Extraction of Data by Crawling Through Facebook Pages

Balvinder#1, Tripatdeep Singh Dua#2

#1Ph.d Scholar, IKG Punjab Technical University, Kapurthla, Punjab, India
#2Assistant Professor Computer Applications, Guru Nanak Institute of Management and Technology, Ludhiana, Punjab, India

btanejaus@gmail.com, tripatdua@yahoo.com

Abstract—Large amount of data available on the web has given opportunities for research. Facebook an online social networking site with over 2 billion people has come an exclusive tool for every updates over the social networking sites. The high variant data present in Facebook pages, groups has increased the prospectus of predictions about specific outcome for the business houses. Extracting such data and analysing is an issue in the present time. With the large amount of time spent by the users on social media, it has put pressure on the business organisations to implement social media across their business. The paper mainly concentrates on extracting data like reviews given the user on Facebook pages for a given post. It gives a comprehensive process of data extraction that helps the business organisation to mine data, which can help for growth.

Keywords—API, Facebook pages, Information extraction, Facebook graph API, Facebook Graph API explorer.

I. INTRODUCTION

Online social network websites [1] like Facebook, Twitter, LinkedIn etc. have large amount of data which is unstructured, but useful for the companies at the same time. Information extraction tools now extract data which is structured in nature. For the unstructured data like reviews, comments etc. research has been going on and few works has been published by the various authors.

Today online social networking sites Facebook has provided many facilities [2] to users, to make the website interactive and to engage the users new features are being added by the Facebook on daily basis. Provision of sharing the posts, videos, photos, chat with friends, playing games, made social media popular. At the same time promotion of business through Facebook pages, interaction of multiple users in Group community, paid promotions, events, jobs at the major benefits of Online social network sites.

With the invention of internet and proliferation of social media in recent times, the business organisation interaction with the customers has entirely changed. Business organisations are also spending on inbound marketing techniques along with the out bound marketing. It helps the users to know about new innovations, new products, new trends happening in the market via posts / infographics.

When we discuss about online social networking websites Facebook comes on the top along with other professional platforms [3] like LinkedIn, Twitter, Pinterest, Instagram. Facebook works on a model of friend of friend. Instagram and Twitter works on the model of follow and follower. LinkedIn allows users to add connections. Pinterest allow users to create & use boards. All online social network platforms are also used to build the business through the concept of pages, which help business organisations to reach to millions of people across the world.

A Online social networking website allows a user to create his/her profile with his/her name, where he/she can grow his/her friends/connections/followers depending upon the platform he/she is using. On Facebook first user has to create his/her profile and then user can communicate with his/he friends directly, they can form a group/community, they can build an application using Facebook APIs and do much more things.

Facebook [4] also allow the user to create a business page where user can invite people to like his/her business page. User can create posts about his business, people who have liked their page can view posts, and they can like the post or give their opinion on the posts. User can get professional insights from Facebook page about post reach, engagement and conversion [5] from the page.

1. Reach: Number of passive users who have just views the post.
2. Engagement: Number of active users who have liked, clicked or involved in any type of activity on the post.
3. Conversion: users who have performed the action desired by business organization.

Source: https://www.facebook.com/btanejaus/insights/?referrer=page_insights_tab_button.

Analysis of data [6] given by the users on Facebook page helps business organizations to know the opinion of people.
about how the product/service is. These online social networking sites are the major point where customers share their messages. It helps business organizations / consumers to raise their standards.

There are two types of data available on the online social networking websites

1. **Structured Data**: structured data is stored directly in the database tables, where output can be obtained and calculations can be performed directly. Information extracted is accomplished using the SQL queries e.g. user can calculate the no. of likes on a post, no. posts shares of posts on different pages etc. this types data comes under the structured data.

2. **Unstructured data**: unstructured data is in the form of free text written in the natural language. No structure can be found on these types of data. User can apply specialised information extraction techniques to extract the data from these pages. Reviews, comments etc. are the example of unstructured data which is the current topic of research in these days.

Diagram showing the Facebook data process:

![Diagram showing the Facebook data process](https://developers.facebook.com/tools/explorer)

**II. METHOD OF EXTRACTING DATA**

Facebook provide graph APIs and graph API explorer [7] to extract [8] the data from own Facebook pages. User can use Facebook URL https://developers.facebook.com/tools/explorer to extract the data. Graph API explorer framework is as shown in the figure

![Fig. 2: Facebook data generation and analysis process](https://developers.facebook.com/tools/explorer)

**Fig 2: Facebook data generation and analysis process.**

**Step 1: Login-Authentication of User**

This is the first step and user is asked to login [9] to his account after passing the authentication with user-id and password provided by the account. Only authenticated user has the rights to access the data.

**Step 2: Access Token**

Facebook allow user to access Graph APIs by providing an access token. User can get access token by three ways:

1. **User Access Token**
2. **App Access Token**
3. **Page Access Token**
Access token is provided to the user after passing the authentication process. User access token are provided with the data access permission like user data permissions, events, groups, pages, Instagram, videos, insights. According to your access token provided user can access the data on Facebook objects.

OAuth 2.0 Tool

OAuth 2.0 [10] is a authorization protocol layer, where an end user, the resource owner can grant an access to his protected data stored at the resource server without sharing his username and password with the third party services. This can be achieved by authenticating user credentials directly with the server trusted by authorization server, which issues the access to data by providing the access token.

OAuth 2.0 has four roles:

1. Resource owner: It refers to the end user entity who can grant endpoints to a protected data/resource.
2. Resource Server: the server which is hosting the end user data/resources and the server will issue the access token to third party to access resources owner data/resources.
3. Client: this can be an application or program which sends requests on the behalf of the resource owner for accessing protected data with its authorization.
4. Authorization server: This is a server issues access token to the client after authenticating the resource owner.

Step 3: Extraction of Data using graph API explorer

After passing the authentication process and getting the access token we can extract data using graph API explorer. API is primary way to get data in and out from the online social networking sites.

Facebook Graph API provides http/1.1 for transferring the data and all endpoints require https. To extract the data from Facebook, reading operations begin with node with its unique ID. E.g. A page node with unique ID can be used to receive the data from Facebook Page.

We can use command:

Get https://graph.facebook.com/<pageID> to retrieve the information about the page.

Nodes in the Facebook have edges, which help in return collection of other nodes connected with them. To extract the data using edge, user can use node ID and edge name in the path URL. E.g. to read the post nodes on a page /feed edge is used.

Get http://graph.facebook.com/<pageID>/feed

It will retrieve the data in JSON format while extracting the data from Facebook with the nodes and edges you can specify the filters on data. To achieve this fields are used. Fields are node properties. With fields you can specify which field you want to retrieve by using field parameter and listing each filed.

e.g. to retrieve about me information one can use:

me/?fields=about

So with the help of the nodes, edges, and fields user can extract the data from Facebook pages in JSON format.

To retrieve the comments from the post on a page we can use:

Get

https://graph.facebook.com/<pageID_postID>/Comments

This command will extract the comments on a particular post given by users as a feedback or review for the post.

Step 4: Conversion of Data

Facebook Graph explorer API return the data in JSON which make it difficult for end user to read the data in proper format and use it for further processing. So, we need to convert the data into XML. XML is Extensible Markup Language and format provided by XML is easy to use, all the web application development languages accept data in XML format which can be formatted with XSL for better presentation.

To convert the JSON data into XML format python is used. Python is high level object oriented programming language [11]. Python is used for the development of data science program development. Python is free and open source language with large standard library for regular expressions, web browsers, unit testing, email, image manipulation, threading, databases and a lot of other functionality.

Python [12] provides modules and packages which is a collection of python programs used to solve any problem by programming. In this paper we have used xmltodict package to convert JSON code into xml code.

Full python code is given below:

### TABLE 1

<table>
<thead>
<tr>
<th>Python Programming Code to Convert JSON file to XML format.</th>
</tr>
</thead>
<tbody>
<tr>
<td>import json</td>
</tr>
<tr>
<td>import xmltodict</td>
</tr>
<tr>
<td>with open('sample.json', 'r') as f:</td>
</tr>
<tr>
<td>jsonString = f.read()</td>
</tr>
<tr>
<td>print('JSON input (input.json):')</td>
</tr>
<tr>
<td>print(jsonString)</td>
</tr>
<tr>
<td>xmlString = xmltodict.unparse(json.loads(jsonString), pretty=True)</td>
</tr>
<tr>
<td>print('XML output(output.xml):')</td>
</tr>
<tr>
<td>print(xmlString)</td>
</tr>
<tr>
<td>with open('output.xml', 'w') as f:</td>
</tr>
<tr>
<td>f.write(xmlString)</td>
</tr>
</tbody>
</table>

This code accepts the input file in JSON format and generates the output file in XML format.

Step 5: Analysis of Data

From the Facebook page different types of data can be extracted with nodes, edges and fields like comments, shares, no of likes, etc. Analysis of data can be performed according
to the need. Analysis is always based on the application of the end user. User can generate different types of data patterns from the extracted data which may be utilized for decision making in any organization.

III. EXPERIMENTATION RESULTS

Step 1: Login-Authentication of User
The first step to use Facebook graph API explorer is to login into the Facebook system with your personal account.

Fig. 5: To show Login screen of Facebook.

Step 2: Access Token
After passing the login process next step is to get the access token, get access token on Balvinder Taneja page.

Fig. 6: To get the access token on a Facebook page

Step 3: Extraction of Data using graph API explorer
After getting the access token on page, following command was executed in the Facebook Graph API explorer:

```
1444109289155226_2078984755667673/comments
```
Where 1444109289155226 is page id and 2078984755667673 is post id for a particular post which make it complete node and /comments edge to retrieve the comments from a post.

Fig 7 Facebook page comments data retrieval.

Step 4: Conversion of Data

```
{"date": []
  
  "created_time": "2018-12-10T14:30:32+0000",
  "name": "Balvinder Taneja",
  "id": "1444109289155226",
  "message": "Learn Full Course",
  "is": "2078984900000091_227312609214593"
}
```
```
{"date": []
  
  "created_time": "2018-02-14T16:48:48+0000",
  "name": "Yo Balvinder I want to learn",
  "is": "2078984900000091_208695937153878"
}
```

Fig. 8: Input.json file retrieved from Facebook page.

After executing the jsontoxml.py program in python out is sent to Output.xml file. Result is shown below:

```
<?xml version="1.0" encoding="UTF-8"?>
<

default
  
  "created_time": "2018-12-10T14:30:32+0000",
  "name": "Balvinder Taneja",
  "id": "1444109289155226",
  "message": "Learn Full Course",
  "is": "2078984900000091_227312609214593"
}
```
```
<default
  
  "created_time": "2018-02-14T16:48:48+0000",
  "name": "Yo Balvinder I want to learn",
  "is": "2078984900000091_208695937153878"
>
```

Fig. 9: Output.xml file after executing the python code.

IV. CONCLUSION

Online Social media websites are very much trending in these days for sharing posts, videos, comments, reviews, audios, photos. It helps to get access to list of friends and their activities through their posts. The process of extraction and analysis of data has become challenging as attribute of online social networking sites is very dynamic and data is unstructured.

This paper is mainly to help the researcher or students in understanding the steps involved in extracting data from Facebook pages and come up with new ideas. Data Mining can be applied to generate different patterns which can help organizations in decision making. This is our basic study on how to extract unstructured data from the review or comments.
Further study will help to mine the different data patterns from Facebook pages integrated with ecommerce websites.

REFERENCES


