International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 2, Issue 2, February 2015

A Study on Content Extraction from Social Networking Site

-Twitter/Facebook as a case study

[1]Narashima S. Purohit, [2]Meghana Bhat, [3]Akshata B. Angadi,[4]Karuna C. Gull
[1] [2] [3] [4] K.L.E. Institute of Technology, Hubli
[1]narashima.purohit@gmail.com, [2] mmeghanabhat@gmail.com, [3]akshata_angadi@yahoo.co.in, [4]karuna7674@gmail.com

Abstract: ***Internet is rich in useful information. Social media has a rising impact on the way the world connects, builds and communicates. Social networks are basically about the people in them and their activities include shares, comments and posts. Facebook, Twitter, LinkedIn are the popular Social networking services where users have personal profiles, add other users as friends. In this paper, we have considered popular social networking sites (Facebook, Twitter) as a case study. The paper helps students to learn how the dynamic and unstructured data can be extracted from the live sites and the concept behind the scene of extraction process is detailed, this helps them to work further to develop an algorithm that suits better to raise the marketing strategies. We have proposed the Application framework in general. The paper says how the open source API will help and shows what all facilities will be provided by it so that they can take a better use of it to build an application. Thus the main moto of writing this paper is to enlighten the students to start with the development of new applications using the data from SNS.***

Index Terms— ***API (Application Programming Interface), JSON (Java Script Object Notation), OAuth, SNS (Social Networking Sites).***

I. INTRODUCTION

Social media are the Internet sites where people interact freely, share and discuss information about each other and their lives. Social Media doesn't mean scrap booking or blogging alone; it includes social networks, wikis, news forums etc. As stated by Boyd and Ellison [1] "SNS can be defined as web-based services that allow users to create profiles and articulate networks that they can share with others within the system".

When a term Social Networking Site is heard we directly think of Facebook, Twitter and LinkedIn. These three are the popular and well known sites or services. Facebook is generally considered the most casual; Twitter and LinkedIn are typically used for professional purposes. LinkedIn allows you to add Connections, Twitter creates Followers and Facebook has Friends [2]. The sites are also used to build the business as social networking sites helps to reach millions of people worldwide.

A social networking web site allows a user to create his/her profile that shows his identity, Increase the contacts or connections by adding friends to his account, Communicate and engage with these users/friends, Form community/group, Build an Application using API, Create Social Graph that shows the influence of users.

In sites like Facebook and twitter each individual have a provision to create their respective profile. [3] SNS provides facilities like:

- Creating a profile will provide personal statistics that include his location, contact information, associations, work history, relation among different friends, friends list.
- Can make friends from all over the world. Thus increase in contacts list.
- Can communicate among friends through messages. Messages may be in the form of text, audio, video or images.
- Can Like/ share/comment on the particular posts sent by friends.

The data exchanged among people is huge when this data is tracked and analyzed systematically people are benefitted.

As shown in Fig.1 the data collected can be tracked/ analyzed and build an application using it. Social media or social networking sites such as Facebook, LinkedIn and Twitter are the platforms where user can publicly post content.

Analysis of exchanged messages, posts or likes helps to know the opinion of people of how the product is. These

sites act as club houses where in the communication messages receive the most attention from customers. This in turn helps the marketers/ consumers to raise their levels. In the future, information extraction from cross-website pages will become more important as we move toward semantic Web.

LITERATURE SURVEY

Now a day's huge amount of information is being transferred through social media, including blogs, Web fora and micro-blogging services like Facebook or Twitter. Social Networking sites allow us to build an application.

The API is developed and is made open source these days so that the users can build new application that add features and give better experience to the user's flexibility. In this paper, the steps involved in the extraction of content of these sites are detailed in brief.

Bernhard Rieder[5], Paper provides an overview over analytical directions opened up by the data made available, discusses platform specific aspects of data extraction via the official Application Programming Interface, and briefly engages the difficult ethical considerations attached to this type of research. Author has described the Netvizz application, a general purpose data-extractor for different subsections of the Facebook platform.

Sean Whitsitt et. al [6], paper describes a transformation based approach to the extraction of a social network graph. Author described methods of extracting a social network from data structures not originally intended to be analyzed using metrics for the behavior of an organization. Novel approaches to graph analysis algorithms are used in order to permit scalability of common metrics, and these approaches leverage the concept of the social network to reduce the size of the graph under consideration.

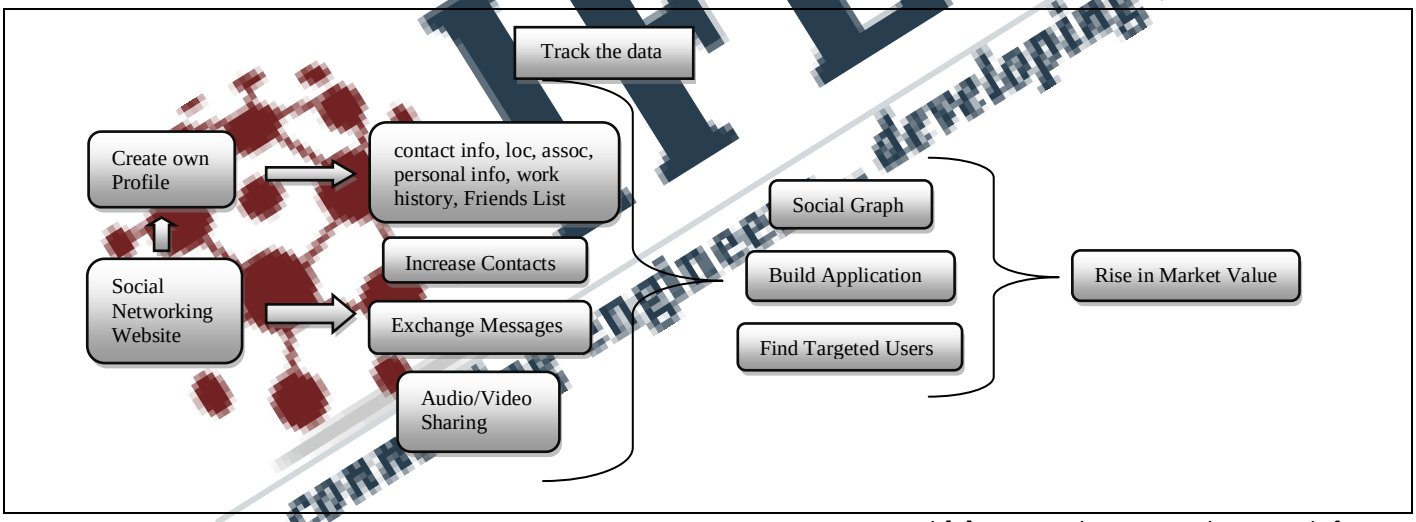


Figure.1. Use of Social Networking Websites

The data in the Social media is of three types: Structured, Unstructured & Semi structured data. Structured data is the data that can be easily organized. When the data is placed in relational tables, data can be easily understood, can know where it is and know how it relates to other data present there. [4]Semantically tagged documents (structured) are easy to analyze when compared with the unstructured data. Semi structured data is typically treated as unstructured data for the purpose of machine processing and analysis. If the object to be stored carries no tags (metadata about the data) and has no established schema, ontology, glossary, or consistent organization it is unstructured. Unstructured data includes: feedbacks, email, news and blog articles, tweets, web pages, other social media as well as audio and video files. 80-85% of enterprise content is in unstructured format whereas 10-15% is in structured. Analysis of such data plays an important role.

. al [7] proposed automated approach for personal data extraction was developed to extract personal details and top friends from MySpace profiles and place them into a repository. An online social network graph was generated from the repository data where nodes represent peoples' profiles. Research has provided opportunities for future research to be carried out: Development of an agent to automate the process of data retrieval. The agent will be able to track the behavior of the profile and report any changes and extracting the data from the online social networking profiles using a Depth First Search (crawling algorithms).

Web Data Extraction architecture based on Intelligent Agents", given by [8], is shown in Fig 2. An architecture consists of three components: (i)a server running the mining agent(s); (ii)a cross-platform Java application, which

implements the logic of the agent; (iii)an Apache interface, which manages the information transfer through the Web. Proposed architecture is able to implement several crawling strategies like BFS or rejection sampling.

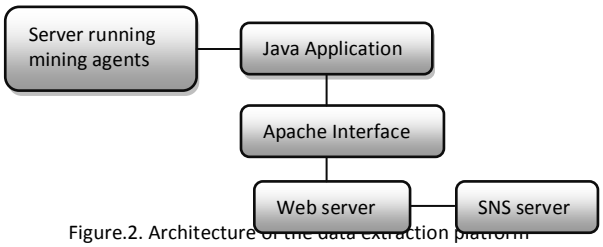


Figure.2. Architecture of the data extraction platform

The sampling procedure in [8] works as follows: an agent is activated and it queries the Facebook server(s) to obtain the list of Web pages representing the list of friends of a Facebook user. Of course, the Facebook account to visit depends on the basis of the crawling algorithm. After parsing list of pages, it is possible to reconstruct a portion of the Facebook network. Collected data is converted into HTML/XML format in such a way as to they can be exploited by other applications.

Earlier Google spreadsheets were used to extract the the tweets. Now the Twitter has introduced twitter API's that helps in extracting the required text from account using OAuth. Taking this as a basis we have designed our methodology to extract the data from SNS.

METHOD OF EXTRACTING DATA

The Framework of Application flow is shown in Fig.3 The process has four main components:

1) Authentication process
2) Extraction of the Data
3) Conversion of Data
4) Analysis of the Data

The process includes following steps:

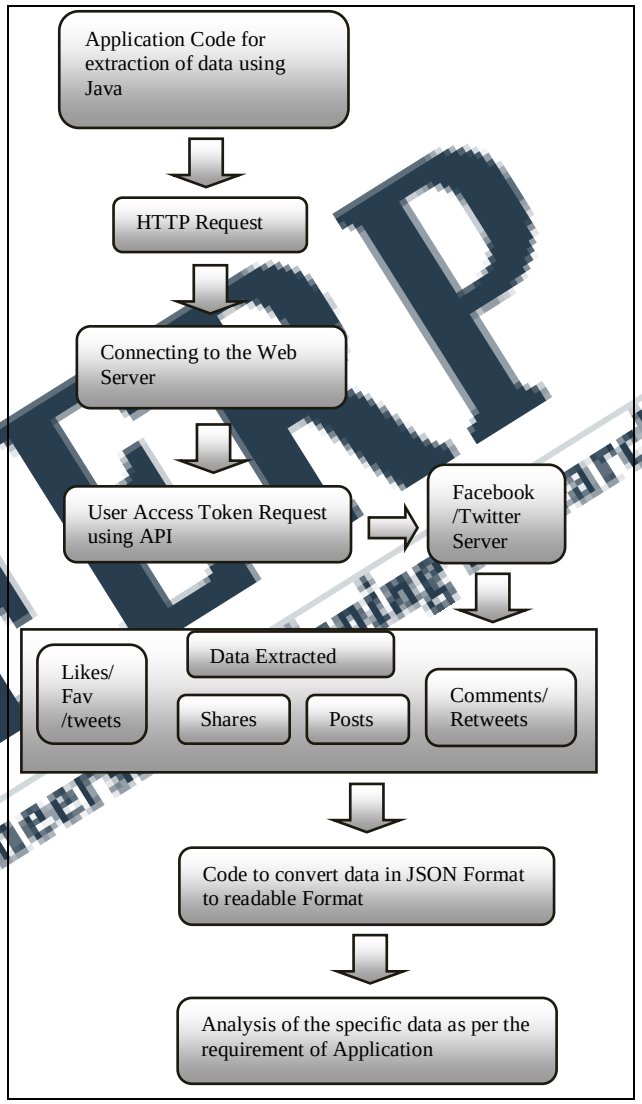


Figure.3. Framework of an Application Flow

Step 1: **Authentication**

Authenticated requests are must to access the API's of SNS's. Each request must be signed with valid user credentials.

Figure.4. Authentication Framework

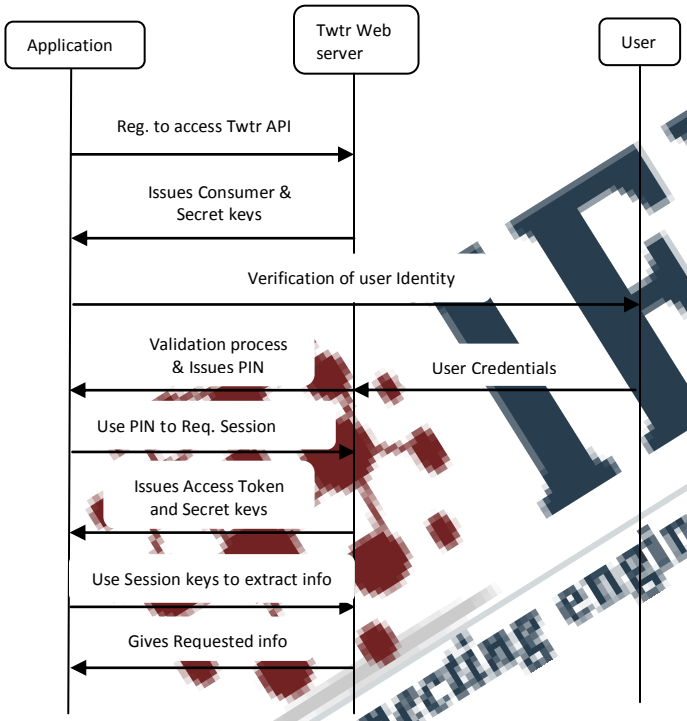
A. OAuth

Organization/Business is using OAuth (Open Authentication) to protect the APIs they offer their partners and customers. OAuth is only component of a full API access control and security solution. It is an open standard for authentication, adopted by Twitter/Facebook to provide access to protected information and the process is carried out using a three-way handshake.

OAuth provides a method of third party authentication that allows Web services to share data through their APIs. A

user establishes an account with one service, and a server from that service can provide other services with tokens that can be used to access that data. So a user with a Twitter/Facebook account, for instance, can approve a third-party application to access some of his or her Facebook/Twitter information, without actually providing the service with a log-in name and password [9].

Fig.5 summarizes the steps involved in using OAuth for authentication process of user and application [10].




```
{
“id”: 156782290,
“from_user_id”: 154677,
“created_at”: “Tue,09 Sept 2014 13:21:22 +0000”
……….
},
{
 “metadata” : [
{
“result_type”:”popular”,
“recent_retweets”: “134”
 }
“source”: “<a href=http://twitter.com/twitter</a>,
“iso_language_code”: “nl”
…….
]
“since_id”: 0,
“max_id”: 132870768,
“page”: 1
}
```

Figure.5. Process of Registration/Authentication of User and Application

Step 2: **Extraction of data from SNS**

Once authentication process is completed, we can extract the data depending on our requirements of an application.

Social networking site i.e. Facebook / Twitter is a structured model. Based on the kind of data present in social media, extraction method is applied. The data present in the webpage of Facebook / Twitter come under structured whereas related data of the same, given by user come under unstructured one. We write the query to get the contents of web page of Facebook / Twitter which is in the JSON format.

Step 3: **Conversion technique**

As said above, the APIs return data as JSON which makes it difficult for clients/developers to interact with/read that data.

We need to transform that data, from JSON to XML as reading data from XML is easier compared to JSON. And there are a rich set of APIs and Tools available to do these transformations. Thus REST and Java client APIs provide full support for loading and querying JSON documents, where the JSON documents are stored and retrieved as XML (Indirectly data is retrieved from JSON format). This allows for fine-grained access to the JSON documents [Source: RestApiTutorial]. There are seven different patterns in XML. The different samples of fields in both forms i.e. XML and JSON are given in Table.1 [Source: XML.com].

Table. 1. Seven Different formats of XML

Pat.	XML	JSON
1	<e/>	"e": null
2	<e>text</e>	"e": "text"
3	<e name="value" />	"e":{"@name": "value"}
4	<e name="value">text</e>	"e": { "@name": "value", "#text": "text" }
5	<e><a>text text </e>	"e": { "a": "text", "b": "text" }
6	<e><a>text <a>text </e>	"e": { "a": ["text", "text"] }
7	<e> text <a>text </e>	"e": { "#text": "text", "a": "text" }

JSON data and values will be in form of name/value pairs, where values may be a string, number, Boolean, object or an array. JSON Objects are written within the curly braces and Arrays are written inside the square braces. The objects returned from most Server APIs are highly nested. The sample data (name/value pairs) are shown in Table.2 taken from Twitter developer’s site.

Table.2. Sample JSON code of Twitter

Write the code / query with field name to extract required data from XML and dump them into either database or file for further processing.

Table .3 Fields of FriendList

Name	Description	Permissions	Returns

Id	The friend list ID	read_friendlists	String
Name	The name of the friend list	read_friendlists	String
list_type	The type of the friends list; Possible values are: close_friends, acquaintances, restricted,user_created, education, work, current_city or family	read_friendlists	String

Table.3 shows the fields that will be available by extracting the information about friends. Similarly different data like posts, tweets, comments, retweets etc can be extracted from Facebook/Twitter using API.

Step 4: **Analysis of data**

Once the data is extracted the process of analysis can be done depending on the need.

FEW CODE SNIPPETS

Following are some of the code snippets written for twitter processing.

1) For connecting to twitter

```
private final static String CONSUMER_KEY =
"nRKrO4pHWwAKwtDlxIjusA";
private final static String CONSUMER_KEY_SECRET =
"mfxnnnyu0tA92NJive4OmLwVb4euZfzrHuwP9j8RD8";
    Twitter twitter=null;
     AccessToken accessToken = null;
     RequestToken requestToken=null;
Private void jButton4ActionPerformed (java.awt.event.ActionEvent
evt) {
 twitter= new TwitterFactory().getInstance();
 twitter.setOAuthConsumer(CONSUMER_KEY,
CONSUMER_KEY_SECRET);
 try  {
 requestToken= twitter.getOAuthRequestToken();
jLabel4.setText("COPY IN BROWSER AND GET KEY:
"+requestToken.getAuthorizationURL());
jTextField6.setText(requestToken.getAuthorizationURL());
 }
```

2) For fetching the tweets of Hashtags

```
Public String[] fetchTweets(String hashTag,
int number_of_messages)throwsTwitterException{
Twitter twitter =
newAuthenticateCredentials().getTwitterInstance();
                Query query = newQuery(hashTag);
                query.setCount(number_of_messages);
                QueryResult result;
                result = twitter.search(query);
                List<Status> tweets = result.getTweets();
                int i = 0;
                tweetsOfHashtag = new String[tweets.size()];
                for (Status tweet : tweets) {
                tweetsOfHashtag[i] = "@" +
tweet.getUser().getScreenName() + " -- " + tweet.getText();
                i++;
                      }
                returntweetsOfHashtag;
}
```

EXPERIMENTAL RESULTS

The Web Interface of our Application should consist of:

Module	Input/Trigger	Expected Output
Login	Username And Password	Successful/Unsuccessful login, Redirect to Main/Login Page
Logout	N/A	Redirect to Login page

And at the Server side the three modules are necessary:

Module	Input/Trigger	Expected Output
Authentication	Username email-id And password	Successful login/ error display
Provide access token	Authenticated user	Access token for Each User
Storing into database	User information from client side	Inserting data into SQL database/File

A. *Login Page Twitter as a case*

This step involves creation of web application on the client side using HTML, Java and JavaScript. Login page is created through which user logs into an application developed by user. When a user logs into client side web application, it will be automatically directed to Twitter log in.

This module gets the access token for each user and also helps in loading various modules of Twitter through Twitter SDK for each user asynchronously. The SDK consists of API's that provides information about the user activities which is publically available. Each time the user logs into application, a new access token is generated by Twitter for that particular user.

B. Authentication process

Initiate by registering our application to Twitter service using consumer key and secret key (written inside code). Twitter uses Rest or Stream API to access the user token as per user convenience. This states that Every SNS or a Service has its own API and a standard authentication process.

OAuth Token is requested initially to get authorization URL. Copy and paste the received URL on the Address bar of the browser. The authorize application page will be opened. Fig. 6a, b, c shows the links. Click on Authorize App to get Pin(OAuth Verifier).

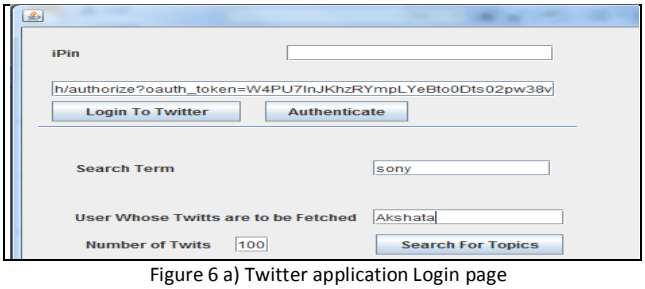
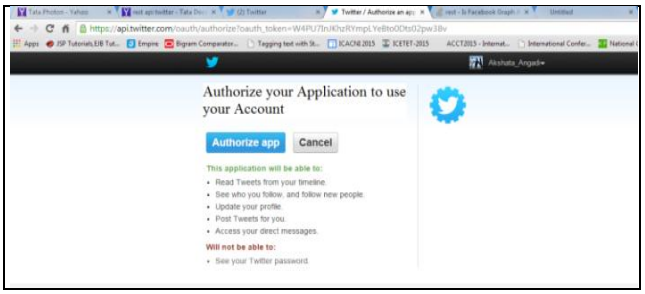
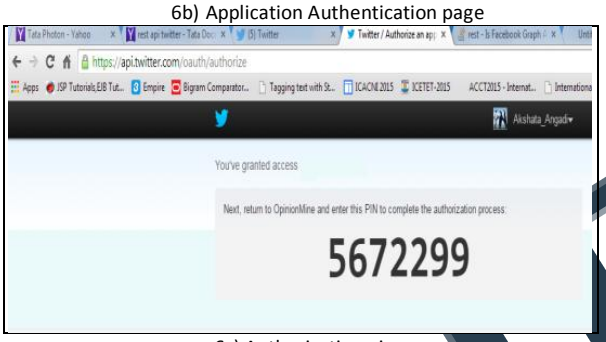


Figure 6 a) Twitter application Login page



6b) Application Authentication page



6c) Authorization pin page

Copy the Pin received and paste it in iPin textbox to authenticate our application to access session keys. Using secret key we can extract data from twitter server depending on topic /choice given.

The Fig.7.a) and b) shows the sample twitter page and the extracted tweets as per the Twitter account web page. If the application requires No.of followers, friends list, No of Likes, favourites of a use we can even extract it by changing the code and query.

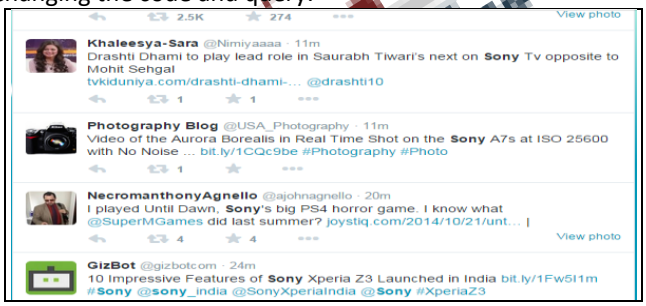
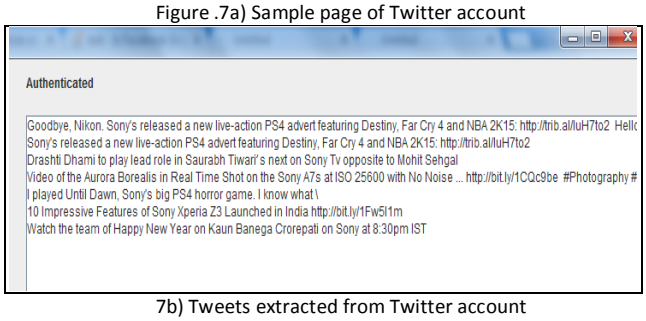


Figure .7a) Sample page of Twitter account



7b) Tweets extracted from Twitter account

CONCLUSION AND FUTURE WORK

SNS helps to get an access to friends and exchange messages. Messages may be in the form of content or content in posts, shares, comments. The Analysis of the content is the challenging attribute in Social Media as the data is dynamic and the format of data will be different in different sites.

The paper helps students to know the steps of extracting content and help them to come up with new innovations in developing better applications by using the dynamic data. The Data Mining Techniques and many better algorithms can be used to develop applications and raise the strategies of market level by developing new applications. Further the strong connections in networks can be known by analyzing the activities of users. This is our initial study which gives basics about how to start with the application development. Further study will be added with improved algorithm and process of application development. I hope this study will help the students to think on the concept and come up with new applications.

REFERENCES

[1] Danah M. Boyd and Nicole B. Ellison, "Social Network Sites: Definition, History, and Scholarship", Journal of Computer-Mediated Communication, Vol 13, Issue 1, pp 210–230, doi:10.1111/j.1083-6101.2007.00393.x, Oct 2007.

Brad Dinerman, "Social networking and security risks" , in GFI White Paper ,2011, pp.1-8.

K. L. James, "The Internet: A User's Guide", Google E-book, April 10, 2010.

www.wikipedia.com

[2] Bernhard Rieder, "Studying Facebook via Data Extraction: The Netvizz Application", WebSci'13, ACM, May 2–4, 2013, Paris, France.

[3] Sean Whitsitt, Abishek Gopalan, Sangman Cho, Jonathan Sprinkle, Srinivasan Ramasubramanian, Liana Suantak, Jerzy Rozenblit, "On the Extraction and Analysis of a Social Network with Partial Organizational Observation" ATRAP work, USA.

[4] Sophia Alim, Ruqayya Abdulrahman, Daniel Neagu and Mick Ridley, "Online social network profile data extraction for vulnerability analysis", Int. J. Internet Technology and Secured Transactions, Vol. 3, No. 2, 2011.

[5] Catanese, S., De Meo, P., Ferrara, E., Fiumara, G., Provetti, A.: Crawling facebook for social network analysis purposes. In: Proc. of the International Conference on Web Intelligence, Mining and Semantics, pp. 52:1-52:8. ACM, 2011.

Joab Jackson, "OAuth 2.0 security used by Facebook, others called weak", IDG News Service, Sept 22, 2010.

Shamanth Kumar, Fred Morstatter, Huan Liu, "Twitter Data Analytics", Springer, Aug 19, 2013.