

**上海交通大學**

**信息安全工程學院**

**IS412 信息內容安全的理論與應用**

**組別論文（Group Paper）**

****

**Facebook爬蟲技術**

**Facebook Crawler**

**呂承祐 715030920002**

**顏澤龍 715030920005**

# 目錄

[**目錄**](#_5aa2xv9bx0ay) **1**

[**背景**](#_4p7tqcjvrlk5) **2**

[**社群網站特點**](#_h83gbglfsoo4) **3**

[多樣性](#_arcbytwxozqy) 3

[URL](#_skmmkqihug24) 3

[資訊重複性](#_tjhs4c5xowpe) 4

[隱私權意識](#_qeop5hj01net) 4

[資訊即時性](#_oitly2psn5tt) 5

[依照時間來決定是否更新檔案內容](#_ehpmduljuuc3) 5

[對比更新檔案內容](#_1gze188ewnrf) 5

[**Facebook Crawler**](#_xyqw0ocy2bo7) **6**

[示範](#_ehvznp90zj7b) 6

# 

# 

# 背景

在世界各地的網絡上，用戶們均可以在各個社群網站上發表個人意見、互相交流資訊、交換意見與觀點，從而發展出一個在網絡上虛擬的社交平台。而社群媒體與一般傳統媒體的不同之處，傳統媒體只能作出單一方向的發放資訊，而社群媒體則可以讓用戶在網絡上留言、讚好，從而大大提高媒體與大眾的互動性。

現今流行的社群媒體網站多不勝數，例如：Facebook（臉書）、Twitter（推特）、YouTube、新浪微博、微信朋友圈，已不再向舊式BBS電子佈告欄一樣，可以透過單獨使用簡單的GET或POST把網站的內容取出，它們往往有自己的API和權限，而本篇論文將介紹一些筆者認為實用的Facebook API參數，並且使用這些參數提出一個Facebook Crawler Model並進行實做，讓初學者能在一天內體驗爬取Facebook資訊的世界。

　　　　　　　　　　

　　　　　

# 社群網站特點

## 多樣性

現今社群網站為了給使用者獨特的使用者體驗，往往會提供自己獨特的API，而Facebook也提供了API給開發者使用，以便讓開發者在Facebook上更為方便地作出相應資訊挖掘。

Facebook Graph API是Facebook提供給開發者的簡單圖形測試工具，該測試工具會自行組合使用者選取的Facebook API參數成URL（超連結）並送出GET然後返回結果。因此使用者只需要了解Facebook API的運作和URL的組成，以及編寫一個可以送出GET的程式，就可以編寫出像Facebook Crawler一樣的工具，從而在Facebook上進行資訊挖掘工程。

### URL

|  |  |
| --- | --- |
| **https://grapth.facebook.com/v2.10/{page}/{access\_token=your\_token}&{fields=your\_fields}** | |
| **{page}** | 用戶ID或Facebook專頁ID |
| **{access\_token=your\_token}** | 權仗 |
| **{fields=your\_fields}** | API參數 |

## 

## 

## 資訊重複性

Facebook提供給使用者貼文的分享功能，為了不讓爬蟲爬取重複內容可以使用parent\_id來作篩選是否爬取該貼文的內容，該parent\_id會顯示該貼文的分享ID、Facebook的每則貼文和留言和使用者皆有自己ID有趣的是貼文和留言ID的前半段為發文使用者的ID，因此可以只需單靠ID便能知道貼文的來源。

Facebook提供使用者的留言功能，在一則貼文下可能有數萬筆的留言，但可能只有數百筆是重要的，因此我們需要有一個方法去篩選關注度高的留言，然當中每則留言中的like\_count會顯示該文的讚好次數，間接顯示用戶們對該留言的關注度。

## 隱私權意識

Facebook API使用了OAuth標準，因此我們需要開發者的token才能存取貼文資料。User Access Token由Facebook用戶授權，可獲取該名用戶貼文和貼文下的留言及個人資訊，但無法獲得設定為非公開的朋友資訊，在Facebook API版本2.10後需要透過Page Access Token才能以POST方法獲取資訊，值得一提的是留言需要用POST發送，因此沒Page Access Token是不能在貼文底下留言。

## 

## 

## 資訊即時性

由於Facebook用戶眾多，貼文留言數目往往以每秒幾十萬個貼文的速度增長，使其Facebook的更新速度極快。由於Facebook的更新速度快，在Facebook Crawler中亦選取了以兩種不同方法來與之比對。

### 依照時間來決定是否更新檔案內容

在此項方法中，我們在更新檔案內容中加入了貼文的更新時間。在每一次在Facebook中進行資訊挖掘時，檢查貼文的更新時間是否為最新，若為最新更新時間，則不會更新檔案內容；若更新時間並不是最新，即該貼文內容、留言及其讚好次數已被更改或有更新，檔案便會將內容作出即時更新，以確保檔案內容為最新。

### 對比更新檔案內容

在此項方法中，在每一次在Facebook中進行資訊挖掘時，如果該貼文是第一次出現，檔案便會即時儲存並不會與之作出比對；如果該貼文已有備份，舊有的更新檔案會儲存在log文件夾之內，而更新的貼文內容便會即時儲存新的檔案，儲存後便會即時作出比對，如內容有不同，系統便會告訴開發者「內容已更新」，這樣便可以讓開發者得知貼文內容、留言及其讚好次數有更改過或有更新。

# 

# 

# Facebook Crawler

然而由於互聯網上傳播資訊速度極快，為了以便作出監控，中國大陸政府於2005年起架設防火長城（Great Firewall of China）封鎖Facebook、Twitter、YouTube、Instagram等外國大型社群媒體網站，以便作出適當的監控。由於在中國大陸境內不能正常瀏覽Facebook，在技術方面我們透過VPN來連接境外伺服器，再透過Facebook Crawler來進行在Facebook上進行資訊挖掘，這樣便可以在中國大陸境內同樣可以在Facebook上進行資訊挖掘。

當同一個IP地址頻繁的送出請求或一次性獲取大量資訊時，這行動很容易被Facebook拒絕訪問及挖掘資訊，因此我們使用了since、until和limit來控制或取得資料的數量。在每次向Facebook請求並獲取資訊後，均會用sleep()來間隔10秒，以減少被拒絕訪問的機率，而上則貼文的更新時間會紀錄到檔案中，以便再次進行接續工作。我們同時利用next參數來獲取下一則貼文的URL，並利用updated\_time來決定是否更新檔案或資料庫。

## 示範

|  |  |
| --- | --- |
| **Brute Force with 55 Posts** | **Update Time with 55 Posts** |
|  |  |
| 運行時間 | |
| 0.05672s | 0.02694s |