УДК 004.827

Студент Тавров Д.Ю.

Національний технічний університет України
«Київський політехнічний інститут»

АНАЛІЗ ДАНИХ ТА ЗАБЕЗПЕЧЕННЯ ПРИВАТНОСТІ В СОЦІАЛЬНИХ МЕРЕЖАХ

Abstract

Dan Tavrov, student

Data Mining and Providing Privacy in Social Networks

Along with various possibilities for self-expression, social networks bring powerful means for people to share their preferences, tastes, and opinions. Appropriate user groups, either formal or informal, can pose large interest for marketing services willing to research the demand. In this paper, we discuss the generic scheme of data analysis in social networks and pay special attention to privacy issues. We study various kinds of privacy breaches likely to occur and propose general solutions to them.

Вступ

Останніми роками спостерігається активний розвиток соціальних мереж. Участь у них бере все більша кількість людей задля висловлення своїх поглядів, установлення нових контактів, обміну релевантною інформацією. У результаті формуються великі масиви даних про користувачів, їхні вподобання, переваги, що можуть становити значний інтерес [1] для служб дослідження ринку.

Однак, доступ до такої інформації може мати небажані наслідки. Наприклад, загальновідомо, що окуляри та контактні лінзи є товари-замінювачі. Разом із тим, лінзи неможливо застосовувати без розчину для їхньої дезінфекції та зберігання. Нескладно уявити ситуацію, за якої виробники лінз та розчинів до них укладають картельну змову з метою витіснення виробників окулярів із відповідного сегменту ринку. Такі результати не можуть заохочуватися, а тому мають бути унеможливлені.

Постановка задачі

У даній роботі ставиться задача розробки загальної схеми аналізу даних в соціальних мережах та визначення необхідних етапів підготовки таких даних до публікування. Пропонується застосовувати методи

індивідуальної [2] та групової [3] анонімності для забезпечення потрібного рівня приватності в даних соціальної мережі до передачі їх третім сторонам.

Представлення даних про користувачів соціальної мережі

Для проведення аналізу дані потрібно представити в доступному форматі. У даній роботі пропонується організовувати їх у вигляді *мікрофайлу* **M**, який зручно подавати в матричній формі (Табл. 1). Так, мікрофайл можна розглядати як матрицю $\|z_{ij}\|_{\mu\times\eta}$, рядки r_i котрої відповідають користувачам соціальної мережі, а стовпці u_j – їх атрибутам (інформації про них).

Табл. 1. Матрична форма представлення мікрофайлу

		атрибути			
		u_1	u_2	…	u_η
користувачі	r_1	z_{11}	z_{12}	…	$z_{1\eta}$
	r_2	z_{21}	z_{22}	…	$z_{2\eta}$
	…	…	…	…	…
	r_μ	$z_{\mu 1}$	$z_{\mu 2}$	…	$z_{\mu\eta}$

Загальна схема підготовки мікрофайлу до публікації

Аналіз мікрофайлу соціальної мережі можна розбити на такі кроки:

1. Подача третьою стороною заявки, яка повинна містити мету майбутніх досліджень, обсяги вибірки, умови її формування, побажання щодо наявності окремих атрибутів, рівень участі власника даних у процесі їх аналізу.
2. Підготовка вибіркового мікрофайлу соціальної мережі.
3. Проведення аналізу даних вибіркового мікрофайлу (крок має місце, якщо це передбачено заявкою).
4. Анонімізація вибіркового мікрофайлу, тобто здійснення перетворення $\varphi:\mathbf{M}\to\mathbf{M}'$ для забезпечення необхідного рівня приватності даних, а також збереження достатнього рівня корисності (досягнення балансу між неможливістю розкриття даних та здатністю отримувати результати аналізу, близькі до початкових).

5. Публікація мікрофайлу.

Анонімізація мікрофайлу соціальної мережі

Під *анонімністю* суб'єкта [4] (користувача соціальної мережі) розумітимемо його властивість бути нерозрізненним у мікрофайлі.

Задачі забезпечення анонімності такого роду можна назвати задачами забезпечення *індивідуальної* анонімності даних. Наряду з ними, існують також задачі забезпечення *групової* анонімності, які полягають [5] у видозміні (для кожної групи користувачів окремо) масиву даних із метою недопущення розкриття конфіденційної інформації.

Різні види інформації в соціальній мережі можна проілюструвати за допомогою схеми, представленої на рис. 1. Існують конфіденційні дані, які становлять однозначну загрозу індивідуальній анонімності суб'єкта. Відповідні атрибути називають *ідентифікаторами* (серія/номер паспорта тощо). Їх неможливо анонімізувати і потрібно вилучати. З іншого боку, існують атрибути, розкриття значень яких не становить загрози для суб'єкта (володіння іноземними мовами чи футбольні уподобання). Їх узагалі не варто модифікувати. До інших атрибутів варто застосовувати методи індивідуальної та групової анонімності:

1. Методи індивідуальної анонімності:

а) *рандомізація* (зашумлення) даних, тобто додавання до деяких атрибутів мікрофайлу інформаційного шуму;

б) *k-анонімізація*: досягнення розподілу значень атрибутів, за якого будь-яка їхня комбінація відповідає щонайменше *k* користувачам;

в) *узагальнення*, яке передбачає заміну деяких конкретних значень атрибутів їхніми узагальненими аналогами;

г) *обмін даними*, що полягає в обміні деякими атрибутовими значеннями між двома або більше користувачами.

2. Методи групової анонімності:

а) *маскування розподілу*: визначають конкретний розподіл значень одних атрибутів іншими, а потім намагаються його замаскувати, прибравши або приховавши екстремальні точки такого розподілу;

б) *розрив залежностей*: визначають факт залежності між значеннями двох (або більше) атрибутів, а потім намагаються розірвати таку залежність шляхом викривлення відповідних розподілів;

в) *маскування інших властивостей*: окремі розподіли можуть мати додаткові властивості (внутрішня циклічність тощо), які можуть стати небажаними для розкриття.

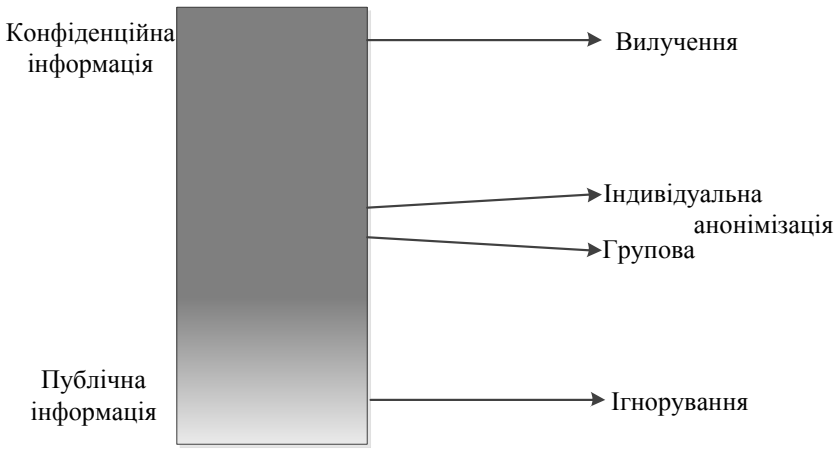

Рис. 1. Класифікація інформації в мікрофайлі соціальної мережі

Висновки

У рамках даної роботи було визначено загальну схему аналізу даних у соціальній мережі та забезпечення їх приватності при публікації чи передачі третій стороні, класифіковано інформацію за критерієм конфіденційності та запропоновано різні шляхи її анонімізації. Відкритими лишаються питання практичної реалізації пропонованої схеми.

Література

1. *Buddy Media*. Reaching Customers in Local Markets [Електронний ресурс]. – 2010. – 5 с. – Режим доступу: http://corpsite2.jillian.dev.buddymedia.com/news-room/wp-content/uploads/2010/08/Buddy-Media-_Harris_interactive_poll.pdf.
2. *Fung B., Wang K., Chen R., Yu P.* Privacy-Preserving Data Publishing: A Survey on Recent Developments // ACM Computing Surveys. – 2010. – Vol. 42(4). – P. 66-118.
3. *Chertov O., Tavrov D.* Data Group Anonymity: General Approach // International Journal of Computer Science and Information Security. – 2010. – Vol. 8(7). – P. 1-8.
4. *Pfitzmann A., Hansen M.* A Terminology for Talking about Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management, Version v0.34 [Електронний ресурс]. – 2010. Режим доступу: dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf.
5. *Chertov O.* Group Methods of Data Processing. – Raleigh: Lulu.com. – 2010. – 156 p.