Web Crawler Practice

Ethical Issues

Dr. Chun-Hsiang Chan

Department of Geography,
National Taiwan Normal University

Outline

- The Definition of Web Crawler
- Legal Issues
- Regulations
- Ethical Problems
- Lawsuit Examples
- Questions

The Definition of Web Crawler

- 網頁爬蟲係指利用自動化的機制將網頁上的資訊擷取下來；在此過程中，我們可能會收集到許多個人訊息，所以不論是靜態網頁爬蟲或是動態網頁爬蟲，都有可能會有一些潛在的法律問題。
- 因此，今天我們就來介紹一下究竟網頁爬蟲會有哪些法律問題需要注意：
  1) 爬蟲可以爬什麼資料？
  2) 資料要怎麼使用？
  3) 資料要如何保存？
  4) 公開＝＝授權

Legal Issues

- 先說在前頭，法律問題就要看你怎麼使用跟怎麼做了！

第三十六章　妨害電腦使用罪 (法源：中華民國刑法；修正日期：民國112年12月27日)	
第358條	無故輸入他人帳號密碼、破解使用電腦之保護措施或利用電腦系統之漏洞，而入侵他人之電腦或其相關設備者，處三年以下有期徒刑、拘役或科或併科三十萬元以下罰金。
第359條	無故取得、刪除或變更他人電腦或其相關設備之電磁紀錄，致生損害於公眾或他人者，處五年以下有期徒刑、拘役或科或併科六十萬元以下罰金。
第360條	無故以電腦程式或其他電磁方式干擾他人電腦或其相關設備，致生損害於公眾或他人者，處三年以下有期徒刑、拘役或科或併科三十萬元以下罰金。
第361條	對於公務機關之電腦或其相關設備犯前三條之罪者，加重其刑至二分之一。
第362條	製作專供犯本章之罪之電腦程式，而供自己或他人犯本章之罪，致生損害於公眾或他人者，處五年以下有期徒刑、拘役或科或併科六十萬元以下罰金。
第363條	第三百五十八條至第三百六十條之罪，須告訴乃論。

Legal Issues

第三章 非公務機關對個人資料之蒐集、處理及利用 (個人資料保護法；修訂日期：民國 112 年 05 月 31 日)	
第19條	1. 非公務機關對個人資料之蒐集或處理，除第六條第一項所規定資料外，應有特定目的，並符合下列情形之一者： 一、法律明文規定。 二、與當事人有契約或類似契約之關係，且已採取適當之安全措施。 三、當事人自行公開或其他已合法公開之個人資料。 四、學術研究機構基於公共利益為統計或學術研究而有必要，且資料經過提供者處理後或經蒐集者依其揭露方式無從識別特定之當事人。 五、經當事人同意。 六、為增進公共利益所必要。 七、個人資料取自於一般可得之來源。但當事人對該資料之禁止處理或利用，顯有更值得保護之重大利益者，不在此限。 八、對當事人權益無侵害。 2. 蒐集或處理者知悉或經當事人通知依前項第七款但書規定禁止對該資料之處理或利用時，應主動或依當事人之請求，刪除、停止處理或利用該個人資料。

Legal Issues

第三章 非公務機關對個人資料之蒐集、處理及利用 (個人資料保護法；修訂日期：民國 112 年 05 月 31 日)	
第20條	1. 非公務機關對個人資料之利用，除第六條第一項所規定資料外，應於蒐集之特定目的必要範圍內為之。但有下列情形之一者，得為特定目的外之利用： 一、法律明文規定。 二、為增進公共利益所必要。 三、為免除當事人之生命、身體、自由或財產上之危險。 四、為防止他人權益之重大危害。 五、公務機關或學術研究機構基於公共利益為統計或學術研究而有必要，且資料經過提供者處理後或經蒐集者依其揭露方式無從識別特定之當事人。 六、經當事人同意。 七、有利於當事人權益。 2. 非公務機關依前項規定利用個人資料行銷者，當事人表示拒絕接受行銷時，應即停止利用其個人資料行銷。 3. 非公務機關於首次行銷時，應提供當事人表示拒絕接受行銷之方式，並支付所需費用。

Legal Issues

第三章 非公務機關對個人資料之蒐集、處理及利用 (個人資料保護法；修訂日期：民國 112 年 05 月 31 日)	
第21條	非公務機關為國際傳輸個人資料，而有下列情形之一者，中央目的事業主管機關得限制之： 一、涉及國家重大利益。 二、國際條約或協定有特別規定。 三、接受國對於個人資料之保護未有完善之法規，致有損當事人權益之虞。 四、以迂迴方法向第三國（地區）傳輸個人資料規避本法。

Legal Issues

第三章 不公平競爭 (公平交易法；修訂日期：民國 106 年 06 月 14 日)	
第21條	事業不得在商品或廣告上，或以其他使公眾得知之方法，對於與商品相關而足以影響交易決定之事項，為虛偽不實或引人錯誤之表示或表徵。 前項所定與商品相關而足以影響交易決定之事項，包括商品之價格、數量、品質、內容、製造方法、製造日期、有效期限、使用方法、用途、原產地、製造者、製造地、加工者、加工地，及其他具有招徠效果之相關事項。 事業對於載有前項虛偽不實或引人錯誤表示之商品，不得販賣、運送、輸出或輸入。 前三項規定，於事業之服務準用之。 廣告代理業在明知或可得而知情形下，仍製作或設計有引人錯誤之廣告，與廣告主負連帶損害賠償責任。廣告媒體業在明知或可得而知其所傳播或刊載之廣告有引人錯誤之虞，仍予傳播或刊載，亦與廣告主負連帶損害賠償責任。廣告薦證者明知或可得而知其所從事之薦證有引人錯誤之虞，而仍為薦證者，與廣告主負連帶損害賠償責任。但廣告薦證者非屬知名公眾人物、專業人士或機構，僅於受廣告主報酬十倍之範圍內，與廣告主負連帶損害賠償責任。 前項所稱廣告薦證者，指廣告主以外，於廣告中反映其對商品或服務之意見、信賴、發現或親身體驗結果之人或機構。

Regulations

- 基本上，每個網頁或多或少都會有些相關規定，這些規定的強度與方式會影響你可以在這個網頁做的事情。綜合的來說，以下6個事情是組機本需要遵守的：
  1) 確認爬文的資料合法性(Robot.txt)
  2) 避免造成網頁伺服器大量的負荷
  3) 遵守網頁上的使用者條款
  4) 尊重知識財產權
  5) 收集到的資料保護與使用規範
  6) 公開資訊**不等於**授權

Regulations

1. **Actually, Facebook disallows any scraper, according to its robots.txt file**
   - When planning to scrape a website, you should always check its robots.txt first. Robots.txt is a file used by websites to let "bots" know if or how the site should be scrapped or crawled and indexed. You could access the file by adding "/robots.txt" at the end of the link to your target website.
   - Enter https://www.facebook.com/robots.txt in your browser, and let's check the robots file on Facebook. These two lines could be found at the bottom of the file:
     User-agent; *
     Disallow: /
   - The lines state that Facebook prohibits all automated scrapers. That is, no part of the website should be visited by an automated crawler.

5 Things You Need to Know Before Scraping Data From Facebook (url: https://medium.com/dataseries/5-things-you-need-to-know-before-scraping-data-from-facebook-f4e84b2ab80)

Regulations

- **Why do we need to respect robots.txt?**
- Websites use the robots file to specify a set of rules on how you or a bot should interact with them. When a website blocks all access to crawlers, the best thing to do is to leave that site alone. To follow the robots file is to avoid unethical data gathering as well as any legal ramifications.

5 Things You Need to Know Before Scraping Data From Facebook (url: https://medium.com/dataseries/5-things-you-need-to-know-before-scraping-data-from-facebook-f4e84b2ab80)

Regulations

Source: https://www.facebook.com/apps/site_scraping_tos_terms.php

Date of Last Revision: April 15th, 2010

Automated Data Collection Terms

1. These terms govern your collection of data from Facebook through automated means, such as through harvesting bots, robots, spiders, or scrapers ("Automated Data Collection"), as well as your use of that data.

2. You will not engage in Automated Data Collection without Facebook's express written permission.

3. By obtaining permission to engage in Automated Data Collection you agree to abide by these Automated Data Collection Terms, which incorporate by reference the Statement of Rights and Responsibilities.

4. You agree that your use of data you collect through Automated Data Collection will be confined solely to search indexing for display on the Internet unless granted separate approval by Facebook for alternative usage and display on the Internet.

5. You agree that you will not sell any data collected through, or derived from data collected through, Automated Data Collection.

6. You agree that you will not transfer data collected through Automated Data Collection in aggregated or bulk form.

7. You agree that you will destroy all data you have collected through Automated Data Collection upon Facebook's written request and that you will certify such destruction under penalty of perjury.

8. You agree that Facebook may revoke any permission granted at anytime for any reason and you agree to immediately cease collection and use of data collected through Automated Data Collection on notice of such revocation.

9. You agree to provide an accounting of all uses of data collected through Automated Data Collection within ten (10) days of your receipt of Facebook's request for such an accounting.

10. You agree that you will not circumvent any measures implemented by Facebook to prevent violations of these terms.

11. You agree that you will not violate the restrictions in any robot exclusion header.

12. You agree that you will only use your own true IP address/useragent identity and will not mask your services under the IP address/useragent string of another service.

13. You agree that you will not transfer any approved IP address or useragent to any party without Facebook's express written consent.

14. You agree that any violation of these terms may result in your immediate ban from all Facebook websites, products and services. You acknowledge and agree that a breach or threatened breach of these terms would cause irreparable injury, that money damages would be an inadequate remedy, and that Facebook shall be entitled to temporary and permanent injunctive relief, without the posting of any bond or other security, to restrain you or anyone acting on your behalf, from such breach or threatened breach. Nothing herein shall be construed as preventing Facebook from pursuing any and all remedies available to it, including the recovery of money damages.

15. Nothing herein shall be construed to confer any grant to, or license of, any intellectual property rights, whether by estoppel, by implication, or otherwise.

Regulations

Source: https://about.fb.com/news/2021/04/how-we-combat-scraping/

Meta | Shop | Our technologies | About us | Build with us

Back to Newsroom

Meta

How We Combat Scraping

April 15, 2021
By Mike Clark, Product Management Director

Last week, we shared more details about a public database containing information about people on Facebook that appeared online and generated a lot of conversation around data scraping. Given the fact that similar stories have emerged since then about public datasets involving information obtained from a number of other companies, including LinkedIn and Clubhouse, we'd like to explain more about what scraping is, how it works and what we're doing to prevent scraping to protect people's information.

What Is Scraping?

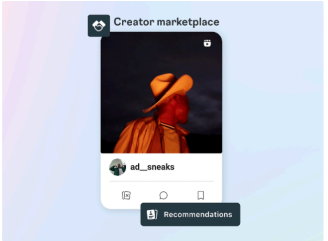
Scraping is the automated collection of data from a website or app and can be both authorized and unauthorized. Every time you use a search engine, for example, you are likely using data which was scraped in automated ways with the consent of the website or app. This is a form of scraping known as crawling and it's what helps make the internet searchable.

Using automation to get data from Facebook without our permission is a violation of our terms. The data itself is not necessarily off-limits; scraped

Topics

Company News
Technology and Innovation
Data and Privacy
Safety and Expression
Combating Misinformation
Economic Opportunity
Election Integrity
Strengthening Communities
Diversity and Inclusion

Featured News



Instagram

Making it Easier for Brands and Creators to Collaborate on Instagram
February 21, 2024

Regulations

- **Source:** https://www.facebook.com/robots.txt

```
# Notice: Collection of data on Facebook through automated means is
# prohibited unless you have express written permission from Facebook
# and may only be conducted for the limited purpose contained in said
# permission.
# See: http://www.facebook.com/apps/site_scraping_tos_terms.php
User-agent: Applebot
Disallow: /*/plugins/*
Disallow: /a/bz?
Disallow: /ajax/
Disallow: /album.php
Disallow: /checkpoint/
Disallow: /contact_importer/
Disallow: /dialog/
Disallow: /fbml/ajax/dialog/
Disallow: /feeds/
Disallow: /file_download.php
Disallow: /job_application/
Disallow: /l.php
Disallow: /login.php?next=
Disallow: /login/?next=
Disallow: /moments_app/
Disallow: /p.php
Disallow: /photos.php
Disallow: /plugins/
Disallow: /share.php
Disallow: /share/
Disallow: /sharer.php
Disallow: /sharer/
Disallow: /tr/
Disallow: /tr?
Disallow: /ufi/reaction/profile/browser/
Disallow: /x/oauth/
```

```
User-agent: baiduspider
Disallow: /*/plugins/*
Disallow: /a/bz?
Disallow: /ajax/
Disallow: /album.php
Disallow: /checkpoint/
Disallow: /contact_importer/
Disallow: /dialog/
Disallow: /fbml/ajax/dialog/
Disallow: /feeds/
Disallow: /file_download.php
Disallow: /job_application/
Disallow: /l.php
Disallow: /login.php?next=
Disallow: /login/?next=
Disallow: /moments_app/
Disallow: /p.php
Disallow: /photos.php
Disallow: /plugins/
Disallow: /share.php
Disallow: /share/
Disallow: /sharer.php
Disallow: /sharer/
Disallow: /tr/
Disallow: /tr?
Disallow: /ufi/reaction/profile/browser/
Disallow: /x/oauth/
```

```
User-agent: Bingbot
Disallow: /*/plugins/*
Disallow: /a/bz?
Disallow: /ajax/
Disallow: /album.php
Disallow: /checkpoint/
Disallow: /contact_importer/
Disallow: /dialog/
Disallow: /fbml/ajax/dialog/
Disallow: /feeds/
Disallow: /file_download.php
Disallow: /job_application/
Disallow: /l.php
Disallow: /login.php?next=
Disallow: /login/?next=
Disallow: /moments_app/
Disallow: /p.php
Disallow: /photos.php
Disallow: /plugins/
Disallow: /share.php
Disallow: /share/
Disallow: /sharer.php
Disallow: /sharer/
Disallow: /tr/
Disallow: /tr?
Disallow: /ufi/reaction/profile/browser/
Disallow: /x/oauth/
```

Ethical Problems

- **範例一：**
  最近某航空公司推出0元機票，大家瘋狂搶票；小明為了可以確保與女朋友出國，於是就撰寫一個網頁爬蟲程式，幫他訂票。試問：是否有違法或是有違反爬蟲倫理道德？

- **範例二：**
  最近某國知名歌手來台舉辦全球巡迴演唱會，小明撰寫一支程式專門作為程式搶票工具，再將搶到的門票轉賣出去(不論金額多少)。試問： 是否有違法或是有違反爬蟲倫理道德？

Ethical Problems

- **範例三：**
  市面上有很多廣告行銷公司與市場調查公司，會利用爬蟲的技術，大量收集每個網紅(KOL)、名人(Celebrity)或是政治人物的相關社群聲量或是計算其好感度等。這類的分析稱為社群聆聽(Social Listening or Social Buzz)。試問：這類的商業分析模組，會經由各種不同的服務架構販售給使用者，是否有潛在的法律爭議？

Lawsuit Examples – 榨取他人努力成果的顯失公平行為

- 曾有人開發及行銷關於戲院電影場次資訊的App「moovy」，卻未經同意而將其他業者經營之「開眼電影網」（@movies）網站上所編製的電影場次資料，混充為自身開發程式之資料內容，推展自己商品或服務，而遭公平會認定為足以影響交易秩序之顯失公平行為（參見公平會公處字第101094號處分書）。

Source: https://www.netadmin.com.tw/netadmin/zh-tw/viewpoint/BC05E753C27943AB93187936A8F25B7E

Lawsuit Examples – 591 vs. 豬豬快租App

- 近年來出現一款名為「豬豬快租」的手機App，與591房屋交易網及其App同屬提供消費者與出租物件刊登者可互通聯繫的平台，且在Google Play Store、Apple App Store均有上架。惟經591業者發現豬豬快租有涉嫌抄襲591的行為，乃委請公證人公證豬豬快租App 操作過程及其與591房屋App出租物件對照資料，發現豬豬快租App近9成內容係直接擷取591房屋網之出租物件資訊，且使用者輸入地區、用途、租金範圍、坪數範圍等搜尋條件後出現之頁面，並無揭露591所屬租屋物件原登載網址等情事，故591業者乃向公平會提出檢舉。
- 經公平會調查後認定：經營豬豬快租App的商家豬豬科技公司，未經同意而自行透過「網路爬蟲技術」擷取591租屋網站出租物件資訊，作為自身App之內容，並以此招攬使用者下載、付費購買加值服務、銷售廣告版位等商業交易行為，構成足以影響交易秩序之顯失公平行為，違反公平交易法第25條規定（參見公平會公處字第106084號處分書）。

Source: https://www.netadmin.com.tw/netadmin/zh-tw/viewpoint/BC05E753C27943AB93187936A8F25B7E

Lawsuit Examples – 美國近年有關網路爬蟲相關訴訟案例

- 2013年Hakikur Rahman與Isabel Ramos所著之「資料探勘在社會經濟發展應用與其倫理議題」一書[1]檢視網路爬蟲可能產生之相關法律問題，包括1. 隱私權保護、2. 網頁內容著作權歸屬與侵害與3. 網頁使用者協議等三大主要法律問題。
- 2013年與2016年JIM SNELL與DEREK CARE著有二篇短文[2]，整理近期美國有關網路爬蟲的相關法律訴訟案件，發現目標網頁控告網頁爬蟲的主要訴因（causes of actions）為：1. 著作權侵害；2.違反使用者協議（website terms）之效力；3. 違反美國電腦詐欺與濫用法案（Computer Fraud and Abuse Act，CFAA）；4.非法入侵動產（Trespass to chattels）；5. 重大新聞（misappropriation of hot news）之不當挪用。

Source: http://www.naipo.com/Portals/1/web_tw/Knowledge_Center/Infringement_Case/IPNC_180627_0501.htm?fbclid=IwAR1KID5rsiuR3gjUaiszycHqSW-g1tCnrI7xN1-NpiAn5hrYT2OeZGGUy0M

Lawsuit Examples – 著作權侵害

- 著作權法保護原創性之著作，而美國著作權法除了要求受保護之著作必須原創性之外，尚必須附著於一定媒介物。2007年美國聯邦第九巡迴上訴法院在Perfect 10, Inc. v. Amazon.com[3]乙案中，指出電腦記憶體、網路伺服器都是一種媒介，因此附著於記憶體或伺服器的軟體程式碼，如有原創性，亦可受著作權保護。
- 網頁上之具有原創性的著作內容物受到著作權保護，因此未經授權爬取網頁上之受保護之著作內容，將構成侵害網頁內容之著作權，並無疑問。然而，如果爬取網頁之內容，為不受著作權保護之事實，則其網頁爬取行為是否仍涉及著作權侵害，不無疑義？2009年在Facebook, Inc. v. Power Ventures, Inc[4]案，北加州聯邦地方法院認為，縱使所爬取的網頁之內容為不受著作權保護之事實，由於其爬蟲軟體必須暫時性複製顯示網頁事實內容之程式碼，因此也可能涉及著作權侵權。
- 儘管法院肯認從事網頁爬取行為涉及重製，但法院也認為後續的利用行為態樣，亦可能有成立合理使用之可能。例如在Kelly v. Arriba Soft Corp案[5]，被告承認搜尋引擎抓爬目標網頁的高解析度影像後而顯示低解析度之影像縮圖，構成對原告網頁資料重製，然而被告抗辯此一行為係屬於高解析度之影像轉換性之合理使用。聯邦第九巡迴上訴法院同意被告之抗辯，認為被告所顯示低解析度之縮圖，有助於一般大眾於網路上獲取資訊，因此被告之使用具有高度轉換性，與目的網頁以高解析度影像所要傳達著作之美感之目的並不相同。
- 據此，網路爬蟲的行為是否構成著作權侵權，仍然不能一概而論，必須依照具體的案件事實分析，重製的內容是否為著作權保護之標的或新的產出是否會影響著作物的市場價值、使用著作物標的之量與質等等問題，以決定是否有成立合理使用之空間。

Source: http://www.naipo.com/Portals/1/web_tw/Knowledge_Center/Infringement_Case/IPNC_180627_0501.htm?fbclid=IwAR1KID5rsiuR3gjUaiszycHqSW-g1tCnrI7xN1-NpiAn5hrYT2OeZGGUy0M

Lawsuit Examples – 違反使用者條款

- 多數商業網站均訂有使用者條款，以規範到訪和或使用網站之條件，用戶必須根據對這些條款之約定到訪或使用網站。儘管，網頁爬取行為展現科技的新用途，然而這種行為可能因為使用者違反使用者條款以抓取網頁資料，因而引發違反使用者協議之爭議。
- 多數情況下，目標網頁主張網路爬蟲違反使用者條款之舉證責任，往往較主張著作權侵權之舉證責任為高。後者，網頁抓爬之目標網頁僅須證明為網頁所有人與抓爬標的為受著作權保護之標的已足。證明違反使用者條款，網頁抓爬之目標網頁不但須證明使用者條款具拘束力且可執行、且必須證明抓爬的行為違反適用者條款、以及抓爬行為構成目標網頁之損害。
- 2007年在Southwest Airlines Co. v. BoardFirst, LLC案[6]，被告BroadFirst的軟體提供一項商業服務，以協助西南航空的客戶，利用西南航空公司的「開放」座位政策與辦理登機手續(check in)以獲得飛機優先座位之利益。在本案由於網頁的使用者條款用語明確限制網頁使用者作為非個人與商業用途之使用，因此，法院認為被告的行為屬於使用者條款所欲規範的範圍且與條款之內容直接相關，因此不同意被告主張該使用者條款欠缺明確而無執行力。本案審理之德州地方法院，因而判定被告BoardFirst使用西南公司網頁之行為，已違反了西南航空的網頁中使用者條款，因為條款禁止使用者利用網頁為個人與非商業目的用途(personal and non-commercial purpose)。

Source: http://www.naipo.com/Portals/1/web_tw/Knowledge_Center/Infringement_Case/IPNC_180627_0501.htm?fbclid=IwAR1KID5rsiuR3gjUaiszycHqSW-g1tCnrI7xN1-NpiAn5hrYT2OeZGGUy0M

Lawsuit Examples – 電腦詐欺與濫用

- 美國法院認為網絡爬取行為如果違反網站使用者條款，同時可能違反電腦詐欺濫用法案（Computer Fraud and Abuse 簡稱CFAA），該法案禁止「未經授權」或「逾越授權」進入電腦、網路、伺服器或資料庫。一般而言，只要電腦是公開可進入，並且不受密碼或其他保護安全措施，法院拒絕認定網路爬蟲任何造訪網頁行為違反CFAA。然而當網路爬蟲進入受保護之網頁，且網頁透過技術措施防止未經授權之進入網頁，或有明確停止未經授權之警告通知，則有可能構成違反CFAA法案。以下有2案涉及進入網頁抓取資料是否違反CFAA為審理。

Source: http://www.naipo.com/Portals/1/web_tw/Knowledge_Center/Infringement_Case/IPNC_180627_0501.htm?fbclid=IwAR1KID5rsiuR3gjUaiszycHqSW-g1tCnrI7xN1-NpiAn5hrYT2OeZGGUy0M

The End

Thank you for your attention!

Email: chchan@ntnu.edu.tw
Web: toodou.github.io

SCAN ME

SCAN ME