บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

บทนี้เป็นการนำเสนอวิธีการค้นหาคำสำคัญในฐานข้อมูลวิธีต่างๆ โดยที่ผู้ใช้ไม่จำเป็นต้องมีความรู้เกี่ยวกับโครงสร้างฐานข้อมูลและภาษาที่ใช้ในการสอบถาม ซึ่งในส่วนที่ 2.1 กล่าวถึงข้อมูลพื้นฐานของการค้นหาคำสำคัญในฐานข้อมูล ส่วนที่ 2.2 กล่าวถึง ระบบการค้นหาคำสำคัญในฐานข้อมูลที่ใช้คำดัชนีที่ตรงกับค่าข้อมูลระบบต่างๆ ส่วนที่ 2.3 กล่าวถึง ระบบการค้นหาคำสำคัญในฐานข้อมูลที่ใช้คำดัชนีที่ตรงกับค่าข้อมูลและข้อมูลฐานข้อมูล และสุดท้าย ส่วนที่ 2.4 เป็นการวิเคราะห์วิธีการค้นหาคำสำคัญในฐานข้อมูลจากระบบต่างๆ

2.1 การค้นหาคำสำคัญในฐานข้อมูล

ระบบการค้นหาคำสำคัญในฐานข้อมูล เป็นระบบที่ช่วยให้ผู้ใช้สามารถค้นหาข้อมูลจากฐานข้อมูลได้สะดวกยิ่งขึ้น โดยที่ผู้ใช้ไม่จำเป็นต้องมีความรู้เกี่ยวกับโครงสร้างฐานข้อมูลและภาษาที่ใช้ในการสอบถาม เช่น ภาษาเอสคิวแอล ซึ่งคล้ายกับการค้นหาข้อมูลบนอินเทอร์เน็ต

การค้นหาคำสำคัญ (Keyword Searching) คือ การที่ผู้ใช้ระบุคำที่มีนัยสำคัญ (Significant Word) หรือคำที่สามารถค้นหาได้จากดัชนี เพื่อค้นหาข้อมูลที่ต้องการ โดยคำดัชนีต้องเป็นคำที่มีนัยสำคัญที่ปรากฏอยู่ในชื่อเรื่อง เนื้อเรื่อง หรือบทความ [8]

ในที่นี้กำหนดให้ คำสำคัญ คือ คำที่ผู้ใช้ระบุในการสอบถามและมีนัยสำคัญ หรือคำที่ปรากฏในดัชนีผกผัน ส่วนคำดัชนี คือ คำหรือดัชนี ในแฟ้มดัชนีผกผัน

การค้นหาคำสำคัญในฐานข้อมูล คือ การค้นหาระเบียนข้อมูลทุกระเบียนจากฐานข้อมูลที่ประกอบด้วยคำสำคัญ และ/หรือระเบียนข้อมูลที่มีชื่อเอนทิตีหรือชื่อแอทริบิวตรงกับคำสำคัญตามที่ผู้ใช้ต้องการ โดยที่ระเบียนเหล่านั้นต้องมีความสัมพันธ์กันตามโครงสร้างฐานข้อมูล คำสำคัญในที่นี้เป็นคำที่ปรากฏอยู่ในข้อมูล 2 ประเภท คือ ค่าข้อมูลที่ได้จากฐานข้อมูล และข้อมูลฐานข้อมูล เช่น ชื่อเอนทิตี และชื่อแอทริบิว เป็นต้น หากคำสำคัญปรากฏในค่าข้อมูล คำสำคัญนั้นจะใช้ในการระบุระเบียนข้อมูล และหากคำสำคัญปรากฏในข้อมูลฐานข้อมูล คำสำคัญนั้นจะใช้ในการระบุเอนทิตีหรือแอทริบิว

ระบบการค้นหาคำสำคัญในฐานข้อมูลในปัจจุบันสามารถแบ่งตามประเภทของคำดัชนีได้ 2 ประเภท คือ ระบบที่ใช้คำดัชนีตรงกับค่าข้อมูล และระบบที่ใช้คำดัชนีตรงกับค่าข้อมูลและข้อมูลฐานข้อมูล

1. ระบบที่ใช้คำดัชนีตรงกับค่าข้อมูล ประกอบด้วย ระบบดาตาสปอต (DTL's DataSpot) [5] ระบบดิสโคเวอร์ (Discover) [7] ระบบดีบีเอกซ์โพลเรอ (DBXplorer) [3] และระบบอีเคเอสโอ (EKSO) [11]

2. ระบบที่ใช้คำดัชนีตรงกับค่าข้อมูลและข้อมูลฐานข้อมูลแบ่งเป็น 2 วิธี ได้แก่

2.1 คำดัชนีที่ตรงกับค่าข้อมูลหรือชื่อแอทริบิว ประกอบด้วย ระบบดีบีเซอร์ฟเฟอร์ (DBSurfer) [13]

2.2 คำดัชนีที่ตรงกับค่าข้อมูลหรือชื่อเอนทิตี/ชื่อแอทริบิว ประกอบด้วย ระบบบีเอ็นเคเอส (BANKS) [1][2] และระบบมราจียาติ (Mragyati) [10]

หลักการทั่วไปของการค้นหาคำสำคัญในฐานข้อมูลมีดังต่อไปนี้

2.1.1 สถาปัตยกรรมของระบบ

สถาปัตยกรรมของระบบการค้นหาคำสำคัญในฐานข้อมูลแบ่งออกเป็น 2 ส่วน คือ ส่วนการเตรียมข้อมูลเริ่มต้น และส่วนการสอบถาม ในส่วนการเตรียมข้อมูลเริ่มต้น เป็นการเตรียมข้อมูลจากฐานข้อมูลเพื่อใช้ในขั้นตอนการค้นหา เรียกว่า แบบจำลองข้อมูล (Data Model) [12] ส่วนที่สอง คือ ส่วนการสอบถาม เป็นขั้นตอนการค้นหาผลลัพธ์ โดยเริ่มจากการส่งผ่านคิวรีไปค้นหาคำสำคัญ แล้วนำคำสำคัญที่ได้ไปพิจารณาหาผลลัพธ์ตามขั้นตอนของระบบ

2.1.2 แบบจำลองข้อมูลและดัชนีผกผัน

แบบจำลองข้อมูลส่วนใหญ่แบ่งเป็น 2 ประเภท คือ แบบจำลองในรูปของกราฟ [1][3][5][7] และแบบจำลองในรูปของเอกสารเสมือน (Virtual Document) [11] [13]

1. แบบจำลองในรูปของกราฟ แบ่งเป็น 2 ประเภท คือ กราฟโครงสร้างฐานข้อมูล และกราฟข้อมูล กราฟทั้งสองประเภทอาจเป็นกราฟแบบมีทิศทางหรือกราฟไม่มีทิศทาง ขึ้นอยู่กับขั้นตอนในการค้นหาผลลัพธ์ของผู้ออกแบบระบบ

1.1 กราฟโครงสร้างฐานข้อมูล กำหนดให้โหนดใช้แทนเอนทิตี และเส้นเชื่อมใช้แทนความสัมพันธ์ระหว่างเอนทิตี

1.2 กราฟข้อมูล กำหนดให้โหนดใช้แทนระเบียนข้อมูล และเส้นเชื่อมใช้แทนความสัมพันธ์ระหว่างระเบียน

แม้ว่ากราฟข้อมูลจะให้ประสิทธิภาพในการค้นหาดีกว่ากราฟโครงสร้างฐานข้อมูล แต่กราฟข้อมูลจำเป็นต้องใช้หน่วยความจำในการเก็บข้อมูลมากกว่า ในกรณีที่ฐานข้อมูลมีขนาดใหญ่กราฟข้อมูลต้องใช้หน่วยความจำมากยิ่งขึ้น ในขณะที่กราฟโครงสร้างฐานข้อมูลยังคงใช้หน่วยความจำเท่าเดิม

2. แบบจำลองในรูปของเอกสารเสมือน คือ แบบจำลองที่ได้จากการนำค่าข้อมูลจากฐานข้อมูลมาแปลงให้อยู่ในรูปของเอกสารเสมือนหรือเว็บเพจ โดยในแต่ละเอกสารเสมือนอาจประกอบด้วยข้อมูลจากระเบียนข้อมูลเดียวกัน หรือข้อมูลจากหลายระเบียนเชื่อมต่อกันตามความสัมพันธ์ของโครงสร้างฐานข้อมูล

ส่วนดัชนีผกผัน คือ คำดัชนีต่างๆ ที่มีรายการอ้างอิงไปยังตำแหน่งของคำดัชนีนั้นในแบบจำลองข้อมูล

2.1.3 ภาษาสอบถาม

ภาษาสอบถามเป็นการกำหนดรูปแบบของภาษาที่ใช้ในการสอบถาม ระบบการค้นหาคำสำคัญในฐานข้อมูลส่วนใหญ่กำหนดรูปแบบให้ผู้ใช้ระบุคำสำคัญเป็นการสอบถามแบบคำเดี่ยว (Single Word Query) การสอบถามแบบวลี (Phrase Query) การสอบถามแบบบูล (Boolean Query) [8] การสอบถามแบบพร็อกซิมิตี (Proximity Query) [6] หรือการสอบถามแบบกำหนดรูปแบบ (Pattern Matching Query)

2.1.4 ผลลัพธ์จากการสอบถาม

ผลลัพธ์ที่ได้จากการสอบถาม ประกอบด้วยระเบียนข้อมูลจากเอนทิตีใดเอนทิตีหนึ่ง หรือการจอยกันของระเบียนข้อมูลจากหลายๆ เอนทิตีที่มีความสัมพันธ์กันตามโครงสร้างฐานข้อมูล โดยที่ระเบียนข้อมูลเหล่านั้นอาจประกอบด้วยคำสำคัญที่ได้จากการสอบถามทั้งหมดหรือไม่ก็ได้

2.1.5 การจัดลำดับผลลัพธ์

เนื่องจากคำสำคัญที่ได้จากการสอบถามอาจตรงกับแบบจำลองข้อมูลในหลายตำแหน่ง ดังนั้นผลลัพธ์จึงอาจมีได้หลายผลลัพธ์ การจัดลำดับความสำคัญของผลลัพธ์เพื่อแสดงผลแก่ผู้ใช้จึงเป็นสิ่งจำเป็น วิธีการจัดลำดับความสำคัญของผลลัพธ์สามารถทำได้โดยการให้คะแนนกับผลลัพธ์ หลักเกณฑ์ในการให้คะแนนสามารถพิจารณาได้จากปัจจัย 3 ปัจจัย คือ

1. คะแนน IR (IR Score) ของแอทริบิว สามารถคำนวณได้จากจำนวนของคำสำคัญที่ปรากฏในแอทริบิวนั้น หรืออาจใช้วิธีการคำนวณหาน้ำหนักของระบบสืบค้นสารสนเทศ (Information Retrieval Weighting Method) อื่นๆ เช่น การให้ค่า tf-idf เป็นต้น

2. โครงสร้างของกราฟคำตอบ คะแนนของผลลัพธ์ขึ้นอยู่กับขนาดของกราฟคำตอบ ตัวอย่างเช่น จำนวนของโหนดหรือเส้นเชื่อมในกราฟคำตอบ

3. ความหมายของเส้นเชื่อม คะแนนของโหนดขึ้นอยู่กับเส้นเชื่อมระหว่างโหนดๆ หนึ่งกับโหนดใดๆ

การให้คะแนนโดยการพิจารณาจากปัจจัยดังกล่าวขึ้นอยู่กับนิยามของผลลัพธ์ในระบบนั้นๆ ถ้าผลลัพธ์ประกอบด้วยโหนดหลายๆ โหนด การให้คะแนนให้คะแนนตามโครงสร้างของกราฟคำตอบ แต่ถ้าผลลัพธ์มีเพียงโหนดเดียว โครงสร้างของกราฟคำตอบก็ไม่จำเป็นต้องนำมาพิจารณา

2.2 ระบบการค้นหาคำสำคัญในฐานข้อมูลที่ใช้คำดัชนีตรงกับค่าข้อมูล

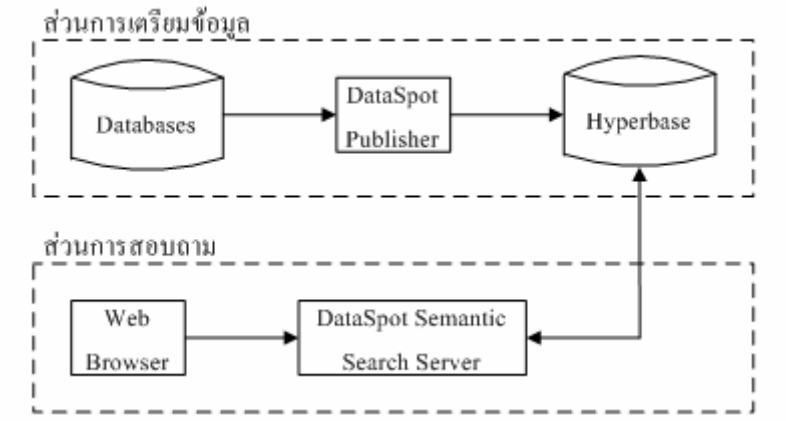
ระบบการค้นหาคำสำคัญในฐานข้อมูลที่ใช้คำดัชนีตรงกับค่าข้อมูล ได้แก่ ดาตาสปอต ดิสโคเวอร์ ดีบีเอกซ์โพลเรอ และอีเคเอสโอ ซึ่งมีรายละเอียดดังต่อไปนี้

2.2.1 ดาตาสปอต

ดาตาสปอต [5] เป็นระบบการค้นหาข้อมูลจากฐานข้อมูลด้วยการระบุข้อความที่เรียบง่าย โดยอาศัยแบบจำลองข้อมูลที่เรียกว่า ไฮเปอร์เบส (Hyperbase) ผลลัพธ์ที่ได้ คือ กราฟย่อยของไฮเปอร์เบส ซึ่งคำสำคัญปรากฏในโหนดใบของกราฟย่อยนั้น

1. สถาปัตยกรรมของระบบ

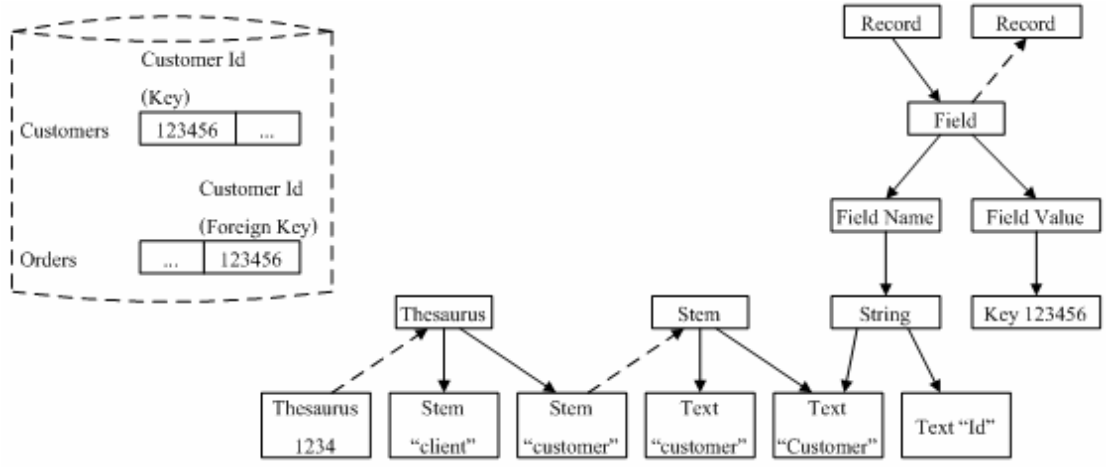
สถาปัตยกรรมของดาตาสปอตเป็นดังรูปที่ 2.1 ซึ่งแบ่งเป็น 2 ส่วน คือ ส่วนการเตรียมข้อมูล และส่วนการสอบถาม ในส่วนการเตรียมข้อมูล ระบบใช้แบบจำลองข้อมูลที่เรียกว่าไฮเปอร์เบส ในส่วนการสอบถาม ระบบทำการค้นหาคำสำคัญจากไฮเปอร์เบสโดยตรง



รูปที่ 2.1 สถาปัตยกรรมของดาตาสปอต

2. แบบจำลองข้อมูลและดัชนีผกผัน

แบบจำลองข้อมูลของดาตาสปอต คือ กราฟโครงสร้างฐานข้อมูลรวมกับกราฟข้อมูลแบบมีทิศทาง เรียกว่า ไฮเปอร์เบส ซึ่งประกอบด้วยโหนดซึ่งใช้แทนเอนทิตี แอทริบิว และค่าข้อมูล นอกจากนี้ยังประกอบด้วยโหนดซึ่งใช้แทนรากศัพท์และคำพ้องความหมายของโหนดค่าข้อมูลด้วย ในส่วนของดัชนีผกผัน ดาตาสปอตไม่จำเป็นต้องใช้ดัชนีผกผันในการค้นหาคำสำคัญ แต่สามารถค้นหาจากไฮเปอร์เบสได้โดยตรง ตัวอย่างฐานข้อมูลและไฮเปอร์เบสเป็นดังรูปที่ 2.2



รูปที่ 2.2 ตัวอย่างฐานข้อมูลและกราฟไฮเปอร์เบส

3. ขั้นตอนวิธีในการหาผลลัพธ์

ระบบทำการส่งผ่านคิวรีไปค้นหาคำสำคัญและตำแหน่งของโหนดใบที่คำสำคัญนั้นปรากฏอยู่จากไฮเปอร์เบส ผลลัพธ์ที่ได้คือ กราฟย่อยที่เชื่อมโยงโหนดดังกล่าวเข้าด้วยกัน และคำสำคัญทุกคำต้องปรากฏในโหนดของกราฟย่อยนั้น ผลลัพธ์ที่ได้เรียงลำดับความสำคัญตามขนาดของกราฟย่อย

4. การวิเคราะห์

การวิเคราะห์ระบบตามหลักการทั่วไปของการค้นหาคำสำคัญในฐานข้อมูล มีดังต่อไปนี้

4.1 สถาปัตยกรรมของระบบ ระบบแบ่งการทำงานเป็น 2 ส่วนหลัก คือ การเตรียมข้อมูลและการสอบถาม

4.2 แบบจำลองข้อมูลและดัชนีผกผัน ระบบใช้กราฟโครงสร้างฐานข้อมูลร่วมกับกราฟข้อมูลเป็นแบบจำลองข้อมูล รวมทั้งนำรากศัพท์ และคำพ้องความหมายมาเป็นส่วนหนึ่งของแบบจำลอง ทำให้ผู้ใช้สามารถค้นหาคำสำคัญจากรากศัพท์ และคำพ้องความหมายได้ แต่ไม่สามารถค้นหาคำสำคัญที่ตรงกับข้อมูลฐานข้อมูลได้ และเนื่องจากระบบทำการคัดลอกค่าข้อมูลไปยังไฮเปอร์เบส และไม่มีการนำดัชนีผกผันมาใช้ ระบบจึงทำการค้นหาคำสำคัญจากไฮเปอร์เบส

โดยตรง นั่นคือ การค้นหาตำแหน่งของคำสำคัญต้องค้นหาจากกราฟแบบจำลองข้อมูล ซึ่งระบบต้องทำงานกับกราฟขนาดใหญ่

4.3 ภาษาสอบถาม ระบบอนุญาตให้ผู้ใช้ใช้ภาษาข้อความอิสระแบบง่ายๆ หรือการสอบถามแบบวลีในการสอบถาม

4.4 ผลลัพธ์จากการสอบถาม คือ กราฟย่อยของไฮเปอร์เบส กราฟย่อยดังกล่าวประกอบด้วยโหนดต่างๆ โดยที่โหนดใบต้องมีคำสำคัญทั้งหมดปรากฏอยู่

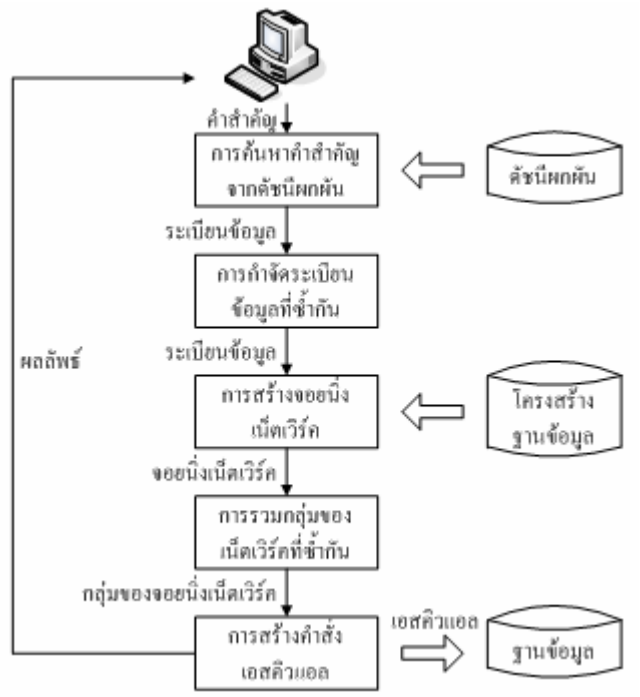
4.5 การจัดลำดับผลลัพธ์ ระบบจัดลำดับความสำคัญของผลลัพธ์ตามโครงสร้างของกราฟคำตอบ หรือตามขนาดของกราฟ ซึ่งกราฟที่มีขนาดเล็กที่สุดเป็นกราฟคำตอบที่มีนัยสำคัญมากที่สุด การจัดลำดับความสำคัญดังกล่าวเพิ่มความความสะดวกให้กับผู้ใช้ในการเลือกผลลัพธ์

2.2.2 ดิสโคเวอร์

ดิสโคเวอร์ [7] เป็นระบบการค้นหาคำสำคัญในฐานข้อมูลเชิงสัมพันธ์ โดยใช้กราฟโครงสร้างฐานข้อมูลในการค้นหาผลลัพธ์ ผลลัพธ์ที่ได้จากการค้นหา คือ ระเบียนข้อมูลที่สัมพันธ์กัน ระเบียนข้อมูลเหล่านี้ประกอบด้วยคำสำคัญที่ได้จากการสอบถามทั้งหมดหรือไม่ก็ได้ ในกรณีที่ผลลัพธ์มีหลายผลลัพธ์ ระบบทำการจัดลำดับความสำคัญของผลลัพธ์ตามจำนวนการจอยกันของระเบียนข้อมูล และคะแนน IR ของแอทริบิว

1. สถาปัตยกรรมของระบบ

สถาปัตยกรรมของระบบดิสโคเวอร์เน้นไปที่ส่วนการสอบถาม กล่าวคือ ระบบใช้กราฟโครงสร้างข้อมูลเป็นแบบจำลองข้อมูล และในส่วนการสอบถามประกอบด้วยขั้นตอนการทำงานหลัก 5 ขั้นตอน ดังรูปที่ 2.3 ซึ่งขั้นตอนการทำงานดังกล่าว ได้แก่ การค้นหาคำสำคัญจากดัชนี การกำจัดระเบียนข้อมูลที่ซ้ำกัน การสร้างจอยนิ่งเน็ตเวิร์ค การรวมกลุ่มของเน็ตเวิร์คที่ซ้ำกัน และการสร้างคำสั่งเอสคิวแอล



รูปที่ 2.3 สถาปัตยกรรมของดิสโคเวอร์

2. แบบจำลองข้อมูลและดัชนีผกผัน

ระบบใช้แบบจำลองในรูปของกราฟโครงสร้างฐานข้อมูล สมมติให้ฐานข้อมูลหนึ่งประกอบด้วยเอนทิตีจำนวน n เอนทิตี $(E_1, E_2, \ldots, E_n)$ แต่ละเอนทิตี E_i มีจำนวนแอทริบิวเท่ากับ m_i แอทริบิว $(a_1^i, a_2^i, \ldots, a_{m_i}^i)$ แบบจำลองข้อมูลของดิสโคเวอร์ถูกเก็บอยู่ในรูปของกราฟแบบมีทิศทาง โดยกำหนดให้โหนด E_i ใช้แทนเอนทิตี E_i และเส้นเชื่อม $E_i \rightarrow E_j$ ใช้แทนความสัมพันธ์ระหว่างเอนทิตีจากคีย์หลักไปยังคีย์นอก ซึ่งเชื่อมระหว่างเซตของแอทริบิว $(a_{b_1}^i, a_{b_2}^i, \ldots, a_{b_l}^i)$ ของ E_i ไปยังเซตของแอทริบิว $(a_{b_1}^j, a_{b_2}^j, \ldots, a_{b_l}^j)$ ของ E_j เมื่อ $a_{b_k}^i \equiv a_{b_k}^j$ โดยที่ k = 1, 2, …, l

ระบบใช้ดัชนีผกผันในการเก็บคำดัชนี แต่ละคำดัชนีมีการอ้างอิงไปยังแบบจำลองข้อมูล โดยกำหนดให้คำดัชนีคือค่าข้อมูลประเภทอักขระ

3. ขั้นตอนวิธีในการหาผลลัพธ์

ในส่วนของขั้นตอนวิธีในการหาผลลัพธ์ จากตัวอย่างข้อมูลในรูปที่ 2.4 เมื่อทำการค้นหาด้วยคำสำคัญ K = {Smith, Miller} ดิสโคเวอร์มีขั้นตอนการทำงานดังต่อไปนี้

ขั้นตอนที่ 1 จากดัชนีผกผัน เมื่อได้ตำแหน่งของคำสำคัญในแบบจำลองข้อมูลแล้วระบบจะทำการดึงระเบียนข้อมูลที่คำสำคัญนั้นปรากฏอยู่ จากตัวอย่างการค้นหา K ได้ระเบียนข้อมูลดังนี้

Smith ได้ระเบียน O_1

Miller ได้ระเบียน O_2 และ O_3

ขั้นตอนที่ 2 พิจารณาหาระเบียน R_i^K ที่เป็นไปได้ทั้งหมด เพื่อกำจัดระเบียนที่อาจซ้ำกันในขั้นตอนที่ 1

ขั้นตอนที่ 3 นำกลุ่มของระเบียนที่ได้มาจอย (Join) กันตามเงื่อนไขที่ได้จากแบบจำลองข้อมูล เรียกว่า จอยนิ่งเน็ตเวิร์ค (Joining Network) จากการสอบถาม K ได้จอยนิ่งเน็ตเวิร์ค J_1 และ J_2 ดังนี้

$$J_1 = \text{ORDERS}^{\text{Smith}} \bowtie \text{CUSTOMER}^{\{\}} \bowtie \text{ORDERS}^{\text{Miller}}$$

$$J_2 = \text{ORDERS}^{\text{Smith}} \bowtie \text{CUSTOMER}^{\{\}} \bowtie \text{NATION}^{\{\}} \bowtie \text{CUSTOMER}^{\{\}} \bowtie \text{ORDERS}^{\text{Miller}}$$

ขั้นตอนที่ 4 พิจารณาส่วนของจอยนิ่งเน็ตเวิร์คที่เหมือนกันในแต่ละเน็ตเวิร์ค เพื่อลดจำนวนการจอยกันของข้อมูลและสามารถใช้ส่วนของจอยนิ่งเน็ตเวิร์คนั้นร่วมกันได้ ผลลัพธ์ที่ได้เรียงลำดับตามจำนวนการจอยกันในแต่ละเน็ตเวิร์ค โดยผลลัพธ์ที่มีจำนวนการจอยกันของค่าข้อมูลน้อยกว่าจะถือว่ามีความสำคัญมากกว่า จากตัวอย่าง ส่วนของจอยนิ่งเน็ตเวิร์คที่เหมือนกัน คือ T_1

$$T_1 \leftarrow J_1 = \text{ORDERS}^{\text{Smith}} \bowtie \text{CUSTOMER}^{\{\}}$$

$$C_1 \leftarrow T_1 \bowtie \text{ORDERS}^{\text{Miller}}$$

$$C_2 \leftarrow T_1 \bowtie \text{NATION}^{\{\}} \bowtie \text{CUSTOMER}^{\{\}} \bowtie \text{ORDERS}^{\text{Miller}}$$

ขั้นตอนที่ 5 สร้างคำสั่งเอสคิวแอล

ORDERS

	ORDKEY	CUSTKEY	ORDSTATUS	TOTALPRICE	ORDDATE	CLERK
O1	100105	12312	complete	$5,000	5/2/2001	John Smith
O2	100111	12312	incomplete	$3,000	5/1/2001	Mike Miller
O3	100125	10001	incomplete	$7,000	5/1/2001	Mike Miller
O4	100110	10002	complete	$8,000	4/2/2001	Keith Brown

CUSTOMER

	CUSTKEY	NAME	ADDRESS	NATKEY
c1	12312	Brad Lou	3811 State Drive, Los Angeles	01
c2	10001	George Wales	43655 Ave, New York	01
c3	10013	John Roberts	3321 Broadways St, San Francisco	01

NATION

	NATKEY	NAME	REGIONKEY
n1	01	USA	N America

รูปที่ 2.4 ตัวอย่างข้อมูลจากฐานข้อมูล TPC-H

4. การวิเคราะห์

การวิเคราะห์ระบบตามหลักการทั่วไปของการค้นหาคำสำคัญในฐานข้อมูล มีดังต่อไปนี้

4.1 สถาปัตยกรรมของระบบ เน้นในส่วนการสอบถาม ประสิทธิภาพของระบบขึ้นอยู่กับจำนวนของคำสำคัญ การวิเคราะห์หาจอยนิ่งเน็ตเวิร์ค จำนวนของจอยนิ่งเน็ตเวิร์คที่ได้ และการพิจารณาส่วนของจอยนิ่งเน็ตเวิร์คที่เหมือนกันในแต่ละเน็ตเวิร์คเพื่อสร้างส่วนของการจอยที่สามารถใช้ร่วมกันได้

4.2 แบบจำลองข้อมูลและดัชนีผกผัน ระบบใช้กราฟโครงสร้างฐานข้อมูลเป็นแบบจำลองข้อมูล และใช้ดัชนีผกผันในการชี้ตำแหน่งของคำดัชนีไปยังแบบจำลองข้อมูล โดยกำหนดให้คำดัชนีคือค่าข้อมูลเท่านั้น ทำให้ผู้ใช้ไม่สามารถค้นหาคำพ้องความหมายของคำดัชนี และไม่สามารถค้นหาด้วยคำสำคัญที่ตรงกับข้อมูลฐานข้อมูลได้

4.3 ภาษาสอบถาม ระบบกำหนดให้การสอบถามเป็นการระบุเซตของคำสำคัญ หรือเซตของคำสำคัญแบบเดี่ยว ไม่สามารถใช้ข้อความอิสระในการสอบถามได้ นอกจากนี้คำสำคัญที่ใช้ค้นหาอาจตรงกับค่าข้อมูล หรือส่วนหนึ่งของค่าข้อมูล (Sub-String Matching)

4.4 ผลลัพธ์จากการสอบถาม คือ ระเบียนข้อมูลที่สัมพันธ์กัน ระเบียนข้อมูลเหล่านี้ประกอบด้วยคำสำคัญที่ได้จากการสอบถามทั้งหมดหรือไม่ก็ได้

4.5 การจัดลำดับผลลัพธ์ ในกรณีที่ผลลัพธ์มีหลายผลลัพธ์ ระบบมีการจัดลำดับความสำคัญของผลลัพธ์โดยพิจารณาจากจำนวนการจอยกันของระเบียนข้อมูล หรือตามโครงสร้างของกราฟคำตอบ โดยพิจารณาร่วมกับการให้คะแนน IR ของแอทริบิว เพื่อให้ผู้ใช้เลือกผลลัพธ์ที่ดีที่สุด

2.2.3 ดีบีเอกซ์โพลเรอ

ดีบีเอกซ์โพลเรอ [3] เป็นระบบการค้นหาคำสำคัญในฐานข้อมูลอีกระบบหนึ่ง การหาผลลัพธ์ทำได้โดยการค้นหาเซตของเอนทิตี และแอทริบิวหรือระเบียนที่ประกอบด้วยคำสำคัญ เอนทิตีทั้งหมดที่ได้เชื่อมต่อกันตามโครงสร้างฐานข้อมูล ผลลัพธ์ที่ได้ คือ กราฟย่อยที่ประกอบด้วยคำสำคัญทั้งหมดจากการสอบถาม จากนั้นจึงแปลงกราฟย่อยเป็นภาษาเอสคิวแอล ในกรณีที่ผลลัพธ์มีหลายผลลัพธ์ ระบบทำการจัดลำดับความสำคัญของผลลัพธ์ตามกราฟย่อยที่มีระยะทางที่สั้นที่สุด นั่นคือ ผลลัพธ์ที่มีความสัมพันธ์ระหว่างข้อมูลใกล้ชิดกันมากที่สุดให้ถือว่าเป็นผลลัพธ์ที่ดีที่สุด

1. สถาปัตยกรรมของระบบ

สถาปัตยกรรมของระบบดีบีเอกซ์โพลเรอแบ่งออกเป็น 2 ส่วน คือ ส่วนการเตรียมข้อมูลที่เรียกว่า พับลิช (Publish Component) และส่วนการสอบถาม เรียกว่า เสิร์ช (Search Component) ในส่วนของพับลิชประกอบด้วยอินเทอร์เฟส (Interface) สำหรับการเลือกฐานข้อมูล การเลือกเอนทิตีหรือแอทริบิว และการปรับปรุง หรือลบ หรือจัดการแบบจำลองข้อมูล ในส่วนของการสอบถามประกอบด้วยอินเทอร์เฟสสำหรับการค้นหาคำสำคัญจากแบบจำลองข้อมูล และการระบุเอนทิตี แอทริบิว หรือระเบียนสำหรับขั้นตอนการหาผลลัพธ์

2. แบบจำลองข้อมูลและดัชนีผกผัน

ระบบดีบีเอกซ์โพลเรอทำการจัดเก็บแบบจำลองข้อมูลในรูปของกราฟโครงสร้างฐานข้อมูลแบบไม่มีทิศทาง โดยกำหนดให้โหนดใช้แทนเอนทิตี และเส้นเชื่อมใช้แทนความสัมพันธ์ระหว่างเอนทิตี ส่วนการจัดเก็บดัชนีผกผัน ระบบทำการเก็บคำดัชนี 2 ระดับ คือ ระดับคอลัมน์ (Column Level Granularity) และระดับเซลล์ (Cell Level Granularity) ในการเก็บคำดัชนีระดับคอลัมน์ คำดัชนีถูกระบุตำแหน่งไปยังคอลัมน์หรือแอทริบิวที่คำดัชนีนั้นปรากฏอยู่ ส่วนการเก็บคำดัชนีในระดับเซลล์ คำดัชนีถูกระบุตำแหน่งไปยังเซลล์ที่คำดัชนีนั้นปรากฏ ดังตารางที่ 2.1

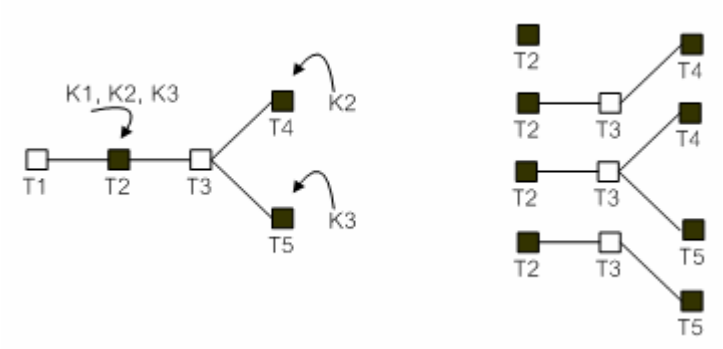
ตารางที่ 2.1 แสดงวิธีการเก็บดัชนีผกผันในระบบดีบีเอ็กซ์โพลเรอ

ระดับการเก็บดัชนี	วิธีการอ้างถึง
ระดับคอลัมน์	table.column
ระดับเซลล์	table.column.rowid

3. ขั้นตอนวิธีในการหาผลลัพธ์

ระบบนำตำแหน่งของคำสำคัญที่ได้จากดัชนีไปตรวจสอบกับกราฟโครงสร้างฐานข้อมูลว่าอยู่ในโหนดใดของกราฟ แล้วจึงทำการตัดโหนดที่ไม่ปรากฏคำสำคัญและโหนดที่ไม่ได้เชื่อม

ระหว่างคำสำคัญทิ้ง กราฟที่เหลืออยู่จะประกอบด้วยเอนทิตีที่คำสำคัญนั้นปรากฏอยู่ทั้งหมด โดยไม่มีคำสำคัญใดปรากฏซ้ำกันมากกว่า 1 โหนด ดังตัวอย่างจากรูปที่ 2.5 สมมติให้ K1, K2, และ K3 เป็นคำสำคัญที่ถูกพบในโหนดดังรูปทางซ้าย ส่วนรูปทางขวาถือเป็นกราฟคำตอบทั้งหมดที่ได้จากคำสำคัญ



รูปที่ 2.5 จอยทรี

4. การวิเคราะห์

การวิเคราะห์ระบบตามหลักการทั่วไปของการค้นหาคำสำคัญในฐานข้อมูล มีดังต่อไปนี้

4.1 สถาปัตยกรรมของระบบแบ่งเป็น 2 ส่วน คือ ส่วนการเตรียมข้อมูลที่เรียกว่า พับลิช และส่วนการสอบถาม เรียกว่า เสิร์ช

4.2 แบบจำลองข้อมูลและดัชนีผกผัน แบบจำลองข้อมูลของดีบีเอกซ์โพลเรออยู่ในรูปของกราฟโครงสร้างฐานข้อมูล โดยใช้ดัชนีผกผัน 2 ระดับ คือ ระดับคอลัมน์ และระดับเซลล์ ซึ่งจากการทดสอบดัชนีทั้งสองระดับใน [3] พบว่าดัชนีระดับคอลัมน์ให้ประสิทธิภาพในการทำงานดีกว่าดัชนีระดับเซลล์ นอกจากนี้คำดัชนีที่ใช้ในระบบ คือ ค่าข้อมูลจากฐานข้อมูล ทำให้ผู้ใช้ไม่สามารถค้นหาคำพ้องความหมายของคำดัชนี และไม่สามารถค้นหาด้วยคำสำคัญที่ตรงกับข้อมูลฐานข้อมูลได้

4.3 ภาษาสอบถาม ระบบกำหนดให้การสอบถามเป็นการระบุเซตของคำสำคัญ หรือเซตของคำสำคัญแบบเดี่ยว ไม่สามารถใช้ข้อความอิสระในการสอบถามได้ นอกจากนี้คำสำคัญที่ใช้ค้นหาอาจตรงกับค่าข้อมูล หรือส่วนหนึ่งของค่าข้อมูล

4.4 ผลลัพธ์จากการสอบถาม คือ กราฟย่อยที่ประกอบด้วยคำสำคัญทั้งหมดจากการสอบถาม

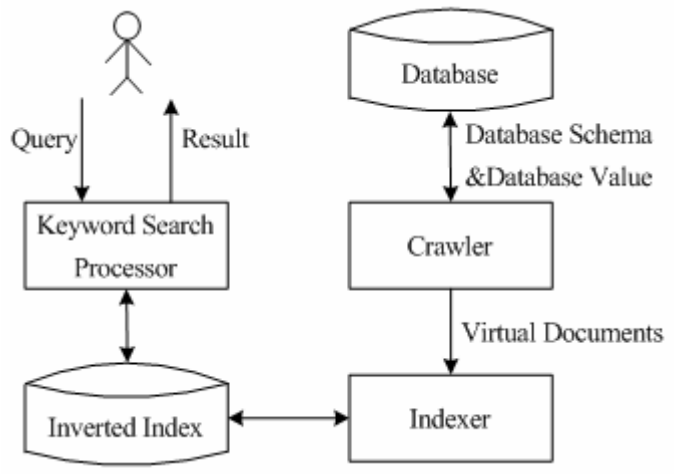
4.5 การจัดลำดับผลลัพธ์ ระบบดีบีเอกซ์โพลเรอใช้โครงสร้างของกราฟคำตอบในการพิจารณาจัดลำดับความสำคัญของผลลัพธ์

2.2.4 อีเคเอสโอ

อีเคเอสโอ [11] เป็นระบบการค้นหาคำสำคัญในฐานข้อมูลเชิงสัมพันธ์ โดยใช้แบบจำลองข้อมูลในรูปของเอกสารเสมือน และใช้ดัชนีผกผันที่มีคำดัชนีตรงกับค่าข้อมูล ซึ่งรายละเอียดของระบบมีดังนี้

1. สถาปัตยกรรมของระบบ

สถาปัตยกรรมของระบบเป็นดังรูปที่ 2.6 ซึ่งประกอบด้วยส่วนสำคัญ 3 ส่วน ส่วนแรก คือ ครอว์เลอร์ (Crawler) มีหน้าที่ในการดึงข้อมูลจากฐานข้อมูลมาสร้างเอกสารเสมือน ส่วนที่สอง คือ ตัวสร้างดัชนี (Indexer) มีหน้าที่สร้างดัชนีผกผันจากเอกสารเสมือน และสุดท้าย คือ กระบวนการค้นหาคำสำคัญ (Keyword Search Processor)



รูปที่ 2.6 สถาปัตยกรรมของระบบอีเคเอสโอ

2. แบบจำลองข้อมูลและดัชนีผกผัน

แบบจำลองข้อมูลของระบบอยู่ในรูปของเอกสารเสมือน จากรูปที่ 2.6 เอกสารเสมือนถูกสร้างโดยครอว์เลอร์ วิธีการสร้างเอกสารเสมือนแบ่งเป็น 2 ขั้นตอน คือ การสร้างเทกซ์ออบเจกต์ และการสร้างเอกสารเสมือน

2.1 การสร้างเทกซ์ออบเจกต์

เทกซ์ออบเจกต์ประกอบด้วยเซตของระเบียนข้อมูลที่สัมพันธ์กับรูททูเพิล (Root Tuple) [9] ซึ่งถูกกำหนดโดยอัตโนมัติหรือกำหนดโดยผู้ดูแลระบบ จากตัวอย่างข้อมูลในรูปที่ 2.7 เทกซ์ออบเจกต์ของรูททูเพิล P1 ประกอบด้วยเซตของระเบียนข้อมูล {P1, PS1, L1, S1, O1, C1}

2.2 การสร้างเอกสารเสมือน

เอกสารเสมือนประกอบด้วยค่าข้อมูลจากแอทริบิวที่มีชนิดข้อมูลเป็นอักขระ โดยที่แอทริบิวเหล่านั้นมาจากระเบียนข้อมูลในเทกซ์ออบเจกต์ จากตัวอย่างเทกซ์ออบเจกต์ดังกล่าว เอกสารเสมือนที่ได้ คือ Discus fish Acme Wisconsin Bob Madison

Part

	PARTKEY	NAME	...
P1	10	Discus fish	
P2	11	Apple PowerBook	
P3	12	Snow shevel	

Supplier

	SUPPKEY	NAME	ADDRESS	...
S1	1000	Acme	Wisconsin	
S2	1001	Vidgets	California	

Partsupp

	PARTKEY	SUPPKEY	...
PS1	10	1000	
PS2	11	1001	
PS3	12	1000	
PS4	12	1001	

Lineitem

	ORDERKEY	PARTKEY	SUPPKEY	LINENUMBER	...
L1	10001	10	1000	2	
L2	10002	11	1001	1	
L3	10003	12	1000	3	

Orders

	ORDERKEY	CUSTKEY	...
O1	10001	100	
O2	10002	100	
O3	10003	101	

Customer

	CUSTKEY	NAME	ADDRESS	...
C1	100	Bob	Madison	
C2	101	Alis	San Francisco	

รูปที่ 2.7 ตัวอย่างข้อมูลจากฐานข้อมูล TPC-H

ในส่วนของดัชนีผกผัน จากรูปที่ 2.6 ตัวสร้างดัชนีทำการสร้างดัชนีผกผันจากเอกสารเสมือนดังกล่าว โดยใช้ดีบีทูเอ็นเอสอี (DB2 NSE: DB2 Net Search Extender)

3. ขั้นตอนวิธีในการหาผลลัพธ์

ในกระบวนการค้นหาคำสำคัญ โปรแกรมค้นหาเอ็นเอสอี (NSE Search Engine) ทำการค้นหาคำสำคัญจากดัชนีผกผัน และคืนค่าผลลัพธ์ซึ่งเป็นรหัสในเทกซ์ออบเจกต์ ซึ่งผู้ใช้สามารถเลือกได้ว่าต้องการค้นดูข้อมูลในฐานข้อมูลจากรหัสเหล่านี้ หรือต้องการสอบถามใหม่

4. การวิเคราะห์

การวิเคราะห์ระบบตามหลักการทั่วไปของการค้นหาคำสำคัญในฐานข้อมูล มีดังต่อไปนี้

4.1 สถาปัตยกรรมของระบบแบ่งเป็น 2 ส่วน คือ ส่วนการเตรียมข้อมูล ซึ่งประกอบด้วย ครอว์เลอร์และตัวสร้างดัชนี และส่วนการสอบถาม

4.2 แบบจำลองข้อมูลและดัชนีผกผัน แบบจำลองข้อมูลของระบบอยู่ในรูปของเอกสารเสมือน โดยในแต่ละเอกสารเสมือนประกอบด้วยค่าข้อมูลและโครงสร้างฐานข้อมูล ทำให้การค้นหามีประสิทธิภาพยิ่งขึ้น ในส่วนของดัชนีผกผันซึ่งสร้างจากเอกสารเสมือนดังกล่าว ประกอบด้วยคำดัชนีที่มาจากข้อมูลประเภทอักขระ ระบบจึงไม่สามารถค้นหาคำพ้องความหมายของคำดัชนี ไม่สามารถจัดการกับคำดัชนีที่เป็นจำนวนหรือวันที่ได้ และไม่สามารถค้นหาคำสำคัญที่ตรงกับข้อมูลฐานข้อมูลได้

4.3 ภาษาสอบถาม ระบบกำหนดให้การสอบถามเป็นการระบุเซตของคำสำคัญ หรือเซตของคำสำคัญแบบเดี่ยว

4.4 ผลลัพธ์จากการสอบถาม คือ ระเบียนข้อมูลที่ปรากฏคำสำคัญทั้งหมด

2.3 ระบบการค้นหาคำสำคัญในฐานข้อมูลที่ใช้คำดัชนีตรงกับค่าข้อมูลและข้อมูลฐานข้อมูล

ระบบที่ใช้คำดัชนีตรงกับค่าข้อมูลและข้อมูลฐานข้อมูลแบ่งเป็น คำดัชนีที่ตรงกับค่าข้อมูลหรือชื่อแอทริบิว คำดัชนีที่ตรงกับค่าข้อมูลหรือชื่อเอนทิตี และคำดัชนีที่ตรงกับค่าข้อมูลหรือชื่อเอนทิตี/ชื่อแอทริบิว ซึ่งระบบที่ใช้คำดัชนีตรงกับค่าข้อมูลหรือชื่อแอทริบิว ได้แก่ ดีบีเซอร์ฟเฟอร์ ระบบที่ใช้คำดัชนีที่ตรงกับค่าข้อมูลหรือชื่อเอนทิตี ได้แก่ บีเอเอ็นเคเอส และระบบที่ใช้คำดัชนีที่ตรงกับค่าข้อมูลหรือชื่อเอนทิตี/ชื่อแอทริบิว ได้แก่ และมราจียาติ

2.3.1 ดีบีเซอร์ฟเฟอร์

ดีบีเซอร์ฟเฟอร์ [13] เป็นระบบการค้นหาคำสำคัญในฐานข้อมูล ที่ใช้แบบจำลองข้อมูลในรูปของกราฟโครงสร้างฐานข้อมูล และดัชนีผกผันที่สร้างจากเอกสารเสมือน ซึ่งคำดัชนีเป็นคำที่ตรงกับค่าข้อมูลหรือชื่อแอทริบิว ผลลัพธ์ที่ได้จากการสอบถามอยู่ในรูปของเทรล (Trail) [14] ซึ่งไม่จำเป็นต้องแปลงให้อยู่ในรูปของภาษาเอสคิวแอล

1. แบบจำลองข้อมูลและดัชนีผกผัน

ดัชนีผกผันของระบบสร้างจากเอกสารเสมือน หรือเว็บเพจ โดยการดึงข้อมูลในแต่ละระเบียนมาสร้างเป็นเอกสารเสมือนซึ่งอยู่ในรูปของภาษาเอกซ์เอ็มแอล (XML: Extensible Markup Language) ดังตัวอย่างจากรูปที่ 2.8 ซึ่งเป็นเอกสารเสมือนที่ได้จากการดึงข้อมูลจากฐานข้อมูลดีบี

แอลพี (DBLP) [14] ระบบทำการเก็บดัชนีผกผันที่ได้จากเนื้อความ (Textual Content) ของเอกสารเสมือนนี้ รวมทั้งค่า tf.idf ของคำดัชนี

```
1. <PUBLICATION>
2.    <row>
3.       <JOURNAL> Advances in Computers </JOURNAL>
4.       <KEY> journals/ac/Dam66 </KEY>
5.       <PAGES> 239-290 </PAGES>
6.       <TITLE> Computer Driven Displays and Their Use in Man/Machine Interaction. </TITLE>
7.       <TYPE> article </TYPE>
8.       <URL> http://dblp.uni-trier.de/db/journals/ac/ac7.html#Dam66 </URL>
9.       <VOLUME> 7 </VOLUME>
10.      <YEAR> 1966 </YEAR>
11.   </row>
12. </PUBLICATION>
```

รูปที่ 2.8 ตัวอย่างเอกสารเสมือนในรูปของภาษาเอกซ์เอ็มแอลที่ได้จากฐานข้อมูลดีบีแอลพี

ส่วนแบบจำลองข้อมูลของระบบ คือ เอกสารเสมือนที่เปรียบได้กับกราฟข้อมูล และกราฟโครงสร้างฐานข้อมูลที่เรียกว่า ลิงค์กราฟ (Link Graph) ซึ่งพิจารณาจากเงื่อนไขบังคับของคีย์นอก (Foreign Key Constraint) โดยมีลิงค์เชื่อมระหว่างเอนทิตีและแอทริบิว ระหว่างระเบียนข้อมูลกับเอกสารเสมือน ระหว่างเอกสารเสมือนกับเอกสารเสมือน และระหว่างเอกสารเสมือนหรือยูอาร์แอล (URL) ของเอกสารเสมือน

2. ขั้นตอนวิธีในการหาผลลัพธ์

ขั้นตอนวิธีในการหาผลลัพธ์มี 4 ขั้นตอน คือ ขั้นตอนแรก ระบบทำการคำนวณคะแนนของแต่ละโหนดหรือเอกสารเสมือนที่ประกอบด้วยคำสำคัญอย่างน้อยหนึ่งคำ ขั้นตอนที่สอง คือ การสร้างเทรลโดยใช้เบสเทรลอัลกอริทึม (Best Trail Algorithm) [14] ขั้นตอนที่สาม คือ การลบเทรลที่ซ้ำกัน ขั้นตอนสุดท้าย คือ การคำนวณคะแนนของเทรล เพื่อจัดลำดับความสำคัญของเทรลหรือผลลัพธ์ และแสดงผลต่อผู้ใช้

3. การวิเคราะห์

การวิเคราะห์ระบบตามหลักการทั่วไปของการค้นหาคำสำคัญในฐานข้อมูล มีดังต่อไปนี้

3.1 แบบจำลองข้อมูลและดัชนีผกผัน ระบบดีบีเซอร์ฟเฟอร์ใช้แบบจำลองข้อมูลในรูปของเอกสารเสมือน โดยเอกสาร 1 เอกสารใช้แทนข้อมูล 1 ระเบียน และอาศัยกราฟโครงสร้างฐานข้อมูลในการเชื่อมโยงเอกสารเหล่านั้นเข้าด้วยกัน ซึ่งเทียบได้กับการเก็บแบบจำลองในรูปของกราฟข้อมูล การหาผลลัพธ์ของระบบจึงต้องทำงานกับกราฟจำนวนมาก และต้องใช้หน่วยความจำในการเก็บเอกสารเหล่านั้นมากขึ้น ส่วนข้อดีของการเก็บแบบจำลองดังกล่าว คือ สามารถเข้าถึงค่าข้อมูลได้โดยผ่านทางเว็บเพจโดยไม่จำเป็นต้องใช้ภาษาเอสคิวแอลในการติดต่อฐานข้อมูล

ในส่วนของดัชนีผกผันซึ่งสร้างจากเอกสารเสมือนดังกล่าว ประกอบด้วยคำดัชนีที่มาจากข้อมูลประเภทอักขระเท่านั้น ระบบจึงไม่สามารถค้นหาคำพ้องความหมายของคำดัชนี ไม่สามารถจัดการกับคำดัชนีที่เป็นจำนวนหรือวันที่ได้ และไม่สามารถค้นหาคำสำคัญที่ตรงกับข้อมูลฐานข้อมูลได้

3.2 ภาษาสอบถาม ระบบกำหนดให้การสอบถามเป็นการระบุเซตของคำสำคัญ หรือเซตของคำสำคัญแบบเดี่ยว นอกจากนี้ระบบอนุญาตให้ผู้ใช้ทำการสอบถามแบบกำหนดรูปแบบเพื่อให้ใช้คำสำคัญที่ตรงกับชื่อแอทริบิวได้ โดยใช้เครื่องหมาย “=” ในการกำหนดแอทริบิว รูปแบบของภาษาสอบถาม คือ x = y หมายถึง แอทริบิว x มีค่าเท่ากับ y ตัวอย่างเช่น การสอบถาม “simon” หมายถึงข้อมูลจากทุกระเบียนที่ปรากฏคำว่า “simon” ส่วนการสอบถาม “name = simon” หมายถึงระเบียนข้อมูลที่แอทริบิว “name” มีค่าเป็น “simon” เป็นต้น

3.3 ผลลัพธ์จากการสอบถาม คือ เซตของเอกสารเสมือนที่ประกอบด้วยคำสำคัญทั้งหมด เอกสารเสมือนดังกล่าวถูกแสดงผลในรูปของเทรล

3.4 การจัดลำดับผลลัพธ์ ระบบจัดลำดับความสำคัญของผลลัพธ์โดยพิจารณาจากจำนวนของคำสำคัญในแต่ละเทรล จำนวนครั้งของคำสำคัญที่ปรากกฏในเทรล และคะแนน IR ของเทรล

2.3.2 บีเอเอ็นเคเอส

ระบบบีเอเอ็นเคเอส [1][2] เป็นระบบการค้นหาคำสำคัญในฐานข้อมูลโดยใช้กราฟข้อมูลที่กำหนดให้โหนดแทนระเบียนข้อมูล และเส้นเชื่อมแทนความสัมพันธ์ระหว่างระเบียน ผลลัพธ์จากการสอบถามเป็นกราฟย่อยที่ประกอบด้วยคำสำคัญทั้งหมด การจัดลำดับความสำคัญของผลลัพธ์ขึ้นอยู่กับผลรวมของน้ำหนักโหนด (Node Score) และน้ำหนักของเส้นเชื่อม (Edge Weight)

1. แบบจำลองข้อมูลและดัชนีผกผัน

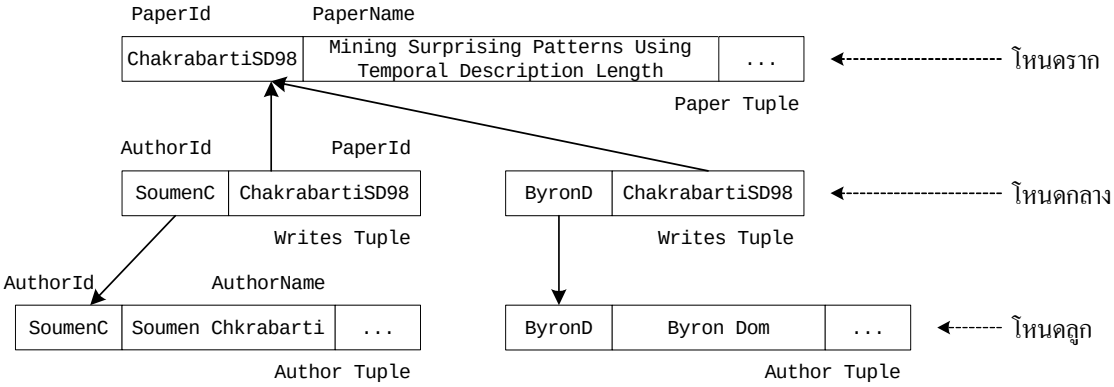
บีเอเอ็นเคเอสมีการเก็บแบบจำลองข้อมูลเป็นกราฟข้อมูลแบบมีทิศทาง โดยการให้โหนดแทนระเบียนข้อมูล และเส้นเชื่อมแทนความสัมพันธ์ระหว่างคีย์หลัก-คีย์นอกของระเบียนข้อมูลนั้น และเก็บดัชนีผกผันเพื่ออ้างไปยังโหนดที่คำสำคัญนั้นปรากฏอยู่

2. ขั้นตอนวิธีในการหาผลลัพธ์

ในการหาผลลัพธ์ บีเอเอ็นเคเอสจะทำการค้นหาเซตของโหนด (S_i) ทั้งหมดที่มีคำสำคัญ (t_i) ตามที่ผู้ใช้ระบุจากดัชนีผกผัน ผลลัพธ์ที่ได้จากการสอบถาม คือ ทรีแบบมีทิศทางที่ประกอบด้วยโหนดราก (Root Node) โหนดใบ (Leaf Node) และ/หรือโหนดกลาง (Intermediate Node) โหนดรากเป็นโหนดที่อยู่ตรงกลางของทรีที่สามารถเชื่อมต่อไปยังทุกๆ โหนดใบได้ โหนดใบเป็นโหนดซึ่งมีคำสำคัญปรากฏอยู่ ส่วนโหนดกลางเป็นโหนดซึ่งเชื่อมระหว่างโหนดรากและโหนดใบ ในแต่ละทรีต้องประกอบด้วยอย่างน้อย 1 โหนดที่มาจากแต่ละ S_i โดยที่ทรีนั้นอาจมี

โหนดที่ไม่จำเป็นต้องเป็นสมาชิกของ S_i ใดๆ ก็ได้เพื่อเป็นโหนดที่เชื่อมระหว่างโหนดใบกับโหนดราก

ตัวอย่างทรีที่เป็นผลลัพธ์จากการการสอบถามด้วย “soumen byron” โดยมีโหนดราก โหนดกลาง และโหนดลูก เป็นดังรูปที่ 2.9



รูปที่ 2.9 ตัวอย่างทรีที่เป็นผลลัพธ์จากการการสอบถามด้วย “soumen byron”

ในกรณีที่ผลลัพธ์มีหลายผลลัพธ์ จะมีเพียงผลลัพธ์ส่วนหนึ่งที่ถูกนำเสนอแก่ผู้ใช้ ผลลัพธ์ดังกล่าว คือ ทรีที่มีน้ำหนักของทรีที่น้อยที่สุดจำนวน n ลำดับ ทรีแต่ละทรีจะถูกนำมาคำนวณน้ำหนัก โดยที่

คะแนนของทรี = ค่า proximity + ค่า prestige

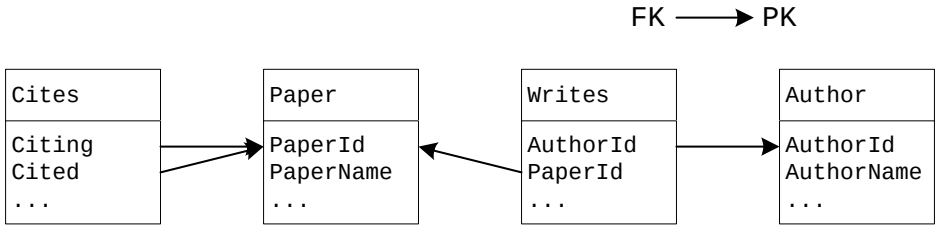
ค่า proximity ของแต่ละทรีได้มาจากน้ำหนักของกิ่งในทรีนั้น ซึ่งสามารถหาได้ ดังต่อไปนี้

proximity = edge score

= log(1/∑ น้ำหนักของกิ่ง)

ในกรณีที่กิ่งนั้นเป็นฟอร์เวิร์ดเอด (Forward Edge) หรือกิ่งที่ชี้ไปยังทิศทางของโหนดราก [3] น้ำหนักของกิ่งจะขึ้นอยู่กับโครงสร้างฐานข้อมูล เช่น จากรูปที่ 2.10 น้ำหนักของกิ่งระหว่างตาราง Cites และ Paper มีค่ามากกว่าน้ำหนักของกิ่งระหว่างตาราง Writes และ Paper ดังนั้นจากรูปที่ 2.9 น้ำหนักของฟอร์เวิร์ดเอดทั้งสองจึงเท่ากัน

ในกรณีที่กิ่งนั้นเป็นแบคเวิร์ดเอด (Backward Edge) หรือกิ่งที่ชี้ไปยังทิศตรงข้ามกับโหนดราก น้ำหนักของกิ่งเท่ากับจำนวนอินดีกรี (Indegree) ของกิ่งที่ชี้มายังโหนดนั้น เช่น จากรูปที่ 2.9 น้ำหนักของกิ่งทั้งสองเท่ากับ 1 เนื่องจากโหนดลูกทั้งสองนั้นมีจำนวนอินดีกรีเท่ากัน



รูปที่ 2.10 ตัวอย่างโครงสร้างฐานข้อมูลของบีเอเอ็นเคเอส

ค่า prestige หาได้จากน้ำหนักของโหนดในทรี โดยพิจารณาเฉพาะโหนดรากและโหนดใบเท่านั้นเพื่อลดผลกระทบจากโหนดกลางที่อาจมีจำนวนมาก ซึ่งสามารถหาได้ดังนี้

prestige = node score

= น้ำหนักของโหนดราก + $\sum$ น้ำหนักของโหนดใบ

โดยที่

น้ำหนักโหนด = log(1+ อินดีกรี)

จากรูปที่ 2.9 โหนดรากมีจำนวนอินดีกรีเท่ากับ 2 โหนดลูกทางซ้ายและโหนดลูกทางขวามีจำนวนอินดีกรีเท่ากัน คือ เท่ากับ 1

3. การวิเคราะห์

การวิเคราะห์ระบบตามหลักการทั่วไปของการค้นหาคำสำคัญในฐานข้อมูล มีดังต่อไปนี้

3.1 แบบจำลองข้อมูลและดัชนีผกผัน เนื่องจากระบบใช้แบบจำลองข้อมูลที่เป็นกราฟข้อมูลทำให้ระบบต้องจัดการกับกราฟจำนวนมาก การทำงานของระบบจึงค่อนข้างช้า และในกรณีที่ฐานข้อมูลมีขนาดใหญ่ ระบบต้องใช้หน่วยความจำในการเก็บข้อมูลมากยิ่งขึ้น ในส่วนของดัชนีผกผัน คำดัชนีที่เก็บ คือ ค่าข้อมูล และชื่อเอนทิตี คำสำคัญที่ใช้ค้นหาจึงเป็นคำที่มีอยู่ในค่าข้อมูลหรือชื่อเอนทิตีเท่านั้น ไม่สามารถค้นหาคำพ้องความหมายของคำดัชนีได้ นอกจากนี้ แม้ระบบจะสามารถค้นหาคำสำคัญที่ตรงกับข้อมูลฐานข้อมูลได้ แต่ยังคงไม่สามารถค้นหาจากชื่อแอทริบิวได้

3.2 ภาษาสอบถาม ระบบกำหนดให้การสอบถามเป็นการระบุเซตของคำสำคัญ หรือเซตของคำสำคัญแบบเดี่ยว ไม่สามารถใช้ข้อความอิสระในการสอบถามได้

3.3 ผลลัพธ์จากการสอบถาม คือ กราฟย่อยที่ประกอบด้วยระเบียนข้อมูลต่างๆ โดยที่ระเบียนเหล่านั้นต้องประกอบด้วยคำสำคัญทั้งหมดที่ได้จากการสอบถาม และมีความสัมพันธ์กันตามโครงสร้างฐานข้อมูล

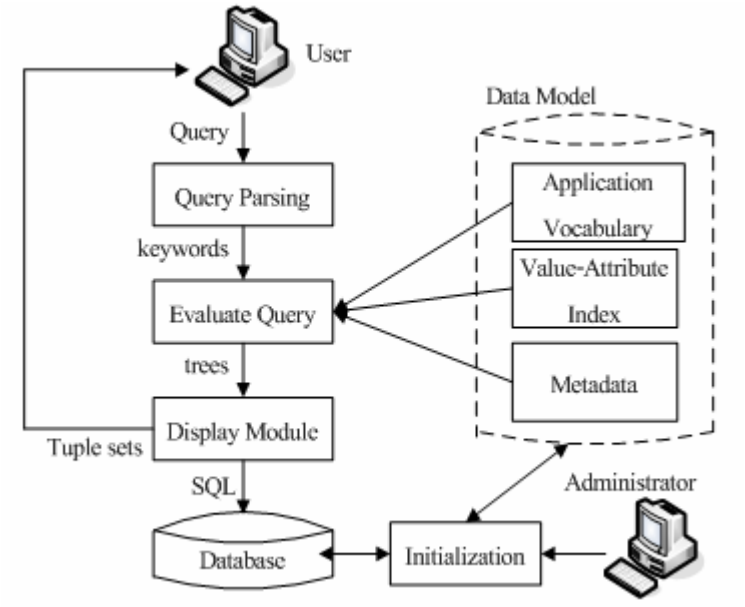
3.4 การจัดลำดับผลลัพธ์ของระบบบีเอเอ็นเคเอสพิจารณาจากโครงสร้างของกราฟคำตอบและความหมายของเส้นเชื่อม

2.3.3 มราจียาติ

มราจียาติ [10] เป็นระบบการค้นหาคำสำคัญในฐานข้อมูลที่ต่างจากวิธีการดังกล่าวข้างต้นทั้งในเรื่องของแบบจำลองข้อมูล ขั้นตอนในการหาผลลัพธ์ และการจัดลำดับความสำคัญของผลลัพธ์ ซึ่งมีรายละเอียดดังต่อไปนี้

1. สถาปัตยกรรมของระบบ

สถาปัตยกรรมของระบบมราจียาติเป็นดังรูปที่ 2.11 ซึ่งประกอบด้วยส่วนสำคัญ 2 ส่วน คือ ส่วนการเตรียมข้อมูลเริ่มต้น และส่วนการสอบถาม ในส่วนการเตรียมข้อมูลเริ่มต้น ผู้ดูแลระบบเป็นผู้จัดการแบบจำลองข้อมูล และในส่วนการสอบถามประกอบด้วยการทำงาน 3 ขั้นตอน คือ การวิเคราะห์การสอบถาม การหาผลลัพธ์ และการแสดงผลลัพธ์



รูปที่ 2.11 สถาปัตยกรรมของระบบมราจียาติ

ในส่วนการวิเคราะห์การสอบถาม ระบบเริ่มค้นหาคำสำคัญจากคำแอบพลิเคชัน และดัชนีค่าข้อมูล-แอทริบิว ตามลำดับ หากไม่พบคำสำคัญจากแบบจำลองข้อมูลดังกล่าว ระบบจึงค้นหาคำสำคัญจากข้อมูลฐานข้อมูลต่อไป

2. แบบจำลองข้อมูลและดัชนีผกผัน

มราจียาติเก็บแบบจำลองข้อมูลในรูปของตารางโครงสร้างฐานข้อมูล โดยแบ่งข้อมูลเป็น 3 ส่วน คือ

- คำแอบพลิเคชัน คือ คำที่มีความหมายสื่อถึงชื่อเอนทิตี ชื่อแอทริบิว และค่าข้อมูลที่อยู่ในรูปของรหัส หรือคำย่อ เช่น ค่าข้อมูล ‘F’ แทนด้วย ‘female’ หรือ ค่าข้อมูล ‘CS’ แทนด้วย ‘computer science’ เป็นต้น
- ดัชนีค่าข้อมูล-แอทริบิว คือ ดัชนีที่ใช้เก็บค่าข้อมูล และตำแหน่งของค่าข้อมูลในฐานข้อมูล ซึ่งจะอยู่ในรูปของ <ค่าข้อมูล, แอทริบิว, เอนทิตี>
- ข้อมูลฐานข้อมูล ประกอบด้วยตาราง 4 ตาราง คือ รายการเอนทิตี รายการแอทริบิว รายการคีย์หลัก และรายการคีย์นอก

3. ขั้นตอนวิธีในการหาผลลัพธ์

ข้อแตกต่างระหว่างระบบมราจียาติและวิธีที่ผ่านมา คือ ผู้ใช้สามารถระบุคำสำคัญที่ตรงกับข้อมูลฐานข้อมูลได้ ซึ่งระบบจะใช้คำสำคัญนี้ในการเจาะจงระเบียนผลลัพธ์เป็นแอทริบิวใดแอทริบิวหนึ่งหรือเอนทิตีใดเอนทิตีหนึ่งได้ ตัวอย่างการสอบถามของผู้ใช้จากตัวอย่างข้อมูลในตารางที่ 2.2 และ 2.3

Q1 : “John” หมายถึง ผู้ใช้ต้องการข้อมูลเกี่ยวกับ John

Q2 : “John Sport” หมายถึง ผู้ใช้ต้องการข้อมูล Sport ของ John

Q3 : “John Activity” หมายถึง ผู้ใช้ต้องการข้อมูล Activity ของ John

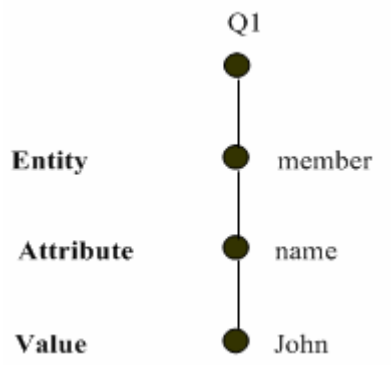
ตารางที่ 2.2 แสดงตัวอย่างข้อมูลในเอนทิตี Member

Name	City	Age
John	New York	45
...	...	...

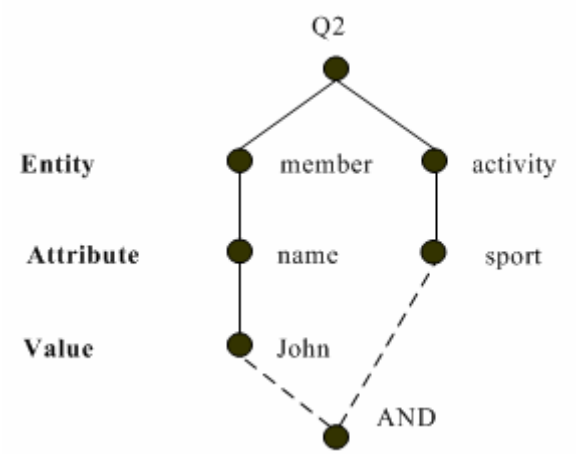
ตารางที่ 2.3 แสดงตัวอย่างข้อมูลในเอนทิตี Activity

Name	Sport
John	Running
John	Swimming
...	...

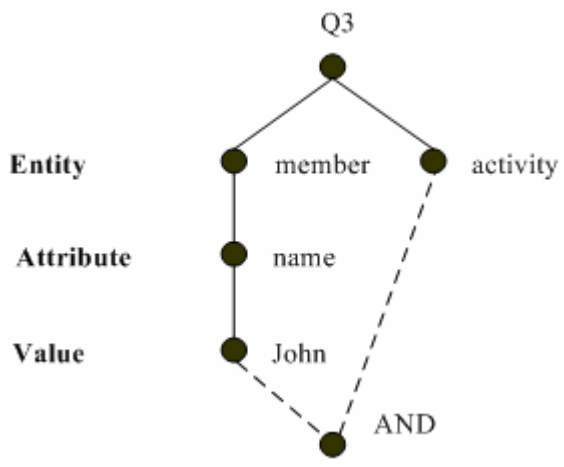
วิธีการหาผลลัพธ์ของมราจียาติแตกต่างจากวิธีการที่ผ่านมาเช่นกัน โดยระบบนำตำแหน่งที่ได้จากดัชนีมาสร้างเป็นทรีที่สามารถเป็นไปได้ทั้งหมด ทรีที่ได้มีความสูง 4 ระดับ คือ ระดับรากของทรี ระดับเอนทิตี ระดับแอทริบิว และระดับค่าข้อมูล โดยการสร้างทรีจะเริ่มจากระดับล่างสุด คือ ระดับค่าข้อมูล ระดับแอทริบิว ระดับเอนทิตี และระดับรากของทรี ตามลำดับ แล้วนำทรีที่ได้มาแปลงให้อยู่ในรูปของภาษาเอสคิวแอล ตัวอย่างทรีที่ได้จากการสอบถามด้วย Q1 Q2 และ Q3 เป็นดังรูปที่ 2.12 2.13 และ 2.14 ตามลำดับ



รูปที่ 2.12 ทรีที่ได้จากการสอบถามด้วย “John”



รูปที่ 2.13 ทรีที่ได้จากการสอบถามด้วย “John sport”



รูปที่ 2.14 ทรีที่ได้จากการสอบถามด้วย “John activity”

4. การวิเคราะห์

การวิเคราะห์ระบบตามหลักการทั่วไปของการค้นหาคำสำคัญในฐานข้อมูล มีดังต่อไปนี้

4.1 สถาปัตยกรรมของระบบ แบ่งเป็น 2 ส่วน คือ ส่วนการเตรียมข้อมูล และส่วนการสอบถาม ในส่วนการสอบถาม วิธีการหาผลลัพธ์โดยการสร้างทรีดังกล่าวทำให้ระบบไม่จำเป็นต้องทำงานกับกราฟจำนวนมาก แต่ทำให้ไม่สามารถค้นหาผลลัพธ์ที่ประกอบด้วยเอนทิตีมากกว่า 2 เอนทิตีได้

4.2 แบบจำลองข้อมูลและดัชนีผกผัน ระบบมราจียาติเก็บดัชนีผกผันร่วมกับแบบจำลองข้อมูล กล่าวคือ แบบจำลองข้อมูลของระบบประกอบด้วยข้อมูล 3 ส่วน คือ คำแอบพลิเคชัน ดัชนีค่าข้อมูล-แอทริบิว และข้อมูลฐานข้อมูล ซึ่งข้อมูลทั้ง 3 ส่วนนี้เก็บอยู่ในรูปของตารางที่แยกจากกัน การค้นหาคำสำคัญจึงต้องทำการค้นหาทีละส่วน ข้อดีของการเก็บแบบจำลองดังกล่าว คือ ผู้ใช้สามารถค้นหาคำสำคัญที่ตรงกับข้อมูลฐานข้อมูล และคำแอบพลิเคชันดังกล่าวได้ ส่วนข้อเสีย คือ ในกรณีที่คำสำคัญตรงกับค่าข้อมูล และข้อมูลฐานข้อมูล ระบบทำการสร้างทรีเฉพาะคำสำคัญที่ตรงกับค่าข้อมูลเท่านั้น ดังนั้นผลลัพธ์ที่ได้ คือ ทรีที่ประกอบด้วยคำสำคัญที่ตรงกับค่าข้อมูลเท่านั้น ส่วนทรีที่ประกอบด้วยคำสำคัญตรงกับข้อมูลฐานข้อมูลจะไม่ถูกสร้าง

4.3 ภาษาสอบถาม ระบบกำหนดให้การสอบถามเป็นการระบุเซตของคำสำคัญ หรือเซตของคำสำคัญแบบเดี่ยว

4.4 ผลลัพธ์จากการสอบถาม คือ ทรีที่ประกอบด้วยคำสำคัญ ซึ่งจำนวนเอนทิตีของทรีต้องไม่เกิน 2 เอนทิตี ทำให้การสอบถามบางการสอบถามไม่สามารถหาผลลัพธ์ได้

4.5 การจัดลำดับผลลัพธ์ของมราจียาติทำได้ 2 วิธี คือ การจัดลำดับผลลัพธ์ตามการจัดเรียงของแอทริบิวใดแอทริบิวหนึ่ง ซึ่งผู้ใช้เป็นผู้กำหนด และการจัดลำดับผลลัพธ์ตามจำนวนของอิน

ดีกรีของโหนดราก ซึ่งการจัดลำดับผลลัพธ์ทั้ง 2 วิธีดังกล่าวไม่ได้จัดเรียงตามความสำคัญของผลลัพธ์

2.4 การวิเคราะห์ระบบการค้นหาคำสำคัญในฐานข้อมูล

โดยทั่วไป การค้นหาคำสำคัญในฐานข้อมูลนำดัชนีผกผันมาใช้ในการเก็บคำสำคัญ ส่วนแบบจำลองข้อมูลถูกเก็บในรูปของกราฟโครงสร้างฐานข้อมูล หรือกราฟข้อมูล การหาผลลัพธ์จึงอาศัยความสัมพันธ์ระหว่างโหนดในกราฟ ซึ่งในการสอบถามแต่ละครั้งอาจมีจำนวนผลลัพธ์มากกว่า 1 ผลลัพธ์ ดังนั้นในการพิจารณาความสำคัญของผลลัพธ์แต่ละผลลัพธ์ จึงอาศัยแนวความคิดความใกล้กันของโหนดในกราฟ หรือระยะทางที่สั้นที่สุด โดยถือว่ากราฟที่มีระยะทางในการเชื่อมต่อกันของโหนดสั้นที่สุดคือผลลัพธ์ที่ดีที่สุด

ข้อจำกัดของวิธีการค้นหาคำสำคัญจากฐานข้อมูลมีดังต่อไปนี้

1. คำสำคัญที่ใช้ค้นหาต้องเป็นคำที่มีอยู่ในค่าข้อมูลหรือเป็นส่วนหนึ่งของค่าข้อมูลเท่านั้น ไม่สามารถค้นหาคำพ้องความหมายของคำสำคัญได้ เช่น ในฐานข้อมูลมีการเก็บค่าข้อมูล “CA” แทนคำว่า “California” เมื่อผู้ใช้ค้นหาด้วยคำว่า “California” ระบบไม่สามารถค้นหาคำสำคัญดังกล่าวนี้ได้ ดังนั้นการเพิ่มความสามารถในการค้นหาด้วยคำพ้องคำสำคัญช่วยให้ผู้ใช้ค้นหาข้อมูลได้แม่นยำยิ่งขึ้น

2. ระบบการค้นหาคำสำคัญในฐานข้อมูลส่วนใหญ่ไม่สามารถใช้คำสำคัญที่ตรงกับข้อมูลของฐานข้อมูลเพื่อช่วยในการสอบถามได้ เช่น การสอบถามด้วย “simon” ในตัวอย่างจากหัวข้อ 2.3.1 ระบบต้องพิจารณาระบียนข้อมูลทุกระเบียนที่มีคำว่า “simon” แต่ถ้าผู้ใช้เจาะจงการสอบถามด้วย “name simon” ระบบจะพิจารณาเฉพาะระเบียนข้อมูลในแอทริบิว Name เท่านั้น การสอบถามด้วยข้อมูลฐานข้อมูลดังกล่าวจึงช่วยในการจำกัดขอบเขตของค่าข้อมูล และสามารถจำกัดผลลัพธ์โดยการระบุเอนทิตีหรือแอทริบิวได้ ทำให้การค้นหาคำสำคัญในฐานข้อมูลให้ผลลัพธ์ที่แม่นยำยิ่งขึ้น

3. ในบางระบบ อนุญาตให้ผู้ใช้สอบถามด้วยคำสำคัญที่ตรงกับข้อมูลฐานข้อมูลได้ แต่ยังคงมีการจำกัดรูปแบบของการสอบถาม เช่น การใช้เครื่องหมาย “=” ใน [13] ดังนั้นการสอบถามด้วยข้อความอิสระแต่ยังคงสอบถามด้วยคำสำคัญที่ตรงกับข้อมูลฐานข้อมูลได้ จึงน่าจะช่วยเพิ่มความสะดวกให้กับผู้ใช้มากยิ่งขึ้น

4. ระบบการค้นหาคำสำคัญส่วนใหญ่กำหนดให้ใช้คำดัชนีที่มีชนิดข้อมูลเป็นอักขระเท่านั้น ไม่สามารถใช้คำดัชนีที่เป็นจำนวนหรือวันที่ได้ เช่น จากตัวอย่างข้อมูลในรูปที่ 2.7 คำดัชนีคือ คำที่มาจากแอทริบิว NAME และ ADDRESS เท่านั้น ไม่สามารถนำคำจากแอทริบิวอื่นที่มีชนิดข้อมูลเป็นจำนวนมาสร้างเป็นคำดัชนีได้

5. ในกรณีที่ผู้ใช้สอบถามข้อมูลโดยใช้คำสำคัญชุดเดียวกัน ระบบส่วนใหญ่ต้องเริ่มกระบวนการค้นหาใหม่ การจัดการเรื่องเคสเบสเลินนิง (Case-Based Learning) และเคสเบสรีซันนิง (Case-Based Reasoning) จะช่วยยืนยันความถูกต้องของผลลัพธ์และเพิ่มประสิทธิภาพในการสอบถามได้