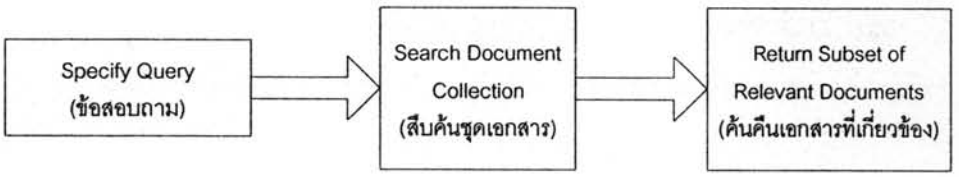
บทที่ 2

วรรณกรรมที่เกี่ยวข้อง

ในบทนี้ ผู้วิจัยจะนำเสนอถึงผลการศึกษาวรรณกรรมที่ผ่านมาในอดีตที่นำมาประยุกต์ใช้ในงานวิจัยและข้อจำกัดของงานวิจัยในอดีต ที่เกี่ยวข้องกับการค้นคืนเอกสาร จะกล่าวถึงทฤษฎีที่สำคัญต่าง ๆ โดยจุดประสงค์หลักของงานวิจัยนี้ต้องการที่จะเพิ่มประสิทธิภาพระบบค้นคืนเอกสารให้ตรงกับความต้องการของผู้ใช้มากขึ้น โดยเปรียบเทียบประสิทธิภาพการค้นคืนด้วยวิธีการวัดความคล้ายคลึงกันของเอกสารและข้อสอบถามด้วยวิธีการวัดเชิงมุม และวิธีการวัดเชิงระยะทาง ดังนั้นในงานวิจัยนี้จึงมีทฤษฎีที่เกี่ยวข้องดังต่อไปนี้

2.1 เทคนิคการค้นคืนสารสนเทศ

ระบบการค้นคืนสารสนเทศ เป็นระบบที่ออกแบบ เพื่อจุดประสงค์ในการตอบสนองต่อผู้ต้องการใช้สารสนเทศ จากเอกสารที่ได้เก็บรวบรวมไว้ ผู้ต้องการสารสนเทศจะใช้ข้อสอบถาม เป็นเครื่องมือเพื่อแสดงความต้องการสารสนเทศ ระบบจะต้องสามารถแสดงถึงเอกสารที่มีสารสนเทศตรงตามความต้องการของผู้ใช้ ผลลัพธ์ของระบบค้นคืนเอกสารที่มีประสิทธิภาพนั้น จะต้องประกอบไปด้วยเอกสารที่ผู้ใช้ต้องการในปริมาณที่มากที่สุด และเอกสารที่ผู้ใช้ไม่ต้องการในปริมาณที่น้อยที่สุด (Baeza-Yates and Ribeiro-Neto 1999: Moulanont 1992) วัตถุประสงค์ของการค้นคืนสารสนเทศสามารถแสดงได้ดังรูปที่ 2.1 (Weiss et al. 2005)

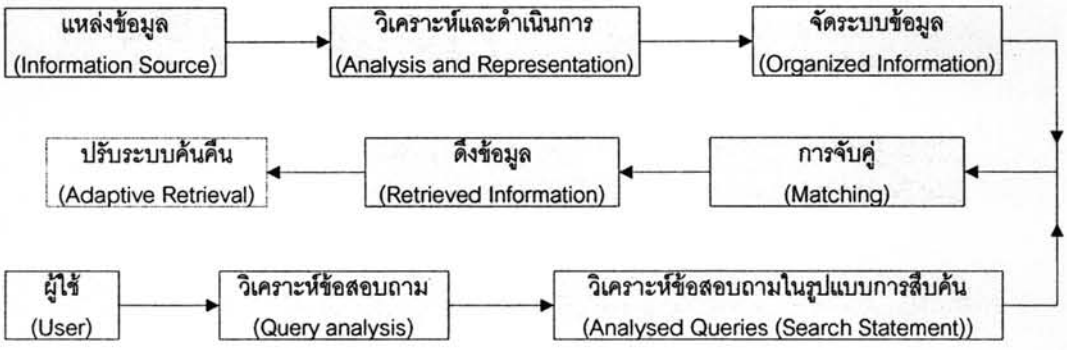


รูปที่ 2.1 รูปแสดงเป้าหมายหลักในการค้นคืนสารสนเทศ (Weiss et al. 2005)

กระบวนการพื้นฐานของเทคนิคการค้นคืนสารสนเทศทั่วไป สามารถแสดงได้ดังรูปที่ 2.2 (Chowdhury 2004) โดยมีกระบวนการดังนี้

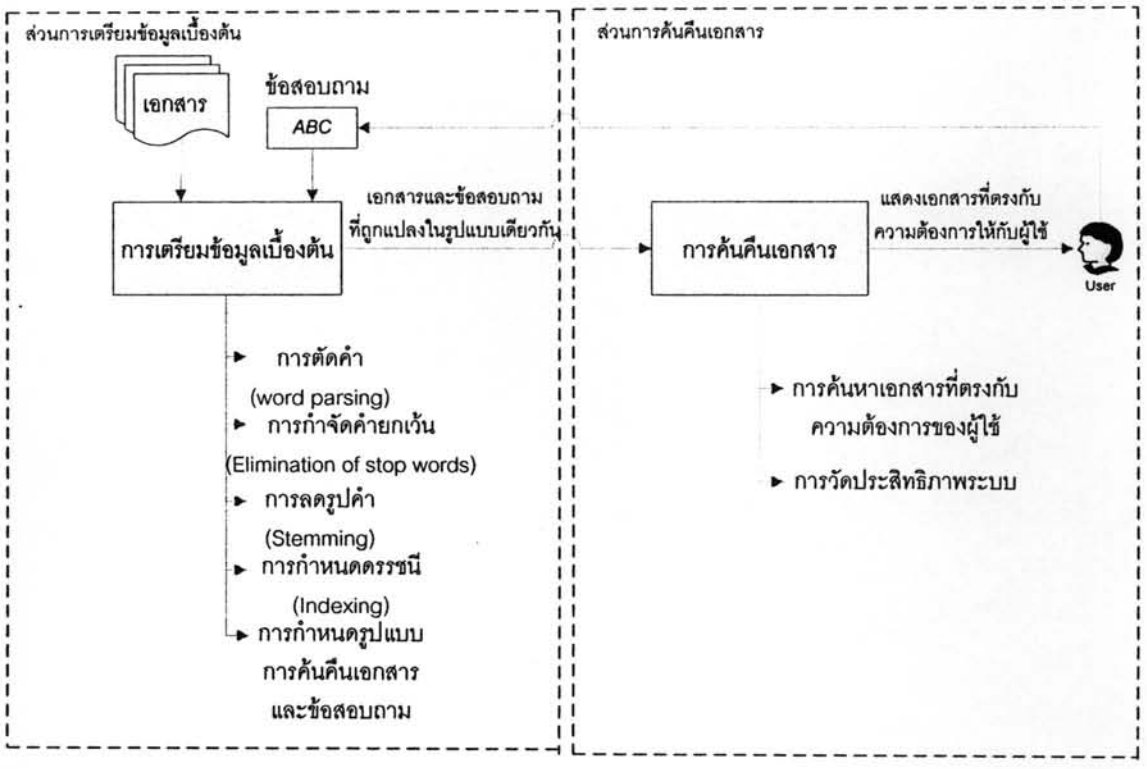
1. กำหนดข้อมูลเพื่อระบุขอบเขตชุดเอกสารที่ผู้ใช้ต้องการค้นหา และเก็บชุดเอกสารนี้เข้าสู่ระบบ

2. วิเคราะห์ความหมายของข้อมูลชุดเอกสารนั้น ๆ
3. แปลงชุดเอกสารในฐานข้อมูลให้อยู่ในรูปแบบที่สามารถเปรียบเทียบกับข้อสอบถามของผู้ใช้
4. วิเคราะห์ข้อสอบถามของผู้ใช้และแสดงข้อสอบถามในรูปแบบที่เหมาะสมกับชุดเอกสารในฐานข้อมูล
5. เปรียบเทียบข้อสอบถามกับชุดเอกสารในฐานข้อมูล
6. ค้นคืนเอกสารที่ตรงความต้องการ
7. ปรับระบบการค้นคืนโดยใช้ผลสะท้อนกลับของผู้ใช้ (ขั้นตอนนี้อาจจะมีหรือไม่มีก็ได้)



รูปที่ 2.2 รูปแสดงกระบวนการพื้นฐานระบบการค้นคืนสารสนเทศ

จากกระบวนการพื้นฐานของเทคนิคการค้นคืนสารสนเทศทั้ง 7 ขั้นตอน ซึ่งแต่ละขั้นตอนมีกระบวนการทำงานที่แตกต่างกันนั้น สามารถที่จะแบ่งการทำงานออกเป็น 2 ส่วนด้วยกันคือ ส่วนของการเตรียมข้อมูลเบื้องต้น และส่วนของการค้นคืนเอกสาร (ชูชาติ หฤไชยะศักดิ์ 2548) ดังรูปที่ 2.3



รูปที่ 2.3 รูปแสดงการทำงานทั้งหมดของระบบการค้นคืนเอกสาร

การทำงานส่วนการเตรียมข้อมูลเบื้องต้นเป็นการทำงานในส่วนของการเตรียมชุดข้อมูลเอกสารและข้อสอบถามให้อยู่ในรูปแบบที่เหมาะสมกับการทำงานในส่วนการค้นคืน การทำงานในส่วนการเตรียมข้อมูลเบื้องต้น มีกระบวนการทำงานดังนี้

1. การสกัดคำสำคัญออกจากเอกสาร ประกอบด้วยขั้นตอน การตัดคำ การกำจัดคำยกเว้น และการลดรูปคำ
2. การกำหนดดรรชนี
3. การกำหนดรูปแบบการค้นคืนเอกสารและข้อสอบถาม

การทำงานส่วนการค้นคืนเอกสาร เป็นการทำงานในส่วนของการค้นคืนเอกสารที่เกี่ยวข้องกับข้อสอบถามที่ผู้ใช้ระบุ และแสดงเอกสารที่เกี่ยวข้องนั้นให้กับผู้ใช้ ซึ่งการทำงานในส่วนนี้มีกระบวนการทำงานดังนี้

1. การค้นหาเอกสารที่ตรงกับความต้องการของผู้ใช้ เป็นส่วนของการเปรียบเทียบความเหมือนระหว่างเอกสารและข้อสอบถาม พร้อมแสดงผลเอกสารที่เกี่ยวข้องกับข้อสอบถามให้กับผู้ใช้
2. การวัดประสิทธิภาพระบบการค้นคืนเอกสาร

จากองค์ประกอบการทำงานทั้ง 2 ส่วนของการค้นคืนเอกสาร ผู้วิจัยได้กล่าวถึงกระบวนการทำงานในแต่ละขั้นตอนของส่วนการเตรียมข้อมูลเบื้องต้น และส่วนของการค้นคืนเอกสารแล้ว ดังนั้นผู้วิจัยขออธิบายขั้นตอนการทำงานทั้ง 2 ส่วนโดยละเอียดในลำดับต่อไป

2.2 การสกัดคำสำคัญออกจากเอกสาร

เป็นขั้นตอนการเตรียมข้อมูลเบื้องต้น เมื่อได้ชุดเอกสารมาแล้ว จะแปลงแต่ละเอกสารให้อยู่ในรูปแบบที่ระบบสามารถทำงานได้ โดยมีขั้นตอนดังต่อไปนี้

2.2.1 การตัดคำ (word parsing)

เป็นขั้นตอนการแยกคำเป็นคำเดี่ยวๆ ออกมาจากเอกสาร โดยอาศัยเครื่องหมายวรรคตอนต่างๆ เป็นตัวช่วยแบ่งคำ เช่น ข้อมูลในเอกสาร "Developing Computer-Based Information Centers" จะสามารถแบ่งคำออกได้เป็น 4 คำ คือ Centers, Computer-Based, Developing และ Information เป็นต้น

2.2.2 การกำจัดคำยกเว้น (Elimination of stop words) (Baeza-Yates and Ribeiro-Neto 1999)

คำยกเว้น เป็นคำที่มักปรากฏขึ้นบ่อยมาก ๆ ในเอกสาร แต่ไม่มีความหมายที่จะเป็นประโยชน์ต่อการสืบค้นมากนัก และไม่สามารถใช้ในการแยกแยะเอกสารได้ เช่น คำว่า "the" , "of", "a" , "and" และอื่น ๆ ซึ่งถือเป็นคำทั่วไปสามารถตัดทิ้งได้ การตัดคำที่เป็นคำยกเว้นนั้นเป็นการลดคำศัพท์ที่เก็บในระบบ งานวิจัยนี้กำหนดใช้รายการของคำที่เป็นคำยกเว้น 571 คำจากสมาร์ท (SMART) (ftp://ftp.cs.cornell.edu/pub/smart/english.stop) ซึ่งเป็นรายการคำยกเว้นที่ถูกพัฒนาโดย Cornell University ตั้งแต่ปี 1960 สำหรับให้นักวิจัยได้นำไปทดลองใช้ในระบบการค้นคืนเอกสารที่พัฒนาขึ้น (Salton 1971) โดยรายละเอียดของสมาร์ทสามารถแสดงได้ดังภาคผนวก ข

2.2.3 การลดรูปคำ (Stemming)

เป็นขั้นตอนการวิเคราะห์หารากศัพท์ของคำ (Stemming) เพื่อเพิ่มประสิทธิภาพในการสืบค้นและลดขนาดของดรรชนี โดยทั่วไปคำที่ปรากฏในเอกสาร มักจะเป็นคำที่มีรูปแบบต่าง ๆ ที่คล้ายคลึงกัน เช่น คำนามที่อยู่ในรูปแบบของพหูพจน์ คำกริยาที่ตามหลังด้วย "ing" คำกริยาที่อยู่ในรูปของอดีต เป็นต้น คำที่อยู่ในรูปแบบที่หลากหลายเหล่านี้จะถูกปรับเปลี่ยนให้อยู่ในรูปแบบเดียวกัน ด้วยวิธีการลดรูปคำ (Stemming) ให้อยู่ในรากศัพท์เดียวกัน โดยการตัดส่วนที่เป็น prefix หรือ suffix ออก เช่น เซตของคำ {computer, computing, computed, compute} เมื่อผ่านกระบวนการนี้แล้ว จะได้เซตของคำ คือ {compute} เป็นต้น (ศิรัตน์ ศิรนานนท์ 2549; Strasberg

et al. 2000) โดยทั่วไปพบว่าขั้นตอนวิธีการลดรูปคำที่เป็นที่นิยมคือ ขั้นตอนวิธีของพอร์ทเตอร์ (Porter algorithm) (Porter 1980) เนื่องจากเป็นวิธีที่เรียบง่ายและได้ผลดี ขั้นตอนวิธีของพอร์ทเตอร์นั้นสามารถช่วยให้ผลลัพธ์การลดรูปคำ (Stemming) ที่มีประสิทธิภาพและสามารถทำงานได้อย่างรวดเร็วอีกด้วย (Baeza-Yates and Ribeiro-Neto 1999) ขั้นตอนวิธีของพอร์ทเตอร์แสดงได้ดังภาคผนวก ค

2.3 การกำหนดดรรชนี (Indexing) (Baeza-Yates and Ribeiro-Neto 1999)

เป็นวิธีการในการจัดทำดรรชนีของคำสำคัญที่พบภายในเอกสาร เมื่อชุดข้อมูลมีขนาดใหญ่มาก ๆ หรือมีเอกสารจำนวนมาก จำเป็นต้องเก็บดรรชนี เพื่อช่วยให้การค้นหามีความรวดเร็วยิ่งขึ้น ดรรชนีของคำสำคัญที่สกัดได้จากเอกสารจะถูกเก็บรวบรวมไว้เป็นฐานข้อมูลขนาดใหญ่เพื่อจัดเตรียมไว้สำหรับการสืบค้น ซึ่งรูปแบบของดรรชนีนั้นมีหลากหลายรูปแบบ เช่น แฟ้มซิกเนเจอร์ (Signature file), ต้นไม้ซัฟฟิก (Suffix tree) และ แฟ้มผกผัน (Inverted file) เป็นต้น ซึ่งเทคนิคการกำหนดดรรชนีที่ได้รับความนิยมนำมากำหนดดรรชนีมากที่สุด คือ เทคนิคแฟ้มผกผัน (Inverted file) ในที่นี้จะขอกล่าวถึงเทคนิคแฟ้มผกผัน ดังต่อไปนี้

2.3.1 แฟ้มผกผัน (Inverted file)

เป็นโครงสร้างที่มีหลักการ คือจะเก็บเอกสารและตำแหน่งของคำที่ปรากฏในเอกสารนั้นด้วย โดยจะประกอบไปด้วยสองส่วน คือส่วนตารางเก็บเอกสาร และส่วนตารางคำศัพท์ที่เก็บคำศัพท์ที่ปรากฏในเอกสารและตำแหน่งของคำนั้นในเอกสาร การระบุตำแหน่งของคำจะคิดจากลำดับของอักขระ โดยคำศัพท์แรกที่ปรากฏในข้อความให้ค่าตำแหน่งอักขระเป็น 1 และคำศัพท์ต่อไปก็จะเรียงตามลำดับการปรากฏของอักขระนั้น ๆ ในเอกสาร (Baeza-Yates and Ribeiro-Neto 1999) ตัวอย่างเช่น

เอกสารหนึ่งมีข้อความปรากฏอยู่ คือ "This is a text. A text has many words." สามารถแสดงเป็นแฟ้มผกผันได้ ดังรูปที่ 2.4

Text

ตำแหน่ง	1	6	9	11	17	19	24	28	33
คำศัพท์	This	is	a	text.	A	text	has	many	words.

Vocabulary	Co-occurrence
many	28
text	11 , 19
words	33

รูปที่ 2.4 รูปแสดงการสร้างแฟ้มผกผัน

จากข้อความที่ปรากฏอยู่ในเอกสารตัวอย่าง สามารถมาสร้างเป็นแฟ้มผกผันได้ดังรูปที่2.4 ซึ่งประกอบไปด้วยตารางเก็บเอกสารที่จะระบุคำในเอกสารและตำแหน่งของแต่ละคำในเอกสาร และตารางคำศัพท์ ซึ่งก่อนจะสร้างตารางคำศัพท์นั้นจะต้องนำข้อความในเอกสารผ่านขั้นตอนของการตัดคำ, การกำจัดคำยกเว้น และการลดรูปคำ ดังนั้นจากเอกสารตัวอย่างจะได้คำสำคัญ 3 คำ คือ “many”, "text" และ “words” ตารางคำศัพท์จะเป็นตารางที่เก็บคำศัพท์ที่ปรากฏในเอกสาร และตำแหน่งของคำสำคัญนั้นในเอกสาร จากรูปข้างต้นจะพบว่าคำสำคัญ “text” จะปรากฏอยู่ในเอกสาร 2 ครั้งด้วยกันและอยู่ในคนละตำแหน่ง ดังนั้นเมื่อสร้างเป็นตารางคำศัพท์ก็จะเก็บค่าตำแหน่งทั้งหมดที่มีคำนั้นปรากฏอยู่ เมื่อได้ทั้ง 2 ตารางแล้วก็แสดงได้เป็นดรรชนีแบบแฟ้มผกผัน

แฟ้มผกผันจะช่วยให้การค้นหารวดเร็วยิ่งขึ้น เพราะเมื่อต้องการค้นหาคำใด ระบบจะไปค้นหาที่แฟ้มผกผัน แทนที่จะลงไปค้นหาในเอกสาร เมื่อมีข้อสอบถามจากผู้ใช้ ระบบการค้นคืนจะไปค้นหาในแฟ้มผกผัน เมื่อคำในข้อสอบถามใดตรงกับในคำในแฟ้มผกผัน ก็จะนำเอาตำแหน่งของคำนั้นไปค้นคืนเอกสารนั้นออกมาแสดงผล

2.4 การกำหนดรูปแบบการค้นคืนเอกสารและข้อสอบถาม

เทคนิคที่นำมาใช้ในการค้นคืนเอกสารนั้นมีอยู่มากมาย ซึ่งวิธีที่ง่ายและใช้กันเป็นส่วนมาก คือ การค้นคืนเอกสารโดยใช้คำสำคัญ (Keyword search) โดยวิธีนี้จะนำคำในข้อสอบถามของผู้ใช้มาเปรียบเทียบกับคำในเอกสารทีละคำ หากพบว่าเอกสารใดมีคำคำนั้นปรากฏ ระบบจะค้นคืนเอกสารออกมาแสดง แต่วิธีนี้ไม่ได้คำนึงถึงคุณภาพของผลลัพธ์ที่ได้ ดังนั้นเอกสารที่แสดง

ออกมาส่วนใหญ่จึงไม่ตรงกับความต้องการของผู้ใช้ ต่อมาวิธีการค้นคืนข้อมูลจึงถูกปรับปรุงให้ดียิ่งขึ้นด้วยวิธีการต่าง ๆ (กฤษณี อริยชาญศิลป์ 2545)

แบบจำลองการค้นคืนสารสนเทศ (Information Retrieval Model) หลายแบบจำลองพยายามสร้างระบบให้ค้นคืนสารสนเทศออกมาให้ตรงตามความต้องการของผู้ใช้มากที่สุด และค้นคืนสารสนเทศที่ไม่เกี่ยวข้องกับความต้องการออกมาน้อยที่สุด เช่น แบบจำลองความน่าจะเป็น (Probabilistic Model) แบบจำลองปริภูมิเวกเตอร์ (Vector Space Model) เป็นต้น ระบบการค้นคืนสารสนเทศ (Information Retrieval) โดยส่วนใหญ่ นิยมใช้แบบจำลองปริภูมิเวกเตอร์ (Vector Space Model) กำหนดรูปแบบการค้นคืนเอกสารและข้อสอบถาม เนื่องจากเป็นแบบจำลองที่มีการคำนวณทางคณิตศาสตร์ที่ไม่ยุ่งยาก และมีการพิจารณาถึงค่าน้ำหนักของคำ (Terms) ในเอกสารและข้อสอบถาม (Query) ซึ่งเป็นแนวคิดพื้นฐานที่สำคัญสำหรับการค้นคืนเอกสาร (Document Retrieval) แบบจำลองปริภูมิเวกเตอร์ (Vector Space Model) จะแบ่งการทำงานออกเป็นสองขั้นตอน ดังนี้

1) การกำหนดเวกเตอร์เอกสารและข้อสอบถาม

แบบจำลองปริภูมิเวกเตอร์เป็นพื้นฐานในการค้นคืนสารสนเทศสำหรับแนวคิดแบบเวกเตอร์ (Vector) ที่ใช้ในการค้นคืนสารสนเทศนั้น จะพิจารณาส่วนต่างๆที่เกี่ยวข้อง คือ คำ (Term) เอกสาร (Document) และข้อสอบถาม (Query) เหล่านี้แทนได้ด้วยรูปแบบเวกเตอร์ (Vector) ดังนี้

ถ้าสมมติระบบมีจำนวนคำ t คำ (หลังจากผ่านขั้นตอนการสกัดคำออกจากเอกสารแล้ว) ดังนั้นเมื่อแทนด้วยรูปแบบเวกเตอร์ของเอกสารและข้อสอบถามจะมี t มิติ (ขนาดของเวกเตอร์ขึ้นอยู่กับจำนวนของคำที่ปรากฏในเอกสารนั้น) สามารถนิยามดังนี้

กำหนดให้ d_j คือ เอกสารที่ j

q คือ ข้อสอบถาม

$k_{i,j}$ คือ คำ i ในเอกสารที่ j

$k_{i,q}$ คือ คำ i ในข้อสอบถาม q

t คือ จำนวนคำทั้งหมดในระบบ (ขนาดมิติของเวกเตอร์)

เวกเตอร์เอกสารแสดงในรูปแบบเวกเตอร์ ขนาด t มิติ

$$d_j = (k_{1,j}, k_{2,j}, \ldots, k_{t,j}) \text{ หรือ } d_j = \sum_{i=1}^{t} k_{i,j} \qquad (2.1)$$

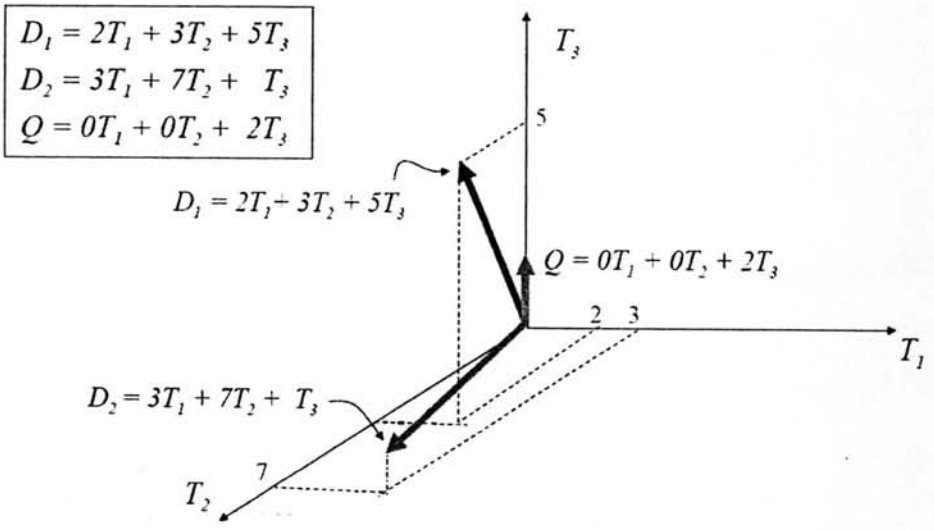
และเวกเตอร์ข้อสอบถามในรูปแบบเวกเตอร์เช่นเดียวกัน

$$q = (k_{1,q}, k_{2,q}, \ldots, k_{t,q}) \text{ หรือ } q = \sum_{i=1}^{t} k_{i,q} \qquad (2.2)$$

จากคำนิยาม 2.1 และ 2.2 แต่ละตำแหน่งมิติของเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถามจะมีค่าก็ต่อเมื่อ

- ถ้าในเวกเตอร์ข้อสอบถามหรือเอกสารมีคำใดปรากฏอยู่ ที่ตำแหน่งมิติของคำนั้นในเวกเตอร์จะมีค่าเท่ากับค่าน้ำหนักของคำในเอกสารนั้น ๆ (จะกล่าววิธีการให้น้ำหนักคำในหัวข้อถัดไป)
- ถ้าในเวกเตอร์ข้อสอบถามหรือเอกสารไม่มีคำใดปรากฏอยู่ ที่ตำแหน่งมิติของคำนั้นในเวกเตอร์จะมีค่าเท่ากับ 0

ตัวอย่างเช่น กำหนดให้ระบบมีคำอยู่ 3 คำ (T_1, T_2, T_3) มีเอกสาร 2 เอกสาร และ 1 ข้อสอบถามแสดงในรูปแบบเวกเตอร์สเปซของระบบมิติ 3 มิติ ดังรูปที่ 2.5 (ชูชาติ หฤไชยะศักดิ์ 2547)



รูปที่ 2.5 รูปแสดงเวกเตอร์สเปซของระบบมิติ 3 มิติ

จากรูปที่ 2.5 รูปแบบเซตของคำในระบบแสดงได้เป็น {T_1, T_2, T_3} ดังนั้นจะได้เวกเตอร์ของเอกสาร D_1 = {2, 3, 5} เวกเตอร์ของเอกสาร D_2 = {3, 7, 0} (เอกสาร D_2 ไม่ปรากฏคำ "T_3") และเวกเตอร์ของข้อสอบถาม Q = {0, 0, 2} (ข้อสอบถาม Q ไม่ปรากฏคำ "T_1" และ "T_2") ในที่นี้สมมติให้แต่ละเอกสารและข้อสอบถามมีค่าน้ำหนักคำต่างกัน

2) **การให้ค่าน้ำหนักคำ**

คำที่มีความสำคัญที่ดีควรจะปรากฏอยู่เป็นจำนวนมากในเนื้อหาของเอกสารเฉพาะฉบับนั้น และปรากฏอยู่น้อยมากในชุดเอกสารที่เหลือ ดังนั้นวิธีการประมาณค่าความสำคัญของคำ โดยการให้น้ำหนักคำคำหนึ่งในเอกสารฉบับหนึ่งจะพิจารณาจากความถี่ของคำที่ปรากฏในเอกสารนั้นและจำนวนของเอกสารทั้งหมดที่มีคำๆนั้นปรากฏอยู่

- **ความถี่ของคำ (Term Frequency)** (Baeza-Yates and Ribeiro-Neto 1999)

ความถี่ในการปรากฏของคำ (Term Frequency: tf) สามารถเป็นสิ่งที่บ่งบอกถึงความสำคัญของคำคำนั้นที่มีต่อเอกสารหนึ่ง ๆ เอกสารที่มีคำที่ผู้ใช้ต้องการจะเป็นเอกสารที่มีประโยชน์ต่อผู้ใช้สูงกว่า เช่น ถ้าผู้ใช้ต้องการคำว่า "computer" เอกสารที่ใช้คำว่า "computer" 10 ครั้ง จะเป็นประโยชน์กว่าเอกสารที่มีคำว่า "computer" เพียงครั้งเดียว โดยที่ค่า tf หาได้จากสมการที่ (2.3)

กำหนดให้ $freq_{i,j}$ คือ ความถี่ที่คำ k_i ปรากฏ ในเอกสาร d_j

$\max_l freq_{l,j}$ คือ ความถี่ของคำใด ๆ ที่ปรากฏในเอกสาร d_j ที่มากที่สุด

$$tf_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \qquad (2.3)$$

- **ความถี่ของเอกสารแบบผกผัน (Inverse Document Frequent : idf)**

เอกสารที่มีคำทั่วไปปรากฏอยู่บ่อยครั้งในทุกเอกสารแต่คำคำนั้นไม่สามารถทำให้เอกสารแตกต่างจากเอกสารอื่นได้ การประเมินค่าความสำคัญของคำนั้นจะมีความสำคัญน้อยกว่าคำที่ใช้น้อยหรือคำที่ปรากฏเฉพาะบางเอกสาร เช่น การพบคำว่า "Indeed" ในเอกสาร แม้ว่าจะพบคำว่า "Indeed" บ่อยในเอกสารก็ตามแต่คำนี้ไม่มีความสำคัญสำหรับเอกสารนั้น ๆ โดยค่า idf หาได้จากสมการที่ (2.4)

กำหนดให้ N คือ จำนวนเอกสารทั้งหมดในระบบ

n_i คือ จำนวนเอกสารที่มีคำ k_i ปรากฏ

$$idf_i = \log \frac{N}{n_i} \qquad (2.4)$$

การให้ค่าน้ำหนักของคำนั้น หากพิจารณาเฉพาะค่าความถี่ของคำ (tf) จะพิจารณาแค่ความถี่ของคำที่ปรากฏในเอกสาร ไม่ได้พิจารณาถึงค่าความสำคัญของคำที่ปรากฏในเอกสาร บางครั้งค่าความถี่ของคำๆหนึ่งมีค่ามาก แต่คำๆนั้นปรากฏอยู่ในทุกเอกสารบ่อยครั้ง และคำๆนั้น

ไม่สามารถที่จะแยกเอกสารที่ตรงกับความต้องการออกจากเอกสารที่ไม่ตรงกับความต้องการได้ ดังนั้นเพื่อแก้ปัญหาคำประเภทเหล่านี้ และช่วยให้การให้ค่าน้ำหนักของคำในเอกสารมีความเหมาะสมยิ่งขึ้น การให้ค่าน้ำหนักของคำในเอกสารจึงพิจารณาค่าความถี่แบบผกผัน (idf) ร่วมด้วย โดยใช้ค่าความถี่ของคำ (tf) และค่าความถี่แบบผกผัน (idf) มาคำนวณร่วมกัน ซึ่งเป็นการเปรียบเทียบความถี่ของคำในเอกสารกับความถี่ของคำในเอกสารอื่น (Salton et al. 1987) ดังสมการที่ 2.5

กำหนดให้ $w_{i,j}$ คือ ค่าน้ำหนักของคำ i ในเอกสารที่ j

$tf_{i,j}$ คือ ความถี่ของคำ i ในเอกสาร j

idf_i คือ ค่าความถี่ผกผันของคำ i ของเอกสารทั้งหมด

$$\text{Document term weight } (w_{i,j}) = tf_{i,j} \times idf_i \qquad (2.5)$$

ตัวอย่างเช่น ในชุดเอกสารหนึ่งในระบบ ประกอบด้วยเอกสาร D_1, D_2 และ D_3 เอกสารแต่ละฉบับหลังผ่านการสกัดคำสำคัญออกจากเอกสารมีลักษณะตามรูปที่ 2.6 จากนั้นหาค่าความถี่ของคำที่ไม่ซ้ำกันในเอกสารแต่ละฉบับตามตารางที่ 2.1 แล้วจึงหาคำที่ไม่ซ้ำกันทั้งหมดในชุดเอกสาร และกำหนดค่าจำนวนเอกสารที่มีคำ ๆ นั้นปรากฏอยู่ และ ค่า idf_i ให้แต่ละคำ ดังตารางที่ 2.2

D_1:	Computer information Computer
	Computer
D_2:	Internet Computer Internet Data
D_3:	System Internet

รูปที่ 2.6 รูปแสดงตัวอย่างชุดเอกสารในระบบ

ตารางที่ 2.1 ตารางแสดงความถี่ของคำในชุดเอกสาร

เอกสาร	คำ	ความถี่
D_1	Computer	3
D_1	Information	1
D_2	Internet	2
D_2	Computer	1
D_2	Data	1
D_3	System	1
D_3	Internet	1

ตารางที่ 2.2 ตารางแสดงค่า idf_i ของคำในชุดเอกสาร

คำ	จำนวนเอกสารที่คำปรากฏ	idf (คำนวณตามสมการ 2.4)
Computer	2	0.18
Information	1	0.48
Internet	2	0.18
System	1	0.48
Data	1	0.48

2.5 การให้ค่าน้ำหนักของคำในข้อสอบถาม

การคำนวณค่าน้ำหนักของคำในข้อสอบถาม เป็นการพิจารณาโดยใช้ค่า tf-idf เช่นกันกับการคำนวณค่าน้ำหนักของคำในเอกสาร ดังที่กล่าวมาข้างต้น จากค่าความถี่ (tf) ของคำจะมีค่าตั้งแต่ 0 ถึง 1 ดังนั้นจึงกำหนดให้ข้อสอบถามมีค่าความถี่ของคำพื้นฐานเป็น 0.5 (ค่ากลาง) และบวกค่าความถี่ของคำโดยพิจารณา 0 ถึง 0.5 เท่านั้น เนื่องจากข้อสอบถามที่ผู้ใช้กรอกเข้ามานั้นถือเป็นคำที่มีนัยสำคัญมาก ดังนั้นค่าความถี่ของคำควรจะมีน้ำหนักไม่ต่ำกว่าครึ่งหนึ่ง (Baeza-Yates and Ribeiro-Neto 1999) จากนั้นนำค่าความถี่ที่ได้คูณด้วยค่าความถี่ของเอกสารแบบผกผัน (idf) ดังนั้นสมการในการคำนวณค่าน้ำหนักของคำในข้อสอบถามจะแสดงดังสมการที่ 2.6

กำหนดให้

$w_{i,q}$	คือ ค่าน้ำหนักของคำลำดับที่ i ในข้อสอบถาม
N	คือ จำนวนเอกสารทั้งหมดในระบบ
n_i	คือ จำนวนเอกสารที่มีคำ k_i ปรากฏ
$freq_{i,q}$	คือ ความถี่ที่คำ k_i ปรากฏ ในข้อสอบถาม q
$max_l freq_{l,q}$	คือ ความถี่ของคำใด ๆ ที่ปรากฏในข้อสอบถาม q ที่มากที่สุด

$$\text{Query term weight } (w_{i,q}) = \left(0.5 + \frac{0.5 freq_{i,q}}{\max_l freq_{l,q}}\right) \times \log \frac{N}{n_i} \qquad (2.6)$$

2.6 การค้นหาเอกสารที่ตรงกับข้อสอบถามของผู้ใช้

ขั้นตอนการค้นคืนสารสนเทศในการค้นหาเอกสารที่ตรงกับข้อสอบถามของผู้ใช้ จะอาศัยการคำนวณความคล้ายคลึงกัน โดยการเปรียบเทียบความเหมือนระหว่างเวกเตอร์เอกสารในระบบและเวกเตอร์ข้อสอบถามที่ผู้ใช้กรอกเข้ามา วิธีการในการค้นหาเอกสารเป็นวิธีการที่สามารถคำนวณระดับความเหมือนระหว่างเวกเตอร์ 2 เวกเตอร์ ซึ่งวิธีการคำนวณความเหมือนมีหลายวิธีการ เช่น วิธีการวัดความเหมือนเชิงมุม(Cosine similarity), การวัดความเหมือนเชิงระยะทาง (Euclidean distance) เป็นต้น วิธีการวัดความเหมือนระหว่างเวกเตอร์ 2 เวกเตอร์นั้น โดยทั่วไปมีการคำนวณอยู่ 4 วิธีด้วยกัน คือ การหาค่าความเหมือนวิธีไดซ์ (Dice coefficient), การหาค่าความเหมือนวิธีเจคคาร์ด (Jaccard coefficient), การหาค่าความเหมือนวิธีโคไซน์ (Cosine coefficient) และการหาค่าความเหมือนวิธีโอเวอร์แล็บ (Overlap coefficient) แต่วิธีการที่นิยมใช้มากที่สุดคือ การหาค่าความเหมือนวิธีโคไซน์ (Cosine coefficient) วิธีวัดความเหมือนอีกวิธีหนึ่งคือ วิธีการวัดความเหมือนเชิงระยะทาง (Distance) ซึ่งเป็นวิธีวัดความเหมือนระหว่าง 2 เวกเตอร์ที่นิยมใช้ในระบบค้นคืนรูปภาพ (Image Retrieval) ซึ่งวิธีการคำนวณการหาค่าความเหมือนเชิงระยะทาง (Distance) ที่ใช้กันอย่างกว้างขวางคือ วิธีการหาค่าความเหมือนระยะห่างยูคลิเดียน (Euclidean distance) โดยวิธีการคำนวณความเหมือนวิธีโคไซน์ (Cosine coefficient) และการคำนวณความเหมือนเชิงระยะห่างยูคลิเดียน (Euclidean distance) แสดงได้ดังนี้

การหาค่าความเหมือนวิธีโคไซน์ (Cosine coefficient) การคำนวณค่าความเหมือนของมุม ระหว่าง 2 เวกเตอร์ สมการของ Cosine coefficient แสดงได้ดังสมการที่ 2.7 (ชูชาติ หฤไชยะศักดิ์ 2547)

กำหนดให้ d_j คือ เอกสารที่ j

q คือ ข้อสอบถาม

$w_{i,j}$ คือ ค่าน้ำหนักของคำที่ i ในเอกสารที่ j

$w_{i,q}$ คือ ค่าน้ำหนักของคำที่ i ในข้อสอบถาม

t คือ จำนวนคำทั้งหมดในระบบ (ขนาดมิติของเวกเตอร์)

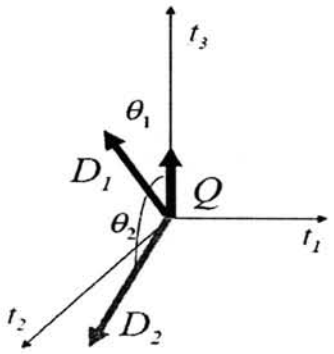
$$\text{CosSim}(d_j,q) = \frac{\sum_{i=1}^{t}(w_{i,j} \times w_{i,q})}{\sqrt{\sum_{i=1}^{t} w_{i,j}^{\,2}} \times \sqrt{\sum_{i=1}^{t} w_{i.q}^{\,2}}} \qquad (2.7)$$

วิธีของการหาค่าความเหมือนวิธีโคไซน์ (Cosine coefficient) มีลักษณะการวัด โดยช่วงค่าความเหมือนจะมีค่าตั้งแต่ค่า 0 ถึง 1 และคำนวณค่าความเหมือนสำหรับเวกเตอร์ที่แต่ละมิติไม่

มีค่าติดลบ (Chowdhury 2004) ค่าความเหมือนระหว่างเวกเตอร์เอกสารใดทำมุมกับเวกเตอร์ข้อสอบถามน้อย เอกสารนั้นจะเป็นผลลัพธ์ของการค้นคืน ซึ่งนำมาแสดงแก่ผู้ใช้ ในงานวิจัยนี้จึงเลือกใช้การหาค่าความเหมือนวิธีโคไซน์ (Cosine coefficient) ในการคำนวณค่าความเหมือนในการค้นคืนเอกสาร เนื่องจากเป็นวิธีได้รับความนิยมในการใช้ในระบบค้นคืนเอกสารส่วนใหญ่ (Chowdhury 2004)

ตัวอย่างการคำนวณค่าความเหมือนวิธีโคไซน์ แสดงได้ดังรูปที่ 2.7 เป็นรูปแสดงเชิงมุมระหว่างเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถามในมิติ โดยใช้ข้อมูลตัวอย่างเดียวกับข้อมูลที่ใช้แสดงในรูปที่ 2.5 คือ ระบบประกอบด้วยคำ 3 คำ (T_1, T_2, T_3) เอกสาร 2 เอกสาร (D_1, D_2) และ 1 ข้อสอบถาม (Q) คือ

$$\left| \begin{array}{l} D_1 = 2T_1 + 3T_2 + 5T_3 \\ D_2 = 3T_1 + 7T_2 + T_3 \\ Q = 0T_1 + 0T_2 + 2T_3 \end{array} \right|$$



รูปที่ 2.7 รูปแสดงการทำมุมระหว่างเวกเตอร์เอกสารและข้อสอบถาม

จากรูปที่ 2.7 สามารถคำนวณค่าความเหมือนเชิงมุมได้ตามสมการ (2.7) ดังนี้

$$\text{CosSim}(D_1, Q) = 10 / (\sqrt{(4 + 9 + 25)} \times \sqrt{(0 + 0 + 4)}) = 0.81$$

$$\text{CosSim}(D_2, Q) = 2 / (\sqrt{(9 + 49 + 1)} \times \sqrt{(0 + 0 + 4)}) = 0.13$$

เนื่องจากการหาค่าความเหมือนวิธีโคไซน์ ได้กำหนดว่าเวกเตอร์เอกสารใดทำมุมกับเวกเตอร์ข้อสอบถามน้อย หรือค่าความเหมือนระหว่างเวกเตอร์เอกสารกับเวกเตอร์ข้อสอบถามที่ได้จากการคำนวณตามสมการที่ 2.7 ออกมามีค่ามากกว่า เอกสารนั้นจะเป็นผลลัพธ์ของการค้นคืน ดังนั้นจากตัวอย่างดังรูปที่ 2.7 สรุปได้ว่าเอกสาร D_1 เป็นเอกสารที่เกี่ยวเนื่องกับข้อสอบถาม Q มากกว่าเอกสาร D_2

การหาค่าความเหมือนวิธีระยะห่างยูคลิเดียน (Euclidean distance) การคำนวณค่าความเหมือนของระยะทางระหว่าง 2 เวกเตอร์ วิธีระยะห่างยูคลิเดียนเป็นวิธีวัดระยะห่าง (Distance) ระหว่างข้อมูล ซึ่งเป็นวิธีการวัดระยะทางที่ใช้ในเทคนิคการแบ่งกลุ่มข้อมูล

(Clustering) เพื่อนำไปพิจารณาระยะห่างของข้อมูลในการจัดกลุ่ม และใช้ในเทคนิควิธีเพื่อนบ้านใกล้สุด (Nearest-Neighbor Methods) ซึ่งเป็นวิธีที่นำหลักการวัดระยะทางระหว่างข้อมูลที่ต้องการแบ่งกลุ่มความเหมือนของชุดข้อมูลที่เป็นเวกเตอร์มาช่วยในการค้นคืนเอกสารที่ตรงกับข้อสอบถาม (Chowdhury 2004) สมการของ Euclidean distance แสดงได้ดังสมการที่ 2.8 (Baeza-Yates and Ribeiro-Neto, 1999)

กำหนดให้ x_i คือ ค่าของเวกเตอร์ x ในตำแหน่งมิติ i

y_i คือ ค่าของเวกเตอร์ y ในตำแหน่งมิติ i

i คือ ตำแหน่งมิติของเวกเตอร์

n คือ ขนาดมิติในเวกเตอร์

$$Eud\ (x_i, y_i) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (2.8)$$

การหาค่าความเหมือนวิธีระยะห่างยูคลิเดียน (Euclidean distance) มีลักษณะการหาความเหมือนโดยวัดระยะห่างระหว่าง 2 เวกเตอร์ เมื่อนำมาช่วยในการค้นคืนเอกสารจะเป็นการหาค่าความเหมือนระหว่างเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถาม (Chowdhury 2004) โดยการคำนวณค่าความเหมือนในการค้นคืนเอกสารวิธีระยะห่างยูคลิเดียน (Euclidean distance) จะพิจารณาความเหมือนระหว่างเวกเตอร์เอกสารใดมีระยะห่างกับเวกเตอร์ข้อสอบถามน้อย เอกสารนั้นจะเป็นผลลัพธ์ของการค้นคืน ซึ่งถูกนำมาแสดงแก่ผู้ใช้ วิธีระยะห่างยูคลิเดียน (Euclidean distance) เป็นวิธีที่นิยมใช้ในการวัดความเหมือนระหว่างระยะห่างเวกเตอร์ 2 เวกเตอร์ในระบบการค้นคืนรูปภาพ (Image Retrieval) (Qian et al. 2004)

ตัวอย่างการคำนวณค่าความเหมือนวิธีระยะห่างยูคลิเดียน สามารถแสดงได้โดยอ้างอิงมาจากรูปที่ 2.5 ประกอบด้วยคำ 3 คำ (T_1, T_2, T_3) มีเอกสาร 2 เอกสาร (D_1, D_2) และ 1 ข้อสอบถาม (Q) วิธีการคำนวณค่าความเหมือนเชิงระยะทางตามสมการ 2.8 แสดงดังนี้

$$Eud\ (D_1, Q) = \sqrt{(4 + 9 + 9)} = 4.69$$

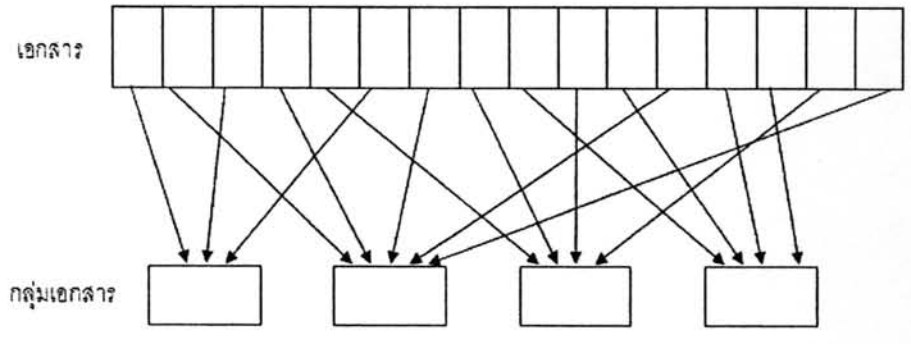
$$Eud\ (D_2, Q) = \sqrt{(9 + 49 + 1)} = 7.68$$

จากหลักการค่าความเหมือนระหว่างเวกเตอร์เอกสารและข้อสอบถามด้วยค่าความเหมือนวิธีระยะห่างยูคลิเดียน ได้กำหนดว่าค่าความเหมือนของระยะห่างระหว่างเวกเตอร์เอกสารไหนที่มีระยะห่างกับเวกเตอร์ของข้อสอบถามที่น้อยที่สุดจากการคำนวณตามสมการที่ 2.8 เอกสารนั้นจะเป็นผลลัพธ์ของการค้นคืน ดังนั้นจากตัวอย่างดังรูปที่ 2.5 แสดงว่าเอกสาร D_1 เป็นเอกสารที่เกี่ยวเนื่องกับข้อสอบถาม Q มากกว่าเอกสาร D_2

2.7 การแบ่งกลุ่มเอกสารเทคนิค K-means Clustering

การแบ่งกลุ่มข้อมูล (Clustering) เป็นเทคนิควิธีการวิเคราะห์ข้อมูลในการทำเหมืองข้อมูล (Data Mining) โดยจะแบ่งข้อมูลซึ่งส่วนใหญ่จะอยู่ในลักษณะของเวกเตอร์ ออกเป็นกลุ่ม (Cluster) โดยข้อมูลที่มีลักษณะเหมือนกันหรือคล้ายกันจะถูกจัดไว้ในกลุ่มเดียวกัน ขั้นตอนวิธีที่ใช้ในการแบ่งกลุ่มจะอาศัยความเหมือน (Similarity) หรือความใกล้ชิด (Proximity) โดยคำนวณจากการวัดระยะระหว่างเวกเตอร์ของข้อมูล โดยการใช้วิธีการวัดระยะแบบต่าง ๆ (วิกิพีเดีย สารานุกรมเสรี 2550) การแบ่งกลุ่มเอกสาร (Document Clustering) เป็นการนำเอาเทคนิคการแบ่งกลุ่มข้อมูล (Clustering) มาประยุกต์ใช้กับข้อมูลเอกสาร โดยมีเป้าหมายเพื่อจัดกลุ่มเอกสารที่เกี่ยวข้องกันอยู่ภายในกลุ่มเดียวกันและเอกสารที่ไม่เกี่ยวข้องกันหรือเกี่ยวข้องกันน้อยอยู่ต่างกลุ่มออกไป (Weiss et al. 2005)

เทคนิค K-means Clustering เป็นขั้นตอนวิธีในการแบ่งกลุ่มข้อมูลออกเป็นกลุ่มย่อย ซึ่งนำมาใช้สำหรับการจัดกลุ่มเอกสาร เทคนิค K-means เหมาะสมสำหรับใช้แบ่งกลุ่มข้อมูลที่มีจำนวน 200 หน่วยขึ้นไป (กัลยา วานิชย์บัญชา 2548) โดยจะต้องกำหนดจำนวนกลุ่มที่ต้องการแบ่ง หรือกำหนดค่า K แนวคิดของเทคนิค K-Means แสดงได้ดังรูปที่ 2.8 ที่เริ่มจากการที่เอกสารกระจายกันอยู่ในกลุ่มเดียวกัน และจากนั้นเอกสารที่เหมือนกันจะถูกกระจายเข้าไปอยู่ในกลุ่มที่เล็กลงไป (Weiss et al. 2005)



รูปที่ 2.8 รูปแสดงแนวคิดของเทคนิค K-means Clustering

การแบ่งกลุ่มเอกสารด้วยเทคนิค K-means จะมีการทำงานวนซ้ำหลายรอบ โดยแต่ละรอบจะมีการรวมของเอกสารให้อยู่ในกลุ่มใดกลุ่มหนึ่ง โดยเลือกกลุ่มที่เอกสารนั้นมีระยะห่างจากค่ากลางของกลุ่มน้อยที่สุด หรืออยู่ในขอบเขต (boundaries) ของกลุ่มนั้น แล้วคำนวณค่ากลางของกลุ่มใหม่ จะทำเช่นนี้จนกระทั่งค่ากลางของกลุ่มไม่เปลี่ยนแปลงหรือครบจำนวนรอบที่กำหนดไว้ การหาค่าระยะห่างจะใช้วิธี Euclidean Distance ขั้นตอนการทำงานของ เทคนิค K-means

Clustering แสดงได้ดังรูปที่ 2.9 (กฤษณี อริยชาญศิลป์ 2545; Han, Karypis and Kumar 2001; Weiss et al. 2005)

1) สุ่มเลือกเอกสารเท่ากับจำนวนกลุ่มที่ต้องการแบ่ง เพื่อนำมาเป็นจุดศูนย์กลาง (center) ของแต่ละกลุ่มเอกสารในรอบแรก
2) หาขอบเขต (boundaries) ระหว่างกลุ่มเอกสาร โดยขอบเขตของกลุ่มเอกสารที่อยู่ติดกันคือกึ่งกลางระหว่างจุดศูนย์กลางของกลุ่มเอกสารทั้งสอง
3) กำหนดกลุ่มให้กับเอกสารแต่ละเอกสาร โดยพิจารณาจากตำแหน่งที่อยู่ของเอกสารนั้นว่าอยู่ภายใต้ขอบเขตของกลุ่มเอกสารใด
4) คำนวณจุดศูนย์กลางของกลุ่มทุกกลุ่มใหม่ วนการทำงานซ้ำไปยังข้อ 2
5) ถ้าจุดศูนย์กลางของกลุ่มเอกสาร หรือขอบเขตระหว่างกลุ่มเอกสารไม่เปลี่ยนแปลงแล้วก็จะหยุดการทำงาน

รูปที่ 2.9 รูปแสดงขั้นตอนการทำงานของเทคนิค K-means Clustering

2.8 การกำหนดค่าความเหมือนในการค้นคืนเอกสารต่อผู้ใช้

ขั้นตอนของการค้นคืนเอกสารต่อผู้ใช้ในระบบการค้นคืนเอกสารนั้น จะต้องกำหนดค่าความเหมือนระหว่างเอกสารและข้อสอบถามในการค้นคืนเอกสาร เมื่อเอกสารและข้อสอบถามใดมีค่าความเหมือนมากกว่าหรือน้อยกว่า(ขึ้นอยู่กับวิธีการที่ใช้คำนวณค่าความเหมือนระหว่างเอกสารและข้อสอบถามดังตัวอย่างวิธีการคำนวณที่ได้กล่าวไปในหัวข้อข้างต้น) ค่าความเหมือนที่ตั้งไว้ ระบบจะค้นคืนเอกสารนั้นออกมาแสดงต่อผู้ใช้ ซึ่งการตั้งค่านี้จะเป็นการตั้งค่าตามความเหมาะสมของระบบค้นคืนเอกสารแต่ละระบบ ไม่มีรูปแบบการคำนวณตายตัว (Baeza-Yates and Ribeiro-Neto, 1999)

งานวิจัยของ Udomchaiporn (2005) ได้เสนอไว้ว่าการตั้งค่าความเหมือนสามารถตั้งได้ตามความเหมาะสมกับระบบค้นคืนเอกสารนั้น ๆ โดยจะสามารถตั้งไว้ที่ค่าเฉลี่ย (Mean) ของค่าความเหมือนของข้อสอบถามกับเอกสาร หรือค่าเฉลี่ยบวกกับค่าเบี่ยงเบนมาตรฐาน (Mean + Standard Deviation) ของค่าความเหมือนข้อสอบถามกับเอกสารหรือมากกว่านี้ได้ตามความเหมาะสม

ในงานวิจัยของ ศิรัตน์ ศิรนานนท์ (2549) ได้ตั้งค่าความเหมือนไว้ที่ค่าเฉลี่ย (Mean) บวกค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) ของค่าความเหมือนทุกข้อสอบถามกับทุกเอกสาร งานวิจัยของศิรัตน์ ศิรนานนท์ใช้วิธีการคำนวณค่าความเหมือนระหว่างเอกสารและข้อสอบถามด้วยวิธีการวัดความเหมือนโคไซน์ (Cosine coefficient) และได้ใช้ฐานข้อมูลนิตยสารไทม์ (TIME Collection) ในการทดสอบงานวิจัย ซึ่งศิรัตน์ ศิรนานนท์ได้ระบุว่าฐานข้อมูลนิตยสารไทม์มีการกำหนดเอกสารที่เกี่ยวเนื่องกับข้อสอบถาม ดังนั้นจึงสามารถหาเปอร์เซ็นต์ของเอกสารที่ไม่เกี่ยวเนื่องถูกดึงออกมาแสดงได้ ดังนั้นศิรัตน์ ศิรนานนท์จึงได้ทดสอบตั้งค่าความเหมือนต่ำสุดเท่ากับค่าเฉลี่ย (Mean) ของค่าความเหมือนทุกข้อสอบถามกับทุกเอกสาร (จากหลักการวิธีการวัดความเหมือนโคไซน์ ระบุว่าเอกสารที่จะถูกค้นคืนจะต้องเป็นเอกสารที่มีค่าความเหมือนมากกว่าค่าความเหมือนที่ตั้งไว้) ผลการทดสอบสามารถจะค้นคืนเอกสารที่ไม่เกี่ยวเนื่องกับข้อสอบถามทั้งหมดในฐานข้อมูลนิตยสารไทม์ออกมามากกว่าการตั้งค่าความเหมือนต่ำสุดไว้ที่ค่าเฉลี่ย (Mean) บวกค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) ของค่าความเหมือนทุกข้อสอบถามกับทุกเอกสาร 5 เท่า ดังนั้นถ้ากำหนดค่าความเหมือนต่ำสุดเท่ากับค่าเฉลี่ย (Mean) เอกสารที่ไม่เกี่ยวเนื่องจะถูกค้นคืนออกมามากจนเกินไป ดังนั้น ศิรัตน์ ศิรนานนท์จึงกำหนดค่าความเหมือนต่ำสุดไว้ที่ค่าเฉลี่ย (Mean) บวกค่าเบี่ยงเบนมาตรฐาน (Standard Deviation)

2.9 การวัดประสิทธิภาพระบบค้นคืนเอกสาร

เมื่อได้ผลลัพธ์จากการค้นคืนเอกสารออกมา การวัดว่าเอกสารมีความถูกต้องตรงกับความต้องการมากน้อยเพียงใด หรือประสิทธิภาพของการค้นคืนมีมากน้อยเพียงใด สามารถกระทำได้หลายวิธี แต่สามวิธีที่นิยมใช้ คือ ค่าความแม่นยำ(Precision) ค่าความระลึก (Recall) และค่าเฉลี่ยฮาร์โมนิค (Harmonic mean)

มีงานวิจัยจำนวนมากที่ใช้ค่าความระลึก (Recall) และค่าความแม่นยำ (Precision) วัดประสิทธิภาพของระบบค้นคืนเอกสาร เช่น งานวิจัยการวัดความคล้ายคลึงระหว่างเอกสารโดยใช้แนวทางด้านความหมาย (พิลาวัณย์ พลับรู้การ และกฤษณะ ไวยมัย 2545) งานวิจัยการประยุกต์ใช้ขั้นตอนวิธีพันธุกรรม (genetic algorithms) ในการค้นคืนสารสนเทศ (Klabbankoh 1999) เป็นต้น และงานวิจัยการส่งเสริมระบบการค้นคืนเอกสารด้วยเทคนิคปริภูมิเวกเตอร์ (Greenwood 2002) เป็นต้น และมีงานวิจัยจำนวนมากที่ใช้ค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) วัดประสิทธิภาพของระบบค้นคืนเอกสาร เช่น งานวิจัยการจัดกลุ่มเอกสารทางเว็บโดยใช้เซตรายการที่มากที่สุด (Zhuang and Dai, 2004) และงานวิจัยการค้นคืนสารสนเทศโดยใช้กฎ

ความสัมพันธ์ร่วมกลับผลสะท้อนกลับจากผู้ใช้ (ศิรัตน์ ศิรนานนท์ 2549) เป็นต้น ค่าความระลึก, ค่าความแม่นยำ และค่าเฉลี่ยฮาร์โมนิค สามารถคำนวณได้ดังนี้

- **ค่าความระลึก (Recall)** เป็นอัตราส่วนของการค้นพบเอกสารที่เกี่ยวเนื่องกับความต้องการ (relevant document) จากจำนวนเอกสารที่เกี่ยวเนื่องทั้งหมดในระบบ ดังสมการที่ (2.9)

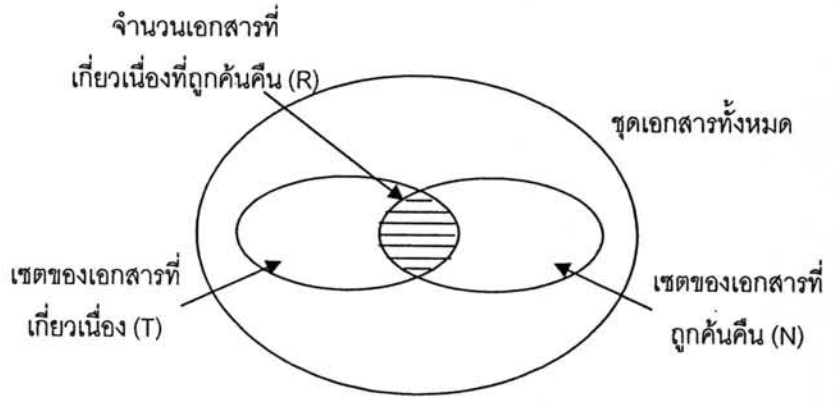
$$\text{ค่าความระลึก (Recall)} = \frac{\text{จำนวนเอกสารที่เกี่ยวเนื่องที่ค้นคืนได้}}{\text{จำนวนเอกสารที่เกี่ยวเนื่องทั้งหมดในฐานข้อมูล}} \quad (2.9)$$

- **ค่าความแม่นยำ (Precision)** เป็นอัตราส่วนของการค้นพบเอกสารที่ถูกดึงขึ้นมาแล้วเกี่ยวเนื่องกับความต้องการ (relevant document) จากจำนวนเอกสารทั้งหมดที่ทำการค้นคืนมาได้ ดังสมการที่ (2.10)

$$\text{ค่าความแม่นยำ (Precision)} = \frac{\text{จำนวนเอกสารที่เกี่ยวเนื่องที่ค้นคืนได้}}{\text{จำนวนเอกสารทั้งหมดที่ค้นคืนมาได้}} \quad (2.10)$$

ซึ่งค่าความแม่นยำจะเป็นค่าที่แสดงว่าการค้นคืนข้อมูลให้ผลลัพธ์ที่มีความถูกต้องมากน้อยเพียงใด เช่น ถ้าค้นคืนเอกสารออกมาได้ N เอกสาร และมีเอกสารที่ค้นคืนมา R เอกสารที่ถูกต้อง ดังนั้นค่าความแม่นยำจะเป็น R / N

แต่ค่าความระลึกจะเป็นค่าที่แสดงว่าผลลัพธ์ที่ได้ครอบคลุมความต้องการมากน้อยเพียงใด เช่น ถ้าเอกสารที่ตรงกับความต้องการมีทั้งสิ้น T เอกสาร และการค้นคืนสามารถดึงเอกสารที่ตรงกับความต้องการได้ R เอกสาร ค่าความระลึกจะเป็น R / T ดังแสดงได้ในรูปที่ 2.10 (Baeza-Yates and Ribeiro-Neto 1999; Goutte and Gaussier 2005)



รูปที่ 2.10 รูปแสดงเซตของเอกสารที่เกี่ยวเนื่องและเซตของเอกสารที่ค้นคืน

ในการค้นคืนสารสนเทศ ถ้าค้นคืนเอกสารที่เกี่ยวข้องกับความต้องการออกมาทั้งหมดและไม่มีเอกสารที่ไม่เกี่ยวข้องออกมาด้วย ค่าของความระลึก (Recall) และค่าความแม่นยำ (Precision) จะมีค่าเท่ากับ 1 โดยปกติทั้งค่าความระลึก (Recall) และค่าความแม่นยำ (Precision) จะมีค่าอยู่ระหว่าง 0 ถึง 1 (นิพนธ์ เจริญกิจการ 2541)

- **ค่าเฉลี่ยฮาร์โมนิค (Harmonic mean)** เป็นการคำนวณจากค่าความแม่นยำและค่าความระลึกมาเฉลี่ยกัน เพื่อแสดงประสิทธิภาพความถูกต้องและความครอบคลุมในการค้นคืนเอกสาร สมการในการคำนวณค่านี้แสดงได้ดังสมการ (2.11) (ศิรัตน์ ศิรนานนท์ 2549)

กำหนดให้ F(j) คือ ค่าเฉลี่ยฮาร์โมนิคของ r(j) และ P(j)

r(j) คือ ค่าความระลึกของเอกสารที่ j ในลำดับ (Ranking)

P(j) คือ ค่าความแม่นยำของเอกสารที่ j ในลำดับ (Ranking)

j คือ หมายเลขเอกสารซึ่งอยู่ในลำดับที่ j

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}} \qquad (2.11)$$

วิธีการวัดค่าเฉลี่ยฮาร์โมนิคเป็นการวัดที่รวมทั้งค่าความระลึกและค่าความแม่นยำ ซึ่งค่าเฉลี่ยฮาร์โมนิคจะมีค่าอยู่ในช่วงค่า 0 ถึง 1 โดยค่าเฉลี่ยฮาร์โมนิคจะมีค่าเป็น 0 เมื่อเอกสารที่ค้นคืนมาได้นั้นไม่มีเอกสารใดเกี่ยวเนื่องกับข้อสอบถามเลย และจะมีค่าเป็น 1 เมื่อทุกเอกสารที่ค้นคืนมาได้เป็นเอกสารที่เกี่ยวเนื่องกับข้อสอบถามทั้งหมด (Baeza-Yates and Ribeiro-Neto, 1999)

2.10 งานวิจัยที่เกี่ยวข้อง

งานวิจัยชิ้นนี้เป็นการพัฒนาระบบการค้นคืนสารสนเทศที่เป็นเอกสาร ด้วยการใช้เทคนิคการวัดความเหมือนระหว่างเอกสารและข้อสอบถามที่ต่างกัน ในอดีตมีงานวิจัยที่เกี่ยวข้องกับระบบการค้นคืนเอกสารคือ งานวิจัยการวัดความคล้ายคลึงระหว่างเอกสารโดยใช้แนวทางด้านความหมาย (พิลาวัณย์ พลับรู้การ และ กฤษณะ ไวยมัย 2545) งานวิจัยนี้ได้เสนอวิธีการวัดความคล้ายคลึงระหว่างเอกสารโดยใช้แนวทางด้านความหมายเข้ามาช่วยแทนวิธีการวัดความคล้ายคลึงระหว่างเอกสารในปัจจุบัน พิลาวัณย์ พลับรู้การ และกฤษณะ ไวยมัย เสนอว่าวิธีที่นิยมใช้ในการวัดความคล้ายคลึงส่วนใหญ่จะใช้การวัดความคล้ายคลึงเชิงมุม ซึ่งใช้การแทนเอกสารด้วยระบบเวกเตอร์และหลักการทางสถิติในการวัดความคล้ายคลึง ซึ่งข้อเสียของการวัดความคล้ายคลึงแบบนี้คือจะไม่มีการคำนึงถึงความหมายของคำ กล่าวคือ ผลงานวิจัยของพิลาวัณย์

พลับรู้การ และกฤษณะ ไวยมัย จะนำโครงข่ายความสัมพันธ์ของคำมาใช้ในการคำนวณค่าความคล้ายคลึงระหว่างเอกสารสองเอกสาร

แนวทางในงานวิจัยของพิลาวัณย์ พลับรู้การ และกฤษณะ ไวยมัย คือการนำความรู้พื้นฐานเกี่ยวกับความสัมพันธ์ระหว่างคำในแง่ความหมายมาใช้ แนวทางคือคำแต่ละคำมีความใกล้เคียงเชิงความหมายกับคำอื่น ๆ ไม่เท่ากัน เอกสารที่คล้ายคลึงกันจึงควรมีคำที่มีความหมายใกล้เคียงกันปรากฏอยู่ งานวิจัยของพิลาวัณย์ พลับรู้การ และกฤษณะ ไวยมัยสร้างอภิธานคำศัพท์ขึ้นซึ่งประกอบด้วยคำและความสัมพันธ์ระหว่างคำ การพิจารณาความคล้ายคลึงระหว่างเอกสาร จะใช้หลักการที่ว่า เอกสารที่มีความคล้ายคลึงกันมาก มักจะมีคำที่มีความหมายใกล้เคียงกันมาก จากผลการทดลองพบว่าวิธีการที่ได้เสนอให้ผลดีกว่าวิธีการที่ใช้กันทั่วไป ผลการวิจัยแสดงค่าความแม่นยำ (Precision) ของเอกสารที่ได้รับที่ค่าความระลึก (Recall) ต่างๆ กัน พบว่าการวัดความคล้ายคลึงโดยใช้แนวทางด้านความหมายให้ผลที่ดีกว่าแนวทางสถิติ (การวัดความคล้ายคลึงเชิงมุม) เกือบทุกตำแหน่งของค่าความระลึก

ข้อจำกัดของงานวิจัยคือ งานวิจัยของพิลาวัณย์ พลับรู้การ และกฤษณะ ไวยมัยได้เลือกเอกสารที่ทดลองในโดเมนวิทยาการคอมพิวเตอร์ (Computer Science) เฉพาะเจาะจงไปที่เอกสารเกี่ยวกับการสืบค้นความรู้จากฐานข้อมูลขนาดใหญ่ (Data Mining) มาจำนวน 100 เอกสาร เหตุผลของงานวิจัยที่จำเป็นต้องจำกัดเนื้อหาเอกสารให้อยู่ในวงแคบ เนื่องจากข้อจำกัดในการสร้างอภิธานคำศัพท์โดยใช้มนุษย์ ซึ่งอาจจะเกิดข้อผิดพลาดในการสร้างอภิธานคำศัพท์ได้ เทคนิคนี้อาจจะไม่เหมาะสมกับฐานข้อมูลเอกสารขนาดใหญ่ นอกจากนี้เทคนิคในงานวิจัยอาจจะเหมาะสำหรับเอกสารในเชิงวิทยาศาสตร์ที่มีการใช้คำศัพท์เฉพาะ เนื่องจากจะเป็นการสะดวกในการสร้างอภิธานคำศัพท์ ถ้านำมาใช้กับเอกสารเพื่อการตัดสินใจทางธุรกิจการวิจัยอาจจะให้ผลที่แตกต่าง เนื่องจากเอกสารเพื่อการตัดสินใจทางธุรกิจไม่มีการใช้คำศัพท์เฉพาะในเอกสารเท่าที่ควร แนวทางการวิจัยในอนาคตที่กำหนดไว้คือ จะเพิ่มการทดสอบกับเอกสารจำนวนมากขึ้น เพิ่มอภิธานคำศัพท์ให้ครอบคลุมในเรื่องที่กว้างขึ้น และทดสอบด้วยการเปลี่ยนข้อมูลเอกสาร นอกจากนี้จะศึกษาการปรับเปลี่ยนค่าถ่วงน้ำหนักที่แตกต่างกันระหว่างความสัมพันธ์แบบต่างๆ ในอภิธาน

ในงานวิจัยของ Qian และคณะ ในปี 2004 (Qian et al. 2004) เปรียบเทียบความเหมือนของการวัดระยะห่างยูคลิเดียน (Euclidean distance) และการวัดระยะห่างเชิงมุม (Cosine angle distance) สำหรับข้อสอบถามเพื่อนบ้านใกล้สุด (nearest neighbor queries) ด้วยข้อมูลทดสอบรูปภาพ ซึ่ง Qian และคณะได้เสนอวิธีการทดลองเปรียบเทียบสองวิธีการวัดระยะทางใน

แบบจำลองเวกเตอร์ (vector model) ด้วยการใช้ระยะห่างยูคลิเดียน (Euclidean distance: EUD) และการวัดระยะห่างเชิงมุม (Cosine angle distance: CAD) สำหรับเพื่อนบ้านใกล้สุดในข้อมูลมิติระดับสูง ในงานวิจัยของ Qian และคณะได้ทดลองการเปรียบเทียบจากการวิเคราะห์ตามทฤษฎี และการทดลองจริงด้วยการประยุกต์ใช้กับการค้นคืนรูปภาพ (Image retrieval) Qian และคณะได้เสนอว่าเหตุผลที่เปรียบเทียบ 2 วิธีการวัดนี้ เนื่องจากทั้งสองวิธีเป็นการวัดระยะห่างเหมือนกัน และใช้ในรูปแบบจำลองเวกเตอร์ (vector model) ที่เหมือนกัน Qian และคณะเสนอว่าในการค้นคืนรูปภาพนั้นนิยมที่จะใช้การวัดความเหมือนด้วยวิธีระยะห่างยูคลิเดียน (Euclidean distance) ซึ่งหากนำการวัดระยะห่างเชิงมุม (Cosine angle distance) มาประยุกต์ใช้ผลการทดลองที่ออกมาจะสามารถหาค่าความเหมือนได้ใกล้เคียงกับวิธีระยะห่างยูคลิเดียนหรือไม่

Qian และคณะได้ทดลองเปรียบเทียบวิธีการหาความเหมือนด้วยระยะห่างทั้ง 2 วิธี ด้วยการทดลองจากการวิเคราะห์ตามทฤษฎีและการทดลองประยุกต์ใช้กับการค้นคืนรูปภาพ วัดประสิทธิภาพผลการทดลองด้วยการหาค่าความแม่นยำ (precision) และค่าความระลึก (recall) การทดลองที่ได้จากการวิเคราะห์ตามทฤษฎีนั้น Qian และคณะได้ทดลองด้วยการใช้ชุดข้อมูลที่ได้จากการสุ่มขึ้นมา 50,000 ชุดข้อมูล และข้อสอบถามทดลองด้วยวิธีการสุ่มขึ้นมา 30 ชุดข้อสอบถาม ด้วยการเปรียบเทียบในมิติการทดลองที่แตกต่างกันคือ 2, 4, 8, 16, 32, 64 และ 128 และการกำหนดข้อสอบถามเพื่อนบ้านใกล้สุด k ตัวที่แตกต่างกันไป คือ 10, 20, 100, 500 และ 1000 การเปรียบเทียบความเหมือนของการวัดระยะห่างยูคลิเดียนและการวัดระยะห่างเชิงมุมตามการวิเคราะห์ทฤษฎี โดยการเปลี่ยนค่าของมิติไปพร้อมๆกับการเปลี่ยนค่าของข้อสอบถามเพื่อนบ้านใกล้สุด k ตัว ผลการทดลองพบว่า ในช่วงแรกของการใช้มิติที่มีค่าต่ำนั้น วิธีการวัดระยะห่างยูคลิเดียน และการวัดระยะห่างเชิงมุมจะมีผลการวัดความเหมือนที่แตกต่างกัน (คำนวณด้วยค่าเปอร์เซ็นต์ของผลความเหมือน (intersection)) โดยวิธีการวัดระยะห่างยูคลิเดียนจะให้ผลการค้นคืนที่ดีกว่าการวัดระยะห่างเชิงมุม แต่เมื่อเพิ่มมิติให้สูงขึ้นไปเรื่อยๆผลการวัดความเหมือนด้วยวิธีการวัดระยะห่างยูคลิเดียน และการวัดระยะห่างเชิงมุม มีค่าความเหมือนที่ออกมาเหมือนกันมาก Qian และคณะได้เสนอว่าจากการวิเคราะห์ตามทฤษฎีนั้นสามารถสรุปได้ว่าเมื่อพิจารณาวิธีการวัดระยะห่างยูคลิเดียน และการวัดระยะห่างเชิงมุมจะมีความเหมือนกันมากเมื่อใช้กับข้อสอบถามเพื่อนบ้านใกล้สุดที่มีมิติระดับสูง

การทดลองเปรียบเทียบวิธีการวัดระยะห่างยูคลิเดียน และการวัดระยะห่างเชิงมุมด้วยการประยุกต์ใช้กับการค้นคืนรูปภาพ Qian และคณะ ได้ทดลองด้วยการใช้ฐานข้อมูลรูปภาพประกอบด้วย 6344 สีของรูปภาพสัตว์และรูปภาพธรรมชาติ และ 18 รูปของสัตว์ถูกเลือกเป็น

ข้อสอบถามรูปภาพ การทดลองจะแสดงประสิทธิภาพของวิธีวัดระยะห่างเชิงมุม (Cosine angle distance) ด้วยการวัดประสิทธิภาพค่าความแม่นยำและค่าความระลึก ผลการทดลองพบว่าการค้นคืนรูปภาพด้วยการหาค่าความเหมือนเมื่อใช้การวัดระยะห่างเชิงมุม (Cosine angle distance) นั้นได้ค่าความแม่นยำและค่าความระลึกออกมาไม่แตกต่างไปจากการใช้วิธีการวัดระยะห่างยูคลิเดียนซึ่งเป็นวิธีที่ใช้ในการค้นคืนรูปภาพอยู่แล้ว Qian และคณะได้เสนอว่าวิธีการวัดระยะห่างเชิงมุม (Cosine angle distance) สามารถนำมาประยุกต์ใช้กับการค้นคืนรูปภาพได้ไม่แตกต่างไปจากวิธีการวัดระยะห่างยูคลิเดียน

ข้อจำกัดงานวิจัยของ Qian และคณะ คือ ข้อมูลที่ใช้ในการทดลอง Qian และคณะ ได้เสนอข้อมูลที่เกิดจากการสุ่มขึ้นมาไม่ได้เป็นข้อมูลที่มีนำมาจากฐานข้อมูลที่มีอยู่จริงหรือฐานข้อมูลมาตรฐานที่เหมาะสำหรับนำมาทดสอบการค้นคืนรูปภาพ ซึ่งผู้วิจัยมีความเห็นว่า ถ้านำงานวิจัยของ Qian และคณะ มาใช้กับข้อมูลซึ่งเป็นข้อมูลในฐานข้อมูลที่เป็นจริงหรือใช้กับฐานข้อมูลที่เป็นข้อความแล้วนั้น ผลการทดลองที่ได้อาจจะมีความแตกต่างกันก็เป็นได้ และ Qian และคณะได้ทดลองเปรียบเทียบความเหมือนของวิธีการวัดระยะห่างยูคลิเดียน และการวัดระยะห่างเชิงมุมด้วยการประยุกต์กับระบบการค้นคืนรูปภาพนั้น ข้อมูลที่ใช้ในการทดลองก็จะเป็นฐานข้อมูลรูปภาพเพียงอย่างเดียวเท่านั้น ซึ่งถ้านำมาใช้กับฐานข้อมูลที่เป็นเอกสาร และนำมาใช้กับระบบการค้นคืนเอกสาร ผลการทดลองที่ได้อาจแตกต่างกันไปจากงานวิจัยของ Qian และคณะ

ในงานวิจัยของ Greenwood ปี 2002 เป็นงานวิจัยส่งเสริมระบบการค้นคืนเอกสารด้วยเทคนิคปริภูมิเวกเตอร์ (Greenwood 2002) Greenwood ได้ทดสอบการค้นคืนเอกสารเทคนิคปริภูมิเวกเตอร์ด้วยการทดลองใช้วิธีการกำจัดคำยกเว้น,การลดรูปคำร่วมกับวิธีการให้ค่าน้ำหนักของคำในเอกสารและข้อสอบถามที่แตกต่างกัน ซึ่งวิธีการเหล่านี้ต่างก็เป็นกระบวนการพื้นฐานที่ใช้ในระบบการค้นคืนเอกสารอยู่แล้ว แต่เพื่อทดสอบการค้นคืนเอกสารให้สามารถค้นคืนเอกสารที่เกี่ยวข้องตรงตามความต้องการของผู้ใช้ด้วยเทคนิคปริภูมิเวกเตอร์ ซึ่งวัดประสิทธิภาพผลการค้นคืนเอกสารด้วยค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ด้วยการใช้ชุดข้อมูลทดสอบเอกสารและข้อสอบถามของ CACM (ftp://ftp.cs.cornell.edu/pub/smart/cacm/)

แนวทางในงานวิจัยของ Greenwood คือ ทดลองเปรียบเทียบวิธีการให้ค่าน้ำหนักของคำในเอกสารและข้อสอบถามร่วมกับวิธีการกำจัดคำยกเว้น (stop words removal), การลดรูปคำ (stemming) โดย Greenwood ได้ทดสอบการให้ค่าน้ำหนักของคำด้วยความถี่ของคำ (Term Frequency: tf) และความถี่แบบผกผัน (Inverse Document Frequent: idf) ในงานวิจัย Greenwood ได้ทดลองการค้นคืนด้วยวิธี 3 วิธีด้วยกันคือ

1.การค้นคืนเอกสารเทคนิคปริภูมิเวกเตอร์ด้วยขั้นตอนวิธีการกำจัดคำยกเว้น (stop words removal) การลดรูปคำ (stemming) และการให้ค่าน้ำหนักของคำด้วยความถี่ของคำ (Term Frequency: tf) โดยการวัดความเหมือนเชิงมุม (Cosine similarity)

2.การค้นคืนเอกสารเทคนิคปริภูมิเวกเตอร์ด้วยขั้นตอนวิธีการกำจัดคำยกเว้น (stop words removal) การลดรูปคำ (stemming) และการให้ค่าน้ำหนักของคำด้วยความถี่ของคำ (Term Frequency: tf) ร่วมกับความถี่แบบผกผัน (Inverse Document Frequent: idf) โดยการวัดความเหมือนเชิงมุม (Cosine similarity)

3.การค้นคืนเอกสารเทคนิคปริภูมิเวกเตอร์ด้วยการให้ค่าน้ำหนักของคำด้วยความถี่ของคำ (Term Frequency: tf) โดยการวัดความเหมือนเชิงมุม (Cosine similarity)

Greenwood ได้ทดลองเปรียบเทียบทั้ง 3 วิธีที่แตกต่างกัน ซึ่งทั้ง 3 วิธีนี้ได้มีขั้นตอนการวัดค่าความเหมือนด้วยการวัดความคล้ายคลึงเชิง Cosine similarity ที่เหมือนกัน และใช้ข้อมูลทดสอบของ CACM (CACM collection) เป็นเอกสารที่เกี่ยวกับบทความในวารสารของ Communications of the ACM โดยบทความจะเฉพาะเจาะจงไปในสาขาวิชาของวิทยาการคอมพิวเตอร์ (Computer Science) ซึ่งประกอบด้วยเอกสาร 3204 เอกสาร และ 64 ข้อสอบถาม ผลการทดลองวัดค่าความแม่นยำและค่าความระลึก พบว่า การทดลองค้นคืนเอกสารด้วยขั้นตอนวิธีการกำจัดคำยกเว้น (stop words removal) การลดรูปคำ (stemming) และการให้ค่าน้ำหนักของคำด้วยความถี่ของคำ (Term Frequency: tf) ร่วมกับความถี่แบบผกผัน (Inverse Document Frequent: idf) ให้ค่าความแม่นยำและค่าความระลึกที่มีประสิทธิภาพมากกว่าทั้งการทดลองอีก 2 วิธี โดยวิธีการค้นคืนเอกสารด้วยการให้ค่าน้ำหนักของคำด้วยความถี่ของคำ (Term Frequency: tf) โดยไม่ใช้วิธีการกำจัดคำยกเว้น (stop words removal),วิธีการลดรูปคำ (stemming)และความถี่แบบผกผัน (Inverse Document Frequent: idf) ร่วมด้วย จะให้ค่าความแม่นยำและค่าความระลึกที่ต่ำที่สุด ดังนั้น Greenwood ได้เสนอว่า ระบบการค้นคืนเอกสารเทคนิคปริภูมิเวกเตอร์ที่มีประสิทธิภาพต้องค้นคืนเอกสารด้วยขั้นตอนวิธีการกำจัดคำยกเว้น (stop words removal) การลดรูปคำ (stemming) และการให้ค่าน้ำหนักของคำด้วยความถี่ของคำ (Term Frequency: tf) ร่วมกับความถี่แบบผกผัน (Inverse Document Frequent: idf) และใช้วิธีวัดความเหมือนเชิงมุม (Cosine similarity) ค้นคืนเอกสาร

ในงานวิจัยของ Leuski ปี 2002 (Leuski 2002) เป็นงานวิจัยการประเมินการจัดกลุ่มเอกสารสำหรับการค้นคืนเอกสาร Leuski พิจารณาปัญหาของการจัดการลำดับของการแสดงผลที่เป็นผลลัพธ์จากการค้นคืน Leuski ได้ศึกษาประสิทธิภาพของการจัดการเอกสารที่ค้นคืนให้กับผู้ใช้

เพื่อที่จะระบุความสัมพันธ์ของเอกสารที่ค้นคืนมาได้อย่างรวดเร็ว โดย Leuski ได้นำเอาวิธีการจัดกลุ่มเอกสาร (Clustering) มาทดลองจัดกลุ่มให้กับเอกสารที่ถูกค้นคืนก่อนการแสดงต่อผู้ใช้ สมมติฐานการจัดกลุ่มของ Leuski คือเอกสารที่มีความสัมพันธ์ใกล้กันจะมีความสัมพันธ์กับข้อสอบถามไปในแนวทางเดียวกัน

แนวทางการวิจัยของ Leuski คือ ทดลองการจัดกลุ่มให้กับเอกสารที่ถูกค้นคืน เพื่อทดสอบประสิทธิภาพการค้นคืนด้วยค่าเฉลี่ยความแม่นยำ โดย Leuski ต้องการเปรียบเทียบประสิทธิภาพของการแสดงผลเอกสารตามกลุ่มที่จัดให้กับการแสดงผลแบบปกติ งานวิจัยของ Leuski ได้ใช้ฐานข้อมูลทดลองจาก TREC ad-hoc และใช้เทคนิคปริภูมิเวกเตอร์ในการค้นคืนเอกสารซึ่งกำหนดค่า threshold ในการค้นคืนเอกสารด้วยค่าที่ทำให้การค้นคืนมีประสิทธิภาพมากที่สุด และเทคนิคที่ใช้ในการจัดกลุ่มเอกสารคือ เทคนิค Group Average Algorithm การทดลองนี้จะรับผลสะท้อนกลับจากผู้ใช้เพื่อให้ผู้ใช้ระบุเอกสารที่มีความเกี่ยวเนื่องกับความต้องการ โดย Leuski พิจารณาว่าการจัดกลุ่มเอกสารค้นคืนสามารถช่วยให้ผู้ใช้ไม่เสียเวลาในการพิจารณาเอกสาร คือ เมื่อผู้ใช้พิจารณาเอกสารที่ปรากฏในลำดับแรกของกลุ่มแล้วพบว่าไม่ใช้เอกสารที่เกี่ยวเนื่องกับความต้องการ ผู้ใช้ก็จะข้ามไปพิจารณาเอกสารกลุ่มต่อไป ในขณะที่ในการแสดงผลการค้นคืนแบบปกติ ผู้ใช้จะต้องพิจารณารายการเอกสารค้นคืนทั้งหมดทีละรายการไป ซึ่งผู้ใช้อาจจะเสียเวลาในการพิจารณา ในงานวิจัย Leuski ได้จำลองสถานการณ์ให้ผู้ใช้กำหนดเอกสารลำดับแรกที่เกี่ยวเนื่องกับความต้องการจากรายการที่แสดงผลการค้นคืนแบบปกติ จากนั้นสลับไปกำหนดเอกสารลำดับแรกที่เกี่ยวเนื่องกับความต้องการจากรายการเอกสารที่แสดงผลตามกลุ่มที่จัดให้ ผลการทดลองสรุปได้ว่าการจัดกลุ่มเอกสารค้นคืนมีค่าเฉลี่ยประสิทธิภาพร่วมกับการให้ผลสะท้อนกลับจากผู้ใช้ที่ดีกว่าการแสดงรายการลำดับเอกสารค้นคืนแบบปกติ