Journal of Engineering and Applied Sciences 12 (16): 4156-4160, 2017
ISSN: 1816-949X

Friend Recommendation System based on Modeling the Communities using Naive Bayes

[1]Huda N. Nawaf, [2]Wafaa. Al-Hameed and [3]Najah Rasheed Jouda
[1]Department of Information Technology Network,
[2]Department of Information Technology Software, Babylon University, Hillah, Iraq
[3]Department of Computer System, Babylon Technical Institute, Babil, Iraq

Abstract: Popular social networks sites such as Facebook and Twitter are still growing significantly. In this regard, a recommender system can be used to provide user experiences. In this study, we try modeling the online communities using naive bayes model. More specifically, the core of this work is modeling the user's past friends by taking into account the centrality measures and the latest friends. The two real datasets Facebook-Ego and Twitter are consider as a test bed for our proposed system then precision and recall measures have been applied to evaluate the accuracy of the system. In addition, a new metric, namely the $R_{top\text{-}list}$ metric is suggested to express the accuracy of prediction. In sum, the empirical results foster the efficiency of the proposed system.

Key words: Friend recommendation system, Naive Bayes, centrality measures, Ego networks, communities, Twitter

INTRODUCTION

Social networks services such as Facebook and Twitter are growing. This growth is very rapid from a few users to 1 billion users which makes it hopeless for a user to look for friends in this huge number of people (Zhang *et al*., 2015).

Customization of the user experience is a robust property of social networks. Hence, recommendation system has an important role to address the experience of the user and so a friend recommendation system appears.

Generally, recommendation systems concentrate on two areas: object recommendation and link recommendation. Amazon and Netflix companies recommend items to users according to their behavior in the past to provide object recommendation. Link recommendation which is presented in this study is concerned with social networks sites such as Facebook and Twitter.

Principally, the path-based and the Friends-of-Friend methods are considered the foundation of the current friend recommendation algorithms (Zhang *et al*., 2015). The dynamic nature of people's concept of friendship is considered as theme or challenge in developing friend recommendations. However, the concept of friendship is different from one individual to another which leads to change in the structure of the social networks over time.

There are several papers involved with this problem such as that by Wesley and Jianming (2010) where the user's interaction with others in social networks gives an idea about his or her choices when creating new friendships. Whereas in the study by Zhou the information of users and their total attributes have been used to propose friend recommendation system. These attributes of elected friends include gender, age, location, interest and number of common-neighbors. A semantic-based friend recommendation system for social networks has been suggested by Rani and Emmanuel. The researchers recommend friends depending on life style of users instead of social graphs. Thy employ sensor-rich smartphones by which a friend book discovers the life styles of users from user-centric sensor data and then friends have been recommended by measuring the similarity of life styles between users.

Ego networks: "Ego" is an individual "focal" node. Egos can be persons, groups or ganizations or whole societies. A network has as many egos as it has nodes. In other words, the ego and all nodes to which ego has a direct connection at some path length is called the "Neighborhood".

In social network analysis, the "neighborhood" of a node represent all of the nodes or egos with direct connections. The boundaries of Ego networks are defined in terms of neighborhoods (Hanneman and Ridddle, 2005). Figure 1 shows example of Ego network.

Corresponding Author: Huda N. Nawaf, Department of Information Technology Network, Babylon University, Hillah, Iraq

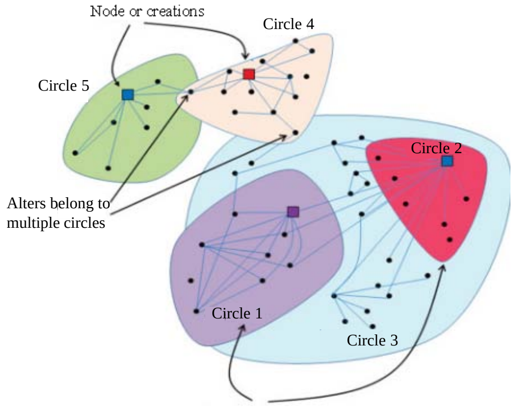

Fig. 1: Ego network

Currently, users in Facebook, Google and Twitter identify their circles either manually or in a naive fashion for example by identifying friends sharing certain features or properties in common.

Measures of centrality: In analysis of extensive scale networks, network centrality is considered a robust tool. Generally, the importance of a node is determined using centrality measure.

Supposed the network graph, the centrality scores are given depending on structural attribute of nodes. Degree, closeness and between are the commonly used centrality measures (Wasserman and Glalaskiewicz, 1994). Degree centrality indicates to the number of edges a node has to other nodes in an undirected network whereas in a directed network there are two different measures in-degree and out-degree centralities (Wasserman and Glalaskiewicz, 1994). As for closeness, it denotes to the inverse of the sum of the shortest distances between each node and every other node in the network graph as stated in Eq. 1:

$$C_c(n_i) = \left[\sum_{j=1}^{g} d(n_i, n_j)\right]^{-1} \tag{1}$$

Where:

$d(n_i, n_j)$ = The distance between node n_i and node n_j
g = The set of all nodes in graph

Eventually, between can be defined by how many times that a node lies on the shortest paths between any pair of nodes in the network graph as Eq. 2 state (Panda *et al*., 2014):

$$C_B(n_i) = \sum_{j<k} \frac{g_{jk}(n_i)}{g_{jk}} \tag{2}$$

Where:

g_{jk} = Total number of shortest paths from node j to k node
$g_{jk}(n_i)$ = The number of those paths that pass through n $_i$

MATERIALS AND METHODS

The proposed method is based on a Naive Bayes prediction model for each community of friends to recommend new friends for each user. Hence, the probability of occurrence F and C events together as follows:

$$P(New_f | Old_f) = P(New_{fi}) \prod_{j:j\neq i}^{m} P(Old_{fi} | Newf_i)$$

Where:

New_f = New friend for target user
Old_f = The friends who the target user has relation with them previously
f_j = The old friend (with highest closeness ($highest_C$-f), highest degree ($highest_D$-f) or latest friend (1_{last}-f). Indeed, f_j is different according to the conditional probability

We try to recommend new friends for the target user by finding the correlation between those new friends and those, Old_f who already has a relationship with them in the same community as the target user. Among those friends who the user already has a relationship with, a friend with high closeness, high degree or latest friend(s) has (have) been chosen with a conditional probability.

Hence, to predict new friends for a user, the relations among m friends and each new friend (New_f) in his community should be looked at which takes a long time. As a result, it should narrow the search domain should be narrowed by choosing friends with distinct properties such as friends with the highest degree, highest closeness or choosing the latest friends. In other words, Old_f can represent friend(s) with highest degree closeness or latest friend(s) in a sequence of friends for each user.

To predict new friends for a user in a community (with m users), the Naive Bayes method calculates a posteriors probabilities for each new friend and assigns to that user the new friends for which the probability is the greatest.

RESULTS AND DISCUSSION

The performance of proposed system has been evaluated on real world datasets Facebook-Ego and Twitter https://snap.stanford.edu/data/. Basically, it has been focused on relationships among users in both datasets to meet the research question. Regarding Facebook-Ego, the experiments have been conducted in two independent ways to get the best results. The first one on community level and the second one on the Ego network level. Table 1 describes both datasets.

It is worth mentioning y that the evaluation measures which have been applied are the popular measures such as recall and precision. In addition, a new measure has been proposed by which we can know the percentage of data appearing at the top of the recommendation list in the test set ($R_{top\text{-}list}$). In other words, this ratio represents the number of times that the first item in the recommendation list (has highest probability) is one of the relevant data (validation or test set).

The first experiment has been applied on the Facebook-Ego dataset on community level. At the beginning, an adjacency matrix has been created for this dataset to group users into communities using the open source software gephi which produces 13 communities as shown in Fig. 2.

Next, the Naive Bayes model has been created for each community. As for the second experiment, it has been accomplished on an Ego network, after using the Naive Bayes Model as previously but for each Ego network. In other words 184 models have been created due to having 184 egos as shows in Table 2.

However, the results have been recorded with each case that is mentioned in Section 4. Additionally, the model has been applied when using one event (one friend) and some combination of two different events (two friends).

For $R_{top\text{-}list}$, the results have been recorded for the top N in the recommendation list, namely 1, 2, 3, 4, 5, 10, 50 and 100 as shown in Table 2. Whereas, the recall and precision measures are calculated as stated in Table 3 and 4 at N values for $N\in\{1, 5, 10, 50, 100\}$. The bold blue and red values are representing the best results in Ego networks and communities, respectively.

It is absolutely logical to get the highest ratio ($R_{top\text{-}list}$) when the length of list is one due to looking for the top item in the list which has one item. However, the ratio decreases as the length of list increases and, therefore, is not surprised (Surprising). Generally, the latest friends event, in particular 3_{last}-f, give the best ratio in an Ego network (bold blue). In contrast, the community level record best ratio at degree centrality (bold red).

Once again, the best recall and precision for Ego networks have been obtained at last friend(s) event(s). Similarly, the performance of the model the in community is the best at latest friend(s) in terms of recall as well however, it is not in terms of precision where it shows the same performance approximately. In general, the performance of the model in the Ego network is superior from that of communities in Rtop-list, recall and precision.

A Twitter dataset is also used to validate the proposed system in a third experiment where it has been clustered into 5 groups using gephi as shows in Fig. 3. It is important to say that only 90000 users have been elected from Twitter dataset.

Table 5-7 state the results of the same evaluation measures which have been used with Facebook-Ego dataset. Roughly, all results related to Twitter dataset indicate to last two or three friends provide the best results, even so some good results relate with Random Friend (R-f) with respect to recall measure as shown in Table 6.

Table 1: Description of Ego-Facebook and Twitter

Vaeiables	Types	Nodes	Edges	No. of circles (Ego)
Ego-Facebook	Undirected	4,039	88,234	184
Twitter	Directed	1,000,000	-	-

Table 2: R_{top_list} versus different characteristics of friends who have relation with target user

Factors	Facebook communities	R@1	R@5	R@10	R@50	R@100
Highest$_C$-f	Ego-Facebook	0.0300	0.08	0.13	0.24	0.35
	communities	0.0010	0.01	0.04	0.16	0.24
Highest$_D$-f	Ego-Facebook	0.0200	0.05	0.10	0.20	0.27
	communities	0.0010	0.01	0.03	0.17	0.24
1_{last}-f	Ego-Facebook	0.0400	0.14	0.21	0.40	0.52
	communities	0.0030	0.02	0.04	0.20	0.30
2_{last}-f	Ego-Facebook	0.0500	0.14	0.21	0.37	0.54
	communities	0.0030	0.01	0.03	0.16	0.25
3_{last}-f	Ego-Facebook	0.0500	0.15	0.21	0.36	0.54
	communities	0.0001	0.02	0.03	0.20	0.30
1_{last}-f and Highest$_C$-f	Ego-Facebook	0.0400	0.11	0.18	0.28	0.40
	communities	0.0000	0.02	0.04	0.17	0.26
Highest$_C$-f and Highest$_D$-f	Ego-Facebook	0.0200	0.06	0.10	0.20	0.27
	communities	0.000	0.01	0.03	0.16	0.24
R-f	Ego-Facebook	0.020	0.05	0.10	0.20	0.27
	communities	0.002	0.01	0.03	0.16	0.25

Table 3: Recall versus different characteristics of friends who have relation with target user

Factors	Facebook communities	$R_{top\text{-}list}$@ 1	$R_{top\text{-}list}$@ 2	$R_{top\text{-}list}$@ 3	$R_{top\text{-}list}$@ 4	$R_{top\text{-}list}$@ 5	$R_{top\text{-}list}$@ 10	$R_{top\text{-}list}$@ 50	$R_{top\text{-}list}$@ 100
Highest$_C$-f	Ego-Facebook	1	0.57	0.45	0.38	0.34	0.20	0.080	0.050
	communities	1	0.50	0.13	0.07	0.05	0.02	0.005	0.003
Highest$_D$-f	Ego-Facebook	1	0.62	0.49	0.42	0.36	0.19	0.070	0.040
	communities	1	0.35	0.32	0.26	0.22	0.06	0.010	0.005
1_{last}-f	Ego-Facebook	1	0.60	0.48	0.42	0.37	0.25	0.140	0.110
	communities	1	0.40	0.21	0.16	0.12	0.05	0.010	0.010
2_{last}-f	Ego-Facebook	1	0.62	0.49	0.42	0.37	0.27	0.170	0.150
	communities	1	0.31	0.30	0.11	0.10	0.04	0.010	0.010
3_{last}-f	Ego-Facebook	1	0.65	0.52	0.44	0.40	0.27	0.180	0.160
	communities	1	0.40	0.30	0.20	0.18	0.03	0.006	0.003
1_{last}-f and Highest$_C$-f	Ego-Facebook	1	0.59	0.45	0.39	0.33	0.22	0.120	0.100
	communities	1	0.33	0.13	0.05	0.03	0.02	0.005	0.003
highest$_C$-f and highest$_D$-f	Ego-Facebook	1	0.60	0.44	0.34	0.29	0.16	0.060	0.040
	communities	1	0.26	0.12	0.05	0.03	0.01	0.004	0.003
R-f	Ego-Facebook	1	0.60	0.44	0.34	0.30	0.16	0.060	0.040
	communities	1	0.55	0.26	0.24	0.13	0.04	0.010	0.010

Table 4: Precision versus different characteristics of friends who have relation with target user

Factors	$R_{top\text{-}list}$@ 1	$R_{top\text{-}list}$@ 2	$R_{top\text{-}list}$@ 3	$R_{top\text{-}list}$@ 4	$R_{top\text{-}list}$@ 5	$R_{top\text{-}list}$@ 10	$R_{top\text{-}list}$@ 50	$R_{top\text{-}list}$@ 100
Highest$_C$-f		0.65	0.50	0.42	0.40	0.30	0.20	0.14
Highest$_D$-f	1	0.62	0.50	0.40	0.40	0.30	0.16	0.14
1_{last}-f		0.65	0.53	0.45	0.40	0.30	0.20	0.14
2_{last}-f	1	0.66	0.53	0.50	0.43	0.35	0.30	0.25
3_{last}-f		0.66	0.54	0.50	0.44	0.36	0.30	0.30
1_{last}-f and highest$_C$-f	1	0.65	0.53	0.45	0.40	0.35	0.24	0.21
Highest$_C$-f and highest$_D$-f	1	0.66	0.55	0.5	0.45	0.35	0.23	0.2
R-f	1	0.66	0.54	0.45	0.43	0.34	0.22	0.22

Table 5: Rtop list versus different characteristics of friends who have relation with target user

Factors	R@1	R@5	R@10	R@50	R@100
Highest$_C$-f	0.01	0.03	0.05	0.12	0.20
Highest$_D$-f	0.01	0.03	0.05	0.13	0.20
1_{last}-f	0.01	0.03	0.05	0.12	0.20
2_{last}-f	0.03	0.13	0.20	0.40	0.45
3_{last}-f	0.04	0.13	0.20	0.36	0.41
1_{last}-f and highest$_C$-f	0.01	0.03	0.06	0.14	0.20
Highest$_C$-f and highest$_D$-f	0.01	0.03	0.05	0.14	0.20
R-f	0.03	0.11	0.20	0.40	0.50

Table 6: Recall versus different characteristics of friends who have relation with target user

Factors	Facebook communities	p@1	p@5	p@10	p@50	p@100
Highest$_C$-f	Ego-Facebook	0.100	0.05	0.05	0.03	0.02
	communities	0.005	0.01	0.01	0.02	0.02
Highest$_D$-f	Ego-Facebook	0.040	0.02	0.02	0.02	0.01
	communities	0.005	0.01	0.01	0.02	0.02
1_{last}-f	Ego-Facebook	0.140	0.13	0.10	0.07	0.04
	communities	0.010	0.02	0.02	0.02	0.02
2_{last}-f	Ego-Facebook	0.200	0.13	0.11	0.06	0.05
	communities	0.010	0.01	0.02	0.02	0.02
3_{last}-f	Ego-Facebook	0.230	0.15	0.12	0.06	0.05
	communities	0.004	0.02	0.02	0.02	0.02
1_{last}-f and Highest$_C$-f	Ego-Facebook	0.120	0.08	0.07	0.04	0.03
	communities	0.001	0.02	0.02	0.02	0.02
Highest$_C$-f and highest$_D$-f	Ego-Facebook	0.040	0.03	0.02	0.02	0.01
	communities	0.001	0.02	0.02	0.02	0.02
R-f	Ego-Facebook	0.100	0.05	0.05	0.04	0.03
	communities	0.01	0.16	0.04	0.01	0.01

Table 7: Precision versus different characteristics of friends who have relation with target user

Factors	P@1	P@5	P@10	P@50	P@100
Highest$_C$-f	0.03	0.03	0.02	0.01	0.01
Highest$_D$-f	0.03	0.03	0.02	0.01	0.01
1_{last}-f	0.04	0.03	0.02	0.01	0.01
2_{last}-f	0.20	0.14	0.11	0.04	0.02

Table 7: Continue

Factors	P@1	P@5	P@10	P@50	P@100
3_{last}-f	0.20	0.14	0.11	0.04	0.02
1_{last}-f and $highest_C$-f	0.05	0.03	0.03	0.01	0.01
$Highest_C$-f and $highest_D$-f	0.04	0.03	0.03	0.01	0.01
R-f	0.10	0.12	0.10	0.04	0.03

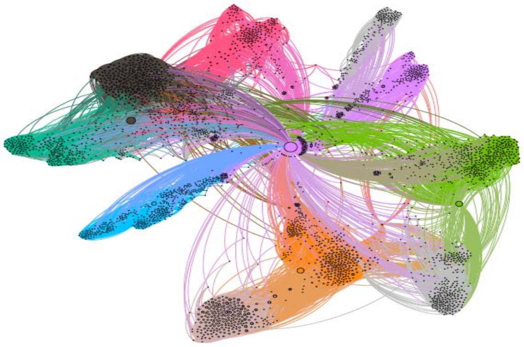

Fig. 2: Facebook communities using Gephi

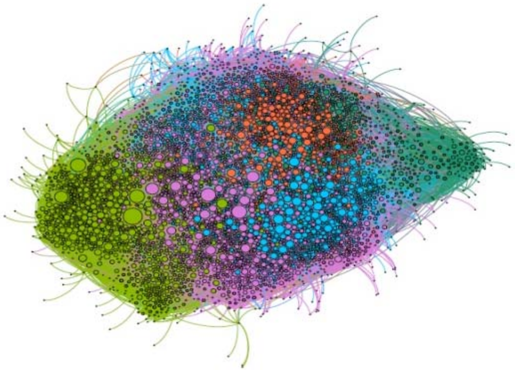

Fig. 3: Gephi

CONCLUSION

In our study, we state that the Naive Bayes Model is affected by several factors, namely the probability condition and the domain of the users which is reflected in the accuracy of the prediction. As a result, a best condition is selected from a set of candidates and a best domain for Facebook at least is selected.

Among of conditions are centrality measures which determine the centrality of a node in a community such as degree, closeness. In addition, the latest friends have been considered as a probability condition too. In sum, we can conclude that the centrality measures such as degree and closeness has no effect on accuracy of the system positively when consider it as conditional probability in naive bayes model but with respect to R_{top}-list of Facebook-Ego at Ego network level the degree and closeness centrality measures are a little better and less than that in community level for R-f and latest friends respectively. From the point of view of this study, the centrality measures have an important role in small communities (Ego network) which is why the results are superior. Additionally, we can conclude that the latest friends have a big role to bring new friends as state in tables for both datasets.

REFERENCES

Hanneman, R.A. and M. Riddle, 2005. Introduction to Social Network Methods. University of California, California, UK.

Panda, M., S. Dehuri and G.N. Wang, 2014. Social Networking: Mining, Visualization and Security. Springer, Berlin, Germany,.

Wasserman, S. and J. Galaskiewicz, 1994. Advances in Social Network Analysis: Research in the Social and Behavioral Sciences. SAGE Publications, Thousand Oaks, USA., ISBN: 9780803943032, Pages: 300.

Wesley, W. and H.C. Jianming, 2010. A social network-based recommender system. Ph.D Thesis, University of California, Los Angeles, California, USA.

Zhang, Z., Y. Liu, W. Ding and W.W. Huang, 2015. A friend recommendation system using users' information of total attributes. Proceedings of the 2nd International Conference on Data Science, August 8-9, 2015, Springer, Sydney, Australia, pp: 34-41.