บทที่ 2

ทฤษฎีที่เกี่ยวข้อง

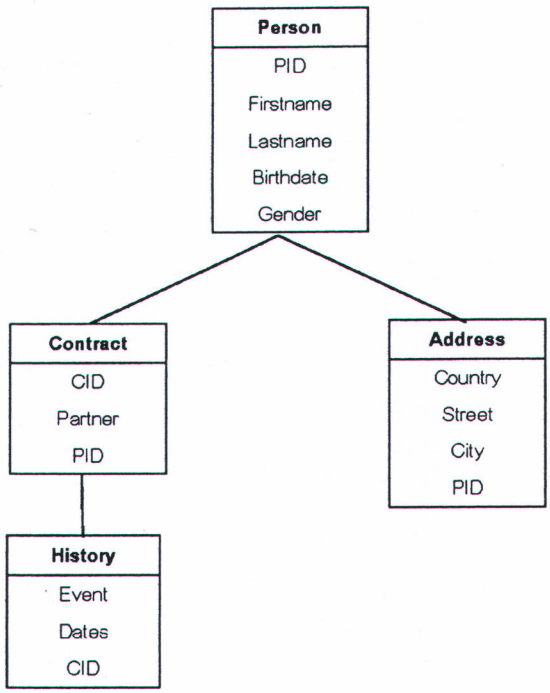
โดยปกติแล้วในฐานข้อมูลของระบบสารสนเทศขนาดใหญ่ อาจมีความซ้ำซ้อนของข้อมูลเกิดขึ้นได้ ความซ้ำซ้อนที่เหล่านั้นสามารถที่จะใช้ขั้นตอนวิธีการต่างๆ เพื่อตรวจสอบความซ้ำซ้อนได้

ส่วนประกอบของกระบวนการการค้นหาความซ้ำซ้อนกันของข้อมูล แบ่งออกได้เป็น 3 ส่วยประกอบหลัก ซึ่งจะถูกนำมาใช้ในงานวิทยานิพนธ์นี้ คือ

1) การเลือกเดสคริปชัน (Description) เพื่อใช้ในการกำหนดแอตทริบิวต์ที่จะนำไปใช้การจำแนก
2) การจำแนก (Classification) ของข้อมูลที่มีความซ้ำซ้อน
3) การปรับปรุงประสิทธิภาพการสอบถาม (Query optimization)

2.1 ทฤษฎีในการเลือกเดสคริปชัน

การเลือกเดสคริปชันคือ การเลือกแอตทริบิวต์ที่เกี่ยวข้องกับแคนดิเดต (Candidates) ซึ่งเป็นตัวแทนของข้อมูลต่างๆ ซึ่งการเลือกเดสคริปชันนั้นคือ การหาแคนดิเดตเพื่อที่จะนำมาใช้เป็นตัวตรวจวัดความซ้ำซ้อนของข้อมูล โดยในขั้นแรกจะใช้แนวคิดเฮอร์ริสติกในระดับสกีมา (Schema-based Heuristics) และมีการใช้เงื่อนไข (Condition) ในการเลือกชุดข้อมูล เพื่อลดการสืบทอดของข้อมูลระหว่างตารางที่มีความสัมพันธ์ของเรนต์-ไชลด์ (Parent-Child Relationship) ที่มากเกินความจำเป็น โดยจะสนใจพียงแค่การสืบทอดในชั้นที่หนึ่งเพียงชั้นเดียว และทำการจำกัดชนิดของข้อมูลที่จะใช้พิจารณา เพื่อลดการพิจารณาที่ซับซ้อน ดังตัวอย่างที่ 1 ซึ่งจะมีการใช้สกีมาจากรูปที่ 1 เพื่อใช้ในการอธิบายเพื่อประกอบความเข้าใจ



Person

PID	Firstname	Lastname	Birthdate	Gender
1	John	Doe	29/03/1970	Male
2	Jonathan	Doe	29/03/1907	Male
3	Michael	Mustermann		Female
4	John	Doe		Male
5	Ken	Mustermann		Male
6	Jane	Smith	12/12/1978	Female

Contract

CID	Partner	PID
123	NZ Bank	1
456	National Bank of New Zealand	2
789	Foreign Bank	3
123	NZ Bank	4
345	NZ Bank	5
123	NZ Bank	6

Address

Country	Street	City	PID
Germany	Main Street 1	Auckland	1
Germany	Large Place 2	Christchurch	1
Germany	Large Place 2	Christchurch	2
Germany			3
Germany	Large Place 2	Christchurch	4
Germany			5
Germany		Queenstown	6

History

Event	Date	CID
Open	29/03/1988	123
Update Student to Pro	01/01/1993	123

รูปที่ 2.1 แสดงรูปแบบการเก็บข้อมูลของบุคคลทั้งในแบบสกีมา และตารางการเก็บข้อมูล

จากรูปที่ 2.1 แสดงรูปแบบการเก็บข้อมูลประวัติของบุคคลประกอบไปด้วย 4 ตาราง คือ ตารางบุคคล (Person) ตารางเก็บการทำสัญญา (Contract) ตารางเก็บที่อยู่ (Address) และตารางการเก็บประวัติในการทำสัญญา (History) ในตารางบุคคลจะทำการเก็บข้อมูลเกี่ยวกับ ชื่อ (Firstname) นามสกุล (Lastname) วันเกิด (Birthdate) และเพศ (Gender) ตารางการเก็บที่อยู่จะเก็บข้อมูลประเทศ (Country) ถนน (Street) และเมือง (City) ตารางการเก็บการทำสัญญาจะเก็บข้อมูล หมายเลยสมาชิกของธนาคาร (CID) ชื่อธนาคาร (Partner) ตารางเก็บประวัติในการทำสัญญาจะเก็บข้อมูลรายการ (Event) และวันที่ใช้บริการธนาคาร (Date) โดยสามารถแสดงความสัมพันธ์ของข้อมูลในตารางได้ เช่น ชื่อ Jonathan Doe เกิดวันที่ 29.03.1907 อยู่ที่ถนน Main Street1 เมือง Auckland ที่อยู่ที่ 2 คือถนน Large Place 2 เมือง Christchurch มีสัญญากับธนาคาร NZBank หมายเลขสมาชิกของธนาคาร 123 ทำการเปิดบัญชี (Open) วันที่ใช้บริการ (29.03.1988) และมีการปรับสมุดบัญชีธนาคาร (Update) วันที่ใช้บริการ (01.01.1993)

ตัวอย่างที่ 1 สมมุติให้ใช้การเลือกตามแนวคิดเฮอร์ริสติก โดยใช้ความสัมพันธ์ของข้อมูลจากรูปที่ 1 สำหรับเลือกแคนดิเดตของบุคคลดังนี้

1 จะตัดตารางที่ไม่ได้เป็นไซลด์ลำดับที่หนึ่งของตารางหลัก เช่น ตัดตารางประวัติความน่าเชื่อถือ (Credit History) เนื่องจากได้ทำการกำหนดให้ตารางบุคลเป็นตารางหลัก ดังนั้นตารางประวัติความน่าชื่อถือจึงเป็นไซลด์ลำดับที่สอง เมื่อเปรียบเทียบกับตารางหลัก

2 ตัดแอตทริบิวต์ที่เก็บข้อมูลเป็นตัวอักษร (Character) ออก เช่น วันเกิด (Birthday)

3 ตัดทุกแอตทริบิวต์ที่เป็นฟอรีนจ์คีย์ (Foreign Key) เช่น Contract.PID

ผลลัพธ์ของข้อมูลที่ได้ตามแนวคิดของเฮอร์ริสติกจะประกอบไปด้วยแอตทริบิวต์

Person.Firstname, Person.Lastname, Person.Gender, Contract.Partner, Address.Country, Address.Street, Address.City

จากการตัดแอตทริบิวต์ตามเงื่อนไขที่ 2 เมื่อพิจารณาจากตัวอย่างที่ 1 จะเห็นได้ว่าเมื่อตัดแอตทริบิวต์ที่เก็บข้อมูลวันเกิดออก จะทำให้สามารถลดการพิจารณาจำนวนของแอตทริบิวต์ลงได้ อย่างไรก็ตามยังมีแอตทริบิวต์ที่เก็บข้อมูลเป็นตัวอักษรอีกจำนวนมาก ที่ไม่สามารถที่จะใช้ในการอธิบายถึงความสัมพันธ์ต่างๆในระบบสารสนเทศได้ และจะเพิ่มการประยุกต์ใช้แนวคิดอินส์แทนเบสเฮอร์ริสติก (Instance Based Heuristics) กับโดเมนของความรู้ (Domain-knowledge) ทั้งนี้อินส

แทนเบสเฮอร์ริสติกจะทำการตัดแอตทริบิวต์มีค่าเหมือนกัน หรือมีค่าว่าง (Null) เป็นจำนวนมาก ซึ่งจะกำหนดด้วยค่าความถี่ของคำ (IDF) [2,8,10] เพื่อลดแอตทริบิวต์ที่สามารถระบุได้ทันทีว่าไม่สามารถที่จะใช้ในการอธิบายว่าเป็นข้อมูลที่มีความซ้ำซ้อนได้ โดยกระบวนการนี้จะตัดแอตทริบิวต์ที่เก็บข้อมูลเกี่ยวกับ ประเทศ และเพศ ออกไป

ในส่วนโดเมนของความรู้จะเป็นขั้นตอนการเลือกแอตทริบิวต์ต่างๆ โดยอาศัยผู้ที่มีความเชี่ยวชาญกับฐานข้อมูลนั้นเป็นพิเศษ เนื่องจากการประยุกต์ใช้ (Application) ที่แตกต่างกันก็ย่อมมีความต้องการที่แตกต่างกันไป

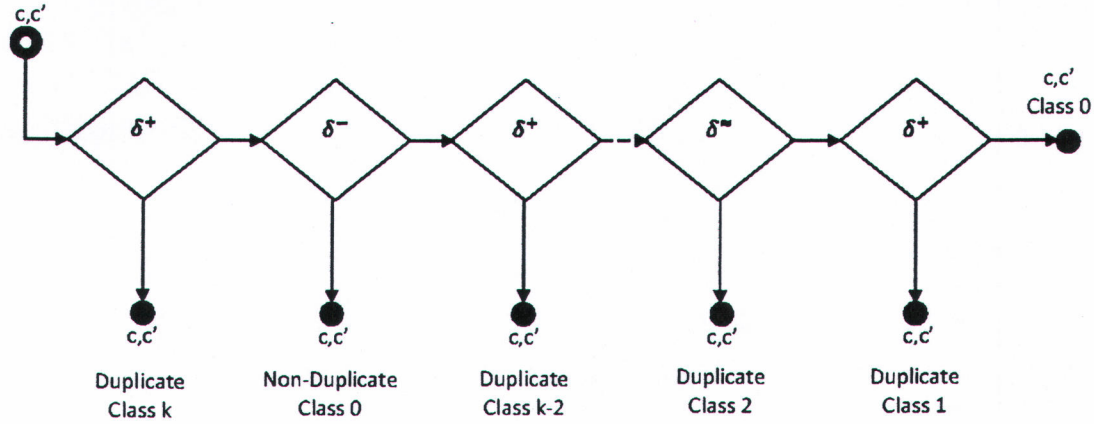
เมื่อผ่านการคัดเลือกในขั้นตอนนี้จะได้ผลลัพธ์ที่ประกอบไปด้วยแอตทริบิวต์ดังนี้ Person.Firstname, Person.Lastname, Contract.Partner, Address.Street, Address.City และเรียกชุดของข้อมูลที่ถูกเลือกนี้ว่า “เคสคริปชัน”

2.2 ทฤษฎีในการจำแนกข้อมูล

การจำแนกของข้อมูลที่มีความซ้ำซ้อนกัน จะนำแคนดิเดตที่ได้จากขั้นตอนการทำเคสคริปชันมาทำการเปรียบเทียบ เพื่อหาว่าเป็นข้อมูลที่ซ้ำซ้อนกัน (Duplicates) หรือไม่ซ้ำซ้อนกัน (Non-duplicates) โดยจะจำแนกข้อมูลจากการจับคู่ของแคนดิเดต เพื่อพิจารณาว่าเป็นข้อมูลที่ซ้ำซ้อนกันหรือไม่

การจำแนกข้อมูลแบ่งเป็น 2 ประเภทคือ (1) การจำแนกข้อมูลจากความคล้ายคลึงกัน (Similarity-base Classifiers) และ (2) การจำแนกข้อมูลตามกฎ (Rule-based Classifiers) ซึ่งจะมี 2 รูปแบบคือ การจำแนกข้อมูลที่เป็นลบ (Negative Classifier) ถ้าจำแนกข้อมูลแล้วให้ผลลัพธ์เป็นจริงแสดงว่าเป็นข้อมูลที่ไม่ซ้ำซ้อนกัน ถ้าผลลัพธ์เป็นเท็จจะได้ผลลัพธ์เป็น -1 ส่วนการจำแนกข้อมูลที่เป็นบวก (Positive Classifier) ถ้าจำแนกข้อมูลแล้วให้ผลลัพธ์เป็นจริงแสดงว่าเป็นข้อมูลที่ซ้ำซ้อนกัน ถ้าผลลัพธ์เป็นเท็จจะได้ผลลัพธ์เป็น -1

โดยการกำหนดลำดับความซ้ำซ้อนของข้อมูล จะใช้ค่า k ในการกำหนดลำดับ โดยให้ k เป็นความสัมพันธ์ร่วมกับคลาส (Class) ซึ่งแทนด้วย c ซึ่งจะมีค่าตั้งแต่ $0 \leq i \leq k$ โดยที่ class $c = i+1$ ถ้าตำแหน่งยิ่งใกล้ค่า k มาก จะแสดงว่าเป็นข้อมูลที่มีความซ้ำซ้อนกันมาก แต่ในการจำแนกข้อมูลที่เป็นลบ ถ้าผลลัพธ์เป็นจริงจะกำหนดให้เป็นส่วนหนึ่งของ Class 0 ทันทีไม่ต้องนำไปหาตำแหน่งคลาสจากความสัมพันธ์ดังกล่าว และจะกำหนดให้ข้อมูลที่มีความซ้ำซ้อนกันมากที่สุดเป็นคลาส k



รูปที่ 2.2 แสดงการจำแนกข้อมูลด้วยการจำแนกในแบบต่างๆ

จากรูปที่ 2.2 แสดงการจำแนกข้อมูล โดยจะทำการจำแนกข้อมูลจากการนำแคนดิเดต 2 แคนดิเดต (c,c') มาทำการเปรียบเทียบกัน ซึ่งแคนดิเดตคู่แรกที่นำมาทำการจำแนกจะได้ผลลัพธ์ว่าเป็นข้อมูลที่มีความซ้ำซ้อนกัน มีคลาสเป็น k แสดงว่าเป็นคู่ของแคนดิเดตที่มีความซ้ำซ้อนกันมากที่สุด และเมื่อทำการจำแนกต่อไปได้ค่าเป็น -1 ทำให้ไม่สามารถบอกได้ว่าเป็นข้อมูลที่มีความซ้ำซ้อนกันหรือไม่ และจะทำการจำแนกข้อมูลต่อไปซึ่งได้ผลลัพธ์ว่าเป็นข้อมูลที่ไม่ซ้ำซ้อนกัน ในการจำแนกครั้งสุดท้ายจะให้ผลลัพธ์เป็น -1 เสมอ และจะมีคลาสเป็น 0

เพื่อให้การจำแนกข้อมูลมีประสิทธิภาพสูง ได้แบ่งการจำแนกข้อมูลออกเป็นทั้งหมด 11 คลาส เพื่อใช้เป็นรูปแบบในการค้นหาข้อมูลที่ซ้ำซ้อนกัน และจะไม่ทำการพิจารณาถึงฟังก์ชันเบื้องหลัง (Function Background) เพราะจะทำให้เกิดความซ้ำซ้อนกัน และความเป็นไปได้ในการร้องขอกระบวนที่แตกต่างกัน เช่น ความซ้ำซ้อนกันของข้อมูลที่เกิดจากเปลี่ยนแปลงที่อยู่ หรือเปลี่ยนนามสกุล และยังสามารถที่จะพัฒนาให้มีประสิทธิภาพเพิ่มมากขึ้นได้ด้วยการเพิ่มรูปแบบการจำแนกให้ครอบคลุมกับความไม่ซ้ำซ้อนกันของข้อมูลที่สามารถรับรู้ได้เอง

2.2.1) การจำแนกข้อมูลที่เป็นลบ

การจำแนกข้อมูลที่เป็นลบจะใช้สมการที่ 1 ในการเปรียบเทียบคู่ของแคนดิเดต ถ้าได้ผลลัพธ์เป็น p หรือจริง แสดงว่าเป็นข้อมูลที่ไม่ซ้ำซ้อนกัน แต่ถ้าได้ออกมาเป็นค่าอื่นจะให้ได้ผลลัพธ์เป็น –1 ไว้ก่อน ซึ่งจะยังไม่สามารถจำแนกได้ว่าเป็นข้อมูลที่มีความซ้ำซ้อน หรือไม่

$$\delta^{-}(c, c') = \begin{cases} 0, & if\ p \\ -1, & otherwise \end{cases} \qquad (1)$$

จะขอทำการยกตัวอย่างจากตัวอย่างที่ 2 เพื่อเพิ่มความเข้าใจในขั้นตอนการจำแนกข้อมูลที่เป็นลบ โดยใช้ฐานข้อมูลจากรูปที่ 1 ซึ่งผ่านขั้นตอนในการทำเดสคริปชันมาแล้ว

ตัวอย่างที่ 2

ชื่อ, นามสกุล, ถนน1, เมือง1

P1 = John, Doe, NZ Bank, Main Street 1,

P2 = Michael, Mustermann, Foreign Bank, Unknown, Unknown

จากข้อมูลในชุด P1 และ P2 ที่ได้นำมาทำการเปรียบเทียบ จะเห็นได้ว่าเป็นคู่ของแคนดิเดตที่ไม่มีแอตทริบิวต์ไหนมีความเหมือนกันเลย จึงทำให้ได้ผลลัพธ์ว่าเป็นจริง และเป็นข้อมูลที่ไม่มีความซ้ำซ้อนกันจากการจำแนกข้อมูลที่เป็นลบ

อย่างไรก็ตามการจำแนกข้อมูลที่เป็นลบในรูปแบบเดิม อาจไม่สามารถที่จะใช้จำแนกความซ้ำซ้อนกันของข้อมูลได้อย่างแท้จริง โดยจะขอยกตัวอย่างที่ 3 เพื่อเพิ่มความเข้าใจดังนี้

ตัวอย่างที่ 3

ชื่อ, นามสกุล, ถนน1, เมือง1

P1 = Michael, Mustermann, Unknown, Unknown

P2 = Ken, Mustermann, Unknown, Unknown

จากตัวอย่างจะเห็นได้ว่าจากชุดข้อมูลที่นำมาใช้ในการเปรียบเทียบ เมื่อทำการจำแนกข้อมูลจากการจำแนกข้อมูลที่เป็นลบแล้ว แทบเป็นไปไม่ได้เลยว่าจะเป็นข้อมูลที่ไม่ซ้ำซ้อนกัน เพราะมีค่าแอตทริบิวต์ที่เหมือนกันถึง 3 ค่าดังนั้นจึงต้องมีการกำหนดรูปแบบในการค้นหาความ

ซ้ำซ้อนที่ถี่ถ้วนมากขึ้น โดยจะกำหนดให้เป็นข้อมูลที่ไม่ซ้ำซ้อนกันถ้าแอตทริบิวต์ที่เก็บข้อมูลในด้านที่อยู่ และแอตทริบิวต์ที่เก็บนามสกุลมีค่าเหมือนกัน ให้ทำการพิจารณาแอตทริบิวต์ที่เก็บข้อมูลเรื่องชื่อด้วยถ้ามีค่าไม่เหมือนกันจะถือว่าเป็นข้อมูลที่ไม่ซ้ำซ้อนกัน ซึ่งเป็นรูปแบบหนึ่งที่จะถูกนำมาใช้ในการจำแนกข้อมูลที่เป็นลบ

2.2.2) การจำแนกข้อมูลที่เป็นบวก

การจำแนกข้อมูลที่เป็นบวกจะใช้สมการที่ 2 ในการเปรียบเทียบคู่ของแคนดิเดต ถ้าได้ผลลัพธ์เป็น p หรือจริง แสดงว่าเป็นข้อมูลที่มีความซ้ำซ้อนกัน แต่ถ้าได้ออกมาเป็นค่าอื่นจะให้ได้ผลลัพธ์เป็น –1 ไว้ก่อน

$$\delta^{+}(c, c') = \begin{cases} c, & if\ p\ (c\ is\ position\ i+1\ in\ profile) \\ -1, & otherwise \end{cases} \quad (2)$$

ใน [1] ได้นำเสนอรูปแบบของการจำแนกข้อมูลที่เป็นบวกจำนวน 9 รูปแบบ โดยในแต่ละการจำแนกจะให้ผลลัพธ์ที่แตกต่างกัน และมีการกำหนดลำดับของความซ้ำซ้อนกันของข้อมูลว่ามีความซ้ำซ้อนกันมากขนาดไหนจากรูปแบบต่างๆ ที่นำมาใช้ในการจำแนกข้อมูลที่เป็นบวก

อย่างไรก็ตามเพื่อเพิ่มความเข้าใจในการจำแนกข้อมูลที่เป็นบวก จะขอยกตัวอย่างรูปแบบการจำแนกข้อมูลที่เป็นบวกมาสักหนึ่งรูปแบบดังในสมการที่ 4

ตัวอย่างที่ 4

ถ้าข้อมูลของ Jane Smith ซึ่งเกิดวันที่ 12/12/1978 อาศัยอยู่ที่ Queenstown แต่งงานกับ John Doe และได้ทำการเปลี่ยนนามสกุลจาก Smith เป็น Doe และย้ายไปอยู่ที่ Christchurch เมื่อทำการส่งรายงานการเปลี่ยนที่อยู่ใหม่ไปด้วยนามสกุลใหม่ แต่ไม่ได้ระบุว่าเป็นคนเดียวกันกับ Jane Smith ด้วยการยืนยันในด้านอื่นๆ เช่น วันเกิด จึงทำให้เกิดการบันทึกข้อมูลใหม่เข้าไปในฐานข้อมูลว่าเป็น Jane Doe วันเกิดเป็น Unknown ที่อยู่แต่ก่อนคือ Queenstown และปัจจุบันอยู่ที่ถนน Large Place 2 ใน Christchurch ซึ่งจากข้อมูลดังกล่าวทำให้ไม่สามารถจำแนกได้ว่าเป็นข้อมูลที่มีความซ้ำซ้อนกัน จึงได้มีการกำหนดเงื่อนไขเพื่อใช้ในการบ่งบอกว่าเป็นข้อมูลที่มีความซ้ำซ้อนกันดังนี้ 1) ชื่อเหมือนกัน 2) มีการเปลี่ยนแปลงที่อยู่ 3) ที่อยู่ก่อนหน้านี้ของบุคคลต้องเหมือนกับที่อยู่ปัจจุบันในฐานข้อมูลเก่า ถ้าเป็นไปตามเงื่อนไขทั้ง 3 ให้ถือว่าเป็นข้อมูลที่ซ้ำซ้อนกัน

2.2.3) การจำแนกข้อมูลที่มีความคล้ายคลึงกัน

การจำแนกข้อมูลที่คล้ายคลึงกัน จะทำการจำแนกข้อมูลจากการคู่ของแคนดิเดตที่มีข้อมูลคล้ายกันในแอตทริบิวต์ชนิดเดียวกัน เช่น การเปรียบเทียบแอตทริบิวต์ที่เก็บข้อมูลเรื่องชื่อ ถ้าข้อมูลในชุดแรกคือ "John" ข้อมูลชุดที่สองคือ "Jonh" จะเห็นได้ว่าข้อมูลทั้งสองไม่เหมือนกันจริงๆ แต่เป็นข้อมูลที่คล้ายกัน ซึ่งแตกต่างกับการเปรียบเทียบในแบบอื่นๆ เพราะแบบอื่นจะเปรียบเทียบแอตทริบิวต์จากข้อมูลทั้ง 2 ชุดว่าเหมือนกันหรือไม่เท่ากัน

$$D^{*}\left(D(c),D(c')\right)=\left\{(d,d')\middle|\begin{matrix}descSim(d,d')>\theta desc\\ \&^{type(d)}=type(d')\end{matrix}\right\} \quad (3)$$

$$D^{\neq}\left(D(c),D(c')\right)=\left\{(d,d')\middle|d,d'\,contradictory@^{type(d)}=type(d')\right\} \quad (4)$$

$$sim(c,c')=\frac{\omega(D^{\approx}(c,c'))}{\omega(D^{\approx}(c,c'))+\omega(D^{\neq}(c,c'))} \quad (5)$$

$$\delta^{*}(c,c')=\begin{cases}c, & \text{if } \mathbf{sim}(c,c')>\theta, c \text{ is position } i+1\\ -1, & \text{otherwise}\end{cases} \quad (6)$$

สมการที่ 3 เป็นสมการที่ใช้ในการเปรียบเทียบชุดของข้อมูลที่เป็นชนิดเดียวกัน และมีค่าข้อมูลที่มีความคล้ายคลึงกันว่า เป็นข้อมูลที่มีค่าที่ความคล้ายคลึงกันมากกว่าค่าที่กำหนดไว้หรือไม่

สมการที่ 4 เป็นสมการที่ใช้ในการเปรียบเทียบชุดข้อมูลที่มีชนิดเดียวกัน แต่ค่าของข้อมูลไม่มีความคล้ายคลึงกันเลย

สมการที่ 5 เป็นสมการที่ใช้ในการหาค่าน้ำหนักความคล้ายกันของชุดข้อมูลที่ได้มาจากการเปรียบเทียบในสมการที่ 3 และ 4

สมการที่ 6 เป็นสมการที่ใช้ในการเปรียบเทียบค่า $sim(c,c')$ ว่ามากกว่าค่าความคล้ายเริ่มต้นที่ได้กำหนดไว้หรือไม่

จะขออธิบายถึงสัญลักษณ์ต่างๆ ที่ใช้ในสมการที่ 3, 4 และ 5 โดยกำหนดให้ D(c) = {d1,d2,….,dn} เป็นชุดของเดสคริปชันจากแคนดิเดต c และให้ ω เป็นฟังก์ชันที่ใช้คำนวนค่าน้ำหนักของข้อมูล θ_{desc} คือค่าความคล้ายกันเริ่มต้นของแอตทริบิวต์ ซึ่งกำหนดไว้ว่าต้องไม่ต่ำกว่า

ค่านี้ descSim เป็นฟังก์ชันที่ใช้คำนวณค่าความคล้ายกันของแอตทริบิวต์ว่ามีความคล้ายกันมากพอจะที่จะใช้การจำแนกข้อมูลที่มีความคล้ายคลึงกันหรือไม่ ถ้ามีค่าน้อยกว่าค่า θ desc จะไม่นำมาเปรียบเทียบ type(d) คือชนิดของเดสคริปชันหรือชนิดของแอตทริบิวต์ ที่ถูกเลือกมาเพื่อทำการเปรียบเทียบกัน D^* คือการเปรียบเทียบชุดของข้อมูลที่มีความคล้ายคลึงกัน และ $D^{\neq}$ คือการเปรียบเทียบชุดของข้อมูลที่แตกต่างกัน และในสมการที่ 6 ถ้าค่าน้ำหนักของชุดข้อมูลที่นำมาเปรียบเทียบมีค่ามากกว่าค่าความคล้ายเริ่มต้น (θ) แสดงว่าเป็นข้อมูลที่มีความซ้ำซ้อนกัน แต่ถ้าน้อยกว่าจะยังไม่รู้ว่าเป็นข้อมูลที่มีความซ้ำซ้อนกันหรือไม่

การจำแนกข้อมูลที่มีความคล้ายคลึงกันจะมีขั้นตอนในการจำแนกทั้งหมด 4 ขั้นตอนด้วยกัน

1. เปรียบเทียบจากแอตทริบิวต์ที่มีความคล้ายกันด้วยสมการที่ 3
2. เปรียบเทียบจากแอตทริบิวต์ที่มีความแตกต่างกันด้วยสมการที่ 4
3. ทำการหาค่าน้ำหนักของชุดข้อมูล $sim(c, c')$ ที่นำมาทำการจำแนกด้วยสมการที่ 5
4. นำค่าน้ำหนักของชุดข้อมูลไปเปรียบเทียบกับค่าความคล้ายเริ่มต้นที่กำหนดไว้ในตอนแรกด้วยสมการที่ 6

จะขอยกตัวอย่างการจำแนกข้อมูลที่ความคล้ายคลึงกัน เพื่อประกอบความเข้าใจจากตัวอย่างที่ 5 โดยจะทำการกำหนดให้ค่าความคล้ายเริ่มต้นมีค่าเท่ากับ 0.7

ตัวอย่างที่ 5

จะใช้ชุดข้อมูล 2 ชุดโดยกำหนดให้แคนดิเดตของ John Doe และ Jonathan Doe เป็น p1 และ p2 ตามลำดับ

D(p1) ={John , Doe , NZ Bank , Main Street 1 , Auckland ,
Large Place2 , Christchurch}

D(p2) ={Jonathan , Doe , National Bank , National Bank of New Zealand ,
Large Place2 , Christchurch}

เมื่อทำการเปรียบเทียบตามขั้นตอนที่ 1 และ 2 ตามลำดับจะได้

D^* D(D(p1),D(p2)) ={(Doe , Doe),(Large Place 2 , Large Place 2),
(Christchurch , Christchurch)}

$D^{\neq}$ D(D(p1),D(p2)) ={(John , Jonathan),

(NZ Bank , National Bank of New Zealand)}

จะเห็นได้ว่าชื่อถนนและชื่อเมืองที่ยังเหลืออยู่ใน p1 จะไม่นำมาใช้ในการพิจารณา เพราะว่าไม่มีคู่ของแคนดิเดตที่จะนำมาใช้ในการเปรียบเทียบความสันพันธ์ระหว่าง D(p1) และ D(p2) ซึ่งจะเห็นได้ว่าการขาดหายไปของข้อมูลไม่มีผลต่อการวัดความคล้ายคลึงกัน และในขั้นตอนสุดท้ายจะคำนวณค่าน้ำหนักของชุดข้อมูลจากสมการที่ 5 ซึ่งจะมีเมื่อแทนค่าตามสมการจะได้ sim(p1,p2) = 3/3+2 = 0.6 และค่าที่ได้ไปเปรียบเทียบกับค่าความคล้ายเริ่มต้นที่กำหนดไว้ที่ 0.7 จะเห็นได้ว่า 0.6 < 0.7 ซึ่งจะแสดงว่าเป็นข้อมูลที่ยังไม่สามารถจำแนกได้ว่าเป็นข้อมูลที่มีความซ้ำซ้อนกัน

2.3 การปรับปรุงประสิทธิภาพการสอบถาม

การปรับปรุงประสิทธิภาพการสอบถามใหม่ ถือว่าเป็นแนวทางที่สำคัญแนวทางหนึ่งที่ใช้ในการพัฒนาในการจัดการระบบฐานข้อมูล โดยในงานวิจัยจะทำการปรับปรุงประสิทธ์ภาพระบบจัดการฐานข้อมูลใหม่ จากแนวคิดการเขียนการสอบถามและการจัดการกับความสัมพันธ์ที่ใช้ในการจัดเก็บข้อมูลต่อไปนี้

2.3.1) การปรับปรุงประสิทธิภาพการสอบถามด้วยการเขียนการสอบถามใหม่

การเขียนการสอบถามใหม่เป็นส่วนหนึ่ง ที่สามารถช่วยเพิ่มประสิทธิภาพในการประมวลผลได้ ดังนั้นหากผู้ใช้เขียนการสอบถามที่ดี ก็จะทำให้การประมวลการสอบถามเป็นไปได้อย่างรวดเร็ว ดังนั้นจะขอแนะนำแนวทางในการเขียนการสอบถามใหม่ดังต่อไปนี้

1. หลีกเลี่ยงตัวดำเนินการทางคณิตศาสตร์ที่จะมาเป็นเงื่อนไขของการสอบถาม ดังต่อไปนี้

```
SELECT Ename, salary
FROM employee
WHERE salary/365>=1000
```

จากตัวอย่างดังกล่าว สามารถที่จะแก้ไขปัญหาได้โดยการสร้างโดยการสร้างความสัมพันธ์ชั่วคราว (Temporary Relation) เพื่อรองรับข้อมูลจากการประมวลผลที่มีการดำเนินการทางคณิตศาสตร์ และจึงสร้างดัชนีที่เหมาะสมกับแอตทริบิวต์นั้นๆ ต่อไป ซึ่งสามารถเขียนการสอบถามใหม่ได้ดังนี้

```
SELECT Ename, salary_temp
FROM employee
WHERE salary_temp >=1000
```

ซึ่งการเขียนการสอบถามดังกล่าว จะทำให้สามารถสร้างดัชนีการประมวลผลในแบบทรีได้ ซึ่งถือว่าเป็นการประมวลผลที่เร็วกว่าแบบลิเนียร์ ที่จะเป็นการประมวลผลที่เกิดขึ้นในการเขียนแบบสอบถามในตอนแรก

2. หลีกเลี่ยงการใช้ OR โดยไม่จำเป็น โดยให้ใช้ UNION ในการเขียนแบบสอบถามใหม่แทน
3. หลีกเลี่ยงการใช้ IN จากตัวอย่างนี้

```
SELECT eID FROM employee
WHERE department_no
IN (SELECT dept_no FROM department
WHERE manager_id='4806128')
```

จากตัวอย่างข้างต้น สามารถแก้ไข้ให้โดยการเขียนการสอบถามใหม่ได้ โดยการจอยความสัมพันธ์กันแทนซึ่งสามารถเขียนการสอบถามใหม่ได้ดังนี้

```
SELECT eID  FROM employee, department
WHERE manager_id='4806128' AND department_no=dept_no
```

4. หลีกเลี่ยงการใช้คำสั่ง DISTINCT เพราะจะทำเกิดการประมวลที่ช้า เนื่องจากจะทำการประมวลผลแบบโปรเจคท์ก่อนหนึ่งเสมอ
5. หลีกเลี่ยงการสอบถามซ้อนการสอบถามแบบมีเงื่อนไขเกี่ยวข้อง (Correlated Sub-query) หากไม่สามารถหลีกเลี่ยงได้ ควรสร้างความสัมพันธ์ชั่วคราว หากต้องมีการจอยกันระหว่างความสัมพันธ์ ควรเลือกแอตทริบิวต์ที่มีการสร้างดัชนีในการจอย

จากตัวอย่างการเขียนการสอบถามใหม่ ที่ได้แสดงไว้ในข้างต้นสามารถสังเกตได้ว่าการสร้างดัชนีเป็นอีกแนวทางหนึ่งที่สามารถเพิ่มประสิทธิภาพ ในการประมวลผลการค้นหาข้อมูลได้ ซึ่งจะขอกล่าวในขั้นถัดไป

2.3.2) การเพิ่มประสิทธิภาพแบบสอบถามด้วยการสร้างดัชนีข้อมูล

การสร้างดัชนีข้อมูลก็เป็นอีกทางเลือกหนึ่งในการเพิ่มประสิทธิภาพในการสอบถามได้ โดยเฉพาะหากข้อมูลมีจำนวนมาก โดยจะช่วยทำให้สามารถเข้าถึงข้อมูลได้เร็ว ซึ่งการสร้างดัชนีข้อมูลสามารถที่จะพิจารณาการสร้างได้เป็นกรณีดังต่อไปนี้

1. ปัจจัยที่เกี่ยวข้องกับการสอบถามและการทำรายการเปลี่ยนแปลงข้อมูลภายในฐานข้อมูล
    1.1 ลักษณะการเก็บไฟล์ของข้อมูล เช่น ไฟล์ที่ไม่เรียงข้อมูล ไฟล์ที่เรียงข้อมูล หรือไฟล์แบบแฮช ซึ่งเป็นส่วนหนึ่งในการพิจารณาการสร้างดัชนีได้
    1.2 การเรียงตัวของข้อมูลในแอตทริบิวต์ที่เกี่ยวข้องกับเงื่อนไขของการสอบถาม หรือการทำรายการเปลี่ยนแปลง ซึ่งควรสร้างดัชนีในแอตทริบิวต์ที่มีการเรียงตัวของข้อมูล
    1.3 หากจะต้องจอยกันระหว่างความสัมพันธ์ การสร้างดัชนีในแอตทริบิวต์ที่ใช้ในการจอย เป็นอีกปัจจัยหนึ่งในการพิจารณาการสร้างดัชนี
    1.4 พิจารณาการสร้างดัชนีจากการสอบถาม โดยเมื่อเทียบค่าคำนวณเชิงคณิตศาสตร์ การสร้างดัชนีจากการสอบถาม (Select) มีค่าคำนวณเชิงคณิตศาสตร์น้อยกว่าการเพิ่ม (Insertion) การลบ (Deletion) และการปรับปรุง (Update)
2. ปัจจัยที่เกี่ยวข้องกับความถี่ที่มีการใช้การสอบถามและการทำรายการเปลี่ยนแปลง

    หากพิจารณาการสร้างดัชนีจากการสอบถาม หรือการทำรายการเปลี่ยนแปลงสามารถพิจารณาการสร้างดัชนีได้จากการเข้าถึงข้อมูลในแอตทริบิวต์นั้นๆ จากความถี่ที่มีการใช้การสอบถามเพื่อเข้าถึงข้อมูล ว่ามีความถี่มากน้อยขนาดไหน โดยการเลือกสร้างดัชนีสามารถพิจารณาจากการเข้าถึงข้อมูลในแอตทริบิวต์นั้นๆ ที่มีความถี่ค่อนข้างสูงเป็นหลัก

    ในการสร้างดัชนีหากเป็นไปได้ ควรสร้างดัชนีตั้งแต่การออกแบบระบบในระยะแรก ซึ่งจะส่งผลดีในการจัดเก็บข้อมูลเป็นอย่างมาก หากไม่สามารถสร้างดัชนี

ดังที่กล่าวมาในเบื้องต้นได้ แต่เมื่อใช้ระบบไปได้ในระยะหนึ่ง การสร้างดัชนีจึงเป็นอีกทางเลือกหนึ่งในการเพิ่มประสิทธิภาพของการสอบถามได้

3. ปัจจัยที่เกี่ยวข้องกับเงื่อนไขเวลาของการทำงานของการสอบถาม และการทำรายการเปลี่ยนแปลง

   จากการเก็บข้อมูลความต้องการของผู้ใช้ การประมวลผลของคำสั่งเอสคิวแอลโดยมีเงื่อนไขของเวลาเกี่ยวข้องด้วย จำเป็นอย่างมากในการสร้างดัชนีเพื่อเพิ่มประสิทธิภาพในการประมวลผล แต่อย่างไรก็ตามหากมีการสร้างดัชนี แต่ไม่มีการใช้ดัชนีนั้นตามเงื่อนไขของเวลาจริง ก็อาจจะทำให้การประมวลผลช้าได้ ดังนั้นการทดลองสร้างดัชนีเพื่อใช้งานในระบบเล็กๆก่อนที่จะสร้างดัชนีจริงในระบบใหญ่ๆ ที่มีการจัดเก็บข้อมูลจริง จึงเป็นแนวคิดที่สิ่งสำคัญอีกประการหนึ่ง

4. ปัจจัยที่เกี่ยวข้องกับเงื่อนไขการไม่ซ้ำกันของค่าในแอตทริบิวต์

   การเก็บข้อมูลแบบไม่ซ้ำกันของข้อมูลในแอตทริบิวต์ ที่มีการเข้าถึงข้อมูลจากการสอบถามหรือจากการทำรายการเปลี่ยนแปลง เป็นอีกปัจจัยหนึ่งในการพิจารณาการสร้างดัชนี ซึ่งข้อมูลที่ไม่ซ้ำกันในแอตทริบิวต์สามารถสร้างดัชนี เพื่อเพิ่มประสิทธิภาพในการค้นหาให้เป็นแบบไบนารีได้

จากปัจจัยการสร้างดัชนีทั้งหมดที่ได้นำเสนอไป เป็นการเพิ่มประสิทธิภาพของการประมวลผลด้วยคำสั่งเอสคิวแอล ซึ่งยังมีปัจจัยที่ควรพิจารณาอีกอย่างหนึ่ง คือ ไม่ควรสร้างดัชนีในแอตทริบิวต์ที่มีการปรับปรุงข้อมูลบ่อยๆ ทั้งนี้เพราะว่าหากสร้างดัชนีในแอตทริบิวต์ที่กล่าวมา เมื่อมีการปรับปรุงข้อมูลทุกครั้ง จะทำให้มีการปรับปรุงไฟล์ดัชนีด้วยเสมอ

จากการสร้างดัชนีข้อมูลที่ได้กล่าวมาในข้างต้น จะเห็นได้ว่าการสร้างดัชนีข้อมูลจะไปเพิ่มประสิทธิภาพในการสอบถาม แต่การสร้างดัชนีข้อมูลก็มีข้อที่ต้องระวังอยู่คือ ไม่ควรสร้างดัชนีข้อมูลในแอตทริบิวต์ที่การเปลี่ยนแปลงข้อมูลบ่อยๆ เพราะเมื่อมีการปรับปรุงข้อมูล ดัชนีข้อมูลก็จะทำการปรับปรุงข้อมูลด้วยทุกครั้ง

จากทฤษฎีบทที่ได้กล่าวมาทั้งหมด จะถูกนำมาใช้ในการค้นหาความซ้ำซ้อนของข้อมูลในฐานข้อมูลซึ่งจะนำเสนอแนวทางในบทต่อไป