บทที่ 3
หลักการและแนวความคิดเกี่ยวกับระบบสืบค้นสารสนเทศ

3.1 หลักการของระบบสืบค้นสารสนเทศ

ในปัจจุบัน มีระบบจัดการสารสนเทศบนคอมพิวเตอร์ ที่สำคัญ ๆ และเป็นที่รู้จักโดยแพร่หลายอยู่หลายประเภท แต่ละประเภทจะมีความแตกต่างกัน ในรายละเอียดเกี่ยวกับ ลักษณะของสารสนเทศที่เก็บ ลักษณะความต้องการค้นหาของผู้ใช้ ลักษณะของการประมวลผล ตลอดจนรูปแบบของการตอบโต้กันระหว่างผู้ใช้กับระบบ ดังนั้น จึงได้มีการจัดแบ่งระบบจัดการสารสนเทศออกเป็นประเภทใหญ่ ๆ ดังนี้ ระบบบริหารฐานข้อมูล หรือ DBMS (Data Base Management Systems) เป็นระบบที่มีการเก็บรวบรวมระเบียนข้อมูลไว้อย่างมีระเบียบแบบแผนเดียวกัน โดยแต่ละระเบียนจะประกอบด้วยเขตข้อมูลต่าง ๆ เหมือนกัน และมีการกำหนดลักษณะของข้อมูลที่จะเก็บในแต่ละเขตข้อมูลไว้ล่วงหน้า การค้นหาข้อมูลจะกำหนดเขตข้อมูลและค่าที่ต้องการแบบตรงตัว ระบบถามตอบ หรือ QA (Question-Answering System) เป็นระบบที่มีการเก็บข้อมูลหลักความจริงที่เน้นหนักไปทางสาขาวิชาใดวิชาหนึ่ง และผู้ใช้สามารถสอบถามหลักความจริงในเรื่องนั้น โดยไม่จำกัดรูปแบบของการสอบถาม ซึ่งในการจัดทำระบบประเภทนี้ จะต้องอาศัยหลักการทางด้านปัญญาประดิษฐ์ หรือ AI (Artificial Intelligent) มาประยุกต์รวมกับหลักการของระบบบริหารฐานข้อมูล ส่วนระบบสืบค้นสารสนเทศ หรือ IR (Information Retrieval Systems) ใช้ในการค้นหาข้อมูลประเภทเอกสารหลักฐาน หรือเอกสารอ้างอิงต่าง ๆ จึงมักมีลักษณะของข้อมูลที่เก็บเป็นข้อความภาษาธรรมชาติ และในการค้นหาผู้ใช้ต้องกำหนดความต้องการค้นโดยย่อเป็นภาษาดรรชนี ซึ่งจะบรรยายรายละเอียดในระบบชนิดนี้ต่อไป

Salton และ McGill (Salton and McGill 1983: 7-12,16-20) ได้บรรยายถึงหลักการของระบบสืบค้าสารสนเทศ ซึ่งสรุปได้ว่า สารสนเทศที่เป็นเป้าหมายในการค้น ของระบบสืบค้นสารสนเทศนั้น จะมีลักษณะเป็นภาษาธรรมชาติ ข้อมูลถูกเก็บคละเคล้ากันไป ไม่เรียงลำดับ และไม่มีระเบียบแบบแผนของข้อมูลอย่างที่เป็นอยู่ในระบบ DBMS ดังนั้น การค้นหาเอกสารอ้างอิงที่เกี่ยวกับเรื่องราวหนึ่ง เช่น ต้องการเรื่องราวของ INFORMATION อาจใช้วิธีไล่ค้นหาคำ INFORMATION ในทุกข้อความของเอกสารทั้งหมด และเอกสารใดที่มีคำนี้ปรากฏอยู่ในบทความ แสดงว่าเอกสารนั้นมีเรื่องราวเกี่ยวกับ INFORMATION ตามต้องการ แต่การค้นหาโดย วิธีเปรียบเทียบกลุ่มตัวอักษร (string matching method) ดังกล่าวนั้น เป็นที่ทราบกันดีว่า แม้จะนำเทคนิคพิเศษใด ๆ เข้าช่วยก็คงไม่ลดการเสียเวลาไปได้เท่าไรนัก จึงได้มีการนำหลักการของ **ดรรชนี** (Indexing) มาประยุกต์ใช้ โดย

การเก็บรวบรวมดรรชนี ที่จะชี้บอกได้ว่า เรื่องราวใด มีอยู่ในเอกสารใดบ้าง และดรรชนีที่ใช้กันมากในระบบสืบค้นสารสนเทศ ก็คือ คำสำคัญ ที่โดยความหมายในทางภาษา ของคำสำคัญนั้น ๆ จะสามารถอ้างอิงถึงเรื่องราวหนึ่ง ซึ่งผู้ใช้เข้าใจได้ในทันที และเมื่อรวบรวม คำสำคัญ ต่าง ๆ ไว้ในแฟ้มดรรชนี โดยมีโครงสร้างเป็น แฟ้มข้อมูลย้อนกลับ แล้ว ก็จะสามารถจัดทำขบวนการสืบค้นสารสนเทศในเรื่องราวใด ๆ ด้วยการนำ คำสำคัญ คำหนึ่งไปค้นหาในแฟ้มดรรชนี เมื่อพบคำที่ตรงกัน ก็จะทราบได้ในทันทีว่า เอกสารใดบ้างที่มีเรื่องราวที่ต้องการ จากนั้นก็สามารถไปดึงสารสนเทศในเอกสารเหล่านั้นมาได้

3.2 ฐานข้อมูลและขอบข่ายการค้นของระบบสืบค้นสารสนเทศทั่วไป

ในหัวข้อนี้ จะได้กล่าวถึง ฐานข้อมูลของระบบสืบค้นสารสนเทศทั่วไป โดยสรุปเรื่องราวใน 2 ประเด็น คือ ข้อมูลหลัก ๆ ที่เก็บในฐานข้อมูล และลักษณะโครงสร้างฐานข้อมูลของระบบสืบค้นสารสนเทศเหล่านั้น พร้อมกันนี้ จะกล่าวถึงขอบข่ายการค้นโดยสรุป ควบคู่กันไปในแต่ละระบบด้วย

3.2.1 ระบบ DIALOG

ระบบ DIALOG (Salton and McGill 1983: 30-34 ; Dialog Information Services Inc. 1979: 3-1 - 3-19) นิยมใช้กันมากในงานห้องสมุด โครงสร้างฐานข้อมูลของระบบ เป็นลักษณะของ แฟ้มข้อมูลย้อนกลับ ข้อมูลที่เก็บประกอบด้วย 2 ส่วน คือ

- **ข้อมูลสำหรับประมวลผลในการค้น**
  แบ่งได้เป็น 2 พวก คือ พวกแรก ได้แก่ ข้อมูลคำสำคัญ ข้อมูลหมายเลขประจำเอกสาร ข้อมูลตำแหน่งของคำสำคัญ ข้อมูลพวกนี้ ถูกรวบรวมมาจากสารสนเทศในส่วนบทความ ส่วนอีกพวกหนึ่ง เป็นข้อมูลที่รวบรวมมาจากสารสนเทศในส่วนกำหนดเขตข้อมูล ได้แก่ ชื่อผู้แต่ง ชื่อวารสาร ภาษา ปีที่พิมพ์ เป็นต้น
- **ข้อมูลสารสนเทศ**
  ประกอบด้วย บทความ และกลุ่มข้อมูลที่กำหนดเขตข้อมูล เช่น ชื่อผู้แต่ง ชื่อวารสาร ภาษา ปีที่พิมพ์ ประเภทเอกสาร วันเดือนปีที่บันทึกข้อมูล เป็นต้น

ระบบ DIALOG สามารถค้นสารสนเทศด้วย คำเดี่ยว ส่วนของคำ และ คำคู่ ซึ่งผลลัพธ์ของการค้นจะถูกเก็บเป็น เซทของหมายเลขเอกสาร ไว้ทุกครั้งที่ค้น พร้อมทั้งจัดหมายเลขประจำให้แก่เซทนั้น ๆ ผู้ใช้สามารถดำเนินปฏิบัติการแบบบูลระหว่างเซทเหล่านี้ ตั้งแต่ 2 เซทขึ้นไป ปฏิบัติการแบบบูล ที่มีให้ใช้ ได้แก่ **AND, OR, NOT** (ในการใช้จะแทนด้วย * + - ตาม

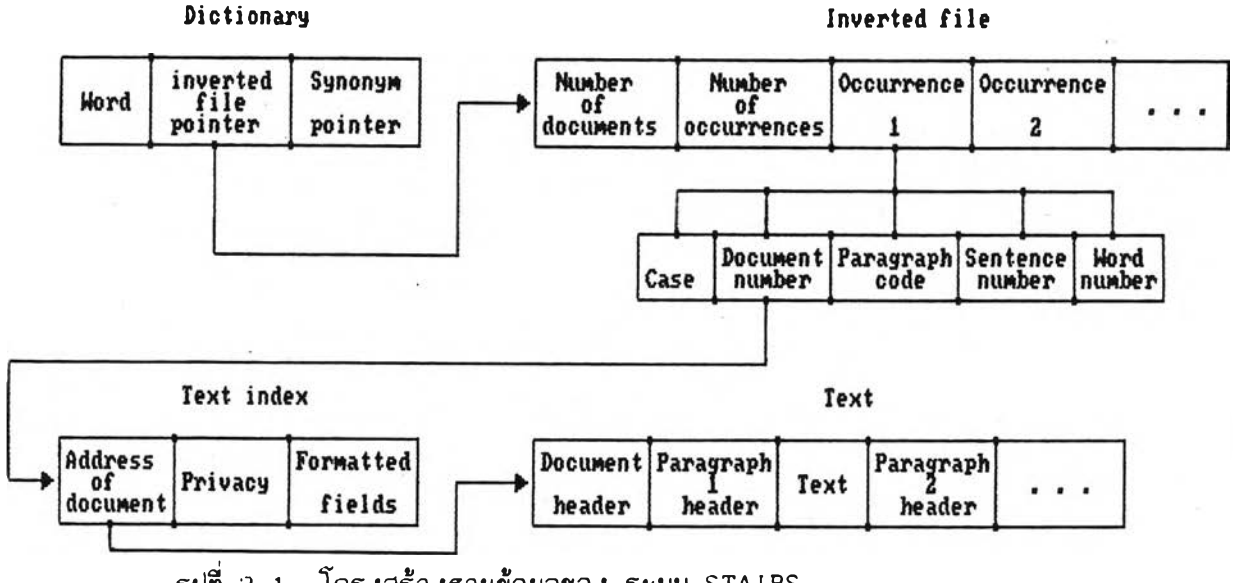
ลำดับ) และกำหนดให้ดำเนินปฏิบัติการ NOT เป็นลำดับแรกสุด ตามด้วย AND และ OR ท้ายสุด หรืออาจใช้วงเล็บกำหนดลำดับใหม่ โดยส่วนที่อยู่ในวงเล็บจะถูกดำเนินการก่อนเสมอ

สำหรับ ส่วนของคำ ต้องใช้เครื่องหมาย "?" เติมท้าย เพื่อกำหนดจำนวนอักษรท้ายคำที่ละไว้ และสามารถเติม "?" ไว้ภายใน เทอมค้น ได้ด้วย ส่วน คำคู่ สามารถกำหนดจำนวนคำ ที่แทรกระหว่างคำคู่ หรือกำหนดเป็นคำประชิดกันก็ได้ นอกจากนั้น ผู้ใช้อาจสั่งให้ค้นหาเฉพาะในเขตข้อมูลหนึ่ง เช่น ผู้แต่ง ชื่อวารสาร ปีที่พิมพ์ เป็นต้น โดยระบุค่าที่ต้องการได้

3.2.2 ระบบ STAIRS

ระบบ STAIRS (Salton and McGill 1983: 34-41 ; IBM Corporation 1977: 9,10,22-24,31) มีโครงสร้างฐานข้อมูลเป็นแบบ แฟ้มข้อมูลย้อนกลับ ซึ่งประกอบด้วย

- **แฟ้มข้อมูลพจนานุกรม**
  ใช้เก็บคำสำคัญ และข้อมูลเกี่ยวกับคำที่มีความหมายเหมือนกัน พร้อมทั้งเก็บตัวชี้ ซึ่งชี้ไปยังที่เก็บ ข้อมูลตำแหน่งคำ ในแฟ้มข้อมูลย้อนกลับ
- **แฟ้มข้อมูลย้อนกลับ**
  ใช้เก็บข้อมูลการปรากฏของคำสำคัญแต่ละคำในสารสนเทศ ได้แก่ จำนวนที่พบคำสำคัญ และข้อมูลตำแหน่งของคำสำคัญในแต่ละตำแหน่ง ซึ่งประกอบด้วย หมายเลขสารสนเทศ รหัสย่อหน้า หมายเลขประโยค และหมายเลขคำ
- **แฟ้มข้อมูลสารสนเทศ**
  ใช้เก็บสารสนเทศทั้งหลาย ซึ่งแต่ละสารสนเทศจะประกอบด้วยข้อมูล 2 ส่วน คือ ข้อมูลลักษณะบทความ และข้อมูลกำหนดเขตข้อมูล



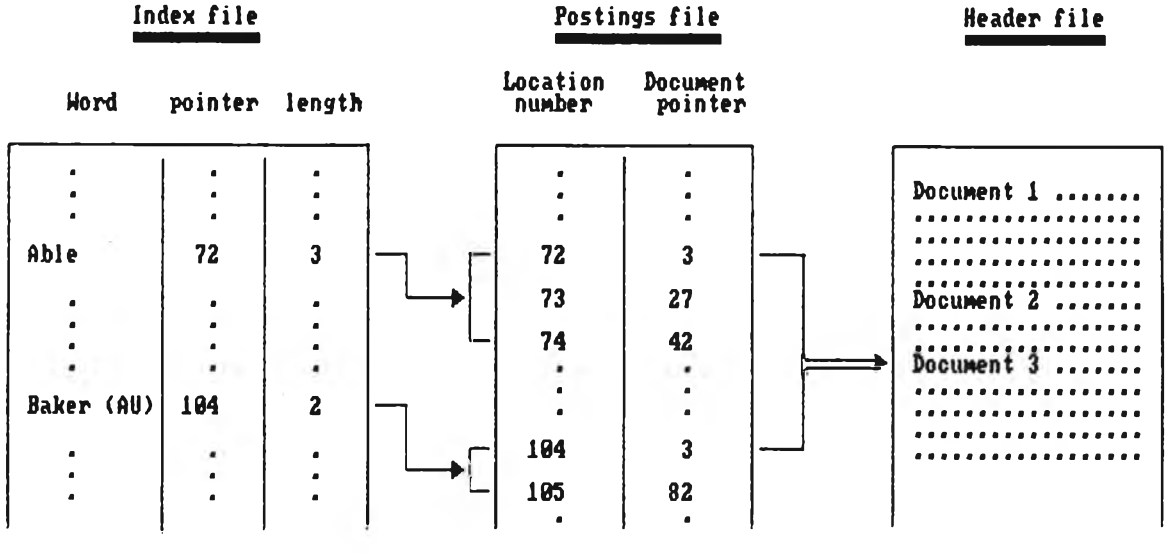
รูปที่ 3.1 โครงสร้างฐานข้อมูลของ ระบบ STAIRS

ในระบบ STAIRS สามารถค้นด้วย คำเดี่ยว ส่วนของคำ และสามารถใช้ปฏิบัติการแบบบูล ปฏิบัติการประชิด ได้แก่ AND, OR, XOR, NOT, ADJ การป้อน ส่วนของคำ ต้องใช้ "$" เติมท้าย เพื่อระบุจำนวนอักษรท้ายคำที่ละไว้ ผู้ใช้ยังสามารถกำหนด คำเหมือน ได้เอง สามารถกำหนดให้ค้นหาเฉพาะในเขตข้อมูลโดยระบุค่าที่ต้องการด้วย ส่วนผลลัพธ์จากการค้นจะถูกเก็บบันทึกและมีหมายเลขกำกับ ซึ่งสามารถนำไปใช้ในการค้นครั้งต่อไป การแสดงสารสนเทศผลลัพธ์สามารถกำหนดให้แสดงตามลำดับค่าแรงค์ ซึ่งคำนวณได้จากน้ำหนักของคำสำคัญ ในสารสนเทศนั้น

3.2.3 ระบบ MEDLARS

ระบบ MEDLARS (Salton and McGill 1983: 42-45) ใช้ได้ดีและมีชื่อเสียงมากที่สุด โครงสร้างฐานข้อมูลของระบบนี้ มีพื้นฐานเป็น แฟ้มข้อมูลย้อนกลับ ภายในฐานข้อมูลประกอบด้วยแฟ้มข้อมูล 3 ชนิด ได้แก่

- **แฟ้มดรรชนี**
  เก็บเทอมที่ยูนิคต่าง ๆ ได้แก่ ชื่อผู้แต่ง คำศัพท์ควบคุม รหัสแยกประเภท ข้อมูลวันเดือนปี เป็นต้น พร้อมทั้งเก็บจำนวนข้อมูลโพสติง ตำแหน่งข้อมูลในแฟ้มโพสติง ควบคู่ไปด้วย
- **แฟ้มโพสติง**
  ใช้เก็บข้อมูลโพสติง ได้แก่ หมายเลข CAN ซึ่งจะระบุเอกสารในแฟ้มเอกสาร
- **แฟ้มเอกสาร**
  ใช้เก็บเอกสารทั้งหมด แต่ละเอกสารจะมีหมายเลขประจำ (หรือ CAN : computer assign number)



รูปที่ 3.2 โครงสร้างฐานข้อมูลของ ระบบ MEDLARS

ระบบนี้ มีคำสั่งในการค้นคล้ายคลึงกับที่มีใช้ใน DIALOG ส่วนที่มีเพิ่มขึ้นได้แก่ การค้นแบบ เทียบอักษร ซึ่งกำหนดให้ค้นได้เฉพาะในกลุ่มข้อมูลที่เป็นผลลัพธ์จากการค้นครั้งก่อนเท่านั้น

จะเห็นได้ว่า ข้อมูลหลัก ๆ ที่เก็บในฐานข้อมูลของระบบสืบค้นสารสนเทศทั่วไป แบ่งได้เป็นกลุ่มข้อมูลเกี่ยวกับการค้นหา ซึ่งได้แก่ ข้อมูลคำสำคัญ ข้อมูลความถี่ในการปรากฏ ข้อมูลตำแหน่งของคำสำคัญ ส่วนข้อมูลอีกกลุ่มหนึ่ง ได้แก่ กลุ่มข้อมูลสารสนเทศ ซึ่งจะมีการกำหนดหมายเลขประจำของแต่ละสารสนเทศไว้ ข้อมูลทั้งสองส่วนจะมีการเชื่อมโยงถึงกันได้ โดยที่ คำสำคัญแต่ละคำ จะสามารถบ่งชี้กลุ่มสารสนเทศที่มีความสัมพันธ์เกี่ยวข้องกัน หรือมีคำสำคัญนั้น ๆ ปรากฏอยู่ในข้อความของสารสนเทศ โดยจะระบุสารสนเทศเหล่านั้น ด้วย หมายเลขประจำสารสนเทศ ที่ได้กำหนดไว้ล่วงหน้าแล้วนั่นเอง ส่วนโครงสร้างฐานข้อมูลของระบบนั้น ข้อมูลจะเชื่อมโยงกันในลักษณะของ **แฟ้มข้อมูลย้อนกลับ** ทั้งสิ้น

3.3 หลักการของขบวนการสืบค้นสารสนเทศกับความต้องการค้นหาของผู้ใช้

ความต้องการค้นหาสารสนเทศของผู้ใช้ อาจแบ่งได้เป็น 2 ระดับ คือ ความต้องการค้นในระดับพื้นฐาน กับ ความต้องการค้นในระดับที่ซับซ้อน ซึ่งจะได้อธิบายรายละเอียดต่อไป

3.3.1 ความต้องการค้นในระดับพื้นฐาน

ความต้องการค้นในระดับนี้ เป็นความต้องการค้นสารสนเทศในเรื่องใดเรื่องหนึ่งโดยเจาะลงไปในแง่มุมเดียว ซึ่งผู้ใช้จะสามารถกำหนด คำสั่งค้น ในรูปของภาษาดรรชนีได้ด้วย **คำเดี่ยว** คือ ระบุคำสำคัญเพียง 1 คำ เพื่อใช้ในการค้นหาสารสนเทศใด ๆ ที่ปรากฏคำสำคัญ ที่ระบุนั้น ข้อมูลที่จำเป็นในการประมวลผล คือ แฟ้มดรรชนี ที่รวบรวมคำสำคัญต่าง ๆ โดยที่แต่ละคำลามารถบ่งชี้สารสนเทศที่สัมพันธ์กันได้ ข้อมูลเพียงเท่านี้ ก็สามารถสนองตอบความต้องการค้นในระดับพื้นฐานได้อย่างมีประสิทธิภาพ

3.3.2 ความต้องการค้นในระดับที่ซับซ้อน

ความต้องการค้นในระดับนี้ เป็นความต้องการค้นสารสนเทศในเรื่องใดเรื่องหนึ่งซึ่งมีลักษณะที่สัมพันธ์กับสารสนเทศมากกว่า 1 แง่มุม ดังนั้น การกำหนด คำสั่งค้น ในรูปของภาษาดรรชนี ก็ไม่อาจใช้ คำเดี่ยว ได้ จำเป็นจะต้องมีการกำหนดคำสำคัญมากกว่า 1 คำ และเนื่องจากลักษณะความต้องการที่ซับซ้อน ค่อนข้างหลากหลายในรูปแบบ ตัวอย่างเช่น ความต้องการค้นหาสารสนเทศในเรื่องราวที่เกี่ยวกับบุคคลสำคัญ 2 คน ก็อาจแจกแจงได้หลายรูปแบบ ดังนี้

- ความต้องการสารสนเทศในเรื่องราวที่เกี่ยวกับความสัมพันธ์ของบุคคล 2 คน
- ความต้องการสารสนเทศในเรื่องราวที่เกี่ยวกับบุคคลใดก็ได้ใน 2 คนที่สนใจ
- ความต้องการสารสนเทศในเรื่องราวที่เกี่ยวกับบุคคลหนึ่ง ในตอนที่ไม่มีอีกบุคคลหนึ่งมาเกี่ยวข้อง
- ความต้องการสารสนเทศในเรื่องราวที่ค่อนข้างจำเพาะเจาะจง ซึ่งการกำหนดความต้องการค้นด้วย คำสำคัญเพียงคำเดียวจะไม่เพียงพอ หรือไม่ได้ใจความตรงกับที่ต้องการ

จากความหลากหลายในรูปแบบความต้องการค้นในระดับซับซ้อน ดังกล่าว ทำให้การกำหนด ข้อความค้นหา ในรูปของภาษาดรรชนี นอกจากจะต้องระบุคำสำคัญมากกว่า 1 คำ แล้ว ยังจำเป็นต้องมีการระบุ รูปแบบความสัมพันธ์ของคำสำคัญกับสารสนเทศที่ต้องการด้วย ซึ่งหนทางแก้ปัญหานี้ก็คือ การนำเอา **ปฏิบัติการแบบบูล** (Boolean Operations) และ **ปฏิบัติการประชิด** หรือ ADJ (Adjacent Operation) มาใช้ ในการนี้ จำเป็นต้องมีฐานข้อมูล ที่มีโครงสร้างเป็นแบบ แฟ้มข้อมูลย้อนกลับ และข้อมูลที่เก็บภายในฐานข้อมูล จะต้องประกอบด้วย แฟ้มสารสนเทศ ที่เก็บสารสนเทศทั้งหมด โดยที่ ทุกสารสนเทศมีหมายเลขประจำ และมี แฟ้มดรรชนีย้อนกลับ สำหรับเก็บรวบรวมดรรชนี ได้แก่ คำสำคัญต่าง ๆ และแต่ละคำ สามารถระบุหมายเลขสารสนเทศต่าง ๆ ที่ปรากฏคำสำคัญนั้น ๆ อยู่ในสารสนเทศ รวมทั้งมีข้อมูลตำแหน่งคำที่บอกถึงรายละเอียดของตำแหน่งแต่ละตำแหน่ง ที่ปรากฏคำสำคัญในสารสนเทศ ว่า อยู่ในข้อความส่วนใด และเป็นคำสำดับที่เท่าไร ข้อมูลต่าง ๆ เหล่านี้ เป็นส่วนที่สำคัญสำหรับใช้ประมวลผลในการสืบค้นสารสนเทศ เพื่อให้ระบบสามารถตอบสนองความต้องการค้นของผู้ใช้ ในระดับที่ซับซ้อนได้อย่างมีประสิทธิภาพ

ลำดับต่อไป จะได้บรรยายถึง ปฏิบัติการแบบบูล และปฏิบัติการประชิด ว่ามีลักษณะ และแนวทางของการดำเนินปฏิบัติการ อย่างไร

ปฏิบัติการแบบบูล

ปฏิบัติการแบบบูล (Salton and McGill 1983: 25-26) เหมาะกับการสืบค้นข้อมูลที่เกี่ยวพันกับเทอมหลายเทอม ในลักษณะต่าง ๆ กันดังกล่าว ดังนั้น ปฏิบัติการแบบบูล จึงแบ่งออกเป็น 3 ชนิด ตามลักษณะการทำงานของแต่ละปฏิบัติการ ดังนี้

- **ปฏิบัติการ AND**

  ใช้ในการสืบค้นสารสนเทศใด ๆ ที่มีคำสำคัญที่ระบุตั้งแต่ 2 คำ ขึ้นไป ปรากฏอยู่ภายในสารสนเทศเดียวกัน การดำเนินปฏิบัติการ AND จะใช้หลักปฏิบัติการทางเซท ที่เรียกว่า **อินเทอเซกชั่น** (Intersection) โดยในทางปฏิบัติ สำหรับการสืบค้นสารสนเทศ

ที่กำหนดความต้องการค้น เป็นภาษาดรรชนีว่า " คำสำคัญ-1 **AND** คำสำคัญ-2 " จะมีแนวทางหลัก ๆ ดังนี้

1. ใช้ คำสำคัญ-1 ค้นหาจากแฟ้มดรรชนีย้อนกลับ ให้ได้หมายเลขสารสนเทศที่ปรากฏคำสำคัญนี้ เรียกกลุ่มหมายเลขสารสนเทศเหล่านี้ว่า Set-1
2. ใช้ คำสำคัญ-2 ค้นหาจากแฟ้มดรรชนีย้อนกลับ ให้ได้หมายเลขสารสนเทศที่ปรากฏคำสำคัญนี้ เรียกกลุ่มหมายเลขสารสนเทศเหล่านี้ว่า Set-2
3. พิจารณาหมายเลขสารสนเทศ ที่เป็นสมาชิกของ Set-1 และ Set-2 ว่า มีหมายเลขใดบ้าง ที่เป็นผลลัพธ์ของ อินเทอเซกชั่น ระหว่างเซททั้งสอง ซึ่งได้แก่ หมายเลขที่เป็นสมาชิกของทั้งสองเซท นั่นเอง เรียกกลุ่มหมายเลขผลลัพธ์ว่า Set-3
4. สารสนเทศต่าง ๆ ที่ถูกระบุหมายเลขใน Set-3 เป็นผลลัพธ์ของการสืบค้นนี้

- **ปฏิบัติการ OR**

ใช้ในการสืบค้นสารสนเทศใด ๆ ที่มีคำสำคัญคำใด คำหนึ่ง ในบรรดาที่ระบุให้ตั้งแต่ 2 คำ ขึ้นไป ปรากฏอยู่ภายในสารสนเทศนั้น การดำเนินปฏิบัติการ OR จะใช้หลักปฏิบัติการทางเซท ที่เรียกว่า **ยูเนียน** (Union) โดยในทางปฏิบัติ สำหรับการสืบค้นสารสนเทศ ที่กำหนดความต้องการค้นเป็นภาษาดรรชนี ว่า " คำสำคัญ-1 **OR** คำสำคัญ-2 " จะมีแนวทางหลัก ๆ ดังนี้

1. ดำเนินการเช่นเดียวกับ ข้อ 1 ของปฏิบัติการ AND
2. ดำเนินการเช่นเดียวกับ ข้อ 2 ของปฏิบัติการ AND
3. พิจารณาหมายเลขสารสนเทศ ที่เป็นสมาชิกของ Set-1 และ Set-2 ว่า มีหมายเลขใดบ้าง ที่เป็นผลลัพธ์ของ การยูเนียน ระหว่างเซททั้งสอง ซึ่งได้แก่ หมายเลขทั้งหมดที่เป็นสมาชิกอยู่ใน Set-1 หรือ Set-2 และถ้ามีหมายเลขที่ซ้ำกัน ให้เก็บไว้เพียงหมายเลขเดียว เรียกกลุ่มหมายเลขผลลัพธ์ว่า Set-3
4. สารสนเทศต่าง ๆ ที่ถูกระบุหมายเลขใน Set-3 เป็นผลลัพธ์ของการสืบค้นนี้

- **ปฏิบัติการ NOT**

ใช้ในการสืบค้นสารสนเทศใด ๆ ที่มีคำสำคัญที่ระบุคำแรก ปรากฏอยู่ภายในสารสนเทศนั้น โดยที่จะต้องไม่มีคำสำคัญอีกคำที่ระบุให้ ปรากฏอยู่ในสารสนเทศเดียวกัน สำหรับการดำเนินปฏิบัติการ NOT จะใช้หลักปฏิบัติการทางเซท ที่เรียกว่า **ดิฟเฟอเรนซ์** (Difference) โดยในทางปฏิบัติ สำหรับการสืบค้นสารสนเทศ ที่กำหนดความต้องการค้นเป็นภาษาดรรชนีว่า " คำสำคัญ-1 **NOT** คำสำคัญ-2 " มีแนวทางหลัก ๆ ดังนี้

1. ดำเนินการเช่นเดียวกับ ข้อ 1 ของปฏิบัติการ AND
2. ดำเนินการเช่นเดียวกับ ข้อ 2 ของปฏิบัติการ AND
3. พิจารณาหมายเลขสารสนเทศ ที่เป็นสมาชิกของ Set-1 และ Set-2 ว่า มีหมายเลขใดบ้าง ที่เป็นผลลัพธ์ของ การหาดิฟเฟอเรนซ์ ระหว่างเซททั้งสอง ซึ่งได้แก่ หมายเลขที่เป็นสมาชิกของ Set-1 แต่ ไม่เป็นสมาชิกของ Set-2 เรียกกลุ่มหมายเลขผลลัพธ์ว่า Set-3
4. สารสนเทศต่าง ๆ ที่ถูกระบุหมายเลขใน Set-3 เป็นผลลัพธ์ของการสืบค้นนี้

ปฏิบัติการประชิด

ปฏิบัติการประชิด (Salton and McGill 1983: 28-30) ใช้ในการค้นหาสารสนเทศด้วยคำสำคัญมากกว่า 1 คำ ที่ปรากฏอยู่ต่อเนื่องกันเป็น **วลี** และเพื่อหลีกเลี่ยงการไล่ค้นวลีสำคัญ ในสารสนเทศโดยตรง ก็จำเป็นต้องจัดทำ ขบวนการตรวจสอบคำประชิด โดยการประมวลผลข้อมูลตำแหน่ง ของคำสำคัญแต่ละคำที่ประกอบเป็น วลีสำคัญ ว่ามีตำแหน่งใดบ้างที่คำสำคัญเหล่านั้น จะเรียงประชิดติดกันไปเป็น วลีสำคัญ ซึ่งเป็นที่มาของ **ปฏิบัติการประชิด** นั่นเอง

การดำเนินปฏิบัติการ ADJ สำหรับการสืบค้นสารสนเทศที่กำหนดเป็นภาษาดรรชนีว่า

" คำสำคัญ-1 **ADJ** คำสำคัญ-2 " มีแนวทางหลัก ๆ ดังนี้

1. ดำเนินการเช่นเดียวกับ ข้อ 1 ของปฏิบัติการ AND
2. ดำเนินการเช่นเดียวกับ ข้อ 2 ของปฏิบัติการ AND
3. จากสารสนเทศหมายเลขเดียวกันของ Set-1 และ Set-2 จะพิจารณาข้อมูลตำแหน่งคำของ คำสำคัญ-1 กับ คำสำคัญ-2 เพื่อหาว่า มีกรณีที่ตำแหน่งของคำทั้งสองเรียงประชิดตามกัน หรือไม่ เรียกกลุ่มหมายเลขสารสนเทศที่พบคำสำคัญทั้งสองอยู่ประชิดกัน ว่า Set-3
4. สารสนเทศต่าง ๆ ที่ถูกระบุหมายเลขใน Set-3 เป็นผลลัพธ์ของการสืบค้นนี้

กฎเกณฑ์ที่จำเป็นสำหรับการค้นในระดับที่ซับซ้อน

ในกรณีที่ความต้องการค้นซับซ้อนมาก ๆ จำเป็นต้องระบุคำสำคัญมากกว่า 2 คำ และต้องใช้ ปฏิบัติการแบบบูล และ/หรือ ปฏิบัติการประชิด พร้อมกันหลายปฏิบัติการนั้น จำเป็นต้องกำหนดกฎเกณฑ์เกี่ยวกับการประมวลผลไว้ล่วงหน้า เพื่อไม่ให้เกิดสับสนในการประมวลผล และผู้ใช้จะต้องรับรู้ถึงกฎเกณฑ์เหล่านี้ด้วย เพื่อจะได้กำหนดความต้องการเป็นภาษาดรรชนีได้ถูกต้อง และได้รับสารสนเทศผลลัพธ์ที่ตรงกับความต้องการ กฎเกณฑ์ที่ต้อง

กำหนดไว้ มีดังนี้

1. การกำหนดลำดับก่อนหลังของการดำเนินปฏิบัติการ
   ถ้ามีปฏิบัติการ AND, OR, NOT, ADJ ถูกระบุมาในความต้องการค้นด้วยกัน ในการประมวลผลจะมีการจัดว่า จะทำปฏิบัติการใดเป็นลำดับแรกสุด ถัดไปเป็นปฏิบัติการใด และทำปฏิบัติการใดหลังสุด
2. การกำหนดทิศทางของการดำเนินปฏิบัติการ
   ในกรณีที่ คำสั่งค้น มีการใช้ปฏิบัติการหลายปฏิบัติการที่มีศักย์เท่ากัน อาจจะให้ดำเนินปฏิบัติการทางซ้ายก่อน แล้วจึงค่อยดำเนินการไล่ไปทางขวา หรือในทิศทางตรงข้าม คือ ดำเนินปฏิบัติการทางขวาก่อน แล้วค่อยไล่ทำมาทางซ้าย ซึ่งต้องกำหนดลงไปว่าจะยึดตามทิศทางใดเป็นหลัก
3. การใช้วงเล็บกำหนดลำดับก่อนหลังในการดำเนินปฏิบัติการโดยตรง
   ถ้ายอมให้มีการใช้วงเล็บ ก็ต้องกำหนดกฎเกณฑ์เกี่ยวกับการใช้วงเล็บไว้ ซึ่งโดยทั่วไปมักจะกำหนดว่า ให้ดำเนินปฏิบัติการที่อยู่ในวงเล็บเป็นลำดับแรกสุด แล้วจึงค่อยดำเนินปฏิบัติการอื่น ๆ ที่อยู่ภายนอกวงเล็บ ตามลำดับของกฎเกณฑ์ในข้อ 1-2 และถ้ามีวงเล็บซ้อนกันหลายชั้น จะดำเนินการภายในวงเล็บในสุด ไล่ออกมาทีละชั้นจนถึงวงนอกสุด
4. การใช้ **ส่วนของคำ** (partial word) ในการสืบค้น
   การใช้ ส่วนของคำ เพื่อให้ครอบคลุมกลุ่มคำสำคัญ ที่มีบางส่วนของคำเหมือนกัน เช่น กลุ่มคำสำคัญที่ขึ้นต้นเหมือนกัน ต่างกันเฉพาะท้ายคำ เป็นต้น ซึ่งในกรณีการระบุด้วยส่วนของคำ เพื่อแทนกลุ่มคำสำคัญ อาจมีให้เลือกใช้หลายรูปแบบ และมักมีการใช้เครื่องหมายพิเศษ เช่น "?" "$" "*" สำหรับเติมไว้กับ ส่วนของคำที่ระบุนั้น เพื่อใช้กำหนดจำนวนอักษรส่วนที่ละไว้ วิธีการต่าง ๆ เหล่านี้ จะต้องมีการกำหนดไว้ด้วย

3.3.3 <u>แนวทางสำคัญในการจัดทำขบวนการสืบค้นสารสนเทศ</u>

จากหลักการต่าง ๆ ที่นำมาใช้สำหรับจัดทำระบบสืบค้นสารสนเทศ ให้สามารถตอบสนองความต้องการค้นของผู้ใช้ได้อย่างกว้างขวาง โดยเฉพาะความต้องการค้นหาข้อมูลในลักษณะที่ซับซ้อน ซึ่งมีแนวทางในการจัดทำที่สำคัญ คือ ในขบวนการสืบค้นสารสนเทศ จำเป็นต้องมีการใช้**แฟ้มข้อมูลพิเศษ** สำหรับใช้เก็บผลลัพธ์ที่ได้จากการดำเนินปฏิบัติการแต่ละขั้นตอนไว้ชั่วคราว เพื่อสามารถนำผลลัพธ์เหล่านั้น มาดำเนินการในขั้นต่อไป จนได้ผลลัพธ์ของการค้นออกมาในที่สุด และในการจัดทำระบบสืบค้น ที่เอื้ออำนวยให้ผู้ใช้มีอิสระในการกำหนดทิศทางการค้นได้ด้วยตนเองนั้น

จำเป็นต้องจัดเก็บผลลัพธ์ที่ได้จากการสืบค้นแต่ละครั้งไว้ใน แฟ้มข้อมูลพิเศษ และให้ผู้ใช้สามารถนำผลลัพธ์นี้ ไปใช้ดำเนินปฏิบัติการค้นในขั้นตอนถัดไปที่ผู้ใช้จะกำหนดให้อีกทีหนึ่ง รวมทั้งจัดให้มีระบบจัดการ แฟ้มข้อมูลเก็บผลลัพธ์เหล่านี้โดยเฉพาะ เพื่ออำนวยความสะดวกแก่ผู้ใช้ด้วย ดังนั้น การออกแบบโครงสร้างของ แฟ้มข้อมูลพิเศษ ให้สามารถใช้งานได้อย่างมีประสิทธิภาพในลักษณะดังกล่าวนั้น จึงเป็นสิ่งสำคัญอย่างยิ่ง ที่ต้องพิจารณาวางแผนอย่างรอบคอบ โดยจะต้องคำนึงถึงการประหยัดเนื้อที่เก็บด้วย

ในบทถัดไปจะได้บรรยายถึง การออกแบบระบบสืบค้นสารสนเทศพระไตรปิฎกระบบใหม่ ซึ่งจะมีการออกแบบ โครงสร้างของขบวนการสืบค้นสารสนเทศของระบบใหม่ ที่ได้ขยายขอบข่ายการค้นออกไปอย่างกว้างขวาง และมีการออกแบบโครงสร้างแฟ้มข้อมูลพิเศษ ที่ใช้เก็บสารสนเทศผลลัพธ์จากการค้น ซึ่งเป็นการเพิ่มเติมเข้าไปเป็นส่วนหนึ่ง ของฐานข้อมูลในระบบใหม่ ซึ่งจะต้องสามารถเชื่อมโยงกับข้อมูลเดิมใน ฐานข้อมูลพระไตรปิฎก ได้ด้วย