บทที่ 2

แนวคิดและทฤษฎีที่เกี่ยวข้อง

ในอดีตการใช้งานคอมพิวเตอร์จะดำเนินการกับงานทางด้านวิทยาศาสตร์ ซึ่งข้อมูลส่วนมากจะเป็นตัวเลข เพื่อช่วยแก้ปัญหาการคำนวณที่ซับซ้อน ต่อมาคอมพิวเตอร์ได้ขยายตัวสู่งานทางธุรกิจ ซึ่งข้อมูลจะประกอบด้วยทั้งตัวเลข และตัวอักษร แต่ทั้งนี้และทั้งนั้นลักษณะข้อมูลยังต้องมีการกำหนดโครงสร้างที่แน่นอน โปรแกรมประยุกต์จะดำเนินการกับข้อมูลที่มีโครงสร้างและผลิตสารสนเทศเพื่อประโยชน์ต่อธุรกิจ ทั้งในระดับปฏิบัติการ ระดับบริหาร และระดับเพื่อการตัดสินใจสั่งการ ลักษณะโครงสร้างของข้อมูลเริ่มตั้งแต่ลักษณะของแฟ้มข้อมูล มีการจัดการแฟ้ม (file management) เกิดขึ้นหลายแบบ เช่น แฟ้มลำดับ แฟ้มสุ่ม แฟ้มสัมพันธ์ และแฟ้มลำดับดรรชนี เป็นต้น ต่อมาได้เกิดการจัดการฐานข้อมูล (database management) เพื่อรวมข้อมูลทั้งองค์กรให้มีความสัมพันธ์แนบแน่น เพื่อลดปัญหาความซ้ำซ้อนของข้อมูล และเพื่อร่วมกันใช้ข้อมูลที่สัมพันธ์กันได้ทั้งองค์กร โครงสร้างของการจัดการฐานข้อมูลก็มีเกิดขึ้นหลายแบบ เช่น ฐานข้อมูลเชิงเครือข่าย (network database) ฐานข้อมูลเชิงลำดับชั้น (hierarchy database) และฐานข้อมูลเชิงสัมพันธ์ (relational database) เป็นต้น

2.1 การจัดการข้อความ และระบบฐานข้อมูลข้อความ

(Text Management and Text Database Systems)

จากที่กล่าวมาข้างต้น จะเห็นว่าระบบการคำนวณแบบดั้งเดิม (Traditional computing systems) เก็บข้อมูลที่มีโครงสร้างที่แน่นอน และใช้โปรแกรมคอมพิวเตอร์เพื่อผลิตสารสนเทศ แต่แนวโน้มในปัจจุบัน การจัดการกับข้อมูลที่เป็นข้อความในลักษณะภาษาธรรมชาติ ซึ่งมีโครงสร้างที่ซับซ้อน และแตกต่างจากเทคนิคการประมวลผลข้อมูลแบบดั้งเดิม กำลังทวีบทบาทยิ่งขึ้น ประเด็นสำคัญที่จะชี้ให้เห็นถึงแนวโน้มดังกล่าวมี 2 ประเด็น คือ [2]

ก. ในปัจจุบันเครื่องประมวลคำ (word processor) จดหมายอิเล็กทรอนิกส์ (electronic mail) และระบบแผงข่าว (bulletin board systems) ได้ถูกนำเข้ามา

ใช้ในสำนักงาน ทำให้มีแนวโน้มที่จะปรับเปลี่ยนวิธีการทำงานในสำนักงาน ให้มีความรวดเร็ว ประหยัด และคล่องตัวมากยิ่งขึ้นกว่าระบบเก่าที่ใช้กระดาษเป็นหลักในการติดต่อสื่อสาร ทำให้มีเอกสารมากมายวนเวียนไปมาในสำนักงาน เป็นภาระที่หนักในการเก็บรักษา และเรียกใช้ในภายหลัง

ข. ปัจจุบันปริมาณสิ่งพิมพ์ และสำนักพิมพ์เพิ่มจำนวนขึ้นมากมาย ประมาณว่ามีหนังสือใหม่พิมพ์ออกมาในภาษาอังกฤษ และภาษายุโรปอื่น ๆ ในแต่ละปีมากกว่า 420,000 เล่มจากสำนักพิมพ์ต่าง ๆ กว่าหนึ่งหมื่นแห่งวารสารในภาษาเดียวกันกว่า 100,000 ฉบับ บทความเฉพาะด้าน เช่น บทความเฉพาะในอุตสาหกรรมทางเทคโนโลยีขั้นสูงในสหรัฐฯ ณ นี้มีบทความเฉพาะด้านนี้เพียงอย่างเดียวถึง 35,000 ชิ้นต่อเดือน ซึ่งถูกตีพิมพ์ในสิ่งพิมพ์ต่าง ๆ ถึง 11,000 ฉบับ นอกจากนั้นยังมีสิ่งพิมพ์อื่น ๆ อีกมาก เช่น ตำรา รายงานวิจัย เอกสารการจัดการบริหาร เอกสารการดำเนินงานทั้งในภาครัฐและเอกชน สิ่งเหล่านี้เป็นสารสนเทศซึ่งอยู่ในรูปตัวหนังสือ ที่มีคุณค่าและความสำคัญ ดังนั้นการให้ได้มาซึ่งสารสนเทศนับเป็นปัญหาร่วมกันของหลาย ๆ ฝ่าย การติดตามสารสนเทศจึงนับเป็นงานหนักและซับซ้อน

มีการตั้งคำถามว่า ทำไมการประยุกต์ทางการจัดการข้อความมีเพียงส่วนน้อยของการประมวลผลทั่ว ๆ ไป และทำไมในปัจจุบันนักวิจัยหลายท่านได้ทำนายไว้ว่า อัตราการเจริญเติบโตของการประยุกต์ทางการจัดการข้อความจะมีมากกว่า 60 เปอร์เซ็นต์ ต่อปี [2]

เหตุผลส่วนหนึ่งที่จะตอบคำถามเกี่ยวกับการขยายตัวของการประยุกต์ทางการจัดการข้อความก็คือ ราคาของสื่อบันทึกข้อมูลที่ลดลง ในทางตรงกันข้าม ความสามารถในการจัดเก็บกลับสูงขึ้นอย่างมากมาย เพียงพอที่จะจัดเก็บข้อมูลข้อความซึ่งต้องการเนื้อที่มาก ฐานข้อมูลทางธุรกิจสำหรับผู้ใช้ 12 คน ต้องการเนื้อที่จัดเก็บข้อมูลประมาณ 10 ล้านตัวอักษร ฐานข้อมูลตัวเลขขนาดใหญ่เริ่มต้นด้วยเนื้อที่จัดเก็บประมาณ 50 ล้านตัวอักษร สำหรับนักพิมพ์ดีดที่มีความชำนาญในเวลาหนึ่งปี จะสามารถใช้เครื่องประมวลคำพิมพ์เอกสารกระดาษ A4 ได้ประมาณ 3,500 ถึง 4,000 แผ่น และความต้องการเนื้อที่ในการจัดเก็บฐานข้อมูลข้อความจะเริ่มที่ขนาดถึง 100 ล้านตัวอักษร เมื่อราคาของอุปกรณ์จัดเก็บข้อมูลลดลงอย่างมากและการเข้าสู่ยุคของแผ่นจานแสงที่สามารถลบได้ (erasable optical discs) ทำให้ลดปัญหาของความต้องการเนื้อที่ขนาดใหญ่ลงได้

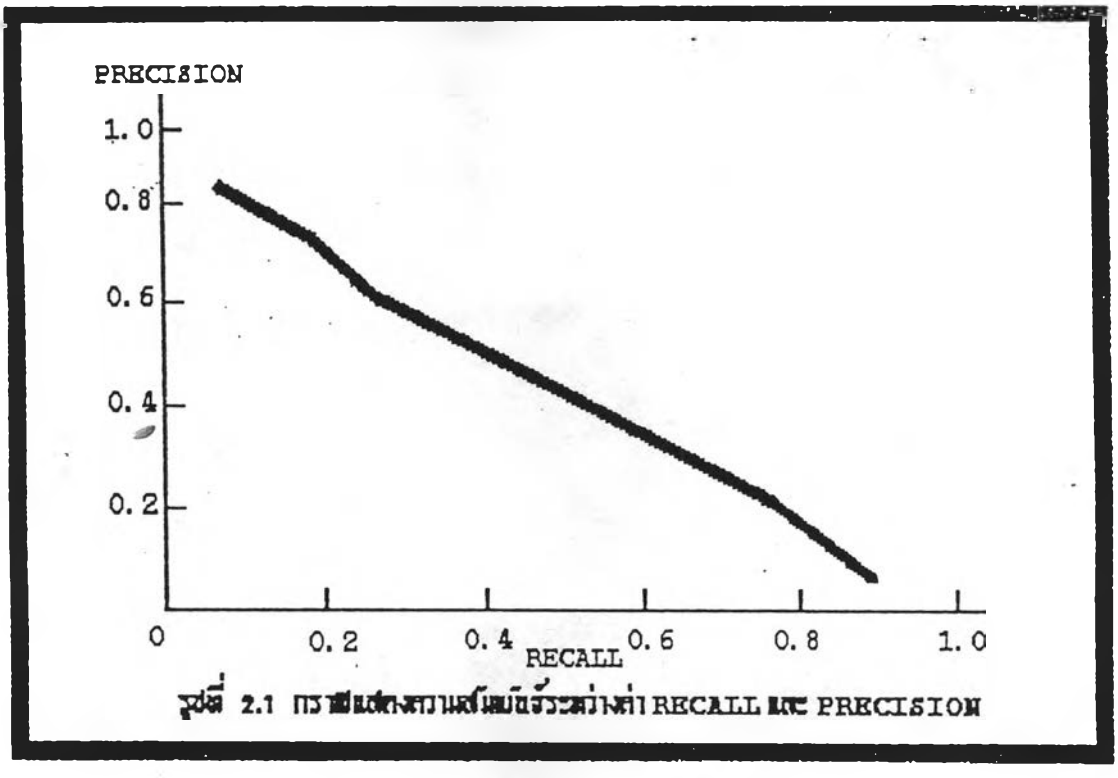
เหตุผลที่สองที่ทำให้การประยุกต์ทางการจัดการข้อความมีเพียงส่วนน้อยในอดีตคือ ข้อความในลักษณะภาษาธรรมชาติ ไม่มีโครงสร้างที่แน่นอนสำหรับภาษาคอมพิวเตอร์ เป็นเรื่องยากที่จะทำให้ข้อความมีความยาว หรือมีโครงสร้างตายตัว ให้เหมือนกับแฟ้มข้อมูลที่ใช้ในภาษาคอมพิวเตอร์ทางธุรกิจ เช่น ภาษา COBOL

เหตุผลที่สามที่หน่วงระบบการจัดการข้อความในหน่วยงานคอมพิวเตอร์ยุคเก่าคือ การค้นคืนข้อความ (text retrieval) เกี่ยวข้องกับความน่าจะเป็น คำตอบของการค้นหามีโอกาสขาดความครบถ้วนสมบูรณ์ (lacks absolute answer) ตรงข้ามกับการค้นคืนข้อมูล (data retrieval) ซึ่งจะมีคำตอบที่สมบูรณ์ (เช่น ต้องการค้นหาพนักงานที่มีเงินเดือนมากกว่า 9,000 บาท) ในการค้นคืนข้อความจะมีการวัดปริมาณที่สำคัญ 2 ค่า คือ [1]

$$\text{Precision} = \frac{\text{จำนวนเอกสารตรงกับกรณีที่ค้นคืนออกมาได้}}{\text{จำนวนเอกสารทั้งหมดที่ค้นคืนออกมา}}$$

$$\text{Recall} = \frac{\text{จำนวนเอกสารตรงกับกรณีที่ค้นคืนออกมาได้}}{\text{จำนวนเอกสารตรงกับกรณีทั้งสิ้นในฐานข้อมูล}}$$

precision จะเป็นปริมาณที่แสดงว่าการค้นคืนเอกสารจะได้ตรงกับกรณีเพียงใด เช่น ถ้าค้นคืนเอกสารออกมาได้ N ฉบับ และมีเอกสารอยู่ R ฉบับที่ตรงกับกรณี ดังนั้นค่า precision จะเป็น R/N หรือเป็นโอกาสของเอกสารที่ค้นคืนออกมาที่ตรงกับกรณี ส่วน recall จะเป็นปริมาณที่แสดงความครอบคลุม (thoroughness) เช่น ถ้าฐานข้อมูลมีเอกสารที่ตรงกับกรณีทั้งสิ้น T ฉบับ และการค้นคืนสามารถดึงเอกสารที่ตรงกับกรณีได้ R ฉบับ ค่า recall จะเป็น R/T ในทางอุดมคติ การค้นคืนสารสนเทศต้องการให้ได้เฉพาะเอกสารที่ตรงกับกรณีออกมาเท่านั้น ซึ่งในกรณีนี้ค่าของทั้ง recall และ precision จะมีค่าเป็น 1 ซึ่งในทางปฏิบัติเป็นไปได้ยาก จากผลการวิจัย [1][3] พบว่าค่า recall และ precision มีความสัมพันธ์เป็นปฏิภาคผกผันและพบว่าผลรวมของค่า recall และ precision มีค่าอยู่ในช่วงประมาณ 1 และ 1.4 ดังแสดงในรูปที่ 2.1 ซึ่งอาจกล่าวสรุปได้ว่า หากต้องการให้ค่า recall สูง ค่า precision ก็จะต่ำ และในทางตรงกันข้าม หากต้องการให้ค่า precision สูง ค่า recall ก็จะต่ำ



รูปที่ 2.1 กราฟแสดงความสัมพันธ์ระหว่าง RECALL และ PRECISION

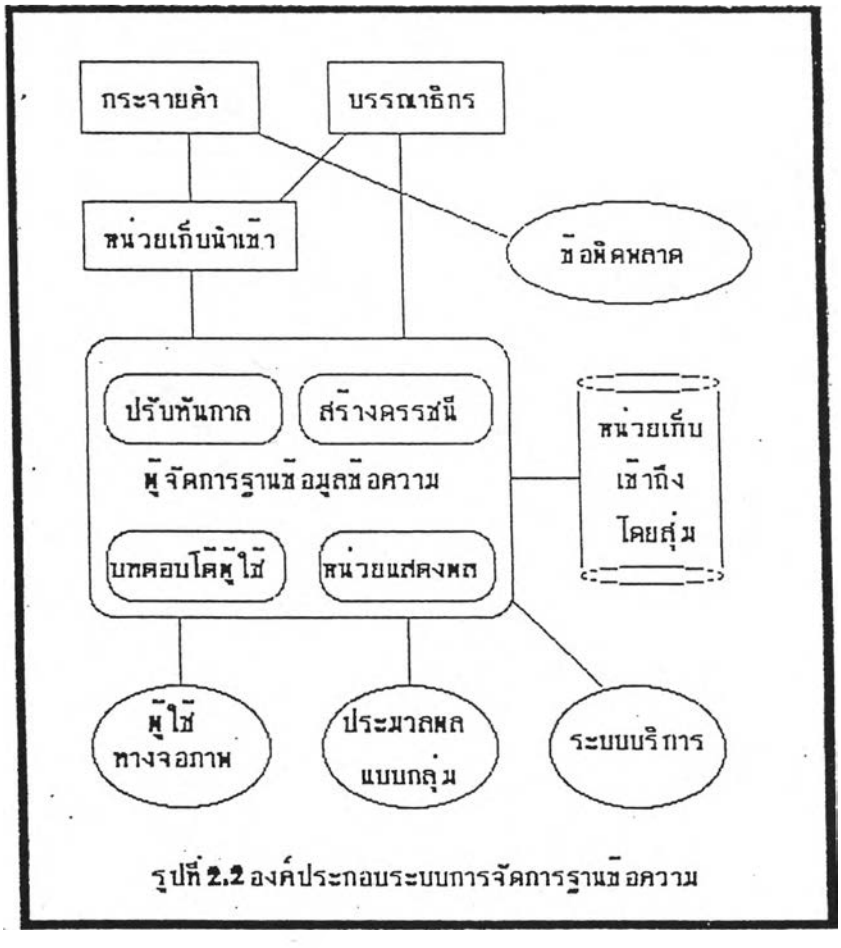
เหตุผลสุดท้ายคือ สารสนเทศบนเอกสารไม่ได้มีเพียงข้อความ แต่จะประกอบด้วยเส้น วาด ตาราง รูปภาพทั้งสีและขาวดำ เหล่านี้ทำให้การจัดการข้อความในอดีตเป็นไปด้วยความยากลำบาก

จากประเด็นดังกล่าว ผนวกกับราคาของอุปกรณ์จัดเก็บข้อมูลได้มีราคาลดต่ำลง ในขณะที่ความสามารถในการจัดเก็บเพิ่มสูงขึ้น รวมทั้งได้เกิดเทคโนโลยีในการจัดเก็บที่ใช้แสง ซึ่งทำให้สามารถเก็บข้อมูลได้ในปริมาณข้อมูลที่สูงขึ้นมาก ทำให้ในหลายปีที่ผ่านมาเกิดการประยุกต์ทางการจัดการข้อความมีอัตราการขยายตัวเพิ่มสูงขึ้น และลงสู่คอมพิวเตอร์ส่วนบุคคลด้วย

ระบบฐานข้อมูลตัวหนังสือเริ่มมีขึ้นมานับแต่ปี 1970 โดยพัฒนาเริ่มแรกจากการเรียงพิมพ์ด้วยคอมพิวเตอร์ของดรรชนีวารสารสิ่งพิมพ์ ในสาขาวิทยาศาสตร์และการแพทย์ และต่อมาได้พัฒนาสร้างเป็นระบบแม่ข่ายให้หมุนเข้ามาเชื่อมตรง (on-line dial up host systems) โดยเป็นโครงการของ NASA ซึ่งดำเนินการโดยบริษัท Lockheed และโครงการของ the National Library of Medical, Washington ดำเนินการโดย Systems Development Corporation ทั้งสองโครงการได้เปิดบริการในนามของ DIALOG และ ORBIT ตามลำดับ

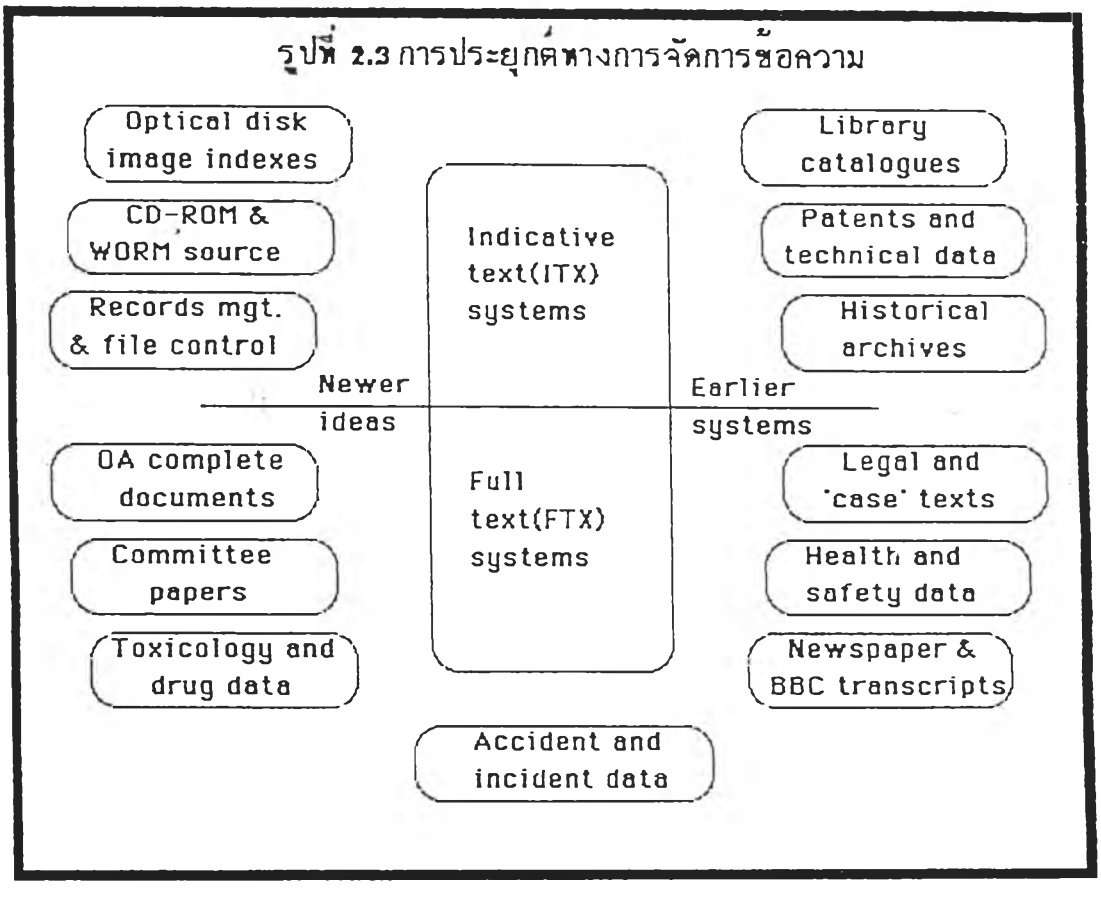
ในอีกทางหนึ่ง ได้มีการศึกษาวิจัยเป็นแนวทฤษฎีพื้นฐานที่ดีสำหรับการจัดการข้อความ โดย AERE Harwell, Battelle Laboratories in Columbus และ IBM ทั้งสามได้สร้างระบบฐานข้อมูลข้อความสำหรับวัตถุประสงค์ของตัวเองขึ้น จนเป็นระบบที่มีชื่อเสียง คือ STATUS, BASIS และ STAIRS ตามลำดับ โดย AERE และ IBM ให้ความสนใจเป็นพิเศษกับลักษณะข้อมูลที่เป็นข้อความเต็ม (full text) ระบบที่น่าสนใจอีกตัวหนึ่งคือกรณีของ BRS Information Technologies ซึ่งได้สร้างระบบ on-line host business โดยใช้ระบบ STAIRS ของ IBM ได้พัฒนาทั้งการจัดเก็บ และประสิทธิภาพของการตอบสนองต่อผู้เรียกใช้ จนในที่สุดกลายเป็นผลิตภัณฑ์ที่มีชื่อเสียงในนามของ BRS/SEARCH

ดังนั้นสามารถแบ่งผลิตภัณฑ์ทางการจัดการข้อความเป็น 2 ระบบใหญ่ ๆ คือ ระบบแม่ข่ายเชื่อมตรง (The on-line host systems) และ ผลิตภัณฑ์ชุดคำสั่งสำเร็จเฉพาะ (in-house package products) ระบบที่เป็นแม่ข่ายเชื่อมตรง จะมีปัญหาที่แตกต่างจากผลิตภัณฑ์ชุดคำสั่งสำเร็จเฉพาะ คือปัญหาเกี่ยวกับ การสอบถามระยะไกล (large scale remote enquiry) และ การคิดชำระเงินค่าบริการ (billing) แต่ถึงอย่างไรก็ตาม ทั้งสองระบบนี้มีองค์ประกอบหลักที่สำคัญเหมือนกัน ดังแสดงในรูปที่ 2.2 [2]



รูปที่ 2.2 องค์ประกอบระบบการจัดการฐานข้อความ

ถึงแม้ว่าระบบฐานข้อมูลข้อความในช่วงแรกมีวัตถุประสงค์ที่สำคัญในการจัดเก็บ และค้นคืน ในลักษณะข้อความเต็มของเอกสาร (full text of documents) ซึ่งในระบบแม่ข่ายเชื่อมตรง ส่วนมากจะมีความสามารถในลักษณะของข้อความเต็ม แต่อย่างไรก็ตามการประยุกต์ที่สำคัญที่มีขึ้นมาก่อนปี 1983 ซึ่งดำเนินการเกี่ยวกับการจัดเก็บ และค้นคืนข้อมูลบรรณานุกรม (the storage and retrieval of bibliographic data) เหล่านี้เป็นระเบียนรายการของหนังสือ วารสารในห้องสมุด หรือระเบียนการอ้างอิงรายงานทางวิชาการ สิทธิบัตร และ อื่น ๆ ซึ่งมีลักษณะการจัดเก็บที่ไม่ได้ใช้ข้อความเต็ม โดยมีลักษณะเป็นข้อความบ่งบอก (indicative text) การประยุกต์ทางการจัดการข้อความสามารถแสดงได้ดังรูปที่ 2.3 [2]



รูปที่ 2.3 การประยุกต์ทางการจัดการข้อความ

2.2 ระบบการค้นคืนสารสนเทศ

(Information Retrieval Systems หรือ IR Systems)

ระบบการค้นคืนสารสนเทศ เป็นระบบที่ออกแบบ เพื่อจุดประสงค์ในการตอบสนองต่อผู้ต้องการใช้สารสนเทศ จากเอกสารที่ได้เก็บรวบรวมไว้ ผู้ต้องการสารสนเทศจะใช้ภาษาสอบถาม (query language) เป็นเครื่องมือเพื่อแสดงความต้องการสารสนเทศ ระบบจะต้องสามารถแสดงถึงเอกสารที่มีสารสนเทศตรงตามความต้องการ ระบบที่มีประสิทธิภาพนั้น จำนวนเอกสารที่ดึงออกมาได้ จะต้องประกอบด้วยเอกสารที่ผู้ใช้ต้องการ หรือเรียกว่าเอกสารที่ตรงกับกรณีให้มากที่สุด (most relevant documents) และเป็นเอกสารที่ผู้ใช้ไม่ต้องการ หรือเรียกว่าเอกสารที่ไม่ตรงกับกรณีให้น้อยที่สุด (very few irrelevant documents)

ในสภาพการณ์ของเอกสารที่เก็บรวบรวมนี้ (เช่น หนังสือ วารสาร รายงานทางวิชาการ เป็นต้น) รายการสารสนเทศที่ต้องการสืบค้น จะไม่อยู่ในรูปแบบระเบียน (records) หรือ ทูเพิล (tuples) ดังเช่นที่พบในระบบการจัดการฐานข้อมูลแบบสัญนิยม (conventional database management systems) ดังนั้นระบบจะต้องมีตัวบ่งชี้ ซึ่งสะท้อนถึงเนื้อหาสาระของตัวเอกสาร เป็นรายการสารสนเทศที่ประกอบอยู่ในตัวของเอกสาร ซึ่งจะถือเป็นตัวแทนของเอกสาร และใช้เป็นกลไกสำหรับการสืบค้น ตัวแทนของเอกสารดังกล่าวนี้ เป็นเรื่องยากเรื่องหนึ่งสำหรับการที่จะหารายการสารสนเทศที่จะสามารถสะท้อนถึงเนื้อหาของเอกสารได้อย่างครบถ้วนสมบูรณ์

ปัญหาอีกข้อหนึ่ง ที่ใกล้เคียงกับการหาตัวแทนของเอกสารคือ ตัวแทนความต้องการสารสนเทศของผู้เรียกใช้ ระบบการค้นคืนสารสนเทศจะมีภาษาสอบถามให้ใช้ แต่ก็เป็นเรื่องยากเช่นกัน สำหรับการที่จะบ่งบอกอย่างครบถ้วนและสมบูรณ์ ถึงความต้องการของผู้ใช้โดยผ่านภาษาสอบถามของระบบ

ดังนั้นอาจกล่าวได้ว่า ในความเป็นจริงแล้ว เป็นไปไม่ได้ที่จะหาตัวแทนของเอกสารและตัวแทนของประโยคสอบถาม เพื่อให้สามารถค้นคืนสารสนเทศให้ออกมาเฉพาะเอกสารที่ตรงกับกรณี

2.3 คุณลักษณะของฐานข้อมูลข้อความ

ฐานข้อมูลข้อความสามารถแบ่งได้เป็น 3 ประเภท [4] คือ

2.3.1 ฐานข้อมูลบรรณานุกรม (Bibliographic database) เป็นฐานข้อมูลที่ใช้แพร่หลายในงานทางห้องสมุด ระเบียนบรรณานุกรมจะเป็นเสมือนตัวแทนของเอกสารหนึ่ง ๆ โดยจะบ่งบอกสารสนเทศต่าง ๆ ของเอกสาร แต่จะไม่เก็บข้อความเต็มของเอกสารไว้ ตัวอย่างดังรูป 2.4 ฐานข้อมูลบรรณานุกรมมีใช้กันมาตั้งเดิมตั้งแต่ยังเป็นระบบที่ทำด้วยมือ ถึงปัจจุบันเทคโนโลยีทางคอมพิวเตอร์สามารถสนองความต้องการ และช่วยแก้ปัญหาต่าง ๆ ของระบบที่ทำด้วยมือได้ จึงมีการนำเทคโนโลยีคอมพิวเตอร์ มาใช้ในการทำฐานข้อมูลบรรณานุกรม

รูปที่ 2.4 ตัวอย่างระเบียนฐานข้อมูลบรรณานุกรม
(Sample Bibliographic Database Record)

AN 84031234

TI Identification and Evaluation of Software for Microcomputer-base Inhouse Database.

AU Tenopir, Carol

CS University Of Hawaii at Manoa

JN Information Technology and Libraries

PY March 1984

CI Vol.3, Pgs. 21-34.

LA English

AB Discusses methods and sources for locating software packages suitable for inhouse textual databases. Describes the procedures for evaluation such software.

SH Information Storage And Retrieval; Computer Software; Bibliographic Database; Database Desige; Microcomputers.

ในลักษณะฐานข้อมูลภายใน (inhouse database) เพราะจะใช้เนื้อที่การจัดเก็บข้อมูลในคอมพิวเตอร์น้อย และสามารถใช้บนเครื่องไมโครคอมพิวเตอร์ได้

2.3.2 ฐานข้อมูลข้อความเต็ม (Full text database) เป็นฐานข้อมูลที่เก็บข้อความเต็มของเอกสารนั้น ๆ ไว้ด้วย เช่น บทความในวารสาร รายงานผลการวิจัย รายงานของบริษัท และจดหมายติดต่อต่าง ๆ เป็นต้น ดังตัวอย่างรูปที่ 2.5 ฐานข้อมูลข้อความเต็มจะมีค่าใช้จ่ายในการแปลงข้อมูลเพื่อให้คอมพิวเตอร์รับรู้ได้ (machine readable) และต้องใช้เนื้อที่ในการจัดเก็บมาก จึงเป็นอุปสรรคในการจัดทำฐานข้อมูลประเภทนี้ อย่างไรก็ตามหน่วยงานส่วนมากในปัจจุบัน ใช้เครื่องประมวลคำสร้างเอกสารต่าง ๆ ภายในหน่วยงาน ซึ่งทำให้สามารถตัดค่าใช้จ่ายในการแปลงข้อมูลลงไปได้ นับเป็นส่วนประกอบที่สำคัญสำหรับการทำฐานข้อมูลข้อความเต็ม

2.3.3 ฐานข้อมูลตัวอ้างอิง (referral or directory database) เป็นฐานข้อมูลสร้างไว้เพื่อจัดเก็บข้อมูลอ้างอิง เช่น ชื่อ ที่อยู่ เบอร์โทรศัพท์ และสารสนเทศอื่น ๆ เกี่ยวกับองค์กรหรือนิติบุคคล ตัวอย่างฐานข้อมูลอ้างอิงดังรูปที่ 2.6 เป็นฐานข้อมูลภายในที่เก็บข้อมูลอ้างอิงของบริษัทต่าง ๆ

คุณลักษณะของฐานข้อมูลข้อความที่แตกต่างจากฐานข้อมูลอื่น ๆ ที่ใช้ในงานธุรกิจ เช่น ฐานข้อมูลสินค้าคงคลัง มีดังต่อไปนี้

1.อักขระในเขตข้อมูลจะมีลักษณะเป็นตัวอักษรเลข (Alphanumeric character) ดังนั้นทุกเขตข้อมูลจะเป็นข้อความหรือสายของอักขระ (string) ถึงแม้เขตข้อมูลที่เป็นตัวเลขก็ถือเป็นตัวอักษร

2. โดยปกติฐานข้อมูลจะมีขนาดใหญ่

3. ระเบียนส่วนใหญ่จะมีหลาย ๆ เขตข้อมูล

รูปที่ 2.5 ตัวอย่างระเบียนฐานข้อมูลข้อความเต็ม
(Sample Full Text Database Record)

AN 87012345

TI Online Education: Planning for the Future.

AU Tenopir, Carol

CS University of Hawaii at Manoa

JN Online

PY January 1987

CI Vol. 11, Number 1, pgs. 65-66.

SH Education; Library and Information Science;
Online Database Searching

TX1 In the last decade, schools of library and information science have recognized the important role they play in education online intermediaries. A survey in 1982 found that at that time 76% of all schools accredited by the American Library Association includeed online searching in their curriculum.

TX2 Most of these courses to date have concentrated on the

. .
. .
. .

TX14 ...prepare students for all aspects of the information industry.

FN Stephen P. Harter and Carol H. Fenichel, "Online Searching in Library Education," Journal of Education for Librarianship 23 Summer 1982: 3-22.

CP The author.

รูปที่ 2.6 ตัวอย่างระเบียนฐานข้อมูลตัวอ้างอิง

(Sample Referral Database Record)

AN 123456

CN The XYZ Corporation

AD 123 Maple Street

CY Los Angeles ST CA ZP 90024

PH (213)555-6767

BU Manufacturer of contact lenses and other eye care products.

YR Started 1899

SC 6794

SA $1.5 million

OF John Smith, CEO; Mary Smith, President; Joe Smith, Vice President

EM 45 employees at this location

SQ 5000 square feet

---

4. แต่ละเขตข้อมูลจะมีความยาวแปรได้ บางเขตข้อมูลจะมีความยาวมาก

5. ระเบียนข้อมูลส่วนมากจะประกอบด้วยเขตข้อมูลที่เหมือนกัน (ในบางระเบียนอาจมีเขตข้อมูลไม่เท่ากันได้)

6. เขตข้อมูลบางตัวจะมีการทบซ้ำค่า (repeating values) เช่น เขตข้อมูลชื่อผู้แต่ง อาจมีได้หนึ่ง สอง หรือหลาย ๆ ผู้แต่ง แตกต่างกันไปในแต่ละระเบียน

7. ส่วนมากมีความต้องการที่จะสืบค้นสารสนเทศในทุก ๆ เขตข้อมูล หรือในบางส่วนก็ได้ หรือผสมกันหลายเขตข้อมูล

8. ผู้ใช้ฐานข้อมูลจะมีเป้าหลักเพื่อค้นหาสารสนเทศ ดังนั้นเทคนิคในการค้นหาที่ตระเตรียมไว้จะต้อง สะดวก ง่าย และฉลาด ต่อผู้เรียกใช้

2.4 กระบวนการกระจายคำ (The parsing process)

ในการออกแบบฐานข้อมูลข้อความ จำเป็นต้องกำหนดเขตข้อมูลที่ประกอบขึ้นเป็นระเบียนข้อมูล และต้องกำหนดว่าเขตข้อมูลใดบ้างที่จะสามารถสืบค้นได้ การกระจายคำเป็นกระบวนการเพื่อแบ่งแยกคำออกจากข้อความในเขตข้อมูลที่สามารถสืบค้นได้ โดยคำที่แยกได้อาจเป็นคำเดี่ยว หรือวลี คำที่ได้นี้จะเป็นข้อมูลที่ใช้ในกระบวนการค้นคืนสารสนเทศจากฐานข้อมูลข้อความ โครงสร้างของการจัดเก็บ และบำรุงรักษาคำที่ได้จะมีความสัมพันธ์กับกระบวนการค้นคืนซึ่งจะได้กล่าวถึงต่อไป ในส่วนนี้จะกล่าวถึงกระบวนการกระจายคำ ที่ใช้ในข้อความภาษาอังกฤษเพื่อประกอบการพิจารณาในการออกแบบระบบในขั้นรายละเอียดต่อไป

Carol Tenopir and Gerald Lundeen [4] ได้กล่าวถึงกระบวนการกระจายคำ โดยแบ่งเป็น 3 แบบ ที่มีใช้กันในโปรแกรมประยุกต์ทางการจัดการฐานข้อมูลข้อความ ซึ่งอาจเลือกกระทำได้ในแบบใดแบบหนึ่ง หรือผสมกันหลายแบบ ขึ้นอยู่กับความสามารถของแต่ละโปรแกรม

2.4.1 Word parsing จะทำการแยกคำเป็นคำเดี่ยว ๆ ออกมา โดยอาศัยเครื่องหมายวรรคตอนต่าง ๆ เป็นตัวช่วยแบ่งแยกคำ เช่น เขตข้อมูลชื่อเรื่อง "Developing Computer-Based Information Centers" จะสามารถแบ่งแยกคำออกได้เป็น 5 คำ เรียงตามลำดับตัวอักษรคือ BASED, CENTERS, COMPUTER, DEVELOPING และ INFORMATION โปรแกรมบางตัวจะไม่ถือยติภังค์ (Hyphen) เป็นเครื่องหมายวรรคตอน ดังนั้นจะสามารถแบ่งคำได้เพียง 4 คำ คือ CENTERS, COMPUTER-BASED, DEVELOPING และ INFORMATION คำบางคำที่แยกออกมาได้แต่ไม่มีความหมายที่จะเป็นประโยชน์ต่อการสืบค้น เช่น of, and, a, the ซึ่งถือเป็นคำทั่ว ๆ ไปสามารถตัดทิ้งได้ โปรแกรมประยุกต์จะสร้างคำทั่วไปนี้ไว้ให้ โปรแกรมประยุกต์บางตัวยอมให้ผู้ใช้เพิ่มเติมคำทั่วไปเพื่อให้เหมาะกับงานนั้น ๆ ทำให้ช่วยตัดคำที่ไม่มีประโยชน์ต่อการสืบค้นออกไปได้มากขึ้น

2.4.2 Phrase parsing คำที่จะใช้ในการสืบค้นบางครั้งจะไม่เป็นคำโดด แต่จะเป็นคำวลี ซึ่งจะมีเครื่องหมายวรรคตอนพวกช่องว่างอยู่ในคำวลี ดังนั้นโปรแกรมประยุกต์ส่วนมากจะอาศัยเครื่องหมายพิเศษใส่ขั้นไว้ที่ส่วนหัวและท้ายของคำวลี เพื่อให้โปรแกรมสามารถแบ่งแยกคำวลีออกมาได้

2.4.3 Combination parsing เป็นกระบวนการที่รวมของทั้ง word parsing และ pharse parsing ซึ่งจะทำให้การค้นหามีความอ่อนตัวมากยิ่งขึ้น แต่ก็ทำให้คำที่ได้จากการกระจายคำมีมากขึ้นด้วย

2.5 กระบวนการค้นคืน

การค้นคืนสารสนเทศจากฐานข้อมูลข้อความ เป็นการพยายามนำกลุ่มของเอกสารหรือเรื่องราวที่เห็นว่ามีความเหมือน หรือตรงกับกรณีกับข้อคำถามหรือประโยคของสารสนเทศจากผู้ต้องการสารสนเทศ ออกมาจากฐานข้อมูล ส่วนสำคัญที่เป็นตัวกำหนดถึงกำลังและความสามารถของระบบการค้นคืนมี 2 ส่วน คือ

ก. ภาษาสอบถาม (the query (command) language)

ข. กระบวนการค้นคืน (the matching (searching) process)

โดยสองส่วนนี้อยู่บนพื้นฐานของโครงสร้างของไฟล์ที่ใช้ ซึ่งจะมีผลกระทบโดยตรงต่อเวลาของการตอบสนอง และเนื้อที่ในการจัดเก็บข้อมูล

ในลำดับต่อไปจะกล่าวถึงเทคนิค และความสามารถทั่วไปที่พัฒนามากว่า 20 ปี สำหรับโปรแกรมสำเร็จรูปทางการจัดการฐานข้อมูลข้อความ รวมถึงข้อจำกัดของระบบการค้นคืนที่ใช้กันอยู่ และงานวิจัยที่จะช่วยแก้ปัญหาข้อจำกัดของระบบเก่า

2.5.1 ระบบการค้นคืนแบบสัญนิยม (Conventional Retrieval System)

เทคโนโลยีที่ใช้ในระบบการค้นคืนสารสนเทศได้เกิดขึ้นมาตั้งแต่ปี 1960 [3] เทคนิคการค้นหาที่ใช้กันในระบบการค้นคืนข้อความมีการเปลี่ยนแปลงที่น้อยมาก ระบบในปัจจุบันเกือบทั้งหมดยังคงอยู่บนพื้นฐานของเทคนิคการค้นคืนแบบบูล นับเป็นระบบการค้นคืนแบบสัญนิยม [1] โดยเทคนิคนี้จะอาศัยดรรชนีผกผัน (inverted indexes) และตรรกแบบบูล (boolean logic) ในการสร้างภาษาสอบถาม (query language)

2.5.1.1 ดรรชนีผกผัน (Inverted Indexes)

ปกติฐานข้อมูลข้อความจะสามารถสืบค้นข้อมูลได้ในหลาย ๆ ทาง คำค้นจะแฝงอยู่ในฟิลด์ต่าง ๆ และอยู่ในตำแหน่งต่าง ๆ กันได้โดยไม่มีกฎเกณฑ์ ดังนั้นในทางปฏิบัติจึงไม่สามารถยึดฟิลด์ใดเป็นคีย์ฟิลด์ได้ ตามธรรมดาเอกสารจะบันทึกลงฐานข้อมูลเป็นลำดับตามการนำข้อมูลเข้า ปริมาณข้อมูลมีแต่จะสะสมเพิ่มมากขึ้น ไม่เหมือนแฟ้มข้อมูลด้านธุรกิจที่มีทั้งการเข้าและออกของข้อมูล ลักษณะฐานข้อมูลข้อความจึงเป็นแฟ้มลำดับ การสืบค้นข้อมูลหากพิจารณาเมื่อฐานข้อมูลมีขนาดไม่ใหญ่นัก ก็จะต้องค้นหาโดยการเปรียบเทียบคำค้นกับฟิลด์ในระเบียนต่าง ๆ เป็นลำดับไปจากต้นจนหมดแฟ้ม ระเบียนที่มีคำค้นบรรจุอยู่ก็จะแสดงต่อผู้ค้น ในขั้นตอนแสดงผลอาจมีกระบวนการจัดลำดับระเบียนตามที่ผู้ค้นต้องการได้

ในกรณีฐานข้อมูลที่มีขนาดใหญ่จะไม่สามารถดำเนินการดังกล่าวข้างต้นได้ เพราะจะใช้เวลามาก วิธีการที่ใช้โดยการสร้างดรรชนีผกผัน

ดรรชนีผกผันคือคำค้นที่แยกออกมาจากระเบียนต่าง ๆ ในฐานข้อมูลโดยใช้กระบวนการกระจายคำที่ได้กล่าวมาแล้ว คำค้นที่แยกออกมาประกอบกับตัวชี้กลับไปตำแหน่งของระเบียนและอาจมีสารสนเทศอื่น ๆ ข้อมูลเหล่านี้นำมาสร้างเป็นแฟ้มดรรชนีผกผัน รูปที่ 2.7 และ 2.8 เป็นตัวอย่างแฟ้มลำดับของฐานข้อมูลบรรณานุกรม และแฟ้มดรรชนีผกผันตามลำดับ

รูปที่ 2.7 ระเบียนในแฟ้มข้อมูลลำดับ (Records in a Linear File)

AN 001

TI The Three Litter Pigs

AU Mother Goose

PU Wee Press

CY London

PY 1899

AB Real-life testing of house construction methods. Demonstrates advantages and disadvantages of straw, sticks, and bricks.

DE Swine, Miniature; Residential Architecture

ID Wolf, Big Bad

AN 002

TI The House That Jack Built

AU Mother Goose

PU Children's Book Company Inc.

CY New York

PY 1985

AB Construction tips for novices. Describes occupants of Javk's house.

DE Residential Architecture; Animals in fiction

ID Jack

รูปที่ 2.8 แฟ้มดรรชนีผกผัน (Inverted Index File)

VALUE	RECORD	FIELD	POSITION
1899	001	PY	01
1985	002	PY	01
advantages	001	AB	08
animals	002	DE	03
animals in fiction	002	DE	03-05
architecture	001	DE	04
	002	DE	02
bad	001	ID	03
big	001	ID	02
bricks	001	AB	15
built	002	TI	05
construction	001	AB	05
	002	AB	01
demonstrates	001	AB	07
describes	002	AB	05
disadvantages	001	AB	10
fiction	002	DE	05
house	001	AB	04
	002	TI	02
	002	AB	09
jack	002	TI	04
	002	ID	01
jack's	002	AB	08
.	.		
.	.		
.	.		

แฟ้มดรรชนีผกผันจะช่วยให้การค้นหารวดเร็วยิ่งขึ้น เพราะเมื่อต้องการค้นหาคำใด ระบบจะไปค้นหาที่แฟ้มดรรชนีผกผัน แทนที่จะลงไปค้นหาในแฟ้มเอกสาร นอกจากนั้นในแฟ้มดรรชนีผกผันยังได้ผนวกสารสนเทศอื่น ๆ เข้าไปอีก ดังเช่นในรูปที่ 2.8 ผนวกเลขเรียกของแต่ละระเบียน ชื่อฟิลด์ และตำแหน่งของคำดรรชนีที่ปรากฏอยู่ในฟิลด์ สารสนเทศเหล่านี้จะมีผลเพิ่มความสามารถในการสืบค้นได้ เช่น สามารถค้นคำที่อยู่ติดกัน หรือคำที่อยู่บนประโยคเดียวกัน เป็นต้น

เมื่อมีคำค้นจากผู้ต้องการค้นหา ระบบการค้นคืนจะไปค้นหาในแฟ้มดรรชนีผกผัน เมื่อคำค้นตรงกับคำดรรชนี ก็จะเก็บเลขเรียกของระเบียนไว้ ผลลัพธ์จึงเป็นรายการของเลขเรียกระเบียนเรียกว่าเซ็ท เซ็ทจะถูกสร้างขึ้นสำหรับแต่ละคำค้นหรือแต่ละประโยคสอบถาม ในการแสดงผลดอนแรกจะบอกเพียงจำนวนระเบียนที่พบ ผู้ค้นสามารถเรียกดูรายละเอียดของแต่ละระเบียนโดยใช้คำสั่งแสดงผล ซึ่งระบบจะใช้เลขเรียกซึ่งเก็บอยู่ในเซ็ทไปดึงระเบียนออกมาจากแฟ้มเอกสารอีกทอดหนึ่ง

2.5.1.2 ตรรกแบบบูล (Boolean logic)

ภาษาสอบถามที่ใช้ในฐานข้อมูลข้อความ สำหรับระบบการค้นคืนแบบสัญนิยมเกือบทั้งหมดจะอยู่บนพื้นฐานการใช้ตรรกแบบบูล ซึ่งสามารถจะสร้างภาษาสอบถามให้ซับซ้อนขึ้นโดยอาศัยตัวดำเนินการพื้นฐานที่สำคัญ 3 ตัว คือ AND OR และ NOT ซึ่งตัวดำเนินการเหล่านี้จะตรงกับการดำเนินการในเรื่องของเซ็ท คือ intersection union และ complement ตามลำดับ

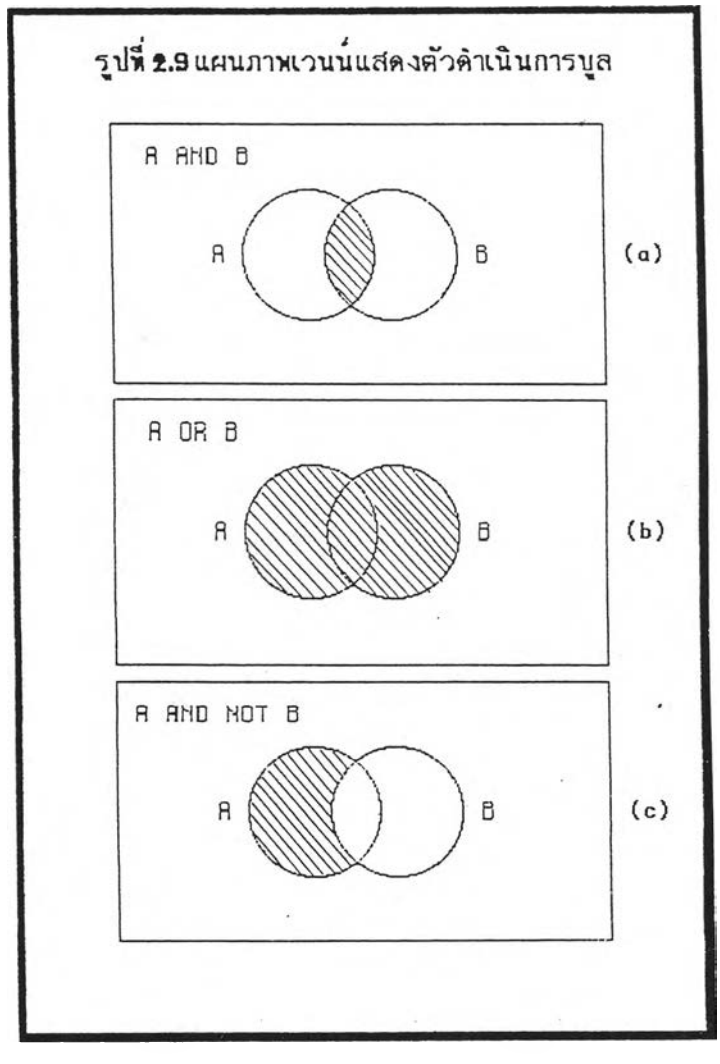
การดำเนินการแบบบูลสามารถแสดงให้เห็นได้อย่างง่าย โดยอาศัยแผนภาพของเวนน์ (Venn diagrams) ดังรูปที่ 2.9 ซึ่งแสดงถึงตัวดำเนินการพื้นฐานทั้ง 3 ตัว คือ

AND เป็นตัวดำเนินการทำให้เซ็ทที่ได้จากการค้นคืนแคบลง รูปที่ 2.9 A (ส่วนแรเงา) เป็นผลจากการดำเนินการ A AND B โดยสมาชิกของเซ็ทที่ได้ต้องเป็นสมาชิกอยู่ทั้งในเซ็ท A และ เซ็ท B

OR เป็นตัวดำเนินการทำให้เซ็ทของการสืบค้นขยายครอบคลุม รูปที่ 2.9 B (ส่วนแรเงา) เป็นผลจากการดำเนินการ A OR B โดยสมาชิกของเซ็ทที่ได้จะครอบคลุมถึงสมาชิกของเซ็ทที่อยู่ใน A หรือ B หรืออยู่ทั้งใน A และ B

NOT เป็นตัวดำเนินการที่จะสร้างเซ็ทซึ่งสมาชิกต้องไม่อยู่ในเซ็ทที่บ่งบอก กรณีนี้มีความหมายเช่นเดียวกับ AND NOT รูปที่ 2.9 C (ส่วนแรเงา) เป็นผลจากการดำเนินการ A AND NOT B โดยสมาชิกของเซ็ทที่ได้ จะอยู่เฉพาะใน A และต้องไม่อยู่ใน B ในบางระบบ NOT จะเป็นตัวดำเนินการที่ใช้นำหน้าได้ เช่น NOT A จะให้ผลลัพธ์ทุกระเบียนในฐานข้อมูลที่ไม่มีเทอม A อยู่

ระบบส่วนใหญ่จะมีความสามารถ ในการใช้ตัวดำเนินการซ้อนกันได้หลายตัว เพื่อสร้างประโยคสอบถามที่ซับซ้อนขึ้น โดยจะใช้วงเล็บเปิดและปิดช่วยในการควบคุมลำดับของการดำเนินการ เช่น (A OR (B AND C) OR (D AND C)) AND (E OR F) เป็นต้น



รูปที่ 2.9 แผนภาพเวนน์แสดงตัวดำเนินการบูล

2.5.1.3 ข้อจำกัดของการสืบค้นแบบบูลีน

ก. ผู้ใช้ทั่วไปจะไม่สามารถสร้างประโยคสอบถามโดยใช้ตัวดำเนินการบูล (AND, OR และ NOT) เพื่อถ่ายทอดออกมาตามความต้องการได้โดยง่าย จำเป็นต้องมีการอบรมผู้ใช้ หรือต้องการคนกลางในการสร้างประโยคสอบถามตามความต้องการของผู้ใช้อีกทอดหนึ่ง

ข. ไม่สามารถควบคุมปริมาณของผลลัพธ์ที่เกิดขึ้นจากการสอบถามได้ เช่น หากผู้ใช้เลือกใช้คำ หรือวลีที่เป็นคำทั่ว ๆ ไปในการสืบค้น ก็จะได้ผลลัพธ์ออกมามาก หากใช้คำที่เฉพาะก็อาจจะไม่มีผลลัพธ์ออกมาเลย ซึ่งในทั้งสองกรณีทำให้ผู้ใช้ต้องพยายามปรับประโยคสอบถามใหม่ จนกว่าผลลัพธ์ที่ได้จะมีจำนวนพอเหมาะที่จะเลือกดูได้

ค. ผลลัพธ์ของกระบวนการค้นคืนแบบบูล จะทำให้แบ่งฐานข้อมูลออกเป็นสองส่วน คือ ส่วนที่ตรง และส่วนที่ไม่ตรงกับประโยคสอบถาม โดยในระเบียนที่ตรงกับประโยคสอบถามก็มีค่าเท่าเทียมกันหมด คือตรงกับกรณี ไม่มีกลไกที่จะให้ลำดับถึงโอกาสของความตรงกับกรณีจากมากไปหาน้อย

ง. ไม่มีสิ่งใดสะท้อนให้เห็นถึงความแตกต่างในส่วนประกอบของประโยคสอบถาม โดยการค้นแบบบูลจะกำหนดเป็นนัยให้ทุก ๆ คำมีน้ำหนักเป็นเอกหนึ่งหรือศูนย์ขึ้นกับว่า ถ้ามีคำในประโยคสอบถามก็มีน้ำหนักเป็นหนึ่ง หากไม่มีก็มีน้ำหนักเป็นศูนย์

2.5.2 ระบบการค้นคืนชั้นสูง (Advanced Retrieval System)

เป็นระบบที่ใช้แบบจำลองปริภูมิเวกเตอร์ เป็นพื้นฐานในการค้นคืนสารสนเทศ สำหรับแนวคิดเวกเตอร์สเปซโมเดล ที่ใช้ในการค้นคืนสารสนเทศนั้น จะพิจารณาส่วนต่าง ๆ ที่เกี่ยวข้องเป็นส่วนของปริภูมิเวกเตอร์ (elements of a vector space) [5] เช่น คำศัพท์ที่ใช้เป็นตรรชนี เอกสาร และการสอบถาม เหล่านี้แทนได้โดยเวกเตอร์ในปริภูมิเวกเตอร์

ให้ $t_1, t_2, \ldots, t_n$ เป็นค่าศัพท์ที่ใช้เป็นดรรชนีของเอกสารกลุ่มหนึ่ง แต่ละค่าศัพท์ t_i กำหนดให้เป็นเวกเตอร์ t_i ($\vec{t}_i$) ในเวกเตอร์สเปซ ดังนั้นเซ็ตของเวกเตอร์ค่าศัพท์คือ $\{ \vec{t}_i : 1 \leq i \leq n \}$ เป็นซับสเปซของเรื่องที่สนใจ ดังนั้นเวกเตอร์ใด ๆ ในซับสเปซสามารถแสดงได้โดยการรวมกันเชิงเส้นของเวกเตอร์ค่าศัพท์ ($\vec{t}_i$)

ให้ $d_1, d_2, \ldots, d_p$ เป็นเอกสารกลุ่มหนึ่งจำนวน p ฉบับ และให้ $\{ \vec{d}_x = (a_{x1}, a_{x2}, \ldots, a_{xn}) : 1 \leq x \leq p \}$ เป็นเซ็ตของเวกเตอร์ที่แทนเอกสาร เมื่อ a_{xi} เป็นเลขจำนวนจริง ดังนั้นเมื่อให้ a_{xi} เป็นส่วนประกอบของ $\vec{d}_x$ ตามทิศทางของเวกเตอร์ค่าศัพท์ ($\vec{t}_i$) จะสามารถแสดงเวกเตอร์ซึ่งเป็นตัวแทนของเอกสารได้ ดังสมการเวกเตอร์ที่ (1)

$$\vec{d}_x = \sum_{i=1}^{n} a_{xi} \vec{t}_i \qquad \ldots\ldots\ldots (1)$$

เช่นเดียวกัน ในการสอบถามก็สามารถสร้างเวกเตอร์ของการสอบถาม ($\vec{q}$) โดยการรวมกันเชิงเส้นของเวกเตอร์ค่าศัพท์ ($\vec{t}_i$) ซึ่งแสดงได้ดังสมการเวกเตอร์ที่ (2)

$$\vec{q} = \sum_{j=1}^{m} q_j \vec{t}_j \qquad \ldots\ldots\ldots (2)$$

เมื่อ q_j เป็นส่วนประกอบของ q ตามทิศทางของ $\vec{t}_j$

สำหรับการแสดงถึงลำดับความสัมพันธ์ หรือความใกล้เคียงกันในระหว่างประโยคสอบถามกับกลุ่มของเอกสาร เพื่อแสดงแก่ผู้เรียกใช้ ซึ่งสามารถจะแสดงออกมาได้ตามลำดับความใกล้เคียงกัน โดยอาศัยการคำนวณค่าของ Cosine similarity function จากสเกลาร์โปรดักท์ระหว่างเวกเตอร์ของเอกสารและการสอบถาม สมการ vector scalar product ดังสมการที่ (3)

$$\vec{d}_x \cdot \vec{q} = \sum_{j=1}^{m} \sum_{i=1}^{n} a_{xi} q_j \vec{t}_i \cdot \vec{t}_j \qquad \ldots\ldots\ldots (3)$$

เมื่อ $x = 1, 2, \ldots, p$

ในการประยุกต์ตามแนวทฤษฎีของเวกเตอร์ดังกล่าวนั้น จะเห็นได้ว่าเมื่อมีการรวบรวมกลุ่มของเอกสารได้กลุ่มหนึ่ง จะต้องสร้างเวกเตอร์เพื่อเป็นตัวแทนของเอกสาร เมื่อพิจารณาจากสมการที่ (1) จะต้องทราบถึงคำศัพท์ที่มีใช้ทั้งหมดในกลุ่มของเอกสารนี้ คือค่าเวกเตอร์ $\vec{t}_i$ สำหรับค่า a_{xi} จะเป็นน้ำหนักที่ให้แก่แต่ละคำศัพท์ ในแต่ละเอกสารแตกต่างกันไป

คำศัพท์ที่จะได้มานั้น จะใช้ทฤษฎีการทำดรรชนีอัตโนมัติ โดยใช้หลักการทางสถิติเพื่อนับความถี่ของคำศัพท์ที่มีในเอกสารนี้ โดยจะตัดคำศัพท์ที่เป็นคำทั่ว ๆ ไปทิ้ง คำศัพท์ที่มีความถี่ที่เหมาะสมตามแนวทฤษฎีการทำดรรชนีอัตโนมัติ จะถูกเลือกเป็นเวกเตอร์ของปริภูมิเวกเตอร์ของกลุ่มเอกสาร คำศัพท์เหล่านี้จะถูกเก็บไว้ในตารางคำศัพท์

สำหรับน้ำหนักที่จะให้แก่แต่ละคำศัพท์ (w) ในแต่ละเอกสาร ซึ่งเป็นปริมาณสเกลาร์ของแต่ละเวกเตอร์คำศัพท์ $\vec{t}_i$ จะได้จากการคำนวณดังแสดงในสมการที่ (4) [1]

$$w = t \log(N/d) \qquad \text{..... (4)}$$

เมื่อ t = ความถี่ของคำศัพท์ใดๆในเอกสารหนึ่ง (term frequency)

d = ความถี่ของเอกสารที่ใช้คำศัพท์นี้ (document frequency)

N = จำนวนเอกสารทั้งหมดในกลุ่มนี้

สำหรับเวกเตอร์ของการสอบถามนั้น เนื่องจากการสอบถามจะใช้เป็นภาษาธรรมชาติ เช่น "ภาวะโภชนาการของหญิงมีครรภ์ในจังหวัดภาคใต้" เป็นต้น ดังนั้นจึงสามารถใช้รูทีนการทำดรรชนีได้เช่นเดียวกับของเอกสารทั่วไป

ในการเรียกคืนสารสนเทศออกมา จะอาศัยการคำนวณความใกล้เคียงกัน ในระหว่างเวกเตอร์ของการสอบถาม กับแต่ละเวกเตอร์ของเอกสารที่รวบรวมอยู่ โดยอาศัยจากการคำนวณ cosine similarity ซึ่งสามารถกระจายออกมาได้จากสมการที่ (3) และเมื่อกำหนดให้แต่ละเวกเตอร์ $\vec{t}_i$ มีคุณสมบัติเป็น orthogonal กันและกัน ดังนั้น

$$\vec{t}_i \cdot \vec{t}_j = 0 \quad \text{เมื่อ} \quad i \neq j \quad \text{และ}$$
$$\vec{t}_i \cdot \vec{t}_j = 1 \quad \text{เมื่อ} \quad i = j$$

ดังนั้นจากสมการที่ (3) จะสามารถแสดงได้ดังสมการที่ (5)

$$\text{COSINE}(\vec{d}_x, \vec{q}) = \sum_{i=1}^{n} (a_{xi} q_i) \Big/ \sqrt{\sum_{i=1}^{n} (a_{xi})^2 \sum_{i=1}^{m} (q_i)^2} \quad \ldots\ldots \quad (5)$$

จากผลการคำนวณ สามารถจะเลือกเอกสารที่ให้ค่าการคำนวณอันอยู่ในเกณฑ์ที่ยอมรับเรียงลำดับความใกล้เคียงกันกับการสอบถามออกมาแสดงได้

ตัวอย่างเวกเตอร์ของเอกสาร เวกเตอร์ของการสอบถาม และ การคำนวณความใกล้เคียงกันของเวกเตอร์เอกสารกับเวกเตอร์การสอบถาม สมมุติให้เอกสารกลุ่มหนึ่งมี 3 ฉบับ คำศัพท์ที่ใช้เป็นดรรชนีมีรวม 4 คำ กำหนดให้เป็น $\vec{t}_1$, $\vec{t}_2$, $\vec{t}_3$ และ $\vec{t}_4$ สำหรับน้ำหนักที่จะให้แก่แต่ละคำศัพท์คำนวณได้โดยสมการที่ (4) สมมุติให้สามารถแสดงเวกเตอร์ของแต่ละเอกสารออกมาได้ดังนี้

$$\vec{d}_1 = 2\vec{t}_1 + 3\vec{t}_2 + 5\vec{t}_3$$
$$\vec{d}_2 = 3\vec{t}_1 + 7\vec{t}_2 + 2\vec{t}_3$$
$$\vec{d}_3 = \vec{t}_1 + 2\vec{t}_2 + \vec{t}_4$$

กำหนดให้เวกเตอร์ของการสอบถาม ดังนี้

$$\vec{q} = \vec{t}_3$$

ให้ S_1, S_2 และ S_3 เป็นค่าความใกล้เคียงกันระหว่างเวกเตอร์ของเอกสารที่ 1, 2 และ 3 กับเวกเตอร์ของการสอบถามตามลำดับ สามารถคำนวณค่า S_1, S_2 และ S_3 โดยใช้สมการ (5) จะได้ผลดังนี้

S_1 = 0.81

S_2 = 0.25

S_3 = 0

ดังนั้นการสอบถามนี้จะสามารถดึงเอกสาร d_1 และ d_2 ออกมาตามลำดับ จากผลการคำนวณ $S_1 > S_2$ และ $S_3 = 0$