Towards an algorithm for efficient use of social network resources by using web scraping techniques

Valon Xhafa, Blerim Rexha and Korab Rrmoku

***Abstract*—**The rapid growth of Web in recent years has changed everyday life. New applications for sharing social activities on the Web, such us Facebook, Twitter, Google+ etc. are becoming more and more popular. The interactivity and behavior among these users in Web is measured by number of likes, comments, posts and shares and their personal perception. But these new social applications have their limits and constraints regarding the usage of their networks resources. This paper presents an algorithm for efficient use of these resources and overcoming these limits. The algorithm is set as outcome of the survey conducted among community members. For better performance, the combined approach using the Web scraping techniques and reverse image search was used.

***Keywords*—**Social network, algorithm, efficiency, Web scraping, privacy, security.

I. Introduction

FOUNDERS of today's Internet, back in '70s, could have hardly imagined the huge impact their four node network would have after many decades of development. Today, this communication network, the Internet, consists of hundreds of millions of connected computing devices, communication links, and packet switching devices used by billions of users all over the world [1]. Everyone and everything tends to be connected, and the trend shows its tendency even to increase more. According to Internet World Stats Internet usage growths for period 2000-2014 was 741% and Internet penetration in World level is about 42% [2].

Main merit of this huge wide spreading and usage of Internet is the introduction in mid '90s of the World Wide Web that uses the Hypertext Transfer Protocol (HTTP), as defined in [3], which is a stateless application protocol for distributed and collaborative information systems. HTTP is implemented at client and server side program. The client sends a HTTP request and server program answers with HTTP response. Client and server are executed on different hosts and talk to each other by exchanging HTTP messages. These HTTP messages are displayed at the client side using the Hyper Text Markup Language (HTML). The HTML uses the predefined tags to present different types of objects in html page such as text, images, audio and video etc., as presented in Fig. 1. The data exchanged between client and servers are plain text, i.e. not encrypted as well as data in HTML code are readable text, as presented in Fig. 1.

Fig. 1: (a) HTML code in Notepad (b) HTML view in browser

Therefore, it would be easy for one to extract these data from Web page, or even more to build an automated intelligence about the collected data. Online price comparison uses these Web scraping techniques.

The Web application changed, and continues to change, everything. Furthermore, new applications are coming everyday on top of this infrastructure. These applications tend to improve our daily life and make our way of doing business easier. Recently, in past ten years, Online Social Networks (OSN) are playing a huge role not only in social field but also in science and economy of each country. The most popular OSN-s, based on [4], are presented in Fig. 2.

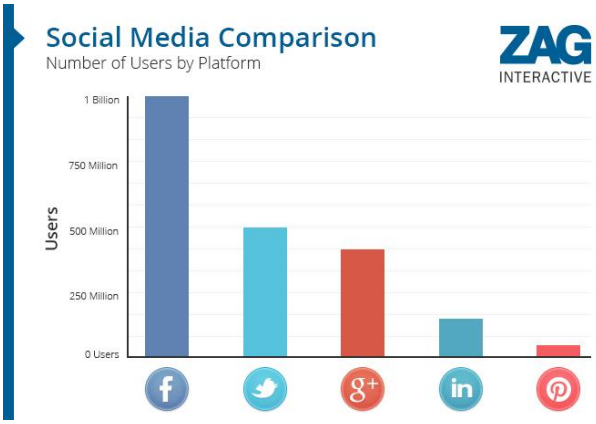


Fig. 2: Most popular OSN [4]

Valon Xhafa is student at Faculty of Electrical and Computer Engineering, University of Prishtina, 10000 Prishtina, Kosovo (e-mail: valon.xhafa@uni-pr.edu).

Korab Rrmoku is software and database developer and external associate as teaching assistant at Faculty of Electrical and Computer Engineering, University of Prishtina, 10000 Prishtina, Kosovo (e-mail: korab.rrmoku@uni-pr.edu).

Blerim Rexha is professor and head of Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Prishtina, 10000 Prishtina, Kosovo (e-mail: blerim.rexha@uni-pr.edu).

All these OSN-s have their limits set to user such as the maximum number of users (friends), and this paper presents an efficient algorithm for using these social network resources using Web scraping techniques and user's behavior in OSN-s. The outcome of this algorithm is the recommendation list to OSN-s user for efficient usage of network resources. Facebook is chosen as main representative OSN for our research.

II. Social Networks

Nowadays, OSN-s connects together people who share almost their entire everyday life. Social networks such as Facebook, Twitter, LinkedIn and Google+ have attracted millions of users last years. Facebook as the largest OSN has limited the number of friends per users. This means that a user can have only up to 5,000 friends. Users, on the other hand, tend to have more active friends (AF). An active friend is a user who has often activity with other users on the OSN. In this case, it came up that it is necessary to have an efficient use of OSN-s resources. As popularity is increasing and due to their open nature OSN-s suffer from abuse in the form of the creation of fake accounts, which do not correspond to real humans. This means that they are particularly vulnerable to the Sybil attacks [5], under which a malicious user can create multiple fake OSN accounts. Bilge et al. [6] investigate the easiness that attackers possess to deploy automated crawling techniques against users in popular OSN-s, in order to gain access to large volumes of personal information. In the first approach, authors' present an automated identity theft technique for enabling the attacker to access the sensitive personal information. The second approach is a more advanced one, and it is based on cross-site profile cloning attack, where a forged profile, in a network where the victim is not registered yet, is created, and afterwards the victim's friends that are registered on both networks are contacted. The actual experiments showed that these techniques are effective and feasible in practice. Gao et al. [7] presents a study about quantifying and characterizing spam campaigns that are launched using user accounts on OSN-s. This study is performed in a large anonymous dataset from Facebook, which contains "wall" messages that are exchanged between users. The authors employ combined techniques to detect correlation between wall messages and to identify the spread of potentially malicious content. These techniques group together the wall posts that share either the same (possibly obscured) destination, or have strong textual similarity. Cao et al. [7] introduce a tool for OSN systems, called SybilRank that ranks user accounts according to their perceived likelihood of being fake. This tool is based on social graph properties and it is computationally efficient by being able to handle graphs with hundreds of millions of nodes. The tool was implemented by using Hadoop and MapReduce parallel frameworks, while its overall complexity is of range O(nlogn). The SybilRank was tested against a real social network system, namely Tuenti, where the actual results showed an accuracy of about 90% of indicating the likelihood of fake user accounts. Krombholz et al. [8] have created fake Facebook profiles and integrated them into existing friendship networks to simulate a data harvesting attack. They analyzed the Facebook user data of profiles that interacted with created fake profiles and human factors to get a deeper understanding of the procedure of successfully integrating a fake profile into an existing friendship network.

A. Social Network Analysis

Having been developed in such a level, Social Networks have raised interest, so in recent years, but especially during last decade, there was a growing interest of the public to understand better "the connection" between modern societies. In the center of this phenomenon is the Network itself – which in reality is a pattern of interconnection between the different sets of things. Social networks that we face today are in fact a collection of social ties between friends, and these connections kept rising during all human history, thus arriving today in a complex condition. This expansion accompanied with complexity is as a result of technology progress and development; hence it has "reduced" traveling distances, global communication and digital relationship between people. Researchers have explored three main properties that we might see as in common within different networks, which are: (i) property of a "small" World, (ii) distribution scale of the networks and (iii) transitive states of the networks [9]. The property of the "small" World represents the shortest possible average distance between connections within one network. The property of distribution scale of a node in a network is a number of connections that one node has with other nodes on the network. Lastly, the property of "transitive" state of the network summarizes two connections that are neighbors to the third connection, which automatically increases the probability of these two nodes to be neighbors with each-other, as well. As an example this property implies the fact that two of your friends have larger probability to know each other, because you are the mutual friend. In order to evaluate the relation between actors on a network, the theory of SNA gives us a set of metrics that we can use, according to the domain and depth of our analysis. As defined in [10], following are some main SNA metrics and their categorization: (i) Global graph metrics: seek to describe the characteristic of a social network as a whole, for example the graphs diameter, mean node distance, the number of components (fully connected subgraphs), cliques, clusters, small-worldness, etc. (ii) Individual actor properties: relate to the analysis of the individual properties of network actors, e.g. actor status as central (degree, closeness, or betweenness centrality) or authoritative (eigenvector, PageRank, SALSA, HITS, or weighted HITS), distance, and position in a cluster.

III. New Algorithm for efficient use of OSN resources

The Facebook, as the most popular OSN, as presented in Fig. 2 allows unlimited number of followers; however it has a hard constrain of 5.000 friends per users. In a real world not all of these friends are active, and even worse in virtual Facebook life, some of them might close their Facebook accounts or unfriend many disliked users. In order to have the most accurate algorithm for managing this constraint, we have decided to ask the Facebook community at our University

about their evaluation of Facebook friendship. Therefore, an online questionnaire was developed. We sent the questionnaire link through email, and we put the link in our Facebook wall. There were 867 respondents, ready to share their Facebook habits.

A. *The Questionnaire*

In order to estimate the real impact of Facebook activities such as: (i) Comment, (ii) Like, (iii) Tag, and (iv) Share, we decided to ask the community through an online questionnaire. The questionnaire was anonymous, but we asked our respondents about their: (i) age, (ii) gender, (iii) time spent on Facebook and (iv) priority of four above mentioned Facebook activities. We received a positive feedback from University community, students and teaching staff, which confirms that Facebook is very popular, especially among young students, as presented in Table I.

Table I – Age group of respondents

Age group	# of respondents	%
Till 18	27	3%
19 - 30	611	70%
31 – 45	169	11%
Above 45	60	7%

The questionnaire response about the priority of activities on Facebook is presented in Fig. 3.
Based on respondent answers we observed that on Facebook, a comment on a user post is highly rated among online community. The overall responses are as follows: Comment has the most value to 37% of respondents, Like has 22%, Tag has 19% and Share just 16%, as presented in Fig. 3. Only 6% of respondents could not decide which of these activities is more important to them.

B. *The Algorithm*

In order to evaluate the importance of user's Facebook friends we implemented the following algorithm: (i) we scraped the number of friends from Facebook profile, and (ii) from every friend we calculated the total number of Likes, Comments, Tags and Shares.

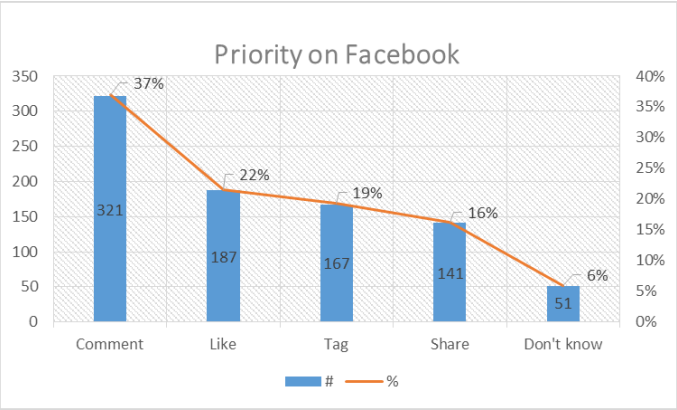


Fig. 3: Priority on Facebook

We multiply these numbers with their respective weights, gained from online questionnaire, as presented in Equation (1).

$$W = N_{Comment} * C_w + N_{Like} * L_w + N_{Tag} * T_w + N_{Share} * S_w \quad (1)$$

Where by: $N_{Comment}$ is the total number of "Comments" on Facebook profile, whereas C_w – is weight of "Comment" activity on Facebook, and based on data presented in Fig. 3, is 37%, thus C_w =0.37, and the same rule applies for "Like", "Tag", and "Share" activity. The sum of all these factors, as presented in Equitation (1), we have denoted as "*weight of relationship"*.
It is obvious that if one makes more comments, he will be more highly rated in friendship list, which could be translated as "close friend" in real world!

IV. System Architecture

The architecture of the system consists of parts developed to gather specific data from specific locations as well as parts in which the algorithm is implemented for efficient use of social network resources. The result of the algorithm, as mentioned is section III, is the recommendation of a list of friends with less activity on Facebook. Fig. 5 shows the scheme of the architecture. In addition the crawler and the reverse image search as part of the architecture will be described.

A. *The Web Scraper*

A Web crawler is a program or automated script that browses the WWW in a methodical and automated manner [6]. A more recent variant of Web crawlers are Web scrapers, which aims at looking for certain information–such as prices of particular goods from various online stores—extracting, and aggregating it into new Web pages [11]. In particular, scrapers are focused on transforming unstructured data and save them in structured databases.

In our scenario, crawler component is responsible for crawling the target social networking site and collect information on users that have chosen to make their profiles public. Our target OSN is Facebook. In this case, to access accounts information, Facebook has created an API (Application Programming Interface) called Facebook Graph API. The Graph API is a way to get data in and out of Facebook's platform. It is a low-level HTTP-based API, that it can be used to query data, post new stories, manage ads, upload photos and a variety of other tasks that an app might need to do. For retrieving more information, not only for a specific account, but also information about specified account friends, it is necessary to use other ways of extracting data from Facebook instead of using Graph API which gives limited access to public data. For this reason, using Web scraping techniques was necessary. It is worth noting that Web scraping may be against the terms of use of some websites. For that, it should be clarified that using Web scraping techniques in our project is only for scientific and research purposes.

The interface of the application is presented in Fig. 4. By default, friend lists are public information on Facebook. For

each analyzed user, we recorded their Facebook user ID, Post ID, comment ID, Friends ID, names, dates and other information needed, especially the profile images of friends which are very important, because they will be used at the Reverse Image Search, a component of the system which finds the authenticity of the profile image. These information are stored in a database.

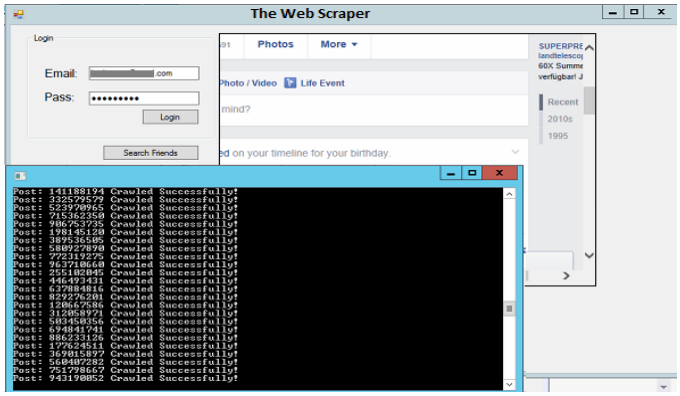


Fig. 4: Web scraping application interface

Taking into account that Graph API is developed specifically for Facebook use, the Web Scraping techniques can be used in other OSN-s, such as in Twitter, LinkedIn, etc.

B. *The reverse image search*

Reverse Image Search (RIS) is a technique that allows users to search for related images just by uploading an image; in terms of information retrieval, the sample image is what formulates a search query. Reverse image search allows users to discover content that is related to a specific sample image, popularity of an image, and discover manipulated versions and derivative works. In our system, the RIS component plays a major role in identifying Facebook user activity in other social networks and also gives the authenticity of an image used as a profile image by Facebook users. If a user has used an original profile image or got one from Internet, this footprint may lead us to find out if a specific user is an active in other OSN-s, too. In our system, an application written in C# uses Google's Search by image system to search for certain images. Google's Search by image is a feature that utilizes reverse image search and allows users to search for related images just by uploading an image or image URL [12].

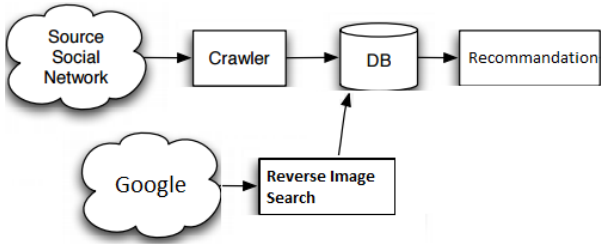


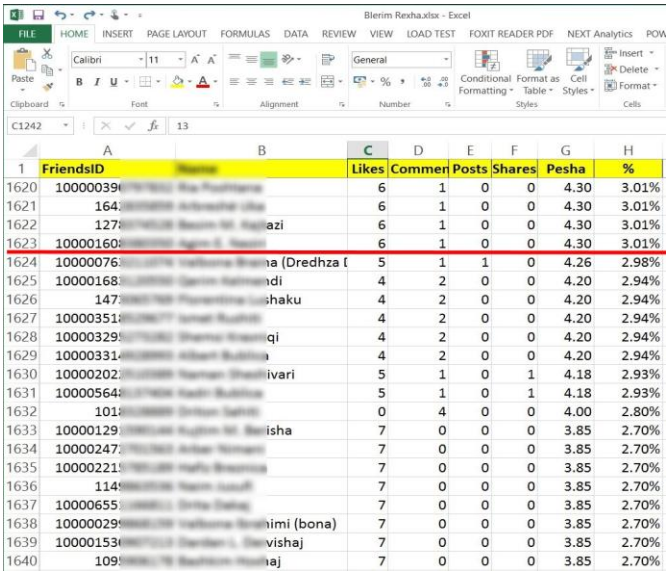
Fig. 5: System Architecture

Google accomplishes this by analyzing the submitted picture and constructing a mathematical model of it using advanced algorithms. It is then compared with billions of other images in Google's databases before returning matching and similar results. It should be noted that when available, Google also uses meta-data about the image such as description. The application mentioned above, for reach user, gets the name and profile image from Database and searches in Google for that image and for that name. In the end, the application gives for every user a result which tells that the user profile image is original was taken from the Internet or that is an original image of that user but found on the Internet.

V. Results

To test the algorithm, we have chosen a number of 53 Facebook users, who have accepted that their profile data's to be a part of the research. Once everyone is logged in application, the process of Web Scraping started crawling data from their profiles. After this step was completed, and the systematic storage of crawled data were saved in the database, then over the following data, the "*weight of relationship"* algorithm is implemented, which as described, is based on the weights obtained by the questionnaire.

A corresponding table in the database is generated for each participating user. Fig. 6 depicts the table containing the list of friends.



	A	B	C	D	E	F	G	H
1	FriendsID		Likes	Commen	Posts	Shares	Pesha	%
1620	10000039		6	1	0	0	4.30	3.01%
1621	164		6	1	0	0	4.30	3.01%
1622	127	azi	6	1	0	0	4.30	3.01%
1623	10000160		6	1	0	0	4.30	3.01%
1624	10000076	a (Dredhza [	5	1	1	0	4.26	2.98%
1625	10000168	ndi	4	2	0	0	4.20	2.94%
1626	147	shaku	4	2	0	0	4.20	2.94%
1627	10000351		4	2	0	0	4.20	2.94%
1628	10000329	qi	4	2	0	0	4.20	2.94%
1629	10000331		4	2	0	0	4.20	2.94%
1630	10000202	ivari	5	1	0	1	4.18	2.93%
1631	10000564		5	1	0	1	4.18	2.93%
1632	101		0	4	0	0	4.00	2.80%
1633	10000129	isha	7	0	0	0	3.85	2.70%
1634	10000247		7	0	0	0	3.85	2.70%
1635	10000221		7	0	0	0	3.85	2.70%
1636	114		7	0	0	0	3.85	2.70%
1637	10000655		7	0	0	0	3.85	2.70%
1638	10000029	imi (bona)	7	0	0	0	3.85	2.70%
1639	10000153	vishaj	7	0	0	0	3.85	2.70%
1640	109	aj	7	0	0	0	3.85	2.70%

Fig. 6: Results of the Algorithm and the recommended list under the threshold

The "*weight of relationship"* is calculated over these data, as presented in the Equation (1). Now, friends of certain user are ranked based on the "*weight of relationship"*. As mentioned in section of Social Networks, a Facebook user can have at most 5.000 friends, a relatively small number given of social networking profile. In this case, one of the main goals of the algorithm is to recommend a list of friends which should be replaced with new friends, as consequence of their relatively low activity.

Towards the recommendation, a threshold should be used over the already ranked list of friends based of their "*weight of relationship"*. Fig. 6 shows that in our case, we have decided as a threshold the value of 3%.

Friends who have a percentage of activity under this value, shown with red line as presented in Fig. 6, are recommended to be replaced with new friends who are waiting in line to be confirmed, but cannot be accepted due to constraints set by the OSN-s, as in our case, Facebook.

VI. Conclusion and Future Work

The right to manage the accounts is the right of every person involved in social networks; however this feature is not offered by OSN-s.

Hence, this paper's focus is in the development of an algorithm for efficiently use of online social network resources. Through the recommendation of a list of friends that weight less than a certain threshold, users can replace the friends in the recommended list with new friends that are waiting for a confirmation of their friend request.

In this paper, these opportunities to manage the accounts efficiently are provided only to certain users and only through a desktop based application, thus the development of an app for mobile device, which would be used by more people, remains a future work. Also, through the mobile app, we aim to invite more users to participate in research. The "*weight of relationship"* algorithm is mainly based on the weights derived from the questionnaire, so the evaluation of the results of the algorithm, also remains a future work. Another aspect which we plan to further expand our research consists in involving Social Network metrics as defined in section II, more specifically, in-degree centrality and authority centrality, in order to better evaluate the results.

References

[1] James. F. Kurosse. &. Keith. W. Ross, Computer Networking - A top down approach, New York: Pearson, 2012

[2] Internet World Stats available at http://www.internetworldstats.com/stats.htm, accessed October 2015

[3] Berners-Lee, Tim. "HyperText Transfer Protocol". World Wide Web Consortium, available at http://www.w3.org/Protocols/HTTP/AsImplemented.html, accessed October 2015

[4] Michelle Brown, "Social Media Comparison: How Your Company Can Benefit from Each Platform", available at: www.zaginteractive.com, accessed October 2015

[5] Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, Abraham Flaxman. "SybilGuard: defending against Sybil attacks via social networks." In Proceeding of SIGCOMM '06, pp. 267-278 ACM, 2006

[6] Bilge, Leyla, Thorsten Strufe, Davide Balzarotti, and Engin Kirda. "All your contacts are belong to us: automated identity theft attacks on social networks." In Proceedings of the 18th international conference on World Wide Web, pp. 551-560. ACM, 2009.

[7] Gao, Hongyu, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y. Zhao. "Detecting and characterizing social spam campaigns." In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, pp. 35-47. ACM, 2010.

[8] Katharina Krombholz, Dieter Merkl, Edgar Weippl. "Fake Identities in Social Media: A Case Study on the Sustainability of the Facebook Business Model." Journal of Service Science Research (2012) 4:175-212

[9] David Easley, Jon Kleinberg -Networks, Crowds, and Markets: Reasoning about a Highly Connected World, Cambridge University Press, 2010
Systems Design and Implementation, pp. 15-15. USENIX Association, 2012.

[10] Ahmedi Lule, Korab Rrmoku, and Kadri Sylejmani. "Tourist Tour Planning Supported by Social Network Analysis." International Conference on Social Informatics (SocialInformatics), IEEE, 2012.

[11] Cao, Qiang, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. "Aiding the detection of fake accounts in large scale social online services." InProceedings of the 9th USENIX conference on Networked

[12] Josip Karpac, Moray Allan, Jakob Verbeek, Frederic Jurie. "Improving Web image search results using query-relative classifiers." Computer Vision and Pattern Recognition (CVPR), IEEE Conference, 2010