Crawling and Mining Social Media Networks: A Facebook Case

Abd El Salam Al Hajjar, Haissam Hajjar, Mazen El Sayed, Mohammad Hajjar
Institute University of Technology
Lebanese University
Lebanon
e-mail: abdsalamhajjar@hotmail.com, haissamh@ul.edu.lb, mazen_elsayed@yahoo.fr, m_hajjar@ul.edu.lb

***Abstract*—Social media is computer-mediated tool that allows people to create, share or exchange information, ideas, and pictures/videos in virtual communities and networks. The most popular social media in these days is Facebook. This paper treats the problems of web crawling and mining. More precisely, we focus our intention on Facebook information extraction, and its importance in gathering data and tracking people. Also, we focus on the crawling mechanism of several Facebook pages and extracting the data contained in it and finally storing it in our database. Specifically, we extract the basic information, such as home town, age, work, education, friends list, events and other information. At first, we describe the content of Facebook web pages and the principle of web crawling and mining. Second, we propose the architecture of our system, which allows extracting the Facebook information, for a specific user logged in. The result of our work is an automated system that takes a Facebook user as an input, extracts recursively the list of friends for this user, and returns the friends information (name, university, etc.).**

Keywords-Web data crawling and mining; social media network; Facebook; HTML; PHP; MySQL; database.

I. INTRODUCTION

The “Social media” axiom is widely used these days; it is a computer-mediated tool that allows people to create, share or exchange information, ideas, and pictures/videos in virtual communities and networks [2]. Social media is defined as "a group of Internet-based applications that build on the ideological and technological foundations of web and that allows the creation and exchange of user-generated content". Furthermore, social media depends on mobile and web-based technologies in order create highly interactive platforms through which individuals and communities share, co-create, discuss, and modify user-generated content [4].

Social media is different from other traditional or industrial media in many ways, including quality, usability, immediacy, and permanence [3]. Internet users spend more time with social media sites than any other type of site. Furthermore , the total time spent on social media in the U.S. increased by 99 percent to 121 billion minutes in July 2012 compared to 66 billion minutes in July 2011 [4][5].

In 2014, the most popular used social network was Facebook, comparing to other networks such as Twitter, Instagram, LinkedIn, Pinterest, etc. [6].

Facebook is an online social network service. Its website was launched on February 4, 2004, by Mark Zuckerberg with his college roommates and fellow Harvard University students Eduardo Saverin, Andrew McCollum, Dustin Moskovitz and Chris Hughes [7]. The founders had initially limited the website's membership to Harvard students, but later expanded it to colleges in the Boston area, the Ivy League, and Stanford University. It gradually added support for students at various other universities and later to high-school students. Facebook now allows anyone who claims to be at least 13 years old to become a registered user of the website [8]. Its name comes from a colloquialism for the directory given to it by American universities students [9]. After registering to use this type of site, users can create a user profile, add other users as "friends", exchange messages, post status and photos, share videos and receive notifications when others update their profiles. Additionally, users may join common-interest user groups, organized by workplace, school or college, or other characteristics, and categorize their friends into lists, such as "people from work" or "close Friends". Facebook had over 1.3 billion active users as of June 2014 [10]. Due to the large volume of data collected about users, the service's privacy policies have faced scrutiny, among other criticisms. Facebook held its initial public offering in February 2012 and began selling stock to the public three months later, reaching a peak market capitalization of 104 billion.

A social media network has a large volume of data; therefore, it will provide a big amount of useful information (people pictures, relations, etc.). So, it is very important to follow a methodology to extract and analyze information from social media web site.

In this work, we collect the data from the social media network, specifically Facebook, organize them in a database, analyze these data, and show the information on screen. The presented information can be a list of all friends at many levels (e.g., level1 gives the only friends of the selected person, level2 presents also the friends of each friends. etc.), and it can be events as like/comment/share for each friends already extracted, etc.

In the next section, we present the web crawling and mining definition and concept. Section 3 presents the social media and Facebook. Section 4 presents the Facebook web crawling and mining methodology, and presents the Facebook content with the architecture of the system. Section 5 presents the result. Section 6 describes the conclusion and the future work.

ISBN: 978-1-61208-412-1

II. Web Crawling and Mining

A web crawler is an Internet bot (a bot is an automated application used to perform simple and repetitive tasks that would be time-consuming, impossible for a human to perform), that systematically browses the World Wide Web, typically, for the purpose of web indexing. A Web crawler may also be called a web spider [11]. Web search engines and some other sites use web crawling to update their web content or indexes of others sites' web content. Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly. Web crawlers can validate hyperlinks and HTML (HyperText Markup Language) code. They can also be used for web scraping (see also data-driven programming) [12]. Within the past few years there has been an increase of free web crawler datasets [13]. The challenges become increasingly difficult when doing this on a larger scale. However, capturing the content of a single web page is quite easy. Below, we present the main steps of the web crawler:

i. Select a URL (Uniform Resource Locator).
ii. Fetch and parse the corresponding page.
iii. Save the important content into database.
iv. Extract URLs from this page.
v. Add URLs to queue.
vi. Select a URL and repeat.

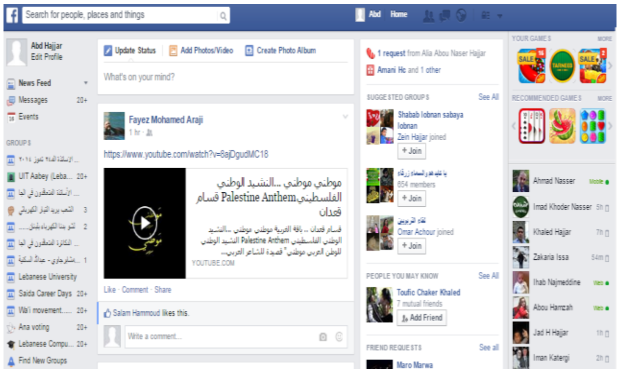
Data-mining is the analysis step of the Knowledge Discovery in Databases (KDD) process [14]; it is allowed to process big amounts of data to provide meaningful and relevant information. The collected information are in an unstructured form, must be transformed into a structured format to be suitable for processing. The data mining technology is coming from a huge evolution; the new and better technique is made available continually to gather whatever information is required [15]. The term “web data mining” is a technique used to crawl through various web resources to collect required information, which enables organizations and individuals to gather information, and to utilize this information in the best interest [16]. The advantage of the web data mining can be shown in the following general example: a company is thinking about launching a new product of cotton shirts, through the client databases founded on web, so they can clearly determine how many clients have placed orders for cotton shirts over the last year and how much profit such orders have brought to the company. The disadvantage is resumed by losing the user privacy when individual information is obtained, used, and distributed, especially if this happens without the user knowledge.

III. Social Media Network, Facebook Case

Social media network is the cooperation of online communications channels dedicated to community-based input, interaction, content-sharing, and collaboration [17]. Social networking is the practice of growing the number of one's business and social contacts by making connections through peoples. While social networking becomes societies themselves, the unparalleled potential of the Internet to promote such connections is only now being fully recognized and exploited, through web-based groups established for that purpose[1][2] (see Figure 1).

Figure 1. Social media networking.

The most popular social media networking sites in the world are Facebook, Google+, LinkedIn, Instagram, and Twitter, etc.



Fayez Mohamed Araji
1 hr

https://www.youtube.com/watch?v=8ajDgudMC18

موطني موطني ...النشيد الوطني الفلسطيني Palestine Anthem قسام قعدان

قسام قعدان .. باقة الغربية موطني موطني ...النشيد الوطني الفلسطيني Palestine Anthem النشيد الوطني للوطن العربي موطني" قصيدة للشاعر العربي...

YOUTUBE.COM

Salam Hammoud likes this.

Figure 2. Facebook general layout

Facebook represents a potentially useful tool in educational contexts (see Figure 2). It allows for both an asynchronous and synchronous, open dialogue via a familiar and regularly accessed medium, and supports the integration of multimodal content, such as student-created photographs, video, and URLs to other texts, in a platform that many students are already familiar with. Further, it allows students to ask more minor questions that they might not feel motivated to visit a professor personally during office hours. It also allows students to manage their own privacy settings, and often work with the privacy settings they have already established as registered users [18].

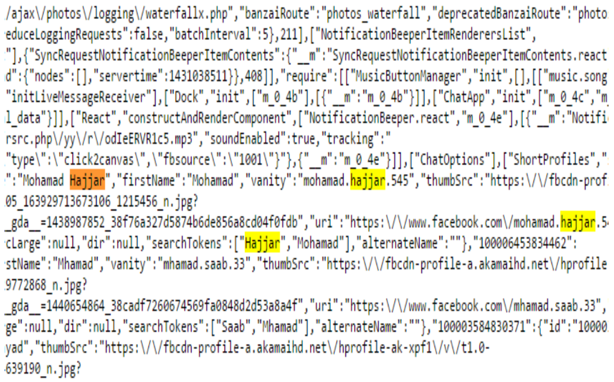
IV. Facebook Crawling and Mining Methodology

In this section, we present the crawling and mining methodology for a social media networking, specifically, a Facebook case. We describe the Facebook web pages content and the system architecture that allows crawling and mining automatically the Facebook information for a specific user logged in.

A. Facebook Content

The first page in Facebook is the login page that allows authenticating a user. If this user is authenticate, the user can navigate different pages in the Facebook website, starting from the profile that includes the following: basic information, messages, photos, friends, notes, status, comments, groups, pages, and the wall. A user is able to search for friends by e-mail address, or just by typing a name of the friend. When people become friends, they are able to see all of each other's profiles including contact information. The user also can create groups. This group allows members who have common interests to find and interact with each other. Also, the user can create Facebook pages that contain many members (other Facebook users). Beside all that, Facebook contains a set of games which allows users to play online, etc.

The target is crawling and extracting data from the Facebook pages, then, saving the data into a database, using a specific web programming language. When, Facebook is not anymore a secure network, we can get the html source of the page using any browser. However, some data, maybe for privacy reasons cannot be viewed. So, we only collect public data organize and view them in the simplest desired form.



```
/ajax\/photos\/logging\/waterfallx.php","banzaiRoute":"photos_waterfall","deprecatedBanzaiRoute":"photo
educeLoggingRequests":false,"batchInterval":5},211],["NotificationBeeperItemRenderersList",
"],{"SyncRequestNotificationBeeperItemContents":{"__m":"SyncRequestNotificationBeeperItemContents.react
d":{"nodes":[],"servertime":1431038511}},408]],"require":[["MusicButtonManager","init",[],[["music.song
"initLiveMessageReceiver"],["Dock","init",["m_0_4b"],[{"__m":"m_0_4b"}]],["ChatApp","init",["m_0_4c","m
l_data"}]],["React","constructAndRenderComponent",["NotificationBeeper.react","m_0_4e"],[{"__m":"Notifi
rsrc.php\/yy\/r\/odIeERVR1c5.mp3","soundEnabled":true,"tracking":"
"type\":\"click2canvas\",\"fbsource\":\"1001\"}"},{"__m":"m_0_4e"}]],["ChatOptions"],["ShortProfiles","
":"Mohamad Hajjar","firstName":"Mohamad","vanity":"mohamad.hajjar.545","thumbSrc":"https:\/\/fbcdn-prof
05_163929713673106_1215456_n.jpg?
_gda__=1438987852_38f76a327d5874b6de856a8cd04f0fdb","uri":"https:\/\/www.facebook.com\/mohamad.hajjar.5
cLarge":null,"dir":null,"searchTokens":["Hajjar","Mohamad"],"alternateName":""},"100006453834462":
stName":"Mhamad","vanity":"mhamad.saab.33","thumbSrc":"https:\/\/fbcdn-profile-a.akamaihd.net\/hprofile
9772868_n.jpg?
_gda__=1440654864_38cadf7260674569fa0848d2d53a8a4f","uri":"https:\/\/www.facebook.com\/mhamad.saab.33",
ge":null,"dir":null,"searchTokens":["Saab","Mhamad"],"alternateName":""},"100003584830371":{"id":"10000
yad","thumbSrc":"https:\/\/fbcdn-profile-a.akamaihd.net\/hprofile-ak-xpf1\/v\/t1.0-
639190_n.jpg?
```

Figure 3. Facebook HTML generated code

In general, most Facebook data is very important, since it allows getting information about someone (e.g., who likes swimming or whom wearing jeans), collecting data for investigations purposes (e.g., crimes, retrace criminals), and for statistics and analysis purpose (e.g., number of likers for a specific event). PHP (Hypertext Preprocessor) presents the main programming language of the Facebook frontend, which is suited for web development and can be embedded into HTML. PHP is an open source, support object-oriented, powerful built in functions. PHP can works and connects with several databases, such as MySQL (My Structured Query Language), Oracle, etc., and it can manipulate XML (Extensible Markup Language) documents. For that, we can use the PHP programming language for the data crawling and mining reasons (see Figure 3).

B. Architecture of Facebook Crawling and Mining system

The login page presents the entrance into Facebook according to a specific user account (username or email, and password); if this user is authenticated, then we can directly access the full information about it, such as news feed, group information, friends, photos, events, comments, etc. All this information is presented in several HTML generated pages; knowing that, the HTML code is generated from many others programming languages. Inside this HTML code, we have all the data viewed on the web browser, so, we can crawl and extract the data from the opened Facebook user account page. This data may be the list of friends, information about each friends, events and notifications.

The data crawling procedure will automatically occur according to an automated system. Our developed automated system allows takes taking as an input a Facebook user, fetching the Facebook HTML pages, splitting the HTML code according to specifics delimiters (rules) into a more manageable portion, removing the unwanted HTML tags, reformatting HTML, adjusting spaces, removing entities, matching content with regular expressions, and storing the pertinent content into a structured MySQL database for future data mining use. The system database structure and algorithm will describe in detail below.

1) Database design

The parsing presents a main step in the data crawling process; it is based on specific characters and symbols that must be defined according to the text to be analyze. In the Facebook crawling, the parsing process must split the HTML code according to specific tags that can be used later as rules. For that, we save these rules in the database (in the "rules" table). Later, for each parsing process, we will select the corresponding rules from the database.

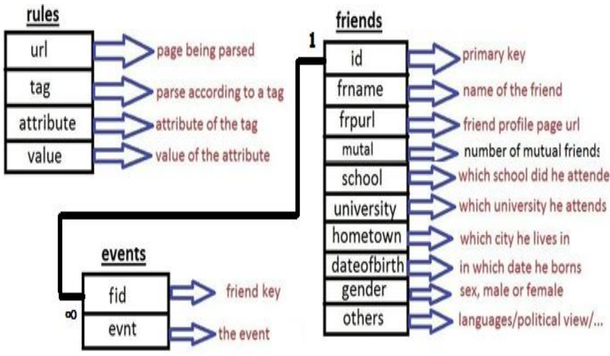


Figure 4. Database diagram

For a specific Facebook account, we can recursively extract a list of friends, and information about these friends (friend name, URL of the friend profile page, number of mutual friend according to the authenticated account, school, university, hometown, dateofbirth, gender, others). For each extracted friend, we can extract their events which saved on the table events (Figure 4).

2) Algorithms

In this section, we will present an algorithm that describes several operations on Facebook crawling and mining. Firstly, the operation allows extracting recursively the list of friends for a specific Facebook user, e.g., for a given Facebook user, we will extract the list of their friends L1, and, for each one in the list L1, we will extract their friends, and so on. Also, in this section, we will present the other operations that allow extracting other information, such as: events, comments, etc.

Figure 5 shows the architecture of the system algorithm for Facebook data crawling automatically starting from the home page of a specific Facebook authenticated user.

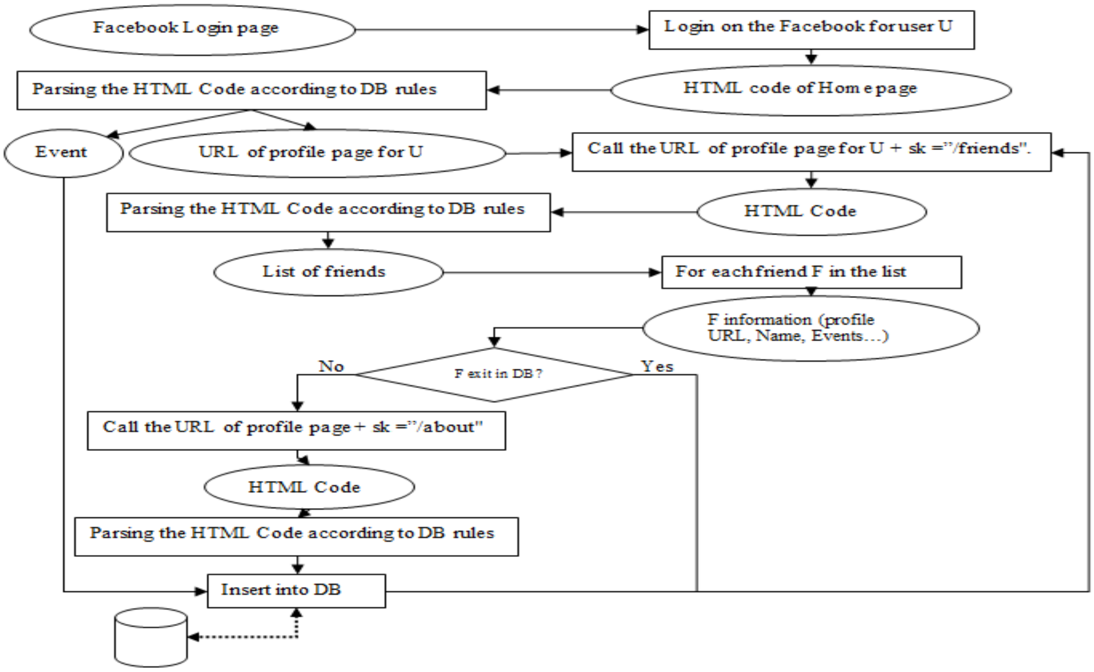

Facebook Login page
Login on the Facebook for user U
Parsing the HTML Code according to DB rules
HTML code of Home page
Event
URL of profile page for U
Call the URL of profile page for U + sk ="/friends".
Parsing the HTML Code according to DB rules
HTML Code
List of friends
For each friend F in the list
F information (profile URL, Name, Events…)
No
F exit in DB ?
Yes
Call the URL of profile page + sk ="/about"
HTML Code
Parsing the HTML Code according to DB rules
Insert into DB

Figure 5. General architecture of Facebook Crawling and Mining system.

This algorithm is composed of several operations, as follow: firstly, we get the HTML of the home page for a specific Facebook user U and parse this page according to the database rules; the objective of parser is to transform an HTML code into data and information. Parsing an HTML code is done in several stages. Firstly, we determine the main delimiters that allow extracting some information; these delimiters are HTML tags. We determine these delimiters by navigation and analysis of the Facebook HTML code, then extract manually these delimiters, and save them as rules in the database; for example in the Facebook HTML code: friends name exist between the two tags <Friendtag1> FriendName </ Friendtag1>, then these 2 tags (<Friendtag1> and </Friendtag1>) must be saved in the table rules. Then, the table rules present the reference of parsing stage in the Facebook crawling system. For example, to extract friend from a Facebook HTML code, firstly, we must get the specific rules of friends from the table rules in our database. The parsing stage allows extracting all the events and the URL of this user profile page, the list of all friends for the user U are presented in a page that have as URL the concatenation between the URL of the user U profile page and "/friends". From this new URL, we can extract, after parsing, the list of all friends. For each friend in list, we can follow the same procedure to getting their information, such as friend name, friend events, friend profile URL page, etc., according to their information (name, date of birth, etc.). We can test whether this friend is already added into DB in the table friends; if it is new, it must go to the About page, the About page URL can be done by concatenating the friend profile URL and "\About"; then, we can extract other information, such as school, university, etc. Finally, all the extracted information (friend profile page URL, friend name, friend events, etc.) will save in the database (in the 2 tables friends and events). Otherwise, if this friend already exists in the database; then, we must get the URL of their profile page, and we will follow the same procedure to get their friends; in this case, we extract all the friends of friends for the user U, and so on.

V. RESULT

The main result of this work is an automated Facebook crawling and mining system. This system take as input a Facebook user and give as output big amount of information about this user such as friends, comments, likers, etc. It is based on set of rules saved manually in the database. This system allows extracting recursively the list of friends for a specific Facebook user, and storing the entire extracted friends' information into a MySQL database for future use. The start point of this system is the home page for a specific Facebook authenticated user. After that, we apply the system operations (see Figure 5) in order to collect the pertinent information for all friends, friends of friends, friends of friends of friends, etc., of the logged user. We evaluate the system on several Facebook users, for each one it returned the demanded information. For that, we will describe the result as an example of the user "Joe", in order to explain that the result of our work is a system that return the pertinent information about a given. Let us consider that we have a Facebook user named

"Joe@hotmail.com", presented as input to our system. The system takes the HTML code of the home page, and selects from the database the rules (or tags). These rules are already saved manually in the rules table. Next, the system allows extracting the profile page link for Joe, according to the selected rules (for example :< a class="_2dpe _1ayn">, <title="Profile">). Then, we apply the parsing operation according to the extracted tags in order to deliver the profile page link "https://www.facebook.com/profile.php?id=1013"; from this URL, we can explore the HTML code in order to extract the friends list, by applying the parsing operation according to the stored rules of friends list. For each friend F in the extracted list, we can apply the same procedure to extract its profile page (profile page of friend F). The profile page includes pertinent information such as friends, photos, posts, etc.

We have several purposes for gathering the information from the system, which focused on extracting many facts about a person. For example, we can view that (Joe) is not friend with (George), but he is related to him on FaceBook because (Joe) is friend with (Mario), (Mario) is friend with (Elie), and (Elie) is friend with (George). Also, we can analyze the stored data and conclude that (Joe)'s hobby is politics, because the collected data describe that (Joe) likes several politic pages, and he likes the politic events.

VI. CONCLUSION AND FUTURE WORK

Social media is becoming an integral part of online life, as social websites and applications proliferate. Most traditional online media include social components, such as comment fields for users. In business, social media is used to market products, promote brands, and connect to current customers and foster new business.

Social media analytics is the practice of gathering data from social media websites and analyzing that data to make business decisions. The most common use of social media analytics is to mine customer sentiment to support marketing and customer service activities.

In this paper, we are interested to study the problem of gathering information from Facebook pages and storing them in a database. More specifically, one gathers the basic information, such as home town, age, work, education, friends, events and much other information. First, we describe the content of Facebook web pages and the principle of web crawling and web data mining. Second, we proposed the architecture of our automated system that allows crawling and mining the Facebook information for a specific user logged in.

In future works, we plan to study the problem of web crawling and web data mining in the cases of other social media websites like YouTube, Instagram, etc. We also plan to study the way in which we use gathering data in the database.

ACKNOWLEDGMENT

This work has been done as a part of the following project: "Automatic information extraction form Arabic texts" supported by the Lebanese National Centre of Scientific Research (CNRSL).

REFERENCES

[1] A. Kaplan and M. Haenle, "Users of the world, unite! The challenges and opportunities of social media", Business Horizons, vol.53 (1), pp. 61, 2010.

[2] J. Kietzmann, K. Hermkens, I. McCarthy, and B. Silvestre, "Social media? Get serious! Understanding the functional building blocks of social media", Business Horizons, vol. 54, no. 3, 2011.

[3] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. "Finding high-quality content in social media", WISDOM – Proceedings of the 2008 International Conference on Web Search and Data Mining: 183–193, 2008.

[4] Nielsen Holdings,"State of the media: The social media report 2012", Featured Insights, Global, Media + Entertainment. Nielsen. Retrieved 9 December 2012.

[5] Q. Tang, B. Gu, Bin, and A. Whinston, "Content Contribution for Revenue Sharing and Reputation in Social Media: A Dynamic Structural Model", Journal of Management Information Systems, vol. 29, no. 2, pp. 41-76, Fall 2012 .

[6] Nielsen Holdings, "The U.S. Digital Consumer Report". Featured Insights, Global, Media + Entertainment. Nielsen [Retrieved November 25, 2014].

[7] N. Carlson,"At Last – The Full Story Of How Facebook Was Founded", Business Insider, http://www.businessinsider.com/, March 5, 2015.

[8] R. E. Cash, "Depression In Young Children: Information For Parents And Educators". Facebook Retrieved, Social/Emotional Development, November 22, 2011.

[9] E. Eldon, "2008 Growth Puts Facebook In Better Position to Make Money", VentureBeat(San Francisco). [Retrieved December 19, 2008].

[10] M. Zuckerberg, "Company Info | Facebook Newsroom", Facebook newsroom, http://newsroom.fb.com/company-info/. September 30, 2014.

[11] J. Wu, P. Teregowda, M. Khabsa, S. Carman, D. Jordan, J. Wandelmer, X. Lu, P. Mitra, and C. Giles "Web crawler middleware for search engine digital libraries: a case study for citeseerX", In proceedings of the twelfth international workshop on Web information and data management pp. 57-64, Maui Hawaii, USA, November 2012.

[12] Y. Sun, "A comprehensive study of the regulation and behavior of web crawlers". A comprehensive study of the regulation and behavior of web crawlers, Publisher: Pennsylvania State University, 2008.

[13] OutWit Technologies, "OutWit Hub - Find, grab and organize all kinds of data and media from online sources". Outwit.com. 2014-01-31. [Retrieved March 20, 2014].

[14] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases". American Association for Artificial Intelligence, fall 1996.

[15] S. Chakrabarti, "Data Mining Curriculum: A Proposal (Version 0.91)". Group of ACM SIGKDD Curriculum Committee, August 5, 2004.

[16] K. Wahlstrom, J. Roddick, R. Sarre, V. Estivill-Castro, and D. de Vries, "Legal and Technical Issues of Privacy Preservation in Data Mining". Legal and Technical Issues of Privacy Preservation in Data Mining, 2007.

[17] D. Boyd and N. Ellison, "Social Network Sites: Definition, History, and Scholarship". Journal of Computer-Mediated Communication, vol.13, pp. 210–230, 2008.

[18] M. Moody, "Teaching Twitter and Beyond: Tip for Incorporating Social Media in Traditional Courses". Journal of Magazine & New Media Research, vol. 11(2): pp. 1-9, 2010.