บทที่ 1

ที่มาและความสำคัญของปัญหา

1.1 ความสำคัญของปัญหา

การขยายตัวทางการใช้คอมพิวเตอร์และอินเตอร์เน็ตในช่วงระยะเวลาที่ผ่านมาเจริญก้าวหน้าอย่างรวดเร็วส่งผลให้เกิดการเก็บข้อมูลหลายชนิดในรูปเอกสารอิเล็กทรอนิกส์ (Electronic Text Document) เช่น เว็บเพจ เอกสารข่าว จดหมายอิเล็กทรอนิกส์ หนังสือ วารสาร เป็นต้น และนับวันข้อมูลประเภทนี้ซึ่งเป็นทรัพยากรที่สำคัญอย่างหนึ่งขององค์กรจะเพิ่มปริมาณสูงขึ้นเรื่อย ๆ ส่งผลให้การค้นหาข้อมูลมีความลำบากมากยิ่งขึ้น ดังนั้นจึงมีการนำเทคนิคการสืบค้นความรู้บางชนิดมาประยุกต์ใช้กับข้อมูลเอกสารอิเล็กทรอนิกส์เพื่อให้ผู้ใช้สามารถสืบค้นเอกสารที่ตรงหรือใกล้เคียงกับความต้องการได้สะดวกมากยิ่งขึ้น (พิลาวัณย์ พลับรู้การและกฤษณะ ไวยมัย 2545; Chowdhury 2004) ระบบการค้นคืนสารสนเทศประเภทนี้จึงเป็นสิ่งจำเป็น โดยเฉพาะในด้านธุรกิจและการศึกษา ในปัจจุบันได้มีการนำคอมพิวเตอร์มาประยุกต์ใช้งานร่วมกับการค้นคืนสารสนเทศอิเล็กทรอนิกส์ที่อยู่ในรูปแบบตัวอักษรหรือข้อความตัวอย่างเช่น โปรแกรมการทำงานสำหรับระบบงานแผนกช่วยเหลือ (Help Desk System) โปรแกรมค้นหาข่าวสาร ห้องสมุดอิเล็กทรอนิกส์ และโปรแกรมค้นคืนเอกสารต่างๆ ในคลังเอกสาร เป็นต้น (ศิรัตน์ ศิรนานนท์ 2549)

การค้นคืนสารสนเทศ (Information Retrieval) มีวัตถุประสงค์เพื่อคืนสารสนเทศที่ผู้ใช้ต้องการได้อย่างครบถ้วน มีประสิทธิภาพและถูกต้องมากที่สุด โดยในการค้นคืนเอกสารแต่ละครั้งผู้ใช้อาจจะได้สารสนเทศที่เกี่ยวข้องกับความต้องการเพียงบางส่วนหรืออาจจะได้สารสนเทศที่ไม่เกี่ยวข้องออกมาในปริมาณมาก ซึ่งเป็นหนึ่งในปัญหาที่สำคัญของการค้นคืนสารสนเทศ ปัญหาสำคัญอีกประการหนึ่งคือ ปัญหาเรื่องคำสำคัญ (Keyword) ที่ผู้ใช้ระบุในข้อสอบถาม (Query) ขาดความเฉพาะเจาะจง ซึ่งอาจส่งผลให้ผู้ใช้ได้รับเอกสารจากการค้นหามากมายมหาศาล โดยเอกสารต่าง ๆ เหล่านั้นอาจจะเป็นเอกสารที่ไม่เกี่ยวข้องกับข้อสอบถาม หรืออีกนัยหนึ่งคือเอกสารที่เกี่ยวข้องกับข้อสอบถามอาจไม่ถูกค้นคืนโดยระบบ ในทางตรงข้ามข้อสอบถามที่มีความเฉพาะเจาะจงอาจทำให้ไม่ได้รับเอกสารใดๆ จากการค้นหาเลย (วันเพ็ญ ชูเมือง 2547; ชูชาติ หฤไชยะศักดิ์ 2548) โดยเฉพาะอย่างยิ่งในกรณีที่ผู้ใช้ขาดความรู้เกี่ยวกับคำเฉพาะที่ใช้ในการอ้างอิงถึงข้อมูลที่ต้องการค้นหา โดยเมื่อผู้ใช้กรอกข้อสอบถามไม่เหมาะสมเข้ามาในระบบ ระบบ

จะค้นคืนข้อมูลที่ตรงกับข้อสอบถามของผู้ใช้ที่กรอกเข้ามา แต่อาจจะไม่สามารถค้นคืนข้อมูลที่ตรงกับความต้องการของผู้ใช้อย่างแท้จริง

ปัญหาในการเลือกใช้คำสำคัญที่มีความเฉพาะเจาะจงจะส่งผลต่อประสิทธิภาพของระบบค้นคืนเอกสารเป็นอย่างมากในกรณีที่เอกสารกลุ่มนั้นเป็นเอกสารที่มีความหลากหลายในหัวข้อ ตัวอย่างหนึ่งของเอกสารประเภทดังกล่าวคือ เอกสารที่ใช้ประกอบการตัดสินใจทางธุรกิจ เนื่องจากในการดำเนินธุรกิจ ผู้บริหารจำเป็นต้องพิจารณาสภาพแวดล้อมทั้งภายในและภายนอกกิจการ โดยที่ตัวอย่างของสภาพแวดล้อมภายในกิจการคือ คณะกรรมการของกิจการ ลูกจ้าง หรือวัฒนธรรมภายในองค์กร ในขณะที่สภาพแวดล้อมภายนอกกิจการจะประกอบไปด้วย อิทธิพลทางการเมือง อิทธิพลทางเศรษฐกิจ อิทธิพลทางเทคโนโลยี อิทธิพลทางสังคมและวัฒนธรรม อิทธิพลจากนานาประเทศ และอื่น ๆ (จินตนา บุญบงการ 2550) ดังนั้นเอกสารที่เกี่ยวข้องกับการตัดสินใจทางธุรกิจจะเป็นเอกสารที่มีความหลากหลาย ซึ่งจะเป็นการยากที่จะกำหนดกลุ่มของคำเฉพาะที่จะใช้ในการค้นหาเอกสารที่ต้องการ

ข้อจำกัดที่เกิดจากระบบค้นคืนข้อมูลส่วนใหญ่คือ ระบบจะค้นคืนเอกสาร (Document) ที่มีคำเหมือนกับคำในข้อสอบถาม (Query) เท่านั้น โดยไม่ได้พิจารณาถึงค่าความสำคัญของคำในเอกสารกับข้อสอบถาม กล่าวคือการเทียบความเหมือนของคำในเอกสารและข้อสอบถาม ระบบจะต้องเทียบความเหมือนของคำทุกคำในข้อสอบถามที่มีปรากฏในเอกสาร ตัวอย่างเช่น ถ้าผู้ใช้ระบุคำในข้อสอบถาม 4 คำ เอกสารที่จะถูกค้นคืนมาแสดงจะต้องมีคำทั้ง 4 คำนี้ปรากฏอยู่ในเอกสารด้วย ซึ่งถ้าเอกสารใดปรากฏคำในข้อสอบถามไม่ครบทั้ง 4 คำ เอกสารนั้นก็จะไม่ใช่ผลลัพธ์ที่จะถูกนำมาแสดงต่อผู้ใช้ นอกจากนี้ในขั้นตอนการค้นคืน การเทียบความเหมือนของคำในเอกสารและข้อสอบถาม ไม่ได้มีการพิจารณาถึงค่าความสำคัญของคำในเอกสาร จะพิจารณาเพียงการปรากฏหรือไม่ปรากฏของคำเท่านั้น เช่น เมื่อผู้ใช้กรอกคำว่า "Paper" ระบบจะดึงเอกสารทุกเอกสารที่มีคำว่า "Paper" ออกมา โดยบางเอกสารอาจจะมีคำว่า "Paper" ปรากฏอยู่เพียงครั้งเดียวในเอกสารเท่านั้น ซึ่งแสดงให้เห็นว่าคำว่า "Paper" ไม่ได้เป็นคำที่มีความสำคัญกับเอกสารมากนัก แต่ระบบก็ค้นคืนเอกสารนี้แสดงต่อผู้ใช้ นอกจากนี้การดำเนินการดังกล่าวยังมีข้อจำกัดในการควบคุมจำนวนเอกสาร เพราะเอกสารที่ได้จากการค้นคืนมีจำนวนมาก ทำให้ผู้ใช้เสียเวลาในการพิจารณาเอกสารที่ค้นคืน ซึ่งอาจส่งผลให้ผู้ใช้ไม่ได้ผลลัพธ์ที่ต้องการภายในระยะเวลาที่เหมาะสม

แบบจำลองการค้นคืนสารสนเทศ (Information Retrieval Model) หลายแบบจำลองได้ถูกสร้างขึ้นเพื่อลดข้อจำกัดในเรื่องการค้นหาเอกสารที่จำเป็นต้องมีคำทุกคำตามที่ระบุในข้อ

สอบถาม และ/หรือเพิ่มความสามารถในการค้นคืนเอกสารโดยใช้ความถี่ของคำที่ปรากฏในเอกสารประกอบการค้นหา โดยแบบจำลองการค้นคืนสารสนเทศที่ได้รับความนิยมมี 3 แบบ คือ

1. แบบจำลองบูลีน (Boolean Search Model) คือแบบจำลองที่อยู่บนพื้นฐานของแนวคิดเชิงตรรกะและพีชคณิตบูลีน (Boolean algebra) กับคำสำคัญที่ถูกรวมกันโดยตัวเชื่อมทางตรรกะพีชคณิต และ (AND) หรือ (OR) และไม่ (NOT) รวมกัน แบบจำลองบูลีนจะพิจารณาว่าคำหนึ่ง ๆ ในข้อสอบถามจะปรากฏหรือไม่ปรากฏในเอกสารเท่านั้น โดยไม่ได้คำนึงถึงค่าน้ำหนักของคำ และไม่มีการจัดลำดับของเอกสาร (Ranking) เมื่อค้นคืนเอกสารออกมา (วันเพ็ญ ชูเมือง 2547; Salton and McGill 1987; Chowdhury 2004)

2. แบบจำลองความน่าจะเป็น (Probabilistic Model) คือ แบบจำลองที่มีจุดประสงค์ในการหาความน่าจะเป็นที่เอกสารตรงกับข้อสอบถาม ซึ่งมีพื้นฐานการคำนวณของสมการความน่าจะเป็นของความเกี่ยวข้องระหว่างข้อสอบถามและเอกสาร เอกสารที่ค้นคืนออกมาต้องผ่านวิธีการคำนวณความน่าจะเป็นทางคณิตศาสตร์ โดยคำนวณความน่าจะเป็นจากข้อมูลความถี่ของคำ โดยข้อมูลความถี่ของคำจะสามารถนำมาใช้คะเนความน่าจะเป็นที่เอกสารจะเกี่ยวข้องกับข้อสอบถามนั้น (Shatkay and Wibur 2000; Chowdhury 2004)

3. แบบจำลองปริภูมิเวกเตอร์ (Vector Space Model) คือ แบบจำลองซึ่งกำหนดให้เอกสารแต่ละอันแสดงโดยเวกเตอร์หรือชุดของคำ (Terms) และมีการเปรียบเทียบความเหมือนกันของเอกสารและข้อสอบถาม (Query) แบบจำลองปริภูมิเวกเตอร์ให้ความสำคัญกับความถี่ของคำที่ปรากฏอยู่ในเอกสาร แบบจำลองนี้ใช้สมการคณิตศาสตร์ที่เรียบง่ายในการคิด โดยการพิจารณาจากความถี่ของคำที่ปรากฏในแต่ละเอกสาร และสามารถจัดลำดับของเอกสาร (Ranking) ลักษณะดังกล่าวทำให้แบบจำลองปริภูมิเวกเตอร์สามารถใช้กับเอกสารที่มีปริมาณคำมากๆได้ดี (วันเพ็ญ ชูเมือง 2547; ชูชาติ หฤไชยะศักดิ์ 2548; Greenwood 2002; Chowdhury 2004)

ในปัจจุบันนี้แบบจำลองปริภูมิเวกเตอร์เป็นแบบจำลองที่ได้รับความนิยมในการพัฒนาระบบค้นคืนสารสนเทศ เนื่องจากเป็นแบบจำลองที่มีการคำนวณทางคณิตศาสตร์ที่มีการพิจารณาถึงค่าน้ำหนักของคำ (Terms) ในเอกสารและข้อสอบถาม (Query) และมีการเปรียบเทียบความเหมือนของเอกสารและข้อสอบถามซึ่งเป็นแนวคิดพื้นฐานที่สำคัญสำหรับการค้นคืนเอกสาร (Document Retrieval) ดังนั้นผู้วิจัยจึงมีความสนใจที่จะนำแบบจำลองปริภูมิเวกเตอร์มาใช้เป็นเทคนิคในการพัฒนาระบบค้นคืนเอกสารในงานวิจัยนี้

กระบวนการที่สำคัญในแบบจำลองปริภูมิเวกเตอร์คือ การวัดความคล้ายคลึง (Measuring Similarity) ระหว่างเอกสารและข้อสอบถาม (Chowdhury 2004) ซึ่งเป็นการจับคู่

พิจารณาค่าความเหมือนของข้อสอบถามและเอกสารทั้งหมด เอกสารใดที่มีค่าความเหมือนกับข้อสอบถามตามที่กำหนดไว้ก็ถูกดึงออกมาแสดงต่อผู้ใช้ ในทางตรงกันข้ามเอกสารใดที่มีค่าความเหมือนกับข้อสอบถามนอกเหนือจากค่าความเหมือนที่กำหนดไว้ก็จะไม่แสดงเอกสารนั้นต่อผู้ใช้ กระบวนการวัดความคล้ายคลึงระหว่างเอกสารส่งผลให้ระบบค้นคืนเอกสารในปัจจุบันมีความสามารถเพิ่มขึ้นในด้านการค้นหาเอกสารที่คล้ายคลึงกับเอกสารตัวอย่าง (Querying by Example) หรือการจัดกลุ่มเอกสาร (Retrieved Document Clustering) ที่ได้จากการสืบค้น เป็นต้น (พิลาวัณย์ พลับรู้การ และกฤษณะ ไวยมัย 2545)

จากงานวิจัยที่ผ่านมาพบว่าวิธีวัดความคล้ายคลึงหรือความเหมือนระหว่างเอกสารและข้อสอบถามที่ใช้กันกว้างขวางเรียกว่า Cosine Similarity หรือ Cosine Angle (Chowdhury 2004) ในงานวิจัยนี้ผู้วิจัยขอเรียกว่า วิธีวัดความคล้ายคลึงเชิงมุม ซึ่งเป็นวิธีการหาความเหมือนระหว่างเอกสารและข้อสอบถามที่ใช้ในแบบจำลองปริภูมิเวกเตอร์ โดยการแทนเอกสารและข้อสอบถามด้วยระบบเวกเตอร์และวัดมุมที่กระทำต่อกันระหว่างเอกสารและข้อสอบถาม วิธีวัดความคล้ายคลึงเชิงมุม (Cosine Angle) จะกำหนดค่าความคล้ายคลึง โดยถ้าเวกเตอร์เอกสารใดทำมุมกับเวกเตอร์ข้อสอบถามน้อยแสดงว่าเอกสารนั้นมีความสำคัญกับข้อสอบถามมาก ในทางตรงกันข้ามถ้าเวกเตอร์เอกสารใดทำมุมกับเวกเตอร์ข้อสอบถามมากแสดงว่าเอกสารนั้นมีความสำคัญกับข้อสอบถามน้อย ส่วนวิธีการวัดความคล้ายคลึงอื่นๆ เช่น วิธีวัดความเหมือนวิธีเจคคาร์ด (Jaccard coefficient) (Salton and McGill 1987) หรือวิธีวัดความเหมือนวิธีไดซ์ (Dice coefficient) (Salton and McGill 1987) ก็มีใช้อยู่เช่นกัน แต่วิธีการวัดความคล้ายคลึงเชิงมุม (Cosine Angle) จะให้ประสิทธิภาพการค้นคืนที่ดีกว่า (Chowdhury 2004; Salton and McGill 1987) วิธีการวัดค่าความคล้ายคลึงอีกวิธีหนึ่งที่สามารถนำมาใช้ในการค้นคืนเอกสารด้วยเทคนิคแบบจำลองปริภูมิเวกเตอร์ คือ วิธีการวัดความคล้ายคลึงระหว่างเวกเตอร์ด้วยการวัดระยะทางแบบยูคลิเดียน (Euclidean Distance) ในงานวิจัยนี้ผู้วิจัยขอเรียกว่า วิธีวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน (Euclidean Distance) ซึ่งเป็นวิธีการวัดความคล้ายคลึงระหว่างเวกเตอร์ที่นิยมใช้ในระบบการค้นคืนรูปภาพ (Image Retrieval) (Baeza-Yates and Ribeiro-Neto 1999)

งานวิจัยที่ผ่านมาเกี่ยวกับการใช้วิธีวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน (Euclidean Distance) โดยส่วนใหญ่จะเป็นงานวิจัยที่ใช้การวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนในระบบการค้นคืนรูปภาพ (Image Retrieval) ตัวอย่างเช่น ในงานวิจัยของ Qian และคณะในปี 2004 (Qian et al. 2004) เปรียบเทียบการวัดความเหมือนในแบบจำลองเวกเตอร์ (Vector model) ด้วยการใช้วิธีวัดระยะห่างยูคลิเดียน (Euclidean distance) และวิธีวัดระยะห่างเชิงมุม

(Cosine angle distance) ในชุดข้อมูลทดสอบที่เป็นรูปภาพ ด้วยการใช้การวิเคราะห์ทางทฤษฎี และการทดลองในระบบการค้นคืนรูปภาพ (Image Retrieval) Qian และคณะศึกษาว่า ในการค้นคืนรูปภาพนั้นนิยมที่จะใช้การวัดความเหมือนระหว่างรูปภาพด้วยวิธีระยะห่างยูคลิเดียน (Euclidean distance) ซึ่งหากนำวิธีวัดระยะห่างเชิงมุม (Cosine angle distance) มาประยุกต์ใช้ ผลการทดลองที่ออกมาจะสามารถหาค่าความเหมือนได้ใกล้เคียงกับวิธีระยะห่างยูคลิเดียนหรือไม่ Qian และคณะ ได้ทดลองด้วยการใช้ฐานข้อมูลรูปภาพ ในการทดลองนี้ประสิทธิภาพของวิธีวัดระยะห่างเชิงมุม (Cosine angle distance) ถูกวัดด้วยค่าความแม่นยำและค่าความระลึก ผลการทดลองพบว่าการค้นคืนรูปภาพด้วยการหาค่าความเหมือนเมื่อใช้การวัดระยะห่างเชิงมุม (Cosine angle distance) นั้น ได้ประสิทธิภาพค่าความแม่นยำและค่าความระลึกออกมาไม่แตกต่างไปจากการใช้วิธีวัดระยะห่างยูคลิเดียน (Euclidean distance) Qian และคณะได้เสนอว่าวิธีการวัดระยะห่างเชิงมุม (Cosine angle distance) สามารถนำมาใช้กับระบบการค้นคืนรูปภาพได้ไม่แตกต่างไปจากวิธีการวัดระยะห่างยูคลิเดียน (Euclidean distance) ซึ่งเป็นวิธีที่ใช้ในระบบการค้นคืนรูปภาพอยู่แล้ว

แม้ว่างานวิจัยที่ผ่านมาจะใช้การวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนในระบบการค้นคืนรูปภาพ แต่ยังไม่มีงานวิจัยใดที่นำวิธีวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน (Euclidean distance) มาใช้ในระบบการค้นคืนเอกสารในเทคนิคแบบจำลองปริภูมิเวกเตอร์ (Vector Space Model) เพื่อค้นคืนเอกสารที่เกี่ยวข้องตรงตามความต้องการของผู้ใช้ ซึ่งสามารถกระทำได้โดยการหาความเหมือนระหว่างเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถามด้วยการคำนวณค่าระยะทางระหว่างข้อสอบถามและเอกสาร โดยถ้าเวกเตอร์เอกสารใดมีค่าระยะห่างจากเวกเตอร์ข้อสอบถามน้อยแสดงว่าเอกสารนั้นมีความสัมพันธ์กับข้อสอบถามมาก ในทางตรงกันข้ามถ้าเวกเตอร์เอกสารใดมีค่าระยะห่างจากเวกเตอร์ข้อสอบถามมากแสดงว่าเอกสารนั้นมีความสัมพันธ์กับข้อสอบถามน้อย

ดังนั้นผู้วิจัยซึ่งมีความสนใจที่จะศึกษาว่าวิธีการวัดความคล้ายคลึงหรือความเหมือนระหว่างเอกสารและข้อสอบถามด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน (Euclidean Distance) เป็นวิธีการวัดที่มีประสิทธิภาพในการค้นคืนเอกสารได้ใกล้เคียงหรือดีกว่าวิธีการวัดความคล้ายคลึงเชิงมุม (Cosine Angle) หรือไม่ เนื่องจากวิธีวัดความคล้ายคลึงทั้ง 2 วิธีใช้แนวคิดเดียวกันในการคำนวณหาค่าระยะทางระหว่างเวกเตอร์ แต่ต่างกันตรงที่วิธีการวัดความคล้ายคลึงโดยวิธีการวัดความคล้ายคลึงเชิงมุมจะหาระยะห่างระหว่างเวกเตอร์ด้วยการคิดค่าเชิงมุมระหว่าง

เวกเตอร์ 2 เวกเตอร์ แต่วิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนจะหาระยะห่างระหว่างเวกเตอร์ด้วยการคิดค่าระยะทางระหว่างเวกเตอร์ 2 เวกเตอร์

อย่างไรก็ตาม Hand (2001) ได้กล่าวว่า ในการวัดความคล้ายคลึงระหว่างเอกสาร วิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน จะให้ค่าประสิทธิภาพที่แย่กว่าวิธีการวัดความคล้ายคลึงเชิงมุม ผู้วิจัยจึงสนใจว่า หากเพิ่มความสามารถให้กับการค้นคืนเอกสารที่ใช้วิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน จะสามารถให้ประสิทธิภาพที่มีความสามารถเทียบเท่ากับวิธีการค้นคืนเอกสารที่ใช้วิธีการวัดความคล้ายคลึงเชิงมุมได้หรือไม่ ทั้งนี้เนื่องจาก ศิรัตน์ ศิรนานนท์ (2549) ได้กล่าวว่า การค้นคืนเอกสารที่ใช้เทคนิคแบบจำลองปริภูมิเวกเตอร์ด้วยการวัดความคล้ายคลึงเชิงมุมนั้น โดยส่วนใหญ่มีการแสดงเอกสารต่อผู้ใช้โดยการกำหนดเงื่อนไข (threshold) หรือกรอบค่าความคล้ายคลึงในการเลือกเอกสารที่เป็นคำตอบด้วยค่าที่คงที่สำหรับทุกข้อสอบถาม ดังนั้นผู้วิจัยจึงสนใจที่จะศึกษาว่า หากเพิ่มความสามารถให้กับการค้นคืนเอกสารที่ใช้วิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน ให้สามารถค้นคืนเอกสารภายใต้กรอบค่าความคล้ายคลึงที่เปลี่ยนแปลงไปตามแต่ละกลุ่มของเอกสารที่มีความใกล้เคียงกับข้อสอบถาม จะสามารถให้ประสิทธิภาพการค้นคืนได้ใกล้เคียงหรือดีกว่าการค้นคืนเอกสารที่ใช้วิธีการวัดความคล้ายคลึงเชิงมุมหรือไม่

ในการศึกษานี้ ผู้วิจัยจะกำหนดกรอบค่าความคล้ายคลึงในการค้นคืนเอกสารด้วยการประยุกต์ใช้เทคนิคการจัดกลุ่มข้อมูล (Clustering) แบบ K-mean Clustering ซึ่งเป็นวิธีการจัดกลุ่มข้อมูลที่นิยมใช้วิธีการวัดความคล้ายคลึงระหว่างเอกสารด้วยวิธีการวัดระยะห่างยูคลิเดียน (K_means Clustering Algorithm 2006) ซึ่งถ้าเอกสารใดที่มีระยะห่างกับข้อสอบถามภายใต้กรอบที่กำหนดจะถูกค้นคืนออกมาแสดงต่อผู้ใช้ และในทางตรงกันข้ามถ้าเอกสารใดมีระยะห่างกับข้อสอบถามนอกเหนือกรอบที่กำหนด เอกสารนั้นก็จะไม่ถูกค้นคืนเพื่อแสดงต่อผู้ใช้ ตัวอย่างเช่น ถ้าผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูลระบุว่า สำหรับข้อสอบถามที่กำหนดควรจะใช้กรอบค่าความคล้ายคลึงเท่ากับ 3 หน่วย เอกสารที่จะถูกค้นคืนมาแสดงต่อผู้ใช้ จะต้องเป็นเอกสารที่มีระยะห่างกับข้อสอบถามเท่ากับหรือน้อยกว่า 3 หน่วยเท่านั้น ซึ่งเอกสารใดที่มีระยะห่างกับข้อสอบถามมากกว่า 3 หน่วยก็จะไม่ถูกนำมาแสดงต่อผู้ใช้

ด้วยเหตุนี้งานวิจัยนี้จึงเป็นการศึกษาเปรียบเทียบประสิทธิภาพการค้นคืนเอกสารโดยใช้เทคนิคปริภูมิเวกเตอร์ (Vector Space Model) ด้วยการวัดความคล้ายคลึงเชิงมุม (Cosine Angle) และการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน (Euclidean Distance) ด้วยการค้นคืนเอกสารต่อผู้ใช้ภายในกรอบค่าความคล้ายคลึงที่กำหนดโดยการประยุกต์ใช้เทคนิคการจัดกลุ่ม

ข้อมูล (Clustering) ว่ามีความแตกต่างกันหรือไม่ ถ้าแตกต่างกันวิธีการวัดความคล้ายคลึงวิธีใดจะให้ประสิทธิภาพการค้นคืนที่ดีกว่ากับชุดเอกสารและข้อสอบถามของคำศัพท์เช่นเดียวกับเอกสารที่ใช้ประกอบการตัดสินใจทางธุรกิจ โดยในการศึกษาครั้งนี้ ผู้วิจัยจะวัดประสิทธิภาพของการค้นคืนด้วยวิธีการวัดค่าความแม่นยำ (Precision) การวัดค่าความระลึก (Recall) และการวัดค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) ซึ่งเป็นวิธีการวัดประสิทธิภาพระบบการค้นคืนเอกสารที่นิยมใช้ (Goutte and Gaussier 2005)

1.2 วัตถุประสงค์ของการวิจัย

1. ศึกษาการประยุกต์ใช้เทคนิคการจัดกลุ่มข้อมูล (Clustering) ในการกำหนดกรอบของค่าความคล้ายคลึง สำหรับเทคนิคปริภูมิเวกเตอร์ที่วัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน
2. เปรียบเทียบประสิทธิภาพของการค้นคืนเอกสาร (Document Retrieval) ที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยการวัดความคล้ายคลึงเชิงมุม กับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering)

1.3 ขั้นตอนโดยสรุปของการทำวิจัย

1. ศึกษารายละเอียดเกี่ยวกับวิธีการของระบบการค้นคืนสารสนเทศที่มีอยู่ในปัจจุบัน
2. ศึกษาแบบจำลองระบบการค้นคืนสารสนเทศในทุกรูปแบบ และศึกษารายละเอียดการค้นคืนเอกสาร (Document Retrieval) ด้วยแบบจำลองปริภูมิเวกเตอร์ (Vector Space Model)
3. ศึกษาเกี่ยวกับวิธีการวัดความคล้ายคลึงระหว่างเอกสาร (Measuring Similarity) ด้วยเทคนิคความคล้ายคลึงเชิงมุมและความคล้ายคลึงเชิงระยะทาง
4. ศึกษารายละเอียดเกี่ยวกับการทำเหมืองข้อมูล (Data Mining) ในเทคนิคการจัดกลุ่มข้อมูล (Clustering)
5. ออกแบบเครื่องมือทดสอบตามที่ได้ศึกษา
6. พัฒนาเครื่องมือทดสอบตามที่ได้ออกแบบ
7. ทดสอบการทำงานของเครื่องมือที่ได้พัฒนา
8. ทำการทดลองโดยใช้เครื่องมือทดสอบที่ได้พัฒนาขึ้น
9. วิเคราะห์ผลการทดลอง และสำรวจข้อมูลเพิ่มเติมจากผลการทดลอง
10. สรุปผลการทดลองและจัดทำเอกสารสรุปงานวิจัย พร้อมข้อเสนอแนะ

1.4 ตัวแปรที่ศึกษา

1. ตัวแปรต้น (Independent Variables)

ระบบค้นคืนเอกสารด้วยเทคนิคปริภูมิเวกเตอร์ในงานวิจัยนี้ สนใจวิธีการวัดความคล้ายคลึงระหว่างเอกสารสองวิธีการด้วยกัน คือ

1) การวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน (Euclidean Distance) ด้วยการค้นคืนเอกสารต่อผู้ใช้ภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering)

2) การวัดความคล้ายคลึงเชิงมุม (Cosine Angle) คือ วิธีการวัดความคล้ายคลึงระหว่างเอกสารและข้อสอบถามด้วยเทคนิคการวัดเชิงมุม (Cosine Coefficient)

2. ตัวแปรตาม (Dependent Variables)

ดังที่ได้กล่าวมาแล้วว่าวัตถุประสงค์ของการค้นคืนเอกสารคือ ระบบสามารถค้นคืนเอกสารที่ผู้ใช้ต้องการได้มากที่สุดและค้นคืนเอกสารที่ไม่ตรงกับความต้องการของผู้ใช้น้อยที่สุด โดยในงานวิจัยนี้ประสิทธิภาพของการค้นคืนจะถูกวัดด้วย

1) ค่าความแม่นยำ (Precision) คือ ค่าสัดส่วนของจำนวนเอกสารที่ถูกค้นคืนแล้วตรงความต้องการเทียบกับจำนวนเอกสารที่ถูกค้นคืนทั้งหมด ซึ่งเป็นค่าที่จะแสดงให้เห็นว่าเอกสารที่ถูกค้นคืนตรงกับความต้องการมากน้อยเพียงใด

2) ค่าความระลึก (Recall) คือ ค่าสัดส่วนของจำนวนเอกสารที่ถูกค้นคืนแล้วตรงความต้องการเทียบกับจำนวนเอกสารที่ตรงความต้องการทั้งหมด ซึ่งเป็นค่าที่จะแสดงความครอบคลุมของการค้นคืน

3) ค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) คือ ค่าการวัดที่รวมทั้งค่าความระลึกและค่าความแม่นยำ ซึ่งเป็นค่าที่แสดงให้เห็นถึงความถูกต้องและความครอบคลุมของการค้นคืนเอกสาร

3. ตัวแปรควบคุม

3.1 เอกสาร คือ ชุดเอกสารในฐานข้อมูลของระบบค้นคืนเอกสารที่นำมาทดสอบ

3.2 ข้อสอบถาม คือ คำหรือประโยคที่ผู้วิจัยกรอกเข้ามายังระบบเพื่อค้นคืนเอกสารที่เกี่ยวเนื่อง โดยมีกำหนดอยู่แล้ว ซึ่งมาพร้อมกับชุดข้อมูลทดสอบ

3.3 ผลเฉลย คือ ชุดผลเฉลยที่มีการกำหนดอยู่แล้วในชุดข้อมูลทดสอบ ซึ่งเป็นการบอกถึงผลเฉลยของแต่ละข้อสอบถาม ว่ามีเอกสารใดบ้างที่เกี่ยวข้องกับข้อสอบถามนั้น

3.4 เครื่องมือสร้างระบบค้นคืนเอกสาร คือ คอมพิวเตอร์ ฐานข้อมูล และภาษาที่ใช้เขียน

1.5 ขอบเขตของการวิจัย

1. การทดสอบประสิทธิภาพการค้นคืนเอกสารนี้จะไม่มีการคำนึงถึงความหมายของคำโดยจะถือว่าคำแต่ละคำเป็นอิสระต่อกัน ซึ่งในความเป็นจริงแล้วคำแต่ละคำอาจมีความใกล้เคียงทางความหมายต่อกัน เช่น คำว่า Cluster analysis จะมีความหมายในทางเดียวกันกับคำว่า Clustering เป็นต้น
2. การทดสอบประสิทธิภาพการค้นคืนเอกสารนี้จะไม่ครอบคลุมถึงคำกำกวม (Ambiguity) เช่น คำว่า "Apple" ซึ่งสามารถแปลความหมายได้ทั้งเป็นชื่อสินค้าคอมพิวเตอร์ หรือผลไม้
3. เอกสารที่จะนำมาใช้ในการทดลองจะเป็นบทความภาษาอังกฤษเท่านั้น
4. เครื่องมือทดสอบเทคนิคการค้นคืนเอกสารนี้สร้างขึ้นเพื่อทดสอบกับชุดเอกสารและชุดข้อสอบถามของไทม์ (ftp://ftp.cs.cornell.edu/pub/smart/time/) เท่านั้น

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1. ผลการทดสอบประสิทธิภาพระบบค้นคืนเอกสารที่ได้ สามารถเป็นแนวทางในการนำเทคนิคการค้นคืนเอกสารโดยเทคนิคปริภูมิเวกเตอร์ด้วยการวัดความคล้ายคลึงเชิงมุมหรือการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนที่เสนอในงานวิจัยนี้ ไปเลือกใช้ได้อย่างเหมาะสม
2. ผู้ที่ต้องการพัฒนาระบบค้นคืนเอกสารสามารถนำเทคนิคการค้นคืนเอกสารภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering) ตามที่เสนอในงานวิจัยนี้ ไปเป็นแนวทางสำหรับประยุกต์ใช้เข้ากับระบบค้นคืนเอกสารที่ต้องการพัฒนา
3. ผู้ที่ต้องการพัฒนาระบบค้นคืนเอกสารสามารถนำวิธีการวัดความคล้ายคลึงระหว่างเอกสารซึ่งเป็นผลจากงานวิจัยนี้ ไปประยุกต์ใช้กับชุดเอกสารทางธุรกิจ เพื่อสามารถค้นคืนเอกสารทางธุรกิจออกมาให้ตรงกับความต้องการ