บทที่ 1

บทนำ

1.1 ปัญหาและที่มาของการศึกษา

ผลจากการสำรวจเว็บไซต์ในประเทศไทยจาก [1] พบว่าข้อมูลที่ถูกเก็บรวบรวมและเกี่ยวข้องกับ เว็บไซต์นั้นมี ตัวบ่งชี้บุคคล (Identifier) เช่น รหัสประจำตัวประชาชน หรือ ชื่อ-สกุล ถึง 64% และมีตัวบ่งชี้บุคคลทางอ้อม (quasi-identifier) เช่น ที่อยู่ อายุ หรือ อาชีพ 21% ไม่เพียงเท่านั้น 1 ใน 3 ข้อมูลเหล่านั้นยังถูกองค์กรที่รวบรวมข้อมูลนำไปตรวจสอบกับแหล่งข้อมูลต้นฉบับว่าเป็นข้อมูลจริงหรือไม่ (Verified Data) และจากการสำรวจยังพบว่า 53% ของนโยบายความเป็นส่วนตัวของข้อมูล (Privacy Policy) มีนโยบายที่อาจเผยแพร่ข้อมูลส่วนบุคคลโดยไม่มีการขออนุญาตจากเจ้าของข้อมูล และ 82% ยังเป็นนโยบายที่ไม่มีนโยบายความปลอดภัยของข้อมูล (Security Policy) ผลสรุปจากการสำรวจจะเห็นได้ว่าในประเทศไทยนั้น ยังไม่ได้ให้ความสนใจในเรื่องของปัญหาความเป็นส่วนตัวของข้อมูลเท่าที่ควร

จาก [1] ผู้เขียนให้ข้อคิดเห็นว่าปัญหาความเป็นส่วนตัวของข้อมูลสามารถแก้ไขได้โดยมีแนวทางหลักในการแก้ปัญหาจำนวน 3 แนวทาง คือ 1) ออกกฎหมายเพื่อแก้ปัญหาดังกล่าว 2) เพิ่มความรับผิดชอบของผู้เก็บข้อมูล 3) การนำเทคโนโลยีเข้ามาช่วยในการแก้ปัญหา ในวิทยานิพนธ์ฉบับนี้จะกล่าวถึงแนวทางที่ 3 เท่านั้น เนื่องจากแนวทางที่ 1 และ 2 นั้นอยู่นอกเหนือขอบเขตงานวิจัยของวิศวกรรมคอมพิวเตอร์

ในการแก้ปัญหาความเป็นส่วนตัวของข้อมูลโดยแนวทางที่ 3) วิธีการที่ง่ายที่สุดที่สามารถปฏิบัติได้ก่อนที่จะทำการเผยแพร่ข้อมูลคือ การลบข้อมูลที่เป็นตัวบ่งชี้บุคคลออกก่อนการเผยแพร่เพื่อให้ข้อมูลในแต่ละระเบียนนั้นไม่สามารถชี้ตัวบุคคลได้ วิธีการนี้ขึ้นอยู่กับการเลือกคอลัมน์ให้ถูกต้องว่าคอลัมน์ใดเป็นตัวบ่งชี้บุคคล

ในตารางที่ 1.1 แสดงตัวอย่างรายละเอียดโครงสร้างข้อมูลของแฟ้มที่ 2 ของข้อมูลสาธารณสุข 12 แฟ้มชื่อ “PAT” ซึ่งเป็นชุดข้อมูลมาตรฐานของการประกันสุขภาพของประเทศไทย ตัวอย่างข้อมูลของโครงสร้างข้อมูลดังกล่าวแสดงในตารางที่ 1.2

ตารางที่ 1.1 โครงสร้างข้อมูลของแฟ้มที่ 2

FIELD NAME	TYPE	LENGTH	DECIMAL	QUALIFICATION
HCODE	C	5	0	รหัสสถานพยาบาล (Left justified)
HN	C	9	0	หมายเลขประจำตัวผู้รับบริการ ควรใช้หมายเลขเดิมให้นานกว่า 5 ปี (Left justified)
CHANGWAT	C	2	0	ตามรหัสมหาดไทย
AMPHUR	C	2	0	ตามรหัสมหาดไทย
DOB	D	8	0	บันทึกวันเดือนปีเกิด ปีมีค่าเป็น ค.ศ.
SEX	C	1	0	1 หมายถึง เพศชาย 2 หมายถึง เพศหญิง
MARRIAGE	C	1	0	รหัสสภาพสมรส
OCCUPA	C	3	0	อาชีพ
NATION	C	2	0	สัญชาติ
PERSON_ID	C	13	0	รหัสประจำตัวประชาชนตามสำนักทะเบียนราษฎร์

ตารางที่ 1.2 ตัวอย่างการลบข้อมูลที่เป็นตัวบ่งชี้บุคคล

HCODE	HN	CHANGWAT	AMPHUR	DOB	SEX	MARRIAGE	OCCUPA	NATION	PERSONID
12	Null	48	1	2/1/1977	1	1	111	99	null
12	Null	48	1	5/2/1970	1	1	112	99	null
12	Null	48	1	9/6/1968	2	3	113	99	null

จากตารางที่ 1.1 ซึ่งเป็นโครงสร้างที่ใช้จริงในโรงพยาบาล ข้อมูลที่อยู่ในตารางเป็นค่ารหัส ซึ่งสามารถตรวจสอบค่าจริงได้จากภาคผนวก 1 จากโครงสร้างดังกล่าวนี้ในการระบุว่าคอลัมน์ใดเป็นตัวบ่งชี้บุคคลอาจขึ้นอยู่กับการวิเคราะห์ของแต่ละงาน ซึ่งในแฟ้มที่ 2 พิจารณาว่า PERSON_ID และ HN เป็นตัวบ่งชี้บุคคล เพราะค่าของทั้งสองคอลัมน์ในแต่ละระเบียนไม่มีการซ้ำกัน จากที่กล่าวมาถ้าทำการลบคอลัมน์ PERSON_ID และ HN ออกจะไม่สามารถชี้ตัวบุคคลได้ ฉะนั้นข้อมูลหลังการลบจึงสามารถเผยแพร่ได้ แต่นั้นหมายถึงการพิจารณาเพียงข้อมูลจากตารางนี้เท่านั้น ซึ่งวิธีการดังกล่าวยังไม่สามารถแก้ปัญหาการระบุตัวบุคคลอีกครั้งได้ (Re-Identification) การระบุตัวบุคคลอีกครั้งนั้นสามารถทำได้โดยการหาข้อมูลจากแหล่งอื่นมาทำการเปรียบเทียบกับข้อมูลในคอลัมน์อื่นที่ยังไม่ถูกลบและเป็นตัวบ่งชี้บุคคลทางอ้อมของตารางที่ถูกปกปิดตัวบ่งชี้บุคคลไปแล้ว โดยข้อมูลจากแหล่งอื่นนั้นอาจเป็นข้อมูลที่มีตัวบ่งชี้บุคคล และถ้า

ข้อมูลทั้งสองแหล่งมีการเหลื่อมซ้อนกัน (Overlap) อาจทำให้ข้อมูลที่ได้ทำการปกปิดไปแล้วนั้นสามารถกลับมาระบุตัวบุคคลอีกครั้งได้

จากการศึกษาข้อมูลสาธารณสุข 12 แฟ้มซึ่งเป็นชุดข้อมูลมาตรฐานของการประกันสุขภาพของประเทศไทย สามารถแสดงให้เห็นว่าหลังจากการตัดตัวบ่งชี้บุคคลออกแล้ว ข้อมูลยังสามารถถูกระบุตัวบุคคลอีกครั้งได้ ให้พิจารณาจากข้อมูลสมมุติต่อไปนี้ สมมุติให้หน่วยงานประกันสังคมต้องการเปิดเผยข้อมูลบางส่วนให้นักวิจัยนำข้อมูลมาวิเคราะห์เพื่อการเฝ้าระวังการทุจริตของผู้ประกันตน โดยข้อมูลอาจเป็นไปตามตารางที่ 1.3

ตารางที่ 1.3 ตัวอย่างข้อมูลจากประกันสังคม

ชื่อ-นามสกุล	ชื่อ นามสกุล ของผู้ประกันตนแล้วมีการเบิกเงิน
อาชีพ	อาชีพของผู้ประกันตน
วันที่เบิก	วันที่ยื่นเรื่องทำการเบิก
จำนวนเงิน	จำนวนเงินที่จ่าย

จากการพิจารณาจะเห็นว่าข้อมูลเหล่านี้ ดูเหมือนไม่ได้เป็นข้อมูลที่มีความสำคัญ การเปิดเผยข้อมูลเหล่านี้ มีความเป็นไปได้ที่จะเกิดขึ้น เมื่อกลับมาพิจารณาข้อมูลสาธารณสุข 12 แฟ้ม ทุกแฟ้มสามารถเชื่อมต่อกันโดยคอลัมน์ HN (Hospital number, หมายเลขประจำตัวผู้ป่วย) หรือคอลัมน์ AN (หมายเลขประจำตัวผู้ป่วยใน) แต่เนื่องจาก HN เป็นตัวบ่งชี้บุคคลทำให้ถูกลบค่าทิ้งไป ส่วน AN นั้นไม่ได้เป็นตัวบ่งชี้บุคคลเนื่องจากสามารถมีระเบียนหลายระเบียนที่มีค่า AN เหมือนกันได้ ดังเช่นในตัวอย่างรูปที่ 1.1 ฉะนั้นสามารถใช้ AN เพื่อเข้าถึงทุกแฟ้มที่เกี่ยวข้อง ตัวอย่างเช่นข้อมูลการวินิจฉัยโรค ข้อมูลการผ่าตัด ข้อมูลการรักษาเป็นต้น ฉะนั้นจากข้อมูลทั้งสองแหล่งถ้านำข้อมูลมาวิเคราะห์แล้วเชื่อมโยงจะสามารถเปรียบเทียบ ให้ AN กลับมาเป็นชื่อและนามสกุลได้ แล้วเชื่อมโยงไปถึงข้อมูลที่ต้องการได้

ต่อไปนี้เป็นตัวอย่างการวิเคราะห์และเชื่อมโยง ให้พิจารณาข้อมูลแฟ้มที่ 11 มาตรฐานแฟ้มข้อมูล การเงินซึ่งมีโครงสร้างตามตารางที่ 1.5 มีคอลัมน์ PTTYPE เป็นสิทธิการรักษาซึ่งมีค่ารหัสตามตารางที่ 1.4 จากการพิจารณาสามารถรู้ได้ว่า AN ที่มี PTTYPE เป็น AI, AJ, AK และ AL อาจมีข้อมูลตรงกับข้อมูลจากประกันสังคม และเนื่องจากแฟ้มที่ 11 มาตรฐานแฟ้มข้อมูลการเงินมีคอลัมน์ DATE, TOTAL (จำนวนเงินค่ารักษา) และ PAID (จำนวนเงินที่ผู้ป่วยจ่ายเอง) จะเห็นว่าถ้าข้อมูลจำนวนเงินที่มาจากประกันสังคมมีจำนวนเงินตรงกับ TOTAL ที่มาจากข้อมูล

แฟ้มที่ 11 เป็นไปได้ว่าเป็นการเบิกเงินจากประกันสังคมทั้งหมดหรือถ้าข้อมูลจำนวนเงินที่มาจากประกันสังคมตรงกับค่า TOTAL ลบด้วยค่า PAID แสดงว่าค่า PAID เป็นการจ่ายส่วนต่างที่ประกันสังคมไม่ได้ครอบคลุมถึง แล้วยังสามารถพิจารณาได้อีกว่าเกิดขึ้นในเดือนเดียวกันหรือไม่ กล่าวคือวันที่เบิกในข้อมูลประกันสังคมกับค่า DATE ในข้อมูลแฟ้มที่ 11 เป็นเดือนเดียวกันหรือไม่เพราะระบบการเบิกประกันสังคมนั้นต้องเบิกภายในเดือนเดียวกัน กับการเรียกชำระเงินจากทางสถานพยาบาล

ตารางที่ 1.4 ตัวอย่างการให้รหัส (ยังไม่ใช้เป็นทางการ)

รหัส	สิทธิการรักษา
A1	ชำระเงินเองโดยไม่มีสิทธิเบิกคืน
A2	ใช้สิทธิเบิกหน่วยงานต้นสังกัดราชการ
A3	สิทธิลดหย่อนประเภท ก. *
A4	สิทธิลดหย่อนประเภท ข. *
A5	สิทธิลดหย่อนประเภท ค. *
A6	สิทธิลดหย่อนประเภท ง. *
A7	ผู้ประกันตนตาม พ.ร.บ.ประกันสังคม
A8	กองทุนเงินทดแทน
A9	ประกันภัยรถ ตาม พรบ.บุคคลที่ 3
AA	เด็ก 0-12 ปี
AB	ผู้มีรายได้น้อย
AC	นักเรียน
AD	ผู้พิการ
AE	ทหารผ่านศึก
AF	ภิกษุ/ผู้นำศาสนา
AG	ผู้สูงอายุ
AH	บัตรชั่วคราว
AI	บัตรประกันสุขภาพ ประชาชนทั่วไป
AJ	บัตรประกันสุขภาพ อาสาสมัครสาธารณสุข
AK	บัตรประกันสุขภาพ ผู้นำชุมชน

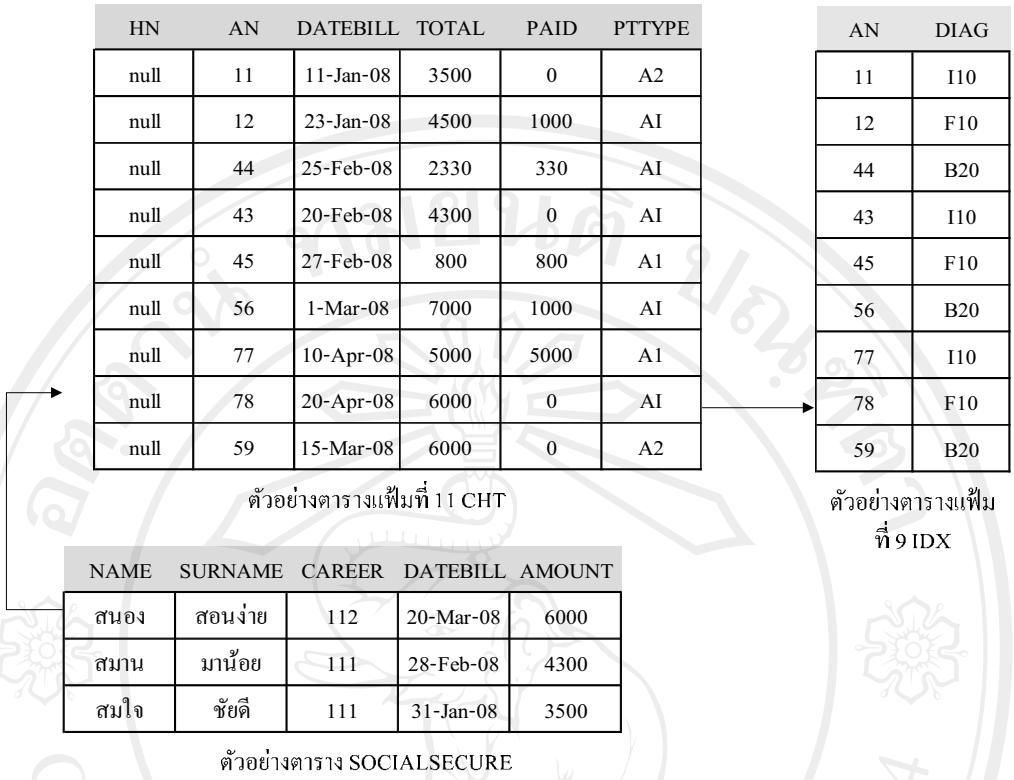
ตารางที่ 1.4 ตัวอย่างการให้รหัส (ต่อ)

AL	บัตรประกันสุขภาพ คนต่างด้าว
UC	บัตรประกันสุขภาพถ้วนหน้า

ตารางที่ 1.5 โครงสร้างข้อมูลแฟ้มที่ 11 มาตรฐานแฟ้มข้อมูลการเงิน

FIELD NAME	TYPE	LENGTH	DECIMAL	QUALIFICATION
HN	C	9	0	หมายเลขประจำตัวผู้รับบริการ ควรใช้หมายเลขเดิมให้นานกว่า 5 ปี (Left justified)
AN	C	9	0	หมายเลขประจำตัวผู้ป่วยใน ไม่ควรใช้หมายเลขนี้ซ้ำ (Left justified)
DATE	DATE	8	0	วันที่คิดค่ารักษา วันจำหน่าย หรือวันที่ผู้ป่วยเปลี่ยนสิทธิการรักษา บันทึก ปีในค่า ค.ศ.
TOTAL	N	7	0	จำนวนเงินค่ารักษารวมเป็นบาท ที่เรียกเก็บ
PAID	N	7	0	จำนวนเงินที่ผู้ป่วยจ่ายเอง ในกรณีที่โรงพยาบาลไม่ได้รับเงินไว้ = 0
PTTYPE	C	2	0	ชนิดการชำระเงิน ถ้าชำระเงินเอง = 10

หมายเหตุ คอลัมน์ AN, PTTYPE, DATE และ PAID เมื่อรวมกันพิจารณาเป็นตัวบ่งชี้บุคคลทางอ้อม (Quasi-Identifier) ซึ่งสามารถนำมาเปรียบเทียบกับตารางข้อมูลจากประกันสังคมเพื่อทำการระบุตัวบุคคลอีกครั้ง ในกรณีนี้คือการเปรียบเทียบค่า AN ให้เป็นชื่อและนามสกุลและเชื่อมโยงไปข้อมูลอื่นที่ต้องการได้



HN	AN	DATEBILL	TOTAL	PAID	PTTYPE
null	11	11-Jan-08	3500	0	A2
null	12	23-Jan-08	4500	1000	AI
null	44	25-Feb-08	2330	330	AI
null	43	20-Feb-08	4300	0	AI
null	45	27-Feb-08	800	800	A1
null	56	1-Mar-08	7000	1000	AI
null	77	10-Apr-08	5000	5000	A1
null	78	20-Apr-08	6000	0	AI
null	59	15-Mar-08	6000	0	A2

ตัวอย่างตารางแฟ้มที่ 11 CHT

AN	DIAG
11	I10
12	F10
44	B20
43	I10
45	F10
56	B20
77	I10
78	F10
59	B20

ตัวอย่างตารางแฟ้มที่ 9 IDX

NAME	SURNAME	CAREER	DATEBILL	AMOUNT
สนอง	สอนง่าย	112	20-Mar-08	6000
สมาน	มาน้อย	111	28-Feb-08	4300
สมใจ	ชัยดี	111	31-Jan-08	3500

ตัวอย่างตาราง SOCIALSECURE

รูปที่ 1.1 ตัวอย่างการระบุตัวตนอีกครั้งเพื่อเชื่อมโยงข้อมูลจากการวิเคราะห์ความสัมพันธ์

จากรูปที่ 1.1 กำหนดให้ ตัวอย่างตาราง SOCIALSECURE เป็นตัวอย่างตารางที่มาจากประกันสังคม ตัวอย่างตารางแฟ้มที่ 11 CHT เป็นตัวอย่างตารางแฟ้มการเงินที่มาจากข้อมูลสาธารณสุข 12 แฟ้มและตัวอย่างตารางแฟ้มที่ 9 IDX และเป็นตัวอย่างตารางแฟ้มการวินิจฉัยโรคที่มาจากข้อมูลสาธารณสุข 12 แฟ้ม ซึ่งในตารางนี้ค่า DIAG เป็นรหัสที่เรียกว่า ICD-10 ซึ่งเป็นรหัสชื่อโรค สามารถเข้าไปดูชื่อโรคได้ที่ [2] ในที่นี้ได้นำตัวอย่างมา 3 โรคคือ I10 = High blood pressure, F10 = ภาวะและพฤติกรรมแปรปรวนเนื่องจากการใช้ alcohol และ B20 = Human Immunodeficiency Virus (HIV)

สมมุติให้มีเพื่อนรวมงานของนายสนอง สอนง่าย ทราบว่านายสนอง ได้ไปโรงพยาบาลมาแล้วต้องการทราบว่านายสนองเป็นโรคอะไร ถ้าข้อมูลทั้ง 3 ตารางในรูปที่ 1 สามารถค้นหาได้จาก Internet เขาอาจใช้หลักการระบุตัวบุคคลอีกครั้งเพื่อทำให้ทราบว่านายสนองเป็นโรคอะไรได้ โดยมีหลักการวิเคราะห์คือ พิจารณาตาราง SOCIALSECURE สนอง สอนง่ายมีค่า AMOUNT เท่ากับ 6000 ซึ่งตรงกับระเบียนที่มีค่า AN เท่ากับ 56, 78 และ 59 ของตัวอย่างตารางแฟ้มที่ 11 CHT โดยระเบียนที่ค่า AN เท่ากับ 78 และ 59 มีค่า TOTAL เท่ากับ 6000 ส่วนระเบียนที่ค่า AN

เท่ากับ 56 มีค่า TOTAL ลบ PAID เท่ากับ 6000 จาก 3 ระเบียนนี้ มีเพียงระเบียนที่ค่า AN เท่ากับ 78 และ 56 ที่มี PTTYPE เป็น AI ซึ่งแสดงว่าเป็นการเบิกเงินจากประกันสังคมและมีเพียงระเบียนที่ค่า AN เท่ากับ 78 เท่านั้นที่ DATEBILL เป็นเดือนเดียวกันกับ DATABILL ของ สนอง สอนง่าย

จากที่กล่าวมานั้นแสดงให้เห็นว่า สนองสอนง่ายนั้นมีค่า AN เท่ากับ 78 และเมื่อนำค่า AN เท่ากับ 78 นี้ไปเชื่อมโยงในตาราง IDX จะได้ว่า สนอง สอนง่ายนั้นเป็นโรค F10 ซึ่งคือโรคภาวะและพฤติกรรมแปรปรวนเนื่องจากการใช้ Alcohol ฉะนั้นจึงสามารถสรุปได้ว่า การปกปิดข้อมูลโดยการลบตัวบ่งชี้บุคคลอย่างเดียวไม่สามารถแก้ปัญหาการระบุตัวบุคคลอีกครั้งได้

เทคนิค *k*-Anonymity

อย่างไรก็ตามจาก [3] และ [4] ผู้เขียนได้นำเสนอวิธีการที่เรียกว่า *k*-Anonymity ซึ่งสามารถแก้ปัญหาการระบุตัวบุคคลอีกครั้งได้ โดยการทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity มีหลักการพื้นฐานอยู่ว่าให้ทำการเปลี่ยนแปลงค่าของข้อมูลในแต่ละระเบียน เมื่อเปลี่ยนแปลงค่าของข้อมูลแล้วต้องมีระเบียนที่มีตัวบ่งชี้ทางอ้อม (Quasi-Identifier) เหมือนกันไม่ต่ำกว่า k ระเบียน ในการเปลี่ยนแปลงค่าของข้อมูลในแต่ละระเบียนอาจทำการกำหนดขั้นตอนการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้น (Hierarchy) ให้กับค่าของข้อมูลในแต่ละคอลัมน์ที่เป็นตัวบ่งชี้บุคคลทางอ้อมไว้ก่อน แล้วจึงทำการเปลี่ยนค่าของแต่ละระเบียนไปทีละคอลัมน์จากระดับปัจจุบันไประดับที่สูงขึ้น จนมีคุณสมบัติ *k*-Anonymity จึงหยุดทำการเปลี่ยนแปลง ตัวอย่างในการกำหนดขั้นตอนการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นเช่นคอลัมน์ DATE ใน ระดับที่ 0 เป็นข้อมูลที่มีหน่วยเป็น วัน/เดือน/ปี ระดับที่ 1 เป็นข้อมูลที่มีหน่วยเป็น สัปดาห์/เดือน/ปี และระดับที่ 2 เป็นข้อมูลที่มีหน่วยเป็น เดือน/ปี เป็นต้น ในการทำข้อมูลให้มีคุณสมบัติ *k*-Anonymity สามารถเลือกทำได้อีก 2 วิธีคือ วิธี Alternative กับวิธี Full-Domain วิธี Alternative คือในแต่ละระเบียนสามารถเลือกตัวบ่งชี้บุคคลทางอ้อมตัวใดก็ได้ ที่อยู่ในระดับเดียวกันในการทำการเปลี่ยนแปลง ส่วนการทำ Full-Domain เป็นการเลือกคอลัมน์ที่เป็นตัวบ่งชี้บุคคลทางอ้อมแล้วทำการเปลี่ยนทุกระเบียนในคอลัมน์นั้นตามขั้นตอนการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นที่กำหนดไว้ ซึ่งในวิทยานิพนธ์นี้สนใจในกรณีของ Full-Domain

จากตัวอย่างในรูปที่ 1.2 ตารางข้อมูล ให้คอลัมน์ AN, DATEBILL, TOTAL และ PAID เป็นตัวบ่งชี้ทางอ้อมสามารถเปลี่ยนแปลงค่าในคอลัมน์ DATEBILL ตามการเปลี่ยนแปลงตามลำดับขั้นของ DATEBILL โดยเปลี่ยนค่าที่อยู่ในตารางข้อมูลซึ่งอยู่ในระดับที่ 0 ให้เป็นระดับ

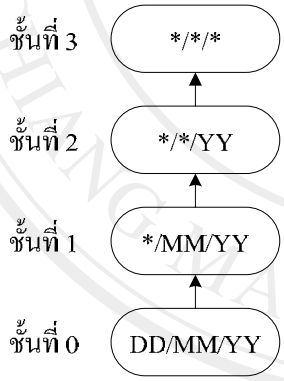
ที่ 1 ส่วนค่าในคอลัมน์อื่นก็สามารถทำได้โดยวิธีการเดียวกัน ในการทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity จากในตัวอย่างได้กำหนดให้ $k = 2$ แสดงว่าต้องทำการเปลี่ยนแปลงค่าในแต่ละคอลัมน์จนมีตัวบ่งชี้ทางอ้อมซ้ำกันเป็นจำนวน 2 ระเบียนขึ้นไป จึงจะถือว่าข้อมูลมีคุณสมบัติ *2*-Anonymity

AN	DATEBILL	TOTAL	PAID	PTTYPE
11	11 ม.ค. 2008	4500	0	A2
12	23 ม.ค. 2008	4500	1000	AI
44	25 ก.พ. 2008	2330	330	AI
43	20 ก.พ. 2008	2300	0	AI
45	27 ก.พ. 2008	2800	2800	A1
56	01 มี.ค. 2008	6000	1000	AI
77	10 เม.ย. 2008	7000	7000	A1
78	20 เม.ย. 2008	7000	0	AI
59	15 มี.ค. 2008	6000	0	A1

ตารางข้อมูล

AN	DATEBILL	TOTAL	PAID	PTTYPE
1*	* ม.ค. 2008	4***	*	A2
1*	* ม.ค. 2008	4***	*	AI
4*	* ก.พ. 2008	2***	*	AI
4*	* ก.พ. 2008	2***	*	AI
4*	* ก.พ. 2008	2***	*	A1
5*	* มี.ค. 2008	6***	*	AI
7*	* เม.ย. 2008	7***	*	A1
7*	* เม.ย. 2008	7***	*	AI
5*	* มี.ค. 2008	6***	*	A1

ตารางข้อมูลหลังมีการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นที่มีคุณสมบัติ *2*-Anonimity



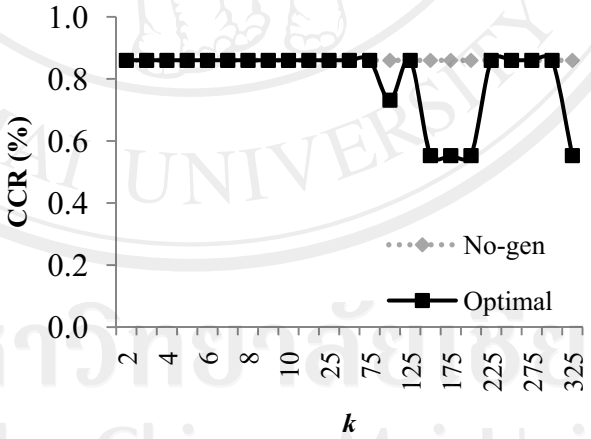
ขั้นตอนการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นของคอลัมน์ DATEBILL

รูปที่ 1.2 ตัวอย่างการทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity

การนำข้อมูลที่มีคุณสมบัติ *k*-Anonymityไปใช้งานต่อนั้น มีปัญหาอยู่ว่าการทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity นั้นสามารถกำหนดระดับของการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นของแต่ละคอลัมน์ที่เป็นตัวบ่งชี้ทางอ้อม หรือ ระดับการเจนเนอรัลไลเซชัน (Generalization Level) ได้หลายรูปแบบ ตัวอย่าง ระดับการเจนเนอรัลไลเซชัน เช่น (AN:2, DATEBILL:1, TOTAL:3, PAID:5, PTTYPE:1) จากรูปที่ 1.2 ถ้าเปลี่ยนแปลงข้อมูลของคอลัมน์

DATEBILL ให้เพิ่มขึ้นไปอีก 1 ระดับ (ระดับการเจนเนอรัลไลเซชัน เท่ากับ (AN:2, DATEBILL:2, TOTAL:3, PAID:5, PTTYPE:1))ข้อมูลก็ยังคงคุณสมบัติ *k*-Anonymity อยู่ ทำให้ยากต่อการเลือกว่าจะใช้ข้อมูลที่มีคุณสมบัติ *k*-Anonymity ที่เกิดจากระดับการเจนเนอรัลไลเซชันที่เท่าไหร่จึงจะเหมาะสมกับงานที่ต้องนำข้อมูลนี้ไปใช้

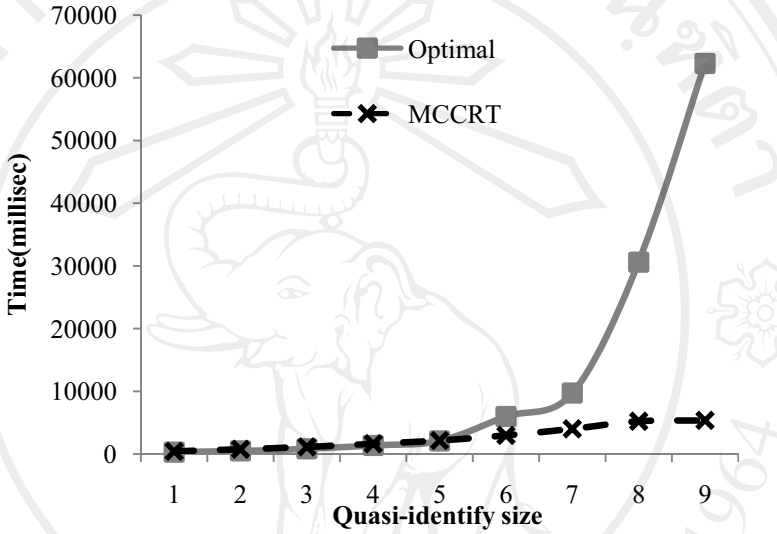
งานวิจัย [5] ได้เสนอวิธีการขั้นตอนวิธีที่เหมาะที่สุด (Optimal Algorithm) ซึ่งเป็นการหาระดับการเจนเนอรัลไลเซชันที่มีผลกระทบต่อคุณภาพข้อมูลน้อยที่สุดโดยการวัดการบิดเบือนของข้อมูล (Distortion Ratio: C_{GM}) และตัววัดคุณภาพข้อมูลสำหรับการจำแนกแบบความสัมพันธ์ (Frequency-based Classification: C_{FCM}) โดยจะเลือกระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ k-Anonymity ที่ให้ค่า C_{GM} และค่า C_{FCM} น้อยที่สุด ข้อมูลที่ทำการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นด้วยระดับการเจนเนอรัลไลเซชันที่ได้จากวิธีนี้เมื่อนำไปทำเหมืองข้อมูลที่เรียกว่าการจำแนกแบบความสัมพันธ์ (Associative Classification) ปรากฏว่าได้ผลการทดลองตามรูปที่ 1.3 ซึ่งเป็นการวัดค่าอัตราความถูกต้องในการจำแนก (Classification Correction Rate: *CCR*) เปรียบเทียบระหว่างการไม่เปลี่ยนแปลงค่าของข้อมูล (No-gen) กับเมื่อเปลี่ยนแปลงค่าของข้อมูลใช้ขั้นตอนวิธีที่เหมาะที่สุด (Optimal)



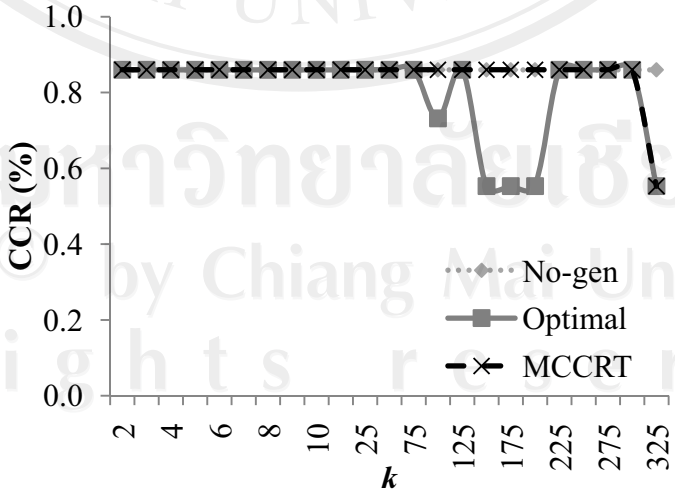
รูปที่ 1.3 ผลกระทบของค่า *k* ต่อค่า *CCR* [5]

แต่ปัญหาการแปลงข้อมูลเพื่อที่จะได้มาซึ่งฐานข้อมูลที่ดีที่สุดโดยยังคงคุณสมบัติ *k*-Anonymity และมีผลกระทบต่อคุณภาพข้อมูลน้อยที่สุด เป็นปัญหาเอ็นพีแบบยาก (NP-hard) โดยขั้นตอนวิธีนี้มีระดับความซับซ้อนเชิงคำนวณที่เป็นเลขชี้กำลัง (Exponential) ใน [6] จึงได้มีการคิดค้น MCCRT (Minimum Classification Correction Rate Transformation

algorithm) ที่มีขั้นตอนวิธีเป็นแบบศึกษาสำนึกที่จะแก้ปัญหานี้ได้อย่างมีประสิทธิภาพ และประสิทธิผล โดยสามารถทำงานได้อย่างรวดเร็วเมื่อมีจำนวนข้อมูลมากขึ้นและยังมีการทำงานที่เร็วรวดเร็วกว่า ซึ่งความซับซ้อนเชิงคำนวณคือ $O(n \log n)$ ในงานวิจัยนี้จึงให้สนใจที่จะปรับปรุงการทำงานของ MCCRT



รูปที่ 1.4 ผลกระทบของจำนวนคอลัมน์ที่เหลื่อมซ้อนกันต่อเวลากระทำการ [6]



รูปที่ 1.5 ผลกระทบค่า k ต่อค่า CCR [6]

ลำดับการทำงานของขั้นตอนวิธี MCCRT

MCCRT มีหลักการทำงานคือการนำอัตราความถูกต้องในการจำแนก (Classification Correction Rate: *CCR*) มาใช้ในการเรียงลำดับเพื่อนำไปเปลี่ยนค่าของแต่ละคอลัมน์โดย จะเรียงจากคอลัมน์ที่มี *CCR* น้อยไปหาคอลัมน์ที่มี *CCR* มาก โดยในกรณีที่คอลัมน์มี *CCR* เท่ากันจะเรียงจากค่าความสูงของขั้นตอนการเปลี่ยนแปลงตามลำดับขั้นของคอลัมน์นั้นๆ จากมากไปหาน้อย ซึ่งเรียกการเรียงตัวกันของ คอลัมน์นี้ว่า *S* (Sequent of novel heuristic algorithm MCCRT) ต่อมาทำการเปลี่ยนแปลงค่าตามลำดับขั้นของคอลัมน์จากตัวแรกของ *S* โดยเริ่มจากระดับที่ 0 ขึ้นไปจนถึงระดับบนสุดแล้วจึงเปลี่ยนแปลงค่าในตัวถัดไปของ *S* ทำซ้ำในรูปแบบเดียวกันจนข้อมูลมีคุณสมบัติ *k*-Anonymity ผลลัพธ์ที่ได้นี้คือระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity จากการทดลองปรากฏว่าเมื่อนำข้อมูลที่ได้จากการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นในระดับการเจนเนอรัลไลเซชันนี้ไปหา Associative Classification ปรากฏว่า Associative Classification ที่ได้มีความแม่นยำที่มากกว่า (ดังรูปที่ 1.5) และมีความเร็วที่มากกว่า (ดังรูปที่ 1.4) ระดับการเจนเนอรัลไลเซชันที่ได้จากแบบขั้นตอนวิธีที่เหมาะที่สุด สาเหตุที่มีความแม่นยำที่มากกว่าเนื่องจาก MCCRT ใช้ค่า *CCR* มาคิดโดยตรงแต่ขั้นตอนวิธีที่เหมาะที่สุดคิดจากค่า C_{FCM} และสาเหตุที่ MCCRT มีความเร็วที่มากกว่ามาจาก การทดสอบคุณสมบัติ *k*-Anonymity ของ MCCRT เป็นการหาระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity เพียงแค่ค่าแรกที่เจอ แต่ของขั้นตอนวิธีที่เหมาะที่สุดต้องทดสอบคุณสมบัติ *k*-Anonymity ในทุกระดับการเจนเนอรัลไลเซชันที่เป็นไปได้เพื่อหาระดับการเจนเนอรัลไลเซชันที่มีคุณสมบัติ *k*-Anonymity ทุกค่า

การประมวลผลแบบหนึ่งการประมวลผลต่อหนึ่งงาน (One-time fashion) ของขั้นตอนวิธี MCCRT

อย่างไรก็ตาม ในการหาระดับการเจนเนอรัลไลเซชันของ MCCRT นั้น ต้องทำการอ่านข้อมูลทั้งหมดหนึ่งครั้ง จึงได้ผลลัพธ์ *S* ออกมาและทำการทดสอบคุณสมบัติ *k*-Anonymity ไปจนได้ค่าระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity ถ้าในภายหลังมีการเพิ่มข้อมูลเข้ามาใหม่จำนวนหนึ่งแล้วต้องการหาค่าระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity ใหม่ ต้องทำการอ่านข้อมูลทั้งหมดและทำการทดสอบ คุณสมบัติ *k*-Anonymity ตั้งแต่เริ่มอีกครั้ง แสดงว่าถ้ามีการเพิ่มข้อมูลใหม่ 10 ครั้งต้อง อ่านข้อมูลทั้งหมดและทำการทดสอบ คุณสมบัติ *k*-Anonymity ตั้งแต่เริ่ม ใหม่ 10 ครั้ง ซึ่งเป็นการเสียเวลาอย่างมาก ใน

วิทยานิพนธ์นี้ได้หาวิธีที่ไม่ต้องอ่านข้อมูลใหม่ทั้งหมดและลดการจำนวนครั้งการทดสอบคุณสมบัติ *k*-Anonymity ให้น้อยลงโดยอ่านข้อมูลที่เข้ามาใหม่และข้อมูลเก่าบางส่วนและทำการทดสอบคุณสมบัติ *k*-Anonymityที่น้อยครั้งลง กล่าวคือไม่ต้องทำการทดสอบตั้งแต่เริ่มต้น ซึ่งในการทำ MCCRT ครั้งแรกได้เก็บข้อมูลระดับการเจนเนอรัลไลเซชัน, ข้อมูลที่ถูกแปลงค่า และ ค่าอัตราความแม่นยำในการจำแนก ซึ่งการนำข้อมูลเหล่านี้มาทำให้เป็นปัจจุบันย่อมเร็วกว่าการคำนวณใหม่และผลลัพธ์ที่ได้เหมือนการอ่านและทำการทดสอบ คุณสมบัติ *k*-Anonymityใหม่ทั้งหมด

1.2 แนวทางการแก้ปัญหา

ในวิทยานิพนธ์นี้จะเสนอขั้นตอนวิธีที่สามารถรักษาความเป็นส่วนตัวของข้อมูลเมื่อมีการเพิ่มขึ้นของข้อมูลในสถานการณ์ที่การวิเคราะห์ข้อมูลเหล่านั้นคือการจำแนกแบบความสัมพันธ์ โดยจะเริ่มจากการวิเคราะห์สถานการณ์ที่จะเกิดขึ้นกับข้อมูลในแง่ของความเป็นส่วนตัวในรูปแบบของ *k*-Anonymity และในแง่ของคุณภาพข้อมูล ซึ่งการวิเคราะห์สถานการณ์ในลักษณะนี้เป็นหลักการพื้นฐานที่ใช้ในการพัฒนาขั้นตอนวิธีแบบเพิ่มขึ้นในงานวิจัย [7] เมื่อเสร็จสิ้นการวิเคราะห์และศึกษาแล้วจะทำการพัฒนาขั้นตอนวิธีแปลงค่าข้อมูลแบบเพิ่มขึ้น โดยขั้นตอนวิธีที่จะพัฒนาขึ้นจะต้องมีความซับซ้อนเชิงคำนวณต่ำกว่าการนำขั้นตอนวิธี MCCRT ที่มีลักษณะเป็นหนึ่งการประมวลผลต่อหนึ่งงาน

จากการศึกษาการทำงานของ MCCRT ใน [6] และแนวทางการแก้ปัญหาของ IncSpan ใน [7] พบว่าเมื่อมีการเพิ่มข้อมูลเข้าใหม่อาจทำให้เกิดการเปลี่ยนแปลงระดับการเจนเนอรัลไลเซชันเพื่อให้ยังคงคุณสมบัติ *k*-Anonymity ซึ่งมีสาเหตุมาจากค่าของข้อมูลที่เพิ่มเข้ามาใหม่มีค่าซ้ำค่าเดิมหรือมีค่าที่ไม่เคยปรากฏมาก่อนหรือเกิดการเปลี่ยนแปลงลำดับใน *S* (Sequent of novel heuristic algorithm MCCRT) เพราะข้อมูลที่เพิ่มเข้ามาใหม่ทำให้ค่า *CCR* เปลี่ยนแปลงไป วิธีการในการแก้ปัญหานั้นสามารถทำได้โดยการทดสอบข้อมูลเพิ่มขึ้นที่มีคุณสมบัติทำให้เกิดการเพิ่มขึ้นหรือการลดลงของระดับการเจนเนอรัลไลเซชันจากสาเหตุกล่าวมาโดยไม่ต้องเริ่มกระบวนการตามขั้นตอนวิธี MCCRT ใหม่ตั้งแต่เริ่มต้น

1.3 สรุปสาระสำคัญจากเอกสารที่เกี่ยวข้อง

บวรศักดิ์ และคณะ [1] ได้ชี้ให้เห็นว่าองค์กรที่เก็บรวมรวมข้อมูลส่วนบุคคลบนระบบอินเทอร์เน็ตในประเทศไทยยังไม่ให้ความสำคัญกับปัญหาความเป็นส่วนตัวของข้อมูลเท่าไหร่นัก และในประเทศไทยกฎหมายที่ใช้ในการปกป้องข้อมูลส่วนบุคคลยังอยู่ในขั้นตอนการร่างกฎหมาย

และยังคงต้องปรับปรุงอยู่ ฉะนั้นในปัจจุบันยังมีความจำเป็นที่งานทางด้านวิศวกรรมคอมพิวเตอร์ควรทำการศึกษาและพัฒนาเทคนิคในการป้องกันข้อมูลส่วนบุคคลเพื่อการเผยแพร่สู่สาธารณะ

Sweeney [3] ได้นำเสนอเทคนิค k-Anonymity ซึ่งเป็นเทคนิคการรักษาความเป็นส่วนตัวของข้อมูล งานวิจัยนี้ถูกงานวิจัยจำนวนมากนำเทคนิค k-Anonymity ไปประยุกต์ใช้ เนื่องจากเป็นเทคนิคที่เข้าใจได้โดยง่ายและสามารถนำไปใช้อย่างมีประสิทธิภาพในการรักษาความเป็นส่วนตัวของข้อมูล

Sweeney [4] ได้นำเสนอเทคนิค k-Anonymity มาใช้ในการรักษาความเป็นส่วนตัวของข้อมูล โดยวิธีการแปลงข้อมูลให้อยู่ในค่าที่ทั่วไป และวิธีการปิดบังข้อมูล และยังได้นำเสนอการกำหนดคุณภาพข้อมูลแบบทั่วไป คือ Precision metric ซึ่งจะใช้กำหนดคุณภาพข้อมูลแบบทั่วไปได้ดีเฉพาะกรณีที่โครงสร้างโดเมนในแต่ละแอตทริบิวต์ที่ถูกแปลงไม่แตกต่างกันมาก

ณัฐพลและคณะ [5] ได้เสนอขั้นตอนวิธีเพื่อใช้ในการเลือกระดับเจนเนอรัลไลเซชันโดยใช้การวัดค่าการบิดเบือนของข้อมูล (Distortion Ratio: C_{GM}) และตัววัดคุณภาพข้อมูลสำหรับการจำแนกแบบความสัมพันธ์ (Frequency-based Classification: C_{FCM}) ของข้อมูลที่ถูกเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นที่มีค่าน้อยที่สุด วิธีการนี้ลดค่าความซับซ้อนเชิงคำนวณในการหาระดับเจนเนอรัลไลเซชันเพื่อนำข้อมูลไปหาค่า Associative Classification

ณัฐพลและคณะ [6] ได้เสนอขั้นตอนวิธีแบบศึกษาสำนึก MCCRT เพื่อใช้ในการหาระดับการเจนเนอรัลไลเซชันที่เหมาะสมกับการนำข้อมูลไปใช้ในการหา Associative Classification โดยได้ลดจำนวนการทดสอบคุณสมบัติ *k*-Anonymity ลงโดยใช้ค่า *CCR* ของแต่ละคอลัมน์มาใช้ในการกำหนดเส้นทางการทดสอบคุณสมบัติ *k*-Anonymity ระดับเจนเนอรัลไลเซชันที่ได้มีความแม่นยำมากกว่าขั้นตอนวิธีที่เหมาะที่สุดและมีค่าความซับซ้อนเชิงคำนวณเพียง $O(n \log n)$

Cheng และคณะ [7] ได้เสนอวิธีการในการหารูปแบบลำดับเมื่อมีการเพิ่มเข้ามาของข้อมูลโดยไม่ต้องทำการอ่านข้อมูลใหม่ทั้งหมด โดยการสร้างโครงสร้างข้อมูลรูปต้นไม้ขึ้นมาแทนการเก็บค่าแบบตาราง และยังเสนอแนวคิดในการแก้ปัญหาโดยการแยกกรณีที่เกิดขึ้นได้เป็น 6 กรณี และทำการแก้ปัญหาไปทีละกรณีจนทำให้สามารถแก้ปัญหาหลักที่ต้องการได้

Wong และคณะ [8] ได้เสนอเทคนิคในการแก้ปัญหาเพิ่มเติมที่ *k*-Anonymity ไม่สามารถแก้ไขได้ คือการมีค่าคลาสเดียวกันของกลุ่มข้อมูลที่ถูกเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้น โดยการเพิ่มค่า α ที่ใช้เป็นตัวชี้วัดค่าคลาสที่ซ้ำกันในกลุ่มข้อมูลเฝ้าระวัง

1.4 วัตถุประสงค์ของการศึกษา

1.4.1 เพื่อวิเคราะห์การเปลี่ยนแปลงของข้อมูลเมื่อมีการเพิ่มขึ้นของข้อมูลในแง่ของความเป็นส่วนตัวข้อมูลและคุณภาพข้อมูลสำหรับการจำแนกแบบสัมพันธ์

1.4.2 เพื่อพัฒนาโครงสร้างข้อมูลที่สามารถลดค่าความซับซ้อนเชิงคำนวณในการหาระดับเจนเนอรัลไลเซชันเมื่อมีการเพิ่มขึ้นของข้อมูล

1.4.3 เพื่อพัฒนาขั้นตอนวิธีในการหาระดับเจนเนอรัลไลเซชันตามขั้นตอนวิธี MCCRT (Minimum Classification Correction Rate Transformation) และขั้นตอนวิธีที่เหมาะที่สุด ซึ่งเมื่อมีการเพิ่มขึ้นของข้อมูล จะมีค่าความซับซ้อนในเชิงคำนวณน้อยกว่าและไม่จำเป็นต้องทำการประมวลผลข้อมูลใหม่ทั้งหมด

1.5 ประโยชน์ที่ได้รับจากการศึกษา เชิงทฤษฎี และ/หรือ เชิงประยุกต์

องค์กรใดๆ ก็ตามที่มีการรวบรวมข้อมูลส่วนบุคคลต้องการเผยแพร่ข้อมูลซึ่งจำเป็นต้องมีการรักษาความเป็นส่วนตัวของข้อมูลและข้อมูลเหล่านี้จะเพิ่มขึ้นเรื่อยๆ เช่นการเผยแพร่ข้อมูลการรักษาโรคของผู้ป่วยในโรงพยาบาลซึ่งจะเพิ่มขึ้นเรื่อยๆในทุกๆ เดือน สามารถใช้ผลลัพธ์จากวิทยานิพนธ์นี้ในการรักษาความเป็นส่วนตัวของข้อมูลได้อย่างมีประสิทธิภาพ

1.6 ขอบเขตการทำวิจัย

1.6.1 การประเมินประสิทธิภาพของขั้นตอนวิธีที่พัฒนาขึ้นจะใช้เวลาจริง (วินาที) และความซับซ้อนเชิงคำนวณ

1.6.2 ข้อมูลที่จะใช้ทดสอบประสิทธิภาพของขั้นตอนวิธีที่พัฒนาขึ้นจะใช้ข้อมูลจาก UCI Repository [9] ซึ่งเป็นแหล่งรวบรวมข้อมูลที่มักจะถูกใช้ในการวิจัยเกี่ยวการทำเหมืองข้อมูล และข้อมูลสังเคราะห์

1.7 วิธีการทำวิจัย

1.7.1 ศึกษาทฤษฎีที่เกี่ยวข้องกับการรักษาความเป็นส่วนตัวของข้อมูล

1.7.2 ศึกษาทฤษฎีที่เกี่ยวข้องกับการสร้างแบบจำลองคุณภาพข้อมูลที่จำเป็นต่อการจำแนกแบบกฎความสัมพันธ์

1.7.3 ศึกษาทฤษฎีที่เกี่ยวข้องกับการแปลงฐานข้อมูลให้มีคุณภาพข้อมูลสูงในสถานการณ์การวิเคราะห์ข้อมูลสำหรับการจำแนกแบบความสัมพันธ์

1.7.4 ศึกษาทฤษฎีที่เกี่ยวข้องกับการแปลงฐานข้อมูลเมื่อมีการเพิ่มขึ้นมาของข้อมูล

1.7.5 ออกแบบการทดลองและพัฒนาแบบจำลองข้อมูลที่ใช้ในการแปลงฐานข้อมูลเมื่อมีการเพิ่มขึ้นมาของข้อมูล

1.7.6 ทำการทดลองและพัฒนาขั้นตอนวิธีการแปลงฐานข้อมูลเมื่อมีการเพิ่มขึ้นมาของข้อมูลตามขั้นตอนวิธี MCCRT และขั้นตอนวิธีที่เหมาะที่สุด

1.7.7 วิจารณ์ สรุปผลการทำวิจัย จัดทำและเสนอรายงานวิทยานิพนธ์

1.8 เครื่องมือในการพัฒนา

ในวิทยานิพนธ์นี้การพัฒนาขั้นตอนวิธีทั้งหมดจะใช้เครื่องมือในการพัฒนาดังนี้

1.8.1 ฮาร์ดแวร์

- 2.0 GHz Intel core 2 Duo notebook
- 4 Gigabytes main memory

1.8.2 วอฟต์แวร์

- Microsoft Windows 7
- JDK 1.6
- NetBean 6.7.1
- WEKA Data Mining Software