of academic affairs, director of department of administration, and director of department of development and planning are recommended to replace those vice-presidents in the IT strategic planning committee.

According to the literature [3][9], the IT administrative planning group should be responsible for analyzing the current Administrative Information Systems (AIS). The task of this group should be to develop planning criteria, schedules, user requirements, an information system plan, and an implementation process. High technology is complex and contains many hidden risks. The scope of administrative and user needs, a commitment of time, an inventory of current management processes and systems, realistic budget limits, and a multitude of vendors with differing technologies are important factors in planning for administrative functions.

Secondly, the IT academic planning group should be responsible for analyzing the research and instructional technology of the plan. According to the literature [6], there are three types of computer activities in academic functions: (a) the study of computer science; (b) the use of the computer as a tool for research or problem solving; and (c) Computer-Aided Instruction (CAI). Planning should recognize the implementation of instructional technology required in the development of resource materials and the modification of curricula to incorporate new materials.

Since research findings indicated that presently there was a lack of an effective committee operations in Thai institutions, it is essential that members of these two working groups should be carefully selected. According to the literature [5], in choosing committee members for each group, it is important to identify individuals with: (a) a past history of willingness to invest time and interest in educational administrative matters for the administrative group and in research and instructions for the academic group; (b) a leadership position in the university; (c) a background knowledge of IT; and (d) the influence of a key stakeholder. A secretary should be appointed to serve on each committee. Committee members of these two working groups should be encouraged to inspect the campus as well as other campuses, compare existing technologies, and envision future trends.

Phase 2: Fact-Finding and Trend Assessment.

The purpose of this phase is to evaluate the current status of IT resources, trends and forecasts of demands, and the environment that affects IT resources. This phase was not considered by these surveyed as an important component of the actual process, but in the preferred process and in the literature [4], it is considered to be a critical research activity of the planning process. Therefore, the proposed process should include this component, starting with data collection from constituents—administrators, faculty members, staff members, students—all end users of the administrative information systems and instructional technology.

[1] CIO (Chief Information Officer): this position should rank at the vice-president level and report to the Chief Executive Office (president) or to the provost. It's a position that manages administrative computing, academic computing, networking, and telecommunications.

their relationships in Phase 2. The first component—back-end—is the background component that affects the IT systems. This back-end component consists of three major parts.

(1) The existing IT systems that describe the current status of hardware, software, infrastructure and network, users' achievement, and other existing situations of IT systems.

(2) An institutional plan that delineates expected constraints on IT resources and specifies the major decisions and policy as to the direction of institutional technological development.

(3) A university and IT environmental assessment including:

(a) PEST (Political, Economical, Social, and Technological including new and emerging technology);

(b) people (characteristics, attitudes, abilities, and capabilities of people);

(c) organizational structure; and

(d) resources (accounting and financial management, capital asset management, and human resource management). Based on the study, this component was not identified as a critical part in the existing IT strategic planning in Thai higher educational institutions. However, after analyzing respondents' perception of preferred planning components, as well as the strategic planning literature, it was agreed (and is highly recommended by the researcher) that the university and IT environmental assessment must be included as a basic component in IT strategic planning because the environmental changes would easily affect the IT systems utilization.

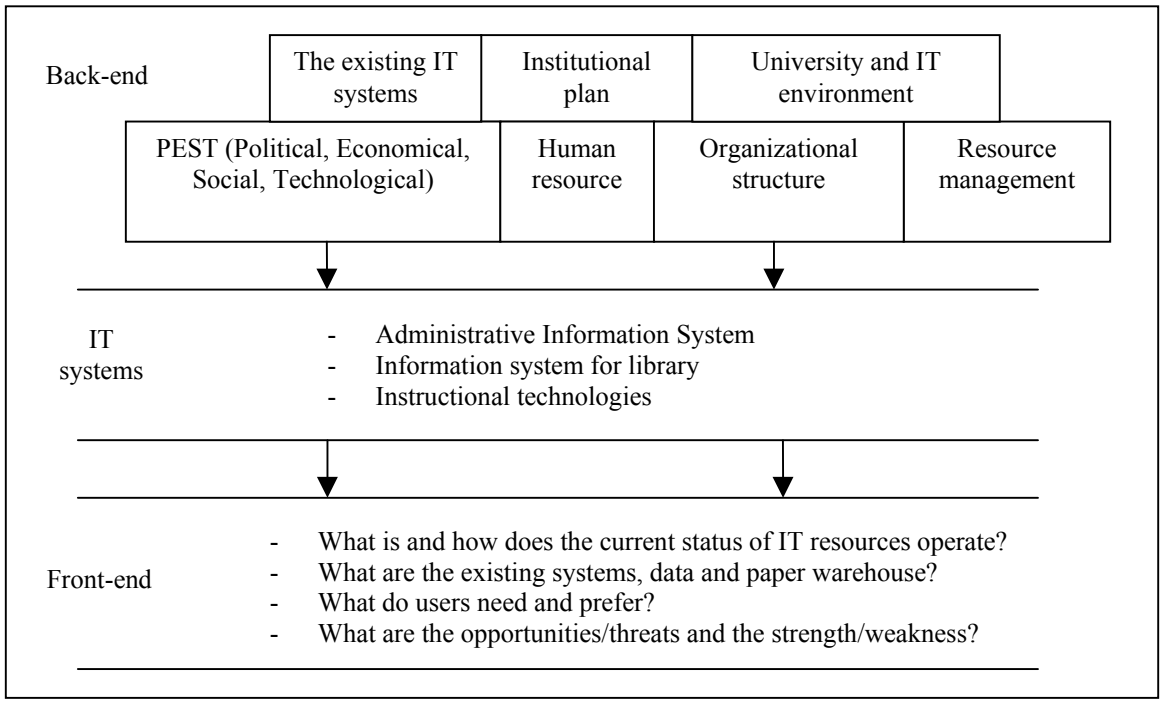


Figure 3. Illustrates Fact-Finding and Trend Assessment (Phase 2)

The back-end component is treated as input because its data affect the current status and trends of IT system implementation on campus.

With regard to the second component, the IT systems include an administrative information system, instructional technology, and the information system used by the library. The IT systems involve input from many parts of the back-end component to produce the output from this process that serve as the front-end component to issues that the university is facing or is expected to face.

Data obtained from this phase would be used to determine the strengths/weaknesses and opportunities/threats, describe what and how the current status of IT resources are operating, determine user needs, and decide how IT can be applied to fulfill these needs. According to the literature [1][4], these data would be obtained through surveys, observations, existing records, and conversations with various individuals at various locations. For example, the IT strategic planning committee might visit all schools and institutes, the dean may extract data from faculty records or student records, as well as conduct formal surveys. The researcher recommends that surveys and structured conversations with selected individuals be conducted because Thai higher education has few records available and there is little sentiment in favor of allowing evaluations by on site observation.

Phase 3: Determination and Dissemination of IT Strategies

As shown in Figure 4, the purpose of the third phase is to build IT strategies. Even though this phase was not considered as a strong component in the actual processes in Thai universities, it was recommended as a preferred process and in the literature [1][6] it is seen as an important component.

Based on the preferred outcome of the study, a vision statement should be developed to express what the institution

desires and sees as the future of technology in the institution and community. Then a mission statement should be prepared to describe the purposes and plans to fulfill the vision of IT in the institution. In addition, goals and objectives should be determined stating what the institution plans to accomplish to achieve the goals.

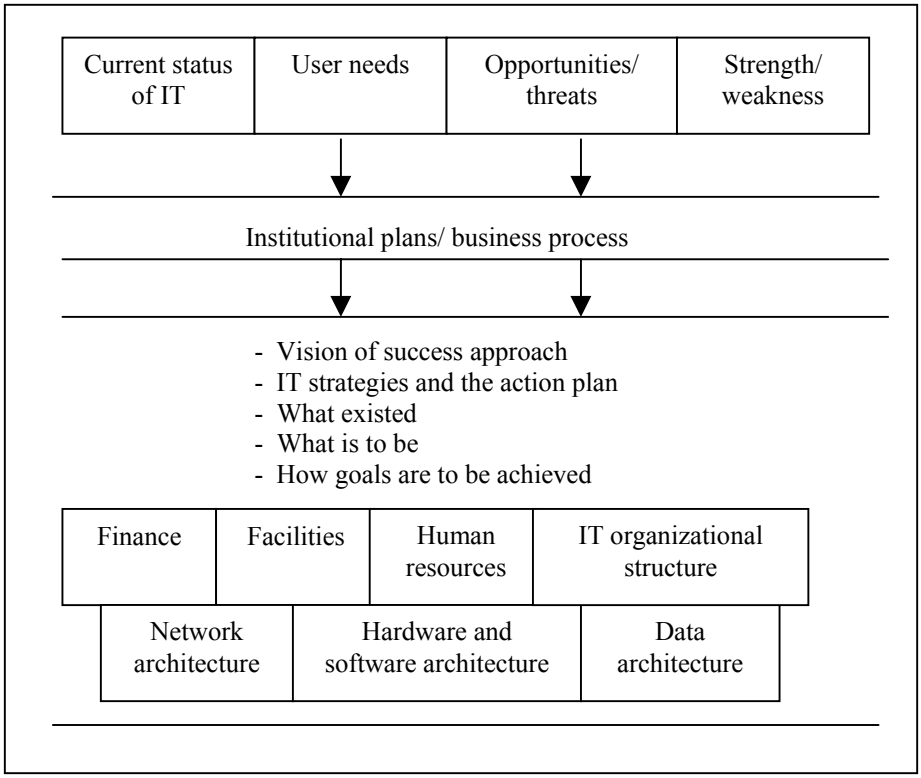


Figure 4. Illustrates Determination and Dissemination of IT Strategies (Phase 3)

According to the preferred process and the literature [1][6], after fact finding and trend assessment, meetings with key people—vice-presidents, deans, directors, and others—should be held to allow people to share their tentative plans and negotiate different points of view and different requirements in order to build a framework of IT strategic objectives. These objectives should then be prioritized by the representatives from all constituencies—faculty, students, staff, library, and the IT organization as well as key financial, academic, and administrative officers—according to costs and benefits. These final strategic objectives should then be disseminated to the university community to ensure that there is sufficient awareness of the strategic benefits to users. Finally, the strategic objectives should be translated into operational strategies for implementation, as shown in Table 1.

Table 1. Shows An Operational Strategies for Implementation

Strategy Areas	Considerations/Tasks
IT Architecture	
Hardware Strategies	- What administrative information systems will be appropriate for use by administrators, faculty, staffs, and students? - What instruction/curriculum will be appropriate for use by faculty and students? - What will be the minimum specifications of functions and capacities for hardware? - What hardware is available in Thailand? - Are companies willing to donate hardware to the institutions? - What maintenance of hardware is provided by vendors?
Software Strategies	- Are other institutions using these software products? - Should the institutions develop their own Administrative information system/instructional technology or buy them? - Is the software user-friendly?

Utilisations of an Embedded System in a Physics Laboratory

Wattanapong Kurdthongmee
Lecturer, Division of Physics, Institute of Science
Walailak University, Tha-sa-la, Nakorn-si-thammarat 80160 Thailand

ABSTRACT - **Majority of laboratory-designed instruments in an experimental Physics still rely on utilizations of standard electronic devices. The instruments absolutely perform with respect to the requirements of the design. They are, however, difficult to re-implement, replicate, maintain, and repair. In the last few years, many modern electronics devices have been commercialized with the benefits of smaller size, low power, low cost/performance ratio, and user-programmability. These devices are microcontrollers, CPLDs (Complex Programmable Logic Devices), FPGAs (Field Programmable Gate Arrays), and digital signal processors (DSPs). Unfortunately, the exploitation of these devices in Physics laboratory is limited. As part of our research, we have improved and/or redesigned the new laboratory instruments used both in teaching and research purposes to employ these modern electronic devices. In this paper we present the two instruments resulted from the utilizations of a microcontroller. These are an embedded system to automatically sample an experimental data and an embedded system to interface between a numerically displayed scientific instrument and a computer.**

Keywords - microcontroller-based system, scientific instrument

บทคัดย่อ - **โดยทั่วไปแล้วเครื่องมือทางวิทยาศาสตร์สำหรับใช้ในงานการทดลองทางฟิสิกส์ที่ถูกออกแบบขึ้นเองในห้องปฏิบัติการยังคงใช้อุปกรณ์อิเล็กทรอนิกส์มาตรฐาน เครื่องมือเหล่านั้นสามารถทำงานได้ตามความต้องการของการออกแบบ แต่การจัดสร้างเครื่องมือเพิ่มเติมตลอดจนการซ่อมแซมนั้นสามารถทำได้ค่อนข้างยาก ในช่วงระยะเวลาไม่นานมานี้อุปกรณ์อิเล็กทรอนิกส์ชนิดใหม่ๆ ได้รับการวางตลาดด้วยจุดเด่นคือ มีขนาดที่เล็กลง ต้องการกำลังไฟฟ้าต่ำ ราคาถูกในขณะที่มีความสามารถในการทำงานสูง และยังมีความสามารถในการโปรแกรมการทำงานได้เองโดยนักออกแบบระบบ อุปกรณ์เหล่านี้ได้แก่ ไมโครคอนโทรลเลอร์ ชิพวงจรรวมชนิด CPLD และ FPGA และชิพประมวลผลสัญญาณดิจิตอล อย่างไรก็ตามอุปกรณ์เหล่านี้แทบไม่ได้รับการนำมาประยุกต์ใช้ในห้องปฏิบัติการทางฟิสิกส์เลย ส่วนหนึ่งของงานวิจัยที่ผู้เขียนได้กระทำอยู่คือ การปรับปรุงประสิทธิภาพการทำงานและการออกแบบเครื่องมือทางวิทยาศาสตร์เพื่อใช้ในงานการทดลองทางฟิสิกส์ทั้งในการเรียนการสอนและงานวิจัย โดยมุ่งเน้นการนำเอาอุปกรณ์อิเล็กทรอนิกส์สมัยใหม่มาใช้งาน ในบทความนี้ ผู้เขียนจะได้นำเสนอผลงานส่วนหนึ่งที่นำเอาไมโครคอนโทรลเลอร์มาใช้ในการควบคุมการทำงานของเครื่องสุ่มอ่านข้อมูลทางวิทยาศาสตร์ และการใช้ไมโครคอนโทรลเลอร์ในการสร้างส่วนเชื่อมต่อระหว่างเครื่องมือวิทยาศาสตร์ที่แสดงผลเชิงตัวเลขกับคอมพิวเตอร์**

คำสำคัญ - ระบบควบคุมโดยไมโครคอนโทรลเลอร์, เครื่องมือวิทยาศาสตร์

1. Introduction

Most of scientific laboratories, especially in experimental Physics which the author involves directly, rely heavily on utilizations of electronic instruments. The examples of such instruments are a data logger, a process controller, a radiation counter, a programmable logic controller (PLC), a thin-film thickness monitoring instrument, etc. While commercial scientific instruments have been constantly developed to keep up in pace with modern electronic devices, laboratory-designed instruments, in contrast, still rely on utilizations of standard electronic devices; i.e. standard TTL 74XX/74LSXX/74HCXX/74HCTXX or CMOS 4XXX families of digital devices and standard family of linear ICs. The instruments perform absolutely properly with respect to the requirement of researchers/experimentors. They are, however, difficult to re-implement, replicate, maintain, repair and are very high power consumption. This arises from the fact that as the instruments are getting more and more complex as a result of the demand for higher functionality of the instruments, the circuit complexity also increases.

In the last few years, many modern electronics devices have been commercialized with the benefits of smaller size, low power comsumption, low cost/performance ratio, and user-programmability. These devices are high performance microcontrollers, CPLDs (Complex Programmable Logic Devices), FPGAs (Field Programmable Gate Arrays), and digital signal processors (DSPs). Unfortunately, the exploitation of these devices in laboratories, both teaching and research, is limited. As part of our research, we have improved and/or redesigned the laboratory instruments used both in teaching and research purposes to exploit the benefits

of these modern electronic devices. In this paper we present two instruments which are the result from the utilizations of a microcontroller.

The organization of this paper is as follow. In section 2, we present a detail of an embedded system to automatically sample an experimental data. An embedded system designed to open a connectivity between a numerically displayed scientific instrument and a computer, another utilization, i presented in section 3. Finally, section 4 concludes this paper and gives the future plan of our work. It is noted that in section 2 and 3, we focus more on the underlying ideas and algorithms to be used rather than specific applications. The given examples is only for demonstrating the proposed ideas. We hope that the readers may exploit these ideas in their own applications.

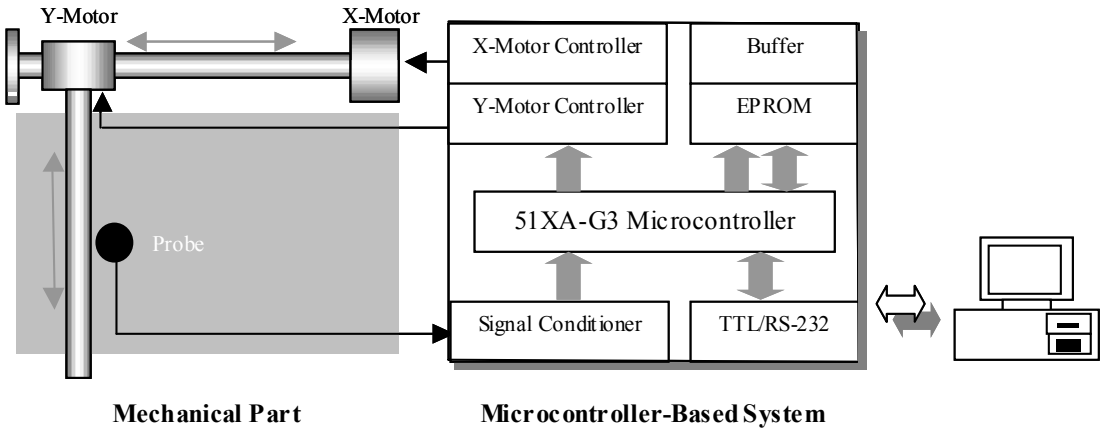


Figure 1. Illustrates the diagram of the microcontroller-based automatic sampling system

2. A Microcontroller–Based Sampling System

The motivation behind the design and implementation of our microcontroller-based sampling system came from the first year introductory physic laboratory: the experiment to study an electric field mapping [3, 4]. According to the experimental procedures [8], students have to prove the existence of electric fields in a medium by sampling values of electric potential from uniform grids with equal separations. These are then used to construct equipotential lines by connecting points with approximately equal potential values. These experimental procedures take time and only one configuration of electrodes can be performed by our students in a three-hour laboratory session.

Luckily, this limitation can be resolved by utilisation of a microcontroller-based system to control a mechanical part that is used to automatically sample potentials from a uniform grid within a medium. A fairly simple implementation of an automatic sampling system, with the capability to change grid separations was designed and constructed with an easy adaptation to another application in mind. In this case, the system, together with a moderate performance computer and a computer program, is successfully used to demonstrate experimental procedures to students.

The automatic sampling system consists of four main parts as illustrated in Figure 1: a mechanical part, a microcontroller-based system to control the mechanical parts, a data acquisition subsystem and an input/output subsystem. We avoid describing the details of the latter two parts which are very specific to this experiment and only focus on the mechanical and microcontroller based system. These are detailed in the following subsections.

2.1 The Mechanical Part

Two stepping motors have been used as mechanical parts to precisely move a probe to locations to sample potentials. To change from a motion around the motors' axis to a linear motion along the x and y arms, a synchronous belt is used to connect between the rotating axis of a motor and a freely-rotating end at the opposite side of the arm. One end of the y-axis arm is attached firmly to the synchronous belt of the x-axis arm and is free to move back and forth along this axis. The y-axis arm is supported by an aluminium rod firmly fixed to the base of the system parallel to the x-axis arm in order to ensure that it does not oscillate during motion. The probe is attached to the y-axis arm at a location that allows it to move freely in the y-direction under control of the y-stepping motor. With this type of connection, one step of rotation, 1.8^0., is changed to 1/256 inch equally in both axis along the arms which are the accuracy of the sampling system.

2.2 Hardware and Software to Control Mechanical Parts

To make the stepping motors produce enough torque and rotate as smoothly and accurately as possible, a unipolar driving circuit [5] is utilised in our prototype. With this type

of circuit, one end of each phase of a stepping motor is connected to an open-collector switching power transistor while the other end is connected to the power supply. The power transistor is logically controlled by an output port of the microcontroller. As we use 4-phase stepping motors, we need 4 bits from the output port to control each transistor separately. In order to rotate stepping motors in a half-phase mode, the following 4-bit binary code sequences need to be sent out to the output port: 1001, 1000, 1010, 0010, 0110, 0100, 0101 and 0001.

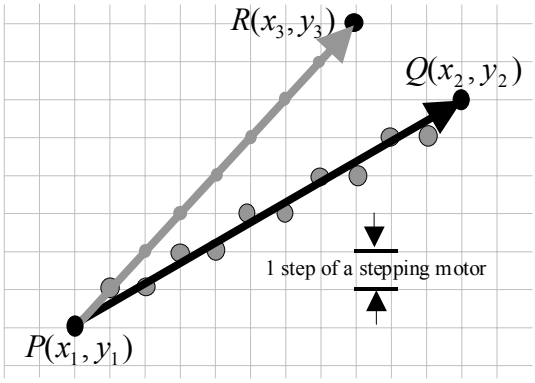


Figure 2. Illustrates Two ideal motion paths from $P(x_1,y_1)$ to $Q(x_2, y_2)$ and from $P(x_1,y_1)$ to $R(x_3, y_3)$

We employ a Brasenham's line algorithm [1] to use in controlling the stepping motors. The algorithm, which is actually a fundamental computer graphics algorithm to scan-convert straight lines on a raster screen, is an efficient method in that it uses only integer addition, subtraction and multiplication by 2 [6]. Therefore, it is fairly easy to code in an assembly language for use in a microcontroller-based control system.

The modified Brasenham algorithm [3, 4], shown in Figure 3, to control the stepping motors of our sampling system works as follows. Assume that we would like to move the probe from $P(x_1,y_1)$ to $Q(x_2, y_2)$ as illustrated in Figure 2. The best approximation to the ideal motion path is described by those coordinates in the discrete domain that fall the least distance from the ideal motion path. Call the distance to those abscissas lying above the ideal path NE_i and the distance to those directly below it E_i. The modified Brasemham's line algorithm identifies the decision variable, $d_i = NE_i - E_i$. When $d_i < 0$, the closest abscissa in the discrete domain will be the abscissa below the ideal motion path. Conversely, when $d_i \geq 0$, the abscissa immediately above the ideal motion path is closest. In figure 2, Two ideal motion paths from $P(x_1,y_1)$ to $Q(x_2, y_2)$ and from $P(x_1,y_1)$ to $R(x_3, y_3)$ are shown by solid lines. The gray dots are the intermediate positions of the probe under a condition of constant increment in the x-axis.
The technique provides an efficient method for moving the probe in directions less than or equal to 45^0.from the x-axis. For directions greater than 45^0 and less than 90^0., the coordinate (x, y) must be interchanged prior to passing it to the procedure and must be interchanged again prior to moving the probe. The result of the interchange will cause just y or both x and y to be incremented at each step.

2.3 The Computer Software

A major role of the computer program to the sampling system is to define a group of coordinates to be sampled, group into a communication protocol and send to the automatic sampling system via a serial communication port.

The program divides the width and height of the sampling region into grids with equal separations along each axis at the required sampling resolution. The coordinate of each grid is coded by combining it with the protocol header and sent out to the sampling system. To avoid unnecessary waiting for incoming data to arrive, which is against the rules of Windows programming, we implemented the serial communication interrupt routine to accept the potential values from the sampling system. By this approach, the computer is free to do another job while data gathering is in progress. The incoming data is also in the form of a communication protocol and needs to be decoded to remove the protocol header from the potential value field. The decoded data is then stored in the two-dimensional array at the appropriate row and column.

2.4 Implementation and Result

All of the approaches and algorithms described in subsection 2.1 and 2.2 were implemented into the 51XA-G3 16-bit microcontroller system as shown in figure 5 by use of the 51XA-specific assembly language. The computer program described in subsection 2.3 were coded in Visual Basic.

The whole system has been successfully used in an introductory physics laboratory to sample potentials from different configurations of electrodes. The number of grid points, which is actually programmable, is set to 600 (30 × 20) with a grid separation of 5 mm. The times taken to sample potential values from all points in all experiments are approximately 5 min. Along with experimental demonstration to students, we also point out the benefit of exploitation of embedded system to physics experiments.

3. An Intelligent Unit to Interface a Computer to a Numerically Displayed Instrument

Most of scientific measurement instruments display data in a numeric form. Some of them, especially obsolete models, do not have a capability to interface to a computer. This factor limits the exploitation of such instruments in an experiment that must be performed automatically under computer-controlled. In this section, we present another approach to open a connectivity between a computer and a numerically displayed instrument. The benefit of this approach is that it can be done without interfering any complex parts inside the instrument. This approach has been successfully employed in order to open a connectivity between a radiation counter, used in nuclear physics laboratory, and a computer.

As our approach relies on a modification of a display unit of the instrument, in the following subsection, therefore, two configurations of this unit of the instrument are briefly

described. Then, the interfacing techniques of both type of display unit are presented.

```
// An array to hold the logical codes, in 4-bit binary format, used to drive motors (send these codes to an output
// port to which the stepping motor connected.
STEP_TABLE : ARRAY(0..7) = (1001, 1000, 1010, 0010, 0110, 0100, 0101, 0001)
// Indicate which member of the STEP_TABLE is currently used to drive the stepping motor
CURRENT_STEP_X, CURRENT_STEP_Y : INTEGER
// Hold the current location (coordinate) of the probe with respect to the sample space
Xcurrent, Ycurrent : INTEGER
PROCEDURE MOVETO_XY(Xend, Yend : INTEGER)
BEGIN
        CURRENT_STEP_X := 0
        CURRENT_STEP_Y := 0
        dX := Xend – Xcurrent
        dY := Yend – Ycurrent
        incE := 2 * dY
        incNE := 2 * (dY - dX)
        d := incE – dX
        X := Xcurrent
        // X is a main axis of motion, loop until the probe has already moved to the end point at Xend
        WHILE X < Xend DO
                // No need to step in the Y direction
                IF d <= 0 THEN
                        d := d + incrE
                // Need to step in the Y direction
                ELSE
                        d := d + incrNE
                        // STEPPING_MOTOR_Y is an address of the output port that controls the Y motor.
                        OUTPORT(STEPPING_MOTOR_Y, STEP_TABLE[CURRENT_STEP_Y])
                        // Update a member in the STEP_TABLE currently used to drive the Y motor.
                        CURRENT_STEP_Y := CURRENT_STEP_Y + 1
                        IF CURRENT_STEP_Y > 7 THEN
                                CURRENT_STEP_Y := 0
                        ENDIF
                ENDIF
                // STEPPING_MOTOR_Y is an address of the output port that controls the X motor
                OUTPORT(STEPPING_MOTOR_X, STEP_TABLE[CURRENT_STEP_X])
                // Update a member in the STEP_TABLE currently used to drive the X motor.
                CURRENT_STEP_X := CURRENT_STEP_X + 1
                IF CURRENT_STEP_X > 7 THEN
                        CURRENT_STEP_X := 0
                ENDIF
                X := X + 1
        ENDWHILE
        // Update the current location of the probe
        Xcurrent := Xend
        Ycurrent := Yend
END
```

Figure 3. Illustrates the modified Brasenham's algorithm to control the stepping motors in our sampling system

3.1 The Configurations of the Display Unit

The electronic circuits of nearly all laboratory instruments are complicate and need to be calibrated appropriately by the manufacturers. Moreover, the details for these circuits vary from one instrument to another. Even two instruments with similar functionality may have completely different electronic circuits. We have found that only the display unit of the instruments is unique and most suitable to be modified and extended in order to interface it with a computer. Broadly, the display unit can be classified into 2 categories: LED (light emitting diode) multiplexing digital display and direct-drive LED display [2].

In LED multiplexing configuration (see the shaded area in figure 4), the same segment pins from all digits of the LED are tied together and connected to *the segment bus* of a microcontroller/microprocessor or special function chips; i.e. 74C911, 74C912 or 74C926. The common pin of each LED is controlled digit by digit separately by a switching transistor. To display a numeric value at any digits of the LED, all switching transistors are inactivated first. This causes every digit of the LEDs to display nothing. The segment data to be shown is then sent out through the segment bus and the switching transistor corresponding to the active digit are activated causing the numeric value to appear. Repeating the same procedures sequentially from the first to the last digit of the display unit at approximately 20 Hz of refresh rate causes all digits to display their numeric values continuously.

The display unit employing the direct-drive technique, on the other hand, uses as many BCD (Binary Coded Decimal) counters and BCD to 7-segment decoders as the required number of digits. Refer to the shaded area in figure 5, the output from a BCD counter is converted to a numeric form by a BCD to 7-segment decoder. The output from the decoder is, in turn, used to drive the corresponding digit of the LED directly. By this technique, every digit of the LED displays it's own numeric value continuously.

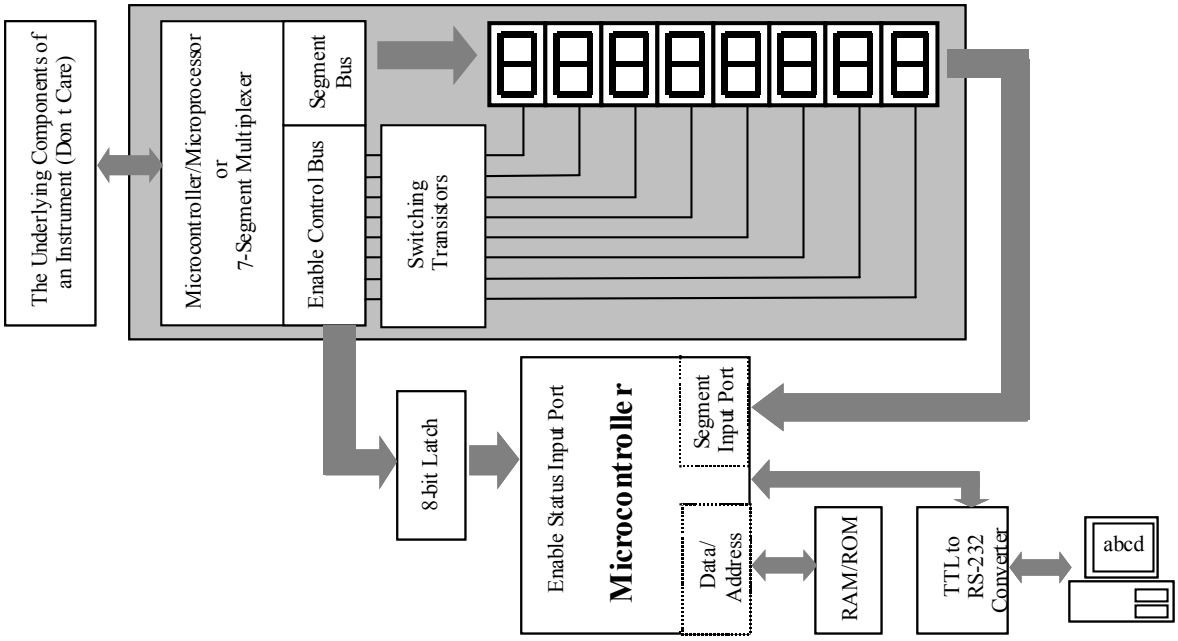


Figure 4. Illustrates The schematic of the instrument utilizing the LED multiplexing configuration and the modification.

3.2 The Modification Approaches

The most suitable electronic device to interface between a numerically displayed instrument, in both configurations of display unit, and a computer is a microcontroller. Employing this device makes the appearance of the interfacing circuit less complex as many components can be substituted by software. In addition, a microcontroller extends the functionalities of the system at no extra cost since many functionalities can be added at a software level. In this section, we present the details of the microcontroller based interfacing unit between the display component of an instrument and a computer in both configurations.

To automatically read data shown on the display unit of the instrument utilizing the LED multiplexing technique, the active digit at any instance of time must be known first. By examining the status of the enable signals, it is possible to point out to which digit the data at the segment bus is currently sent (only one digit of the LEDs is active at any instance of time). Then, the numeric data of the active digit is read from the segment bus and stored appropriately. Refer to the schematic in Figure 4, the enable signals from the enable control bus are stored in the 8-bit latch before connecting to the microcontroller. In order to avoid unnecessary hooking the microcontroller only to check, most of the time, unchanged state of the 8-bit latch, the microcontroller's timer interrupt technique is employed. The interruption causes the microcontroller to periodically identify the active digit by sampling the logical state from the enable control bus. During the interrupt process, the microcontroller samples data from the segment bus via the segment input port. This data, in 7-segment format, is converted to the corresponding BCD format and store in a suitable memory location.

The pseudo-code of the interrupt service routine described earlier is as follow.

```
DigitData : ARRAY [0..7] OF BYTE
// Determine the active digit, read the segment data and
// store in an appropriate location in the array DigitData
TIMER INTERRUPT readActiveDigit
    // Read the current active digit from the 8-bit latch.
    ActiveDigit := INPORT(8_BIT_LATCH)
    // Use look-up table (LUT) to change to BCD format.
    DigitIndex := BINARY_TO_BCD(ActiveDigit)
    // Read seven-segment code from segment input port.
    DigitDataSegmentCode :=
```

```
    INPORT(SEGMENT_INPUT_PORT)
  // Change from seven-segment code to the corresponding
  // BCD and store in the array.
  DigitData[DigitIndex] :=
    SEGMENT_TO_BCD(DigitDataSegmentCode)
END INTERRUPT
```

Although, a radiation counter with the direct-drive LED display component appears to have fairly complicate circuit, the microcontroller interfacing system as shown in Figure 5 is easy to understand and implement. From the figure, the BCD signals from every digit of the BCD counters are input to the 8-to-1 multiplexer. There are four multiplexers in the schematic, each of them accepts only one bit of the BCD signals; D (2^3), C (2^2), B (2^1) and A (2^0). Each bit must be arranged in order from digit 1 to 8. The output signals from the multiplexers are, then, connected to the BCD input port of the microcontroller. The digit to be read into the microcontroller's BCD input port is controlled by the selection signals.

In this configuration, the procedure to read data is trivial and can be implemented without utilizing any interrupt service routine. The pseudo-code of this procedure is shown below:

```
DigitData : ARRAY [0..7] OF BYTE

PROCEDURE ReadBCDData
  // All digits of display
  FOR digit := 0 TO 7
    // Send out digit selection signal to the selection
    // signal port of the microcontroller.
    OUTPORT(SELECTION_SIGNAL_PORT, digit)
    // Read BCD of selected digit.
    dataIn := INPORT(BCD_INPUT_PORT)
    // Store data in an array
    DigitData[digit] := dataIn
  END FOR
END PROCEDURE
```

Apart from acquiring data from the display unit and storing it in the array, the additional roles of the microcontroller in both configurations is to transmit these data to a computer via an appropriate communication channel; e.g. RS-232. Furthermore, the microcontroller can be programmed to partially/fully control the instrument. For example, the microcontroller in our sample implementation, a radiation counter interfacing system, has additional tasks to:

- release a command to the radiation counter in order to start gathering data at the specified, sampling time,
- release a command to stop gathering data and possibly reset the radiation counter,
- store data in its buffer if it detects that the computer is fail to receive data.

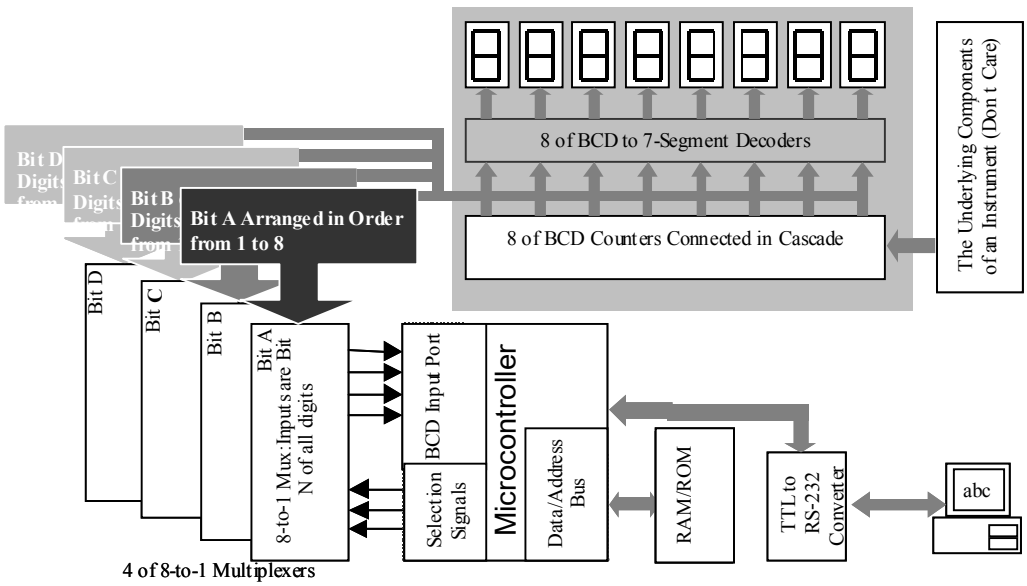


Figure 5. Illustrates the schematic of the instrument utilizing the direct-drive configuration and the modification.

3.3 Implementation and Result

We have implemented the proposed interfacing techniques in an embedded system under control of the MCS-51 family of 8-bit microcontrollers. The system was designed to interface to the ST-350 radiation counter [5] in an experiment to study a decay time of radio active materials. Apart from being able to read data from the radiation counter, we added features to make the system possible to directly and automatically control the radiation counter; i.e. start/stop collecting data at a specified time and period, without any extra hardware. The system is used in conjunction with a moderate performance notebook. In the experiment, we programme the system to gather data once a day. As the system has its own memory, data are, therefore, mostly stored in the memory and transferred to the computer only once a week. It is, however, definitely obvious that employing a microcontroller system and a computer and a few modifications of the instruments, only in the least complex unit, is viable and can ease experiment designs.

4. Conclusion and Future Work

We have presented two examples that exploit the benefits of microcontrollers to Physics experiments. The examples are: the automatic sampling system and the intelligent unit to interface between a computer and a numerically displayed instrument. Obviously, the application of microcontrollers to these experiments not only eases the design process of the instrument and it also increases the functionalilty of the instrument. This comes from the fact that many complex parts of the instrument can mostly be substituted by software instead of hardware implementation which is a basic building block in nearly all traditional laboratory-designed instruments. It might be argued that the learning curve of modern electronic devices is very steep as most of them rely on using a computer program to design and simulate. This is in constrast to the traditional hardware-based approach to use only standard TTL/CMOS/OP-AMP devices. However, we have found that although learning to use the modern devices is fairly difficult at the first place, this difficulty will be paid off in a short term.

In order to further adopt modern electronic devices to Physics experiment, we have planned to employ CPLDs (Complex Programmable Logic Devices), FPGAs (Field Programmable Gate Arrays) and microcontroller with Ethernet connectivity in the new implementation of many laboratory-designed instruments. Currently, we have finished the CPLD implementation of an electronic accelerometer. In the implementation, a dozen of standard TTL devices is substituted by only a medium density CPLD with higher functionality/programmability than the former implementation. The result of this implementation will be presented elsewhere.

References

[1] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics: Principle and Practice*, 2nd Edition, Addison-Wesley, Reading, 1992.

[2] P. Horowitz and W. Hill, *The Art of Electronics*, 2nd Edition, Cambridge University Press, 1989.

[3] W. Kurdthongmee, *Experimental Study of an Electric Field Mapping by Use of a Computer and an Automatic Sampling System*, European Journal of Physics, September 2000, pp. 441-450.

[4] W. Kurdthongmee, *Design and Implementation of an Automatic System to Construct Electric Field Maps by Use of a Microcomputer and Microcontroller*, Songklanakarin Journal of Science and Technology, Vol. 23, No. 1, January-March 2001.

[5] J. B. Peatman, *Design with Microcontrollers*, McGraw-Hill, New York, 1988.

[6] R. A. Plastock and G. Kalley, *Schaum's Outline Series: Theory and Problems of Computer Graphics*, McGraw-Hill, New York.

[7] *ST-350 Radiation Counter Instruction Manual*, Spectrum Techniques, November 1994.

[8] H. D. Young and R. A. Freedman, University Physics, 9th Edition, Addison-Wesley, Reading, 1996.

Wattanapong Kurdthongmee received his Ph.D. degree in Computer Science (Computer Graphics and Solid Modelling) from Brunel University, UK. in 1998. He received his bachelor degree and master degree both in Physics from Prince of Songkla University, Hatyai campus, Thailand in 1987 and 1994, respectively. Currently, Dr. Kurdthongmee is a faculty member in Division of Physics, Institute of Science, Walailak University, Thailand. His research interests include real-time embedded system, utilizations of electronic devices in physic experiment, and computer graphics.

การสร้างโฟโตดีเทคเตอร์แกลเลียมอาเซไนด์ฟอสไฟด์โครงสร้างโลหะ-สารกึ่งตัวนำ-โลหะ

Fabrication of Metal-Semiconductor-Metal GaAsP Photodetectors

อภิชาติ สังข์ทอง นิศาพร เกียรติไพศาลโสภณ และ จิติ หนูแก้ว
ห้องปฏิบัติการวิจัยควอนตัมและสารกึ่งตัวนำทางแสง ภาควิชาฟิสิกส์ประยุกต์
คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ABSTRACT - We present the fabrication and characterization of metal-semiconductor-metal (MSM) GaAsP photodetector grown by Molecular Beam Epitaxy (MBE). Metal-Semiconductor-Metal structure is fabricated using lithography technique. A thermal evaporator is used for growing gold electrodes. The electrode width and gap spacing of this device are in the same value of 100 μm, and the active area is 2100 x 1300 μm^2. The calculated capacitance by conformal mapping technique is approximately 1 pF. At a bias voltage of 8 V, a dark current is in order of nanoampare. This experimental data can be purposed for a potential optical communication devices.

KEYWORDS - MSM Photodetector, GaAsP Semiconductor, and Lithography

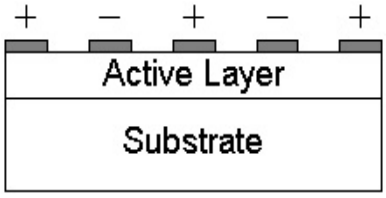
บทคัดย่อ - งานวิจัยนี้เป็นการสร้างโฟโตดีเทคเตอร์โครงสร้างโลหะ-สารกึ่งตัวนำ-โลหะของสารกึ่งตัวนำแกลเลียมอาร์เซไนด์ฟอสไฟด์ที่ได้จากการปลูกผลึกด้วยวิธีจัดเรียงผลึกด้วยลำโมเลกุล (Molecular Beam Epitaxy ; MBE) โครงสร้างโลหะ-สารกึ่งตัวนำ-โลหะสร้างโดยวิธีการลิโทกราฟฟีและสร้างขั้วโลหะทองโดยวิธีการระเหยสารด้วยความร้อนในระบบสุญญากาศ ขนาดความกว้างของขั้วโลหะและช่องว่างระหว่างขั้วเท่ากับ 100 ไมครอน และมีขนาดพื้นที่เท่ากับ 2100 x 1300 ตารางไมครอน ใช้เทคนิคคอนฟอมอลแมบปิงคำนวณค่าความจุได้เท่ากับ 1 พิโกฟารัด จากการวัดกระแสที่แรงดันไบแอส 8 โวลต์ ได้ค่ากระแสมืดอยู่ในระดับนาโนแอมแปร์ จากผลการทดลองที่ได้สามารถนำไปประยุกต์ใช้เป็นอุปกรณ์ด้านการสื่อสารทางแสงได้

1. บทนำ

สารกึ่งตัวนำโครงสร้าง MSM เริ่มมีประดิษฐ์ขึ้นในปี 1979 โดยอาศัยแนวคิดของโครงสร้างที่ที่มีกำแพงศักย์แบบชอตต์กีสองข้าง (two Schottky barrier) [1] เป็นโครงสร้างที่ง่ายต่อการประดิษฐ์และสามารถลดค่าเวลาเคลื่อนย้ายพาหะ (carrier transit time) และ ค่าคงที่ RC (RC constant) ซึ่งเป็นพารามิเตอร์ที่กำหนดความสามารถของสารกึ่งตัวนำ [2] ปัจจุบันได้รับความสนใจในการนำมาประยุกต์ใช้ในการทำโฟโตดีเทคเตอร์ในงานสื่อสารทางแสง โดยสารกึ่งตัวนำที่นิยมใช้จะเป็นจำพวกสารประกอบกึ่งตัวนำ ได้แก่ แกลเลียมอาร์เซไนด์ซึ่งอยู่ในกลุ่ม III-V สารประกอบกึ่งตัวนำแกลเลียมอาร์เซไนด์มีค่าความคล่องตัว (mobility) สูงสามารถใช้ในงานการสื่อสารความเร็วสูงได้ดี ในปัจจุบันมีการปลูกผลึกที่มีส่วนประกอบของสารกึ่งตัวนำตั้งแต่สองชนิดขึ้นไปอย่างแพร่หลาย เช่น AlGaAs, InGaAs, GaAsP เป็นต้น จุดประสงค์เพื่อเปลี่ยนแปลงคุณสมบัติทางแสงให้เหมาะสมกับการสื่อสารในปัจจุบัน

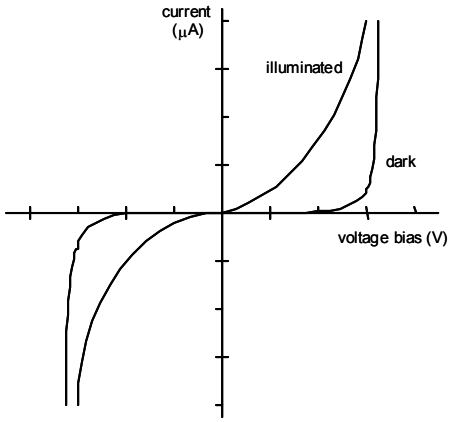
2. ทฤษฎีและหลักการ

ลักษณะทั่วไปของโฟโตดีเทคเตอร์โครงสร้าง MSM แสดงในรูปที่ 1 ประกอบด้วยแผ่นฐานรองที่เป็นสารกึ่งฉนวน (semi-insulating substrate) บนแผ่นรองเป็นชั้นของสารกึ่งตัวนำที่ทำหน้าที่เป็นชั้นแอกทีฟ (active layer) ซึ่งได้จากการปลูกด้วยวิธีต่างๆ เช่น MBE, OMCVD หรือการระเหยสารด้วยความร้อน (thermal evaporator) เป็นต้น ด้านบนสุดเป็นขั้วโลหะที่มีโครงสร้างแบบอินเตอดิจิเตจ (interdigitated electrode) รอยต่อระหว่างขั้วโลหะกับชั้นแอกทีฟเกิดกำแพงศักย์แบบชอตต์กี (schottky barrier) ในลักษณะประกบด้านหลัง (back-to-back) พื้นที่รับแสงจะอยู่ระหว่างช่องว่างของขั้วโลหะ การสร้างขั้วโลหะจะใช้วิธีการ optical lithography ซึ่งสามารถทำขั้วให้มีขนาดเล็กระดับไมครอนได้



รูปที่ 1 แสดงโครงสร้างโฟโตดีเทคเตอร์โครงสร้างโลหะ-สารกึ่งตัวนำ-โลหะ

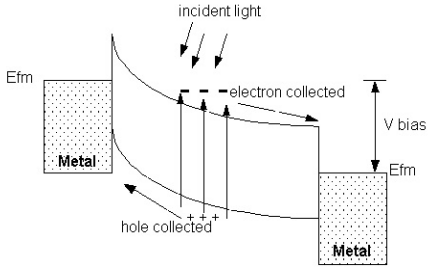
เนื่องจากโฟโตดีเทคเตอร์โครงสร้าง MSM นี้ กำแพงศักย์แบบชอตต์กีประกบกันแบบ back–to-back เมื่อทำการไบแอส ไม่ว่าจะเป็นกรณีไบแอสตรง (forward bias) หรือไบแอสกลับ (reverse bias) จะมีผลทำให้กำแพงศักย์ด้านหนึ่งมีทิศไปข้างหน้า (forward direction) และอีกด้านหนึ่งมีทิศย้อนกลับ (reverse direction) พิจารณาเป็นไดโอด 2 ต่อกลับด้านกันซึ่งได้กราฟคุณสมบัติทาง I-V ดังรูปที่ 2 สังเกตว่าไดโอดตัวหนึ่งจะต่อแบบไบแอสกลับเสมอ



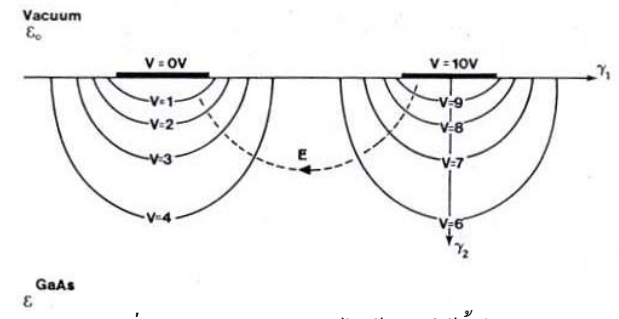
รูปที่ 2 แสดงสมบัติทางกระแส-แรงดันของโฟโตดีเทคเตอร์โครงสร้างโลหะ-สารกึ่งตัวนำ-โลหะ

การศึกษาคุณสมบัติทาง I-V ของดีเทคเตอร์แบบ MSM จะพิจารณาทั้งในส่วนของกระแสมืด (dark current) และในส่วนของกระแสโฟโต (photo current) ในรูปที่ 3 แสดงคุณสมบัติของ MSM ดีเทคเตอร์ภายใต้ค่าความเข้มแสงค่าต่างๆ ในส่วนของกระแสมืดจะขึ้นอยู่กับกลไก 2 อย่าง อย่างแรกคือการสร้างคู่อิเล็กตรอน-โฮลแบบสุ่ม (spontaneous electron-hole pairs) ภายในบริเวณแอกทีฟ (active region) เนื่องจากพลังงานความร้อน (thermal Energy) อีกอย่างหนึ่งเกิดจากพาหะที่สามารถข้ามกำแพงชอตต์กีได้ ไดอะแกรมของแถบพลังงานของดีเทคเตอร์แบบ MSM ภายใต้การไบแอสแสดงในรูปที่ 4 ประกอบด้วยตำแหน่งระดับเฟอร์มิ (Fermi Level; E_f) ของโลหะ แถบวาเลนซ์ (valence band; E_v) และแถบนำ (conduction band; E_c) ของสารกึ่งตัวนำ

เมื่อแสงตกกระทบดีเทคเตอร์ เกิดการดูดกลืนแสงที่ผิวของสารกึ่งตัวนำ เกิดคู่อิเล็กตรอน-โฮลภายในบริเวณแอกทีฟ ถ้าระหว่างขั้ว 2 ขั้ว ถูกไบแอสขั้วหนึ่งเป็นลบ (cathode) อีกขั้วเป็นบวก (anode) คู่อิเล็กตรอน-โฮลจะถูกแยกออกจากกันโดยโฮลจะเคลื่อนที่ไปยังขั้วลบ ส่วนอิเล็กตรอนจะเคลื่อนที่ไปยังขั้วบวก (ภายใต้อิทธิพลของสนามไฟฟ้า) ในรูปที่ 4 แสดงเส้นสนามไฟฟ้า (electric field lines) ภายใต้ขั้วโลหะ เมื่อพาหะที่เกิดจากแสงเคลื่อนที่ไปถึงขั้ว เกิดการไหลของกระแสที่วงจรภายนอก กระแสโฟโตนี้จะเพิ่มตามความเข้มแสงที่ตกกระทบ



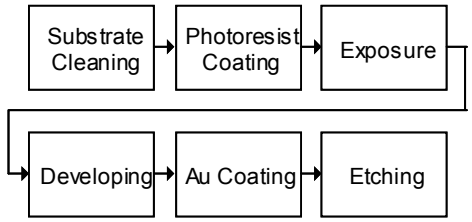
รูปที่ 3 แสดงไดอะแกรมแถบพลังงานของโฟโตดีเทคเตอร์ที่มีกำแพงศักย์แบบชอตต์กี



รูปที่ 4 แสดงการเกิดสนามไฟฟ้าภายใต้ขั้วโลหะ

3. การสร้าง

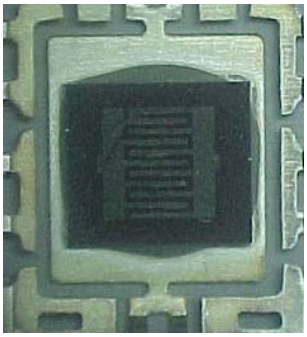
สารกึ่งตัวนำที่ใช้ในการทำโฟโตดีเทคเตอร์เป็นแกลเลียมอาร์เซไนด์ฟอสไฟด์ที่ได้จากการปลูกผลึกบนแผ่นรองรับแกลเลียมอาร์เซไนด์ด้วยระบบ MBE กรรมวิธีการสร้างขั้วโลหะบนสารกึ่งตัวนำมีขั้นตอนดังรูปที่ 5



รูปที่ 5 แสดงกรรมวิธีการสร้างขั้วโลหะ [3]

กรรมวิธีการสร้างขั้วโลหะประกอบด้วยขั้นตอนการทำความสะอาดแผ่นสารกึ่งตัวนำ (substrate cleaning) ซึ่งเป็นขั้นตอนที่สำคัญในการลดจุดบก

พร่องของรอยต่อระหว่างสารกึ่งตัวนำกับขั้วโลหะหลังจากนั้นเข้าสู่กระบวนการสร้างหน้ากากบนแผ่นสารกึ่งตัวนำโดยทำการเคลือบน้ำยาโฟโตรีซีส (photoresist coating) บนแผ่นสารกึ่งตัวนำ หมุนแผ่นสารกึ่งตัวนำที่ความเร็ว 3000 รอบต่อนาทีเพื่อให้โฟโตรีซีสกระจายอย่างสม่ำเสมอ จากนั้นทำการฉายแสง (exposure) ตามลักษณะขั้วที่ออกแบบไว้ หลังจากผ่านขั้นตอนการฉายแสงทำการดึงโฟโตรีซีสส่วนที่ไม่ต้องการออก (developing) ซึ่งจะได้หน้ากากรูปขั้วโลหะตามที่ออกแบบไว้ หลังจากนั้นทำการปลูกขั้วโลหะทอง (Au coating) โดยใช้วิธีการระเหยสารด้วยความร้อนในสุญญากาศ ความหนาของขั้วโลหะประมาณ 100 นาโนเมตร การปลูกขั้วโลหะให้บางๆ เพื่อให้แสงที่ตกกระทบสามารถผ่านชั้นของขั้วโลหะได้ จากนั้นกำจัดโฟโตรีซีสส่วนที่เหลือออก (etching) ซึ่งได้โครงสร้างขั้วโลหะบนสารกึ่งตัวนำมีขนาดความกว้างของขั้วและระยะระหว่างขั้วเท่ากับ 100 ไมครอน ขนาดพื้นที่เท่ากับ 2100×1300 ตารางไมครอน ลักษณะขั้วที่ได้แสดงในรูปที่ 6



รูปที่ 6 แสดงลักษณะขั้วโลหะทองบนสารกึ่งตัวนำ

การหาค่าความจุของโฟโตดีเทคเตอร์แบบ MSM ใช้วิธีการคำนวณแบบ 2 มิติ ที่เรียกว่าคอนฟอมอลแมบปิง (conformal mapping) [4,5] เป็นวิธีการประมาณค่าความจุที่เกิดจากเขตปลอดพาหะโดยพิจารณาใน 2 มิติ ซึ่งสามารถคำนวณได้จาก

$$C = \frac{K(k)}{K(k')}\varepsilon_0(1+\varepsilon_r)\frac{A}{\text{finger}\quad\text{period}}$$

เมื่อ A เป็นพื้นที่รับแสง

ε_o เป็นสภาพยอมในสุญญากาศ ($8.854\text{x}10^{-34}$ F/m)

ε_r เป็นค่าคงที่ไดอิเลิกตริกสัมพัทธ์ (ประมาณ 12 สำหรับแกลเลียมอาร์เซไนด์) ระยะคาบของขั้ว (finger period) คือระยะที่มีขั้วและไม่มีขั้วในหนึ่งคู่

และ K(k) และ K(k′) เป็นฟังก์ชันการอินทิกรัลเชิงวงรี (complete elliptic integral of the first kind) ซึ่งหาได้จาก

$$K(k) \equiv \int_0^{\pi/2} \frac{d\phi}{\sqrt{(1-k^2\sin^2\phi)}}$$

เมื่อ

$$k \equiv \tan^2\frac{\pi}{4}\frac{L}{W+L}$$

และ

$$K(k') \equiv \int_0^{\pi/2} \frac{d\phi}{\sqrt{(1-k'^2\sin^2\phi)}}$$

เมื่อ

$$k' \equiv \sqrt{(1-k^2)}$$

โดย W และ L เป็นความกว้างของขั้วโลหะและระยะระหว่างขั้วโลหะ ตามลำดับ

จากโครงสร้างที่ออกแบบไว้สามารถคำนวณค่าความจุได้ประมาณ 1 พิโกฟารัด

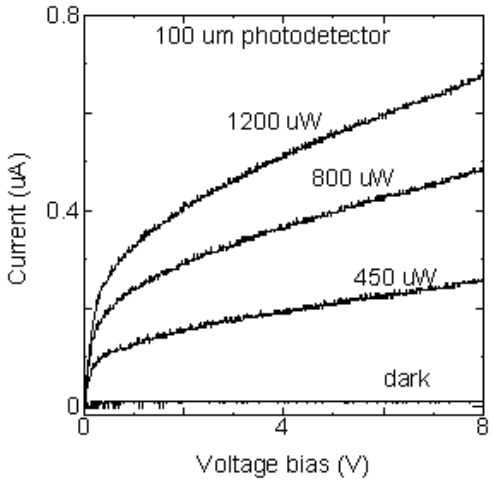
4. การทดลองและการอธิบาย

ทำการศึกษาสมบัติทางแสงของโตดีเทคเตอร์โดยจัดระบบดังรูปที่ 7 แสงความยาวคลื่น 632 นาโนเมตรจากฮีเลียมนีออนเลเซอร์ส่งผ่านอุปกรณ์ปรับความเข้มแสงก่อนเข้าโฟโตดีเทคเตอร์ ลอกอินแอมปลิไฟเออร์ใช้ในการวัดค่ากระแสโฟโตและควบคุมแรงดันไบแอสผ่านโปรแกรมคอมพิวเตอร์

รูปที่ 7 แสดงระบบการวัดสมบัติกระแส-แรงดัน

จากการทดลองได้กราฟในรูปที่ 8 แสดงให้เห็นว่ากระแสมืด (dark current) อยู่ในระดับที่ต่ำมากเนื่องจากกำแพงศักย์ของซอตต์กี และกระแสเพิ่มมากขึ้นตามค่าความเข้มแสงที่ตกกระทบ พิจารณาในแต่ละความเข้มแสงจะพบว่ากระแสเพิ่มขึ้นตามแรงดันไบแอส ซึ่งสารถอธิบายได้ว่าพาหะที่เกิดจาก

แสงในชั้นแอกทีฟใช้เวลาในการเคลื่อนที่ไปยังขั้วไม่เท่ากัน กรณีที่แรงดันดันไบแอสน้อยๆ ทำให้เกิดสนามไฟฟ้าไม่แรงพอที่จะทำให้พาหะเคลื่อนที่ด้วยความเร็วอิ่มตัวซึ่งพาหะบางส่วนเกิดการรวมตัวกันก่อนที่จะไปถึงขั้ว แต่เมื่อเพิ่มแรงดันไบแอสมากขึ้นเกิดสนามไฟฟ้ามากขึ้นตามทำให้พาหะเคลื่อนที่ไปถึงขั้วก่อนที่จะมีการรวมตัวกัน นั่นคือทำให้เกิดกระแสมากขึ้นตามค่าแรงดันไบแอสซึ่งสอดคล้องกับทฤษฎี



รูปที่ 8 แสดงกราฟการเปลี่ยนแปลงของกระแสโฟโต ที่ความเข้มแสงค่าต่างๆ

สรุป

จากงานวิจัยนี้แสดงให้เห็นว่าโครงสร้าง MSM เป็นโครงสร้างที่ประดิษฐ์ได้ง่าย ค่าความจุที่ได้มีค่าต่ำซึ่งส่งผลทำให้แบนด์วิดธ์สูงขึ้นเหมาะสำหรับการสื่อสารที่ต้องการความเร็วสูง รวมทั้งกระแสมืดที่ได้มีค่าต่ำ โดยโฟโตดีแทคเตอร์แกลเลียมอาร์เซไนด์ฟอสไฟด์สามารถนำไปใช้ในงานสื่อสารทางแสงในย่านสีแดงจนถึงอินฟาเรดโดยการเปลี่ยนแปลงอัตราส่วนของสารกึ่งตัวนำแต่ละชนิด

กิตติกรรมประกาศ

งานวิจัยนี้ได้รับทุนวิจัยจากเนคเทครวมทั้งได้รับความร่วมมือเกี่ยวกับลิโทกราฟฟีจากห้องปฏิบัติการออปโตอิเล็กทรอนิกส์และศูนย์วิจัยไมโครอิเล็กทรอนิกส์เป็นอย่างดีในด้านเครื่องมือรวมทั้งอาจารย์นักศึกษาของห้องปฏิบัติการวิจัยควอนตัมและสารกึ่งตัวนำทางแสงที่ให้คำแนะนำที่เป็นประโยชน์ต่องานวิจัยนี้ จึงขอขอบคุณทุกฝ่ายไว้ ณ ที่นี้ด้วย

เอกสารอ้างอิง

[1] Paul R. Berger, " MSM photodiodes", IEEE Potential, April / May 1996.

[2] Julian B., D. Soole and Hermann Schumacher, "Transit-Time Limited Frequency Response of InGaAs MSM Photodetectors", IEEE Transactions on Electron Devices, Vol. 37, No. 11, November 1990, pp. 2285-2291.

[3] L.F. Thopson, C.G. Willson, and M.J.Bowden, "Introduction to Microlitho graphy", American Chemical Society, Washington, D.C., 1983, pp. 161-214.

[4] Sybil P.Parker, "McGraw-Hill Encyclo- pedia of Physics", Second Edition, McGraw-Hill, Inc., 1993, pp. 212-214.

[5] Murry R. Spiegel, "Theoretical Mechanics", McGraw-Hill book Company, 1989, pp 106.

อภิชาติ สังข์ทอง สำเร็จการศึกษาปริญญาตรีทางวิศวกรรมไฟฟ้าสื่อสาร สาขาวิศวกรรมโทรคมนาคม และปริญญาโทสาขาฟิสิกส์ประยุกต์ จากสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ได้รับทุนยกเว้นค่าเล่าเรียนตลอดการศึกษาระดับปริญญาตรี ทุนพัฒนาอาจารย์จากสถาบันราชภัฏราชนครินทร์ มีผลงานวิจัยเรื่อง wavelength division multiplexing ตีพิมพ์ในวารสาร IEEE

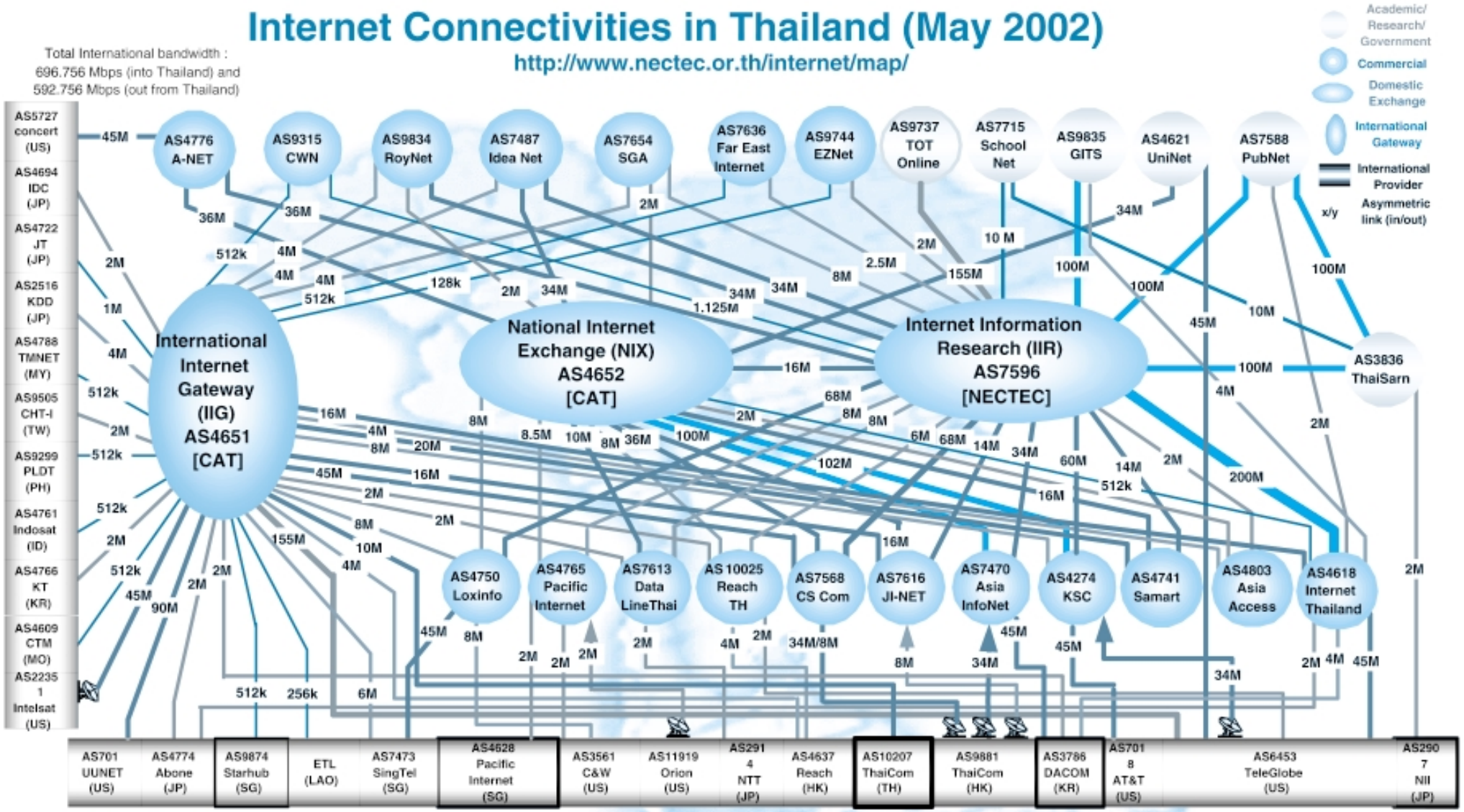
นิศาพร เกียรติไพศาลโสภณ จบการศึกษาปริญญาตรี (การศึกษาวิทยาศาสตร์-ฟิสิกส์) มหาวิทยาลัยศรีนครินทรวิโรฒสงขลา พ.ศ. 2533 จบปริญญาโท (ฟิสิกส์ประยุกต์) สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง พ.ศ. 2536 ทำงานในตำแหน่งนักวิจัย 1 งานวิจัยอิเล็กโทรออปติกส์ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ในช่วง 2537-2544 ปัจจุบันศึกษาปริญญาเอก (ฟิสิกส์ประยุกต์/ optoelec tronics) ณ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

จิตติ หนูแก้ว จบการศึกษาระดับปริญญาตรีสาขาฟิสิกส์ จากมหาวิทยาลัยศรีนครินทรวิโรฒสงขลา พ.ศ. 2526 จบการศึกษาระดับปริญญาโทสาขาฟิสิกส์ จากมหาวิทยาลัยเชียงใหม่ พ.ศ. 2532 และจบปริญญาเอกสาขา Material Sciences and Engineering จากมหาวิทยาลัยนาโกยา พ.ศ. 2541 ปัจจุบันดำรงตำแหน่งอาจารย์ประจำภาควิชาฟิสิกส์ประยุกต์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง หัวข้องานวิจัยที่สนใจได้แก่ Quantum Well Device, Semiconductor Physics และ Optical Instrumentation.

Internet Connectivities in Thailand (May 2002)

http://www.nectec.or.th/internet/map/

Total International bandwidth :
696.756 Mbps (into Thailand) and
592.756 Mbps (out from Thailand)



DISCLAIMER

Chart Date: 2002-05-01

ThaiSam