KKU Engineering Journal

http://www.en.kku.ac.th/enjournal/th/

การทบทวนงานวิจัยเกี่ยวกับเทคนิคเคแอนโนนิไมเซชันในการแปลงข้อมูลเพื่อรักษาความเป็นส่วนตัว

Review of research on k-anonymization technique to transform data for privacy preservation

นรารัตน์ เรืองชัยจตุพร*

Nararat Ruangchaijatupon*

ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น จังหวัดขอนแก่น 40002

Department of Electrical Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, Thailand, 40002.

Received December 2013

Accepted March 2014

บทคัดย่อ

ข้อมูลความเป็นส่วนตัวอาจรั่วไหลผ่านการเปิดเผยฐานข้อมูล แม้จะปกปิดแอททริบิวท์ที่เป็นตัวบ่งชี้บุคคลแล้วการระบุตัวบุคคลอีกครั้งสามารถทำได้โดยการเชื่อมโยงแอททริบิวท์ที่เหลืออยู่ไปยังหลายแหล่งข้อมูล การแปลงข้อมูลด้วยเทคนิคเคแอนโนนิไมเซชันเป็นวิธีการที่สามารถยืนยันระดับความเป็นส่วนตัวได้ บทความนี้เป็นบทความวิชาการที่ทบทวนการศึกษาและวิจัยเกี่ยวกับเทคนิคเคแอนโนนิไมเซชัน โดยเรียบเรียงความรู้เบื้องต้น ปัญหาการระบุตัวบุคคลอีกครั้ง ศัพท์เฉพาะที่เกี่ยวข้องและชนิดของแอททริบิวท์ อีกทั้งได้ทบทวนวิธีประเมินการสูญเสียสารสนเทศและอัลกอริธึมต่างๆ ที่ใช้แปลงข้อมูลเพื่อให้มีคุณสมบัติเคแอนโนนิมิที นอกจากนี้ยังได้อธิบายการโจมตีที่เป็นไปได้และเสนอแนะแนวทางสำหรับศึกษาและวิจัยในอนาคตไว้ด้วย

คำสำคัญ : การรักษาความเป็นส่วนตัว การแปลงข้อมูล เคแอนโนนิไมเซชัน

Abstract

Private information may leak through data disclosure. Although the identifier attributes are suppressed, re-identification is possible by linking remaining attributes to several sources. Transforming data using the k-anonymization technique can ensure levels of privacy. This paper reviews the studies and researches of k-anonymization technique. Firstly, the paper presents basic knowledge, the re-identification problem, related terminologies, and attribute types. The paper also reviews information loss estimations and transforming algorithms to achieve k-anonymity property. Lastly, possible attacks are described, and future studies and researches are suggested.

Keywords : Privacy preservation, Data transformation, k-anonymization

* Corresponding author. Tel.: +66-4320-2353 ; fax: +66-4320-2836

Email address: nararat@kku.ac.th .

1. บทนำ

ปัจจุบันหลายองค์กรมีการรวบรวมข้อมูลบุคคลจำนวนมากไว้ในฐานข้อมูลขององค์กร เช่น ฐานข้อมูลลูกค้าของธนาคาร บริษัทประกัน บริษัทบัตรเครดิต เวชระเบียนอิเล็กทรอนิกส์ของโรงพยาบาล [1, 3, 10] บางครั้งองค์กรเหล่านั้นมีการเปิดเผยหรือแลกเปลี่ยนฐานข้อมูลดังกล่าวด้วยเหตุผลทางธุรกิจหรือเหตุผลทางกฎหมาย [2] การเปิดเผยฐานข้อมูลที่บรรจุข้อมูลส่วนบุคคล เช่น วัน เดือนปีเกิด รหัสบัตรประชาชน ระดับรายได้ โรคที่รักษา ก่อให้เกิดความเสี่ยงที่บุคคลจะถูกละเมิดความเป็นส่วนตัว แม้การเปิดเผยฐานข้อมูลนั้นจะมีการปกปิด (Suppression) แอททริบิวท์ที่เป็นตัวบ่งชี้บุคคล (Identifier) ไปแล้ว บุคคลยังสามารถถูกระบุตัวตนได้ด้วยการเชื่อมโยงจากแอททริบิวท์ที่เหลืออยู่ [3-4] ไปยังหลายแหล่งข้อมูล

เพื่อป้องกันการละเมิดความเป็นส่วนตัว ก่อนการเปิดเผยฐานข้อมูลจึงจำเป็นต้องปกปิดหรือแปลงข้อมูลการทำให้ข้อมูลยุ่งเหยิง (Data perturbation) โดยการเพิ่มข้อมูลรบกวน (Noise) และสลับข้อมูล (Swapping) [5] โดยคงไว้ซึ่งคุณสมบัติทางสถิตินั้นไม่เป็นที่นิยมเนื่องจากไม่สามารถรับรองความสมบูรณ์ของข้อมูล (Data integrity) [6] อีกทั้งยังมีความเสี่ยงในการถูก*ระบุตัวบุคคลอีกครั้ง* (*Re-identify*) ซึ่งสามารถวิเคราะห์ได้ด้วยวิธีทางสถิติ [7-9] อีกแนวทางหนึ่งซึ่งได้รับความสนใจคือการใช้เทคนิคเคแอนโนนิไมเซชัน [10] โดยจะแปลงข้อมูลเพื่อให้มี*คุณสมบัติเคแอนโนนิมิที* (*k-anonymity*) ซึ่งทำให้สามารถยืนยันได้ว่า หลังจากการแปลงข้อมูลแล้ว จะมีข้อมูลจำนวนไม่น้อยกว่า k ระเบียน (Record) ที่ไม่สามารถแยกความแตกต่างได้ ดังนั้นความน่าจะเป็นที่จะสามารถระบุตัวบุคคลได้คือ $1/k$ โดย k มีค่ามากกว่าหรือเท่ากับ 1 หาก k มีค่ามากขึ้น แสดงถึงความเป็นส่วนตัวที่มากยิ่งขึ้น [11]

บทความนี้นำเสนอการทบทวนเอกสารเกี่ยวกับการแปลงข้อมูลด้วยเทคนิคเคแอนโนนิไมเซชันเพื่อรักษาความเป็นส่วนตัว โดยเริ่มจากปัญหาการระบุตัวบุคคลจากการเชื่อมโยงแอททริบิวท์ที่เหลืออยู่และนำเสนอศัพท์เฉพาะที่เกี่ยวข้อง นอกจากนี้บทความยังได้ทบทวนวิธีประเมิน*การสูญเสียสารสนเทศ* (*Information loss*) จากการที่ข้อมูลถูกแปลงและอัลกอริธึมต่างๆ ที่ใช้ในการแปลงข้อมูล พร้อมทั้งอธิบายวิธีโจมตีที่เป็นไปได้ และเสนอแนะแนวทางการศึกษาและวิจัยในอนาคต

2. การระบุตัวบุคคลและศัพท์เฉพาะ

ยกตัวอย่างปัญหาการระบุตัวบุคคลดังนี้ หากมีการเปิดเผยฐานข้อมูลเวชระเบียนอิเล็กทรอนิกส์โดยปกปิดแอททริบิวท์ชื่อ (Name) ซึ่งเป็นตัวบ่งชี้บุคคล แต่เปิดเผยวันเดือนปีของใบเสร็จรับเงินที่ออกโดยโรงพยาบาล (Date-bill) จำนวนเงินค่ารักษา (Paid) และโรคที่เข้ารับการรักษา (Disease) ดังตารางที่ 1 ขณะที่ฐานข้อมูลผู้ประกันตนจากสำนักงานประกันสังคมซึ่งดูเหมือนไม่ได้เป็นข้อมูลส่วนตัวที่มีความอ่อนไหว จึงเปิดเผยชื่อ (Name) วันเดือนปีของใบเสร็จรับเงิน (Date-bill) และจำนวนเงินที่สำนักงานจ่าย (Paid) ดังตารางที่ 2 [12]

ตารางที่ 1 ตัวอย่างตารางฐานข้อมูลเวชระเบียนอิเล็กทรอนิกส์ที่ปกปิดตัวบ่งชี้บุคคล

Name	Date-bill	Paid	Disease
*	24/03/2552	3300	กระดูกขาหัก
*	10/12/2556	300	ท้องเสีย
*	28/09/2551	320	หวัด
*	21/04/2556	2800	มือเท้าปาก

ตารางที่ 2 ตัวอย่างตารางฐานข้อมูลผู้ประกันตน

Name	Date-bill	Paid
กนก	24/03/2552	3300
ขจร	10/12/2556	300
คมสัน	28/09/2551	320
งามตา	21/04/2556	2800

จะเห็นว่าแอททริบิวท์ “21/04/2556” (Date-bill) และ “2800” (Paid) เป็น*ตัวบ่งชี้ทางอ้อม* (*Quasi-identifier*) เมื่อนำไปเชื่อมโยงกับแอททริบิวท์ที่เหลืออยู่ในตารางที่ 1 ทำให้สามารถทราบว่างามตาเป็นโรคมือเท้าปาก ซึ่งอาจส่งผลทางสังคมต่องามตา หากมีการแปลงข้อมูลด้วยเทคนิคเคแอนโนนิไมเซชันจากตารางที่ 1 เป็นตารางที่ 3 (k = 2) ซึ่งทุกแอททริบิวท์ที่เปิดเผยมีอย่างน้อย

2 ระเบียนที่ไม่แตกต่างกัน จะปกป้องความเป็นส่วนตัวของบุคคลที่มีข้อมูลในตารางได้

การแปลงข้อมูลด้วยเทคนิคเคแอนโนนิไมเซชันมีศัพท์เฉพาะซึ่งอธิบายด้วยตัวอย่างดังนี้ จากตารางที่ 3 จะเห็นว่าการแปลงข้อมูลสามารถกระทำได้ 2 แบบคือการปกปิดซึ่งแทนชื่อผู้ป่วยด้วยเครื่องหมาย (*) และ*การเจเนรัลไลเซชัน* (*Generalization*) [13] เช่น การแปลงวันที่ (21/04/2556) เป็นช่วงวันที่ (2556-2560) การแปลงจำนวนเงิน (2800) เป็นช่วงของจำนวน (2001-4000) หรือการแปลง “หวัด” เป็น “หวัดหรือมือเท้าปาก” ซึ่งทำให้ข้อมูลยังคงให้สารสนเทศบางส่วน ทำให้นิยมใช้การเจเนรัลไลเซชันมากกว่าการปกปิดซึ่งไม่ให้สารสนเทศใดๆ นอกจากนี้ ยังสามารถแบ่งแอททริบิวท์เป็น 2 ชนิดคือ แอททริบิวท์*เชิงตัวเลข* (*Numerical attribute*) ซึ่ง*สามารถเรียงลำดับ* (*Sortable*) มากน้อยหรือก่อนหลังได้ เช่น จำนวนเงิน วันเดือนปี เวลา และแอททริบิวท์*เชิงลักษณะ* (*Categorical attribute*) เช่น อาชีพ โรคที่เป็น เพศ สถานภาพ เป็นต้น

ตารางที่ 3 ตัวอย่างตารางฐานข้อมูลเวชระเบียนอิเล็กทรอนิกส์ที่มีคุณสมบัติ 2-anonymity

Name	Date-Bill	Paid	Disease
*	2551-2555	2001-4000	กระดูกขาหัก หรือ
*	2556-2560	0-2000	ท้องเสีย
*	2551-2555	0-2000	หวัด หรือ
*	2556-2560	2001-4000	มือเท้าปาก

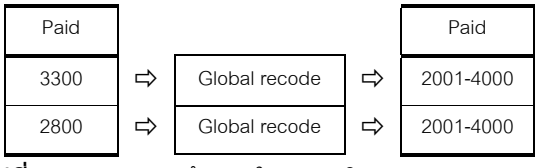
การแปลงข้อมูลแต่ละแอททริบิวท์ของแต่ละระเบียนเรียกว่า*การเข้ารหัสใหม่* (*Recoding*) ซึ่งมี 2 แบบดังนี้

การเข้ารหัสใหม่แบบ*โดยรวม* (*Global recoding*) จะแปลงแอททริบิวท์ที่เข้าข่ายการแปลงเดียวกันไปเป็น*ค่าโดยทั่วไป* (*Generalized value*) แบบเดียวกันเท่านั้น โดยกระทำกับทุกระเบียนในตารางที่เข้าข่ายการแปลงนั้นๆ เช่น ในรูปที่ 1 แอททริบิวท์จำนวนเงิน “3300” และ “2800” เข้าข่ายช่วง “2001-4000” เหมือนกัน จะถูกแปลงไปเป็น “2001-4000” แบบเดียวเท่านั้น

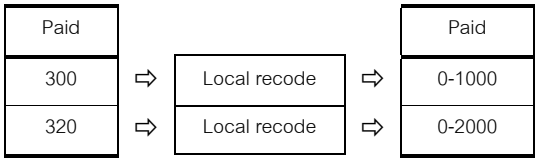
ส่วนการเข้ารหัสใหม่แบบ*เฉพาะแห่ง* (*Local recoding*) อนุญาตให้แปลงแอททริบิวท์ที่เข้าข่ายการแปลงเดียวกันไปเป็นค่าโดยทั่วไปที่ต่างกันได้ เช่น ในรูปที่ 2 แอททริบิวท์จำนวนเงิน “300” ถูกแปลงไปเป็น “0-1000” ส่วนแอททริบิวท์จำนวนเงิน “320” ถูกแปลงไปเป็น “0-2000” แม้ว่าทั้ง “300” และ “320” จะเข้าข่ายการแปลงเดียวกัน (คือเข้าข่ายทั้ง “0-1000” และ “0-2000”) แต่ค่าที่ถูกแปลงแล้วนั้นต่างกัน

การเข้ารหัสใหม่แบบเฉพาะแห่งมีความยืดหยุ่นกว่าและมีโอกาสในการสูญเสียสารสนเทศน้อยกว่า [14] การเข้ารหัสใหม่ทั้งสองแบบสามารถกระทำกับแอททริบิวท์เชิงตัวเลขหรือเชิงลักษณะได้ทั้งสองชนิด

แอททริบิวท์ต่างๆ ของตารางสามารถมีค่าได้ต่างๆ กัน เซตของค่าที่เป็นไปได้ของแอททริบิวท์เรียกว่า*โดเมน* (*Domain*) [15] การเจเนรัลไลเซชันคือการแปลงค่าแอททริบิวท์ไปยังโดเมนที่เจาะจงน้อยกว่าและสามารถกระทำเป็นลำดับชั้นดังรูปที่ 3และ 4



รูปที่ 1 ตัวอย่างการเข้ารหัสใหม่แบบโดยรวม

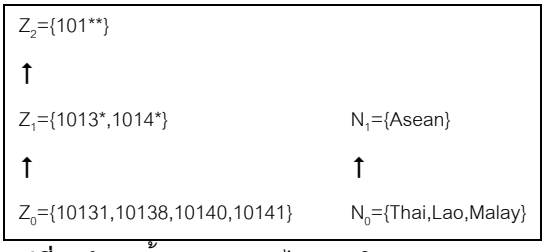


รูปที่ 2 ตัวอย่างการเข้ารหัสใหม่แบบเฉพาะแห่ง

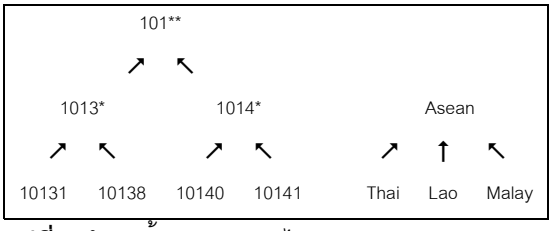
รูปที่ 3 แสดง*ลำดับชั้นการเจเนรัลไลเซชันโดเมน* (*Domain generalization hierarchy*) [16] ของรหัสไปรษณีย์ซึ่งเป็นแอททริบิวท์เชิงตัวเลข จากลำดับชั้นที่ Z_0 ไปยัง Z_1 และ Z_2 ซึ่งในแต่ละลำดับชั้นจะทำให้ความเจาะจงของแอททริบิวท์นั้นลดลงหรือสูญเสียสารสนเทศมากขึ้น นอกจากนั้นรูปที่ 3 ยังแสดงลำดับชั้นการเจเนรัลไลเซชันของแอททริบิวท์สัญชาติ (N_0 ไปยัง N_1) ซึ่งเป็นแอททริบิวท์เชิงลักษณะ ส่วนรูปที่ 4 แสดง*ลำดับชั้นการเจเนรัลไลเซชันค่า* (*Value generalization hierarchy*) ของแอททริบิวท์ทั้งสองเช่นกัน การเจเนรัลไลเซชันโดเมนจะกำหนดเซตของโดเมนที่สามารถ*จัดระบบได้โดยสมบูรณ์*

(*Totally ordered*) [3] ส่วนการเจเนรัลไลเซชันค่าจะกำหนดว่าแต่ละค่าในลำดับชั้นต่ำกว่าจะสอดคล้องกับค่าใดในลำดับชั้นที่สูงขึ้น การเจเนรัลไลเซชันทั้งสองเป็นการแปลงค่าให้กับแอททริบิวท์ซึ่งทำให้ได้ผลลัพธ์ของการแปลงเหมือนกัน

เมื่อข้อมูลถูกแปลงค่าแล้วจะถูกจัดให้อยู่ในเซตเดียวกันเรียกว่า *คลาสเท่าเทียม* (Equivalence class) [18] ซึ่งก็คือเซตของทุกระเบียนในตารางที่มีแอททริบิวท์ตัวบ่งชี้ทางอ้อมทุกตัวใน*เซตตัวบ่งชี้ทางอ้อม* (*Quasi-identifier set*) มีค่าเท่ากันหรือเหมือนกันทั้งหมด



รูปที่ 3 ลำดับชั้นการเจเนรัลไลเซชันโดเมน



รูปที่ 4 ลำดับชั้นการเจเนรัลไลเซชันค่า

3. การสูญเสียสารสนเทศ

แม้ว่าเทคนิคเคแอนโนนิไมเซชันจะช่วยรักษาความเป็นส่วนตัว แต่ก็มีข้อเสียคือทำให้เกิดการสูญเสียสารสนเทศ เพราะการเจเนรัลไลเซชันทำให้ข้อมูลมีความเจาะจงน้อยลง หากแอททริบิวท์ถูกเจเนรัลไลซ์ไปยังลำดับชั้นที่สูงมากขึ้นเท่าใด ก็จะสูญเสียสารสนเทศมากขึ้นเท่านั้น และหากจำนวนแอททริบิวท์ที่ถูกเจเนรัลไลซ์มีมาก การสูญเสียสารสนเทศก็จะมากตามไปด้วย มีงานวิจัยหลายงานที่เสนออัลกอริธึมซึ่งแปลงข้อมูลโดยพยายามทำให้เกิดการสูญเสียสารสนเทศน้อยที่สุด ในขณะที่ทำให้ตารางข้อมูลที่แปลงแล้วยังคงมีคุณสมบัติเคแอนโนนิมิตีโดยไม่ทำให้เกิดการเจเนรัลไลซ์เซชันมากเกินไป [11, 21] อัลกอริธึมเหล่านั้นพยายามหาคำตอบที่ดีที่สุด (Optimal solution) นั่นคือพยายามทำให้จำนวนระเบียนในคลาสเท่าเทียมมีค่าเท่ากับ k และรักษาสารสนเทศในข้อมูลให้มากที่สุด อัลกอริธึมเหล่านั้นจึงจำเป็นต้องวัดค่าการสูญเสียสารสนเทศ วิธีพื้นฐานในการวัดค่าการสูญเสียสารสนเทศคือการนับจำนวนข้อมูลที่ถูกปกปิดหรือถูกเจเนรัลไลซ์ [6] นอกจากวิธีนับแล้ว ยังมีวิธีวัดค่าการสูญเสียสารสนเทศอีกหลายวิธี ดังจะอธิบายต่อไปนี้

หากกำหนดให้ตาราง T มีจำนวนระเบียนเท่ากับ n และเซตตัวบ่งชี้ทางอ้อม Q_T มีจำนวนสมาชิกเท่ากับ d ให้ $R_i(j)$ เป็นแอททริบิวท์ที่ j ของระเบียนที่ i และให้ $R_i'(j)$ เป็นแอททริบิวท์ที่ถูกแปลงค่าแล้ว ให้ตาราง T' คือตารางที่ถูกแปลงข้อมูลแล้ว งานวิจัยของ Iyengar [2] เสนอ*เมตริกการสูญเสีย* (*Loss metric*: *LM*) เพื่อคำนวณการสูญเสียสารสนเทศเฉลี่ยของตาราง T' ดังสมการที่ (1)

$$LM\left(T,\left(T'\right)\right)=\frac{1}{nd}\cdot\sum_{i=1}^{n}\sum_{j=1}^{d}\frac{\left|R'_i(j)\right|-1}{\left|R_i(j)\right|-1} \qquad (1)$$

เมื่อให้ฟังก์ชัน $|\ .\ |$ คืนค่าจำนวนแอททริบิวท์ที่แตกต่างกันสำหรับแอททริบิวท์เชิงลักษณะ ส่วนแอททริบิวท์เชิงตัวเลข $|R_i'(j)|$ จะคืนค่าช่วงค่า+1 และ $|R_i(j)|$ จะคืนค่าพิสัย+1

ในงานวิจัยเดียวกัน Iyengar [2, 18] ก็ได้เสนอ*เมตริกการจำแนก* (*Classification metric*: *CM*) ซึ่งพิจารณาการจำแนกข้อมูลโดยใช้แอททริบิวท์เป็นตัวแบ่งแยก แอททริบิวท์ที่ใช้นั้นรวมไปถึงแอททริบิวท์ที่ไม่เป็นตัวบ่งชี้ทางอ้อมหรือมีศักยภาพในการบ่งชี้บุคคลด้วย ดังนั้นเมื่อถูกปกปิดหรือถูกเจเนรัลไลซ์ จะทำให้การจำแนกทำได้ยากขึ้น นอกจากนั้นยังพิจารณาว่าระเบียนต่างๆ ที่มีค่าแอททริบิวท์ตัวบ่งชี้ทางอ้อมเหมือนกันจัดอยู่ในกลุ่ม (Group) เดียวกัน ดังนั้นในการจำแนกจึงเป็นการดีกว่าที่ระเบียนจะอยู่ในทั้งกลุ่มและคลาสเดียวกัน ระเบียนที่อยู่ในกลุ่มเดียวกันแต่ต่างคลาสกันจะได้รับค่าโทษ (Penalty) เท่ากับ 1 ในการคำนวณค่าเมตริกการจำแนก สามารถคำนวณได้ด้วยสมการที่ (2)

$$CM\left(T,\left(T'\right)\right)=\frac{1}{n}\cdot\sum_{i=1}^{n}penalty(i) \qquad (2)$$

ฟังก์ชัน $penalty(i)$ จะคืนค่า 1 เมื่อระเบียนที่ i มีอย่างน้อยหนึ่งแอททริบิวท์ที่ถูกปกปิดหรือถูกเจเนรัลไลซ์ หรือระเบียนที่ i อยู่ต่างคลาสกับกลุ่มส่วนใหญ่ (Majority group) จะเห็นว่าค่า *CM* มากแสดงถึงการสูญเสียสารสนเทศมาก

Bayardo และ Agrawal [20] เสนอ*เมตริกความสามารถวินิจฉัย* (*Discernibility metric*: *DM*) โดยเน้นการสูญเสียสารสนเทศไปที่จำนวนระเบียนของตาราง T' ที่ไม่สามารถแยกความแตกต่างจากระเบียนอื่นในตารางได้ ค่า *DM* คำนวณได้ด้วยสมการที่ (3)

$$DM(T') = \sum_{i=1}^{n} D(i) \quad (3)$$

ฟังก์ชัน $D(i)$ คือค่าจำนวนระเบียนของตาราง T' ที่ไม่สามารถแยกความแตกต่างจากระเบียนที่ i ได้ จะเห็นว่าหากตาราง T' ที่ถูกแปลงข้อมูลจนใกล้เคียงคำตอบที่ดีที่สุดแล้ว จำนวนระเบียนในคลาสเท่าเทียมจะมีค่าเข้าใกล้ค่า k ซึ่งใกล้เคียงกับค่า *DM* จึงทำให้ไม่นิยมใช้ค่า *DM* ในการวัดการสูญเสียสารสนเทศ [24]

Nergiz และ Clifton [18] เสนอให้ใช้*เมตริกความคลุมเครือ* (*Ambiguity metric*: *AM*) ในการวัดการสูญเสียสารสนเทศโดยคำนวณจากจำนวนวิธีในการรวมกลุ่ม (Combination) ที่เป็นไปได้ทั้งหมดของแอททริบิวท์ของระเบียนที่ถูกเจเนรัลไลซ์ไปเป็นแอททริบิวท์ที่เหมือนกัน โดยมีการคำนวณดังสมการที่ (4)

$$AM(T,(T')) = \frac{1}{n} \cdot \sum_{i=1}^{n} \prod_{j=1}^{d} |R'_i(j)| \quad (4)$$

จะเห็นว่าเมตริกความคลุมเครือคือค่าเฉลี่ยของผลคูณคาร์ทีเซียน (Cartesian product) ของทุกแอททริบิวท์ที่ถูกเจเนรัลไลซ์ของแต่ละระเบียน ทำให้มีข้อด้อยคือพจน์ผลคูณนับรวมเอาจำนวนแอททริบิวท์ที่อาจไม่มีอยู่ในตารางดั้งเดิม T เข้าไว้ด้วย [24]

Xu และคณะ [21] เสนอ*นอร์มัลไลซ์เซอร์เทนทีพีนัลที* (*Normalized certainty penalty*: *NCP*) ซึ่งคล้ายกับเมตริกการสูญเสียแต่มีความละเอียดกว่า โดยคำนวณค่า *NCP* ของแต่ละแอททริบิวท์ที่ถูกเจเนรัลไลซ์ แอททริบิวท์ต่างชนิดกันจะใช้วิธีคำนวณค่า *NCP* ต่างกัน โดยค่า *NCP* ของแอททริบิวท์เชิงตัวเลขจะคำนวณดังสมการที่ (5)

$$NCP^E(R'_i(j)) = \frac{\max^E(R'_i(j)) - \min^E(R'_i(j))}{\max(R(j)) - \min(R(j))} \quad (5)$$

เมื่อกำหนดให้ E เป็นคลาสเท่าเทียมและ $R(j)$ เป็นแอททริบิวท์ที่ j ซึ่งอยู่ในคลาส E ค่า *NCP* ในสมการที่ (5) คือผลหารระหว่างพิสัยของแอททริบิวท์ที่ j ของระเบียนที่ i ที่ถูกแปลงค่าแล้วและพิสัยของแอททริบิวท์ที่ j ของระเบียนเดียวกันที่ยังไม่ถูกแปลงค่า ส่วนแอททริบิวท์เชิงลักษณะจะใช้วิธีการคำนวณค่า *NCP* ดังสมการที่ (6)

$$NCP^E(R'_i(j)) = \begin{cases} 0 & ,|u| = 1 \\ \dfrac{|u_i(j)|}{|R(j)|} & ,otherwise \end{cases} \quad (6)$$

ในการคำนวณค่า *NCP* ของแอททริบิวท์เชิงลักษณะนั้นจะมีการกำหนด*แผนภาพต้นไม้แสดงหมวดหมู่* (*Taxonomy tree*) [17] ซึ่งคล้ายกับลำดับชั้นการเจเนรัลไลเซชันค่า โดยแผนภาพต้นไม้จะประกอบด้วย u ซึ่งเป็น*โหนด* (*Node*) ในลำดับชั้นการเจเนรัลไลเซชัน และ*โหนดใบ* (*Leaf node*) ของ u ซึ่งต่อยอดจาก u จะแสดงค่าที่เป็นไปได้ทั้งหมดของ u ต่อเนื่องไปคล้ายกิ่งก้านของต้นไม้

ในสมการที่ (6) $|u_i(j)|$ จะคืนค่าจำนวนโหนดใบ ของ u ในลำดับชั้นที่แอททริบิวท์ที่ $R_i'(j)$ อยู่ ส่วน $|R(j)|$ คือคาร์ดินาลิตี้ (Cardinality) ของเซต $R(j)$ จะเห็นว่าวิธีการคำนวณค่า *NCP* ข้างต้น ครอบคลุมถึงการเข้ารหัสใหม่แบบเฉพาะแห่งที่อนุญาตให้แปลงแอททริบิวท์ที่เหมือนกันไปเป็นค่าโดยทั่วไปที่ต่างกัน ซึ่งอาจทำให้ค่า *NCP* ที่คำนวณได้มีค่าต่างกัน นอกจากนี้หากต้องการให้น้ำหนักกับบางแอททริบิวท์ก็สามารถคำนวณค่า*เวจท์เซอร์เทนทีพีนัลที* (*Weighted certainty penalty*) โดยคูณค่า *NCP* ด้วยค่าน้ำหนัก (w_j) ของแอททริบิวท์ที่ j

จากนั้นจึงคำนวณค่า *NCP* เฉลี่ยของ T' ดังนี้

$$NCP(T,(T')) = \frac{1}{nd} \cdot \sum_{i=1}^{n} \sum_{j=1}^{d} NCP^E(R'_i(j)) \quad (7)$$

4. อัลกอริธึมที่ใช้แปลงข้อมูล

มีการเสนออัลกอริธึมในการแปลงข้อมูลหลายอัลกอริธึมซึ่งต่างพยายามหาคำตอบที่ดีที่สุดและทำให้ค่าการสูญเสียสารสนเทศต่ำที่สุด อัลกอริธึมเหล่านั้นใช้วิธีหลายวิธี เช่น พยายามเจเนรัลไลซ์แอททริบิวท์จำนวนน้อยที่สุดเท่าที่จำเป็น พยายามเจเนรัลไลซ์ไปยังลำดับชั้นต่ำที่สุดเท่าที่จำเป็น หรือพยายามจัดข้อมูลที่มีความใกล้เคียงกันไว้ด้วยกัน

เมื่อพิจารณาจากวิธีการเจเนรัลไลเซชันแล้ว จะสามารถจำแนกอัลกอริธึมเหล่านั้นออกเป็น 3 หมวด [17] ได้ดังนี้

4.1 การเจเนรัลไลเซชันโดยอาศัยลำดับชั้น (Hierarchy-based generalization)

การเจเนรัลไลเซชันโดยอาศัยลำดับชั้นหมายถึงในการเจเนรัลไลเซชันจะพิจารณาเซตของแอททริบิวท์ที่จะเจเนรัลไลซ์จากลำดับชั้นการเจเนรัลไลเซชัน [16, 17] งานวิจัย [3, 16, 22] เสนอการเจเนรัลไลเซชันโดยอาศัยลำดับชั้น โดยกำหนดลำดับชั้นการเจเนรัลไลเซชันสำหรับแต่ละแอททริบิวท์ j ในเซตตัวบ่งชี้ทางอ้อม Q_T ไว้ล่วงหน้า จากนั้นอัลกอริธึมจะพยายามหาคำตอบที่ดีที่สุด Samarati [3] หาคำตอบที่ดีที่สุดโดยพิจารณาจากค่าที่ใช้วัดความแตกต่างของแอททริบิวท์ในตาราง ซึ่งในงานวิจัยหลายงานมักเรียกว่า*ฟังก์ชันระยะทาง* (*Distance function*) [1, 17] และแต่ละงานวิจัยให้นิยามของฟังก์ชันระยะทางไว้แตกต่างกัน ในงานของ Samarati ได้พิจารณาฟังก์ชันระยะทางของแต่ละแอททริบิวท์ใน Q_T หรือเรียกว่าส่วนสูง (Height: h) [16] ซึ่งเป็นระยะทางนับจากชั้นล่างสุดของลำดับชั้นการเจเนรัลไลเซชันไปจนถึงลำดับชั้นสูงสุดที่ทำการเจเนรัลไลเซชัน หากส่วนสูงที่สูงสุดของตาราง T คือ h แล้ว อัลกอริธึมจะเริ่มตรวจสอบแต่ละแอททริบิวท์ j จากส่วนสูง $h/2$ ว่ามีจำนวนระเบียนผ่านคุณสมบัติเคแอนโนนิมิทีหรือไม่ หากผ่านจะตรวจสอบที่ส่วนสูง $h/4$ หากไม่ผ่านจะตรวจสอบที่ส่วนสูงกึ่งกลางระหว่างลำดับชั้นที่ไม่ผ่านและลำดับชั้นที่ผ่านก่อนหน้านั้น (เช่น $3h/4$ ซึ่งอยู่ระหว่าง $h/2$ และ $h/4$) และจะทำเช่นนี้ต่อไปโดยลดส่วนสูงลงครึ่งหนึ่ง สุดท้ายอัลกอริธึมจะแปลงทุกแอททริบิวท์ในเซต Q_T ไปตามลำดับชั้นต่ำสุดที่หาคำตอบได้ ข้อด้อยของการเจเนรัลไลเซชันของ Samarati คือต้องตรวจสอบทุกแอททริบิวท์ และยังอาจทำให้เกิดการสูญเสียสารสนเทศเกินความจำเป็น เนื่องจากหากมีเพียง 1 ระเบียนที่ต้องแปลงค่าแอททริบิวท์ที่ j แล้ว ระเบียนอื่นทั้งหมดในตารางก็ต้องถูกแปลงค่าแอททริบิวท์ที่ j ไปที่ลำดับชั้นคำตอบเดียวกัน แม้ว่าระเบียนอื่นนั้นอาจมีแอททริบิวท์อื่นที่ผ่านคุณสมบัติเคแอนโนนิมิทีแล้ว

Fung และคณะ [23] เสนอการเจเนรัลไลเซชันโดยอาศัยลำดับชั้นเช่นกัน แต่อนุญาตให้มีการยุบระเบียนที่ซ้ำกันแล้วแสดงโดยใช้ความถี่ (Frequency) จากนั้นอัลกอริธึมจะเริ่มจากลำดับชั้นเจเนรัลไลเซชันสูงสุด (ซึ่งสูญเสียสารสนเทศมากที่สุด) จะถูก*สเปเชียลไลซ์* (*Specialize*) คือการแทนด้วยค่าในลำดับชั้นถัดลงมา การสเปเชียลไลซ์จะเลือกแอททริบิวท์ใดก่อนขึ้นอยู่กับค่าคะแนน (Score) ซึ่งคำนวณจาก*อัตราการได้สารสนเทศ* (*Information gain*) และ*การสูญเสียแอนโนนิมิที* (*Anonymity loss*) รวมทั้งจำนวนระเบียนที่จะต้องผ่านคุณสมบัติเคแอนโนนิมิที ทำให้เพิ่มความยืดหยุ่นในการเจเนรัลไลเซชันเนื่องจากผู้ใช้สามารถกำหนดค่าอัตราการได้สารสนเทศและการสูญเสียแอนโนนิมิทีได้ในทุกขั้นตอน แต่ยังคงข้อด้อยเรื่องการสูญเสียสารสนเทศจากการแปลงแอททริบิวท์ในเซตตัวบ่งชี้ทางอ้อมของทุกระเบียนไปตามลำดับชั้นคำตอบเดียวกัน การเจเนรัลไลเซชันโดยอาศัยลำดับชั้นที่กล่าวมาทั้งสองคล้ายกับ*การเจเนรัลไลเซชันแบบมิติเดียว* (*Single dimensional generalization*) ซึ่งจะพิจารณาเจเนรัลไลซ์แอททริบิวท์ในเซตตัวบ่งชี้ทางอ้อมแต่ละแอททริบิวท์แยกกัน [16] นอกจากนี้ยังมีหลายงานวิจัยที่เสนออัลกอริธึมซึ่งถูกพัฒนาจากการเจเนรัลไลเซชันแบบมิติเดียว แต่ไม่ใช้วิธีกำหนดลำดับชั้นการเจเนรัลไลเซชันไว้ล่วงหน้า [20]

4.2 การเจเนรัลไลเซชันโดยอาศัยการแบ่งกลุ่มข้อมูล (*Partition-based generalization*)

อัลกอริธึมที่แปลงข้อมูลโดยอาศัยการแบ่งกลุ่มข้อมูล [16] ในขั้นแรกจะมีการแบ่งกลุ่มข้อมูลโดยอาศัยเกณฑ์ต่างๆ กัน ขั้นที่สองจึงทำการเจเนรัลไลเซชัน เช่น Meyerson และ Williams [6] เสนอให้คำนวณค่าเส้นผ่าศูนย์กลาง (Diameter) ซึ่งคำนวณจากระยะทางสูงสุดระหว่างสองข้อมูลที่ยังไม่ถูกเจเนรัลไลซ์ เนื่องจาก [6] พิจารณาว่าข้อมูลคือกลุ่มของตัวอักษร ระยะทางระหว่างสองข้อมูลจึงคำนวณจากจำนวนตัวอักษรที่แตกต่างกัน Meyerson และ Williams แบ่งระเบียนออกเป็นกลุ่มหรือซับเซต (Subset) ที่มีเส้นผ่าศูนย์กลางหลายๆ ขนาดโดยใช้ค่าผลรวมเส้นผ่าศูนย์กลางน้อยที่สุด (*Minimum diameter sum*) แต่ละซับเซตอาจมีสมาชิก (ระเบียน) ซ้ำกันก็ได้ จากนั้นพิจารณาซับเซตที่มีระเบียนซ้ำกันและแบ่งโดยไม่ให้เกิดระเบียนซ้ำ โดยแต่ละซับเซตจะต้องมีจำนวนสมาชิกอยู่ระหว่าง k ถึง $2k$ -1 แล้วจึงทำการเจเนรัลไลเซชัน

LeFevre และคณะ [14] ได้อธิบาย*การแบ่งกลุ่มแบบมิติเดียว* (*Single-dimensional partitioning*) ว่าเป็นการแบ่งกลุ่มโดยพิจารณาทีละแอททริบิวท์ในเซตตัวบ่งชี้ทางอ้อม ซึ่งเมื่อแบ่งแล้วทำให้ค่าสถิติของแอททริบิวท์นั้นๆ ของกลุ่มเท่ากับค่าสถิติก่อนการแบ่งกลุ่ม พร้อมทั้งได้เสนอว่าในขั้นแรก ให้ใช้*การแบ่งกลุ่มหลายมิติแบบเคร่งครัด* (*Strict multi-dimensional partitioning*) ที่พิจารณาแบ่งกลุ่มโดยใช้แอททริบิวท์ในเซตตัวบ่งชี้ทางอ้อมหลายแอททริบิวท์ กลุ่มที่แบ่งได้จะต้องสามารถแม็พ (map) แต่ละสมาชิกในกลุ่มไปยังค่าสถิติสรุป (Summary statistic) ซึ่งเป็นตัวแทนของกลุ่ม ค่าสถิติสรุปที่ใช้ใน [14] ได้แก่ค่าพิสัย (Range) และค่าเฉลี่ย (Mean) เนื่องจากสามารถนำไปคำนวณค่าสถิติอื่นๆ ได้เช่นค่าสูงสุด-ต่ำสุด (Max-Min) ผลรวม (Sum) ค่าเฉลี่ยเลขคณิต (Average) และเมื่อแบ่งกลุ่มแล้วแต่ละกลุ่มต้องมีสมาชิกอย่างน้อย k ระเบียน จากนั้นในขั้นที่สองจึงทำการเจเนรัลไลเซชันระเบียนในแต่ละกลุ่มโดยใช้วิธีเข้ารหัสใหม่แบบเฉพาะแห่ง

งานวิจัย [19, 21, 24] เสนอให้ใช้ค่าการสูญเสียสารสนเทศในการแบ่งกลุ่มข้อมูล Xu และคณะ [21] วัดการสูญเสียสารสนเทศของกลุ่มด้วยค่า *NCP* เฉลี่ย และเสนออัลกอริธึมบนสู่ล่าง (Top-down algorithm) ในการแบ่งกลุ่มระเบียน โดยเริ่มจากจำนวนระเบียนทั้งหมดของตาราง หากยังมีค่ามากกว่า k ให้แบ่งเป็น 2 กลุ่มโดยพิจารณาความใกล้เคียงกันของระเบียนในกลุ่มด้วยค่า *NCP* และมีเงื่อนไขว่ากลุ่มใหม่หนึ่งกลุ่มต้องมีอย่างน้อย k ระเบียน หากกลุ่มใหม่ใดมีจำนวนระเบียนมากกว่า k ให้แบ่งต่อไป ดังแสดงในรูปที่ 5 อัลกอริธึมบนสู่ล่างจะทำงานร่วมกับอัลกอริธึมล่างสู่บน (Bottom-up algorithm) ซึ่งได้อธิบายไว้ในหมวดอัลกอริธึมที่ใช้วิธีการจัดกลุ่มข้อมูล

Top-down (Input T, k) {
  WHILE | T | > k DO {
    Partition T into T_1, T_2 such that T_1, T_2 are more local and either T_1 or T_2 has at least k records;
    IF | T_1 | > k THEN Recursively partition T_1 ;
    IF | T_2 | > k THEN Recursively partition T_2 ;
Output (T_1, T_2, ...) ; }

รูปที่ 5 อัลกอริธึมบนสู่ล่าง (Top-down algorithm)

Greedy (Input T, k) {
  WHILE | T | >= k DO {
    Find largest group G in Q_T with minimum Information loss where | G | >= k ;
    Put to Equivalence class E with grouping condition;
    Add grouping condition ; Generalize records; }
Output (T') ; }

รูปที่ 6 อัลกอริธึมกรีดดี (Greedy algorithm)

Huda และคณะ [24] ทบทวนอัลกอริธึมกรีดดี (Greedy algorithm) [6, 19, 21] ซึ่งจะค้นหากลุ่มที่มีขนาดใหญ่ที่สุดจากแต่ละแอททริบิวท์ในเซต Q_T เนื่องจากกลุ่มที่มีขนาดใหญ่ที่สุดจะมีการสูญเสียสารสนเทศต่ำสุดเมื่อถูกเจเนรัลไลซ์ และกลุ่มที่ค้นหาได้จะต้องมีขนาดตั้งแต่ k ระเบียนขึ้นไป แล้วจัดไว้ในคลาสเท่าเทียมเดียวกัน จากนั้นจะทำซ้ำโดยค้นหากลุ่มอื่นในแอททริบิวท์ในเซต Q_T ต่อไป ดังรูปที่ 6

4.3 การเจเนรัลไลเซชันโดยอาศัยการจัดกลุ่มข้อมูล (*Clustering-based generalization*)

มีหลายอัลกอริธึมที่แปลงข้อมูลโดยอาศัยการจัดกลุ่มข้อมูล เช่น อัลกอริธึมล่างสู่บน [21] ซึ่งจัดกลุ่มระเบียนโดยเริ่มจากกำหนดให้แต่ละระเบียนเป็นหนึ่งกลุ่ม หากกลุ่มใดยังมีจำนวนระเบียนน้อยกว่า k จะนำไปรวมกับกลุ่มอื่น โดยมีเงื่อนไขว่าเมื่อรวมกันแล้วค่า *NCP* เฉลี่ยของกลุ่มใหม่จะต้องมีค่าต่ำสุด นอกจากนี้หากกลุ่มใดมีจำนวนระเบียนในกลุ่มมากกว่า $2k$ จะถูกแบ่งย่อยโดยให้กลุ่มย่อยนั้นมีจำนวนระเบียนอย่างน้อย k ระเบียน ดังรูปที่ 7 อัลกอริธึมล่างสู่บนจะถูกใช้งานร่วมกันอัลกอริธึมบนสู่ล่าง โดยหลังจากอัลกอริธึมบนสู่ล่างได้แบ่งกลุ่มระเบียนแล้ว สำหรับกลุ่มใดที่มีสมาชิกน้อยกว่า k จะใช้อัลกอริธึมล่างสู่บนในการรวมกลุ่มระเบียน เพื่อให้ตารางผลลัพธ์มีคุณสมบัติเคแอนโนนิมิที

Bottom-up (Input T, k) {
Initialize: Create a group G for each record;
WHILE existing G such that $|G| < k$ DO {
FOR each group G DO {
Find group H such that *NCP*($G \cup H$) is minimum;
Merge G and H; }
FOR each group G such that $|G| > 2k$ DO {
Split G into G's which has at least k records; }
Output (T') ; }

รูปที่ 7 อัลกอริธึมล่างสู่บน (Bottom-up algorithm)

Byun และคณะ [17] เสนอว่าเมื่อจัดกลุ่มข้อมูลโดยแต่ละกลุ่มต้องมีอย่างน้อย k ระเบียนแล้ว จะต้องทำให้ค่าฟังก์ชันระยะทางระหว่างระเบียนภายในกลุ่มมีค่าน้อยที่สุด และมีการกำหนดให้ค่าการสูญเสียสารสนเทศทั้งหมดของการจัดกลุ่มต้องมีค่าไม่เกินค่าคงที่ (Constant: c) ที่กำหนดไว้ ส่วน Nergiz และ Clifton [18] เสนออัลกอริธึมที่คล้ายอัลกอริธึมเคมีน (k-means) คือเสนอให้จัดทีละระเบียนเข้ากลุ่มที่เหมาะสมที่จะทำให้ค่าการสูญเสียสารสนเทศต่ำที่สุด หากไม่พบกลุ่มที่เหมาะสมจะสร้างกลุ่มใหม่จากระเบียนนั้น เมื่อในกลุ่มมี k ระเบียนแล้วจะเจเนรัลไลซ์ระเบียนในกลุ่มไปยังโครงสร้างลำดับชั้นโดเมนที่ถูกกำหนด เมื่อจัดทุกระเบียนเข้ากลุ่มหมดแล้ว กลุ่มที่ยังมีจำนวนระเบียนยังไม่ครบจะถูกนำมาจัดรวมกันโดยพิจารณาจากค่า *AM* ที่ต่ำที่สุด หากยังเหลือกลุ่มที่ยังมีจำนวนระเบียนยังไม่ครบอีก ระเบียนสมาชิกในกลุ่มเหล่านั้นจะถูกแปลงข้อมูลด้วยวิธีการปกปิด (Suppression)

5. การโจมตีที่เป็นไปได้

แม้ว่าเทคนิคเคแอนโนนิไมเซชันจะรับรองระดับความเป็นส่วนตัวได้ แต่ยังมีวิธีการอื่นๆ ที่สามารถโจมตีเพื่อละเมิดความเป็นส่วนตัวของข้อมูลได้อีก เช่นการโจมตีโดยสังเกตจากลำดับของข้อมูล (Unsorted matching attack) [10] ซึ่งอาศัยช่องว่างเนื่องจากการเปิดเผยตารางข้อมูลเดียวกันแต่ละครั้งอาจมีการเจเนรัลไลเซชันไว้ต่างกัน ดังตัวอย่างในรูปที่ 8

Date-bill	Paid
//2552	3300
//2556	300
//2551	320
//2556	2800

Date-bill	Paid
**/03/255*	33**
**/12/255*	3**
**/09/255*	3**
**/04/255*	28**

รูปที่ 8 การโจมตีโดยสังเกตจากลำดับของข้อมูล

หากนำข้อมูลในรูปที่ 8 (ซ้ายและขวา) มาสังเกตก็จะพบว่ามีลำดับตรงกันและทำให้สามารถได้ข้อมูลตัวบ่งชี้ทางอ้อมเพิ่มขึ้น ดังนั้นหากสลับลำดับของข้อมูลในการเปิดเผยตารางข้อมูลเดียวกันแต่ละครั้ง จะทำให้การโจมตีโดยสังเกตลำดับทำได้ยากขึ้น

นอกจากนี้ยังพบว่าโอกาสการโจมตีมักเกิดจากการเปิดเผยตารางข้อมูลเดียวกันต่างเวลากัน (Sequential data release) [25, 26, 27] เช่น ในการเปิดเผยข้อมูลครั้งแรกบางแอททริบิวท์อาจไม่ถูกเจเนรัลไลซ์หรือไม่ถูกปกปิด ในการเปิดเผยข้อมูลครั้งต่อมาแอททริบิวท์ที่ไม่ถูกเจเนรัลไลซ์นี้สามารถถูกใช้เป็นตัวบ่งชี้ทางอ้อมได้ ดังนั้นในการเจเนรัลไลเซชันครั้งต่อมา ควรรวมแอททริบิวท์นี้ไว้ในเซตตัวบ่งชี้ทางอ้อมด้วย

เนื่องจากฐานข้อมูลเองก็มีการเปลี่ยนแปลงอย่างสม่ำเสมอ เช่น มีการเพิ่มระเบียนและเปลี่ยนแปลงข้อมูลในระเบียน ดังนั้นในการเปิดเผยตารางข้อมูลแต่ละครั้งควรนำเอาตารางข้อมูลเดียวกันที่เปิดเผยไปแล้วมาพิจารณาร่วมด้วย [28] นั่นคือ หากให้ตาราง T_0 เป็นตารางดั้งเดิมและตาราง T_0' คือตารางที่ถูกแปลงข้อมูลและเปิดเผยไปแล้ว ให้ตาราง T_t เป็นตารางดั้งเดิมที่เวลาต่อมาการเจเนรัลไลเซชันครั้งหลังจึงควรพิจารณา $T_0' \cup (T_t - T_0)$ แทนที่จะพิจารณาเฉพาะ T_t

6. บทสรุปและข้อเสนอแนะ

บทความนี้เป็นบทความวิชาการที่ทบทวนการศึกษาและวิจัยเกี่ยวกับเทคนิคเคแอนโนนิไมเซชันเพื่อรักษาความเป็นส่วนตัวของข้อมูลบุคคล โดยได้อธิบายปัญหาการระบุตัวบุคคลอีกครั้ง และทบทวนวิธีต่างๆ ที่ใช้ประเมินการสูญเสียสารสนเทศและอัลกอริธึมต่างๆ ที่ใช้แปลงข้อมูลเพื่อให้มีคุณสมบัติเคแอนโนนิมิที อัลกอริธึมเหล่านั้นสามารถแบ่งได้เป็น การเจเนรัลไลเซชันโดยอาศัยลำดับชั้นซึ่งพิจารณาจากลำดับชั้นการเจเนรัลไลเซชัน การ

เจเนรัลไลเซชันโดยการแบ่งกลุ่มข้อมูล และการเจเนรัลไลเซชันโดยการจัดกลุ่มข้อมูล ซึ่งสองหมวดหลังอาศัยหลักการพิจารณาแอททริบิวท์ในข้อมูล

เมื่อข้อมูลถูกแปลงโดยการเจเนรัลไลเซชันแล้ว จะทำให้เกิดการสูญเสียสารสนเทศ หลายอัลกอริธึมที่ใช้แปลงข้อมูลจึงพยายามรักษาสารสนเทศในข้อมูลโดยใช้วิธีคำนวณการสูญเสียสารสนเทศเข้าร่วมในขั้นตอนการแปลงข้อมูล เพื่อให้ข้อมูลที่แปลงแล้วคงไว้ซึ่งสารสนเทศสูงสุด ขณะเดียวกันยังคงรักษาความเป็นส่วนตัวด้วยคุณสมบัติเคแอนโนนิมิที

นอกจากอัลกอริธึมและการประเมินการสูญเสียสารสนเทศจะมีความสำคัญต่อความเป็นส่วนตัวและสารสนเทศในข้อมูลแล้ว การเลือกค่า k ให้เหมาะสมกับการใช้งานและระดับความเป็นส่วนตัว ยังเป็นเรื่องที่ต้องศึกษาร่วมกับการประเมินความเสี่ยงและการกำหนดนโยบายความเป็นส่วนตัวของแต่ละองค์กรที่มีการรวบรวมข้อมูลบุคคลเอาไว้ การจัดเก็บข้อมูลของแต่ละองค์กรเองก็มีความหลากหลาย ขึ้นอยู่กับการใช้งาน การออกแบบฐานข้อมูลและลักษณะของข้อมูลที่จัดเก็บ เนื่องจากการปกป้องความเป็นส่วนตัวของบุคคลที่มีข้อมูลในตารางมีความสำคัญอย่างยิ่ง ดังนั้นในการดำเนินการใดๆ เกี่ยวกับข้อมูลบุคคลจึงควรคำนึงถึงการรักษาความเป็นส่วนตัวในทุกขั้นตอน

7. เอกสารอ้างอิง

[1] Gionis A, Mazza A, Tassa T. k-Anonymization revisited. Proceeding of the 2008 IEEE 24th International Conference on Data Engineering; 2008 Apr 7-12; Cancun, Mexico. p. 744-753.

[2] 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.

[3] Samarati P. Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering archive; Volume 13 Issue 6, 2001 November. p. 1010-1027.

[4] Duncan G, Lambert D. Disclosure-limited data dissemination. Journal of the American Statistical Association, 81; 1986. p. 10-28.

[5] Agrawal R, Srikant R. Privacy-preserving data mining. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data; 2000 May 16-18; Dallas, Texas, USA. p. 439-450.

[6] Meyerson A, Williams R. On the complexity of optimal k-anonymity. Proceeding of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems; 2004 June 13-18; Paris, France. p. 223-228.

[7] Kim J, Winkler W. Masking microdata files. Proceeding of the Section on Survey Research Methods, American Statistical Association; 1995. p. 114-119.

[8] Domingo-Ferrer J, Mateo-Sanz J, Torra V. Comparing SDC methods for microdata on the basis of information loss and disclosure risk. Proceedings of ETK-NTTS; 2001; Luxemburg. p. 807-826.

[9] Yancey W, Winkler W, Creecy R. Disclosure risk assessment in perturbative microdata protection. Inference Control in Statistical Databases: From Theory to Practice. Lecture Notes in Computer Science, vol. 2316; 2002. p. 135-152.

[10] 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.

[11] Liang Z, Wei R. Efficient k-anonymization for privacy preservation. Proceeding of the 12th International Conference on Computer Supported Cooperative Work in Design; 2008 April 26-18; Xi'an, P.R. China. p. 737-742.

[12] Seisungsittisunti B. Incremental data transformation algorithm development for privacy preserving of associative classification [MSc thesis]. Chiang Mai: Chiang Mai University; 2010. (In Thai).

[13] Aggarwal G, Feder T, Kenthapadi K, Motwani R, Panigrahy R, Thomas D, Zhu A. Approximation algorithms for k-anonymity. Journal of Privacy Technology. 2005.

[14] LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian multidimensional k-anonymity. Proceeding of the 22nd International Conference on Data Engineering, 2006 April 3-7; Atlanta, GA, USA. p. 25.

[15] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information. Proceedings of the 7th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems; 1998 June 1-4; Seattle, WA, USA. p. 188.

[16] LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: efficient full-domain k-anonymity. Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data; 2005 June 13-17; Baltimore, MD, USA. p. 49-60.

[17] Byun JW, Kamra A, Bertino E, Li N. Efficient k-anonymization using clustering techniques. Proceeding of the 12th International Conference on Database Systems for Advanced Applications; 2007 April 9-12; Bangkok, Thailand. p.188-200.

[18] Nergiz ME, Clifton C. Thoughts on k-anonymization. Data & Knowledge Engineering, 2007; vol. 63, no. 3: p. 622–645.

[19] Huda MN, Yamada S, Sonehara N. An efficient k-anonymization algorithm with low information loss. Recent Progress in Data Engineering and Internet Technology, Lecture Notes in Electrical Engineering; 2013 January. p. 249-254.

[20] 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.

[21] Xu J, Wang W, Pei J, Wang X, Shi B, Fu WA. Utility-based anonymization using local recoding. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2006 August 20-23; Philadelphia, PA, USA. p. 785-790.

[22] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems; Volume 10 Issue 5, 2002 October. p. 571-588.

[23] Fung B, Wang K, Yu P. Top-down specialization for information and privacy preservation. Proceedings of the 21st International Conference on Data Engineering, 2005 April 5-8; Tokyo, Japan. p. 205-216.

[24] Huda MN, Yamada S, Sonehara N. On enhancing utility in k-anonymization. International Journal of Computer Theory and Engineering, 4(4); 2012 August. p. 527-532.

[25] 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.

[26] Riboni D, Pareschi L, Bettini C. JS-reduce: defending your data from sequential background knowledge attacks. IEEE Transactions on Dependable and Secure Computing, Volume 9 Issue 3, 2012 May. p. 387-400.

[27] Anjum A, Raschia G. Anonymizing sequential releases under arbitrary updates. Proceeding of the Joint International Conference on Extending Database Technology/International Conference on Database Theory 2013 Workshops; 2013 March 18-22; Genoa, Italy. p.145-154.

[28] Seisungsittisunti B, Natwichai J. Incremental privacy preservation for associative classification. Proceeding of the ACM 1st International Workshop on Privacy and Anonymity for Very Large Databases, CIKM-PAVLAD; 2009 November 6; Hong Kong, P.R. China. p. 37-44.