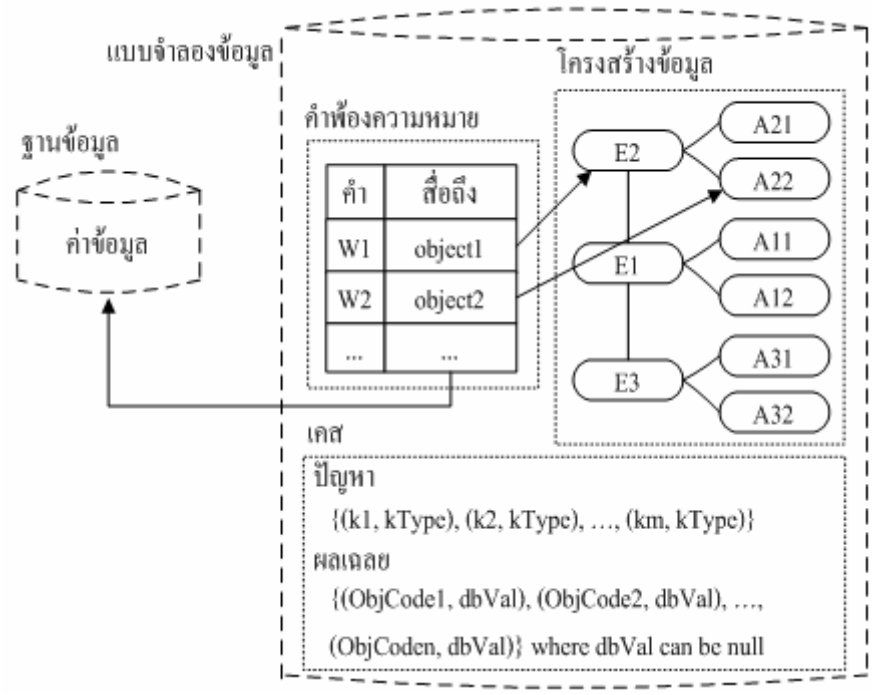
บทที่ 3
การค้นหาคำสำคัญในฐานข้อมูลโดยใช้การค้นข้อมูลฐานข้อมูล

บทนี้เป็นการนำเสนอระบบการค้นหาคำสำคัญในฐานข้อมูล ซึ่งผู้ใช้สามารถค้นหาด้วยคำสำคัญที่ปรากฏในค่าข้อมูล หรือข้อมูลฐานข้อมูล (ชื่อเอนทิตี หรือชื่อแอทริบิว) และสามารถค้นหาจากคำพ้องความหมายได้ ผลลัพธ์ที่ได้จากการค้นหา คือ ระเบียนข้อมูลที่ประกอบด้วยคำสำคัญและ/หรือชื่อเอนทิตีหรือชื่อแอทริบิวตรงกับคำสำคัญตามที่ผู้ใช้ต้องการ โดยระเบียนเหล่านั้นมีความสัมพันธ์กันตามโครงสร้างฐานข้อมูล

ในส่วนที่ 3.1 เป็นการนำเสนอแบบจำลองข้อมูล ส่วนที่ 3.2 นำเสนอสถาปัตยกรรมของระบบ ส่วนที่ 3.3 กล่าวถึงอัลกอริทึมที่ใช้ในการค้นหาผลลัพธ์ และส่วนสุดท้าย กล่าวถึงการจัดการเคส

3.1 แบบจำลองข้อมูล

แบบจำลองข้อมูลจัดเก็บข้อมูลที่จำเป็นสำหรับการค้นหาผลลัพธ์ แบ่งออกเป็น 3 ส่วน คือ โครงสร้างฐานข้อมูล เคส และคำพ้องความหมาย ดังรูปที่ 3.1 ซึ่งมีรายละเอียด ดังต่อไปนี้



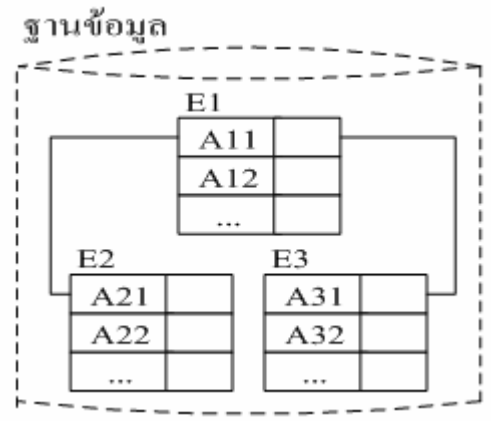
รูปที่ 3.1 แบบจำลองข้อมูล

3.1.1 โครงสร้างฐานข้อมูล

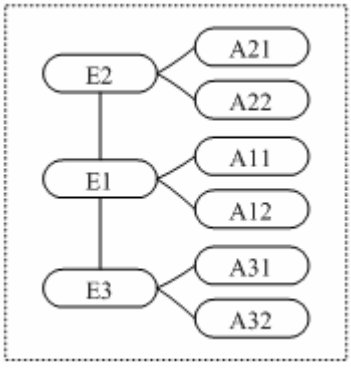
โครงสร้างฐานข้อมูลถูกจัดเก็บในรูปของกราฟแบบไม่มีทิศทางที่กำหนดให้โหนดใช้แทนเอนทิตีและแอทริบิว ส่วนเส้นเชื่อมใช้แทนความสัมพันธ์ระหว่างเอนทิตีกับเอนทิตี และความสัมพันธ์ระหว่างเอนทิตีกับแอทริบิว ซึ่งเส้นเชื่อมระหว่างเอนทิตีกับเอนทิตีอาศัยความสัมพันธ์ระหว่างคีย์หลักและคีย์นอก กราฟโครงสร้างข้อมูลมีนิยาม ดังนี้

นิยามที่ 1 กราฟโครงสร้างข้อมูล SG (Schema Graph) คือ กราฟ <V, E> โดยที่ V คือเซตของจุด 2 ประเภท คือ เอนทิตีและแอทริบิว ส่วน E คือ เซตของข้อบังคับในการเชื่อมต่อระหว่างจุด ซึ่งได้แก่ การเชื่อมต่อระหว่างแอทริบิวกับเอนทิตี (E^{ae}) และเอนทิตีกับเอนทิตี (E^{ee})

ตัวอย่างกราฟโครงสร้างฐานข้อมูล สมมติให้ฐานข้อมูลประกอบด้วยเอนทิตี 3 เอนทิตี และมีความสัมพันธ์ระหว่างคีย์หลักและคีย์นอกเป็นดังรูปที่ 3.2 กราฟโครงสร้างฐานข้อมูลเป็นดังรูปที่ 3.3

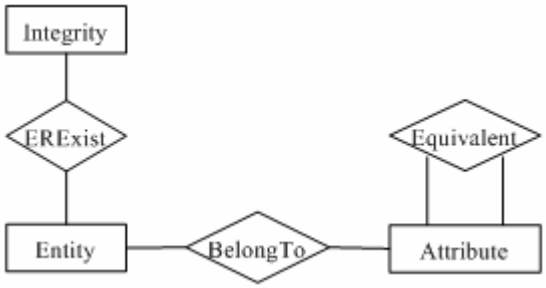


รูปที่ 3.2 ตัวอย่างโครงสร้างฐานข้อมูล



รูปที่ 3.3 ตัวอย่างกราฟโครงสร้างฐานข้อมูล

การจัดเก็บกราฟโครงสร้างข้อมูลดังตัวอย่างในรูปที่ 3.3 สามารถแสดงความสัมพันธ์ระหว่างข้อมูลฐานข้อมูลได้ดังรูปที่ 3.4 ซึ่งประกอบด้วยเอนทิตี Entity สำหรับเก็บข้อมูลเอนทิตีต่างๆ ในฐานข้อมูล เอนทิตี Attribute สำหรับเก็บข้อมูลแอทริบิว เอนทิตี Equivalent สำหรับเก็บแอทริบิวที่เท่าเทียมกัน และเอนทิตี Integrity เก็บข้อมูลความสัมพันธ์ระหว่างเอนทิตี



รูปที่ 3.4 แผนภาพแสดงความสัมพันธ์ระหว่างข้อมูลฐานข้อมูล

3.1.2 เคส

ระบบนำเคสในเคส-เบสรีซันนิง [4] มาใช้ในการพิจารณาหาผลลัพธ์ที่เหมาะสม โดยกำหนดให้เคสประกอบด้วยปัญหา (Problem) และผลเฉลย (Solution) ที่ได้จากการสอบถามหนึ่งๆ ในส่วนของปัญหาคือเซตของคำสำคัญที่ได้จากการสอบถาม และผลเฉลยสร้างจากกราฟคำตอบ โดยที่

$$Case = (Problem, Solution)$$

ซึ่ง Problem คือ เซตของจุดที่ค้นพบในกราฟคำตอบ และ Solution คือ กราฟคำตอบ โดยที่

$$Problem = \{(v_1, kType), (v_2, kType), \ldots, (v_m, kType)\}$$

เมื่อ v_i แทน จุดที่ค้นพบคำสำคัญ i โดยที่ $v_i \in$ กราฟคำตอบ

kType แทน ประเภทของคำสำคัญ โดย kType $\in$ {A: Attribute, E: Entity, V: Value}

m แทน จำนวนคำสำคัญที่ได้จากการสอบถาม

$$Solution = (V^{ae}, E^{ee})$$

เมื่อ V^{ae} คือ เซตของคู่ลำดับระหว่างจุดแอทริบิวหรือจุดเอนทิตี และจุดค่าข้อมูล กล่าวคือ $V^{ae} = \{(ObjCode_1, dbVal_1), (ObjCode_2, dbVal_2), \ldots, (ObjCode_n, dbVal_n)\}$ โดยที่ n แทน จำนวนจุดแอทริบิวและจุดเอนทิตีที่ไม่ปรากฏจุดแอทริบิวของผลลัพธ์ในเคส และ $ObjCode_i$ แทนจุดแอทริบิว หรือจุดเอนทิตีของกราฟคำตอบ

E^{ee} คือ เซตของเส้นเชื่อมระหว่างเอนทิตีกับเอนทิตีในกราฟคำตอบ

3.1.3 คำพ้องความหมาย

คำพ้องความหมาย คือ คำย่อ คำหรือวลีที่ใช้แทนหรือมีความหมายสื่อถึงเอนทิตี แอทริบิว หรือค่าข้อมูล เช่น คำพ้องความหมาย “California” ใช้แทนค่าข้อมูล “CA” คำพ้องความหมาย “Writer” ใช้แทนเอนทิตี “Author” เป็นต้น คำพ้องความหมายมีรูปแบบการสื่อถึงเอนทิตี แอทริบิว หรือค่าข้อมูล ดังต่อไปนี้

คำพ้องความหมาย s สื่อถึง (sType, Object)

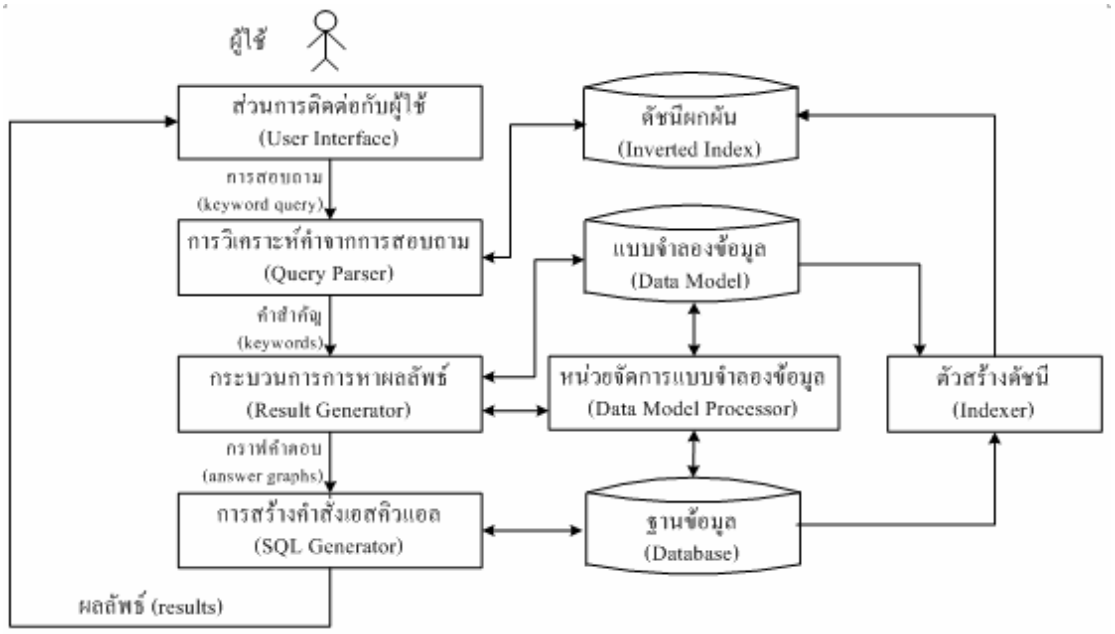
เมื่อ sType คือ ประเภทของคำพ้องความหมาย โดย sType $\in$ {A: Attribute, E: Entity, V: Value}

Object แทน คำที่คำพ้องความหมายสื่อถึง

นิยามที่ 2 เซตของคำพ้องความหมาย S_{i^t} คือ เซตของคำที่มีความหมายสื่อถึงคำดัชนี i^t เมื่อ t คือ ประเภทของคำดัชนี โดยที่ t = A E หรือ V ซึ่งใช้แทนแอทริบิว เอนทิตี และค่าข้อมูลตามลำดับ กล่าวคือ สมาชิกของ S_{i^t} คือ $\{s_1, s_2, \ldots, s_j\}$ เมื่อ j คือ จำนวนคำพ้องความหมายที่สื่อถึง i^t และแต่ละ s_j สื่อถึง i^t

3.2 สถาปัตยกรรมของระบบ

สถาปัตยกรรมของระบบการค้นหาคำสำคัญในฐานข้อมูลโดยใช้การค้นข้อมูลฐานข้อมูลเป็นดังรูปที่ 3.5 ส่วนประกอบที่สำคัญของระบบประกอบด้วยส่วนสำคัญ 5 ส่วน คือ หน่วยจัดการแบบจำลองข้อมูล (Data Model Processor) ตัวสร้างดัชนี (Indexer) ตัววิเคราะห์คำจากการสอบถาม (Query Parser) กระบวนการหาผลลัพธ์ (Result Generator) และตัวสร้างคำสั่งเอสคิวแอล (SQL Generator) ในแต่ละส่วนมีรายละเอียดดังต่อไปนี้



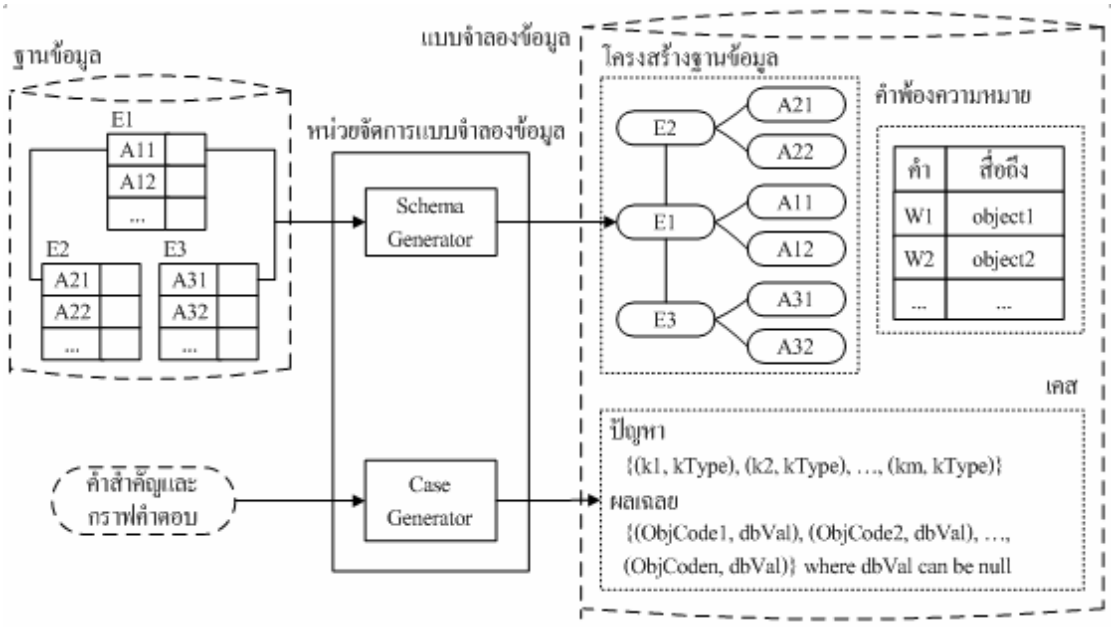
รูปที่ 3.5 สถาปัตยกรรมของระบบการค้นหาคำสำคัญในฐานข้อมูลโดยใช้การค้นข้อมูลฐานข้อมูล

3.2.1 หน่วยจัดการแบบจำลองข้อมูล

หน่วยจัดการแบบจำลองข้อมูลมีหน้าที่จัดการข้อมูล 2 ส่วน คือ ข้อมูลโครงสร้างฐานข้อมูล และข้อมูลเกี่ยวกับเคส ดังรูปที่ 3.6

1. ตัวสร้างโครงสร้างฐานข้อมูล (Schema Generator) มีหน้าที่ทำการวิเคราะห์ฐานข้อมูล เพื่อนำข้อมูลฐานข้อมูลซึ่งได้แก่ ชื่อเอนทิตี ชื่อแอทริบิว ความสัมพันธ์ระหว่างเอนทิตีกับเอนทิตี และความสัมพันธ์ระหว่างเอนทิตีกับแอทริบิวสร้างเป็นแบบจำลองโครงสร้างฐานข้อมูลที่อยู่ในรูปของกราฟแบบไม่มีทิศทาง โดยกำหนดให้โหนดใช้แทนเอนทิตีและแอทริบิว ส่วนด้านใช้แทนความสัมพันธ์ระหว่างเอนทิตีกับเอนทิตี และความสัมพันธ์ระหว่างเอนทิตีกับแอทริบิว

2. ตัวสร้างเคส (Case Generator) มีหน้าที่สร้างเคสจากการสอบถามหนึ่งๆ ซึ่งประกอบด้วยปัญหาและผลเฉลย ในส่วนของปัญหาสร้างจากจุดที่ค้นพบของคำสำคัญที่ได้จากกราฟคำตอบ และผลเฉลยสร้างจากจุดแอทริบิวและจุดเอนทิตีที่ไม่ปรากฏจุดแอทริบิวในกราฟคำตอบ



รูปที่ 3.6 หน่วยจัดการแบบจำลองข้อมูล

3.2.2 ตัวสร้างดัชนี

ตัวสร้างดัชนีทำการสร้างดัชนีผกผันจากแบบจำลองข้อมูลและค่าข้อมูลจากฐานข้อมูล ในดัชนีผกผัน ประกอบด้วย คำดัชนี และรายการอ้างอิง ดังตัวอย่างจากตารางที่ 3.1 ซึ่งมีรายละเอียดดังต่อไปนี้

1. คำดัชนี คือ คำหรือวลี 4 ประเภท คือ ค่าข้อมูลจากฐานข้อมูล ข้อมูลจากแบบจำลองข้อมูล และคำที่ไม่มีนัยสำคัญ

1.1 ค่าข้อมูลในฐานข้อมูล คือ ค่าข้อมูลทั้งหมดจากฐานข้อมูล ทั้งข้อมูลประเภทอักขระ จำนวน และวันที่ โดยที่ค่าข้อมูลจากแอทริบิวเดียวกันจะไม่ซ้ำกัน ซึ่งตัวสร้างดัชนีจะทำการสร้างคำดัชนีประเภทนี้อัตโนมัติ

1.2 ข้อมูลจากแบบจำลองข้อมูล ได้แก่ ข้อมูลจากโครงสร้างฐานข้อมูล ข้อมูลจากเคส และข้อมูลจากคำพ้องความหมาย คำดัชนีที่มาจากโครงสร้างฐานข้อมูลประกอบด้วยชื่อเอนทิตีและชื่อแอทริบิว คำดัชนีที่มาจากเคสประกอบด้วยคำสำคัญจากปัญหาในแต่ละเคส ซึ่งตัวสร้างดัชนีจะทำการสร้างคำดัชนีจากแบบจำลองข้อมูลโดยอัตโนมัติ

1.3 คำที่ไม่มีนัยสำคัญ คือ คำที่ไม่ปรากฏอยู่ในคำดัชนีทั้ง 3 ประเภทดังกล่าว ซึ่งถือว่าเป็นคำที่ไม่มีนัยสำคัญในการค้นหา เนื่องจากผู้ใช้สามารถใช้ข้อความอิสระในการสอบถามได้ จึงอาจมีคำที่ไม่มีนัยสำคัญปรากฏอยู่ เช่น “give me detail of smith” คำที่ไม่มีนัยสำคัญ คือ “give” “me” “detail” และ “of” คำที่ไม่มีนัยสำคัญนี้ช่วยเพิ่มประสิทธิภาพในกระบวนการวิเคราะห์คำจากการสอบถามได้ ซึ่งคำดัชนีประเภทนี้ผู้ดูแลระบบจะเป็นผู้กำหนด

2. รายการอ้างอิง คือ การอ้างถึงตำแหน่งที่อยู่ของคำดัชนี สามารถอ้างอิงได้ดังนี้

รายการอ้างอิง = {(ประเภทของคำดัชนี, รหัสอ็อบเจกต์)}

โดยที่ ประเภทของคำดัชนี (Index Type) ประกอบด้วย A C E N S และ V

A (Attribute) หมายถึง คำดัชนีที่อ้างไปยังแอทริบิว

C (Case) หมายถึง คำดัชนีที่ปรากฏอยู่ในปัญหาของเคส

E (Entity) หมายถึง คำดัชนีที่อ้างไปยังเอนทิตี

N (Not use) หมายถึง คำดัชนีที่ไม่มีนัยสำคัญ

S (Synonym) หมายถึง คำดัชนีที่เป็นคำพ้องความหมาย

V (Value) หมายถึง คำดัชนีที่อ้างไปยังค่าข้อมูลในฐานข้อมูล

รหัสอ็อบเจกต์ (Object Code) ถูกกำหนดตามประเภทของคำดัชนี ดังนี้

A : กำหนดให้ใช้รหัสของแอทริบิว

C : กำหนดให้ใช้รหัสของเคส

E : กำหนดให้ใช้รหัสของเอนทิตี

N : กำหนดให้เป็นค่าว่าง

S : กำหนดให้ใช้รหัสของคำพ้องความหมาย

V : กำหนดให้ใช้รหัสของแอทริบิว

ตารางที่ 3.1 แสดงตัวอย่างดัชนีผกผัน

คำดัชนี	**รายการอ้างอิง**
w_1	$\{(iType_{11}, objCode_{11}), (iType_{12}, objCode_{12}), \ldots, (iType_{1n}, objCode_{1n})\}$
w_2	$\{(iType_{21}, objCode_{21}), (iType_{22}, objCode_{22}), \ldots, (iType_{2n}, objCode_{2n})\}$
…	
w_m	$\{(iType_{m1}, objCode_{m1}), (iType_{m2}, objCode_{m2}), \ldots, (iType_{mn}, objCode_{mn})\}$

3.2.3 การวิเคราะห์คำจากการสอบถาม

การวิเคราะห์คำจากการสอบถามเป็นกระบวนการวิเคราะห์หาคำสำคัญและตำแหน่งของคำสำคัญในแบบจำลองข้อมูล โดยถือว่าคำหรือวลีจากการสอบถามที่สามารถจับคู่กับคำดัชนีในดัชนีผกผันได้คือคำสำคัญ ดังนิยามต่อไปนี้

นิยามที่ 3 กำหนดให้การสอบถาม Q (Query) เป็นการสอบถามด้วยข้อความอิสระ ที่มีการแบ่งคำด้วยอักขระว่าง โดยที่

$Q = w_1, w_2, \ldots, w_m$ เมื่อ w คือ คำจากการสอบถาม และ m คือ จำนวนคำจากการสอบถาม

นิยามที่ 4 คำสำคัญ K (Keyword) ของ Q คือ ส่วนของ Q ที่สามารถจับคู่กับคำดัชนีจากดัชนีผกผันได้ โดยที่

$K = \{ k_1, k_2, \ldots, k_n \}$ เมื่อ n คือ จำนวนคำสำคัญที่ค้นพบ โดยที่ $n \leq m$

นิยามที่ 5 แต่ละ k_i เมื่อ i = 1, 2, …, n ประกอบด้วยรายการอ้างอิง L_{k_i} (Reference List) ดังที่กล่าวไว้ในส่วนที่ 3.2.2

3.2.4 กระบวนการหาผลลัพธ์

การค้นหาคำสำคัญในฐานข้อมูลโดยใช้การค้นข้อมูลฐานข้อมูลอาศัยแบบจำลองข้อมูลดังที่กล่าวไว้ในส่วนที่ 3.1 เพื่อความเข้าใจในการอ้างถึงคำศัพท์ต่างๆ ในกระบวนการหาผลลัพธ์จึงได้นิยามคำศัพท์ไว้ ดังต่อไปนี้

นิยามที่ 6 กราฟเสมือน G (Virtual Graph) คือ กราฟโครงสร้างฐานข้อมูล SG ที่มีการเชื่อมต่อจุดแอทริบิวด้วยจุดค่าข้อมูล กล่าวคือ กราฟ G คือกราฟ <V, E> โดยที่ V คือเซตของจุด 3 ประเภท คือ เอนทิตี แอทริบิว และค่าข้อมูล ส่วน E คือ เซตของข้อบังคับในการเชื่อมต่อระหว่างจุด 3 ประเภท ซึ่งได้แก่ การเชื่อมต่อระหว่างเอนทิตีกับเอนทิตี (E^{ee}) ระหว่างแอทริบิวกับเอนทิตี (E^{ae}) และระหว่างแอทริบิวกับค่าข้อมูล (E^{va})

นิยามที่ 7 เซตของจุดที่ค้นพบ (Searchable Vertex Set) ของคำสำคัญ k_i หรือ V_{k_i} คือ เซตของจุดในกราฟ G ที่มีค่าสอดคล้องกับ k_i หรือมีคำพ้องความหมายสอดคล้องกับ k_i กล่าวคือ สมาชิกของ V_{k_i} คือ จุดที่ค้นพบ $v^t_{k_i}$ เมื่อ t = E A หรือ V โดยที่ t คือ ประเภทของจุดซึ่งแทนด้วย เอนทิตี แอทริบิว และค่าข้อมูลตามลำดับ

นิยามที่ 8 กำหนดให้จำนวนคำสำคัญเท่ากับ n เมื่อ $n \neq 0$ รูปจำลองการสอบถาม (Query Image) คือ เซตของจุดที่ค้นพบ v^t_i เมื่อ i = 1, 2, …, n และ $v^t_i \in v^t_{k_i}$ ตามลำดับ

นิยามที่ 9 เส้นทางพื้นฐาน P_i (Basic Path) ของ v_i คือ เซตของจุดที่น้อยที่สุดในกราฟ G ที่ประกอบด้วย v_i และจุดเอนทิตี โดยที่จุดเอนทิตีนั้นสามารถเชื่อมต่อกับ v_i นั้นได้

นิยามที่ 10 เซตของจุดระหว่างกลาง V^C (Intermediate Vertex) คือ เซตของจุดที่เชื่อมระหว่างเส้นทางพื้นฐาน P_i กับ P_j โดยที่ $i \neq j$ กล่าวคือ $v^C \in G$ และ $v^C \notin P$ และมี E^{ee} เชื่อมระหว่าง v^C กับ v^E โดยที่ $v^E \in P$

นิยามที่ 11 กราฟที่เป็นไปได้ FG (Feasible Graph) ของรูปจำลองการสอบถามที่มีจำนวนคำสำคัญที่ค้นพบเท่ากับ n เมื่อ $n \neq 0$ คือ เซตของจุดในเส้นทางพื้นฐาน $P_1 \cup P_2 \cup \ldots \cup P_n$ และจุดระหว่างกลาง โดยที่ P_i คือ เส้นทางพื้นฐานที่ถูกกำหนดโดยจุดที่ค้นพบ v_i ของรูปจำลองการสอบถาม

นิยามที่ 12 กราฟคำตอบ (Answer Graph) คือ กราฟที่ได้จากกราฟ FG โดยที่ $FG \subset G$ ซึ่ง FG ประกอบด้วย <V, E> เมื่อ V คือ เซตของจุดเอนทิตี จุดแอทริบิว และจุดค่าข้อมูล และ E คือ เซตของเส้นเชื่อมต่อระหว่างจุดดังกล่าว

ในกระบวนการหาผลลัพธ์แบ่งเป็น 2 กรณี คือ การหาผลลัพธ์โดยพิจารณาจากเคส และการหาผลลัพธ์จากการค้นข้อมูลฐานข้อมูล ในการหาผลลัพธ์โดยพิจารณาจากเคสระบบทำการเปรียบเทียบเซตของคำสำคัญที่ได้กับปัญหาของเคส หากเซตของคำสำคัญสอดคล้องกับปัญหาของเคส ผลลัพธ์ที่ได้คือผลเฉลยของเคส หรือการประยุกต์ใช้ผลเฉลยของเคสนั่นเอง ในกรณีที่เป็นการหาผลลัพธ์จากการค้นข้อมูลฐานข้อมูล ระบบทำการค้นหาเซตของจุดที่ค้นพบ (นิยามที่ 7) จากกราฟ G (นิยามที่ 6) และกำหนดรูปจำลองการสอบถาม (นิยามที่ 8) ทั้งหมด จากนั้นจึงพิจารณาหากราฟที่เป็นไปได้ (นิยามที่ 11) จากรูปจำลองการสอบถามโดยอาศัยเส้นทางพื้นฐาน (นิยามที่ 9) ซึ่งกราฟคำตอบ (นิยามที่ 12) คือ กราฟที่ได้จากกราฟที่เป็นไปได้จำนวน r กราฟ โดยเรียงลำดับความสำคัญของกราฟตามต้นทุนของกราฟ กราฟคำตอบที่ได้ถูกแปลงให้อยู่ในรูปภาษาเอสคิวแอลและนำเสนอต่อผู้ใช้ต่อไป ซึ่งหากผู้ใช้ยืนยันคำตอบที่ถูกต้องของผลลัพธ์และต้องการเก็บผลลัพธ์นั้นเป็นเคส ระบบจะนำเซตของคำสำคัญและกราฟคำตอบที่ได้เก็บให้อยู่ในรูปของเคสดังกล่าวข้างต้น

3.2.5 การสร้างคำสั่งเอสคิวแอล

จากกราฟคำตอบสามารถสร้างคำสั่งเอสคิวแอลได้ด้วยการพิจารณาความสัมพันธ์ระหว่างข้อมูลในกราฟ และเงื่อนไขที่ได้จากการสอบถาม ซึ่งได้แก่ รายการเอนทิตี (Entity List) รายการ

แอทริบิว (Attribute List) เงื่อนไขการจอย (Join Condition) และเงื่อนไขการเลือก (Selection Condition)

1. รายการเอนทิตี คือ เอนทิตีที่ปรากฏอยู่ในกราฟคำตอบทั้งหมด เป็นรายการที่บ่งบอกถึงระเบียนผลลัพธ์มาจากเอนทิตีใด

2. รายการแอทริบิว คือ แอทริบิวทุกแอทริบิวที่ปรากฏในกราฟคำตอบ และแอทริบิวจากเอนทิตีที่ไม่ปรากฏจุดแอทริบิวในกราฟคำตอบ

3. เงื่อนไขการจอย คือ การกำหนดความสัมพันธ์ระหว่างเอนทิตีโดยอาศัยความสัมพันธ์ของคีย์หลักและคีย์นอกของรายการเอนทิตี

4. เงื่อนไขการเลือก คือ การกำหนดระเบียนข้อมูล ในกรณีที่เงื่อนไขการเลือกมีหลายเงื่อนไข ซึ่งเป็นเงื่อนไขที่มีแอทริบิวต่างกัน ให้เชื่อมเงื่อนไขนั้นด้วยตัวดำเนินการ “AND” แต่ถ้าเป็นเงื่อนไขที่มีแอทริบิวเดียวกัน ให้เชื่อมเงื่อนไขนั้นด้วยตัวดำเนินการ “OR” ดังตัวอย่างจากตารางที่ 3.2

ตารางที่ 3.2 แสดงตัวอย่างการเชื่อมเงื่อนไขของเงื่อนไขการเลือก

เงื่อนไขการเลือก	การเชื่อมเงื่อนไข
{PUBLICATION.TYPE = ‘misc’, PUBLICATION.YEAR = 1999}	WHERE PUBLICATION.TYPE = ‘misc’ AND PUBLICATION.YEAR = 1999
{ PUBLICATION.TYPE = ‘misc’, PUBLICATION.TYPE = ‘article’}	WHERE PUBLICATION.TYPE = ‘misc’ OR PUBLICATION.TYPE = ‘article’

3.3 อัลกอริทึมที่ใช้ในการหาผลลัพธ์

กำหนดให้การสอบถาม Q = $w_1, w_2, \ldots, w_m$ เมื่อ w คือ คำจากการสอบถาม และ m คือ จำนวนคำจากการสอบถาม อัลกอริทึมที่ใช้ในการหาผลลัพธ์มีดังต่อไปนี้

ขั้นตอนที่ 1 วิเคราะห์คำจากการสอบถาม

1.1 หาเซตของคำสำคัญ K จาก Q โดยการค้นหาจากดัชนีผกผัน โดยที่คำสำคัญอาจเป็นคำ หรือวลีจาก Q ที่มีค่าตรงกับคำดัชนีหรือส่วนหนึ่งของคำดัชนี ดังนั้น

$K = \{ k_1, k_2, \ldots, k_n \}$ เมื่อ n คือ จำนวนคำสำคัญที่ค้นพบ โดยที่ $n \leq m$

1.2 หาเซตของจุดที่ค้นพบ V_{k_i} ของคำสำคัญ k_i เมื่อ i = 1, 2, …, n ถ้า $k_i \in$ Synonym แล้ว $V_{k_i} \in G$ ดังนั้นเซตทั้งหมดของจุดที่ค้นพบ คือ $V_K = \{ V_{k_1}, V_{k_2}, \ldots, V_{k_n} \}$

ขั้นตอนที่ 2 กำหนดรูปจำลองการสอบถาม

ทำการกำหนดรูปจำลองการสอบถาม QI โดย QI_j คือ เซตของจุดที่ค้นพบในแต่ละ k_i กล่าวคือ $QI_j = \{ v_{k_1,j} \in V_{k_1}, v_{k_2,j} \in V_{k_2}, \ldots, v_{k_i,j} \in V_{k_i} \}$ เมื่อ i = 1, 2, …, n และ j = 1, 2, …, m โดยที่ n คือ จำนวนคำสำคัญ และ m คือ จำนวนรูปจำลองการสอบถาม ซึ่ง $m = |V_{k_1}| |V_{k_2}| \ldots |V_{k_i}|$

ขั้นตอนที่ 3 หาเส้นทางพื้นฐาน

หาเส้นทางพื้นฐาน $P_{i,j}$ ของ $v_{i,j}$ ในแต่ละรูปจำลองการสอบถาม เมื่อ i = 1, 2, …, n และ j = 1, 2, …, m ดังเงื่อนไขต่อไปนี้

3.1 ถ้า $v_{i,j} = v^D_{i,j}$ แล้ว $P_{i,j} = \{(v^E, v^A, v^D_{i,j}) | \ v^E, v^A \in G$ และมี $E^{v^D_{i,j} v^A}$, $E^{v^A v^E} \in G\}$

3.2 ถ้า $v_{i,j} = v^A_{i,j}$ แล้ว $P_{i,j} = \{(v^E, v^A) | \ v^E \in G$ และมี $E^{v^A_{i,j} v^E} \in G\}$

3.3 ถ้า $v_{i,j} = v^E_{i,j}$ แล้ว $P_{i,j} = \{ v^E_{i,j} \}$

ขั้นตอนที่ 4 หากราฟคำตอบ

4.1 พิจารณาหากราฟที่เป็นไปได้ FG_j จากแต่ละ QI_j ซึ่ง FG_j เป็นกราฟย่อยของกราฟ G โดยที่ FG_j ประกอบด้วยเซตของจุดใน $P_{i,j}$ เมื่อ i = 1, 2, …, n และ v^C ซึ่ง $v^C \in G$ และ v^C เป็นจุดระหว่างกลาง

4.2 พิจารณากราฟคำตอบ โดยการเรียงลำดับกราฟคำตอบจากต้นทุนของกราฟ กล่าวคือ $\text{cost}(FG_j) = (\text{number of vertices of } FG_j) - 1$ กราฟที่มีต้นทุนต่ำที่สุดคือกราฟคำตอบที่มีนัยสำคัญมากที่สุด

ขั้นตอนที่ 5 สร้างคำสั่งเอสคิวแอล

SELECT [Attribute List]

FROM [Entity List]

WHERE [Join Condition] [AND] [Selection Condition]

3.4 การจัดการเคส

การจัดการเคส ประกอบด้วย กระบวนการในการสร้างเคส กระบวนการพิจารณาเคสที่เหมาะสมกับการสอบถามเพื่อนำมารียูส และกระบวนการรียูสเคส

3.4.1 กระบวนการในการสร้างเคส

เมื่อผู้ใช้ยืนยันความถูกต้องของผลลัพธ์ และต้องการเก็บเคสนั้นเพื่อการค้นหาครั้งต่อไป ระบบ หรือตัวสร้างเคสจะทำการสร้างเคสจากขั้นตอน ดังต่อไปนี้

ขั้นตอนที่ 1 กำหนดปัญหา

ให้ Problem = $\{(v_{k_1}, k_1\text{Type}), (v_{k_2}, k_2\text{Type}), \ldots, (v_{k_n}, k_n\text{Type})\}$ เมื่อ v_{k_i} คือ เซตของจุดที่ค้นพบคำสำคัญ k_i และ $v_{k_i} \in FG$ โดยที่ n คือจำนวนคำสำคัญ และ $k_i\text{Type} \in \{A, E, V\}$

ขั้นตอนที่ 2 กำหนดผลเฉลย

2.1 กำหนดจุดของกราฟคำตอบให้เป็นจุดผลเฉลย โดยที่

สำหรับแต่ละ v_i^t โดยที่ $v_i^t \in FG$ และ $t \in \{A, E\}$

ถ้า v_i^t โดยที่ $t = A$ แล้ว V^{ae} += $\{(v_i^A, v_j^D)\}$ เมื่อ $v_i^D \in FG$ และ v_i^D มีเส้นเชื่อมไปยัง v_i^A หรือ V^{ae} += $\{(v_i^A, -)\}$

มิฉะนั้น ถ้า v_i^t โดยที่ $t = E$ และ v_i^t ไม่ปรากฏ v_i^A กำหนดให้ V^{ae} += $\{(v_i^E, -)\}$

2.2 กำหนดเส้นเชื่อมระหว่างเอนทิตีกับเอนทิตีให้เป็นผลเฉลย โดยที่ สำหรับแต่ละ e_i^{ee} โดยที่ $e_i^{ee} \in FG$ กำหนดให้ E^{ee} += $\{e_i^{ee}\}$

3.4.2 กระบวนการพิจารณาเคสที่เหมาะสมกับการสอบถามเพื่อนำมารียูส

วิธีการพิจารณาเคสที่เหมาะสมอาศัยค่าความคล้ายคลึงระหว่างคำสำคัญที่ได้จากการสอบถาม และปัญหาของเคส โดยใช้ค่าเฉลี่ยน้ำหนัก [4] ซึ่งมีวิธีการ ดังต่อไปนี้

ขั้นตอนที่ 1 พิจารณาหาเคสที่อาจเป็นไปได้

พิจารณารายการอ้างอิงของแต่ละ k_i เมื่อ i = 1, 2, …, n เมื่อ n คือจำนวนคำสำคัญ ถ้า $i\text{Type}_{k_i}$ = C แล้ว เซตของเคสที่อาจเป็นไปได้สามารถพิจารณาได้จาก

$\{ objCode_1 \cap objCode_2 \cap \ldots \cap objCode_m \}$ โดยที่ เซตดังกล่าวไม่เท่ากับเซตว่าง เมื่อ $objCode_i$ แทนรายการอ้างอิง และ m แทนจำนวนรายการอ้างอิงที่มากที่สุดที่สามารถอินเตอร์เซกชันกันได้ ซึ่ง $m \leq n$

ขั้นตอนที่ 2 พิจารณาหาเคสที่เหมาะสม

2.1 หาค่าความคล้ายคลึงระหว่างการสอบถามกับเคส โดยพิจารณาจากเซตของจุดที่ค้นพบ V_k กับปัญหาของแต่ละ C_i โดยที่

$$\text{ค่าความคล้ายคลึง } (V_k, C_i) = \frac{\text{จำนวนจุดเทียบเท่า}}{\text{จำนวนจุดพื้นฐาน}}$$

เมื่อ จำนวนจุดเทียบเท่า คือ จำนวนจุดจาก V_k ที่ตรงกับจุดจาก C_i

จำนวนจุดพื้นฐาน คือ จำนวนจุดที่ได้จาก $V_k \cup C_i$

2.2 เคสที่ให้ค่าความคล้ายคลึงมากที่สุด คือ เคสที่เหมาะสม

ค่าความคล้ายคลึงมีค่าตั้งแต่ 0-1 ถ้าค่าความคล้ายคลึงมีค่าเท่ากับ 1 แสดงว่าผลเฉลยของเคส คือ ผลลัพธ์ที่ได้จากการค้นหา ถ้าค่าความคล้ายคลึงมีค่ามากกว่าหรือเท่ากับ θ แสดงว่าผลเฉลยของเคสสามารถนำมารียูสได้ โดยค่า θ คือ ค่าการตัดสินใจที่ระบบยอมรับได้สำหรับการนำเคสมารียูส

3.4.3 กระบวนการรียูสเคส

การรียูสเคส คือ การปรับปรุงผลเฉลยของเคสเพื่อนำมาเป็นส่วนหนึ่งกราฟคำตอบ ซึ่งสามารถแบ่งเป็น 2 กรณี คือ กรณีแรก คือ การรียูสเคสเมื่อจุดแอทริบิวจากเป็นเซตย่อยของผลเฉลยจากเคส และกรณีที่สอง คือ การรียูสเคสเมื่อผลเฉลยจากเคสเป็นเซตย่อยของจุดที่ค้นพบ การรียูสเคสทั้งสองกรณีมีขั้นตอน ดังต่อไปนี้

1. การรียูสเคสเมื่อจุดที่ค้นพบเป็นเซตย่อยจุดจากเคส สามารถทำได้โดยกำหนดจุดผลเฉลยที่ได้จาก $V_k \cap C_i$ และทำการหากราฟที่เป็นไปได้จากผลเฉลยภายในเคส
2. การรียูสเคสเมื่อผลเฉลยจากเคสเป็นเซตย่อยของจุดที่ค้นพบ สามารถทำได้โดยหาเส้นทางพื้นฐานของจุด $V_k - (V_k \cap C_i)$ แล้วนำเส้นทางพื้นที่ได้มาเชื่อมกับผลเฉลยของเคส C_i