บทที่ 3

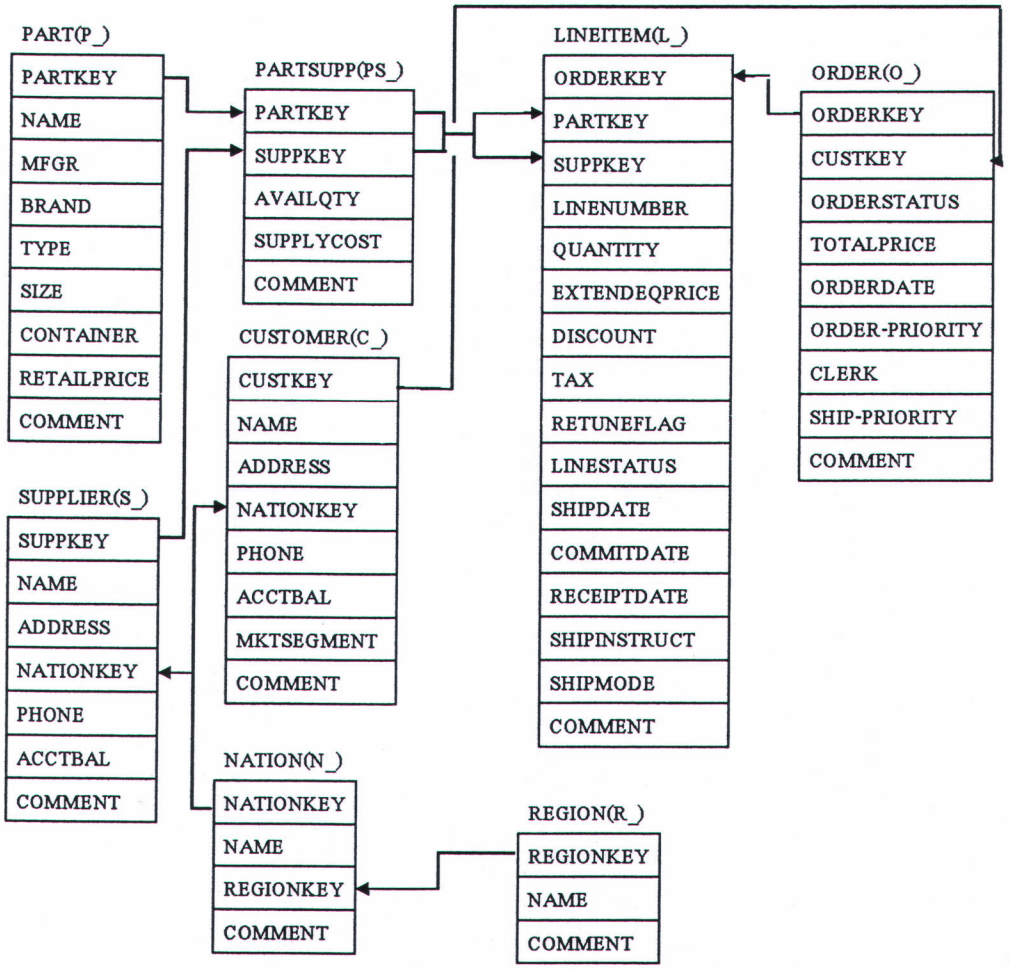
ปัญหาและวิธีการแก้ปัญหา

จากบทที่ 2 ได้นำเสนอขั้นตอนการค้นหาความซ้ำซ้อนของฐานข้อมูลไปแล้วนั้น จะเห็นได้ว่ากระบวนการค้นหาความซ้ำซ้อนของข้อมูลในฐานข้อมูลนั้นยังไม่มีประสิทธิภาพที่ดีพอ เนื่องจากขั้นตอนวิธีที่ใช้ในการค้นหาข้อมูลยังคงมีความยุ่งยากอยู่ เพราะขั้นตอนในการค้นหาความซ้ำซ้อนของข้อมูลต้องมีการนำข้อมูลออกมาจากฐานข้อมูล เพื่อนำมาค้นหาความซ้ำซ้อน และเมื่อทำการค้นหาความซ้ำซ้อนเสร็จแล้วจึงทำการนำข้อมูลกลับเข้าไปในฐานข้อมูลใหม่ ซึ่งทำให้ใช้วลาในขั้นตอนการค้นหาความซ้ำซ้อนมาก

แนวทางของการแก้ปัญหาของงานวิจัยนี้ คือการปรับปรุงประสิทธิภาพขั้นการค้นหาความซ้ำซ้อนของข้อมูล โดยจะทำการค้นหาความซ้ำซ้อนภายในระบบจัดการฐานข้อมูล ซึ่งสามารถแบ่งออกได้เป็น 2 วิธีด้วยกันคือ 1) การใช้วิธีการอิมพลีเมนต์บนระบบจัดการฐานข้อมูลด้วยภาษาพีแอลเอสคิวแอลทั้งหมด โดยจะนำเทคนิคในการปรับปรุงประสิทธิภาพขั้นพื้นฐานที่ได้รับความนิยมเกี่ยวกับการจัดการฐานข้อมูลมาใช้ ได้แก่ การเขียนการสอบถามใหม่ (Query Rewriting) และการสร้างดัชนีข้อมูล (Indexing) 2) การใช้วิธีการที่ให้ผู้ใช้สร้างฟังก์ชันการทำงานขึ้นมา (User-Define Function) ซึ่งเป็นภาษาบนระบบจัดการฐานข้อมูลที่ออราเคิล (Oracle) สร้างขึ้นมา และมีการใช้เทคนิคการสร้างดัชนีข้อมูลช่วยด้วย ซึ่งผลลัพธ์ของเวลาจากการปรับปรุงประสิทธิภาพด้วยเทคนิคที่กล่าวมาทั้งหมด จะนำไปเปรียบเทียบกันในรูปแบบของกราฟเปรียบเทียบประสิทธิภาพซึ่งจะทำการอธิบายในบทต่อไป

ในการค้นหาความซ้ำซ้อนของข้อมูล จะต้องมีขั้นตอนการเตรียมข้อมูลที่จะนำมาใช้ในการค้นหาความซ้ำซ้อนของข้อมูล เพื่อทำการเลือกชุดของข้อมูลในขั้นตอนการเลือกเดสคริปชัน โดยใช้หลักการของการทำเดสคริปชันที่ได้กล่าวไว้ในบทที่ 2 โดยจะใช้สกีมาของฐานข้อมูลทีซีพีเอช (TPC-H) ซึ่งเป็นฐานข้อมูลที่ได้นำมาใช้ในการศึกษา และทำการทดลองจริงในวิทยานิพนธ์ฉบับนี้

ซึ่งในบทนี้จะทำการอธิบายถึงการนำทฤษฎีต่างๆ ที่ได้อธิบายไว้ในบทที่ 2 ว่าได้นำมาประยุกต์ใช้อย่างไรเพื่อที่จะทำให้สามารถเพิ่มประสิทธิในการค้นหาความซ้ำซ้อนได้



รูปที่ 3.1 แสดงสกีมาของฐานข้อมูลทีซีพีเอช

จากรูปที่ 3.1 แสดงสกีมาของฐานข้อมูลทีซีพีเอช โดยจะอธิบายการรูปแบบในการจัดเก็บข้อมูลของฐานข้อมูลทีพีซีเอช จากสกีมาของฐานข้อมูลทีพีซีเอช มีการแบ่งตารางในการเก็บข้อมูลออกเป็น 8 ตาราง ได้แก่ ตารางเก็บข้อมูลสินค้า (PART) ตารางเก็บข้อมูลบริษัทผู้ผลิตสินค้า (SUPPLIER) ตารางเก็บข้อมูลความสัมพันธ์ระหว่างสินค้าและบริษัทผู้ผลิตสินค้า (PARTSUPP) ตารางการเก็บข้อมูลของลูกค้า (CUSTOMER) ตารางการเก็บข้อมูลประเทศที่มีการผลิตสินค้า (NATION) ตารางเก็บข้อมูลการซื้อขายสินค้า (LINEITEM) ตารางเก็บข้อมูลภมิภาคที่มาการผลิต

สินค้า (REGION) และตารางเก็บข้อมูลการสั่งสินค้า (ORDER) โดยกำหนดให้ตารางเก็บข้อมูลของสินค้า และตารางเก็บข้อมูลบริษัทผู้ผลิตสินค้า เป็นตารางหลักของชุดข้อมูลในสกีมา และกำหนดให้ตารางเก็บความความสัมพันธ์ระหว่างสินค้าและบริษัทผู้ผลิตสินค้า ตารางการเก็บข้อมูลของลูกค้า และตารางการเก็บข้อมูลประเทศที่มีการผลิตสินค้า เป็นไชลด์ (Child) ลำดับที่หนึ่งของตารางหลัก กำหนดให้ตารางเก็บข้อมูลการซื้อขายสินค้า และตารางเก็บข้อมูลภมิภาคที่มาการผลิตสินค้า เป็นไชลด์ลำดับที่สองของตารางหลัก และกำหนดให้ตารางเก็บข้อมูลการสั่งสินค้า เป็นไชลด์ลำดับที่สามของตารางหลัก

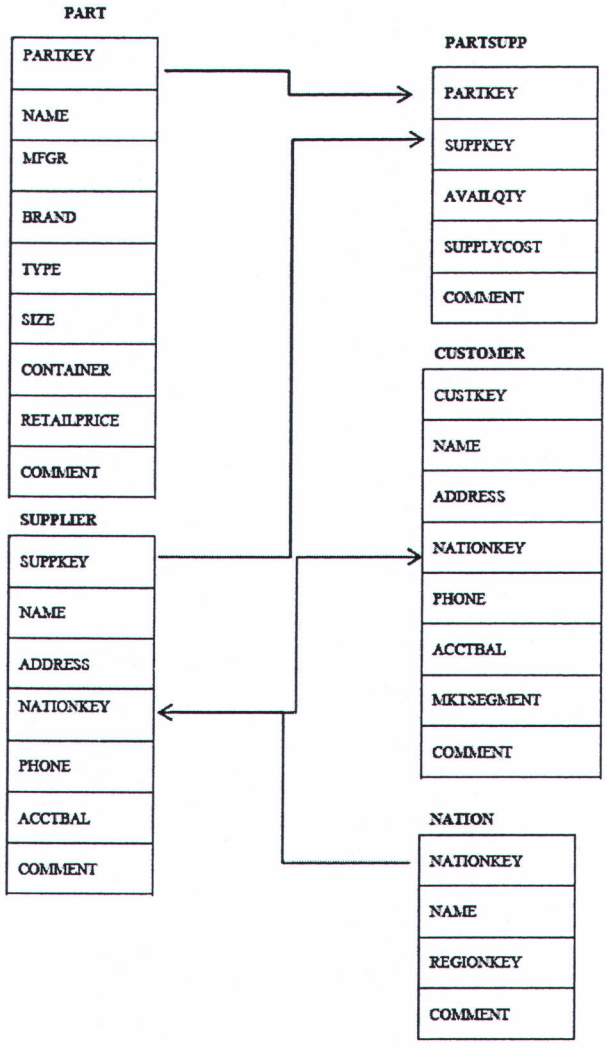
3.1 การเลือกเดสคริปชัน

จากสกีมาของฐานข้อมูลทีซีพีเอชในรูปที่ 3.1 เมื่อนำมาผ่านขั้นตอนการเลือกเดสคริปชันสามารถโดยใช้หลักการในการเลือกเดสคริปชันในบทที่ 3.2 โดยจะขออธิบายกระบวนการในการเลือกเดสคริปชันเป็นขั้นตอนๆ ดังนี้

3.1.1 การเลือกตามเลือกตามแนวคิดเฮอร์ริสติกในระดับสกีมา (Schema-based Heuristics) และการกำหนดเงื่อนไข (Condition)

แนวคิดเฮอร์ริสติกในระดับสกีมา คือแนวคิดที่ใช้ในการเลือกแอตทริบิวต์โดยจะให้ความสนใจไปที่ตัวสกีมาของฐานข้อมูลเพียงอย่างเดียวไม่ได้สนใจไปถึงตัวข้อมูลที่อยู่ภายใน โดยจะทำการตัดตารางที่ไม่ได้เป็นไชลด์ลำดับที่ 1 และตารางที่ไม่ใช่ตารางหลักของฐานข้อมูลออก

จากสกีมาของฐานข้อมูลในรูปที่ 3.1 เมื่อนำมาเข้ากระบวนการตามแนวคิดเฮอร์ริสติกจะได้สกีมาของฐานข้อมูลตามรูปที่ 3.2



รูปที่ 3.2 สกีมาของฐานข้อมูลทีซีพีเอชที่ผ่านกระบวนการตามแนวคิดเฮอร์ริสติกในระดับสกีมา

จากรูปที่ 3.2 สกีมาที่ได้มาหลังจากผ่านกระบวนการตามแนวคิดเฮอร์ริสติกในระดับสกีมา จะเห็นว่ามีการตัดตารางเก็บข้อมูลการซื้อขายสินค้า ตารางเก็บข้อมูลภมิภาคที่มาการผลิตสินค้า และตารางเก็บข้อมูลการสั่งสินค้าออกไปทั้งหมด 3 ตาราง เนื่องจากตารางทั้ง 3 เป็นตารางที่ไม่ได้เป็นไซลด์ลำดับที่ 1 ของตารางหลัก

การกำหนดเงื่อนไขในที่นี้เป็นการกำหนดเงื่อนไขขึ้นมา เพื่อใช้ประกอบการพิจารณาในการเลือกแอตทริบิวต์ ที่จะนำมาใช้เป็นตัวจำแนกในขั้นตอนการค้นหาความซ้ำซ้อนของข้อมูลในฐานข้อมูลต่อไป โดยจะทำการกำหนดเงื่อนไขซึ่งจะทำการพิจารณาเพียงแค่ในระดับสกีมา ซึ่งมีเงื่อนไขดังต่อไปนี้

1. ตัดแอตทริบิวต์ที่เก็บข้อมูลเป็นตัวอักษรออก
2. ตัดทุกแอตทริบิวต์ที่เป็นฟอรีนจ์คีย์ออก

จะขออธิบายถึงการตัดแอตทริบิวต์ต่างๆ เป็นลำดับไปตามเงื่อนไขที่ได้บอกไปในข้างต้นดังนี้ ตามเงื่อนไขที่ 1 ตัดแอตทริบิวต์ที่ได้เก็บข้อมูลเป็นตัวอักษรออก

1) ในตารางเก็บข้อมูลสินค้า จะทำการตัดแอตทริบิวต์ที่ชื่อว่า SIZE, RETAILPRICE ออก
2) ในตารางเก็บข้อมูลบริษัทผู้ผลิตสินค้า จะตัดแอตทริบิวต์ที่มีชื่อว่า ACCTBAL ออก
3) ในตารางเก็บข้อมูลความสัมพันธ์ระหว่างสินค้าและบริษัทผู้ผลิตสินค้า จะทำการตัดแอตทริบิวต์ที่มีชื่อว่า AVAILQTY, SUPPLYCOST ออก
4) ในตารางการเก็บข้อมูลของลูกค้า จะตัดแอตทริบิวต์ที่มีชื่อว่า ACCTBAL ออก
5) ในตารางการเก็บข้อมูลประเทศที่มีการผลิตสินค้า จะไม่มีการตัดแอตทริบิวต์ออกเลย

ตามเงื่อนไขที่ 2 จะตัดแอตทริบิวต์ที่เป็นฟอรีนจ์คีย์ออก ซึ่งจะทำการตัดแอตทริบิวต์ต่างๆ ออกดังนี้

1) ในตารางเก็บข้อมูลสินค้า จะทำการตัดแอตทริบิวต์ที่ชื่อว่า PARTKEY ออก
2) ในตารางเก็บข้อมูลบริษัทผู้ผลิตสินค้า จะตัดแอตทริบิวต์ที่มีชื่อว่า SUPPKEY, NATIONKEY ออก

3) ในตารางเก็บข้อมูลความสัมพันธ์ระหว่างสินค้าและบริษัทผู้ผลิตสินค้า จะทำการตัดแอตทริบิวต์ที่มีชื่อว่า PARTKEY, SUPPKEY ออก

4) ในตารางการเก็บข้อมูลของลูกค้า จะทำการตัดแอตทริบิวต์ที่มีชื่อว่า CUSTKEY, NATIONKEY ออก

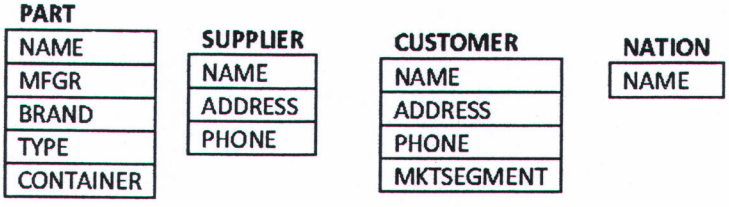
5) ในตารางการเก็บข้อมูลประเทศที่มีการผลิตสินค้า จะทำการตัดแอตทริบิวต์ที่มีชื่อว่า NATIONKEY, REGIONKEY ออก

3.1.2 การประยุกต์ใช้ตามแนวคิดอินสแทนเบสเฮอร์ริสติก (Instance Based Heuristics) และการเลือกโดยใช้โดเมนของความรู้ (Domain-knowledge)

การประยุกต์ตามแนวคิวอินสเทนเบสเฮอร์ริสติก คือ การเลือกแอตทริบิวต์ที่ให้ความสนใจในเรื่องของข้อมูลที่อยู่ภายในฐานข้อมูล ว่าเป็นข้อมูลที่สามารถใช้ในการค้นหาความซ้ำซ้อนของข้อมูลได้หรือไม่

การเลือกแอตทริบิวต์ที่จะนำมาใช้ในการค้นหาความซ้ำซ้อนโดยใช้โดเมนของความรู้ คือ การเลือกที่ต้องอาศัยความรู้ และความสามารถในการพิจารณาของผู้ที่มีความเชี่ยวชาญ หรือผู้ดูแลระบบฐานข้อมูลนั้นๆ

ซึ่งจากหลักการที่ใช้ในการพิจารณาที่ได้กล่าวมา จะทำการตัดแอตทริบิวต์ที่มีชื่อว่า COMMENT ออกในทุกๆตารางของฐานข้อมูลที่ได้ผ่านกระบวนการที่ 3.1.1 มาแล้ว



รูปที่ 3.3 สกีมาที่ได้ผ่านกระบวนการในการเลือกเดสคริปชันแล้ว

รูปที่ 3.3 เป็นสกีมาที่ได้ผ่านกระบวนการต่างๆ ที่ใช้ในการเลือกเดสคริปชันมา ทั้งกระบวนการที่ 3.1.1 และ 3.1.2 แล้ว ซึ่งตารางเก็บข้อมูลสินค้าจะเหลือแอตทริบิวต์ที่มีชื่อว่า NAME, MFGR, BRAND, TYPE, CONTAINER ในตารางเก็บข้อมูลบริษัทผู้ผลิตสินค้าจะเหลือแอตทริบิวต์ที่มีชื่อว่า NAME, ADDRESS, PHONE ในตารางเก็บข้อมูลความสัมพันธ์ระหว่างสินค้าและบริษัทผู้ผลิตสินค้า จะไม่เหลือแอตทริบิวต์เลย ในตารางการเก็บข้อมูลของลูกค้าจะเหลือแอตทริบิวต์ที่มีชื่อว่า NAME, ADDRESS, PHONE, MKTSEGMENT ส่วนในตารางการเก็บข้อมูลประเทศที่มีการผลิตสินค้า จะเหลือเพียงแค่แอตทริบิวต์ทีมีชื่อว่า NAME แค่นั้น

3.2 การจำแนกข้อมูลภายในฐานข้อมูล

ขั้นตอนการจำแนกข้อมูลที่มีความซ้ำซ้อนกัน เป็นขั้นตอนในการเลือกแอตทริบิวต์ที่จะนำมาใช้กำหนดเป็นเงื่อนไข ในการจำแนกความซ้ำซ้อนของข้อมูลว่าเป็นข้อมูลที่มีความซ้ำซ้อนหรือไม่ โดยสามารถแบ่งลักษณะของเงื่อนไขออกเป็น 3 ประเภทใหญ่ๆ คือ การจำแนกข้อมูลที่เป็นลบ การจำแนกข้อมูลที่เป็นบวก และประเภทสุดท้าย การจำแนกข้อมูลที่มีความคล้ายคลึงกัน

3.2.1 การจำแนกข้อมูลที่เป็นลบ

การจำแนกข้อมูลที่เป็นลบ จะทำการจำแนกข้อมูลภายในฐานข้อมูลว่าเป็นข้อมูลที่มีความซ้ำซ้อนกันหรือไม่ โดยจะช่วยในการตัดสินใจว่าเป็นข้อมูลที่ไม่มีความซ้ำซ้อนกันตามเงื่อนไขที่ได้กำหนดไว้ ถ้าผ่านการจำแนกตามเงื่อนไขแล้วได้ผลลัพธ์เป็นจริง จะแสดงว่าเป็นข้อมูลที่ไม่มีความซ้ำซ้อนกัน แต่ถ้าได้ผลเป็นอย่างอื่นจะยังไม่สามารถยืนยันได้ว่าเป็นข้อมูลที่มีความซ้ำซ้อนกันต้องทำการเปรียบเทียบข้อมูลในเงื่อนไขต่อๆไปก่อน

เงื่อนไขที่ใช้ในการจำแนกข้อมูลที่เป็นลบของฐานข้อมูลทีซีพีเอช คือ

1) ถ้าแอตทริบิวต์ที่มีชื่อว่า PART.CONTAINER, PART.TYPE, PART.BRAN เหมือนกันแต่แอตทริบิวต์ที่ชื่อ PART.NAME ต่างกันถือว่าเป็นข้อมูลที่ไม่มีความซ้ำซ้อนกัน เพราะ

สินค้าประเภทเดียวกัน ยี่ห้อเดียวกันถูกเก็บในคลังสินค้าเดียวกัน อาจจะไม่ใช่สินค้าชิ้นเดียวกันก็ได้เพราะอาจมีหลายๆรุ่น

2) ถ้าแอตทริบิวต์ที่มีชื่อว่า PART.NAME, PART.BRAND เหมือนแต่แอตทริบิวต์ที่ชื่อ PART.TYPE ต่างกันถือว่าเป็นข้อมูลที่ไม่มีความซ้ำซ้อนกัน เพราะสินค้าที่มีชื่อเดียวกันยี่ห้อเดียวกัน อาจจะไม่สินค้าชนิดเดียวกันก็ได้

3) ถ้าไม่เหมือนกันเลยซักแอตทริบิวต์ ให้ถือว่าเป็นข้อมูลที่ไม่มีความซ้ำซ้อนกัน

3.2.2 การจำแนกข้อมูลที่เป็นบวก

การจำแนกข้อมูลที่เป็นบวก จะทำการจำแนกข้อมูลภายในฐานข้อมูลว่าเป็นข้อมูลที่มีความซ้ำซ้อนกันหรือไม่ โดยจะช่วยในการตัดสินใจว่าเป็นข้อมูลที่มีความซ้ำซ้อนกันหรือไม่ ตามเงื่อนไขที่ได้กำหนดไว้ ถ้าผ่านการจำแนกตามเงื่อนไขแล้วได้ผลลัพธ์เป็นจริง จะแสดงว่าเป็นข้อมูลที่มีความซ้ำซ้อนกัน แต่ถ้าได้ผลเป็นอย่างอื่นจะยังไม่สามารถยืนยันได้ว่าเป็นข้อมูลที่ไม่มีความซ้ำซ้อนกันต้องทำการเปรียบเทียบข้อมูลในเงื่อนไขต่อๆไปก่อน

เงื่อนไขที่ใช้ในการจำแนกข้อมูลที่เป็นบวกของฐานข้อมูลทีซีพีเอช คือ

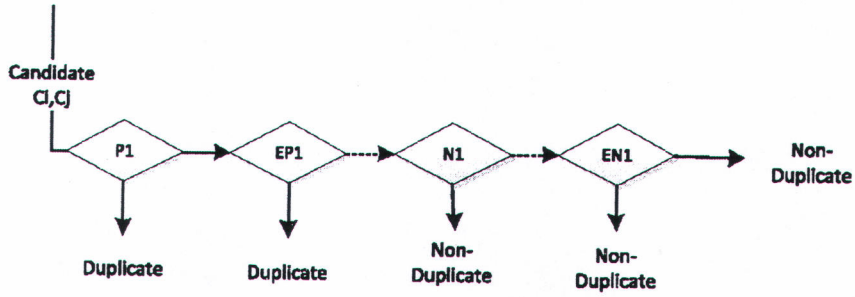
1) ถ้าเหมือนกันหมดทุกแอตทริบิวต์ จะแสดงว่าเป็นข้อมูลที่มีความซ้ำซ้อนกัน

2) ถ้าแอตทริบิวต์ที่มีชื่อว่า SUPPLIER.NAME, SUPPLIER.ADDRESS เหมือนแต่แอตทริบิวต์ที่ชื่อ SUPPLIER.PHONE ต่างกัน ถือว่าเป็นข้อมูลที่มีความซ้ำซ้อนกัน เพราะการที่ชื่อของสินค้า และที่อยู่ที่เหมือนกัน แต่การที่มีเบอร์โทรศัพท์ที่แตกต่างกันอาจเกิดจากการที่มีการเปลี่ยนแปลงเบอร์โทรศัพท์

3) ถ้าแอตทริบิวต์ที่มีชื่อว่า SUPPLIER.NAME เหมือนกันแต่แอตทริบิวต์ที่ชื่อ SUPPLIER.ADDRESS ต่างกัน ถือว่าเป็นข้อมูลที่มีความซ้ำซ้อนกัน เพราะการที่ชื่อของสินค้าเหมือนกันแต่มีที่อยู่ที่แตกต่างกันอาจจะผลิตจากบริษัทเดียวกัน แต่ต่างสาขากันก็ได้

3.2.3 การจำแนกข้อมูลที่มีความคล้ายคลึงกัน

การจำแนกข้อมูลที่มีความคล้ายคลึงกัน จะใช้เงื่อนไขในการพิจารณาทั้งในส่วนของ การจำแนกข้อมูลที่เป็นลบ และการจำแนกข้อมูลที่เป็นบวกควบคู่กันไป แต่จะมีเรื่องของการใช้ฟังก์ชันการหาค่าความต่างกันของตัวอักษรเข้ามาช่วยในการพิจารณา ซึ่งถ้าเมื่อผ่านการพิจาราณาแล้วมีความแตกต่างกันของตัวอักษรน้อยกว่าค่าที่ได้ทำการกำหนดไว้ในตอนเริ่มต้น จะถือว่าเป็นข้อมูลที่ไม่มีความซ้ำซ้อนกัน



รูปที่ 3.4 แสดงลำดับของเงื่อนไขที่ใช้ในการจำแนกความซ้ำซ้อนของข้อมูล

จากรูปที่ 3.4 แสดงลำดับของเงื่อนไขที่ใช้ในการจำแนกความซ้ำซ้อนของข้อมูลโดยให้ C แทนคาร์ดิเดตของข้อมูล และให้ i, j แทนลำดับของชุดข้อมูลที่จะนำมาทำการเปรียบเทียบ ให้ P แทนการจำแนกข้อมูลที่เป็นบวก ให้ N แทนการจำแนกข้อมูลที่เป็นลบ และให้ EP แทนการจำแนกข้อมูลที่มีความคล้ายคลึงกัน ส่วนตัวเลขใช้แทนเงื่อนไขลำดับที่ 1,2,3 ไปจนครบทุกเงื่อนไขของการจำแนกในแต่ละแบบ โดยมีลำดับในการจำแนกข้อมูลดังนี้คือ ทำการจำแนกข้อมูลที่เป็นบวกตามเงื่อนไขลำดับที่ 1 ก่อน (P1) และจำแนกข้อมูลที่มีความคล้ายคลึงกันตามเงื่อนไขลำดับที่ 1 ต่อ และทำการจำแนกข้อมูลที่เป็นบวกตามเงื่อนไขลำดับที่ 2 ต่อ (P2) โดยจะทำการจำแนกสลับกันไปแบบนี้จนเสร็จทุกเงื่อนไขในการจำแนกที่เป็นบวก แล้วจึงทำการจำแนกข้อมูลที่เป็นลบสลับไปการจำแนกข้อมูลที่มีความคล้ายคลึงกันต่อจนเสร็จครบทุกเงื่อนไข

3.3 การทำดัชนีข้อมูล

การทำดัชนีข้อมูลจะถูกแบ่งออกเป็น 2 รูปแบบด้วยกัน คือ 1) แบบที่ใช้สำหรับวิธีการที่ให้ผู้ใช้สร้างฟังก์ชันการทำงานขึ้นมา (UDF) และ 2) แบบที่ใช้สำหรับการค้นหาความซ้ำซ้อนโดยการใช้วิธีการอิมพลีเมนต์บนระบบจัดการฐานข้อมูลด้วยภาษาพีแอลเอสคิวแอลทั้งหมด โดยจะขออธิบายการทำดัชนีข้อมูล เพื่อที่ใช้ในวิธีการค้นหาความซ้ำซ้อนของข้อมูลต่างๆ ดังนี้

3.3.1) การทำดัชนีข้อมูลเพื่อใช้สำหรับวิธีการที่ให้ผู้ใช้สร้างฟังก์ชันการทำงานขึ้นมาเอง

จากสกีมาที่ได้ผ่านขั้นตอนการเลือกเดสคริปชันแล้วในรูปที่ 3.3 เมื่อนำมาทำการจอยกันซึ่งจะได้สกีมาดังรูปที่ 3.5

TT1
NUM_REC
P_NAME
P_MFGR
P_BRAND
P_TYPE
P_CONTAINER
S_NAME
S_ADDRESS
S_PHONE
C_NAME
C_ADDRESS
C_PHONE
C_MKTSEGMENT
N_NAME

TT2
NUM_REC
P_NAME
P_MFGR
P_BRAND
P_TYPE
P_CONTAINER
S_NAME
S_ADDRESS
S_PHONE
C_NAME
C_ADDRESS
C_PHONE
C_MKTSEGMENT
N_NAME

รูปที่ 3.5 แสดงสกีมาที่ได้ผ่านการจอยกันแล้ว

จากรูปที่ 3.5 แสดงสกีมาที่ได้ผ่านขั้นตอนการทำเดสคริปชันแล้ว และนำมาจอยกันแล้ว โดยกำหนดให้มีชื่อว่า TT1 และ TT2 ซึ่งให้ตาราง TT1 เก็บข้อมูลในชุดที่หนึ่ง และตาราง TT2 จะเก็บข้อมูลในชุดที่สอง เพื่อที่จะนำมาเปรียบเทียบกัน

จากหลักการในการทำดัชนีข้อมูลต่างๆที่ได้อธิบายเอาไว้ในบทที่ 2 เมื่อทำการพิจารณากับสีกมาในรูปที่ 3.5 สำหรับในตาราง TT1 จะต้องทำดัชนีข้อมูลในแอตทริบิวต์ NUM_REC และในตาราง TT2 ก็จะต้องทำดัชนีข้อมูลในแอตทริบิวต์ NUM_REC เหมือนกัน

สำหรับแนวทางให้ผู้ใช้สร้างฟังก์ชันการทำงานขึ้นมาเองนั้น จะต้องทำการสร้างตารางขึ้นมาเพื่อที่จะทำการรองรับข้อมูลที่ได้ผ่านการค้นหาความซ้ำซ้อนแล้วมาเก็บเอาไว้ เพื่อที่จะเก็บค่าว่าเป็นรูปแบบที่มีความซ้ำซ้อนรูปแบบไหน

Sim

Key_A	Key_B	KeySim_of_AB

รูปที่ 3.6 แสดงสกีมาที่จะใช้การเก็บค่ารูปแบบที่มีความซ้ำซ้อนกันของข้อมูล

รูปที่ 3.6 แสดงสกีมาที่จะใช้การเก็บค่าว่าได้การค้นหาความซ้ำซ้อนในรูปแบบไหน โดยกำหนดให้แอตทริบิวต์ Key_A จะเก็บค่าว่าเป็นลำดับที่เท่าไหร่ของข้อมูลในตาราง TT1 กำหนดให้แอตทริบิวต์ Key_B จะเก็บค่าว่าเป็นลำดับที่เท่าไหร่ของข้อมูลในตาราง TT2 และกำหนดให้แอตทริบิวต์ KeySim_of_AB จะเก็บค่าว่าเป็นความซ้ำซ้อนในรูปแบบไหนเมื่อได้ผ่านขั้นตอนการค้นหาความซ้ำซ้อนมาแล้ว

3.3.2) การทำดัชนีข้อมูลเพื่อใช้สำหรับวิธีการอิมพลีเมนต์บนระบบจัดการฐานข้อมูลด้วยภาษาแอลเอสคิวแอลทั้งหมด

การทำดัชนีข้อมูลในแนวทางที่ใช้การอิมพลีเมนต์บนระบบจัดการฐานข้อมูลด้วยภาษาแอลเอสคิวแอลทั้งหมด จะต้องทำการทำดัชนีข้อมูลของสกีมาในรูปที่ 3.5 ซึ่งเป็นสกีมาที่ต้องใช้ในการขั้นตอนการค้นหาความซ้ำซ้อนตามแนวทางนี้เหมือนกัน แต่สำหรับในแนวทางนี้จะต้องมีการทำดัชนีข้อมูลเพิ่มในส่วนของตารางที่ต้องมีการสร้างขึ้นมาเพิ่มเติม เพื่อที่จะใช้ในขั้นตอนการจำแนกความซ้ำซ้อนตามรูปแบบการจำแนกข้อมูลที่มีความคล้ายคลึงกัน

ในการจำแนกข้อมูลที่มีความคล้ายคลึงกัน ตามแนวทางการค้นหาความซ้ำซ้อนของข้อมูลในรูปแบบที่ใช้การอิมพลีเมนต์บนระบบจัดการฐานข้อมูลด้วยภาษาแอลเอสคิวแอลทั้งหมด ซึ่งในการจำแนกข้อมูลในรูปแบบนี้จะต้องใช้ค่าความแตกต่างกันของตัวอักษรมาช่วยในการพิจารณา ซึ่งการที่จะใช้ค่าความแตกต่างของอักษรมาพิจารณานั้น ต้องทำการตัดคำในแอตทริบิวต์ต่างๆออกมาโดยแบ่งออกเป็นทีละสามตัวอักษร ซึ่งจะต้องทำการตัดคำในทุกๆแอตทริบิวต์ที่ได้ผ่านขั้นตอนการเลือกเดสคริปชันมาแล้วซึ่งจะแสดงเป็นสกีมาได้ดังนี้

รูปที่ 3.7 แสดงสกีมาของแอตทริบิวต์ที่ได้ผ่านการตัดคำแล้ว

จากรูปที่ 3.7 จะเห็นได้ว่าจะมีคำว่า “TT1 และ TT2” นำหน้าชื่อในตารางซึ่งหมายความเป็นแอตทริบิวต์ที่ได้ผ่านขั้นตอนในการตัดคำมาจากตารางไหน ถ้านำหน้าด้วย TT1 แสดงว่าเป็นแอตทริบิวต์ที่ได้ผ่านการตัดคำมาจากตาราง TT1 และถ้าเป็น TT2 ก็แสดงว่าเป็นแอตทริบิวต์ที่ได้ผ่านการตัดคำจากตาราง TT2 สำหรับแอตทริบิวต์ต่างๆในตารางจะทำการเก็บค่าเหมือนกัน คือ แอตทริบิวต์ ID จะเก็บค่าลำดับของแอตทริบิวต์ที่อยู่ในตาราง TT1 สำหรับตารางที่มีชื่อขึ้นต้นด้วย TT1 และจะค่าลำดับของแอตทริบิวต์ที่อยู่ในตาราง TT2 สำหรับตารางที่มีชื่อขึ้นต้นด้วย TT2 ส่วนในแอตทริบิวต์ POS จะเก็บค่าลำดับที่มีการตัดคำว่าเป็นลำดับที่เท่าไหร่ที่ทำการตัดคำในคำๆนั้น เช่น คำว่า “Position” จะต้องทำการตัดคำทั้งหมด 10 ครั้ง ครั้งที่ 1 ได้ “##P” ครั้งที่ 2 ได้ “#Po” ครั้งที่ 3 ได้ “Pos” ครั้งที่ 4 ได้ “osi” ครั้งที่ 5 ได้ “sit” ครั้งที่ 6 ได้ “iti” ครั้งที่ 7 ได้ “tio” ครั้งที่ 8 ได้ “ion” ครั้งที่ 9 ได้ “on$” ครั้งที่ 10 ได้ “n$$” ซึ่งในที่นี้ลำดับครั้งในการตัดคำคือค่าที่จะถูกนำไปเก็บในแอตทริบิวต์ POS และในส่วนของแอตทริบิวต์ QGRAM จะทำการเก็บค่าตัวอักษรที่ได้จากการตัดคำโดยจะเริ่มต้นการตัดคำด้วยสัญญาลักษณ์ “#” และจะจบการตัดคำด้วยสัญญาลักษณ์ “$”

จากสกีมาของข้อมูลในรูปที่ 3.7 นั้นเมื่อนำมาทำการพิจาราณาในการทำดัชนีข้อมูล จะต้องทำการทำดัชนีข้อมูลในแอตทริบิวต์ที่มีชื่อว่า “ID” ในทุกๆตารางข้อมูลเนื่องจากเป็นแอตทริบิวต์ที่ต้องมีการจัดลำดับของข้อมูล

จากแนวทางในการแก้ปัญหาต่างๆที่ได้กล่าวมาข้างต้นนั้น จะถูกนำไปทำการทดลองเพื่อวิเคราะห์ผลลัพธ์ที่ได้ว่าเป็นวิธีการที่สามารถเพิ่มประสิทธิภาพในการค้นหาความซ้ำซ้อนของข้อมูลในฐานข้อมูลได้หรือไม่ ซึ่งจะทำการทดลองในบทต่อไป