RUM Extractor: A Facebook Extractor for Data Analysis

Rehab M. Duwairi
Department of Computer Information Systems
Jordan University of Science and Technology
Irbid 22110, Jordan
rehab@just.edu.jo

Mosab AlFaqeeh
Department of Computer Science
Jordan University of Science and Technology
Irbid 22110, Jordan
Mos3b.faqeeh@gmail.com

***Abstract*— Social Network Analysis (SNA) is a field of study that focuses on analyzing user profiles and participations on social network channels in order to model relationships between people and to predict certain behaviors or knowledge. To achieve their goals, researchers, interested in SNA, have to extract content and structure from the numerous social networks available today. Existing tools, which help in this task, often require substantial pre-processing or good programming skills which may not be available for all SNA researchers. This paper describes RUM, a data extraction tool which allows researchers to easily extract several types of content and structure that are available on Facebook pages. Consequently, the extracted data can be saved and analyzed. RUM Extractor is easy to set up and use, and it gives flexible options to users to specify the type and amount of content and structure they want to retrieve. The paper also demonstrates how RUM can be exploited by collecting and further analyzing data collected from two popular Arabic news pages.**

Keywords*— *Social Network Analysis, Data Extraction, Crawlers, Facebook Posts, Data Analysis

I. INTRODUCTION

The increasing popularity of Social Networks platforms is witnessed by the huge number of people that have accounts on Twitter, Facebook, and so on. Social network platforms became the tools of choice for connecting people. Sociologists anticipate that the structure of social networks will mirror real-life society and relationships [17]. For example, Facebook has an estimated 13 million transactions per second at peak time. Also, Facebook as of October 2015 has 1.44 billion monthly active users [7, 13]. This makes Facebook a favorable target to researchers who are interested in SNA [11]. Obviously, Facebook research is challenging as it introduces several subtle issues which researchers should consider such as data collection, optimization, scalability, and robustness [13, 15]. This paper presents a software tool, called RUM[1], designed to help researchers to easily extract data, for analysis, from Facebook.

Data published on the web, via social network channels, provide excellent opportunities for researchers. These data can be exploited to analyze social and cultural aspects of society. This line of research, which relies on digital data and on computer analysis, has several advantages over traditional field study approaches. The former approach is superior in terms of cost, speed, detail, comprehensiveness and size [5, 10].

RUM Extractor is a tool which facilitates data extraction from social networks; Facebook in particular. RUM is capable of collecting several types of data which is available on Facebook pages such as posts, comments on posts, shares of posts and so on. All of this is achieved without requiring substantial programming skills from users. RUM also respects privacy by not collecting any data related to user profiles. Only public posts and comments are collected. The details of RUM design and functionality are described in Section 3 of this paper.

This paper is organized as follows: Section 1 has provided an introduction about social network analysis. Section 2, on the other hand, provides a description to related work and places RUM in its proper place in the research community. Section 3 describes the architecture of RUM and explains its design decisions. Section 4 presents a case study where RUM was used to extract data from the Facebook pages of two popular Arabic news pages. Finally, Section 5 concludes this paper.

II. BACKGROUND AND RELATED WORK

The task of extracting data from Facebook and other social networks has captured the attention of many researchers [1, 3, 9, 18]. In this section, we review a number of related researches. We will present some methods of crawling large social networks, such as Facebook. Collected data are typically mapped onto graph data structures with the aim of analyzing their structural ingredients. The eventual objective of these efforts is possibly best described by the work reported by Kleinberg [12]. This study shows that node degree distribution works based on a power law, especially in social networks. That aspect points to the fact that mainly social network participants are regularly inactive. Little key users create large segments of data/traffic.

The massive widespread of Facebook has encouraged the appearance of a large number of analytics software for marketing purposes, which often focus on pages, groups even public figures profiles. As these tools are usually created for monitoring marketing campaigns, they target all available data.

[1] RUM is a well-known and beautiful valley in Jordan.

978-1-4673-8103-1/15 $31.00 © 2015 IEEE
DOI 10.1109/FiCloud.2015.116

There are many tools that function as common purpose data extractors for researchers and marketing purposes. One of these tools is the work reported in [6]. They created a plug-in for NodeXL network analysis and visualization tool which provides good functionality for downloading personal networks and, extracting data from Facebook pages, together with networks for users and posts.

The authors of [4] describe their work in the gathering and analysis of enormous data that describes connections among users of online social networks. They used ad-hoc crawlers which preserve privacy. Huge samples which consist of millions of connections have been collected. The data is anonymous and is organized as an undirected graph. They introduce a tool to analyze specific properties of social-network graphs.

In [2], the authors describe the Netvizz application. It is a standalone web application, unlike NodeXL, for extracting information from online social networks. As social network datasets are not publically accessible, often researchers interested in collecting data from social networks are forced to adopt extraction techniques typically used to crawl the web [8].

The authors of [14] investigated sampling techniques from large social graphs. The main goal of the research was to explore several graph mining algorithms to overcome bias in sampling. Their results state that the social graph preserves most of its properties when only a sample of size equals 15% of the graph is maintained.

The work reported in [16] describes in-depth analysis of the topological properties of social networks such as link symmetry and power node degrees on data crawled from Orkut, Flicker and Live Journal. The reported work also addresses challenges of large-scale data extraction from social networks.

Ye et al. [18] addressed crawling techniques of online social networks. They focused on quantitative aspects such as the efficiency of the visiting algorithms, and bias of data produced by different crawling algorithms.

The work of Gjoka et al [11] is similar to the work reported in this paper in the sense that is addressed data extraction from Facebook. The authors explored several crawling algorithms to sample and analyze Facebook friendship graphs. In particular, the BFS, Random Walk, and Metropolis-Hastings Random Walks algorithms were used in Gjoka's paper. Their main objective was to obtain consistent samples which preserve the properties of the whole graph. Their crawling strategy requires knowing the degree of nodes in advance so that nodes with high degrees are visited. In RUM there is no need to know any prior information about the visited nodes from the users' side. Such information can be easily extracted from the profile of Facebook pages.

III. METHODOLOGY

The RUM Extractor provides "raw" data for pages. It runs as a Web application. RUM does not require to be integrated within software such as NodeXL and therefore it is fast and usable with Facebook's data.

The Rum Extractor can collect several types of data which are typically available on Facebook pages such as posts, comments on posts, shares of posts, likes of posts and so on. This provides a multitude of possibilities for analyzing the crawled data. For example, one can identify the posts that possess the highest degree of engagement.

The data, collected by RUM, are provided in a data file as CSV, fully prepared for statistical analysis. To make content analysis easier, two files are maintained: one for the posts and one for the comments. RUM is flexible and allows its users to specify whether they want to retrieve posts made by users or the page owner.

A. Data Access via the Facebook API

A Facebook application is a program that directly integrates into Facebook platform and is often provided by a third party. Access through APIs makes use of Web 2.0 services provided to third-party developers. These APIs often have restrictions on the size of retrieved data and on the frequency of retrieval. Facebook is very restrictive with respect to these two aspects.

In order to do analysis on a Facebook page data, it is necessary to cache it temporarily on RUM servers. The data is acquired through the RUM Connection App using the Facebook API. RUM respects the privacy settings which the users have specified on their Facebook accounts. Our system is set up to cache user data on RUM servers for one hour, which allows users to download the data.

RUM is a simple Facebook application written in PHP which runs on a dedicated server. It is part of Facebook's app directory and can be accessed just by the admin until now. RUM requires users to log onto with an existing Facebook account to be able to access any data.

B. RUM Extractor Architecture

RUM is composed of the following elements:

1. A server running the crawling agents;
2. A cross-platform PHP application, which implements the logic of the agents;
3. An Apache interface that manages the data flow through the web. While running, the agents query the Facebook servers obtaining the pages ID's that have been submitted by the crawler. It is running as queue first input first output. Collected data are stored on the server.

Figure 1 shows the architecture of RUM. RUM is comprised of the following three components:

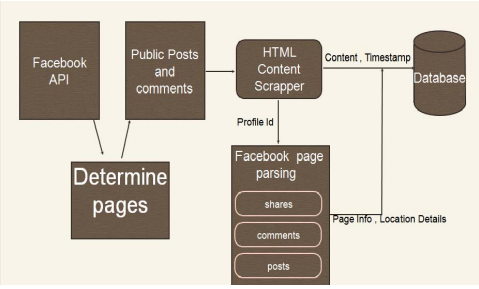

Fig. 1. The Architecture of RUM

1) Identifying the Crawler: Before using RUM, users have to log onto their Facebook accounts. Afterwards, the extractor can be invited by using either of the following two strings to allow smooth communication with Facebook API:

facebookexternalhit/1.1,

(+http://www.facebook.com/externalhit_uatext.php)

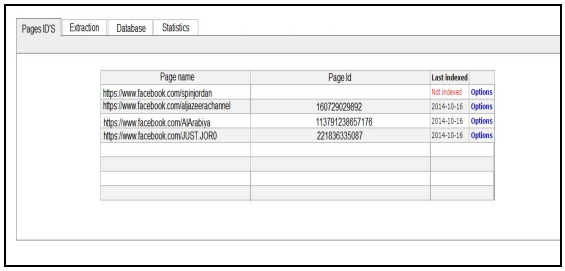

Fig. 2. A sample waiting list of pages to be processed

After the collection of posts of a given page is completed the second process starts. This second phase or process is concerned with collecting information about the posts. So, it starts by harvesting the post IDs and then for every post the following information will be extracted: total number of comments, total number of shares, total number of likes either of the post or of its comments, and text of each comment.

2) Content Scrapping: The user can choose from two alternatives when extracting data from Facebook pages: (1) To-Depth. This is an integer number which specifies depth that RUM must follow to collect the required posts which should be extracted. (2) To-Date which means retrieve all posts that fall within a given time interval. Figure 3 illustrates how the user can set these parameters.

3) Handling the Extracted Data : RUM does not use the extracted data for anything beyond providing the app functionality. No data other than the data aggregated in the output files is accessed. RUM does not store any user profiles. The extracted data, based on user queries, are deleted from RUM server at regular intervals without being submitted to any form of analysis or aggregation. No data is stored or shared with anybody.

Fig. 3. Content scrapping in RUM

C. *Characteristics of RUM*

RUM has many desirable features that makeit favorable by users seeking data posted on Facebook. The following points will briefly describe these characteristics.

- RUM has an easy to deal with GUI. RUM enables users to interact with system features like selecting the number of retrieved posts or comments and entering the pages' IDS to extract the data. RUM has consistent interface which allows users to master the roles of the several buttons, tabs, icons and other elements. It is fairly easy for users to customize RUM and add new features.
- RUM is flexible in the sense that it will allow users to extract data based on multiple factors. For example, it is possible to specify the Date-Time interval that extracted post/comments should achieve. Further, RUM allows its users to specify the depth or levels that the extractor should follow to extract the required information.
- RUM is cross-lingual. RUM can collect posts/comments/likes and so on regardless of the language.
- RUM supports user privacy. Online social networks, including Facebook, are recent additions, and therefore protocols of research ethics are not fully regulated. To make RUM comply with ethical standards, two decisions have been made: First, the user can only extract data from existing liked pages and cannot extract any data from other non-liked pages.Second, no anonymous users are allowed to use RUM – users of RUM must be registered Facebook users and the extracted data is first stored on RUM servers before being direct to users.

IV. CASE STUDY

In order to demonstrate how RUM can be used, we have collected data from two popular Arabic news sites, namely, Aljazeera and Alarabyia. The data were collected from their respective pages on Facebook. Posts and comments, which were posted on both pages from January 2014 to June 2014, were collected using RUM. Table 1 shows statistics about

collected posts on Aljazeera and Alarabyia pages. As it can be seen from Table 1, Alarabyia is more active in terms of the total number of posts, comments, shares and so on. Table 2, by comparison illustrates details of comments number of comments, number of people who commentated and number of replies. Again, Table 2 shows that visitors of Alarabyia page are more active that Aljazeera users for the same interval.

TABLE I. ALJAZEERA AND ALARABYIA STATISTICS ON POSTS

Statistics on Posts			
Post statistics	***Description***	***Total for Aljazeera***	***Total for Alarabyia***
type_post:	Refers to type of the post: text, image, audio or video	All types	All types
post_published:	Date of publishing the post. Can be used to determine number of posts published in a given interval.	7,951	19,462
likes_count_fb:	Likes to the post itself not to its comments.	9,211,579	17,237,549
comments_all:	All comments to a given post including replies	998,466	1,829,453
comments_base:	Comments without replies	96,329	436,396
comments_replies:	Replies to comments	88,462	160,469
comment_likes:	Likes on the comments	910,171	1,444,605
shares:	Shares of posts	358,889	196,670
engagement:	Total number of likes, comments and shares	32,979,736	17,237,549

TABLE II. ALJAZEERA AND ALARABYIA STATISTICS ON COMMENTS

Statistics on Comments			
Comments statistics	***Description***	***Total for Aljazeera***	***Total for Alarabyia***
post_id:	number of posts	7,951	19,462
comment_id:	number of comments	587,808	1,829,453
comment_by:	Number of people who commentated	359,917	1,922,425
is_reply:	how many replies on the comments	88,462	160,469
comment_like_count:	number of likes on the comments	910,171	1,444,605

V. CONCLUSIONS

This paper has introduced RUM Extractor. RUM is a dedicated data extractor which is specialized in extracting data from Facebook. These data include posts, comments, shares, and likes. RUM Extractor has an easy interface which allows users to extract the type of data they are interested in easily and efficiently. To prove the usability of RUM, the paper, also, has presented a case study of using RUM to collect data from the Facebook pages of two popular news agencies, namely, Aljazeera and Alarabyia. The collected data covers posts, comments, shares, replies and likes that were published on Aljazeera and Alarabyia Facebook pages from January 2014 to June 2014. Collected data show that Alarabyia publishes more posts when compared to Aljazeera. Also, the collected data shows that readers of Alarabyia page are more enthusiastic about interacting with the page by writing comments on posts and by sharing the posts among their networks. Of course, these finding are not conclusive to all intervals and not to all posts. The results shown in this study only address a six months interval and the judgment was based on the total number of posts, comments, shares and likes. In the future we plan to carry out more sophisticated analysis such as determine popular posts, active users, loyal users and so on.

REFERENCES

[1] Albert R., Barab´asi A.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**(1), 47–97 (2002).

[2] BernhardR..Studying Facebook via data extraction: the Netvizz application. Proceedings of the 5th Annual ACM Web Science Conference.ACM, 2013.

[3] Catanese S., De Meo P., Ferrara E., Fiumara G. Analyzing the Facebook friendship graph. In: Proceedings of the 1st International Workshop on Mining the Future Internet, pp. 14–19. Berlin, Germany (2010).

[4] Catanese S., De Meo P., Ferrara, E.Fiumara G., and ProvettiA..Crawling Facebook for social network analysis purposes.In Proceedings of the international conference on web intelligence, mining and semantics.ACM, 2011.

[5] Chau D., Pandit S., Wang S., Faloutsos C.: Parallel crawling for online social networks. In: Proceedings of the 16th International Conference on the World Wide Web, pp. 1283–1284. Banff, AB, Canada (2007).

[6] Derek H., Shneiderman B., and Smith M. Analyzing social media networks with NodeXL: Insights from a connected world. Morgan Kaufmann, 2010.

[7] Facebook Key Facts.http://newsroom.fb.com/Key-Facts.

[8] Ferrara, E., Fiumara, G., Baumgartner, R. Web data extraction, applications and techniques: a survey. Technical Report (2010).

[9] Garton L., Haythornthwaite C., Wellman B. Studying online social networks. J. Comput. Med. Commun. **3**(1), 75–105 (1997).

[10] Gjoka M., Kurant M., Butts C., and Markopoulou A. Practical recommendations on crawling online social networks. Selected Areas in Communications, IEEE Journal on, 29(9), (2011): 1872-1892.

[11] Gjoka M., Kurant M., Butts C., Markopoulou A. Walking in Facebook: a case study of unbiased sampling of OSNs. In: Proceedings of the 29th Conference on Information Communications, pp. 2498–2506. IEEE, San Diego, CA, USA (2010).

[12] Kleinberg, J. The small-world phenomenon: an algorithm perspective. In: Proceedings of the 32nd Symposium on Theory of Computing, pp. 163–170. ACM, Portland, OR, USA (2000).

[13] Kross E., Verduyn P., Demiralp E., Park J., Lee D.S., Lin N, Shablack H., Jonides J., Ybarra O. Facebook use predicts declines in subjective well-being in young adults. PLoS One.8(8).doi: 10.1371/journal.pone.0069841, 2013.

[14] Leskovec J., Faloutsos C. Sampling from large graphs. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 631–636. Philadelphia, PA, USA (2006).

[15] McAndrew F. T., Jeong H. S. Who does what on Facebook? Age, sex, and relationship status as predictors of Facebook use. Computers in Human Behavior. 28(6), (2012): 2359-2365.

[16] Mislove A., Marcon M., Gummadi K., Druschel P., Bhattacharjee B. Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, pp. 29–42. ACM, San Diego, CA, USA (2007).

[17] Wilson R, Gosling S., and Graham L. A review of Facebook research in the social sciences. Perspectives on Psychological Science. 7(3) (2012): 203-220.

[18] Ye S., Lang J., Wu F. Crawling online social graphs. In: Proceedings of the 12th International Asia-Pacific Web Conference, pp. 236–242. IEEE, Busan, Korea (2010)