HybridRank: Ranking in the Twitter Hybrid Networks

Jianyu Li
Department of Computer Science
University of Maryland, College Park
jli@cs.umd.edu

ABSTRACT

User influence in social media may depend on multiple modes of communication. Research has identified the importance of a hybrid network on Twitter, comprising of the follower, retweet and mention networks. HybridRank is an extension of PageRank which considers all three networks. Using a longitudinal panel dataset of 8K Twitter users, we evaluated the performance of HybridRank and predicted the influence of users. We used the number of times that a user was retweeted or mentioned as a proxy for influence. HybridRank outperforms a (good) baseline method for mentions outside the focal user's follower network, as well as for novel mentions (users who have not mentioned the focal user in the past).

General Terms

rankings, PageRank, network, Twitter, hybrid networks

1. INTRODUCTION

Social media has grown in volume, usage and impact. A user can post tweets, contribute to information diffusion by retweeting, and can engage in a conversation by mentioning other users. An active user may engage in all of the above behavior in an attempt to attract more followers.

We are interested in a measure of a user's influence. The size of the follower count could be one such measure; however, this may be a better reflection of popularity. A more accurate measure may be reflected by the count of how many times a user was retweeted, or how many times a user was mentioned. An additional signal of influence is when a user is mentioned by users outside their follower network, or when a user attracts a novel mention (by a user who has not mentioned the focal user in the past).

Social influence analysis and rank analysis has drawn a lot of research interests. Previous research studies include PageRank analysis [3] on web document, objectRank analysis [1] on database system, futureRank on scientific article [4], and personalized recommendation analysis for focal users based on individual level influence [6]. K.Subbian proposed a supervised rank aggregation technique to rank user's influence in Twitter [5]. The paper proposed a model which generates high AUC and AP scores for the second month data. However, high AUC and AP score was expected because the data they used are in two consecutive month. A user who was active in the first month has a high chance to be active in the second month. There is no guarantee that their model will predict well for long term data.

Using a snowball sampling mechanism, we created a network of 15K active users [6]. We then sampled out a 8K users network, in which each user has at least 2.4% of his followers and friends. We collected these 15K users' tweets for 60 consecutive days in summer 2011, and subsequently collected their tweets after 12 months.

We have successfully shown that a hybrid network of retweet and mention network in Twitter has more power in predicting a long term rank based on HybridRank scores. It also performs better in predicting focal user's future mentioned count by users who was not in focal user's network and did not mention focal user before. In addition, we found that adding follower network to hybrid network would not help in measuring user influences because it makes prediction result even worse.

2. DEFINITION FOR TRANSITION MATRIX

2.1 PageRank

We briefly describe the PageRank algorithm and then propose our transition matrix for the three hybrid networks in Twitter.

The underlying assumption of PageRank is that links between pages confer authority. A link from page i to page j is evidence that i is suggesting that j is important. The importance contributed to page j from i is inversely proportional to the out degree of i. Let D_i be the out degree of page i. The corresponding random walk on the directed web graph can be expressed by a transition matrix A as follows:

$$A[i,j] = \begin{cases} \frac{1}{D_i} & \text{if there is an edge from i to j} \\ 0 & \text{otherwise} \end{cases}$$

When calculating PageRank, suppose R be the PageRank vector, and let damping factor ϵ be the probability that a web surfer follows the hyperlinks and let $1-\epsilon$ be the probability of a surfer making a random jump. Let P denote the base set probability, in PageRank, it is an $n \times 1$ vector, and each entry has the value $\frac{1}{n}$ (equally probability). The following is the way we calculate original PageRank:

$$R = \epsilon A^T R + (1 - \epsilon)P$$

3. HYBRID NETWORK

Three networks, namely Retweet network, Mention network and Follower network, exist in Twitter. Every two users may have a relationship in any of these three networks.

Three types of edges, namely Retweet(R) edge, Mention(M) edge and Follower(F) edge, are introduced in hybrid network. Edges are directed. A Retweet edge exists from user i to user j if i has retweeted of j. Similarly, Mention edge exists from i to j if i has mentioned of j, Follower edge exist from i to j if i follows j, in other words, j is a friend of i.

The edge weight between i and j indicates the number of times i retweet/mention j, and is 1 if i follows j. Each matrix entry in transition matrix follows the rule of original PageRank. To be more specific, the entry for a Retweet edge from i to j would be calculated by the number of times i retweeted of j, divided by the total number of retweet by user i. Similarly, the entry for a Mention edge from i to j is the number of times i mentioned of j, divided by the total number of mentions by user i. The entry for a Follower edge from i to j is 1 divided by the number of friend i has, if i follows j, otherwise, it is 0.

3.1 Three Types (R,M,F) of Transition Matrix

In Twitter network, three transition matrixes, namely Retweet A_R, Mention A_M, and Follower A_F exist. Let R_{ij} denotes the total number of times i retweet of j, and R_i denotes the total number of retweets by user i. M_{ij} denotes the total number of times i mentions of j, and M_i denotes the total number of mentions by user i. F_{ij} denotes whether user i follows user j, if i follows j, then the value of F_{ij} is 1, otherwise, it is 0. And F_i denotes the total number of users i follows. We have the following transition matrix entry definition.

$$A_R[i,j] = \begin{cases} \frac{R_{ij}}{R_i} & \text{if } R_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$A_M[i,j] = \begin{cases} \frac{M_{ij}}{M_i} & \text{if } M_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$A_F[i,j] = \begin{cases} \frac{1}{F_i} & \text{if i follows j} \\ 0 & \text{otherwise} \end{cases}$$

For example, in figure 1, User A has three types of links (R,M,F) to or from other users. A follows B, C and D, A mentions B and D, A retweet D, and A got mentioned by B. We could easily figure out the transition matrix entry value of R,M,F edges from A to D according to the entry definition for each edge type. For example, for M edge type, A mentions D 10 times, and B 90 times, so the M edge entry from A to D would be $\frac{10}{10+90}$, which is 0.1. For R edge type, consider the R edges coming from A, the R edge entry from A to D would be $\frac{20}{20}$, which is 1. For the F edge, each F edge would have value 1 to indicate the follower/friend relationship, so the F edge entry from A to D is $\frac{1}{3}$ because A follows D, and A has three friends.

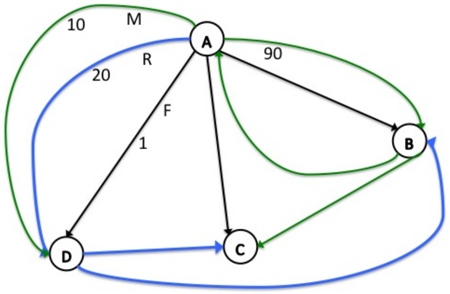


Figure 1: Black, green, blue indicates Follower, Mention, Retweet type edge correspondingly; The weight of edge indicates the number of times user i retweet/mention user j; 1 in Follower edge indicates user i follows user j

3.2 Hybrid Transition Matrix

In hybrid networks, we combine the R,M,F transition matrixes between two users and construct our hybrid transition matrix A_{hybrid} as follows:

$$A_{hybrid}[i,j] = \alpha A_R[i,j] + \beta A_M[i,j] + \gamma A_F[i,j]$$

$$\alpha + \beta + \gamma = 1$$

In our experiment, we set α to 0.5 and β to 0.5 to get hybrid networks of Retweet (R) and Mention (M) network; and we set α to 0.33, β to 0.33 and γ to 0.33 to get hybrid networks of R,M,F.

3.3 Hybrid Global Rank

To evaluate HybridRank, we introduced the ranking vector R_{hybrid}. Therefore, we have the following:

$$R_{hybrid} = \epsilon A^T_{hybrid} R_{hybrid} + (1 - \epsilon)P$$

We set ϵ to 0.85 in our experiment.

4. DATA

4.1 15,000 user network

We want to construct a dataset that reflected a comprehensive history of user interaction and tweet content, over two collection periods, for the same set of significant number of active users, given the strict limitations imposed by the Twitter API. We constructed 15,000 significant active users, and as well as their follower(friend) subnetwork within these 15,000 active users. By saying active, we mean an active user must satisfy the following constrains:

1) At least have one retweet during his/her latest 100 tweets.

2) Tweet frequency is not less than 1 tweet per day in the latest 100 tweets.The calculation of the time span of the latest 100 tweets is from the oldest time in the 100 tweets to the time when the 100 tweets are collected.

We used Twitter API to construct the network in the following way: pick a seed user from focal user list, and used BFS to explore seed user's follower (friend) network. If seed user's follower (friend) is an active user, we will add him/her to the focal user list. We kept doing this until no more users could be expanded, or stopped by hand once we get 15K focal user.

4.2 8k user network

To capture a more active subset, we want to construct a subset of 15k users in which each user would have more than a threshold $X\%$ of his total number of follower/friend in this subset. We used a threshold $X\%$ to filter out those users in the subset by the following way:

1) First get a set of the users who has at least $X\%$ of friends and also at least $X\%$ followers from the 15K users. Label this set of users as S.

2) Repeat the following loop until the number of users in S is stable, i.e., $|S|$ does not change: For each user in S, if the number of her friends or the number of her followers from S is less than $X\%$ of the total number of his friends or the total number of his followers, remove this user from S. 3) Return the set of users S.

We set the threshold $X\% = 2.4\%$, and finally obtained 7780 users in this subset.

4.3 Short Term Data description

Once we have collected the 15K focal user, we collected their two-month tweets.

The data statistics for the two-month tweets data are the following:

1) Time range: 04/25/2011 to 06/25/2011

2) Number of tweets: 10,979,278

3) Number of retweet: 2,366,211

4) Number of mention (not a reply): 6,678,227

4.4 Long Term Data description

After over a year, we collected the same 15K focal users' tweets data. The data time range is from 2012-09-07 13:00:04 to 2012-12-11 06:56:23. Because only 5000 user queries can be active at once with the Twitter API, we had to split data into three databases. As the limited time, We randomly picked one database to fully extract retweet and mention data. The statistics of the selected database data are:

1) Time range: 2012-09-07 13:00:05 to 2012-12-11 06:50:55

2) Number of tweets: 2,176,608

3) Number of retweet: 519,224

4) Number of mention (not a reply): 551,780

4.5 Diffusion based authority flow on 8K network

We want to measure the influence in the 8k active network. From the definition of our hybrid adjacency matrix and the way we collected the data, we could provide an accurate and interesting PageRank score for users in 8K nodes subgraph. Define our 8K network as G_l, and the rest of network in Twitter as G_e. It would be possible to have very popular users in G_e that would retweet or mention users in G_l, however, we are only interested in diffusion based authority flow, and therefore, retweet or mention happens with high probability between two users if there is a following path between them. Because of the way we collected the data (expand the active follower/friend network), and also the limitation to get all inflow data from G_e to G_l, the way we calculated the PageRank score in 8K graph on diffusion based authority flow is accurate.

5. MEASUREMENT

5.1 Assumptions

The retweets or mentions user received may come from two parts: steady part and random part. Steady means that for a person, if the environment keeps the same, he/she should usually receive some stable numbers of retweets/mentions during a specific time window size. Random means suddenly for whatever reason, a person receives more or less retweets or mentions than he/she should receive.

Therefore, the retweeted count R and mentioned count M could be expressed as $R = R_s + R_r$ and $M = M_s + M_r$. R_r and M_r could be positive or negative. We assume R_s and M_s have more relationship to future retweeted count R_{future} and future mentioned count M_{future}, and hence have more prediction power. Our baseline method is the retweeted/mentioned count from 04/25/2011 to 5/25/2011. Since it uses R or M for prediction, the method might be impacted by R_r and M_r greatly.

We assume that the hybrid network PageRank would capture more features of R_s and M_s, therefore would make a better prediction for R_{future} and M_{future}, outperformed the baseline method. Besides, hybrid network PageRank would capture whether the user has the influence to get retweeted or mentioned by 'newbies' who is not in his/her follower network and did not retweeted or mentioned him/her before.

We also assume that baseline method R and M in the 1st month would perform well on predicting the 2nd month R and M because consecutive month may not cause too much random part, while it would perform much worse for year later R and M because of too many possible randomness affecting the prediction.

5.2 Short Term and Long Term Ground Truth

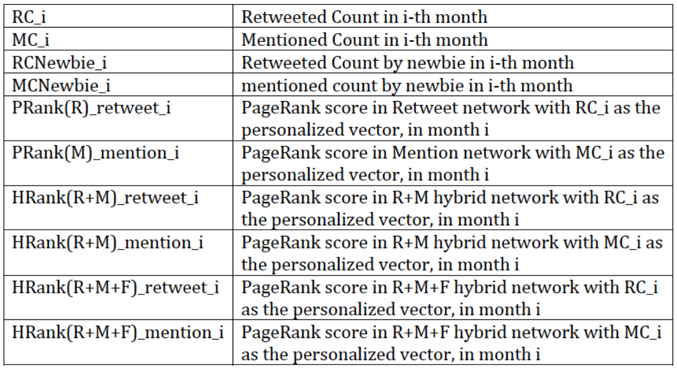
Based on our assumption, we picked three ground truth ranking methods for both short term and long term in order to test our assumptions.

1) Retweeted/Mentioned Count: number of times get retweeted or mentioned by users in the 8k network. We only consider mention which is not a reply as an effective mention. This ground truth measures the focal user's influence in the 8k network.

2) Retweeted/Mentioned Count by Newbies, number of times get retweeted/mentioned by 'newbies' who were not in user's direct follower network, and did not retweet/mention the user in the first month. This ground truth measures the focal user's influence in attracting newbies outside of his/her follower network and who were not a fan of focal user before.

3) Hybrid network (R+M network) PageRank scores. We are interested in the hybrid network rank scores in future short term and long term because PageRank scores in a hybrid network would also be an effective measure of a user's influence.

5.3 Metric



RC_i	Retweeted Count in i-th month
MC_i	Mentioned Count in i-th month
RCNewbie_i	Retweeted Count by newbie in i-th month
MCNewbie_i	mentioned count by newbie in i-th month
PRank(R)_retweet_i	PageRank score in Retweet network with RC_i as the personalized vector, in month i
PRank(M)_mention_i	PageRank score in Mention network with MC_i as the personalized vector, in month i
HRank(R+M)_retweet_i	PageRank score in R+M hybrid network with RC_i as the personalized vector, in month i
HRank(R+M)_mention_i	PageRank score in R+M hybrid network with MC_i as the personalized vector, in month i
HRank(R+M+F)_retweet_i	PageRank score in R+M+F hybrid network with RC_i as the personalized vector, in month i
HRank(R+M+F)_mention_i	PageRank score in R+M+F hybrid network with MC_i as the personalized vector, in month i

Figure 2: Definition of different measuring matrices

The metric we picked here is Spearman's rank correlation. Spearman's rank correlation is a measure of how two ranked sets are related to each other. Correlation result ranges from -1 to 1. If two ranked lists are positively correlated, the value of Spearman's rank correlation result would be close to 1. If they are negatively correlated, the value would be close to -1. In general, if Spearman's rank correlation value is around 0, we may conclude that the two ranked lists are not related to each other.

6. EVALUATION RESULT

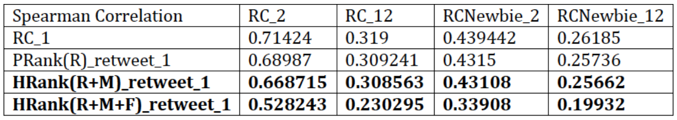
We measured user influence in matrices including retweeted and mentioned count, PageRank score of a single network and hybrid network in the first month in short term. Our test set includes the second month data in the short term and the year later data in the long term.

The definition for different matrices are in figure 2.

From figure 3, we found that all measures in the first month have high correlation scores with the second month data, while relatively low correlation scores with the year later data. More interestingly, the count measure in retweet network outperforms the PageRank measures. Besides, the hybrid network of R and M performs much better than the hybrid network of R, M, F. We could see the noise that the follower network brings in when comparing these two measures.

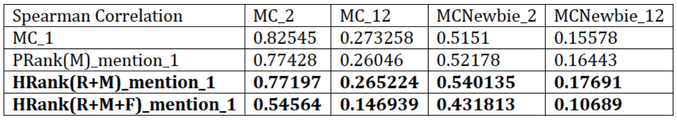
The result verified our original assumption that the correlation result with long term data would be worse because the data would be affected by much more random noise than the short term data. However, we did not see PageRank measure in Retweet network or the hybrid network of Retweet and Mention performs better than Count measure, which could be explained by that too much noise in a year window made term R_r in $R = R_s + R_r$ dominant, and dampened the effect of R_s.

We found almost the same phenomenon from figure 4, with some exceptions. In this case, the hybrid network of Retweet and Mention outperforms the count measure and the single network (M) measure by yielding better correlation results with future newbies mention count, which indicates a hybrid network measure has more prediction power on the future attraction of newbies in respect of the number of mentions. Hybrid network, when including F network, still performs the worst among these measures.



Spearman Correlation	RC_2	RC_12	RCNewbie_2	RCNewbie_12
RC_1	0.71424	0.319	0.439442	0.26185
PRank(R)_retweet_1	0.68987	0.309241	0.4315	0.25736
HRank(R+M)_retweet_1	**0.668715**	**0.308563**	**0.43108**	**0.25662**
HRank(R+M+F)_retweet_1	**0.528243**	**0.230295**	**0.33908**	**0.19932**

Figure 3: Spearman's rank correlation result with future Retweet Count



Spearman Correlation	MC_2	MC_12	MCNewbie_2	MCNewbie_12
MC_1	0.82545	0.273258	0.5151	0.15578
PRank(M)_mention_1	0.77428	0.26046	0.52178	0.16443
HRank(R+M)_mention_1	**0.77197**	**0.265224**	**0.540135**	**0.17691**
HRank(R+M+F)_mention_1	**0.54564**	**0.146939**	**0.431813**	**0.10689**

Figure 4: Spearman's rank correlation result with future Mention Count

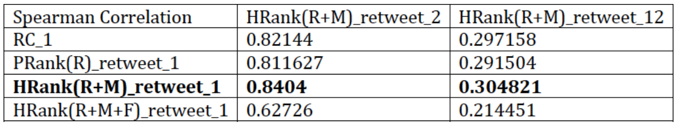
From figure 5 and figure 6, using hybrid rank as ground truth, we found that the hybrid PageRank in 1st month data outperforms both a single network (R or M) PageRank and the Count measure in predicting the long term hybrid PageRank scores. It indicates that in Twitter network, the PageRank score of retweet and mention hybrid network could make better predictions on user influence scores in the future, even for long term prediction.

7. CONCLUSION AND FUTURE WORK

We presented different ranking measures for the 8k diffusion based active network, and their correlation results with future three ground truths in both short term and long term. From the result, we verified our assumption that the performance of the measure works well for 2nd month data, however, relatively worse for the long term data. We found that a hybrid network of Retweet and Mention performs well on ranking users influence based on future attractions to newbies, and outperforms other measures in predicting long term user influence ranks. Besides, we observed that when adding the Follower network to Hybrid network, the prediction power decreases. The reason for that could be million follower fallacy Cha pointed out in her work [2]. Follower network may not be a good measure of influence, adding it to hybrid network would make noises in measuring user influences.

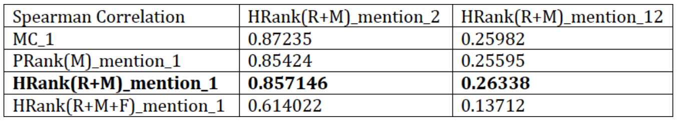
The result from figure 3 shows that PageRank measure did not outperform Count measure in measuring future retweet count. Reason could be due to too much random retweet count happened in the year window. In future work, we could shrink the window size to half a year, and to check whether PageRank could perform better when there is less random noise. We would also want to incorporate topic analysis into Twitter network, which would need personalized rankings on different topics.

8. REFERENCES



Spearman Correlation	HRank(R+M)_retweet_2	HRank(R+M)_retweet_12
RC_1	0.82144	0.297158
PRank(R)_retweet_1	0.811627	0.291504
HRank(R+M)_retweet_1	**0.8404**	**0.304821**
HRank(R+M+F)_retweet_1	0.62726	0.214451

Figure 5: Spearman's rank correlation result with future HybridRanks in retweet domain



Spearman Correlation	HRank(R+M)_mention_2	HRank(R+M)_mention_12
MC_1	0.87235	0.25982
PRank(M)_mention_1	0.85424	0.25595
HRank(R+M)_mention_1	**0.857146**	**0.26338**
HRank(R+M+F)_mention_1	0.614022	0.13712

Figure 6: Spearman's rank correlation result with future HybridRanks in mention domain

[1] A. Balmin, V. Hristidis, and Y. Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *VLDB*, pages 564–575, 2004.

[2] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, 2010.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[4] H. Sayyadi and L. Getoor. Futurerank: Ranking scientific articles by predicting their future pagerank. In *SDM*, pages 533–544, 2009.

[5] K. Subbian and P. Melville. Supervised rank aggregation for predicting influencers in twitter. In *SocialCom/PASSAT*, pages 661–665, 2011.

[6] S. Wu, L. Gong, W. Rand, and L. Raschid. Making recommendations in a microblog to improve the impact of a focal user. In *RecSys*, pages 265–268, 2012.