บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการพัฒนาต้นแบบแชตโรบอทภาษาไทย กรณีศึกษา เรื่องอัญมณีพบว่ามีแนวคิดและทฤษฎีด้านต่าง ๆ ที่ต้องนำมาประกอบการพัฒนาระบบ ดังนี้ระบบแฟ้มข้อมูล ระบบฐานข้อมูล ระบบจัดการฐานข้อมูล และการประยุกต์ใช้งานฐานข้อมูลอินเทอร์เน็ต

2.1. แนวคิดเกี่ยวกับปัญญาประดิษฐ์
2.2. แนวคิดเกี่ยวกับการตัดคำและการค้นหาข้อมูล
2.3. แนวคิดเกี่ยวกับระบบฐานข้อมูลและระบบจัดการฐานข้อมูล
2.4. งานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดเกี่ยวกับปัญญาประดิษฐ์

ปัญญาประดิษฐ์ (Artificial intelligence : AI) หมายถึง การทำให้คอมพิวเตอร์สามารถคิดหาเหตุผลได้ เรียนรู้ได้ ทำงานได้เหมือนสมองมนุษย์ หรือการพัฒนาให้ระบบคอมพิวเตอร์มีลักษณะการทำงานใกล้เคียงกับระบบการประมวลผลและการตอบสนองของมนุษย์ที่มีต่อแต่ละสถานการณ์ เพื่อให้คอมพิวเตอร์สามารถปฏิบัติงานแทนที่มนุษย์ได้อย่างมีประสิทธิภาพ เช่น หุ่นยนต์ หรือ robot เป็นต้น

ปัญญาประดิษฐ์ (Artificial intelligence : AI) คือ ความพยายามในการพัฒนาระบบคอมพิวเตอร์ (ทั้งฮาร์ดแวร์ และ ซอฟต์แวร์) ให้มีพฤติกรรมเลียนแบบมนุษย์ ระบบต่างๆจะต้องมีความสามารถเข้าใจภาษามนุษย์ ทำงานที่ต้องใช้การประสานงาน ระหว่างส่วนต่างๆ (โรโบติก - robotics) ใช้อุปกรณ์ที่สามารถรับทราบ และตอบสนอง ด้วยพฤติกรรม และภาษา (ระบบการมอง และ การออกเสียง) การเลียนแบบความเชี่ยวชาญและการตัดสินใจของมนุษย์ (ระบบผู้เชี่ยวชาญ) ระบบดังกล่าวยังต้องแสดง ความสามารถทางตรรกะ การใช้เหตุผล สัญชาตญาณ และใช้หลักการสมเหตุสมผล (common sense) ที่มีคุณภาพ ในระดับเดียวกับมนุษย์

ระบบปัญญาประดิษฐ์ที่ประสบความสำเร็จ สร้างขึ้นมาบนพื้นฐานของความเชี่ยวชาญ ความรู้ และรูปแบบการใช้เหตุผลบางแบบของมนุษย์ แต่ก็ไม่ได้แสดงความชาญฉลาดของมนุษย์ ระบบปัญญาประดิษฐ์ที่มีอยู่ในปัจจุบัน ไม่สามารถสร้างคำตอบที่แปลกใหม่ หรือคำตอบที่เป็นของระบบๆ เองได้ คือเป็นเพียงการเพิ่มขีดความสามารถให้แก่ผู้เชี่ยวชาญ แต่ไม่ได้สร้างขึ้นมาทดแทน หรือแม้กระทั่งลอกเลียนแบบความสามารถทั้งหมดของผู้เชี่ยวชาญนั้น กล่าวคือ ระบบยังขาด

ความสามารถในการใช้หลักการสมเหตุสมผล และความสามารถในการนำไปใช้งานได้ทั่วไป ที่มีอยู่ในความชาญฉลาด ของมนุษย์

ความชาญฉลาดของมนุษย์มีความซับซ้อน และกว้างขวางกว่าความชาญฉลาดของเครื่องคอมพิวเตอร์ องค์ประกอบที่แยกความแตกต่างของมนุษย์ ออกจากสัตว์ชนิดอื่นๆ คือ ความสามารถในการพัฒนาความเกี่ยวข้อง และ การใช้ตัวแทนความรู้ และการเปรียบเทียบ ซึ่งสามารถนำมาใช้ในการสร้างความรู้ใหม่ๆ กฎเกณฑ์ใหม่ การใช้กฎเกณฑ์เก่ากับสถานการณ์ใหม่ และในบางครั้งก็ใช้สัญชาตญาณในการแก้ปัญหา โดยไม่มีกฎเกณฑ์ใดเข้ามาเกี่ยวข้อง ความชาญฉลาดของมนุษย์ยังรวมไปถึง ความสามารถในการใช้ประสาทสัมผัส ในการรับรู้เหตุการณ์ที่เกิดขึ้นรอบๆ ตัว

ปัจจุบัน นักวิทยาศาสตร์และวิศวกรคอมพิวเตอร์พยายามพัฒนาอุปกรณ์และชุดคำสั่งที่สามารถลอกเลียนความฉลาดของมนุษย์ แต่เป็นที่น่าเสียดายที่ยังไม่มีอุปกรณ์ใดสามารถทำงานเลียนแบบการทำงานของมนุษย์ได้อย่างสมบูรณ์ ดังนั้นเมื่อต้องการพัฒนาระบบความฉลาดที่สามารถทำงานใกล้เคียงกับมนุษย์ในด้านใด เราต้องกำหนดขอบเขต (Domain) และหน้าที่เฉพาะ (specific functions) ที่อุปกรณ์และชุดคำสั่งต้องปฏิบัติเพื่อให้การดำเนินงานของระบบเป็นไปตามความต้องการ ซึ่งเราสามารถยกตัวอย่างสาขา AI ที่มีผู้ศึกษากันอยู่ 5 สาขา ได้แก่

1. การประมวลภาษาธรรมชาติ (natural language processing) ภาษาที่เกี่ยวกับการติดต่อสื่อสารโดยตรงกับคอมพิวเตอร์ เช่น ภาษาอังกฤษ หรือไทย ซึ่งมีปัญหาในเรื่องการวางตำแหน่งคำผิดได้หรือบางครั้งเรียกว่าภาษาธรรมชาติ เป็นการพัฒนาให้ ระบบคอมพิวเตอร์สามารถเข้าใจภาษาที่มนุษย์ใช้ในชีวิตประจำวัน โดยคอมพิวเตอร์อาจจะสามารถอ่าน พูด ฟังและความเข้าใจในแต่ละภาษา ทำให้การทำงานติดต่อสื่อสาร ระหว่างมนุษย์และเครื่องคอมพิวเตอร์ดำเนินไปอย่างสะดวก ทั่วถึง และมีประสิทธิภาพ ประการสำคัญ เทคโนโลยีสารสนเทศง่ายขึ้น และกระจายตัวไปในวงกว้างกว่าปัจจุบัน ประเด็นที่ยกตัวอย่างในการพัฒนาระบบการประมวลผลภาษาธรรมชาติก็คือ จะทำอย่างไรให้ระบบสามารถวิเคราะห์และแก้ปัญหาเกี่ยวกับความกำกวมของภาษาธรรมชาติได้ หรือในบางครั้งสิ่งที่เป็นความหมายเดียวกัน สามารถพูดได้ในหลายลักษณะ จะทำอย่างไรให้ระบบสามารถวิเคราะห์ได้ว่าประโยคเหล่านั้นมีความหมายเดียวกัน

ยืน ภู่วรวรรณ และชัยยงศ์ วงศ์ชัยสุวัฒน์ (2535) ได้ให้คำจำกัดความของระบบภาษาธรรมชาติ ดังนี้ ระบบภาษาธรรมชาติเป็นระบบซอฟท์แวร์ที่กล่าวถึงกันอย่างกว้างขวางเป็นศาสตร์สาขาหนึ่งของปัญญาประดิษฐ์ มีลักษณะสำคัญดังนี้

- ส่วนของระบบอินพุตและเอาท์พุตที่ใช้ในการสั่งหรือติดต่อกับคอมพิวเตอร์จะอยู่ในลักษณะที่เป็นภาษาธรรมชาติ

- การประมวลผลภายในระบบจะใช้หลักการพื้นฐานของฐานความรู้ที่เกี่ยวกับไวยากรณ์ ความหมายและความเข้าใจของภาษาธรรมชาติ

การประมวลผลภาษาธรรมชาติ เป็นการติดต่อสื่อสารระหว่างผู้ใช้กับเครื่องคอมพิวเตอร์ โดยที่ผู้ใช้สามารถนำเข้าเสียงพูด ซึ่งอาจจะเป็นภาษาอังกฤษ หรือภาษาใดก็ได้ตามที่ผู้พัฒนาได้โปรแกรมเอาไว้

2. หุ่นยนต์ (robotics) เกี่ยวกับการออกแบบ การสร้าง และการนำไปใช้งาน ให้เคลื่อนไหวได้คล้ายมนุษย์ โดยเฉพาะใช้งานที่เสี่ยง อันตรายแทนมนุษย์ เป็นสาขาสำคัญของปัญญาประดิษฐ์ที่ได้รับความสนใจจากบุคคลทั่วไป หุ่นยนต์ถูกพัฒนาขึ้น เพื่อจำลองการทำงานของมนุษย์ โดยศึกษากระบวนการในการปฏิบัติงาน แต่ละประเภท แล้วพยายามออกแบบอุปกรณ์ และกำหนดคำสั่งให้หุ่นยนต์สามารถปฏิบัติงานนั้นอย่างมีประสิทธิภาพ

3. ระบบเครือข่ายเส้นประสาท (neural networks) ระบบที่อาศัยความรู้เป็นพื้นฐาน สร้างเรียนแบบเส้นใยประสาทของมนุษย์ เป็นการทำกิจกรรมแบบขนานที่ทำงานพร้อม ๆ กันเพื่อให้ได้คำตอบเดียว เป็นระบบคอมพิวเตอร์ที่ถูกพัฒนาให้จำลองการทำงานของสมอง และระบบเส้นประสาทของมนุษย์ โดยระบบเครือข่ายเส้นประสาทจะมีความสามารถในการสังเกต การเรียนรู้ การจดจำ การทำซ้ำ และการแยกแยะถึงสิ่งต่าง ๆ ได้ ทำให้ระบบคอมพิวเตอร์ในรูปแบบนี้ สามารถพัฒนาศักยภาพในการปฏิบัติงานของตนอย่างต่อเนื่อง โดยมีประสิทธิภาพมากขึ้นและมีข้อบกพร่องน้อยลง

4. ระบบผู้เชี่ยวชาญ (expert system) โปรแกรมคอมพิวเตอร์ที่แสดงความสามารถได้เหมือนกับผู้เชี่ยวชาญ ในสาขาต่าง ๆ หรือในงานเฉพาะอย่างหรือระบบคอมพิวเตอร์ที่ถูกพัฒนาให้สามารถรับรู้ และทำงานเฉพาะด้าน ได้อย่างผู้เชี่ยวชาญ ปัจจุบันระบบผู้เชี่ยวชาญเริ่มได้รับความนิยมและนำมาใช้ทางธุรกิจและการดำเนินงานของหลายองค์การ ระบบความฉลาด ได้รับความสนใจมากขึ้น จากบุคคลหลายกลุ่ม โดยเฉพาะในภาคธุรกิจที่ต้องการระบบสารสนเทศ ที่มีศักยภาพสูง และสามารถปฏิบัติงานที่ซับซ้อนได้แทนบุคคล จึงมีความเป็นไปได้สูง ที่ศาสตร์ในสาขานี้ จะได้รับการพัฒนา และนำมาใช้งานอย่างเต็มที่ในอนาคตอันใกล้นี้

5. ระบบภาพ (vision system) การที่คอมพิวเตอร์สามารถทำงานได้โดยอาศัยการมอง และการจดจำรูปแบบ เช่นการตรวจหาชิ้นส่วนที่บกพร่อง การให้คอมพิวเตอร์แข่งเตะบอล เป็นต้น ถูกพัฒนาขึ้นเพื่อลอกเลียนการมองเห็นของบุคคล โดยมีส่วนรับสัญญาณภาพที่ทำหน้าที่รับสัญญาณแสง เพื่อทำการแปรรูปและประมวลผลตามหน้าที่ที่ถูกกำหนด เช่น การใช้คอมพิวเตอร์ตรวจสอบความบกพร่องของอุปกรณ์ โดยคอมพิวเตอร์จะมีข้อมูลเกี่ยวกับลักษณะต่าง ๆ ของความบกพร่องเก็บไว้ในหน่วยความจำ ถ้าระบบภาพสามารถตรวจพบความบกพร่องก็จะรายงานแก่ผู้ควบคุม โดย

เราสามารถนำระบบภาพ ไปใช้งานในสถานที่ที่มนุษย์ ไม่สามารถ เข้าไปตรวจสอบได้ด้วยตนเอง เนื่องจากข้อจำกัดของขนาด หรืออันตรายที่มีอยู่ในงาน

ประโยชน์ของปัญญาประดิษฐ์ในเชิงธุรกิจ

1. ข้อมูลจะถูกเก็บไว้ในลักษณะคล้ายกับเป็นหน่วยบันทึกความจำของขององค์กร กลายเป็นฐานความรู้องค์กร ที่พนักงานสามารถเข้าสืบค้น และหาคำปรึกษาได้ทุกเวลา ซึ่งถ้าเป็นผู้เชี่ยวชาญจริงๆ แล้วก็จะสามารหาคำปรึกษา ได้เฉพาะ ในช่วงเวลาปฏิบัติงานเท่านั้น

2. ระบบงานประยุกต์ทางปัญญาประดิษฐ์ จะช่วยสร้างกลไกที่ไม่นำความรู้สึก ความเหนื่อยล้า หรือความกังวล เข้ามาเป็นองค์ประกอบ ซึ่งจะมีประโยชน์เป็นอย่างมากกับงานประเภทที่อันตรายต่อมนุษย์ ไม่ว่าจะเป็นด้านสิ่งแวดล้อมที่ไม่ปลอดภัย ด้านร่างกาย หรือด้านจิตใจ ก็ตาม ระบบฯนี้ยังอาจทำหน้าที่ให้คำปรึกษาที่มีประโยชน์ในช่วงเวลาคับขันได้

3. ระบบงานประยุกต์ทางปัญญาประดิษฐ์ จะถูกนำมาทำงานในส่วนที่เป็นงานจำเป็น หรือเป็นงานที่น่าเบื่อหน่ายสำหรับมนุษย์

4. ระบบงานประยุกต์ทางปัญญาประดิษฐ์ จะช่วยเพิ่มความสามารถในฐานความรู้ขององค์กร ด้วยการเสนอวิธีแก้ปัญหาสำหรับงานเฉพาะด้าน ซึ่งมีปริมาณมากหรือ มีความซับซ้อนมากเกินไปสำหรบมนุษย์ โดยเฉพาะเมื่อต้องการทำงานนั้นให้สำเร็จภายในระยะเวลาอันสั้น

2.2 แนวคิดเกี่ยวกับการตัดคำและการค้นหา

2.2.1 วิธีที่ใช้ตัดคำและเทคนิคที่ช่วยในการตัดคำ

ยืน ภู่วรวรรณ และวิวรรธน์ อิ่มอารมณ์ (2535) ได้ให้คำจำกัดความของการตัดคำภาษาไทยเป็นกระบวนการพื้นฐานของการประมวลผลภาษาธรรมชาติ ได้รับการพัฒนาขึ้นมาโดยใช้วิธีการต่าง ๆ ที่ต่างกัน โดยแบ่งประเภทของการตัดคำได้ ดังนี้

1. การใช้กฎ

การตัดคำโดยการตรวจสอบกฎเกณฑ์ทางอักขระวิธีที่กำหนดลักษณะการประสมอักษร ลักษณะการเว้นวรรคและการขึ้นย่อหน้า เพื่อใช้เป็นเกณฑ์ในการกำหนดขอบเขตของคำ วิธีการนี้จะมีข้อจำกัดในการทำงาน คือ ความถูกตองของการตัดคำในระดับพยางค์สูงแต่ความถูกต้องของการตัดคำค่อนข้างต่ำ แต่ข้อดีของวิธีนี้คือมีความรวดเร็วในการทำงานและใช้ทรัพยากรน้อย

2. การใช้พจนานุกรม

การตัดคำโดยพจนานุกรมเป็นการตัดคำโดยใช้ใช้สายอักขระ(String) มาเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรม ซึ่งวิธีนี้ทำให้ได้ความถูกต้องในการตัดคำสูงกว่าการใช้กฎแต่จะใช้เวลามากกว่า

3. การใช้คลังข้อความ

การตัดคำโดยใช้คลังข้อมูลเป็นการตัดคำโดยนำวิธีทางสถิติเข้ามาประมวลผลภาษา โดยใช้คลังข้อมูลทางภาษาเป็นฐานความรู้เก็บค่าความถี่ที่ใช้ในการตัดคำ ซึ่งการตัดคำโดยใช้คลังข้อมูลแบ่งออกเป็น 2 วิธี คือการตัดคำโดยอาศัยความน่าจะเป็น (Probabilistic word segmentation) และวิธีการตัดคำโดยอาศัยคุณลักษณะของคำ (Feature-base word segmentation)

วิธีการตัดคำโดยอาศัยความน่าจะเป็น จะเป็นการตัดคำโดยใช้แบบจำลองเอยแกรม (Word n-gram model) ในการหารูปแบบของการตัดคำและลำดับคำที่เป็นไปได้มากที่สุด โดยวิธีการนี้ต้องมีการใช้คลังข้อมูลที่มีการตัดคำและกำกับหมวดคำที่เตรียมเอาไว้แล้ว ซึ่งวิธีการนี้ผลลัพธ์ที่ได้จะเป็นการเลือกรูปแบบการตัดคำที่มีความน่าจะเป็นมากที่สุด

ตัวอย่างของแบบจำลองไตรแกรม

“การพัฒนาระบบถาม-ตอบ” จะได้ว่า

การ / ารพ / รพั / พัฒ / ัฒน / ฒนา / นาร / าระ / ระบ / บบถ / ถาม / ามต / มตอ / ตอบ

หลังจากนั้นจะทำการเลือกคำที่เป็นไปได้เพื่อทำการประมวลผลต่อไปอย่างไร

วิธีการตัดคำโดยอาศัยคุณลักษณะของคำ จะเป็นการแก้ข้อผิดพลาดของการตัดคำโดยอาศัยความน่าจะเป็นของการจะเป็นของการจำกัดหมวดคำที่จะเป็นแบบจำลองในการตัดคำ ซึ่งวิธีการตัดคำโดยอาศัยคุณลักษณะของคำจะเป็นวิธีการแบบผสม (Hybrid approach)

เทคนิคที่ช่วยในการตัดคำ

เทคนิคที่ใช้ในการตัดคำที่นิยมใช้กันทั่วไปคือ วิธีการเทียบคำที่ยาวที่สุด วิธีการเทียบคำที่สั้นที่สุด วิธีการตัดคำที่ใช้ความถี่ของคำและวิธีการย้อนรอบกลับ

1. วิธีการเทียบคำที่ยาวที่สุด (Longest word pattern matching)

วิธีนี้จะทำการตรวจสายอักขระ (String) ที่นำเข้ามาจากซ้ายไปขวา จากนั้นนำไปเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรม หากตรวจสอบพบว่าพบพยางค์มากกว่า 1 พยางค์ในพจนานุกรม จะทำการเลือกพยางค์ที่ยาวที่สุดแล้วทำต่อไปเรื่อง ๆ ขนจบสายอักขระ

ตัวอย่างคำว่า “กองกลาง”

การตัดคำโดยวิธีนี้จะนำสายอักขระไปเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรมจะพบคำว่า ก , กอ และคำว่า กอง ส่วนคำว่า กองก ไม่พบในพจนานุกรม ดังนั้นจึงได้คำว่า กอง ซึ่งเป็นคำที่ยาวที่สุดที่หาพบ ส่วนที่เหลือคือ กลาง เมื่อนำไปค้นในพจนานุกรมจะได้คำว่า

ก , กล กลาง ดังนั้นจึงเลือกคำว่า กลาง

คำที่ได้จากการตัดคำโดยวิธีการนี้จึงเป็น กอง กลาง

วิธีการนี้ให้ความถูกต้องหลังการตัดคำคำสูงกว่าวิธีการอื่นโดยเฉพาะเมื่อใช้ร่วมกับวิธีย้อนรอยกลับ

2. วิธีการเทียบคำที่สั้นที่สุด (Shortest word pattern matching)

วิธีการนี้คล้ายกับวิธีการเทียบคำที่ยาวที่สุด เพียงแต่จะเลือกคำที่สั้นที่สุดที่พบก่อน แต่วิธีนี้พบว่าได้จำนวนคำมากที่สุดแต่ความถูกต้องของคำหลังทำการตัดคำน้อยกว่าใช้วิธีการเทียบคำที่ยาวที่สุด

ตัวอย่างคำว่า “โคลงเรือ”

การตัดคำโดยวิธีนี้จะเลือกเอาคำแรกที่ค้นหาเจอจากพจนานุกรม ดังนั้นจะได้ว่า

โค ลง เรือ (โดยไม่เลือกคำว่า “โคลง” ที่จะพบต่อไปภายหลังหากทำการค้นหาต่อ)

วิธีนี้ใช้เวลาน้อยกว่าการเทียบคำที่ยาวที่สุด แต่ความถูกต้องที่ได้การตัดคำแบบเทียบคำยาวที่สุดมากกว่า

3. วิธีการตัดคำที่ใช้ความถี่ของคำ (Word usage frequency)

วิธีการนี้เป็นแนวทางหนึ่งในการแก้ปัญหาคำกำกวมของประโยคภาษาไทยโดยการวิเคราะห์ความถี่ของการใช้คำในชีวิตประจำวันโดยจัดเรียงคำในพจนานุกนตามความถี่ที่พบ และใช้วิธีการตัดคำแบบเดียวกับ 2 วิธีข้างต้น

ตัวอย่างคำว่า “ก็อด”

ในกรณีนี้หากใช้ความถี่ของคำจะได้ว่า

ก็ อด (เนื่องจาก ก็ มีความถี่สูงกว่าคำว่า ก็อด)

4. วิธีการย้อนรอยกลับ (Back tracking)

เมื่อทำการเปรียบเทียบคำที่นำมาตัดคำกับคำที่มีอยู่ในพจนานุกรม อาจพบกรณีที่คำที่พบมีมากกว่า 1 คำแล้วทำการเลือกคำที่ยาวที่สุดทำให้สายอักขระที่ตามมาจากคำนั้นไม่สามารถตัดคำได้ เนื่องจากไม่พบตามพจนานุกรม กรณีนี้จะทำการย้อนไปอีกคำที่ไม่ถูกเลือกแล้วทำการตัดคำต่อไป

ตัวอย่างเช่นคำว่า “เมื่อยามนี้” การเปรียบเทียบกับพจนานุกรมจะได้ว่า

เมื่อ , เมื่อย ดังนั้นจึงเลือกคำที่ยาวที่สุดจะได้คำว่า เมื่อย

ส่วนที่เหลือ คือ –ามนี้ ซึ่งไม่พบอยู่ในพจนานุกรม ดังนั้นจะทำการย้อนกลับไปเพื่อเลือกอีกคำหนึ่งคือ เมื่อ จะได้เป็น

เมื่อ ยาม นี้ (โดยคำว่า ยาม เกิดจากการเลือกคำที่ยาวที่สุดระหว่าง ยา และ ยาม)

2.2.2 การค้นหาข้อมูล

การค้นหาคำตอบ หรือการค้นหาข้อมูลในทางคอมพิวเตอร์มักจะกระทำบนโครงสร้างข้อมูลแบบต้นไม้ และกราฟ ทั้งนี้เพราะโครงสร้างข้อมูลในลักษณะนี้สามารถทำให้การค้นหาทำได้สะดวกและสามารถพลิกแพลงการค้นหาได้ง่าย ในความเป็นจริงแล้ว การค้นหาข้อมูลบางครั้งสามารถกระทำบนโครงสร้างข้อมูลชนิดอื่นก็ได้เช่น อาเรย์ สแตก และคิว แต่การจัดข้อมูลในโครงสร้างเช่นนี้ มีข้อจำกัดในการค้นหาข้อมูลมาก การค้นหาทำได้แบบเรียงลำดับ (Sequencial Search) เท่านั้น ซึ่งใช้ได้กับข้อมูลที่มีขนาดเล็ก ดังนั้นในการค้นหาข้อมูลที่มีขนาดใหญ่ ก่อนการค้นหา หรือระหว่างการค้นหา ข้อมูลที่จะถูกค้นจะต้องถูกจัดให้อยู่ในรูปแบบของต้นไม้ หรือกราฟเท่านั้น การค้นหาข้อมูลบนโครงสร้างต้นไม้และกราฟสามารถจำ าแนกได้ 2 แบบคือ การค้นหาแบบไบล์ด (Blind Search) และการค้นหาแบบฮิวริสติก (Heuristic Search)

การค้นหาแบบฮิวริสติก (heuristic search) มีความความแตกต่างจากการค้นหาข้อมูลแบบธรรมดาและแบบฮิวริสติกนั้นอยู่ที่การค้นหาข้อมูล ธรรมดาผู้ที่ทำการค้นข้อมูลจะต้องตรวจสอบข้อมูลทีละตัวทุกตัวจนครบ แต่ฮิวริสติกจะไม่ลงไปดู ข้อมูลทุกตัว วิธีการนี้จะเลือกได้คำตอบที่เหมาะสมให้กับการค้นหา ซึ่งมีข้อดีคือ สามารถทำการ ค้นหาคำตอบจาก ข้อมูลที่มีขนาดใหญ่มากๆ ได้ แต่มีข้อเสียคือคำตอบที่ได้เป็นเพียงคำตอบที่ดี เท่านั้นไม่แน่ว่าจะดีที่สุด แต่เนื่องจากว่าปัญหาในบางลักษณะนั้นใหญ่มาก และเป็นไปไม่ได้ที่จะทำ การค้นหาด้วยวิธี ธรรมดากระบวนการของฮิวริสติกจึงเป็นสิ่งที่จำเป็นในเรื่องของฮิวริสติกนั้น นอกจากจะมีการค้นหาแบบฮิวริสติกแล้วยังมีอีกสิ่งหนึ่งที่สำคัญคือ ฮิวริสติกฟังก์ชัน (heuristic function) ซึ่งหมายถึงฟังก์ชันที่ทำหน้าที่ในการวัดขนาดของความเป็น ไปได้ในการแก้ปัญหาซึ่งจะแสดงด้วยตัวเลข วิธีการดังกล่าวจะกระทำได้โดยการพิจารณาถึงวิธีการ (aspects) ต่าง ๆ ที่ใช้ในการแก้ปัญหา ณ สถานะหนึ่งว่าจะสามารถแก้ปัญหาได้ตามที่ต้องการหรือไม่ โดยกำหนดเป็นน้ำหนักที่ให้กับการแก้ปัญหาของแต่ละวิธี น้ำหนักเหล่านี้จะถูกแสดงด้วยตัวเลขที่กำกับไว้กับโหนดต่าง ๆ ในกระบวนการ ค้นหา และค่าเหล่านี้จะเป็นตัวที่ใช้ในการประมาณความเป็นไปได้ว่าเส้นทางที่ผ่านโหนดนั้นจะมี ความเป็นไปได้ในการนำไปสู่หนทางการแก้ปัญหาได้มากน้อยแค่ไหนจุดประสงค์ที่แท้จริงของฮิวริสติก ฟังก์ชันก็คือการกำกับทิศทางของกระบวนการค้นหา เพื่อให้อยู่ในทิศทางที่ได้ประโยชน์สูงสุด โดยการบอกว่าเราควรเลือกเดินเส้นทางไหนก่อน ในกรณีที่มีเส้น ทางมากกว่าหนึ่งเส้นทางต้องเลือกกระบวนการ

ค้นหาแบบฮิวริสติก โดยปกติแล้วจะต้องอาศัยฮิวริสติกฟังก์ชัน ทำให้การแก้ปัญหาหนึ่ง ๆ จะดีหรือไม่ ก็ขึ้นอยู่กับฮิวริสติกฟังก์ชันดังนั้นการค้นหาแบบนี้จึงไม่มีอะไรเป็นหลักประกันว่าจะได้สิ่งที่ไม่ดีออกมาด้วยเหตุนี้เอง เราจึงเรียกการ ค้นหาแบบฮิวริสติกนี้ว่า Weak Methods หรือจะกล่าวอีกนัยหนึ่งคือ Weak Methods เป็นกระบวนการควบคุมโดยทั่วไป (general-purpose control stategies) ซึ่งการค้นหาแบบนี้ สามารถแบ่งได้เป็น

การค้นหาแบบ A*

การค้นหาแบบ A* เป็นอีกแบบของการค้นหาแบบดีที่สุดก่อน วิธีการเลือกโหนดที่จะใช้ในการดำเนินการต่อจะพิจารณาจากโหนดที่ดีที่สุด แต่ในกรณีของ A* นี้จะมีลักษณะพิเศษกว่าคือ ในส่วนของฮิวริสติกฟังก์ชัน ในกรณีของการค้นหาแบบดีที่สุดก่อนนั้น ค่าที่ได้จากฮิวริสติก ฟังก์ชันจะเป็นค่าที่วัดจาก โหนดปัจจุบัน แต่ในกรณีของ A* ค่าของฮิวริสติก ฟังก์ชัน จะวัดจากค่า 2 ค่าคือ ค่าที่วัดจากโหนดปัจจุบันไปยังโหนดราก และจากโหนดปัจจุบันไปยังโหนดเป้าหมาย ถ้าเราให้ตัวแปร f แทนค่าของฮิวริสติก ฟังก์ชัน g เป็นฟังก์ชันที่ใช้วัดค่า cost จากสถานะเริ่มต้นจนถึงสถานะปัจจุบัน h' เป็นฟังก์ชันที่ใช้วัดค่า cost จากสถานะปัจจุบันถึงสถานะเป้าหมาย ดังนั้น

$$f = g + h'$$

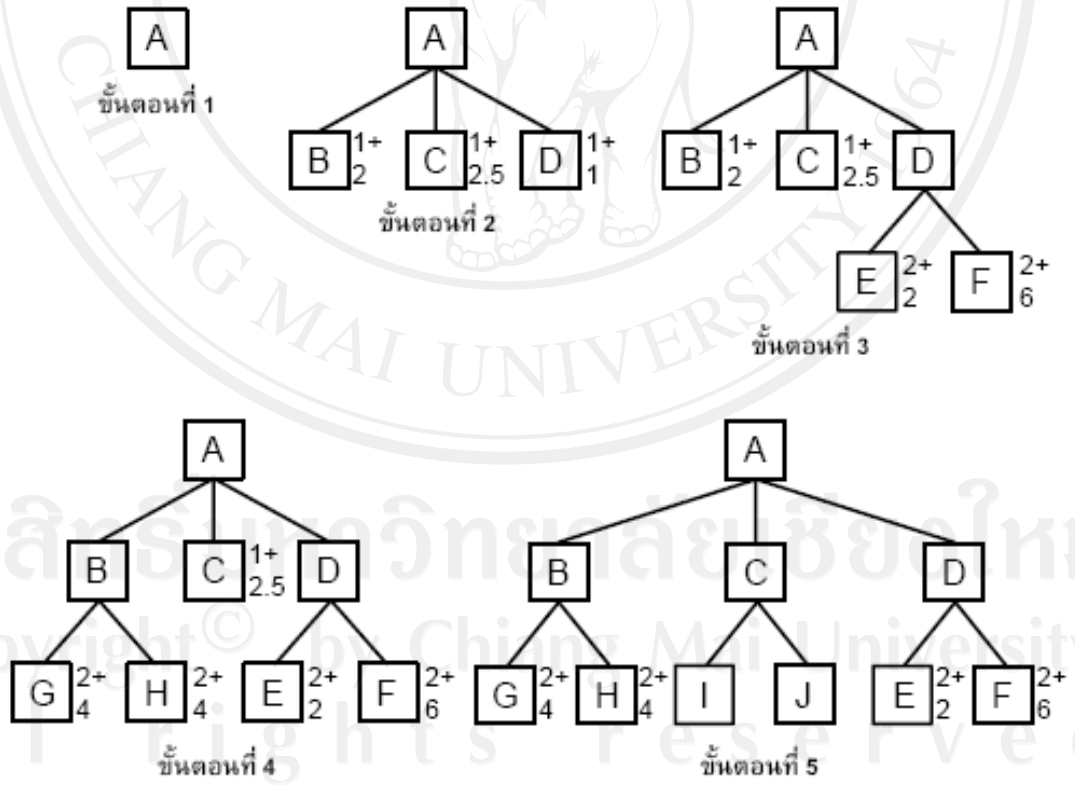
อัลกอริทึม A* (A* Search) เป็นการขยายอัลกอริทึมดีสุดก่อนโดยพิจารณาเพิ่มเติมถึงต้นทุนจากสถานะเริ่มต้นมายังสถานะปัจจุบันเพื่อใช้คำนวณค่าฮิวริสติกด้วย ในกรณีของอัลกอริทึม A* เราต้องการหาค่าต่ำสุดของฟังก์ชัน f' ของสถานะ s นิยามดังนี้

$$f'(s)=g(s)+h'(s)$$

โดยที่ g คือฟังก์ชันที่คำนวณต้นทุนจากสถานะเริ่มต้นมายังสถานะปัจจุบัน h' คือฟังก์ชันที่ประมาณต้นทุนจากสถานะปัจจุบันไปยังคำตอบ ดังนั้น f' จึงเป็นฟังก์ชันที่ประมาณต้นทุนจากสถานะเริ่มต้นไปยังคำตอบ (ยิ่งน้อยยิ่งดี) เรามองได้ว่าฟังก์ชัน h' คือฟังก์ชันฮิวริสติกที่เราเคยใช้ในการค้นหาอื่น ๆ ก่อนหน้านี้เช่นอัลกอริทึมปีนเขา อัลกอริทึมดีสุดก่อน เป็นต้น ในที่นี้เราใส่เครื่องหมาย ' เพื่อแสดงว่าฟังก์ชันนี้เป็นฟังก์ชันประมาณของฟังก์ชันจริงที่ไม่รู้ (เราทำได้แค่ประมาณว่า h' คือต้นทุนจากสถานะปัจจุบันไปยังคำตอบ เราจะรู้ต้นทุนจริงก็ต่อเมื่อเราได้ทำการ

ค้นหาจริงจนไปถึงคำตอบแล้ว) ส่วน g เป็นฟังก์ชันที่คำนวณต้นทุนจริงจากสถานะเริ่มต้นมายังสถานะปัจจุบัน (จึงไม่ได้ใส่เครื่องหมาย ') เพราะเราสามารถหาต้นทุนจริงได้เนื่องจากได้ค้นหาจากสถานะเริ่มต้นจนมาถึงสถานะปัจจุบันแล้ว ส่วน f' ก็เป็นเพียงแค่ฟังก์ชันประมาณโดยการรวมต้นทุนทั้งสอง คือ h' กับ g

อัลกอริทึม A* จะทำการค้นหาโดยวิธีเดียวกันกับอัลกอริทึมดีสุดก่อนทุกประการ ยกเว้นฟังก์ชันฮิวริสติกที่ใช้เปลี่ยนมาเป็น f' (ต่างจากอัลกอริทึมดีสุดก่อนที่ใช้ h') โดยการใช้ f' อัลกอริทึม A* จึงให้ความสำคัญกับสถานะหนึ่ง ๆ 2 ประการ คือ (1) สถานะที่ดีต้องมี h' ดีคือต้นทุนเพื่อจะนำไปสู่คำตอบหลังจากนี้ต้องน้อย และ (2) ต้นทุนที่จ่ายไปแล้วกว่าจะถึงสถานะนี้ (g) ต้องน้อยด้วย เราจึงได้ว่า A* จะค้นหาเส้นทางที่ให้ต้นทุนโดยรวมน้อยที่สุดตามค่า f' ซึ่งต่างจากอัลกอริทึมดีสุดก่อน ที่เน้นความสำคัญของสถานะที่ต้นทุนหลังจากนี้ที่จะนำไปสู่คำตอบต้องน้อยโดยไม่สนใจว่าต้นทุนที่จ่ายไปแล้วกว่าจะนำมาถึงสถานะนี้ต้องเสียไปเท่าไหร่



รูปที่ 2.1 แสดงการค้นหาด้วยอัลกอริทึม A*

การค้นหาด้วยอัลกอริทึม A* กันสถานะในรูปที่ 8 โดยสมมติให้ต้นทุนหรือระยะห่างระหว่างสถานะพ่อแม่ไปยังสถานะลูกเท่ากับ 1 หน่วย เช่นต้น ทุนจริง (g) จาก A ไปยัง B,C หรือ D มีค่าเท่ากับ 1 หน่วย

จากรูปจะเห็นได้ว่าในขั้นตอนที่ 4 สถานะ C จะถูกเลือกมากระจายโดยอัลกอริทึม A* เนื่องจากมีค่า f' น้อยสุดเท่ากับ 3.5 ซึ่งน้อยกว่า E ที่มีค่าเท่ากับ 4 แม้ว่าค่า h' ของ E จะน้อยกว่าซึ่งต่างจากการสร้างสถานะของอัลกอริทึมดีสุดก่อน

ทฤษฎีความน่าจะเป็นของเบย์

ในวงการวิชาการความน่าจะเป็น มีสำนักหรือแนวความคิดที่สำคัญอยู่ 2 สำนัก คือ สำนักที่เชื่อว่าความน่าจะเป็นของเหตุการณ์หรือสมมติฐานใดควรจะเป็นความถี่ของการเกิดเหตุการณ์หรือสมมติฐานนั้น โดยเทียบจากเหตุการณ์ที่เกิดขึ้นทั้งหมดจำนวนมาก ๆ ตัวอย่างเช่น ความน่าจะเป็นของการออกหัวจากการโยนเหรียญเท่ากัน 0.5 ตัวเลข 0.5 นี้ได้จากการโยนเหรียญจำนวนมากครั้งแล้วปรากฏว่าเป็นหัวครึ่งหนึ่งของจำนวนการโยนเหรียญทั้งหมด ปัญหาของสำนักแนวความคิดนี้คือเหตุการณ์ทั้งหมดเกิดขึ้นซ้ำ ๆ กันจำนวนมาก ๆ ได้หรือไม่ ส่วนอีกสำนักหนึ่งนั้นเชื่อว่าความน่าจะเป็นของเหตุการณ์ใดเป็นการวัดความเชื่อส่วนตัวถึงความเป็นไปได้ในการเกิดของเหตุการณ์นั้น ตามความเชื่อของสำนักนี้ เราไม่จำเป็นต้องทำเหตุการณ์ให้เกิดขึ้นซ้ำ ๆ แล้ววัดความถี่ของเหตุการณ์ สำนักนี้มี

ขนาดของความเชื่อในสมมติฐานนั้นอาจจะเปลี่ยนแปลงเมื่อเราได้รับหลักฐานใหม่ ดังนั้น จึงจำเป็นต้องแยกความแตกต่างๆ ระหว่างความน่าจะเป็นของสมมติฐานก่อนและภายหลังได้รับหลักฐาน

2.3 แนวคิดเกี่ยวกับระบบฐานข้อมูลและระบบจัดการฐานข้อมูล

2.3.1 ความหมาย

กิตติ ภักดีวัฒนะกุล และ จำลอง ครูอุตสาหะ (2542) กล่าวไว้ว่า ระบบฐานข้อมูล หมายถึง การจัดเก็บข้อมูลอย่างมีระบบ และความสัมพันธ์ระหว่างข้อมูลประกอบด้วย รายละเอียดของข้อมูลที่เกี่ยวข้องกัน ซึ่งถูกนำมาใช้งานด้านต่าง ๆ ไม่ว่าจะเป็นการเพิ่มข้อมูล การลบ การแก้ไข การเรียกดูข้อมูล เช่น ด้านสถาบันการศึกษา จะมีฐานข้อมูลที่เกี่ยวข้องกับ ข้อมูลอาจารย์ ข้อมูลนักศึกษา และข้อมูลเจ้าหน้าที่ เป็นต้น ซึ่งข้อมูลเหล่านี้จัดเก็บไว้อย่างเป็นระบบ เพื่อประโยชน์ในการจัดการและเรียกใช้ข้อมูลได้อย่างมีประสิทธิภาพ

โครงข่ายสารสนเทศเพื่อพัฒนาการศึกษา ทบวงมหาวิทยาลัย (2544) ได้กำหนดความหมายของระบบฐานข้อมูล ความสำคัญของระบบฐานข้อมูล การบริหารฐานข้อมูล และหน้าที่ของผู้บริการฐานข้อมูลไว้ดังนี้

ระบบฐานข้อมูล (Database) หมายถึง กลุ่มของข้อมูลที่ถูกเก็บไว้ โดยมีความสัมพันธ์ซึ่งกันและกัน โดยไม่ได้บังคับว่าข้อมูลทั้งหมดนี้จะต้องเก็บไว้ในแฟ้มข้อมูลเดียวกัน หรือแยกเก็บหลายๆ แฟ้มข้อมูล นั่นคือการเก็บข้อมูลในฐานข้อมูลนั้นเราอาจจะเก็บไว้ในหลายๆ แฟ้มข้อมูล ที่สำคัญเราจะต้องสร้างความสัมพันธ์ระหว่างระเบียนและเรียกใช้ความสัมพันธ์ระหว่างระเบียนและเรียกใช้ความสัมพันธ์นั้นได้ มีการกำจัดความซ้ำซ้อนของข้อมูลออกและเก็บแฟ้มข้อมูลเหล่านี้ไว้ที่ศูนย์กลาง เพื่อที่จะนำข้อมูลเหล่านี้มาใช้ร่วมกัน ควบคุมดูแลรักษาเมื่อผู้ต้องการใช้งานและผู้มีสิทธิ์จะใช้ข้อมูลนั้นสามารถดึงข้อมูลที่ต้องการออกไปใช้ได้ ข้อมูลบางส่วนอาจใช้ร่วมกันผู้อื่นได้ แต่บางส่วนผู้มีสิทธิ์เท่านั้นจึงจะสามารถใช้ได้ โดยทั่วไปองค์กรต่างๆ จะสร้างฐานข้อมูลไว้เพื่อเก็บข้อมูลต่างๆ ของตัวองค์กร โดยเฉพาะอย่างยิ่งข้อมูลในเชิงธุรกิจ เช่น ข้อมูลลูกค้า ข้อมูลสินค้า ข้อมูลของลูกจ้าง และการจ้างงาน เป็นต้น การควบคุมดูแลการใช้ฐานข้อมูลนั้น เป็นเรื่องที่ยุ่งยากกว่าการใช้แฟ้มข้อมูลมาก เพราะเราจะต้องตัดสินใจว่าโครงสร้างในการจัดเก็บข้อมูลควรจะเป็นเช่นไร การเขียนโปรแกรมเพื่อสร้างและเรียกใช้ข้อมูลจากโครงสร้างเหล่านี้ ถ้าโปรแกรมเหล่านี้เกิดทำงานผิดพลาดขึ้นมา ก็จะเกิดความเสียหายต่อโครงสร้างของข้อมูลทั้งหมดได้ เพื่อเห็นการลดภาวะการทำงานของผู้ใช้ จะได้มีส่วนของฮาร์ดแวร์และโปรแกรมต่างๆ ที่สามารถเข้าถึงและจัดการข้อมูลในฐานข้อมูลนั้น เรียกว่า ระบบจัดการฐานข้อมูล หรือ DBMS (Database Management System)

ระบบจัดการฐานข้อมูล หมายถึง ซอฟต์แวร์ที่เปรียบเสมือนสื่อกลางระหว่างผู้ใช้และโปรแกรมต่างๆ ที่เกี่ยวข้อกับการใช้ฐานข้อมูล ซึ่งมีหน้าที่ช่วยให้ผู้ใช้เข้าถึงข้อมูลได้ง่ายสะดวกมีประสิทธิภาพ การเข้าถึงข้อมูลของผู้ใช้อาจเป็นการสร้างฐานข้อมูล การแก้ไขฐานข้อมูล หรือการตั้งคำถามเพื่อให้ข้อมูลมา โดยผู้ใช้ไม่จำเป็นต้องรับรู้เกี่ยวกับรายละเอียดภายในโครงสร้างของฐานข้อมูล เปรียบเสมือนเป็นสื่อกลางระหว่างผู้ใช้และโปรแกรมต่างๆ ที่เกี่ยวข้องกับการใช้ฐานข้อมูล

2.3.2 ความสำคัญของระบบฐานข้อมูล

การจัดการข้อมูลให้เป็นระบบฐานข้อมูลทำให้ข้อมูลมีส่วนดีกว่าการเก็บข้อมูลในรูปแบบแฟ้มข้อมูลดังนี้

1) ลดการเก็บข้อมูลที่ซ้ำซ้อน ข้อมูลบางชุดที่อยู่ในรูปแบบแฟ้มข้อมูลอาจมีปรากฏอย่างหลายๆ แห่ง เพราะมีผู้ใช้ข้อมูลชุดนี้หลายคน เมื่อใช้ระบบฐานข้อมูลแล้วจะช่วยให้

ความซ้ำซ้อนของข้อมูลลดน้อยลง เช่น ข้อมูลอยู่ในแฟ้มข้อมูลของผู้ใช้หลายคน ผู้ใช้แต่ละคนจะมีแฟ้มข้อมูลเป็นของตนเอง ระบบฐานข้อมูลจะลดการซ้ำซ้อนของข้อมูลเหล่านี้ให้มากที่สุด โดยจัดเก็บในฐานข้อมูล ทำให้ไม่เปลืองเนื้อที่ในการเก็บข้อมูลและลดความซ้ำซ้อนลง

2) รักษาความถูกต้องของข้อมูล เนื่องจากฐานข้อมูลมีเพียงฐานข้อมูลเดียว ในกรณีที่มีข้อมูลชุดเดียวกันปรากฏอยู่หลายแห่ง ในฐานข้อมูล ข้อมูลเหล่านี้จะต้องตรงกัน ถ้ามีการแก้ไขข้อมูลนี้ทุกๆ แห่งที่ข้อมูลปรากฏอยู่จะแก้ไขให้ถูกต้องตามกันหมดโดยอัตโนมัติด้วยระบบจัดการฐานข้อมูล
3) การป้องกันและรักษาความปลอดภัยให้กับฐานข้อมูลระบบฐานข้อมูลทำได้อย่างสะดวก การป้องกันและรักษาความปลอดภัยกับข้อมูลระบบฐานข้อมูลจะให้เฉพาะผู้ที่เกี่ยวข้องเท่านั้น จึงจะมีสิทธิ์เข้าไปใช้ฐานข้อมูลได้ เรียกว่ามีสิทธิส่วนบุคคล (Privacy) ซึ่งก่อให้เกิดความปลอดภัย (Security) ของข้อมูลด้วย ฉะนั้นผู้ใดจะมีสิทธิ์ที่จะเข้าถึงข้อมูลได้จะต้องมีการกำหนดสิทธิ์กันไว้ก่อนและเมื่อเข้าไปใช้ข้อมูลนั้นๆ ผู้ใช้จะเห็นข้อมูลที่เก็บไว้ในฐานข้อมูลในรูปแบบที่ผู้ใช้ออกแบบไว้
4) สามารถใช้ข้อมูลร่วมกันได้ เนื่องจากในระบบฐานข้อมูลจะเป็นที่เก็บรวบรวมข้อมูลทุกอย่างไว้ ผู้ใช้แต่ละคนจึงสามารถที่จะใช้ข้อมูลในระบบได้ทุกข้อมูล ซึ่งถ้าข้อมูลไม่สามารถที่จะใช้ข้อมูลไม่ได้จัดเก็บให้เป็นระบบฐานข้อมูลแล้ว ผู้ใช้ก็จะใช้ได้เพียงข้อมูลของตนเองเท่านั้น ถ้าเก็บไว้ในฐานข้อมูลก็จะสามารถใช้ร่วมกันได้
5) มีความเป็นอิสระของข้อมูล เมื่อผู้ใช้ต้องการเปลี่ยนแปลงข้อมูลหรือนำข้อมูลมาประยุกต์ใช้ให้เหมาะสมกับโปรแกรมที่เขียนขึ้นมาก จะสามารถสร้างข้อมูลนั้นขึ้นมาใช้ใหม่ได้ โดยไม่มีผลกระทบต่อระบบฐานข้อมูล เพราะข้อมูลที่ผู้ใช้นำมาประยุกต์ใช้ใหม่นั้นจะไม่กระทบต่อโครงสร้างที่แท้จริงของการจัดเก็บข้อมูล นั่นคือ การใช้ระบบฐานข้อมูลจะทำให้เกิดความเป็นอิสระระหว่างการจัดเก็บข้อมูลและการประยุกต์ใช้
6) สามารถขยายงานได้ง่าย เมื่อต้องการจัดเพิ่มเติมข้อมูลที่เกี่ยวข้องจะสามารถเพิ่มได้อย่างง่ายไม่ซับซ้อน เนื่องจากมีความเป็นอิสระของข้อมูล จะไม่มีผลกระทบต่อข้อมูลเดิมที่มีอยู่
7) ทำให้ข้อมูลบูรณะกลับสู่สภาพปกติได้เร็วและมีมาตรฐาน เนื่องจากการจัดพิมพ์ข้อมูลในระบบที่ไม่ได้ใช้ฐานข้อมูล ผู้เขียนโปรแกรมแต่ละคนมีแฟ้มข้อมูลของตนเองเฉพาะ ฉะนั้นแต่ละคนจึงต่างก็สร้างระบบการบูรณะข้อมูลให้กลับสู่สภาพปกติในกรณีที่ข้อมูลเสียหายด้วยตนเองและด้วยวิธีการของตนเอง จึงขาดประสิทธิภาพและ

มาตรฐาน แต่เมื่อมาเป็นระบบฐานข้อมูลแล้ว การบูรณะข้อมูลให้กลับคืนสู่สภาพปกติ จะมีโปรแกรมชุดเดียวที่ดูแลทั้งระบบ ซึ่งย่อมต้องมีประสิทธิภาพและเป็นมาตรฐานเดียวกันแน่นอน

2.4 งานวิจัยที่เกี่ยวข้อง

ผศ.ดร.กานดา รุณนะพงศา และ นางสาวปโยธร อุราธรรมกุล (2548) ได้ศึกษาและจัดทำงานวิจัยในหัวข้อเรื่อง การตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่ โดยมีวัตถุประสงค์เพื่อสร้างศึกษาและพัฒนาอัลกอริทึม ในการตัดคำเอกสารภาษาไทยโดยการใช้กฎ (Rule-base) ให้มีประสิทธิภาพมากขึ้น และพัฒนาขั้นตอนวิธีในการตัดคำภาษาไทยโดยการใช้กฎให้มีความถูกต้องสูงขึ้น โดยใช้สถานที่วิจัยคือ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น

การศึกษาและพัฒนาขั้นตอนวิธีการตัดคำในภาษาไทยโดยวิธีการใช้กฎ (Rule-base) เพื่อแก้ไขปัญหาคำเฉพาะและคำที่มาจากภาษาต่างประเทศโดยเน้นที่คำมาจากภาษาต่างประเทศร่วมกับการใช้พจนานุกรมแบบใหม่ โดยมีขั้นตอนคือ การตัดอนุประโยคโดยอาศัยช่องว่างและอักขระพิเศษ การตัดคำโดยอาศัยกฎการผสมอักษรในภาษาไทย การแบ่งประเภทอนุประโยค การวิเคราะห์คำที่มีอยู่ในพจนานุกรมและคำที่ไม่มีอยู่ในพจนานุกรมที่มีความเป็นไปได้ที่จะเป็นภาษาต่างประเทศหรือคำเฉพาะ

ประสิทธิภาพของการตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่มุ่งเน้นเพื่อแก้ไขปัญหาการตัดคำด้วยกฎที่มีอยู่เดิมซึ่งไม่ครอบคลุมคำในภาษาไทยซึ่งมีคำที่มีลักษณะที่แตกต่างซับซ้อนออกไปจากเดิมโดยเฉพาะคำเฉพาะ และคำที่มาจากภาษาต่างประเทศโดยการเพิ่มกฎเพื่อความยืดหยุ่นมากขึ้น อีกทั้งเพิ่มกฎความเป็นไปได้ที่จะเป็นภาษาต่างประเทศหรือคำเฉพาะเพื่อเพิ่มความถูกต้องของการตัดคำโดยเฉพาะคำที่ไม่มีอยู่ในพจนานุกรม อีกทั้งการตัดคำโดยอาศัยคำภาษาอังกฤษที่ปรากฏในเอกสารสามารถลดความผิดพลาดจากการตัดคำอ่านภาษาไทยที่บางส่วนของคำหรือทั้งหมดของคำไม่ปรากฏ

รายงานนี้ได้ปรับปรุงเพื่อตัดคำที่มาจากภาษาต่างประเทศและคำที่ไม่พบในพจนานุกรม แต่คำเฉพาะบางคำที่มีการสะกดอ่านได้หลายพยางค์ไม่สามารถรวมให้เป็นคำเดียวได้นอกจากทำการบันทึกคำนั้นไว้ในพจนานุกรม ซึ่งทำได้เพียงบางส่วนจากภาษาอังกฤษที่อยู่ในเอกสารนำมาแปลงให้เป็นคำอ่านในภาษาไทย ส่วนคำกำกวมสามารถแก้ปัญหาโดยการใช้วิธีตัดคำที่ยาวที่สุดร่วมกับวิธีการย้อนกลับแต่ยังไม่สามารถตัดคำได้ถูกต้องทุกครั้งเนื่องจากการเลือกคำที่ยาวที่สุด

ไม่ใช่กรณีที่ดีที่สุด ดังนั้นควรมีการเพิ่มเติม ปรับปรุงสำหรับตัดคำประเภทนี้ด้วยการใช้ค่าสถิติของคำที่พบในเอกสารทั่วไปร่วมกับการใช้ความถี่ของคำที่พบในเอกสารที่นำมาตัดคำ