

การพัฒนาระบบการปกปิดตัวตนสำหรับข้อมูลสุขภาพ

Development of Deidentification Tool for Health Information

อดุลย์ นามาลี (Adull Namalee)* ดร.วนิดา แก่นอากาศ (Dr.Wanida Kanarkard)**

บทคัดย่อ

การปรับปรุงคุณภาพและการลดค่าใช้จ่ายในการดูแลผู้ป่วยได้นำเอาระบบเวชระเบียนอิเล็กทรอนิกส์มาใช้ในการแก้ปัญหา อย่างไรก็ตามการรักษาความปลอดภัยและความเป็นส่วนตัวยังเป็นความเสี่ยงที่มีแนวโน้มที่เพิ่มขึ้นในระบบการดูแลสุขภาพ งานวิจัยนี้มีวัตถุประสงค์เพื่อปกป้องข้อมูลที่สามารถระบุข้อมูลส่วนบุคคลเกี่ยวกับสุขภาพโดยการจำกัดการใช้และการเปิดเผยข้อมูล การพัฒนาระบบประมวลผลข้อมูลอัตโนมัติโดยใช้เทคโนโลยีเคแอนอนนิไมเซชัน ได้ทำให้เกิดกลไกป้องกันความเป็นส่วนตัวที่เกี่ยวข้องกับข้อมูลส่วนบุคคล ผลของการศึกษาได้พิสูจน์ให้เห็นว่าเครื่องมือปกปิดตัวตนที่สามารถใช้ได้อย่างมีประสิทธิภาพและไม่มีค่าใช้จ่าย

ABSTRACT

Improving the quality and reducing the cost of patient care depends on moving to electronic patient records. However, the security and privacy risks associated with sensitive information are increased by several growing trends in healthcare. This research intended to protect individually identifiable health information by limiting its use and disclosure. The development of automatic data processing using k-anonimisation has made it necessary to consider privacy protection in relation to personal data. The results of the study are the effective Deidentification tool that is available as open source software.

คำสำคัญ: การปกปิดตัวตน การปกป้องสิทธิ์ส่วนบุคคล ข้อมูลสุขภาพ

Key Words: Deidentification, Privacy protection, Health information

* นักศึกษา หลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น

** รองศาสตราจารย์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น

บทนำ

การบันทึกเวชระเบียนผู้ป่วยในรูปของอิเล็กทรอนิกส์ มีวัตถุประสงค์เพื่อก่อให้เกิดความต่อเนื่องในการดูแลรักษาผู้ป่วย เกิดการสื่อสารที่ดีระหว่างทีมผู้ให้บริการ การบันทึกข้อมูลทางคลินิกของผู้ป่วยเป็นความรับผิดชอบของแพทย์ผู้ดูแลรักษาผู้ป่วย ซึ่งจะต้องทำการบันทึกข้อมูลนี้ด้วยตนเองหรือกำกับตรวจสอบให้มีการบันทึกที่ถูกต้อง หากแต่ข้อมูลผู้ป่วยมีการเผยแพร่ออกไปโดยไม่ได้ตั้งใจอาจส่งผลให้เกิดความเสียหายในด้านสังคม เศรษฐกิจ

วิธีการหนึ่งที่สำคัญสำหรับการสร้างความมั่นใจในความเป็นส่วนตัวในขณะที่การประมวลผลข้อมูลจำนวนมากคือกระบวนการปกปิดข้อมูลการแสดงตัวตน (Iyengar, 2003) ข้อมูลที่ถูกปกปิดการแสดงตัวตนหมายถึงข้อมูลไม่สามารถผ่านการเชื่อมโยงไปยังข้อมูลตัวตนของบุคคล องค์ประกอบที่สามารถระบุข้อมูลส่วนบุคคลได้ผ่านกลไกการปรับต่าง (scrubbing) ทำให้ปลอดภัยในแง่ความเป็นส่วนตัว (privacy protection) [1, 7] ในขณะที่เดียวกันก็พยายามที่จะรักษามูลค่าการค้าและทางวิทยาศาสตร์

งานวิจัยนี้จึงมีวัตถุประสงค์หลักในการพัฒนาเครื่องมือที่ช่วยให้เกิดกลไกการปกป้องข้อมูลส่วนบุคคลกับข้อมูลด้านสุขภาพซึ่งมีโอกาสดังกล่าวในการเกิดความเสียหายสูง (Machanavajjhala, Kifer, Gehrke, Venkitasubramaniam, 2007) เพื่อเป็นทางเลือกในการจัดการด้านความปลอดภัยให้เกิดขึ้นในระบบการจัดเก็บข้อมูลอิเล็กทรอนิกส์

งานวิจัยที่เกี่ยวข้อง

องค์ประกอบหลักของเวชระเบียนอิเล็กทรอนิกส์ (Fraser, Biondich, Moodley, Choi, Mamlin, Szolovits, 2005)

1. ข้อมูลทั่วไปเกี่ยวกับผู้ป่วย
2. บันทึกประวัติการรักษาของแพทย์
3. การสั่งยาผ่านคอมพิวเตอร์

4. การสั่ง Lab ผ่านคอมพิวเตอร์
5. การดูผล Lab ผ่านคอมพิวเตอร์
6. การบันทึก
7. การบันทึก
8. การสรุป
9. การดูผลการทำ Diagnostic Test ต่างๆ ผ่านคอมพิวเตอร์
10. การดูผล X-ray ผ่านคอมพิวเตอร์

ในองค์ประกอบข้อมูลสุขภาพในระบบเวชระเบียนอิเล็กทรอนิกส์ทำให้เกิดความเสี่ยง อาจถูกนำไปใช้ในทางที่ไม่ถูกต้อง จะส่งผลกระทบต่อผู้ป่วยได้ นักวิจัยหลายท่านได้พัฒนาแนวคิดในการแก้ปัญหาการเปิดเผยข้อมูลส่วนบุคคลหลายวิธี หนึ่งในเทคโนโลยีที่นักวิจัยให้ความสนใจในกลไกการปกปิดข้อมูลส่วนบุคคลคือ เทคนิคเคแอนโนนิไมเซชัน (k-Anonymisation) (Nergiz, Clifton, 2007) โดยมีการกำหนดค่า k หมายถึงจำนวนเรคอร์ด ที่กำหนดขึ้น โดยมีแนวคิดให้มีจำนวนไม่น้อยกว่า k เพื่อไม่ให้สามารถแยกความแตกต่างได้ ดังนั้นความน่าจะเป็นที่จะสามารถระบุตัวบุคคลได้คือ $1/k$ โดย k มีค่ามากกว่าหรือเท่ากับ 1 หาก k มีค่ามากขึ้น แสดงถึงความเป็นส่วนตัวที่มากยิ่งขึ้น (Machanavajjhala, Kifer, Gehrke, Venkitasubramaniam, 2007)

ปัจจุบันได้มีโปรแกรมที่สามารถทำงานได้ตามทฤษฎีนี้แต่ต้องเสียค่าใช้จ่ายสูงในการนำมาใช้ (PARAT Tool) ดังนั้นงานวิจัยนี้จึงต้องการพัฒนาเครื่องมือที่หน่วยงานในไทยไม่ต้องเสียค่าใช้จ่ายและมีการทำงานได้มีประสิทธิภาพเช่นเดียวกันกับโปรแกรมที่มีขายอยู่ในท้องตลาด โดยได้ออกแบบระบบที่นำเอาเทคโนโลยีเคแอนโนนิไมเซชันมาใช้กับข้อมูลสุขภาพ โดยมีผลที่คาดว่าจะได้รับคือข้อมูลที่ผ่านกลไกเทคโนโลยีเคแอนโนนิไมเซชันจะถูกปกปิด ได้ถูกต้องตามทฤษฎี

row #	Age	Sex	Zipcode	Disease	guarding node
1 (Andy)	5	M	12000	gastric ulcer	stomach disease
2 (Bill)	9	M	14000	dyspepsia	dyspepsia
3 (Ken)	6	M	18000	pneumonia	respiratory infection
4 (Nash)	8	M	19000	bronchitis	bronchitis
5 (Joe)	12	M	22000	pneumonia	pneumonia
6 (Sam)	19	M	24000	pneumonia	pneumonia
7 (Linda)	21	F	58000	flu	0
8 (Jame)	26	F	36000	gastritis	gastritis
9 (Sarah)	28	F	37000	pneumonia	respiratory infection
10 (Mary)	56	F	33000	flu	flu

(a) Microdata

Name	Age	Sex	Zipcode
Andy	5	M	12000
Bill	9	M	14000
Ken	6	M	18000
Nash	8	M	19000
Mike	7	M	17000
Joe	12	M	22000
Sam	19	M	24000
Linda	21	F	58000
Jane	26	F	36000
Sarah	28	F	37000
Mary	56	F	33000

(b) Voter registration list

row #	★Age	Sex	★ Zipcode	Disease
1	[1, 10]	M	[10001, 15000]	gastric ulcer
2	[1, 10]	M	[10001, 15000]	dyspepsia
3	[1, 10]	M	[15001, 20000]	pneumonia
4	[1, 10]	M	[15001, 20000]	bronchitis
5	[11, 20]	M	[20001, 25000]	pneumonia
6	[11, 20]	M	[20001, 25000]	pneumonia
7	[21, 60]	F	[30000, 60000]	flu
8	[21, 60]	F	[30000, 60000]	gastritis
9	[21, 60]	F	[30000, 60000]	pneumonia
10	[21, 60]	F	[30000, 60000]	flu

(c) A 2-anonymous table

รูปที่ 1 ตัวอย่างการแปลงข้อมูลโดยเลือก $k=2$ โดยเลือก
ปัจจัย age และ zip code ที่ปรับแต่งทำให้ไม่
สามารถเชื่อมโยงตัวตนของผู้ป่วยได้

ในรูปที่ 1 มีชุดข้อมูลสองชุดที่แสดงอยู่ใน 2(a) และ
2(b) ข้อมูลใน 2(a) มีข้อมูลเรื่องโรคที่ไม่ควรถูกเปิดเผย
ข้อมูลใน 2(b) มีรายชื่อบุคคล เมื่อทำการเชื่อมโยงข้อมูล
สองตารางที่ไม่ได้ถูกจัดเก็บที่เดียวกัน ปัจจัย Age, Sex,
Zipcode ซึ่งอยู่ในตาราง 2(b) เป็นปัจจัยบ่งชี้ทางอ้อม
(Quasi-identifier) (Bayardo, Agrawal, 2005) เมื่อนำไป
เชื่อมโยงกับปัจจัยที่อยู่ในตาราง 2(a) ก็สามารถเชื่อมโยง
กันได้ทำให้ความเป็นส่วนตัวของบุคคลที่อยู่ในตาราง
2(a) สามารถถูกระบุชื่อได้เป็นบุคคลใดที่มีข้อมูล
วินิจฉัยโรค ซึ่งหากข้อมูลนี้ถูกไปเผยแพร่ก็จะทำให้เกิด

ความเสียหายได้ แต่เมื่อได้มีการปรับแต่งด้วย
เทคโนโลยีเคแอนโนนิไมเซชันใน 2(c) ก็จะทำให้ไม่
สามารถเชื่อมโยงได้ว่าเป็นข้อมูลวินิจฉัยโรคของบุคคล
ใดบ้าง

วิธีการวิจัย

ผู้วิจัยได้พัฒนาโปรแกรมปกปิดความเป็นส่วนตัว

ของข้อมูลสุขภาพ (Electronic Medical Record

Anonymisation: EMRA) ด้วยภาษา Java โดยมี 6 ฟังก์ชัน
หลักดังนี้

1. นำเข้าข้อมูล (data input) สามารถทำได้ 2 รูปแบบ

แบบที่ 1 กำหนดค่าลงใน methods

```
DefaultData data = Data.create();
data.add("age", "gender", "zipcode");
data.add("55", "male", "67890");
```

แบบที่ 2 อ่านจากไฟล์

```
Data data = Data.create("data.csv", ',');
```

2. กำหนดข้อมูลที่ต้องการแปลง (data transformations
settings)

รองรับข้อมูลประเภท "String" "Decimal" และ "Date"

โดยกำหนดประเภทของ attributes และค่าความสูงใน

การทำ generalisation hierarchies

ประเภทของปัจจัย (attributes) โดยมีตัวเลือก 4 ประเภท

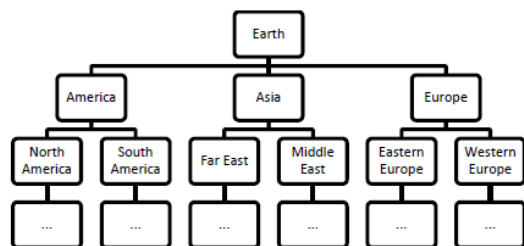
- Insensitive attributes ไม่มีการเปลี่ยนแปลงค่า
 - Directly-identifying attributes จะถูกลบทิ้งออกจากข้อมูล
 - Quasi-identifying attributes จะถูกตัดแปลงโดยใช้หลักการ generalization hierarchies
 - Sensitive attributes จะไม่ถูกเปลี่ยนแปลงค่า แต่สามารถปรับเปลี่ยนโดยใช้เทคนิค l-diverse หรือ t-close transformations ได้
3. ปรับแก้ไขและคำนวณข้อมูลตามทฤษฎี
anonimisation ที่ประกอบด้วย k-anonymity และ l-diversity anonymity (data alteration and manipulation)

4. สั่งให้อัลกอริทึมทำงานกับข้อมูลที่ต้องการปกปิด (anonimisation algorithm processing) โดยกำหนดตัวแปร data ค่า k และค่า suppressionRate
5. ปรับปรุงคุณภาพผลการปกปิดด้วยการประเมินด้วยค่าการสูญเสียของข้อมูล (information loss)
6. แสดงผลลัพธ์การปกปิดข้อมูลส่วนบุคคล

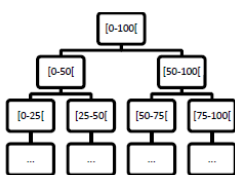
ในงานวิจัยนี้ได้ทดสอบระบบโดยเลือกใช้ข้อมูล adult ที่เป็นข้อมูลมาตรฐาน (benchmark) (UCI) ขนาดปัจจัย 15 attributes และจำนวนเรคอร์ดทั้งหมด 31,062 รายการ

ผลการวิจัย

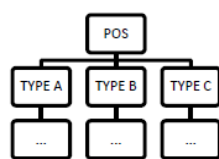
การเจเนรัลไลเซชัน (general hierarchy) คือการแปลงค่าปัจจัยไปยังโดเมนที่เจาะจงน้อยกว่าและสามารถกระทำเป็นลำดับชั้น ซึ่งข้อมูลมาตรฐาน adult สามารถจัดรูปแบบได้ในรูปที่ 3



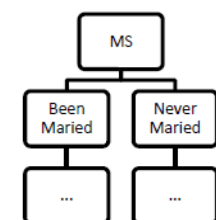
(a) ปัจจัยประเทศ



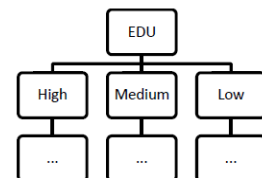
(b) ปัจจัยอายุ



(c) ปัจจัยตำแหน่ง (position)

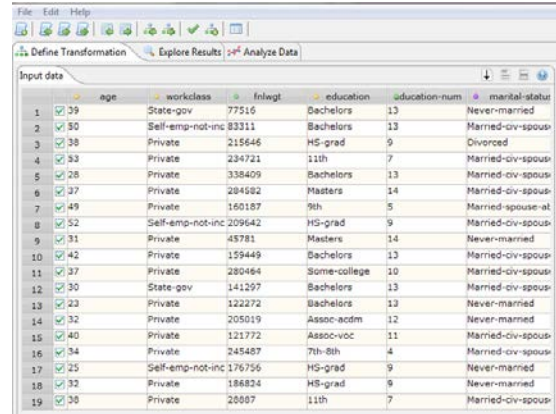


(d) ปัจจัยสถานะสมรส

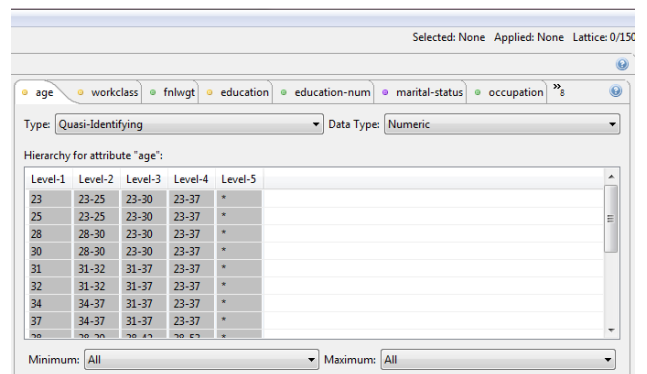


(e) ปัจจัยด้านการศึกษา

รูปที่ 2 โครงสร้างผังสัมพันธ์เชิงระดับ (general hierarchy) ของค่าตัวแปรในแต่ละปัจจัยของข้อมูล adult โปรแกรม EMRA

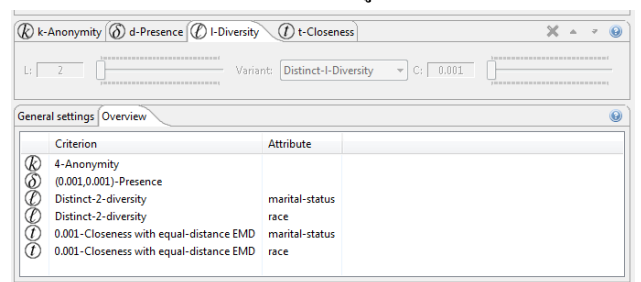


รูปที่ 3 หน้าต่างหลักการนำเข้าข้อมูลที่ต้องการปรับแต่งในโปรแกรม EMRA



รูปที่ 4 หน้าต่างแสดงข้อมูลปัจจัย การกำหนดประเภทของปัจจัย (Type) และประเภทของข้อมูล (Data Type)

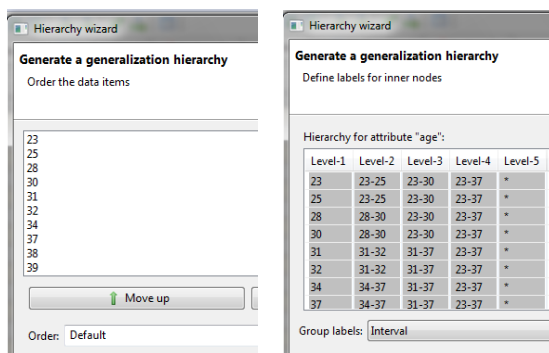
ในรูปที่ 5 ฟังก์ชัน k-Anonymity และ d-Presence ทำการแปลงกับข้อมูลประเภท quasi-identifier ส่วน l-Diversity, t-Closeness ในการแปลงข้อมูลประเภท sensitive



Criterion	Attribute
4-Anonymity	
(0.001,0.001)-Presence	
Distinct-2-diversity	marital-status
Distinct-2-diversity	race
0.001-Closeness with equal-distance EMD	marital-status
0.001-Closeness with equal-distance EMD	race

รูปที่ 5 หน้าต่างฟังก์ชันเทคโนโลยีเคแอนโนนิไมเซชันในการกำหนดค่า k-Anonymity, d-Presence, l-Diversity, t-Closeness

ในการทดลองนี้ กำหนดปัจจัย age marital-status และ education เป็นปัจจัยบังชี้ทางอ้อม ปัจจัย age สามารถจัดโครงสร้างผังสัมพันธ์เชิงระดับได้ 4 ระดับ ดังแสดงในรูปที่ 6 ปัจจัย marital-status สามารถจัดได้ 2 ระดับและ education จัดได้ 3 ระดับ ฟังก์ชันเคแอนโนนิไมเซชันจะทำให้เกิดการสูญเสียสารสนเทศ วิธีพื้นฐานในการวัดค่าการสูญเสียสารสนเทศคือการนับจำนวนข้อมูลที่ถูกลบปิดหรือถูกเจเนรัลไลซ์ (information loss) (Nergiz, Clifton, 2007)



(a) ข้อมูลก่อนการจัด

(b) ข้อมูลที่เปลี่ยนแปลงไปด้วย

โครงสร้างความสัมพันธ์

ความสัมพันธ์แบบ interval

Hierarchy for attribute "age":

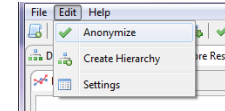
Level-1	Level-2	Level-3	Level-4	Level-5
23	2-2	2-2-2-3	2-2-2-3-3-3-3	*
25	2-2	2-2-2-3	2-2-2-3-3-3-3	*
28	2-3	2-2-2-3	2-2-2-3-3-3-3	*
30	2-3	2-2-2-3	2-2-2-3-3-3-3	*
31	3-3	3-3-3-3	2-2-2-3-3-3-3	*
32	3-3	3-3-3-3	2-2-2-3-3-3-3	*
34	3-3	3-3-3-3	2-2-2-3-3-3-3	*
37	3-3	3-3-3-3	2-2-2-3-3-3-3	*

Group labels: Prefixes of length 1

(c) ข้อมูลที่เปลี่ยนแปลงไปด้วยความสัมพันธ์แบบ

Prefixes of length 1 ได้มากถึง 5 ระดับ

รูปที่ 6 การเจเนรัลไลเซชันปัจจัย age



(a) ฟังก์ชันเคแอนโนนิไมเซชันด้วยเมนู Anonymize

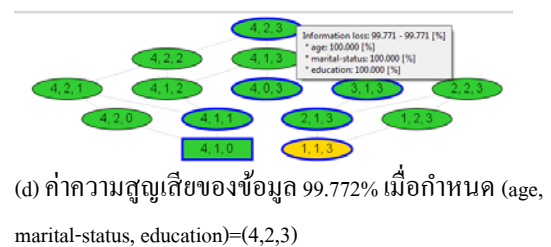
Lattice List

Transformation	Anonymity	Min. Info. Loss	Max. Info. Loss
[1, 1, 3]	ANONY...	5.5170356E7 ...	5.5170356E7 ...
[4, 1, 0]	ANONY...	8.889068E7 [...]	8.889068E7 [...]
[2, 1, 3]	ANONY...	1.0457772E8 ...	1.0457772E8 ...
[4, 1, 1]	ANONY...	1.6231213E8 ...	1.6231213E8 ...
[3, 1, 3]	ANONY...	1.92609378E...	1.92609378E...
[4, 0, 3]	ANONY...	3.11880088E...	3.11880088E...
[4, 2, 0]	ANONY...	8.889068E7 [...]	9.09746244E8...
[1, 2, 3]	ANONY...	5.5170356E7 ...	9.09746244E8...
[4, 2, 1]	ANONY...	1.6231213E8 ...	9.09746244E8...

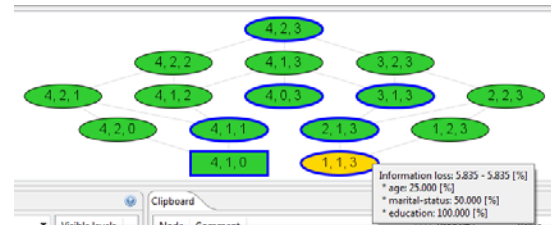
(b) ผลลัพธ์ของการปรับแต่งด้วย k=2 แสดงให้เห็นค่าต่ำสุดสูงสุดโอกาสในการสูญเสียข้อมูล



(c) โครงสร้างการเลือกระดับเจเนรัลไลเซชันของปัจจัยทั้ง 3



(d) ค่าความสูญเสียของข้อมูล 99.772% เมื่อกำหนด (age, marital-status, education)=(4,2,3)



(e) ค่าความสูญเสียของข้อมูล 5.83% เมื่อกำหนด (age, marital-status, education)=(1,1,3)

รูปที่ 7 แสดงผลลัพธ์ในการแปลงข้อมูลด้วยฟังก์ชันเคแอนโนนิไมเซชันที่ควรเลือก generalize hierarchy (1,1,3) ที่ให้ค่าการสูญเสียที่น้อยที่สุด

เมื่อเสร็จสิ้นขั้นตอนการกำหนดค่าฟังก์ชันเคแอนโนนิไมเซชันกดปุ่ม Edit และเลือกคำสั่ง Anonymize ดังแสดงในรูปที่ 7(a) สามารถดูข้อเสนอของการปรับแต่งในรูปของตารางแสดงการสูญเสียข้อมูล 7(b) หรือโครงสร้างผังความสัมพันธ์ 7(c) รูปที่ 8 แสดงผลลัพธ์

ฟังก์ชันเคแอนโนนิไมเซชันกับปัจจัย age marital-status
และ education เปรียบเทียบกับข้อมูลก่อนการปรับแต่ง

Input data

	age	race	marital-status	education
1	88	White	Married-civ-spouse	Prof-school
2	90	White	Married-civ-spouse	Bachelors
3	90	White	Married-civ-spouse	Masters
4	90	White	Married-civ-spouse	Bachelors
5	90	White	Married-civ-spouse	11th
6	90	White	Married-civ-spouse	Prof-school
7	90	White	Married-civ-spouse	Bachelors
8	90	White	Married-civ-spouse	Some-college
9	90	White	Married-civ-spouse	HS-grad
10	90	White	Married-civ-spouse	10th
11	90	Asian-Pac-Islander	Married-civ-spouse	Assoc-acdm
12	90	White	Married-civ-spouse	Bachelors
13	90	White	Married-civ-spouse	HS-grad
14	90	White	Married-civ-spouse	HS-grad

(a) ข้อมูลก่อนการปรับแต่ง

Output data

	age	race	marital-status	education
1	85-89	White	spouse present	*
2	85-89	White	spouse present	*
3	85-89	White	spouse present	*
4	85-89	White	spouse present	*
5	85-89	White	spouse present	*
6	85-89	White	spouse present	*
7	85-89	White	spouse present	*
8	85-89	White	spouse present	*
9	85-89	White	spouse present	*
10	85-89	White	spouse present	*
11	85-89	Asian-Pac-Islander	spouse present	*
12	85-89	White	spouse present	*
13	85-89	White	spouse present	*
14	85-89	White	spouse present	*

(b) ข้อมูลหลังการปรับแต่งด้วย generalize hierarchy
(1,1,3)

รูปที่ 8 หน้าต่างการทำงานโปรแกรมทดลองการปกป้อง
ข้อมูลส่วนบุคคล

อภิปรายและสรุปผลการวิจัย

ผลลัพธ์ที่ได้จากงานวิจัยนี้คือ ได้เครื่องมือที่สามารถทำงานตามทฤษฎีเทคนิคเคแอนโนนิไมเซชันเป็นแนวทางเลือกให้กับหน่วยงานด้านสาธารณสุขหรือหน่วยงานที่เกี่ยวข้องกับการนำข้อมูลด้านสุขภาพไปใช้ โดยมีจุดมุ่งหมายในการปกป้องความเป็นส่วนตัวให้กับเจ้าของข้อมูล โดยโปรแกรมสามารถมีฟังก์ชันที่ปรับเลือกได้ตามสภาพของข้อมูลที่มีความแตกต่างกัน และนำเสนอในรูปของความสัมพันธ์ที่ผู้ใช้สามารถใช้งานได้ง่ายขึ้นและสามารถนำไปใช้งานได้หลาย platform ลดปัญหาการติดตั้งใช้งานและลดค่าใช้จ่าย

เอกสารอ้างอิง

- Bayardo R, Agrawal R. Data privacy through optimal k-anonymization. Proceedings of the 21st International Conference on Data Engineering, 2005 April 5-8; Tokyo, Japan. p. 217-228.
- Fraser HS, Biondich P, Moodley D, Choi S, Mamlin BW, Szolovits P. Implementing electronic medical record systems in developing countries. Inform Prim Care. 2005;13(2): 83-95.
- Iyengar VS. Transforming data to satisfy privacy constraints. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2003 August 24-27; Washington, DC, USA. p. 279-288.
- Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-diversity: privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data, Volume 1 Issue 1, 2007 March. Article No. 3.
- Nergiz ME, Clifton C. Thoughts on k-anonymization. Data & Knowledge Engineering, 2007; vol. 63, no. 3: p. 622-645.
- PARAT Tool, <http://www.privacyanalytics.ca/software/>
- Sweeney L. k-Anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems; Volume 10 Issue 5, 2002 October. p. 557-570.
- UCI, <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult>