社交媒體用戶基本信息曝露之隱私保護

劉妍 *陳榮靜 *王樹義

天津師範大學管理學院 朝陽科技大學資訊管理系 天津師範大學管理學院

Lindayan5588@163.com *crching@yut.edu.tw *nkwshuyi@gmail.com

摘要

互聯網的快速發展，為用戶帶來極大便利的同時，也給用戶信息隱私曝露帶來了很多的困擾。本文選取FACEBOOK社交媒體上的100位用戶的基本信息進行分群分析。通過實驗將用戶的資料集分為五群，並將分群結果進行分類討論。通過對用戶基本信息的分析，尋找用戶基本信息曝露中的危險因素。給用戶提供實質性的建議，為社交媒體的健康發展提供重要的保障。

關鍵詞：社交媒體；用戶基本信息曝露；隱私保護

Abstract: The rapid development of the Internet has brought great convenience to the users, but also brought a lot of troubles to the users' information privacy exposure. This paper selects the basic information of 100 users on FACEBOOK social media for cluster analysis. The data sets of users are divided into five groups by experiment, and the clustering results are classified and discussed. Through the analysis of the basic information of the users, the risk factors of the users' basic information exposure are search. To provide users with substantive recommendations for the healthy development of social media to give an important protection.

Keywords: social media; user basic information exposure; privacy protection

1. 簡介

隨著互聯網，移動互聯網愈發深刻地影響著人們的社會行為，用戶暴露在網絡環境中的維度和幾率大大增加，用戶信息曝露的威脅已經成為關乎每個網絡用戶切身利益的重要問題。特別是在強大的智能終端，豐富多彩的服務應用，日益增長的網絡關聯等因素作用下，用戶越來越多信息的隱私不可避免地主動，無意或者惡意被洩露。根據《中國國民權益保護調查報告2016》中指出54%的網民認為個人信息洩露嚴重，其中21%的網民認為非常嚴重，84%的網民親身感受到了由於個人的信息洩露帶來的不良影響。在接受調查的網民中，有 47%的網民遇到過在“社交軟件上冒充親朋好友進行詐騙”的情況，並且有相當一部分的網民因收到詐騙信息而遭受到錢財的損失[1]。

計算機、網絡等信息技術的快速發展，社交媒體運營商所擁有的數據以指數規模增長。這些微觀的數據日益以公開發佈的形式成為一種公共產品，但是由於微觀數據發佈之後數據發佈者既無法控制數據用戶對數據的使用方式，是否會惡意使用數據。也不了解用戶信息曝露的內容是否會對用戶自身產生一定的危險[2]。而作為網絡應用提供商則希望利用盡可能多的隱私信息獲得更多的利益，其他利益相關方也想從中賺取更多的價值，用戶信息曝露的內容被多方所持有。因此用戶需要將控制隱私的權利把握在自己手中，更好的保證自己的隱私安全。由此可見對隱私問題的深入研究具有重大的現實意義。

社交媒體的出現為用戶隱私問題提供了很好的研究環境，它不僅擁有與現實社會極為接近的用戶規模，社會關係，複雜結構，行為規律。還擁有比現實社會無法相比的易獲取，易分析，易驗證等特徵。隨著人們對於隱私信息越來越重視，用戶在進行信息曝露的過程中，在進行安全社交的過程中用戶的隱私安全如何能得到有效地保證是現在研究的熱點問題。用戶量超過一千萬的社交媒體高達上百種。例如：Facebook，Twitter，LinkIn，Wechat，Weibo 等。并且隨

著“互聯網+”的推廣，小眾社交媒體產品的不斷湧現，社交媒體的來源不明確，使得用戶的隱私曝露安全更加沒有保障。因此面向隱私保護的社交媒體用戶基本信息曝露研究顯得尤為重要。

2.文獻探討

國內外對於社交媒體的研究也是十分重視。所涉略的領域也十分的廣泛。從用戶行為習慣視角來看，王娜，許大辰在《移動社交網絡中個人信息保護現狀的調查與分析》中運用問卷星，跨越多省對移動社交網絡中個人信息安全問題進行抽樣調查。其中有效問卷 535 份。從調查結果的數據中分析發現，影響個人信息安全的用戶行為習慣主要五個方面。分別是，使用與用戶有關聯的個人信息做密碼並且在多個社交網站上使用同一個密碼；安裝使用未知來源的移動社交應用的同時不經常更新系統安全軟件；不使用終端系統的隱私控制功能；擔心未知信息泄露但卻經常開啟 GPS 功能；用戶更加傾向於上傳隱私數據至網盤。這五個方面用戶行為習慣會增強用戶隱私曝露的可能性[3]。

王樹義，朱娜在《移動社交媒體用戶隱私保護對策研究》中對移動互聯網的社交，本地及移動屬性對社交媒體用戶隱私提出新的挑戰，除了傳統的用戶身份，身體狀況，家庭背景，社會關系等隱私泄露外，還包括移動社交媒體用戶的位置隱私，用戶發表狀態暴露自己隱私等用戶行為，服務提供商引發的隱私泄露和黑客攻擊引發的隱私泄露等四個。基於以上隱私泄露問題，作者提出從政府，行業協會，服務提供商和用戶自身等多方面協同合作共同保護用戶隱私綜合化方案[4]。

從社交媒體用戶隱私視角來看，朱侯在《社交媒體用戶隱私關註的心理機制研究》中以國內外重要期刊的論文為研究對象，對用戶的隱私心理的相關文獻進行梳理和分析，並建立隱私關註的集成模型。分成前因變量研究(隱私經歷，隱私意識，個性差異，人口統計學特徵，文化），結果變量研究（規章，行為反應）和調節變量研究（隱私計算，信任）三類[5]。

郭龍飛在《社交網絡用戶隱私關註動態影像因素及行為規律研究》中以突發事件和媒體報道為考察對象。構建社交網絡隱私關註動態影像因素模型，在動態環境中觀察用戶隱私關註的變化，並且對比個人和群體對隱私關註的不用作用效果，探討不同功能社交網絡對隱私關註的不用作用結果，發現用戶社交網絡隱私關註的影響規律。除此之外，還構建了社交網絡突發事件影響下隱私行為演化博弈模型。考慮在突發事件擾動下，用戶群體和網站群體在不同利益策略下，用戶的選擇和突變行為演化規律，以此討論用戶與網站以及第三方策略對社交網絡發展所起的作用[6]。

苗振興在《基於社交媒體的人物分析技術研究》中，研究微博媒體上人物在交互過程中產生的關系和內容。通過關系分析挖掘出用戶的密友，校友和同事等。通過內容分析用來做話題識別，話題分類，興趣挖掘等針對關系分析，運用了壹種基於機構識別的社區劃分算法。針對話題識別，使用了壹種基於主題模型的相似度計算方法。利用多維度推薦的人物推薦算法和信息推薦算法。通過實驗驗證，根據人物分析的方法開發出基於新浪微博的人物分析系統[7]。

從研究方法視角出發，陳希在《基於 R 語言數據挖掘的社交網絡客戶細分研究》中，最早運用數據挖掘工作中的聚類和決策樹算法，對社交網絡的客戶細分進行了深入的探討並最終得出可指導實踐的社交網絡客戶細分規則[8]。

廉捷在《基於用戶特徵的社交網絡數據挖掘研究》結合交叉學科的研究方法，首先優化了基於 Nutch 的分布式網絡爬蟲系統，之後提出基於規則與基於 Wrapper 的網頁解析模型獲取數據，實證分析微博社交網絡特徵與用戶特徵。對新浪微波在線數據進行了多維度的分析。包括用戶特徵、微博特徵、時間與演化特徵等，研究了社交網絡中的用戶個性化推薦算法，最後提出了基於機器學習的信息預測方法。結合微博社交網絡的真實數據，分析影響用戶連接關系以及微博傳播的主要特徵因素，建立基於 SVM 的用戶連接關系模型與機遇邏輯回歸的用戶微博轉發模型[9]。

陳玉英在《基於機器學習的社交網絡用戶特徵分析》中基於社交網絡研究多學科交叉結合的特點，采用機器學習，文本處理等手段對以微博為代表的社交網絡平臺整體特性、屬性特徵、關系特徵、文本特徵進行綜合研究，同時對

社交網絡用戶的網絡特性實現可視化[10]。

王麗文在《基於社交網絡的數據挖掘研究》中先用Python從社交網絡新浪微博上提取熱門話題和參與話題的用戶信息，然後采用分級聚類算法對提取的熱門話題進行聚類。同時使用協作型過濾算法對提取的數據進行分析。通過對兩種算法的結果進行分析比較。通過得出的結果可對用戶推薦屬於同壹類的話題，或者推薦經過分析得出來的用戶最感興趣的話題，從而可以使得新浪微博的話題功能使用率更廣[11]。

從社交媒體運營商的視角出發，Cheng-Hung Tsai，Han-Wen Liu 等人提出在網絡信息爆炸時代，用戶使用各種社交網站，通過社交平臺進行互動。用戶連接到各種形式的社交網站隨時隨地地通過互聯網連接，社交媒體運營商可以較清晰的知道用戶的信息以及社會關系。通過了解用戶的個人喜好和興趣，推薦用戶的個人喜好廣告，產品推薦，推薦文章等多元化的個人社會服務。這樣的方式，既可以滿足用戶的需求，又可以增加產品的點擊率和曝光率。此文通過人們在社交網絡上，對用戶彼此之間互動信息的語氣進行分析所產生的技術發展趨勢，獲取三方面的數據，用戶粉絲頁面，用戶塗鴉墻消息信息，朋友的主頁等三種信息，用於用戶個人喜好的分析，並且通過個人的數據量，找出相應的不同群體的個人喜好類別，從而更好的推薦個人廣告和產品推薦服務。通過實際的驗證，研究效果顯著提高[12]。

從用戶隱私曝光的視角來看，Mohamed Bourimi, Ricardo Tesoriero 等人提出探討關於社交媒體應用的多模式用戶界面的建模的隱私和安全問題。所提出的方法描述了隱私和安全問題是如何從用戶界面的角度建模，這個模型是如何與開發多模式，多平臺用戶界面的四層概念框架。該方法還解釋了如何將這些模型適應社交媒體應用程序的開發。最後作者將這個模型作為壹個社交媒體應用程序，以顯示其可行性的壹個示例模型[13]。

Elena Zheleva，Lise Getoor 提出為了解決隱私問題，很多社會媒體網站允許用戶向社群公眾隱瞞自己的個人檔案。數據利用者如何利用與公共和私人用戶配置文件混合的在線社交網絡來預測用戶的私有屬性。依據關系分類的關系，通過公開的信息來推測用戶的敏感信息。除了朋友之間鏈接，群組之間的顯著信息載體。研究發現在幾個著名的社交媒體網站，使用基於鏈路和基於組的分類研究社交網絡的隱私問題混合的公共和私人用戶配置文件，可以準確地恢復私人配置文件的用戶的信息[14]。

Eden Litt 提出每天都有數以億計的人登錄到社交網站和產生 TB 級的數據，用戶分享狀態更新照片甚至更多。基於此社交網站如何運用技術來保護用戶的隱私。研究結果表明，社交網站上的技術為中介的溝通，即用戶與運營商之間的數據商。隱私權界限的制定以及技術的把控均由人來掌控。最終確定隱私不平等現象，其中某些群體更有可能采取的技術優勢，以保護自己的隱私，但是也存在壹些個人信息和聲譽可能面臨風險比其他人更多[15]。除此之外 Sweeny 等早在 2002 年提出的在發布之前對微觀數據進行 K-匿名（K-anonymity）處理是降低數據發布過程隱私泄露概率的有效方法[16]。

Priyadarshini Lamabam，Kunal Chakma 提出通過代碼混合和代碼切換，以及兩種或多種語言在交流過程中同時被使用。基於前人的研究，對於這樣的混合代碼進行語言識別時，由於其存在非正式文本（例如創意拼寫和拼音輸入），這樣的文本很難進行準確的辨認。作者運用自然語言處理方法，自動識別混合型的社交媒體文本。作者在研究跨語言的社交媒體中，選用 Twitter 和 facebook 帖子中展示英語和 Manipuri 之間的代碼混合。對比基於 Trigram 模型和基於條件隨機場（CRF）模型，那種模型能夠更準確的識別語言[17]。

Vidhyabhushan Dasondi, Milap Pathak 等人提出，運用數據挖掘中的文本分析方法進行社交媒體文本中的情感分析。將收集到的數據集分別進行詞語與句子兩類進行分析，之後計算兩類的權重，形成最終的文本模式。 運用 JAVA 技術對性能進行評價，文章中通過不同的性能參數及其描述其中所提出的系統通過準確性，出錯率，存儲空間，訓練時間，測試時間等幾個方面對形成的模型進行訓練。作者試圖通過以此模型通過以後的研究推廣自動識別模型[18]。

綜合國內外學者的研究現狀，根據已有的研究成果來看，國內雖然有壹些學者使用數據

挖掘進行社交媒體文本的分析，但是選用 Facebook 作為研究範例的較少，並且進行用戶信息曝露方面的研究較少。國外的研究現在有壹些學者運用 data mining,機器學習等方法構建模型進行數據集的研究，從用戶與社交媒體運營商獲得用戶信息數據不平等，跨語言文本分析,多種社交媒體文本分析,運營商獲用戶信息進行商品營銷等方面進行研究。根據學者現有的研究成果來看，多種方式的研究方向可知用戶隱私無懈可擊的暴露在公眾的視野中。在社交媒體用戶隱私方面還沒有較成熟的研究體系。並且現有的用戶信息曝露研究中實證研究較為鮮見。因此基於社交媒體用戶信息曝露研究這個問題對用戶使用社交媒體保證用戶隱私不被泄露具有非常重要的意義。

3.研究進行

本文選取國內外具有影響力的社交媒體 Facebook 作為切入點 ,根據 2016 年 7 月 26 日，Facebook 發佈的第二季度財報，Facbook 的月活躍用戶突破 17 億，日活躍用戶也突破了 11 億。龐大的用戶活躍量,大量的社交媒體文本，公開的 API 獲取數據方式。

本文選取社交媒體用戶註冊時的基本信息作業研究的數據集。將用戶曝露的基本信息曝光程度進行人為定義。

對 Facebook 用戶基本信息進行瀏覽時，用戶的基本信息如表 1 所示，危險等級依序 1> +0 > 0，是人為設定。

表 1 Facebook 上用戶記載的資料信息

Facebbook 用戶基本信息		危險等級
工作經歷		0+
職業技能		0
學歷		0+
生活過的地方		0+
聯繫方式	郵箱	1
	聯繫電話	1
基本信息	生日	1
	性別	1
	血型	0
	性取向	0
其他社交網站		0
家庭成員		0+
感情狀況		1
詳細資料	關於用戶信息	0
	喜愛的格言	0
大事記		0
用戶頭像		1

根據用戶曝露的基本信息，選取最基本的 10 個變量作為特徵值。分別是工作經歷 ,學歷，生活過的地方，聯繫方式中的郵箱和聯繫電話，生日，性別，家庭成員，感情狀況，用戶頭像等十個欄位作為研究變量。其中針對工作經歷和學歷以 0 為定義量度的起點。以用戶曝露的的信息量為準。試圖建立一種模型 ,反應用戶在信息曝露的過程中，歸納發佈數據可能遭受的攻擊類型，所可能給用戶帶來危害性影響的因素。通過實證研究,以期為用戶提供優化策略 ,減輕用戶遭受攻擊的可能性，為社交媒體用戶長久健康的發展提供基礎。

4.實驗結果輸出

將 100 筆用戶的基本信息數據讀入 R 軟體中，運用 K-means 進行分群。

將收集到的 100 筆用戶的數據，運用 k-means 分群的算法，將其分為五群。將這 100 位用戶的數據，分為五群，五群的數據量分別為 14；3；40；8；35。

通過對分群結果進行分析可以得出以下結果，如表 2 所示。

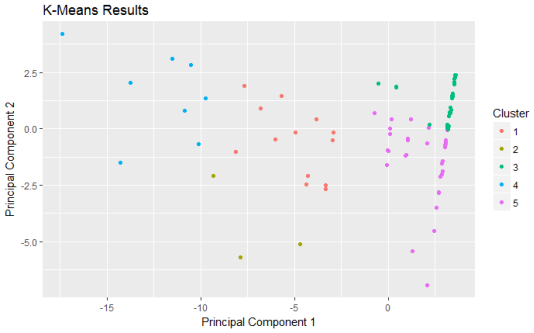
表 2 分群結果

分群特徵	頭像	工作	教育背景	居住地	Email	電話	生日	性別	家庭成員	情感
第一群	1	1--4	2--7	1--4	0	0	1	1	6--11	0
第二群	1	4--8	4--7	1--2	0	0	0	1	7--12	0
第三群	1	基本沒有	0--3	1--2	0	0--1	1	1	基本沒有	0
第四群	1	1--3	1--6	2	0	0	1	1	13--21	0--1
第五群	1	1--7	1--6	2	0	0	1	1	1--4	0--1

表 2 分群結果特徵統計

```
> set.seed(278613)
> AITK5<-kmeans(x=AIT,centers=5)
> AITK5
K-means clustering with 5 clusters of sizes 14, 3, 40, 8, 35

Cluster means:
  head.portrait     work background     live    e.mail     iphone  birthday
1     0.7857143 2.000000   3.500000 1.785714 0.1428571 0.07142857 0.7500000
2     1.0000000 6.333333   5.333333 1.000000 0.0000000 0.00000000 0.3333333
3     0.7375000 0.300000   1.000000 1.300000 0.0000000 0.40000000 0.7000000
4     1.0000000 1.375000   3.000000 1.750000 0.0000000 0.00000000 0.7500000
5     0.7571429 1.771429   3.342857 1.742857 0.0000000 0.05714286 0.4571429
     gender    member emotional
1 0.9285714  8.000000 0.2142857
2 1.0000000  9.666667 0.3333333
3 0.9000000  0.275000 0.1000000
4 0.7500000 15.500000 0.6250000
5 0.8000000  1.085714 0.5142857

Clustering vector:
  [1] 2 1 5 5 1 5 1 5 3 1 1 5 5 5 1 2 5 4 5 5 3 3 3 5 1 2 3 5 5 3 1 5 3 3 3 5 5 5
 [39] 3 5 5 3 3 3 5 5 3 5 3 3 4 1 3 3 5 3 3 3 3 1 1 5 3 3 4 3 5 3 3 3 3 3 3 1 5 3 1
 [77] 3 3 5 1 5 3 4 5 5 5 3 5 4 5 4 3 3 5 5 3 4 4 3 5

Within cluster sum of squares by cluster:
[1] 112.01786  29.33333 138.46875  94.25000 275.58571
 (between_SS / total_SS =  79.1 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

(a)R K-means 分群結果



(b)分群圖示

圖 1 K-means 分群結果與分群圖示

第一群的特徵主要表現在工作，教育背景，居住地，生日，家庭成員等信息曝露較多。

第二群的特徵主要表現在工作，教育背景，家庭成員等信息曝露較多。

第三群的特徵主要表現在工作和家庭成員信息基本沒有曝露，教育背景曝露較少，電話號碼少量曝露。

第四群的特徵主要表現在工作和教育背景曝露較少，家庭成員信息曝露較多。

第五群的特徵主要表現在工作，教育背景，家庭成員曝露信息較少。

將這五群進行用戶信息曝露危險等級定義。由危險到安全等級。第四群為第五級，第一群為第四級，第二群為第三級，第五群為第二級，第三群為第一級即為較安全的信息曝露方式。

5. 結論及未來的研究重點

本文根據100位用戶的數據進行分群分析，可以看出屬於第三群和第五群的用戶較多，根據我們定義的危險等級，也分別為第一級和第二級的用戶人群較多。但是也是有一部分用戶處於更高級別的危險等級，用戶在進行信息曝露的時候，應該注意第三、四、五群主要集中曝露的信息，例如工作，教育背景，生日，家庭成員等基本信息的洩露。本文僅僅選取 100 位用戶的基本信息進行實證研究，在數據集的研究上具有一定的局限性。同時利用非監督學習方式對於數據進行標記也會產生一定的人為誤差同時用戶基本信息數據集的選擇可能也具有一

些不可抵抗的因素。在接下來的研究中，著手進行用户基本信息曝露模型的構建的同時擴充訓練數據集的錄入。在今後的研究中，不僅僅局限於用户的基本信息曝露，還包括用户的評論，用户的地理位置信息等方面。以期通過這樣的方式能為用户的社交提供更加安全的保證。

6. 致謝

本文受大陸國家社科基金青年項目“基於信息價格動態揭示的社交媒體用户隱私保護研究”（項目批准號：15CTQ017）及台灣科技部計畫補助編號 MOST-104-2221-E-324-019-MY2, MOST-103-2632-E-324-001-MY3，特此申謝。

參考文獻

[1]***中國互聯網協會・中國網民權益保護調察報告*** [EB/OL]. http://www.isc.org.cn/zxzx/xhdt/listinfo-33759.html，2016.06.26/2016.10.08，2016

[2]徐勇、王浩、李東勤，“數據發布領域匿名隱私保護相關技術研究”，***情報雜志***，PP:30(8)：128-133,2011.

[3]王娜、許大辰，“移動社交網絡中個人信息保護現狀的調查與分析”，***情報雜志***，PP:34(1)：185-194,2015.

[4]王樹義、朱娜，“移動社交媒體用户隱私保護對策研究”，***情報理論與實踐***，PP:36(7)：36-40,2013.

[5]朱侯，“社交媒體用户隱私關注的心理機制研究”，***圖書情報知識***，PP:(2):75-82,2016.

[6]郭龍飛，“社交網絡用户隱私關注動態影像因素及行爲規律研究”，***北京郵電大學：經濟管理學院***，2013.

[7]苗振興，“基于社交媒體的人物分析技術研究”，***哈爾濱工業大學計算機科學與技術學院***，2013.

[8]陳希，“基于 R 語言數據挖掘的社交網絡客户細分研究”，***北京郵電大學：經濟管理學院***，2011.

[9]廉捷，“基于用户特征的社交網絡數據挖掘研究”，***北京交通大學***，2013.

[10]陳玉英，“基于機器學習的社交網絡用户特征分析”，***北京交通大學：電子信息工程學院***，2015.

[11]王麗文，“基于社交網絡的數據挖掘研究”，***電子科技大學***，2014.

[12]Cheng-Hung Tsai，Han-Wen Liu，et al, “Social persona preference analysis on social networks” ,***2015 International Conference on Connected Vehicles and Expo (ICCVE),*** taiwan，pp.32-39，2015.

[13]Mohamed Bourimi，Ricardo Tesoriero，et al，“Privacy and Security in Mult-modal User Interface Modeling for Social Media”，***2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)***，pp:1364-1367.2011.

[14]Elena Zheleva，Lise Getoor，“To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles ”***Proceedings of the 18th international conference on World wide web***, New York,PP:531-540.

[15]Eden Litt，“ Understanding social network site users’privacy tool use”***Computers in Human Behavior***，pp: 1649–1656，2013.

[16] Samarati P,Sweeney L. ，“Generalizing Data to Provide Anonymity When DisclosingInformation”，***Proc of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems.New York:ACM Press***,pp:188-189,1998.

[17]Priyadarshini Lamabam,Kunal Chakma，“A Language Identification System for Code-Mixed English-Manipuri Social Media Text”,***2016 IEEE International Conference on Engineering and Technology (ICETECH)***，Coimbatore，2016.

[18]Vidhyabhushan Dasondi, Milap Pathak, Narendra Pal Singh, ”An implementation of graph based text classification technique for social media,” ***2016 Symposium on Colossal Data Analysis and Networking***, indore, 2016.