International Journal of Management, IT & Engineering
Vol.9 Issue 4, April2019,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Profile Mining: Analysing Tools & Techniques of Data Extraction from Social Media Platforms

Priti Bali*
Dr. Anoop Sharma**

Abstract

Advancement in technologies has turned the way of life. Huge amount of information is available on web. Information is available but it is in scattered form and different formats such as text, images, videos and audios. Also, the desired information is available in parts on different web pages not on a single page as different applications/social media platforms display different characteristics of a person.Extraction and presentation of this information in a proper format is a great challenge. This problem can be addressed with the help of Profile Mining - by designing an effective and efficient profile extractor that can extract the maximum online available information related to a person and present it in a proper format. Hence, automated data extraction is required that can produce the desired results. Various tools/techniques are available to facilitate data extraction such as APIs (Application Programming Interfaces), web scraping, spy or caller ID applications and penetration testing tools. Some tools are open-source whereas other are paid.This research paper focuses on analysing tools and techniques of data extraction from social media platforms.

Keywords:

Profile Mining;
APIs;
Web Scraping;
Penetration Testing;
HTML Parsing.

Author correspondence:

*Priti Bali
Research Scholar
Dept. of Computer Science
Singhania University
Pacheri Bari, Jhunjhunu, Rajasthan
jmd_priti@rediffmail.com

**Dr. Anoop Sharma
Research Guide
Dept. of Computer Science
Singhania University
Pacheri Bari, Jhunjhunu, Rajasthan
sharmaanoop001@gmail.com

1. Introduction

Profile mining is defined as the automatic extraction of characteristics of a person depending upon the given tokens (inputs). Nowadays, huge amount of information is available online.Almost every person has web presence but information is available in parts on different web pages not on a single page.Rather it is scattered on different applications/social media platforms and display different characteristics of a person. Various apps and tools are available for extracting this information but they provide very less information about a person and that too not in a proper format. Also, techniques like manual copy/pasteand using search engines for extracting the information is a very time-consuming process.Social media like Facebook and Twitter offer APIs for providing access to their data. These tools offer many features but has some limitations

also. Data in SocialMedia is accessible through respective APIs but due to increasing cyber crimes and frauds, comprehensive access to Social Media data is not possible. Social media like Facebook does not allow comprehensive access to data for researches whereas others charge premium for accessing the data. So, for effective data extraction some combination of existing tools and techniques may be usedthat can extract the maximum characteristics of a person and present that information in a proper way. Profile mining is very useful in many areas such as employers need complete profile of a person before hiring a person, parents need to know the complete profile of a person to whom kids are chatting or talking, advertising agencies need user profiling to choose the advertisements that will be shown to user etc. This detailed information is very helpful for combating cyber-crimes and minimizing cyber-criminal activities.[1]

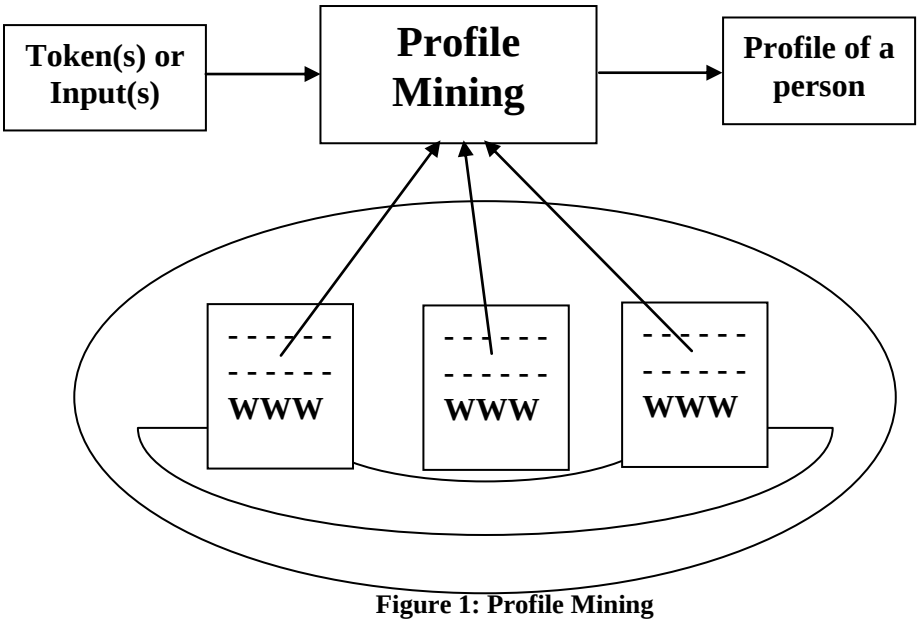

Figure 1: Profile Mining

Relevance of Profile Mining:

- Useful for parents: It will help the parents in getting the detailed relevant information about their kid's online activities. Also, some apps like SafeTeen offers ultimate parental control service that locks the teen's mobile phone while driving so that kid does not get distracted.
- Useful for employers: It will help the employers in getting the detailed relevant information about their employees or before hiring a person.
- Useful for customers: It will help the customers in verifying the profile of a person by extracting the detailed information about that person who is claiming that he/she is a sales executive (company representative) of a renowned company.
- Useful for advertising agencies: It will help the advertising agencies in extracting the customer's profile. On the basis of extracted information, they can choose the advertisements that will be shown to user.
- It will be helpful for all the researchers involved in profile mining process.
- It will be helpful for keeping track of any person (spouse, old friends, old teachers, marital status)
- Useful for cyber cells (for combating cyber-crimes)

2. Literature Review

Various researches have been done on the tools and techniques for extracting the data from social media and web pages. Some of them are mentioned below:

Saurkar, Pathare and Gode (2018) in their research paper have suggested that web scraping technique can be used to crop information from web pages. Web pages are represented by DOM (Document Object Model)

tree and HTML is used for specifying the text format. They have also mentioned that many software tools such as Mozenda, Visual Web Ripper, Web Content Extractor, Import.io, Scrapy, Scraper and ParseHub are available nowadays to automate web-scraping, there is no need to write web-scraping code manually. These tools can extract the data from targeted website and further data can be downloaded in CSV, Excel, JSON format etc. [2]

Mitchell (2018) has covered in her book about "Reasons why an API might not exist or may not be used" that there may be case where source of data does not have the required infrastructure or technical ability for creating an API, other reasons are if the required data is fairly small that the webmaster did not think warranted an API and if data is to be extracted from a collection of sites that do not have a cohesive API. Author has also mentioned that even when an API does exist, request volume and rate limits, the types of data or the format of data that it provides might be insufficient for your purposes. Author has mentioned that web scraping is very useful and efficient for extracting the unstructured data from websites and transforming it into structured format. Author has also mentioned that if data can be accessed through browser, it can be accessed via Python Script and if it can be accessed via Python Script, it can be stored in a database and if it can be stored in a database, virtually anything can be performed with that data. [3]

Dawar, Purwar, Anand and Singla (2018) in their research paper have stated different tools such as Twitter Archive, Bird Song Analytics, Cyfe, NodeXL, TWChat, REST API and Streaming APIs for extracting the data from twitter database. They have also mentioned that APIs take an instruction and performs an action for the user. They have also explained the difference between REST API and Streaming API. REST API makes use of authorization account for performing actions on any user's account, this API provides real time data and allow to perform queries on it whereas Streaming API is a long running request that extracts the data whenever it is available. [4]

Scarno (2018) in his research paper has mentioned that WWW contains huge amount of data but this data is available in unstructured form. Author has stated that web scraping techniques can extract and transform unstructured data to structured data. He has also mentioned that web scraping is a process of automated data extraction from one or more websites and simulate human's browsing process. [5]

Umair, Nanda and He (2017) concluded that there are a lot of challenges in terms of data extraction from Facebook. Tools which are capable of extracting in-depth information requires targeted ID's authorization. Another challenge is Facebook rapidly changes its API and format because of these rapid changes, various research-based data extraction applications stop working or their functionality becomes limited. They also stated that these changes are made for ensuring and maintaining the user's privacy and security. Also, it has been mentioned that changes in API are inevitable because privacy of users is must. [6]

Teixeira and Laureano (2017) in their research paper proposed a personalized tool for extracting the data from Facebook. They have used Facebook Graph API along with a Java client RestFB. It has been stated that RestFB client automates the extraction process and simplifies the access to the core functions of Facebook Graph API. It has also been stated that RestFB is best suitable for custom implementation. Certain limitations of Facebook Graph API have been mentioned such as API gets blocked when called a large number of times consecutively. Data can be extracted yearwise (or some other suitable time intervals) due to the limits of number of calls that can be made to the Facebook API. [7]

Salloum, Emran and Shaalan (2017) in their research paper have mentioned that Facebook is a major source of data. They have also stated that Facebook API can be used for recovering the data from Facebook database. Application Secret Key and Facebook API Key are executed by Facebook API for the recovery of data. In case of Twitter, Twitter's Streaming API is used for capturing streaming tweets. [8]

Zhao (2017) in his research paper have mentioned that modern web contains vast amount of heterogeneous data and this data is being generated constantly on WWW. Author has mentioned that web scraping is widely acknowledged as an efficient and powerful technique for collecting the big data. Author has explained the process of web scraping in detail. He has stated that web scraping process starts with a HTTP request to acquire data from the targeted website by using URL containing a GET query or HTTP message containing a POST query. If the request is successfully received and processed then requested data is sent back to the web scraping program. Data can be in multiple formats such as XML, JSON, images, audio or video files. So, this data is formatted in the required structured format. Hence, there are two main modules of a web scraping process: one module deals with HTTP requests with the help of Urllib package which preprocess the URLs and another for parsing and data extraction from HTML code with the help of BeautifulSoup. BeautifulSoup

automatically detects the encoding of parsing under processing and convert it into a client-readable encode. Author has also mentioned that Selenium is also used to deal with web browsers, it is a web browser wrapper that helps in automating the process of browsing a website via code. Selenium supports almost all modern browsers. Webdriver package enables browser automation for many programming languages such as Python, C# and Java. Author has stated that in order to mitigate the complexity of performing web scraping through program, import.io crawler can be used for extracting data without writing any code. Extracted data is stored in dedicated cloud server and can be exported in CSV or any other required format. [9]

Naseem (2017) in his research paper has stated that data on social media is in scattered form and for extraction process, this data needs to be organized. He has mentioned that NLP (Natural Language Processing) techniques can be used to analyze the scattered data to extract the information such as Event, Category, Date, Place and Time Period. Such extracted information can be stored in database for further use. [10]

Syrien and Hanumanthappa (2016) in their research paper have mentioned that data from social media can be collected through several data crawler tools such as WebHarvest (Java) and Crawler4j (Java). WebHarvest tool focuses on HTML/XML websites whereas Crawler4j is executed with apache ant. These tools can extract the data like birthdate, general location, habits, likes, personal tags, re-twitter counts etc. [11]

Meschenmoser, Meuschke, Hotz and Gipp (2016) in their research paper have stated that dynamic content can be accessed via network logs which is available in browsers. In these logs, URLs that are processed during POST and GET requests can be accessed easily. They have identified various challenges of web scraping like size limitations in result sets, dynamic content of websites and access barriers. They have proposed solutions for addressing these challenges but these solutions cause large performance overhead. [12]

Brooker, Barnett and Cribbin (2016) in their research paper have stated that Twitter's API allow users to extract a range of data entities and associated values. They have mentioned two approaches for the collection of data. First is "Query Keyword Search" which uses hashtags, words and URLs. Second is, "User Following Strategy" based on capturing user-driven data. [13]

Sirisuriya (2015) in his research paper has mentioned that web scraping is a process of automatic data extraction. Author has also mentioned that information on WWW is available in different formats so indexing becomes difficult. For addressing this problem, web scraping is used. Web scraping tools and techniques extract unstructured data and transform it into structured data like spreadsheets, CSV files or database. Author has also mentioned various techniques of web scraping such as traditional copy and paste, text grapping and regular expression matching, HTTP programming, HTML parsing, DOM parsing, Webscraping software, Vertical aggregation platforms, Semantic annotation recognizing and Computer vision web-page analysers. [14]

Alves, Damasio and Correia (2015) in their research paper have stated that Facebook provides a Graph HTTP-based API that provides access to the Facebook Social Graph. Facebook graph represents objects such as people, photos, events and pages along with connections between them such as friend relationships, shared content and photo tags. They have also mentioned that Facebook has its own query language FQL (Facebook Query Language) which is quite similar to SQL. They have also stated that Graph API returns the data in JSON format. [15]

Purohit, Bhat, Angadi and Gull (2015) in their research paper have stated that three different type of web pages (unstructured, structured and semi-structured) are available from which data can be extracted. It has been mentioned that APIs are open-source and can be used for data extraction but authenticated requests are must to access the APIs. Several tools such as ECON, VIPS and POLYPHONET have been mentioned. They have stated that APIs are the primary way to extract data from social medias. Twitter uses two APIs: REST API and STREAM API whereas Facebook uses Open Graph API. They have also mentioned that OAuth (Open Authorization) which is an open standard for authentication has been adopted by Facebook/Twitter for providing access to protected information and the whole process is carried via three-way handshake. [16]

Batrinca and Treleaven (2014) in their research paper have mentioned that social media data can be accessed and extracted through freely available databases, tools and APIs. They have also stated that most useful source of social media data is that which provides programmable access via APIs using HTTP-based protocols. Almost all social medias like Facebook and Twitter provide access to a proportion of their data via

APIs. Social medias like Bing, LinkedIn and Skype does not provide access via APIs. They have also stated that a large number of websites provide access to data via RSS feeds. Also, blog scraping can be used for getting website's source code via Java's URL class which can be parsed further via Regular Expressions for capturing the required information. News feeds and Geospatial feeds can also be used for extracting the required information. They have concluded that social media data can be accessed via APIs but due to commercial value of data major sources like Facebook and Google does not provide comprehensive access to their raw data. They restrict the data access to monetize their data, very few social medias are providing affordable data access to academia and researchers. [17]

Cuesta, Barrero and Moreno (2014) in their research paper have mentioned that tweets can be accessed via a public API provided by Twitter. This access is available in both forms, static and streaming. They have mentioned that MongoDB database has been selected for storing tweets as the data is being extracted in JSON format. [18]

Rieder (2013) in his research paper stated that user's data from their social media profiles can be collected by using either traditional empirical methods such as interviews, observations, experiments or data crawlers (without active participation of users). He has also stated that there are three ways for extracting the data from social profiles. First, Direct Database Access in which access is reserved for in-house researchers. In this case, data is generally well-structured and very large. Second, Access through APIs in which data is accessed through APIs. In this case, data is well-structured but limited in terms of how much data, which data and how often (time intervals) data can be retrieved. Third, User interface Crawling in which data is extracted manually, mostly bots or spiders are used for reading HTML. He has also stated that these data extraction techniques can overcome the limitations of APIs but custom programming and lot of manual work is required. To overcome these problems several data extraction tools have been developed over the years, using APIs. He has developed a data collection and extraction tool "Netvizz" which can extract and export data in standard file formats from different sections of Facebook. Proposed tool eliminates custom programming and manual collection of data.Netvizz data extractor has been written in PHP that extracts data from personal networks, groups and pages. It extracts data such as count for posts and likes, more active users, posts that produced maximum engagement, comments etc. [19]

Various open-source and paid penetration testing tools, Spy or CallerID apps have been studied. Some of them are mentioned below:

Penetration testing is an offensive testing method that determines how much the network and web applications are vulnerable, what are the loop holes that can be exploited to get into the system for extracting the information and what effort is required to break into the system for restricted (no access privileges) as well as unrestricted access (in case of employees with all access privileges). [20]Following are some popular penetration testing tools which are used for the extraction of data from various sources:

Bhatt (2018) in his research paper has discussed the Metasploit framework for data extraction.Metasploit framework of Kali Linuxis most widely used by the pen-testers for infiltrating an organization's network, finding and exploiting the vulnerabilities in that network. Nmap is also used with Metasploit for finding network details. Metasploit framework has a number of built-in exploits that can be used for conducting various kinds of pen-test as these exploits can surpass the security measures to get into a system. This tool comes with "Meterpreter" session which automatically starts at the host machine as soon as the victim installs an exploit. Exploit may be in the form of an application, executable file or image file containing the malicious code. Pen-testers use this tool for analyzing the weaknesses of target system (an organization's network in this case). On the basis of identified vulnerabilities, pen-testers formulate the defense strategies that must be implemented to safeguard the IT infrastructures.For example: To perform mobile penetration testing, msfvenom (which is a combination of two tools msfpayload and msfencode) is used for generating payloads in various formats such as application file (apk files), exe files etc. After generating payload, listener/handler is started using msfconsole to exploit the payload for listening to the victim's mobile (exploit means a successful attack). Generated payload (apk file) is sent to the victim's device using any social engineering method and as soon as victim installs the apk file, meterpreter session starts on the host machine and almost all the information of victim's mobile is accessible to pen-testers. In this way, mobile device is penetrated with exploit and pen-testers can find the vulnerabilities. Metasploit is capable of creating backdoor in the target system. This backdoor can be used for extracting the required information later on. [21]

Samantha and Phanindra (2018) in their research paper have discussed about Sqlmap tool.Sqlmap is again a good open-source Pen-Testing tool. This command-line tool automates the process of identifying and exploiting SQL injection flaws in web applications and hacking over of web database servers. This tool is capable of fetching data, accessing the vulnerable file system and executing OS commands. This tool comes with database fingerprinting feature and supports almost all the databases such as Oracle, Hyper SQL Database, MS Access, Sybase, MySQL, Firebird, Microsoft SQL Server and many more. The information retrieved by this tool can be used further for extraction of required information. [22]

Geethamani and Priyanka (2017) in their research paper have discussed about HconSTF tool. HconSTF is very useful penetration testing tool that allows for creating web exploits that can be used for information gathering in the areas of passwords, databases, networks etc. [23]

Gupta and Anand (2017) in their research paper have mentioned some tools such as SuperScan, Angry IP Scanner, Nikto, Unicornscan, AutoScan, Wireshark, Sqlninja, Netsparker, BeEF, Dradis, Nessus, OpenVAS and Retina that are used by ethical hackers. They have also discussed various cyber-attacks such as DoS (Denial-of-Service), MiTM (Man-in-The-Middle) using Ettercap Tool and Wi-Fi Phishing launched by hackers for extracting target's data. [24]

Dmitry (Deepmagic Information Gathering Tool) is used in the initial steps of pen-testing for gathering information such as subdomains, email addresses etc. This tool extracts detailed information about target websites, this information can be further used for extracting the required information. It is available in both command-line and GUI interface. [25]

theHarvester is an information gathering tool which is used by the pen-testers in the early stages of pen-testing for finding the detailed information about a network such as subdomains, e-mail addresses, IP addresses, open ports, user names, virtual hosts, URLs, banner (name and version information of software) and many more. This tool retrieves the data from various sources such as popular search engines, social medias, shodan database and servers. This tool uses methods such as DNS brute force and DNS reverse lookup for information gathering. IP addresses, user names and passwords can be exploited further for the extraction of required information. [26]

Maltegois a powerful OSINT(Open-Source Intelligence)tool which is capable of tracking names, email-addresses, passwords and other pieces of information. This tool comes in various versions such as eXtra Large (paid), Commercial Version (paid), Community Edition (free) etc. In all the versions of Maltego tool, various transforms are available. Some transforms are free whereas some transforms are paid. For example: "Have I Been Pwned" transform is free, this transform requires E-mail address as an input and provides personal information such as passwords, locations as an output. This tool can be used on any operating system, it is pre-installed in Kali Linux operating system. [27]

Social Engineering Toolkit (SET) attempts to make a person reveal sensitive information like password, business-critical data etc. These attacks are mostly done through phone or internet and it targets certain helpdesks, employees & processes. Social Engineering is the most popular methodology for extracting sensitive data. [28]

Some other penetration tools are SuperScan, Angry IP Scanner, Nikto, Unicornscan, AutoScan, Wireshark, Sqlninja, Netsparker, BeEF, Dradis, Nessus, OpenVAS, Retina, Netdiscover, Security onion, SearchSploit, RouterSploit, WPScan, CMSMap, Nessus, Netsparker, John the Ripper, OpenVAS, Aircrack-ng, SQLsus, SQLmap, Havij, jSQL, Unicornscan, Fluxion, Acunetix, Wapiti, Arachni, Burp Suite, IronWASP, Ettercap, Canvas, Indusface, Dradis, Hping, Websecurify, Cain and Abel, Google Dorks and many more.

3. Results and Analysis

On reviewing the literature, it is clear that tools and techniques of data extraction provide automated support for gathering publicly available information. These tools save a lot of time, on the other hand manual techniques and using search engines for extracting the information is a very time-consuming process. Search Engines also provide certain tools whereas Social Media provide APIs for facilitating the data extraction but most of the tools offer very limited functionality in free versions. These tools are capable of extracting different pieces of information related to a particular person. For example, some tools are capable of

extracting email addresses whereas others are capable of extracting names, locations and other personal information. Outputs produced by these open-source tools can be combined and presented in a proper format.

Many social media like Facebook and Twitter offer API (Application Programming Interface) for providing access to their data. APIs also facilitate the data extraction process but APIs provide access to data repository via computer programs but in this case owner of the data has the overall control over who can access and how much data can be accessed. There are various limitations of using APIs for data extraction such as targeted website does not provide an API, API provided is paid, API is rate limited (specified number of times per minute or per day) and API does not give access to all the data. Web scraping can be used for effective and efficient data extraction from multiple sources. Web scraping tools and techniques can harvest any kind of data from web pages whether it is structured or unstructured. Web scraping technologies not only harvest the data from web pages, these technologies are capable of analysing the web pages for acquiring the desired data and converting that extracted data into structured format. Web scraping is more effective and flexible than APIs as it can be used on more web pages in comparison with APIs because all websites do not offer APIs. Hence, combination of above-mentioned tools and techniques can be used for efficient and effective automateddata extraction.

On the basis of literature review, following model (Figure 2) has been proposed after analysingthe tools and techniques of data extraction. Profile Extractor can extract the maximum publicly available information of a person by using tools and techniques mentioned in the following diagram (Figure 2) and present it in a proper format.

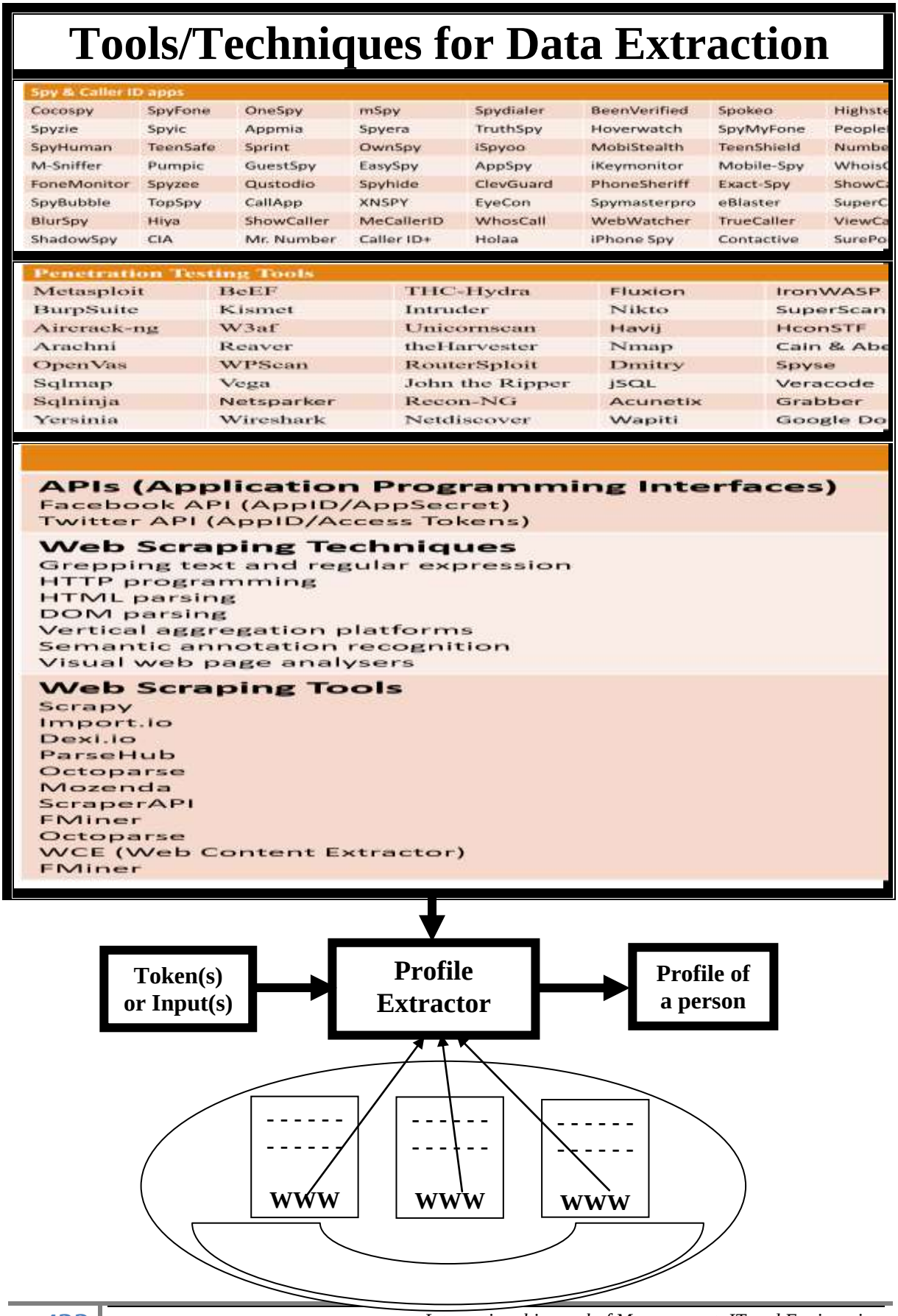
Tools/Techniques for Data Extraction
Spy & Caller ID apps
Cocospy SpyFone OneSpy mSpy Spydialer BeenVerified Spokeo
Spyzie Spyic Appmia Spyera TruthSpy Hoverwatch SpyMyFone
SpyHuman TeenSafe Sprint OwnSpy iSpyoo MobiStealth TeenShield
M-Sniffer Pumpic GuestSpy EasySpy AppSpy iKeymonitor Mobile-Spy
FoneMonitor Spyzee Qustodio Spyhide ClevGuard PhoneSheriff Exact-Spy
SpyBubble TopSpy CallApp XNSPY EyeCon Spymasterpro eBlaster
BlurSpy Hiya ShowCaller MeCallerID WhosCall WebWatcher TrueCaller
ShadowSpy CIA Mr. Number Caller ID+ Holaa iPhone Spy Contactive
Penetration Testing Tools
Metasploit BeEF THC-Hydra Fluxion IronWASP
BurpSuite Kismet Intruder Nikto SuperScan
Aircrack-ng W3af Unicornscan Havij HconSTF
Arachni Reaver theHarvester Nmap Cain & Abe
OpenVas WPScan RouterSploit Dmitry Spyse
Sqlmap Vega John the Ripper jSQL Veracode
Sqlninja Netsparker Recon-NG Acunetix Grabber
Yersinia Wireshark Netdiscover Wapiti Google Do
APIs (Application Programming Interfaces)
Facebook API (AppID/AppSecret)
Twitter API (AppID/Access Tokens)
Web Scraping Techniques
Grepping text and regular expression
HTTP programming
HTML parsing
DOM parsing
Vertical aggregation platforms
Semantic annotation recognition
Visual web page analysers
Web Scraping Tools
Scrapy
Import.io
Dexi.io
ParseHub
Octoparse
Mozenda
ScraperAPI
FMiner
Octoparse
WCE (Web Content Extractor)
FMiner
Token(s) or Input(s)
Profile Extractor
Profile of a person
WWW
WWW
WWW

Figure 2: Profile Extractor

4. Future Directions

Proposed Profile Extractor can be used as a base for Web Application Development by researchers and developers. Proposed extractor takes person's name or mobile numberor any other token as an input, uses a combination of tools/techniques of data extraction for retrieving the informationand produces the desired information in a proper format. For better results, regular expressions could be used for matching the extracted information with user's preferences so that the user gets relevant information.

5. Conclusion

Various researches have been done on the tools and techniques for extracting the data from social media and web pages. Most of the tools have been developed to support academia and research but most of the companies like Facebook and Twitter keep on changing the APIs for ensuring and enhancing the security and privacy of their user's social media account. Also, Governments introduce and amend data protection bills and laws time to time for ensuring and enhancing the information security. So, companies are forced to do the corresponding changes in their security policies, procedures and measures. Hence, if data extraction is not possible via APIs then techniques such as search engines, social networks, email address technique, username technique, real name technique, location technique, IP address technique, domain name technique can be used but with automation otherwise it will become a very tedious task if performed manually. Data can be extracted through websites also. Web scraping uses the process of HTML parsing, structured HTML page is parsed into a DOM (Document Object Model) tree. DOM makes a logical hierarchical structure out of HTML page so that node that contains the required information can be identified easily. DOM tree facilitates the process of finding the specific location of the required data.Web scraping tools and techniques facilitate the automated data extraction to a great extent. Hence, combination of above-mentioned tools and techniques can be used for efficient and effective automated data extraction. For example, AppID generated by the Facebook's API can be used further in various web scraping techniques for sending several requests to Facebook for data.

References

[1] W. Dennis, A. Erwin, M. Galinium, "Data Mining Approach for User Profile Generation on Advertisement Serving", 2016, 8th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia, 978-1-5090-4139-8/16

[2] A.V. Saurkar, K. G. Pathare and S. A. Gode, "An Overview On Web Scraping Techniques And Tools", International Journal on Future Revolution in Computer Science & Communication Engineering (IJFRCSCE), April 2018, ISSN: 2454-4248, Volume: 4, Issue: 4, p.p. 363 – 367.

[3] R. Mitchell, Web Scraping with Python: Collecting Data from the Modern Web, Second Edition, April 2018, O'Reilly Media, USA

[4] A. Dawar, A. Purwar, N. Anand and C. Singla, "TweetRush: A Tool for Analysis of Twitter Data", I.J. Education and Management Engineering, March 2018, 2, Published in MECS (http://www.mecs-press.net) p.p. 31-40.

[5] M. Scarno, "Use of Artificial Intelligence and Web Scraping methods to retrieve information from the World Wide Web", V. Surekha. Int. Journal of Engineering Research and Application, www.ijera.com, ISSN: 2248-9622, Vol. 8, Issue 1, (Part -II), January 2018, p.p. 18-25, Research Gate, https://www.researchgate.net/publication/322520038

[6] A. Umair, P. Nanda and X. He, "Online Social Network Information Forensics: A survey on use of various tools and determining how cautious facebook users are?", published in proceedings of IEEE Trustcom/BigDataSE/ICESS conference, August 2017, https://ieeexplore.ieee.org/document/8029567

[7] A. Teixeira and Raul M.S. Laureano, "Data extraction and preparation to perform a sentiment analysis using open source tools: The example of a Facebook fashion brand

page, published in proceedings of 12th Iberian conference on Information Systems and Technologies (CISTI), June 2017, p.p. 2064-2069, https://ieeexplore.ieee.org/document/7975879

[8] S. Salloum, M. Emran and K. Shaalan, "A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives", Advances in Science, Technology and Engineering Systems Journal (ASTESJ), Vol. 2, No. 1, Jan 2017, ISSN: 2415-6698, p.p. 127-133.

[9] B. Zhao, "Web Scraping", Springer International Publishing AG (outside the USA) 2017 L.A. Schintler, C.L. McNeely (eds.), Encyclopedia of Big Data, DOI 10.1007/978-3-319-32001-4_483-1.

[10] S. Naseem (2017), "Extracting events from social media using NLP", VFAST Transactions on Software Engineering http://vfast.org/journals/index.php/VTSE@ 2017 ISSN(e): 2309-3978; ISSN(p): 2411-6246 Vol 5, No. 1, January-December 2017, p.p. 44-49.

[11] A. Syrien and M. Hanumanthappa, "International Journal of Advanced Research in Computer and Communication Engineering" (IJARCCE), Oct 2016, ISSN: 2278-1021, Vol. 5, Issue 2, p.p. 39-42.

[12] P. Meschenmoser, N. Meuschke, M. Hotz, B. Gipp, "Scraping Scientific Web Repositories: Challenges and Solutions for Automated Content Extraction", DOI: 10.1045/september2016-meschenmoser, https://www.dlib.org/dlib/september16/meschenmoser/09meschenmoser.html

[13] P. Brooker, J. Barnett and T. Cribbin, "Doing social media analytics", Big Data & Society, bds.sagepub.com, July–December 2016, p.p. 1–12.

[14] SCM de S Sirisuriya, "A Comparative Study on Web Scraping", Proceedings of 8th International Research Conference, KDU, Published November 2015

[15] M. B. Alves, C. V. Damasio and N. Correia, "Extracting Metadata from Multimedia Content on Facebook as Media Annotations", Conference Paper, Springer International Publishing Switzerland, Oct 2015, p.p. 243-252. https://link.springer.com/chapter/10.1007/978-3-319-24543-0_18

[16] N. S. Purohit, M. Bhat, A. B. Angadi and K. C. Gull, "Crawling through Web to Extract the Data from Social Networking Site – Twitter", published in National Conference on Parallel Computing Technologies (PARCOMPTECH), Date of Conference: 19-20 Feb. 2015, https://ieeexplore.ieee.org/document/7084522

[17] B. Batrinca and P. Treleaven, "Social media analytics: a survey of techniques, tools and platforms", Springer, AI & Society, July 2014, p.p. 89-116, https://link.springer.com/article/10.1007/s00146-014-0549-4

[18] A. Cuesta, D. F. Barrero and M. D. R. Moreno, "A Framework for massive Twitter data extraction and analysis", Malaysian Journal of Computer Science. Vol. 27(1), 2014, p.p. 50-67.

[19] B. Rieder, "Studying Facebook via Data Extraction: The Netvizz Application", published in proceedings of the 5th annual ACM web science conference, 2013, p.p. 346-355.

[20] N. Bhingardeve, S. Franklin, "A Comparison Study of Open Source Penetration Testing Tools", IJTSRD, ISSN No: 2456-6470, Vol. 2, Issue – 4, May-Jun 2018, p.p. 2595-2597.

[21] D. Bhatt, "Modern Day Penetration Testing Distribution Open Source Platform - Kali Linux - Study Paper", International Journal of Scientific & Technology Research Volume 7, Issue 4, Apr 2018, ISSN 2277-8616, p.p. 233-237.

[22] B. S. Samantha, M.V. Phanindra, "An overview on the utilization of Kali Linux Tools", IJRAR-International Journal of Research and Analytical Reviews, Vol. 5, Issue 2, Apr – June 2018, e ISSN 2348 –1269, Print ISSN 2349-5138, p.p. 104-113. http://www.hcon.in/uploads/1/8/1/9/1819392/hconstf_manual.pdf

[23] G. S. Geethamani, G. Priyanka, "A Study of Ethical Hacking with best penetration testing tools applied in cyber security", International Journal for Research & Development in Technology", Vol.-8, Issue-6, Dec-17 ISSN(O)- 2349-3585, p.p. 194-198.

[24] A. Gupta and A. Anand, "Ethical Hacking and Hacking Attacks", International Journal of Engineering and Computer Science, ISSN:2319-7242, Vol. 6, Issue 4, April 2017, p.p. 21042-21050.

[25] https://www.mynotepaper.com/how-to-use-dmitry-tool-to-gather-website-information-on-kali-linux, accessed on Jan 19

[26] https://www.darknet.org.uk/2012/01/theharvester-gather-e-mail-accounts-subdomains-hosts-employee-names-information-gathering-tool/, https://securitytrails.com/blog/osint-tools, accessed on March 19

[27] https://www.cyberpratibha.com/information-gathering-with-maltego/, accessed on May 2018

[28] https://securitytrails.com/blog/the-social-engineering-toolkit, accessed on Jan 19