Social Network Influencer Ranking Based On Link Analysis

Cheulsoon Baek
cbaek@my.yorku.ca
216880445

Amirhossein Razavi
ahr91@my.yorku.ca
216715963

Lu Zhuang
lukikylu@my.yorku.ca
215665797

Myeongwook Lee
kylelee@my.yorku.ca
216100901

ABSTRACT

Nowadays, Social Network Services are significantly affecting the world. People communicate with others, read news, express their thoughts, and even earn money on it. The impact of SNS will increase as time goes on since the new generations more likely tend to depend on it. Hence, the industries integrate the SNS into their marketing by creating the industrial account on it or putting advertisements on the well-known famous influencers on the network. This new trend is happening because it is known that "Marketing through conventional channel is four times more expensive than marketing through the Internet" (Kotler et al., 2012) [1]

KEYWORDS

SNS, social networking, influencer, ranking, pagerank, weighted page rank

1 INTRODUCTION

This project is about ranking people in social network services based on their influential power, which is essential as many industries use Social Media Marketing. This project can help them choose which influencers to pick for their marketing to maximize profits with the least money spent on advertising.

Social Media Marketing is a relatively new approach in the marketing industry that can be effective and affordable when done right. Many businesses failed because they could not find the customers they needed, and customers who needed this business to be functioning did not know it existed. This problem arises from not advertising correctly, and this project can help them to solve it. This project is critical as it can help have a more precise advertisement system targeting people based on their interests and eventually help many businesses to grow and prevent them from failing.

Applications of this project are to the areas in which the flow of information is both sensitive and essential such as advertising, politics and news reports.

1.1 Follower Counts

It is possible to get the ranking of the users simply based on their followee and follower count. More follower count obviously means that the user has more influence on others. This is simple and strong logic; however, the count can result fake results.

Since the simple count means that the account has more power, some people increases their follow count using fake accounts. Those fake accounts obviously do not affect the real-life influence power of users but the people can be tricked by the number.

Hence, the follower count is obviously one option to measure the influence power of user, it is not sufficient to get the real-life result.

1.2 Analyzing Social Network

The social network, which is the follow relationship graph between users, can be used to analyze the more specific influence ranking. Using the social graph, the PageRank algorithm can be applied and the user ranking can be obtained with its centrality values.

This PageRank algorithm result is much specific and correct analysis on user influence than just counting followers. However, this simple result can sometimes wrong. In social network, even if one user follows another user, they do not read all posts they post on the service. Also, even if one person has good PageRank result in network, the user is not a good influencer if the person do not post anything.

This project addresses these issues and tries to figure out better way to calculate the actual proper influencer rankings.

2 PROBLEM DEFINITION

This project focuses on the influence ranking of users in social network services. Unlike world-wide web, the information in social network do not simply distributed around the network. Some people can retweet or mention other random person's tweets even though they do not have any social network connection between them. Hence, it is important to integrate all factors which occur the information flow between users into the result.

Problem 1: Given a network from SNS services, rank the nodes based on their influential power.

Problem 2: The influence of a user cannot be defined directly from the social network; instead, the other factors such as retweet, mention, and reply network should be considered.

Problem 3: Given multiple networks, the method to integrate each network into one final result should be required.

3 RELATED WORK

The project will base on the PageRank algorithm and integrate the recent post count and the interest keyword of each user. Since the basic PageRank algorithm already considers the keyword as an important factor, no further modification would be required.

The social network already has a format similar to the web page link network. Each node will represent the user and the directed edge will be the "follow" relationship between users. Since the edges are directed and each edge has the same importance, the social network is directed and unweighted.

On top of the basic PageRank algorithm, the number of retweets, reply, and mentions will be integrated with each different weights. With those factors, it is possible to measure how much impact the user have.

3.1 Weighted PageRank Algorithm

Xing and Ghorbani introduce the weighted PageRank algorithm [4]. Web structure mining is trying to find the connection by analysis the page structure. Pagerank algorithm is one of most popular algorithm that use in this mining. Weighted Pagerank Algorithm is the expansion of the PageRank algorithm and could identify a larger number of relevant pages to the given query page compare to the standard PageRank. In this project, we do the similar things that give the pagerank with different weighted to find out the connection.

3.2 Competitivity Groups in Social Network

Francisco Pedroche Sánchez's research main idea is using the personalization vector to bias the PangRank to user base on some specific characteristics, such as number of followers and how many activities they have [3]. What they are finding is about what extend cand they bias the pageRank by using the personalization vector. They using two group which is competitivity groups and leadership groups to compare, and create a model to classify the user of SNS by using the personalization vector to bias the node of pageRank. And they believe this model is applicable to the general network.

4 METHODOLOGY

4.1 Dataset

The snapshot of the Twitter network will be used for the project. The dataset includes social network, retweet network, reply network, and mention network graph. The social-network is the only "unweighted" graph, and the other three graphs are "weighted" graphs. The weight in each graph means the count of reply, retweet, and mention accordingly. The Twitter ID or nicknames are not presented in the network data. All user ID are substituted with the randomly assigned numeric ID. The dataset is provided in "edgelist" file, which is a simple text file containing edge data. The node and edge count is provided in Table 1

Table 1: Graph Specification

Network	Node Count	Edge Count
Social	456,626	12,508,436
Retweet	256,491	327,374
Reply	38,918	29,895
Mention	11,6408	145,774

4.2 Hardware Specification

Since the dataset have large number of nodes and edges, the experiment process will require too much memory. To avoid the memory outage during experiment, each graph will be cleared from memory after each PageRank operation. By doing this, the experiment can be completed in 8gb RAM environment without any issue. The CPU performance will also affect the total completion time; however, the lower performance CPU can still complete the experiment without any problem.

4.3 Python

Python version 3.8.10 is used to calculate and manage the result data from the Graph. Using Python, the result data are correctly sorted and outputted to the file.

4.4 NetworkX Library

For the graph manipulation, the NetworkX library for Python will be take in place. The library contains the algorithm to manage the graphs. On top of that, the famous network-related algorithms are already implemented, and users just need to call the existing method to run the algorithm.

4.5 PageRank Algorithm

The basic PageRank algorithm from NetworkX will be applied without any modification in order to identify the influence ranking of the page, just using the "follow" network. The PageRank algorithm in NetworkX can be performed on both weighted or non-weighted graphs. The algorithm returns the dictionary containing each network nodes as key and the PageRank result for each nodes as values. The PageRank algorithm should be run four times because the four graphs will all create each PageRank dictionary and be integrated together for final result.

4.6 PageRank Result Integration

As a result of PageRank algorithm, the four dictionary results will be generated. Each of the dictionaries contain the pagerank results for each Twitter users. However, as shown in Table 3, not all users(nodes) are present in retweet, reply, and mention network. Therefore, for the final result data, the social network PageRank results will be used as a base. Then the other results will take part in the simple integration using custom weight. The focus of this project is to check how the retweet, reply, and mention data affect to the influence ranking of the users. Therefore, the three different kind of weight distribution is used to calculate the final ranking result. There must be a difference in results, and it will be discussed later.

The process to integrate weighted code list is provided in Listing 1. _WEIGHT variables represents weight value for each graphs in final result. This values will be modified for each trials as in Table 3. _pr variables contain the PageRank result from each graphs. For each nodes in the social network PageRank result, the result from other graphs are integrated if it exist in the graph.

4.7 How to Measure Influence

Now we have our three types of weight distribution ready, it is time to find out those weight distribution actually represents that those users have more influence than others.

One way to measure the influence is to use the reply and retweet network. When the statement "The user is influencer" is true, it means that the user's post can easily spread through network. This kind of spread of the post can be seen in reply and retweet network. Since the reply and retweet network from higgs data set already

have the count saved as the weight of edges, we just need to run the independent cascade to measure how much the information can spread through network.

First, we combined the reply and retweet network to be one graph, with its edge weight is equal to the sum of edge weight from reply and retweet network. The resulting graph specification is described in Table 2. Using this new graph, for each node, we should run the independent cascade model in order to get how much will the user post will spread through the network.

One major problem rises here. Since the network have too many nodes inside, it is not easy to run the whole cascade process for each node in reasonable time space. Hence, we decided to use top 100 nodes from the PageRank result of the followship network. These nodes are chosen because the nodes less than rank 100 had relatively low PageRank values.

Using the independent cascade model, we iterate the model until it stops spreading. Then we counted the infected nodes and used it to sort out the nodes. This result can provide us some expectation of the information flow through the Twitter network, and we can compare this result to our PageRank algorithm result.

Table 2: Retweet+Reply Graph Specification

Network	Node Count	Edge Count
Reply+Retweet	352,632	274,063

Table 3: Weight Distribution Types

Type	Social	Mention	Retweet	Reply
1	0.6	0.1	0.1	0.2
2	0.5	0.15	0.15	0.2
3	0.4	0.2	0.2	0.2

Listing 1: PageRank Inetegration

```
SN_WEIGHT  = 0.4 # social network weight
MN_WEIGHT  = 0.2 # mention network weight
RTN_WEIGHT = 0.2 # retweet network weight
RPN_WEIGHT = 0.2 # reply network weight

for key in SN_pr:
    SN_pr[key] *= SN_WEIGHT
    if key in MN_pr:
        SN_pr[key] += (MN_pr[key] * MN_WEIGHT)
    if key in RPN_pr:
        SN_pr[key] += (RPN_pr[key] * RPN_WEIGHT)
    if key in RTN_pr:
        SN_pr[key] += (RTN_pr[key] * RTN_WEIGHT)
```

5 EVALUATION

5.1 Datasets

For the analysis, we chose to use a dataset called "Higgs Twitter Dataset"[2] provided by Stanford University. The dataset has a total number of 456,626 users and 14,855,842 followerships. Initially, the dataset consists of five separate graphs. The social network graph represents all users as nodes and their followerships as directed edge. Then, the other four record different types of interactions among the users. By calculating the influence level of nodes with the modified version of the PageRank algorithm, we will integrate these graphs into a user-to-user graph where each node has a value representing its influence level.

5.2 Statistics

Table 4 shows the IDs of the top 10 nodes of high centrality for each network from the dataset. It is interesting to see that some nodes are placed in the top 10 for all the networks. For example, Node 88 and Node 677 are both placed in the top 10 for all the networks. However, if we look at Node 1503, it is the node with the highest degree centrality for Social Network, but it is placed nowhere in the top 10 for all the other networks. As we proceed to the further experiments and analyses, we will discover more potential clues like this, and look into it in detail to find any correlations among the networks.

Table 4: Top 10 Nodes of Centrality

Rank	Social	Mention	Retweet	Reply
1	1503	88	88	677
2	206	677	14454	88
3	88	2417	677	220
4	138	59195	1988	3549
5	1062	3998	349	317
6	677	7533	283	349
7	352	383	3571	1988
8	220	1988	6948	7690
9	317	13813	14572	3369
10	301	519	68278	16460

5.3 Experiment Setup

All experiments will be conducted on a PC with Intel(R) Core(TM) i5-10400CPU @ 2.90GHz 2.90GHz and 16 GB RAM with OS Windows 10 Pro x64, using Python 3.9.7, NetworkX Library, and some other basic tools for snapping the experiment results. If the experiments were conducted on dynamic data, multiple trials of the same experiment would help us increase accuracy. In our case, we will not conduct the same experiment multiple times as the results would be identical due to the static dataset.

5.4 Experiments & Analyses

Degree Centrality, used to calculate the rankings for Table 4, could be a good measure of connection levels of nodes, but higher degree centrality does not necessarily indicate a higher influence level because connections have different levels of importance. Therefore, we generated Table 5 using the PageRank algorithm which counts the importance of connections in the calculations. As expected, the nodes in the results are quite different from the ones in the rankings by degree centrality. What is notable though is that Mention, Retweet, and Reply networks share many nodes in common but overall, Social network has quite different rankings from the other interaction networks. From this, we learned that the number of

followers and the number of interactions are not always proportional. For example, node #1 has the most followers but it's not listed anywhere anywhere in the interactions ranking.

Influence levels calculated using just the number of followers can possibly include influencers who are not active at the time. We can filter these influencers and extract the list of active influencers by considering interaction factors. For this matter, we integrated the factors using different ratios and then visualized how the top 10 influencers based on the followership are distributed in the rankings of the integrated versions as figures below.

Table 5: Top 10 Nodes of PageRank Scores

Rank	Social	Mention	Retweet	Reply
1	1	88	88	677
2	88	677	14454	88
3	1503	2417	677	220
4	138	59195	1988	3549
5	220	7533	283	317
6	317	383	349	349
7	206	3998	68278	3369
8	352	1988	6948	7690
9	667	3369	3571	1988
10	301	11792	3549	16460

Each figure has the top 10 nodes of the PageRank scores by followership on the left and a respective versions with the weight distribution types mentioned in Table 3 on the right. If we look at Figure 1, most of the nodes are common in both tables since the integrated version still put 60% weight on the followership. As a result, there are only two nodes who are not ranked in the top 10 of the integrated version. However, as we put more weight on the other factors in Figure 2, more than half of the top 10 nodes on the left are ranked outside the top 10 on the right for both type 2 and type 3 integrations. When we consider these interactions as the spread of information, this experiment implies that the influence level does not always grow in proportion to the followership.

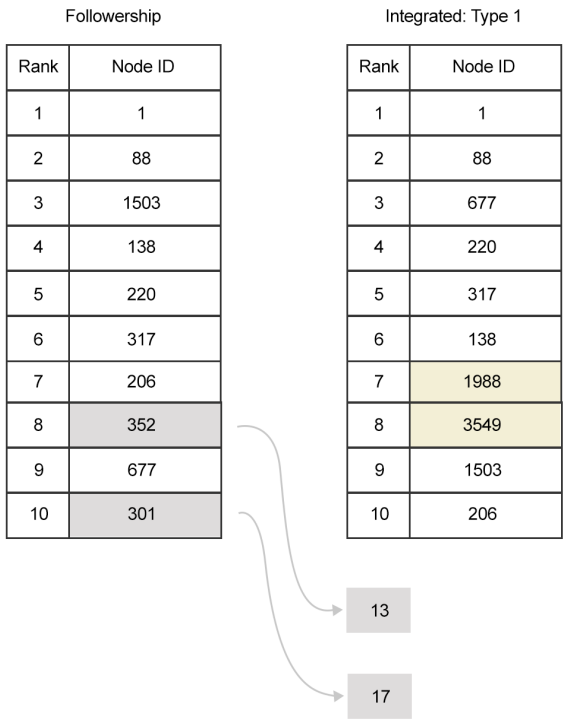

Figure 1: Distribution of Top 10 Nodes in Type 1 Integration

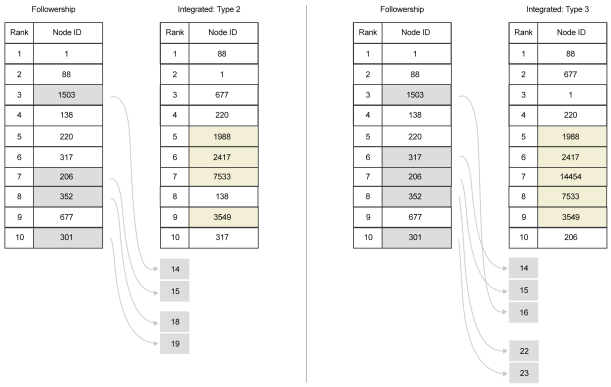

Figure 2: Distribution of Top 10 Nodes in Type 2 & 3 Integrations

5.5 Independent Cascade Result

As the last stage of our experiments, we generated an independent cascade model based on the integration of the original reply and retweet networks. The independent cascade model represents the

information flows over the network, and our goal was to use this model to calculate the influence levels of the top 100 nodes of the PageRank results for the social network so that we could compare to the influence rankings based on the weight distributions and find a correlation. Since the cascade model produced different result every time we ran, we standardized the rankings by calculating the means of the multiple results. Below, Table 6 shows the top 20 nodes by the cascade model and the comparison shows that each ranking based on type 1, type 2, and type 3 weight distribution respectively had 35%, 40%, and 45% of their top 20 nodes listed in the top 20 of the ranking by the cascade model. This might not be 100% correct result since there were random factors in the process which could bring some changes to the result each time. However, it is good enough to confirm that the experiments were heading in the right direction to a certain extent.

Table 6: Top 20 Nodes of Influence levels By Independent Cascade Mode

Rank	Node ID	Rank	Node ID
1	6948	11	407
2	1988	12	283
3	4214	13	2670
4	1267	14	1274
5	327	15	349
6	1	16	6951
7	308	17	34719
8	352	18	68278
9	960	19	31957
10	5226	20	2662

5.6 Visualization

There are various social network analysis and visualization software such as Gephi, R igraph and kuma, all of them could help us to make the data easy to see. For intuitive navigation, we will sequentially provide both the initial graph of the network and the result graph of the integration process which has its base in the PageRank algorithm by using one of the visualization software. In this way, we will be able to show, in a clear manner, the difference between influence levels of the users measured simply by the number of followers and the ones measured by the algorithm which counts in other valid factors for higher precision. In addition to it, we will present a diagram which consists of tables of nodes and PageRank scores of the networks, and arrows from the top 10 nodes of a table to their rankings in the other tables. The visualization will present distribution of the rankings of the top influencers of a network in the other network. This will help us observe how different the results can be depending on the factors used for the calculation of the rankings.

6 CONCLUSION

The result of the experiment showed us that using just the friendship network is not the perfect way to measure the influencer ranking. The PageRank result from follow, mention, reply, and retweet was similar, but obviously all different. Some user was not shown in follow ranking, but shown in reply ranking. It is not yet clear which type of weight distribution is the best option to choose based on this result. One important aspect we can conclude is that the weighted PageRank result would be much accurate than using just the followership graph.

Throughout the project, we looked into social network graphs and observed that influencer rankings calculated simply using a single factor "followerships" could bring us divergent rankings from the actual influence levels of users. In addition, using other factors which represent the interactions between users, we were able to calculate more accurate influencer rankings which also reflect their activeness. In the experiments, we repeatedly observed that the number of followers are not always proportional to the actual influence level. Furthermore, this gave us an idea that followership could actually be used to filter out fake or inactive influencers. A possible future work could be about working on the weight distributions for calculation. For viral marketing where the spread of information via sharing feature is the key, a higher weight on retweet factor or anything equivalent would be ideal to find the most fitting influencers. Finding ideal weight ratios of factors for the calculation per scenario could be a complicated and expansive task to be done.

REFERENCES

[1] R. S. Achrol and P Kotler. 2012. Frontiers of the marketing paradigm in the third millennium. *Journal of the Academy of Marketing Science* 50, 1 (2012), 35–52. https://doi.org/10.1007/s11747-011-0255-4

[2] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.

[3] F. P. Sánchez. 2010. Competitivity groups on social network sites. *Mathematical and Computer Modelling* 52, 7-8 (2010), 1052–1057. https://doi.org/10.1016/j.mcm.2010.02.031

[4] W. Xing and A. Ghorbani. 2004. Weighted PageRank algorithm. *Proceedings. Second Annual Conference on Communication Networks and Services Research* (2004), 305–314. https://doi.org/10.1109/DNSR.2004.1344743