A Clustering Approach for Evaluation of User Interaction on Facebook Social Network

Doaa Hassan
Computers and Systems Department
National Telecommunication Institute
Cairo, Egypt
Email: doaa@nti.sci.eg

***Abstract*—Social Networks analysis has been an important source of gathering information due to the large amount of data that can be generated from users' discussions and participation on social media. One way to analyze social networks is by estimating the amount of user interaction and participation in them. This paper addresses this issue by applying the machine learning clustering technique for categorizing users of Facebook social network based on their participation and interaction on Facebook. Two main features have been used for performing clustering: The first, is the number of links established between a user and others on Facebook through friendship relations. The second, is the number of posts written by a user to the walls of others on Facebook through posting. Therefore, the proposed approach in this paper aims to obtain different clusters of users that are categorized based on their level of interaction on Facebook. Hence we can estimate the amount of user interaction on Facebook by determining to which cluster he/she belongs.**

***Index Terms*—Facebook social networks, user interaction, clustering.**

I. Introduction

Currently, online social networks such as Facebook [14], LinkedIn [15], and Twitter [16] have become extremely a popular mean for interactions among billions of people all over the world, since those networks provide an easy and fast way for daily communication among them [7]. Due to this, the analysis of social network has been a great methodology for gathering and collecting information about people including, but not limited to their jobs, interests, hobbies and political directions.

Recently, mining social networks [13] has been a promising approach for discovering various interesting patterns that can be obtained by extensive social networks analysis. One of the research work done in that direction is using data mining and machine learning techniques to evaluate the user interaction and participation in social networks [5], [21], [22]. This paper extends this direction by presenting a clustering approach for evaluation of user interaction on Facebook social network. We have used clustering for categorizing users of Facebook based on their level of interaction and participation, where clustering groups users based on the features that describe their interaction (where users within the same group are similar to each other and different from those that fall in different cluster [10]). Two main features have been used for generating clusters of different interaction levels: The first, is the number of links that have been established between a user and others on Facebook through friendship relations. The second is the number of posts that have been written by a user to the walls of others on Facebook through posting. The generated clusters are obtained using two clustering techniques: the K-means clustering [11] and EM algorithm [12]. Both algorithms are common clustering techniques for grouping similar samples of dataset in one cluster that distinguishes them from other samples that fall in different clusters. The proposed approach has been applied to a dataset created by extracting the two aforementioned features from two real Facebook datasets. The former consists of a list of Facebook links, while the later consists of a list of Facebook wall posts. The choice of Facebook among other social network is due to its popularity, where users can communicate with each other by establishing a friendship relations, joining and creating groups, and assigning to many social events. Our experimental results show that a small number of users fall in the clusters of a high interaction levels in comparison to those that fall in the clusters of medium and low levels of interaction.

The structure of this paper is organized as follows: In Section 2, we provide an overview of K-means and EM clustering algorithms. In Section3, we present our proposed approach, the experimental settings, and the evaluation results. In Section 4, we discuss the related work. In Section 5, we conclude the paper with some directions for future work.

II. Background

A. K-means Clustering

K-means clustering is one of the most common unsupervised learning algorithms [19]. This algorithm works by initially specifying the number of clusters that are going to be generated. This is called the K parameter. Next, each cluster is assigned randomly a center point called centroid, which is the mean of the points in the cluster. Then all samples are assigned to the closest cluster center by measuring the ordinary Euclidean distance [17] between all samples in the dataset and each cluster center. After this step, the centroid or mean of the instances is computed for each cluster. Therefore, in the end, a number of means equal to the number of their respective clusters is obtained, which is called means. Next, the previous procedure is repeated with the new cluster centers until the same points are assigned to each cluster in successive rounds (i.e, the cluster centers remain the same without change).

B. The Expectation-Maximization Clustering Algorithm

EM is a probabilistic clustering mechanism that computes probabilities of cluster memberships [20]. In this algorithm, the procedure of K-means clustering is adopted and iterated. Initially, it starts with guesses for the parameters of the mixture, which is a set of K probability distributions, representing K clusters that govern the values of features for samples that are members of a certain cluster. Each of these distributions gives the probability that a particular instance would have a certain set of feature values, if it was known to be a member of a certain cluster. Next, the parameters of the mixture are used to calculate the cluster probabilities for each instance, where those probabilities are used for re-estimating the parameters of the mixture and then the procedure is repeated. The cluster probabilities are calculated first by expectation then by maximization of the likelihood of the distributions of the dataset, and hence algorithm is called Expectation-Maximization.

III. Proposed Approach and Experiment

The goal of the proposed approach is to categorize Facebook users into groups according to their level of interaction. Such a level is expressed in terms of the number of friendship links that each user has or the number of posts that he/she has written to the walls of other users. Therefore, the proposed approach uses the machine learning clustering technique for dividing users into groups when they interact on Facebook social network through friendship linking or posting, where each group has a specific interaction level.

The clustering is performed on a dataset that has been created by extracting the number of established friendship links and the number of posts for various users from two real Facebook datasets. Both datasets were presented in [3] and are available at [4]. The first dataset is the list of Facebook links that contains a list of all of user-to-user links from the Facebook New Orleans networks. All links are undirected on Facebook, but are treated as directed. Each dataset example has three features including two anonymized user identifiers, meaning the second user has a friendship link with the first one and a UNIX timestamp that refers to the time of link establishment. This dataset observes 3,646,662 friendship links between about 90,269 users. The second dataset is the list of Facebook wall posts that contains a list of all of the wall posts from the Facebook New Orleans networks. Each wall post entry in the data set contains information about the wall owner, the user who made the post, the time at which the post was made, and the post content. The wall post data observes 838,092 wall posts, for an average of 13.9 wall posts per user. This represents communication between 188,892 distinct pairs of users, which are 12.2% of the links in the New Orleans networks. The remaining 87.8% of the link in the social network did not show any wall activity. Both datasets have been anonymized to protect the privacy of the users themselves represented by the communication between users via friendship links and the wall posts features.

The proposed approach starts with preprocessing both datasets in order to extract the number of established links from the Facebook links dataset and the numbers of posts from the Facebook wall posts for each user. In the next step, the K-means and EM clustering algorithms are applied to the experimental dataset that consists of users records. Each record consists of the number of friendship links (link_count) and the number posts (post_count) for a certain user, that have been extracted from the previous step. The next two subsections provide more details about both steps.

We have run our approach on a windows laptop machine with 2.6 GHZ processor Intel core (TM)i5 and 4 G Memory Rams. We have used Weka [1], a free open source software machine learning tool for generating different clustering models from different categories, training them on the experimental dataset, and visualizing the generated clusters.

A. Dataset Preprocessing

The Facebook links and wall posts datasets have been processed in such a way that makes one single record for each user, whose attributes are all the remaining users. As for the Facebook links dataset, each attribute has a value of ”1” in case if one user has a link with other users and ”0” in case if he/she has not. For the wall posts dataset, each attribute has a value of ”1” in case if one user posts to the walls of other users and ”0” in case if he/she does not. For this reason, a program has been written in python [9] that preprocess the data of both datasets by translating it into this format. This format facilitates extracting the number of links that each user has with others and the number of posts that he/she has written to the walls of others. This is done simply by just counting the number of 1s for each user’s record in each case. Figure 1 shows a part of user’s record in this format that can be obtained after preprocessing each dataset.

In the end, we obtain a new dataset that has two columns, one that represents the number of friendship links that a user has established with others, and another one that represents the number of posts written by that user to the wall of others. Due to the memory limitations we have taken users from 0 to 10000 and examined the number of links that they have with other 1000 users (from 0 to 1000) and the number of posts that they have written to the walls of these users.

```
u1,u2,u3,u4,u5,u6,u7,u8,u9,u10,u11,u12,u13,u14,u15,u16,u17,u18
0,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,
```

Fig. 1. A part of user record obtained after preprocessing either Facebook links or wall posts datasets .

B. Generated Models and Results

We performed two experiments: In the first, we have run the K-means algorithm, while in the second one, we have run the EM algorithms. The number of clusters has been assigned to 5. This allows us to cluster users into five groups of different interaction level including: very low interactive users, low interactive users, medium interactive users, high interactive users and very high interactive users. Figure 2, 3 show the five clusters that have been generated using either K-means

or EM clustering algorithms respectively. As shown from the results, the following can be observed:

```
kMeans
======

Number of iterations: 20
Within cluster sum of squared errors: 15.482240225569806
Missing values globally replaced with mean/mode

Cluster centroids:
                          Cluster#
Attribute     Full Data          0          1          2          3          4
                 (9997)     (4971)     (3288)       (82)      (403)     (1253)
=============================================================================
post_count       0.7368     0.0117     0.5344    15.2195     4.3722     2.0279
link_count       3.0095     0.3408     2.6946    38.1341    17.9231      7.328



Time taken to build model (full training data) : 0.41 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      4971 ( 50%)
1      3288 ( 33%)
2        82 (  1%)
3       403 (  4%)
4      1253 ( 13%)
```

Fig. 2. The clusters generated automatically using K-means clustering algorithm with no of clusters=5.

```
EM
==

Number of clusters: 5


             Cluster
Attribute          0       1       2       3       4
                (0.61)  (0.07)  (0.3)     (0)  (0.02)
=========================================================
post_count
  mean             0  3.2021   1.035 21.6284  7.8059
  std. dev.   0.0063  2.2297  0.8447  9.0656  4.9654

link_count
  mean        0.7222 11.6848  3.8175 44.0927 26.7433
  std. dev.    0.907  4.8751  2.3179 16.4321 10.9759


Time taken to build model (full training data) : 29.2 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      6450 ( 65%)
1       763 (  8%)
2      2612 ( 26%)
3        24 (  0%)
4       148 (  1%)
```

Fig. 3. The clusters generated automatically using EM clustering algorithm with with no of clusters=5.

- For K-means clustering, we found that 4971 of dataset samples (i.e. users) fall in cluster 0, 3288 samples fall in cluster 1, 82 samples fall in cluster 2, 403 samples fall in cluster 3, and 1253 samples fall in cluster 4. Each cluster shows us the level of user interaction according to the mean values of link_count and post_count (where the lowest interaction level is obtained with the lowest mean values of link_count and post_count). Table 1 shows the K-means clusters, sorted in ascending order according to the average value of their mean values of link_count and post_count, and the corresponding interaction level for each cluster.
- For EM clustering, we found that 6450 of dataset samples fall in cluster 0, 763 samples fall in cluster 1, 2612 samples fall in cluster 2, 24 samples fall in cluster 3, and 148 samples fall in cluster 4. Similar to K-means clustering, each EM cluster shows us the level of user interaction according to the mean values of link_count and post_count. Table 2 shows the EM clusters, sorted in ascending order according to the average value of their mean values of link_count and post_count, and the corresponding interaction level for each cluster.

Figures 2, 3, 4 and 5 show also a visualization of the distribution of users within the 5 K-means and EM clusters respectively, when clustering is performed using either link_count or post_count features respectively. From which, we can draw the following conclusion:

- Clearly, we see that in case of K-means clustering, cluster 2 (the very high interactive users group) and cluster 3 (the high interactive users group) have the lowest number of samples respectively. This means that when clustering is performed using either link_count or post_count features, a very small number of users can be classified as very high interactive users, while a larger number of users, but still small number can be classified as high interactive users.
- As for EM clustering, we see that cluster 3 (the very high interactive users group) and cluster 4 (the high interactive users group) have the lowest number of samples respectively. This means that when clustering is performed using either link_count or post_count features, a very small number of users can be classified as very high interactive users, while a larger number of users, but still small number can be classified as high interactive users. Clearly this matches the same observation obtained using K-means clustering.

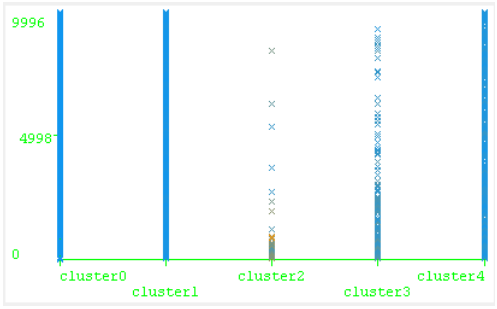

Fig. 4. A visualization of the distribution of users within the 5 K-means clusters based on the mean values of link_count of those clusters .

TABLE I
K-MEANS CLUSTERS IN ASCENDING ORDER ACCORDING TO THE AVERAGE VALUE OF THEIR MEAN VALUES OF LINK_COUNT AND POST_COUNT, AND THE CORRESPONDING INTERACTION LEVELS FOR EACH CLUSTER.

	Cluster0	Cluster1	Cluster4	Cluster3	Cluster 2
link_count	0.34	2.6	7.33	17.92	38.13
post_count	0.01	0.53	2.03	4.37	15.21
Interaction level	Very low interactive	Low interactive	Medium interactive	High interactive	Very high interactive

TABLE II
EM CLUSTERS IN ASCENDING ORDER ACCORDING TO THE AVERAGE VALUE OF THEIR MEAN VALUES OF LINK_COUNT AND POST_COUNT, AND THE CORRESPONDING INTERACTION LEVELS FOR EACH CLUSTER.

	Cluster0	Cluster2	Cluster1	Cluster4	Cluster 3
link_count	0.72	3.81	11.68	7.81	21.63
post_count	0	1.03	3.2	26.74	44.09
Interaction level	Very low interactive	Low interactive	Medium interactive	High interactive	Very high interactive

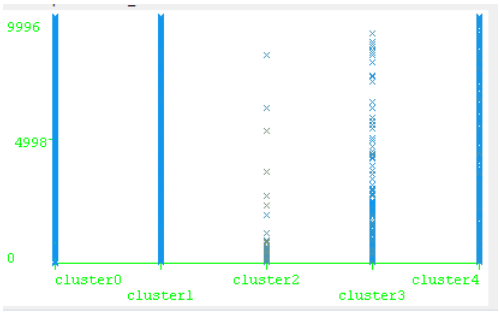

Fig. 5. A visualization of the distribution of users within the 5 K-means clusters based on the mean values of post_count of those clusters.

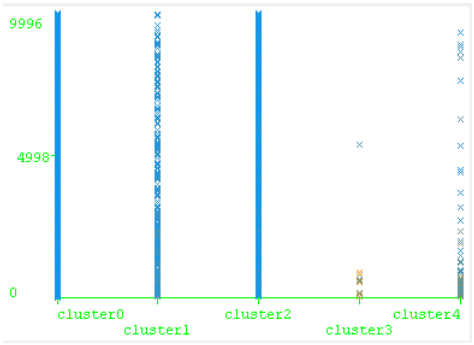

Fig. 6. A visualization of the distribution of users within the 5 EM clusters based on the mean values of link_count of those clusters.

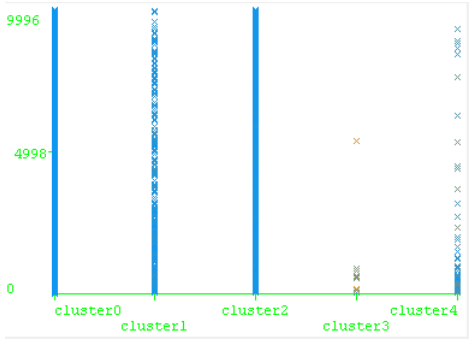

Fig. 7. A visualization of the distribution of users within the 5 EM clusters based on the mean values of post_count of those clusters.

IV. RELATED WORK

C. Wilson [6] et el. presented an approach for quantifying user interactions by analyzing interaction graphs derived from Facebook user traces to define a meaning to online social links. The main target of their approach was to discover if social links are considered as valid indicators of real user interaction on Facebook social network. They showed that interaction activity on Facebook is skewed towards a small portion of each users social links. Their observation raised some doubts regarding the assumption that all social links imply equal expressive friend relationships.

E. Trandafili [2] presented a machine learning approach that uses EM clustering algorithm for producing an accurate profiling of real-world social network users. Their approach clustered users into groups, then it uses decision tree learning for discovering interesting patterns among users that have no direct friendship links.

M. Eslami [8] presented an automated friend grouping tool that can be applied to a Facebook friendship network for creating groups of friends. Their tool uses three clustering algorithms for grouping friends. The main aim of their tool was to discover which clustering algorithm is more convenient for social network groupings and how to integrate the three clustering algorithms into a grouping interface

F. Erlandsson [22] proposed a machine learning approach based on association rule mining for predicting of user participation in online social networks discussions. The prediction mechanism was done based on the activeness of users within current posts. However their approach was limited to user interactions on a subset of Facebook users through posts with a similar topic.

V. Conclusions

In this paper, we have investigated how the machine learning clustering technique, particularly the K-means and EM clustering algorithms can be used for evaluating the interactions of users on Facebook social network. The number of friendship links that the user has and the number of posts that he/she has written to the wall of other users have been used as two features for training the clustering algorithms. The paper experiment has been conducted on a dataset extracted from two real Facebook datasets that contain data about links and wall posts of various users. This allow us to identify the values of the two aforementioned features for each user. The experimental results show that our approach is effective in categorizing the Facebook users into clusters based on their level of interaction, defined in terms of the link count and post count that each user has. Therefore, those clusters can represent various levels of interaction of users on Facebook social network. Thus, we can determine the user interaction level on Facebook by determining in which cluster this user falls. Moreover, the results show that a very small percentage of users on Facebook can be classified as as high interactive users.

As a future work, we are looking forward to applying our proposed approach for evaluating the interaction of users on other types of social networks. Also, we are looking forward to discovering the patterns that can help in determining the most preferred time for interaction of users on social network, where the level of user interaction is defined based on the time of posting. This can also be very useful for identifying the current active users on social networks.

References

[1] I. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc., 2005.

[2] E. Trandafili, M. Biba, and A. Xhuvani Profiling Social Network Users with Machine Learning *In Proceedings of BCI'13*, September 19-21, Thessaloniki, Greece, 2013.

[3] B. Viswanath, A. Mislove, M. Cha and K. P. Gummadi. On the Evolution of User Interaction in Facebook *In Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)*, August, 2009.

[4] WOSN 2009 Data Sets. Avilable at:http://socialnetworks.mpi-sws.org/data-wosn2009.html

[5] Q. Kong, W. Mao, and D. Zeng Predicting User Participation in Social Networking Sites *In proceedings of 2013 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2013.

[6] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B.Y. Zhao User Interactions in Social Networks and their Implications *In proceedings of EuroSys09, April 13*, Nuremberg, Germany, 2009.

[7] L. Jin, Y. Chen, T. Wang, P. Hu, and A. V. Vasilakos Understanding User Behavior in Online Social Networks: A Survey *IEEE Communications Magazine*, September 2013.

[8] M. Eslami, A. Aleyasen, R. Z. Moghaddam, and K. Karahalios Friend Grouping Algorithms for Online Social Networks: preference, bias, and implications *In proceedings of SocInfo 2014*, Springer International Publishing Switzerland 2014.

[9] Python Software Foundation. Python Language Reference, version 2.7. Available athttp://www.python.org

[10] P. Andritsos. Data Clustering Techniques-Qualifying Oral Examination Paper. University of Toronto, Department of Computer Science, March11, 2002.

[11] K. Chen. K-means Clustering. COMP24111 Machine Learning course, 2016. Avilable at:https://studentnet.cs.manchester.ac.uk/ugt/COMP24111/

[12] C. B Do and S. Batzoglou What is the expectation maximization algorithm?. *Nature Biotechnology*, volume 26, number 8, August 2008.

[13] J. Kleinberg Challenges in mining social network data: processes, privacy, and paradoxes *In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '07)*, San Jose, California, USA August 12 - 15, 2007.

[14] http://www.facebook.com/.

[15] https://www.linkedin.com/

[16] https://twitter.com/

[17] Chapter 4: Measures of distance between samples: Euclidean Avialable at: http://84.89.132.1/~michael/stanford/maeb4.pdf

[18] F. Erlandsson, A. Borg, H. Johnson, and P. Brodka. Predicting User Participation in Social Media *In proceedings of NetSci-X 2016*, LNCS 9564, pp. 126135, 2016.

[19] P. Dayan Unsupervised Learning. The MIT Encyclopedia of the Cognitive Science, 1999.

[20] L. Rokach, O. Maimon. Clustering methods. Data Mining and Knowledge Discovery Handbook, pp. 321352, Springer, New York, 2005,

[21] W. Ponchai, B. Watanapa, and K. Suriyathumrongkul, Finding Characteristics of Influencer in Social Network using Association Rule Mining. *In Proceedings of the 10th International Conference on e-Business (iNCEB2015)*, November 23rd - 24th 2015.

[22] F. Erlandsson, P. Brdka, A. Borg, H. Johnson Finding Influential Users in Social Media Using Association Rule Learning *Entropy Journal*, Volume 18, issue 64, 2016.