HYBRID METHODS FOR RECOMMENDING CONNECTIONS IN ONLINE SOCIAL

Ruth Olimpia García Gavilanes

MASTER THESIS / 2010

Dr.Xavier Amatriain

Research Scientist (Telefonica I+D)

UNIVERSITAT POMPEU FABRA

ACKNOWLEDGMENTS

I want to thank my tutor Xavier Amatriain for all the patience and help provided even at very late hours and for giving me the opportunity to do an internship in Telefonica where I have learned a lot. I also want to thank *Fundación Carolina* for giving me a scholarship to pursue my Masters, without which this research could not have been possible. Finally, I want to thank God and my parents for the support.

TABLE OF CONTENTS

HYBRID METHODS FOR RECOMMENDING CONNECTIONS IN ONLINE SOCIAL NETWORKS

Ruth García Gavilanes
UPF
Barcelona-España

Dr. Xavier Amatriain
Telefonica – UPF
Barcelona-España

1. ABSTRACT

Online Social Networks have an important role in the life of millions of active internet users and an even stronger growth and penetration is expected in the following years. Nowadays more than 60% of active Internet users read blogs and more than 50% leave comments. Twitter is capturing a lot of attention and participation from people and it is becoming a strong source to spread real time information. For this reason it is important to find ways to provide recommendations for users and help them find people that they might be interested to follow.

This research carries out in-depth analysis of a large dataset to study the behavior and preferences of users. It focuses on reciprocal connections to study how their popularity influences good recommendations. In this regard, tests are implemented to observe this influence. This thesis also suggests studying the importance of other features in reciprocal connections such as activity, location, socialgraph and content of tweets in recommendations and provides algorithms to be used in future tests. The results of this paper can and should be complemented with more research.

2. INTRODUCTION

One of the main functionalities of the Internet, since its very first appearance, is to connect people easily and fast. It started with e-mails but then it expanded to social networks on the web. The emergence of online social networks is the response to the invisibility problem of "social networks" in the real world. Without the Internet, social connections are hidden and can even be lost due to distance, lack of communication and time, regardless its potential. The creation of profiles containing pictures and personal/public information allow seeing connections at the touch of a mouse click and therefore it has been massively adopted by people in the world.

Nowadays, Online Social Networks have an important role in the life of millions of active internet users and an even stronger growth and penetration is expected in the following years. According to Universal MacCann (Wave 4)[1], 62.5% of Active Internet Users in the world belong to at least one O.S.N by March of 2009 compared to the 57% in the preovious year. From this first group, 56.4% answered that they try to "Find New Friends" in these networks.

[1] (Mccam, Oct , 2009)

Users of O.S.N share many kinds of information with their friends, from videos to blogs, pictures, etc. An interesting factor is that by 2009 more than 60% of users read blogs and more than 50% leave comments.

Universal MacCann suggest that in the future Social media and Social Networks will become synonymous due to the information sharing activities happening in O.S.N.

Twitter is an example of these changes since this site has been capturing major attention and its use has exploded in many countries of the world. In contrast to networks such as Facebook, Orkut and Myspace where privacy is a major issue and people aim to share information only with a closed group of people, Twitter in the majority of the times shares information with everybody by means of status updates of 140 characters long maximum. People have learned and continue learning to use markup signs to help categorizing tweets: RT stands for retweet, '@' followed by the Twitter log-in name identifies the user, and # followed by a word represents a hashtag to show the topic which the tweet is referred to.

Another difference from other common social networks is that in Twitter there is no need of reciprocation. Individuals can *follow* others and be *followed* by others with no need of reciprocation. If a user follows someone the tweets of their *followings* will appear in their walls. In this way, a normal Twitter user has followers, followings and reciprocal connections (followings and followers at the same time appearing both in the list of followers and followings).

Moreover, another interesting feature of twitter is that information is being uploaded and changed constantly from several places in the world and millions can have public and free access to the stream of tweets created every day in the so called *public timeline.* This allows the fast spread of very new information. In fact, Asur and Hubernan from HP labs[2] showed that Twitter can be used to predict the future box-office takings of blockbuster films. In another study[3], Kwak et al. suggests that Twitter is not properly a social network since it lacks reciprocity but instead it is seen more as a news medium. This possibility opens the door to interesting marketing strategies and it represents the microphone of important social events such as the protests in Iran and Cuba and source of information for tragedies such as the earthquake in Chile in 2010. Nevertheless, the majority of tweets seem not to be valuable according to PearAnalytics but despite those findings the valuable information is certainly significant since several major search engines are including tweets as search results[4].

Given all the importance of this micro-blogging site, it is appropriate to find ways to recommend interesting connections to the user. Much has been done already in this area. For example, people have found methods to rank people in twitter while others have focused in avoiding spammers or assessing the quality and diffusion speed of information. For this reason, this study aims to contribute to these methods by exploring a new way of recommendations: the possibility of recommending people based only on their reciprocal connections.

[2] (Asur, et al., 2010)

[3] (Kwak, et al., 2010)

[4] (Gayo-Avello, 2010)

This research uses Twitter as a basis to analyze how reciprocal connections influence recommendations. This paper focuses on the importance of their popularity in recommendations. The author also implements tests based on the popularity of these links to predict the likelihood of two pairs being reciprocally connected. Activity, location, socialgraph and content of tweets in recommendations are also suggested to be studied in future research. After reviewing several papers, this study seems to be the only one combining all these features based on the reciprocal connections of users to provide recommendations.

For the research a dataset comprising a sample of 2,476 users with 815,554 reciprocal connections was provided. The profile details of these users are found in a document comprising 455,000 profiles. The dataset also has 27 million tweets from around 281,825 users but these last data was not explored in this study.

This paper is organized as follows. First, it summarizes the algorithms that the author thinks are the most relevant rank prestige algorithms then several studies on Twitter and some of their contributions are mentioned. Then, the research questions and objectives are listed. After that, general information such as terminology and database are explained. Next, the dataset analysis is presented by means of several Figures and Statistics but no analysis is made for the content of tweets. Afterwards, the evaluation method adopted for the tests is explained. Subsequently, all the classifiers are presented. For the case of popularity the way classifiers are calculated is presented as well as the results of the tests. For the case of activity, location, social graph and content, the algorithms are introduced to be implemented in the future. Finally, the results are discussed and recommendations for future research are proposed.

3. MOTIVATION

Several attempts, as the next section will show, have been done to rank users of Twitter efficiently so that users that provide "valuable information" are ranked first and spammers or heavily advertisers ranked last. Nevertheless, if these rankings are considered for recommendations, the top results will be based on the "elite" of users, that is, those active users that have many valuable followers and less followings. But would all people in Twitter prefer to follow users of the elite? Are there users looking for people that are not in the elite but that somehow suit better their information needs?

The author believes that much can be learned about user preferences from their reciprocal connections and therefore this information should be used in recommendations. It is true that spammers and heavy advertisers look for reciprocal connections to increase their rankings and sometimes they can massively un-follow people to increase even more their popularity. However, active responsible users aiming to find interesting information in twitter will not be likely to follow these spammers and advertisers and if they do for "politeness" the tendency is to unfollow them once they realize they are filling their walls with tweets that are of no interest. On the other hand, the reciprocal connections of heavy advertisers will likely be other heavy advertisers.

Studies have found homopholy in Twitter network showing that users "engaged in a social activity seem to be associated more closely with ones who are similar to them along a certain dimension such as location, age, political view or organization affiliation, compared to ones who are dissimilar.[5]" This shows that despite the lack of reciprocity in Twitter there is a tendency for people to get together with similar people (similar in different aspects). Thus this study aims to explore this homopholy by studying how different contextual features in reciprocal connections can be used for recommendations.

In this regard, the **objectives** of this paper are:

1. Study the dataset to find interesting patters about the behavior and preferences of users
2. Propose different ways to calculate popularity in Twitter
3. Observe how similarities in the popularity level among reciprocal connections can be used to predict connections
4. Suggest ways to test the role of activity, location, socialgraph and content in the prediction of connections

This study does not aim to propose a "good" or "better" way to recommend people in Twitter but it rather tries to explore new possibilities that could be helpful in future research.

4. RELATED WORK

Since the appearance of Twitter, researchers have been very interested in the study of its contents and network structure. The possibility of accessing millions of profiles and contents without restrictions of privacy provided enough data and means to run several experiments, develop applications and write many papers especially concerning Recommendations.

It is now an accepted belief that the number of followers does not determine the *prestige* or popularity of a twitter user since spammers and/or aggressive marketers have ways to have a lot of followers and some of them even more followers than followings[6]; therefore, many studies have been trying to determine effective ways to rank users and provide better recommendations for other users and even for search engines like Google[7] . There are several approaches to rank users varying from considering followers and followings, location, centrality, prestige, content, hash tags and retweets. Some of the most important applications regarding Twitter and recommendations are presented below and are based on the order of Gayo-Avello's literature review8 and other sources.

[5] (De Choudhury, et al., 2010)

[6] (Gayo-Avello, 2010)

[7] (Gayo-Avello, 2010)

[8] (Gayo-Avello, 2010)

4.1. PAGERANK

Pagerank is considered as one of the most important methods to rank prestige because it is based on Google search Engine. Pagerank ranks higher those pages that are heavily linked and that receive links from relevant pages. According to Sobek [9], the basic version of pagerank is the following (it differs a little from Gayo-Avello):

$$\boldsymbol{PR(A) = (1 - d) + d\left(\frac{PR(T1)}{C(T1)} + \cdots \ldots + \frac{PR(Tn)}{C(Tn)}\right)} \qquad \text{(Formula 1)}$$

Where

PR (A) is the PageRank of page A,

PR(Ti) is the PageRank of pages Ti which link to page A,

C(Ti) is the number of outbound links on page Ti and

d is a damping factor which can be set between 0 and 1 (calculated as 0.85 in this case).

Although this formula has been changed several times, it can be used to rank twitter users when making tests. The Formula is implemented in several iterations until it converges.

For a better understanding of this algorithm where examples and clear explanations are found I suggest visiting Sobek webpage.

4.2. HITS

Hyperlink-Induced Topic Search – HITS is also used to calculate the relevance of a document. It assumes that the web is made of hubs and authorities. Authorities are heavily linked documents and hubs are documents linking to several authorities.

In this regard, all the documents will have a hub and authority score but this score is suitable only for query “dependent subgraphs composed of those documents already satisfying a given query.” Nevertheless, when HITS is applied to Twitter in Gayo-Avello’s study, he applies it to his complete Twitter dataset.

HITS values are also initialized (i.e 1) and iterated until it converges. The corresponding formula is the following:

[9] (Sobek, 2002/2003)

$\boldsymbol{aut(p)} = \sum_{\boldsymbol{q:(q,p)\epsilon E}} \boldsymbol{hub(q)}$ -sum of hub scores of pages q linking to p- (Formula 2)

$\boldsymbol{hub(p)} = \sum_{\boldsymbol{q:(q,p)\epsilon E}} \boldsymbol{auth(q)}$ - sum of the authority weights for pages q linked from p (Formula 3)

p= give web page

q= another web page different than p

E= set of edges in the graph

4.3. TUNKRANK

Tunkrank is a very interesting and simple algorithm created by Daniel Tunkelang and implemented by Jason Adams[10] as a Web App. Adams considers two aspects in this algorithm:

a) The amount of attention you can give is spread out among all those you follow. The more you follow, the less attention you can give each one.
b) Your influence depends on the amount of attention your followers can give you.

Likewise, in his blog, Tunkelang specifies the algorithm as[11]:

$$Influence\ (X) = \sum_{Y\varepsilon Followers\ \ (X)} \frac{1+p*Influence\ \ (Y)}{|Following\ \ (Y)|} \qquad \text{(Formula 4)}$$

- Influence(X) = Expected number of people who will read a tweet that X tweets, including all retweets of that tweet. For simplicity, it is assumed that if a person reads the same message twice (because of retweets), both readings count.
- If X is a member of Followers(Y), then there is a 1/||Following(X)|| probability that X will read a tweet posted by Y, where Following(X) is the set of people that X follows.

If X reads a tweet from Y, there's a constant probability p that X will retweet it. So the final formula comes down to:

The recursion is infinite over a graph with directed cycles, but rapidly converges as high powers of *p* approach zero.

The implementation of this algorithm can be found in the web (tunkrank.com).

[10] (Adams, 2009)

[11] (Tunkelang, 2009)

4.4. TWITTERRANK

Twitterrank[12] is a more complex algorithm proposed that is based on PageRank as well. It takes into account the similarity of users based on the content of their tweets. In that sense, users are ranked according to the topics of their tweets.

It is considered a complicated algorithm because "the transition probability among connected users heavily relies in both the topical similarity between users and the number of tweets published not only by the followee [called following in this paper] but by all the followees the follower is connected to". [13]

Gayo-Avello implemented this algorithm changing some variables for sake of simplicity; the results obtained were disappointing although the conceptually appealing methodology used suggested better results. In the conclusions, Gayo-Avello said that this algorithm is 1) more computationally expensive and 2) dependent of much more data than other methods.

After analyzing the results and reading the corresponding study on TwitterRank by Weng et al, I have decided not to specify the algorithm.

4.5. OTHER STUDIES

There are many studies analyzing different aspects of twitter. For example, De Choudhury et al. [14] observe how different sampling methods can influence the level of diffusion of information. They found that sampling techniques incorporating context (activity or location) and topology have better diffusion than only considering context or topology. They also observed the presence of homophily showing that users get together with "similar" users but that the diffusion of tweets will also depend of the themes. They see reciprocal connections as a more useful way to avoid spammers in the context of Twitter because they believe that these links are more robust to spam "a normal user is less likely to follow a spam-like account."

On the other hand, Cha et al.[15] proved that influence is in fact not related to the number of followers. After analyzing 1.7 billion tweets from 54 million users, they claim that it is more influential to have an active audience who retweets or mentions the user. They showed that the most influential users can have significant influence over a variety of topics but that in order to acquire that level of influence effort and great personal involvement is needed. They concluded that it is probable that influential users can be more predictable than expected by other theories.

[12] (Weng, et al., 2010)

[13] (Gayo-Avello, 2010)

[14] (De Choudhury, et al., 2010)

[15] (Cha, et al., 2010)

Furthermore, Kwak et al.[16] analyze 1.47 billions social relations, 4,262 trending topics and 106 million tweets and found that there are low reciprocity among users but that homopholy is present for those who reciprocate. They also analyzed the ranking of Page rank and in-degree number (number of followers) and found the rankings to be similar; however, ranking users by retweets shows very different results than those found in in-degree. They showed that independently of the number of followers, once a tweet is forwarded the tweet reaches an average of 1000 users. It is interesting to notice that some of the results found in the analysis of their dataset are very similar and sometimes almost identical to some of the results found in this study although the datasets were collected during different times and by different people.

Finally an interesting study made by Daniel Gayo-Avello[17] analyzed the performance in detecting spam and abusive marketeres by several ranking algorithms. He also proposed a discounted version of Pagerank that did not consider reciprocal connections if the ratio followers/followings <0, he argued that separating reciprocal links is a good way to separate those users contributing valuable contents to the global system from those with little to no value at all. Opposite to Choudhury et al. [18], he does not see reciprocal connections as a way to avoid spammers. After performing tests, he concludes that TunkRank seems to be the best option but that his proposal needs further research since it offers a very distinctive curve and ranks spammers with lower grades than the other methods.

5. GENERAL INFORMATION

5.1. TERMINOLOGY

For this paper I used the following terminology, some of it is based on the Twitter Glossary Page[19].

Tweet : Status update of 140 words maximum.

***Tweeter*:** A person who uses twitter also called user.

Follower: A follower is a tweeter who has chosen to follow you by subscribing on your tweets updates.

Following: A following is a tweeter who you have chosen to follow by subscribing on his/her tweets updates on the site.

[16] (Kwak, et al., 2010)

[17] (Gayo-Avello, 2010)

[18] (De Choudhury, et al., 2010)

[19] (Twitter, 2010)

Reciprocal connections (R.C): term used to represent a following who is also your follower. They are also called Reciprocal pairs.

Popularity Index (P.I): A P.I of a Tweeter is the value *followers-followings*. For each user in the testing set, the most efficient P.I possible will be found using the training set of the corresponding user.

5.2. TOOLS

MySql 5.1.36 and PHP 5.3.0 (WAMP SERVER 2.0) and Matlab 2007, FILEZILLA

5.3. DATASET

5.3.1. Original Information

The dataset was taken from the Web site of PHD student Munmon de Choudhury.[20] According to her, Tweets were collected between ***2006 and 2009***, the dataset used in this study comprised a set of about 27 million tweets from around 281,825 users and provided details of around 455,000 profiles and a sample of reciprocal connections for 2,476 users with 815,554 links (details in the following sections). In total, three files were available: user profile, social graph and tweets content.

User Profile
According to de Choudhury, the information provided for each user was: tweeter name, time zone, status count, favorite count, followers and followings count.

Social graph
In a separate file, a sample of the social graph of users was taken and a list of 2-tuples made of tweeters names was provided.
The names in the tuples are found in the usertable and they represented reciprocal connections, that is, connections with those users who are followers and followings at the same time. The number of connections for a single user varied from 1 to 500.

Tweet Content
The last file contained the information of the tweets. Each entry contained the name of the tweeter, the content of the tweet and the time when it was posted.

5.3.2. Fixes and Complementary Information

[20] (Choudhury, 2006 - 2010)

After migrating the data to Mysql, there were three tables: user, friend and tweet. I found some inconsistencies between the user profile and the socialgraph table:

Fixes

a) Repetitions: there were **626** repetitions of tuples. All repetitions repeated the same tuple at most 2 times. I fixed the problem so that every pair appears only once.

b) Inverses: in certain cases, tuples in the socialgraph did not have their corresponding inverse. Since I know that if User B is follower and following of User A then the inverse should also be true. I fixed the problem by providing the missing inverses but the resulting table was stored in a separate table in order to keep the original information. The name of the generated table is ***completefriend***

c) Inconsistence: On the social graph table there were users who had more reciprocal connections than followers or followings. I found **1,754** tweeters with consistent amount of reciprocal connections (reciprocal connections <= #followings and #followers) summing a total of 632,783 links and ***722*** tweeters with inconsistent amount of reciprocal connections (reciprocal connections >#followings or #followers) summing a total of 182,771 links. This problem was not fixed because I did not know what to consider as true, the number of tuples in the socialgraph or the number of followers and followings in the user profile. After consulting with de Choudhury, I concluded that changes in the socialgraph may have occurred during the collection of data

Complements

a) ***Sample*** : A sample taken from the table ***completefriend*** made of around 20,000 reciprocal connections was taken in order to execute tests without demanding too much processing and memory use. The training set and the testing set are obtained from this table.

b) ***NOuserAfriend :*** It is all the possible Nonreciprocal connections among the users of the table Sample. Every time a Sample table is generated a NoUserAfriend table is also generated. Right after a testing set is generated from the Sample table, the same amount of nonreciprocal pairs is also inserted.

c) ***Results:*** This table stores the results of the Classifiers performance , it records the name of the classifier testes, the average number of recommendations, the median number of recommendations, the recall, the precision and the f-measure .

5.3.3. Resulting DataBase

After the fixes and complements, the following database was generated:

Resulting Database

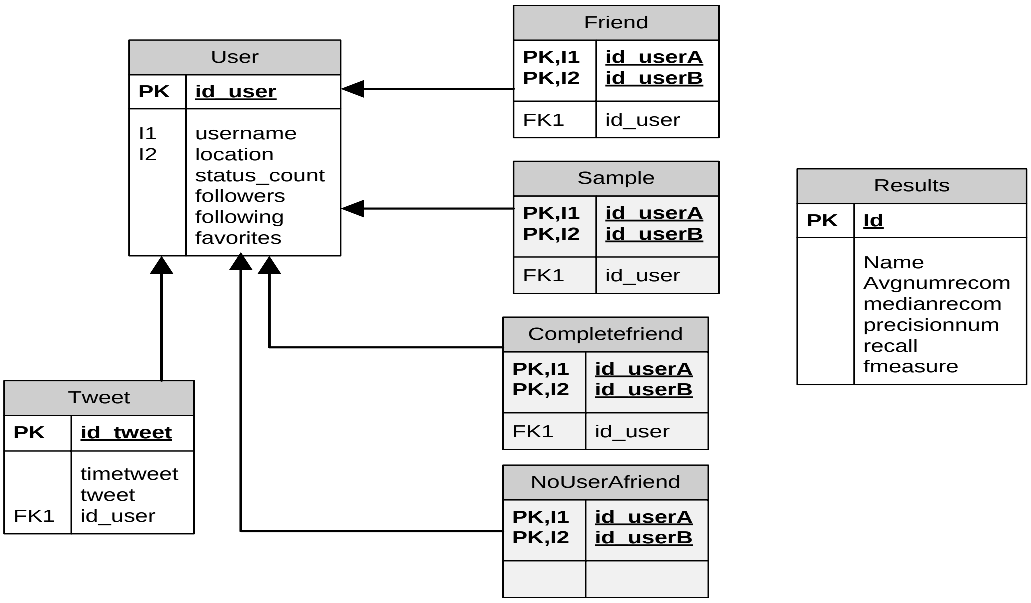


Figure 1 : Resulting database of Twitter dataset after fixes and complements

5.3.4. Database

The database content for each table is the following

Table name : property	**Number**
User : number of users	456,107
Friend : id_userA	2,476
Friend : Reciprocal connections	815,554
Completefriend: id_userA	2,476
Completefriend: Reciprocal connections	821,967
Sample : unique id_userA	61
Sample: Reciprocal connections	20,180
NOUserAfriend : id_userA	2,475
NOUserAfriend : Reciprocal connections	10,230
Tweet : Number of tweets	26,675,204
Tweet : Users who posted tweets in the sample	281,825

Table 1: Contents of database for tables User, Friend, Completefriend, UserAfriend, NOuserAfriend and tweet

6. DATASET ANALYSIS

6.1. USER PROFILE

In the first part of this study, I will analyze the USER profile in terms of the following features:

- Number of tweets
- Number of Followers and Followings
- Popularity Index Followers -Followings
- Location
-

6.1.1. Followers and Followings

Figure 2 shows similar findings of paper Kwak & Lee paper,[21] there are two glitches, the first one appears around x=20 and the other around x=2000. According to Kwak & Lee, this is explained because Twitter suggests 20 followings at the moment of registration and because after 2009 the limit of 2000 followings was eliminated.

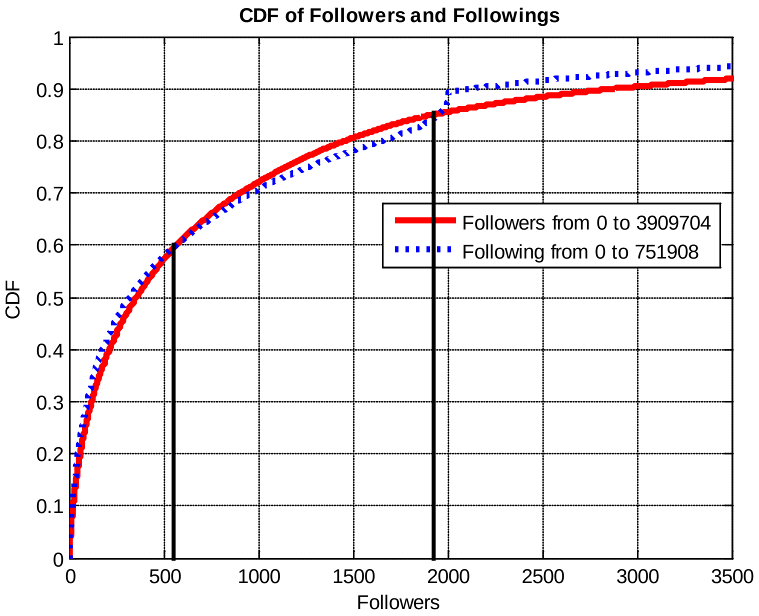


Figure 2: CDF of Followers and Followings for all tweeters in USER PROFILE. X scale from 0 to 3000 in order to have a better view of Graph

[21] (Kwak & Lee, 2010)

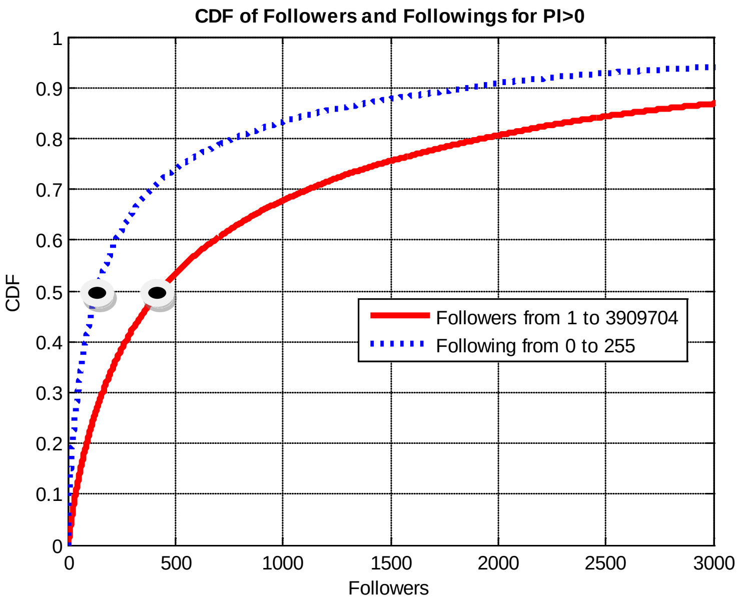


Figure 3: CDF of Followers and Followings for user with P.I >0. X scale from 0 to 3000 in order to have a better view of Graph

In figure Figure 2, around 50% of users have less than ***400 followers and followings,*** whereas in Figure 3, 50% of users have less than 400 followers but approx. 50% have less than 200 followings. I can see three interesting stages in this graph.

a) For numbers less than 600, there are more users having less number of followings than followers (i.e 30% have less than 100 followings but 30% have less than 120 followers).

b) For numbers between 600 and 2000, the are more users having less followers than followings (i.e 80% have less than 1460 followers but 80% have less than 1655 followers)

c) For numbers higher than 2000, there are more users having less number of followings than followers. The maximum number of followings in the sample is 751,908 whereas the maximum number of followers is 3,909,704

Notice in Figure 2 that maximum value of followings is much smaller than the maximum number of followers. There are *30 people* having more followers that the max number of followings varying from 751908 to 3909704. Probably this is because of many celebrities who tend to have many followers without proportional reciprocation. I will analyze those cases later.

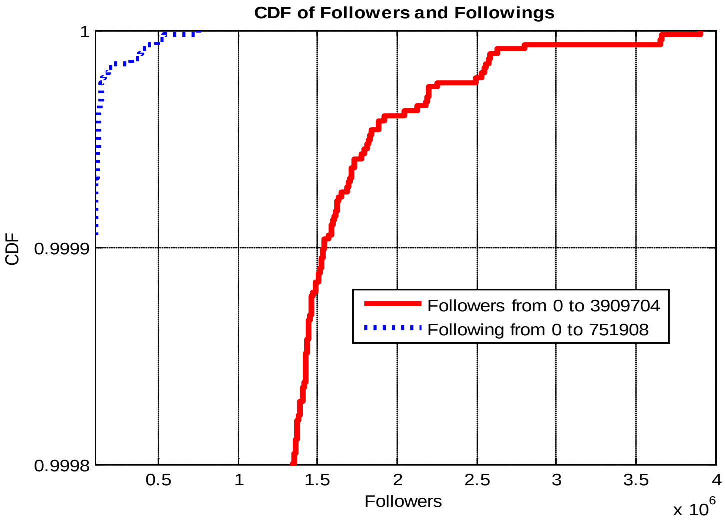


Figure 4: Tail of CDF of Followers and Followings showing the maximum numbers for each case

6.1.2. POPULARITY INDEX : followers - followings

I face the following cases for P.I:

a) Followers with larger number of followings
b) Followers with equal number of followings
c) Followers with less number of followings

The difference followers-followings determines who goes to which group. Ignoring cases where followers = 0 or null, I arrived to **454,959** tweeters to be analyzed:

a) Followers - followings > 0 = Tweeters with larger number of followers than followings (188,311 users)
b) Followers- followings = 0 Followers with equal number of followings (3,493 users)
c) Followers - followings < 0 Followers with less number of followings (263,155 users)

This index will be considered at the moment of recommending connections based on Popularity Index.

6.1.2.1. Calculation of Popularity Index

To better illustrate the importance of considering a Popularity index, let´s look at the following graphs.

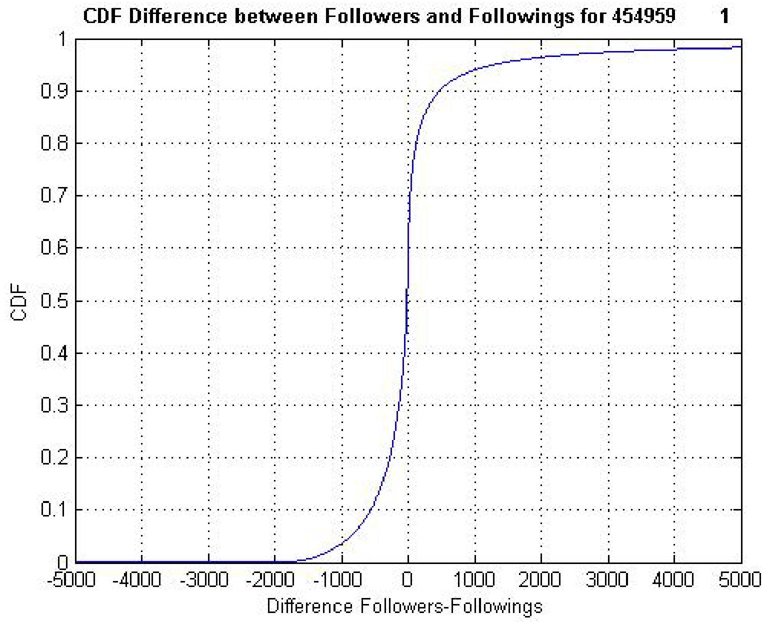


Figure 5: CDF of Popularity Index Followers-Followings, the larger the index the better the recommendation.

Figure 5 shows that around 60% of users have a P.I less than 0 meaning that there are more people with more followings than followers. The lowest difference is **-43,160**. On the other hand, the maximum P.I index is **3,909,449**.

Figure 6 shows that around 50% of users having P.I >0 have a ***P.I > 100*** and less than 15% have >1000 whereas Figure 7 shows that 50% of users with P.I <0 have around ***P.I >- 150*** but less than 10% have a P.I < -1000**.**

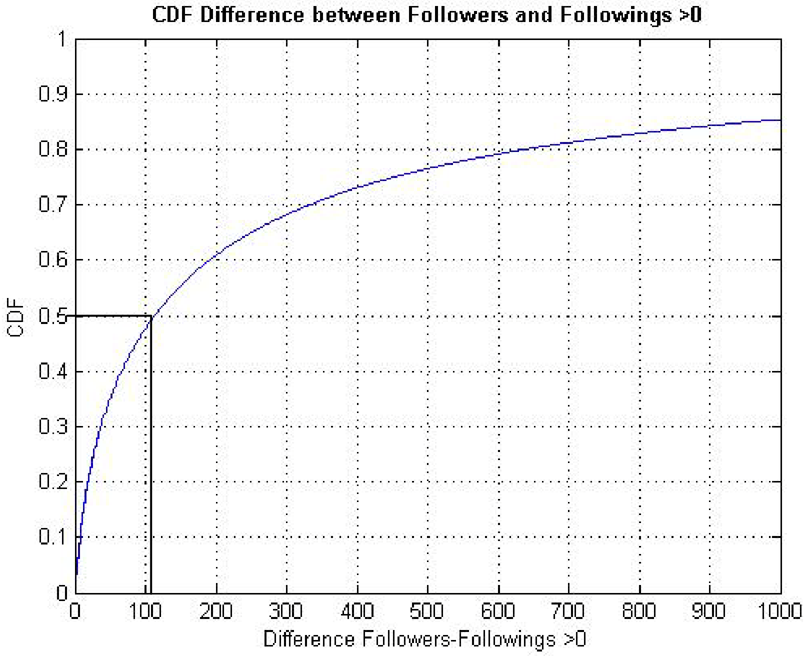


Figure 6: CDF of P.I> 0 users. Graph shows Cumulative Distribution for users with Followers > Followings.

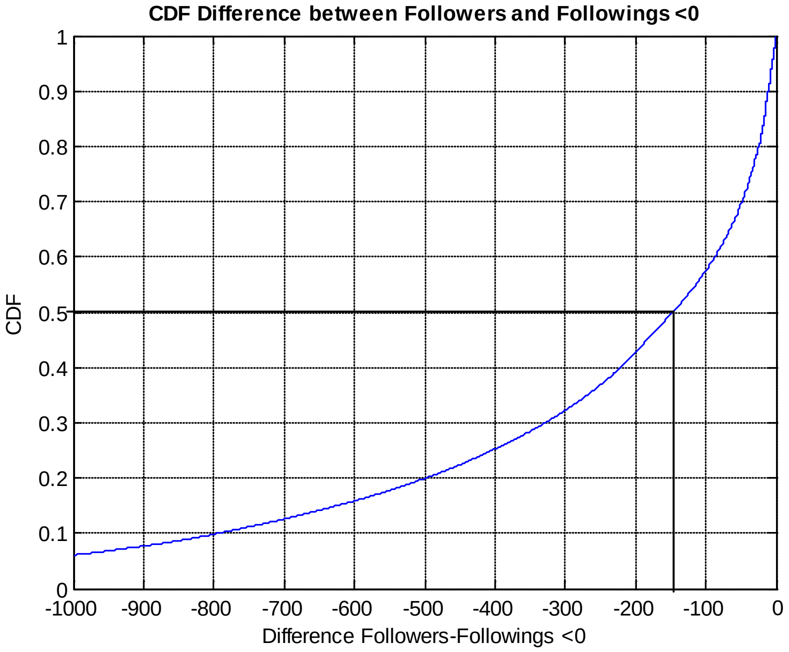


Figure 7: CDF of P.I< 0 users. Graph shows Cumulative Distribution for users with Followers < Followings.

Figure 8 and Figure 9 shows the distribution of P.I for users with different amount of followers. Figure 8 only shows cases with P.I > 0, in this figure I can see an increasing P.I as the number of followers increases; however, from 1 to more or less 10,000 followers, there are big variances in P.I for the same amount of followers (from 1 to nearly 10,000). These big variances seem to decrease at the tail of the Figure, showing a more uniform distribution. On the other hand, Figure 9 shows a less uniform distribution even at the tail. Users with very low P.I will be expected to be spammers.

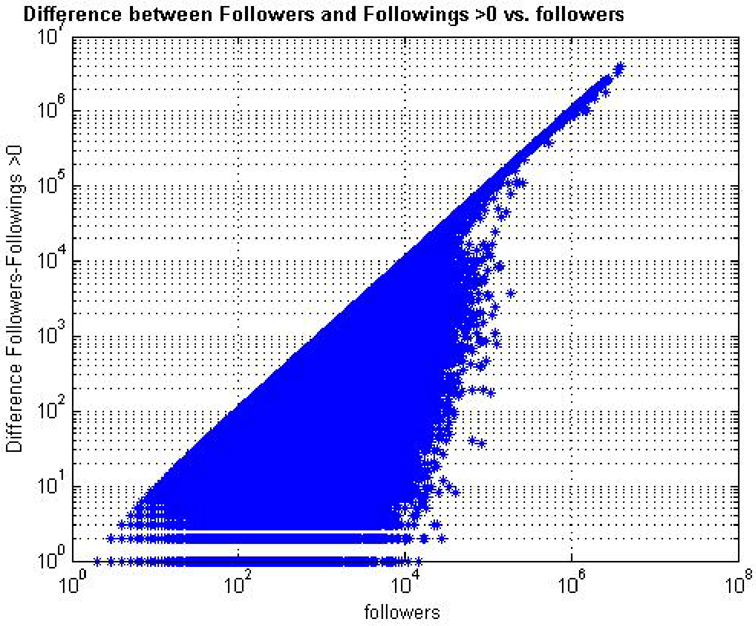


Figure 8: Log log Figure of P.I vs Followers considering only P.I > 0

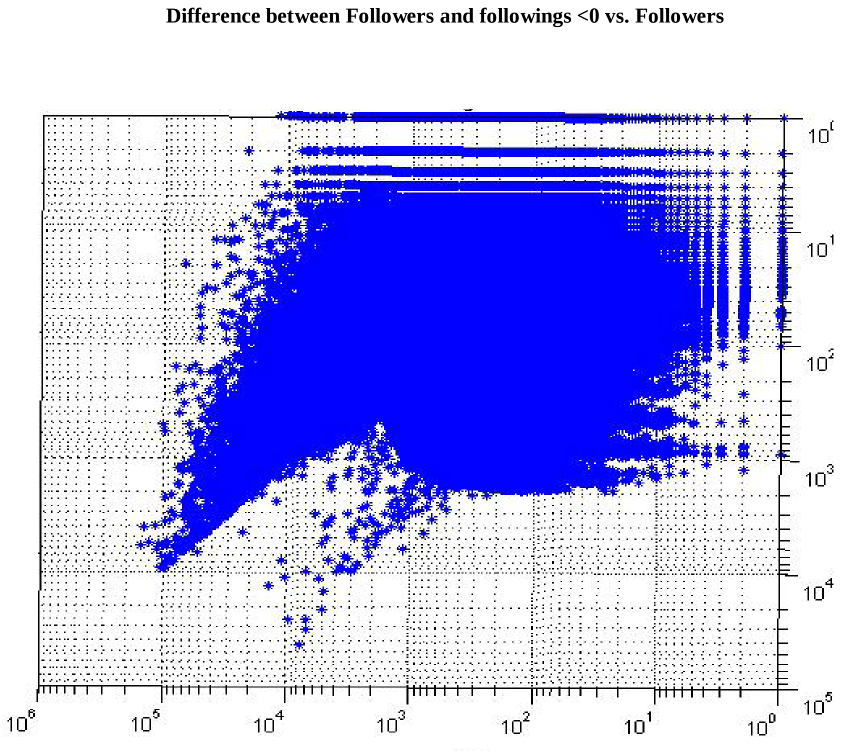


Figure 9: log-log Figure of P.I vs Followers considering only P.I < 0. For results in log scale the values where transformed in positive numbers and then the graph was turned upside-down.

6.1.3. Tweets

The following Figure shows the CDF of tweets for all type of users with an approximate average of 200 tweets.

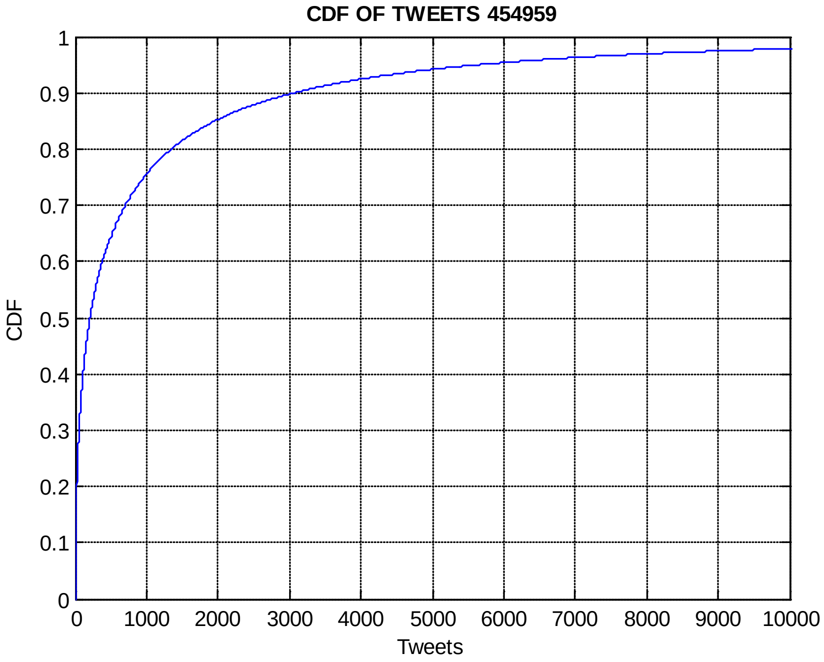


Figure 10: CDF of Tweets for 454,959 users. Average around 200 users

Interesting enough, this number almost doubles (around 460) if I consider only users with P.I >0 and decreases (around 130) if considering P.I <0. This means that users with followers >followings tweet more than users with followings > followers.

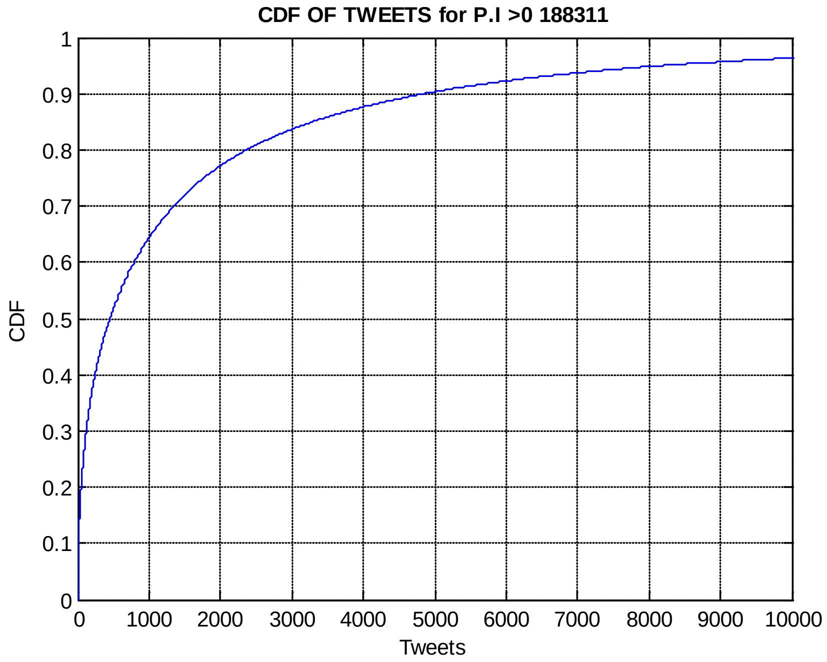


Figure 11 : CDF of Tweets for users with followers > followings. Max number of tweets is **496,338** and min number 0 and median of around 460

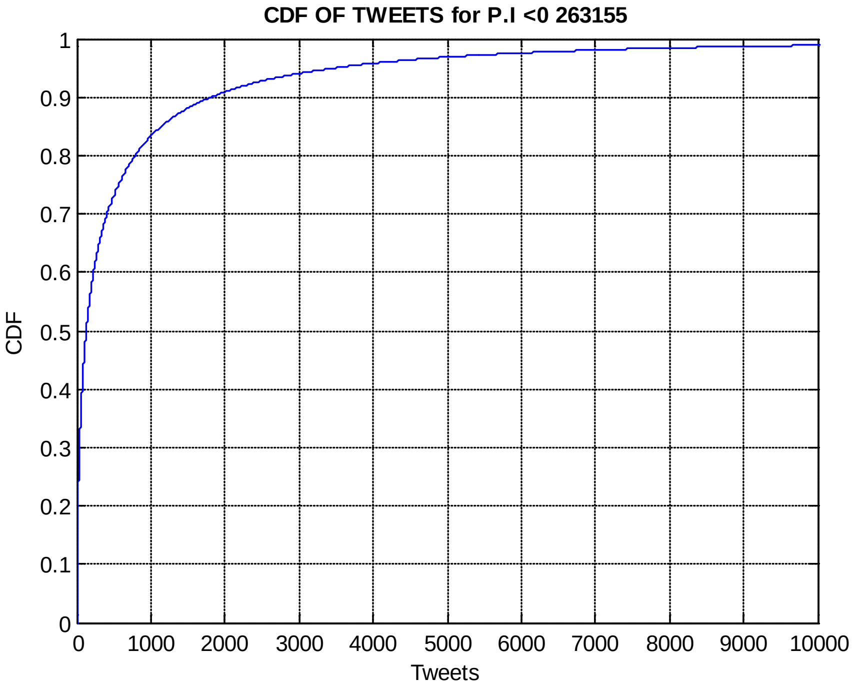


Figure 12: CDF of Tweets for users with followers > followings. Max number of tweets is **212,926** and min number of tweets is 0 and median of around 130.

6.1.4. Relationship between Followers and tweets

Besides the number of followers and the index, recommendations should also consider active tweeters who post tweets often. The tweets were captured from 2006 and 2009 and the profile information was taken by the end of the capture of tweets.

6.1.4.1. *Do people with more followers tweet more?*

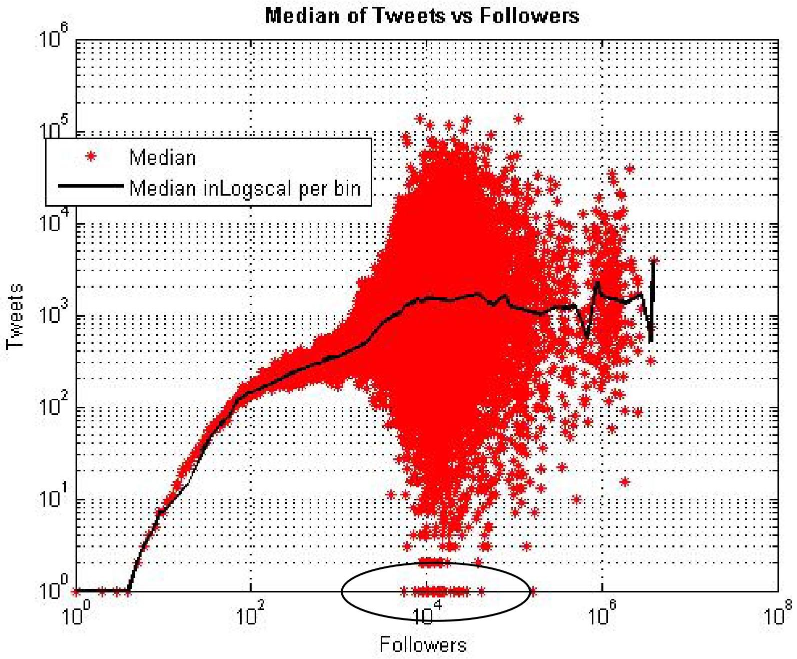


Figure 13: Median of Tweets vs Followers and the median per bin - logscale[22]

[22] Bins are calculated per power of bin 1 =10, bin 2 = 10, bin 3= 100, bin 4= 1000, etc.

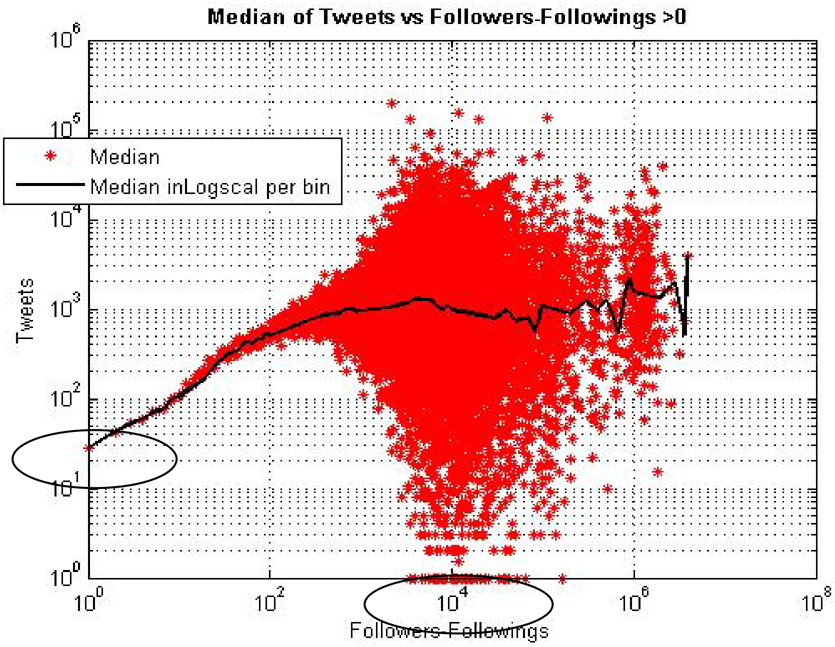


Figure 14 : Media of Tweets vs follower-followings >0 and the median per bin in logscale (black line)

Figure 13 considers both users with negative and positive P.I while Figure 14 only considers positive P.I. The differences are appreciated at the beginning, P.I >0 users tweet even if their P.I is low. Both figures present the highest variances between x values 1000 and 10,000 and it is interesting to see that the values captured by the ellipse are practically the same in both Figures, which means that in these cases Followings are very low. Due to the high number of followers, it is probable that these are celebrities, important people or causes that do not tweet a lot. The possibility of private accounts is excluded because tweets=0 are not considered in log-log scale (private accounts do not show the number of tweets).

Despite the variances in both figures and some plunges in the black line, both figures show that the tendency is to tweet more as the number of P.I and followers increase. Figure 14 shows less variance in the median number of tweets (more or less flat) between x values of 100 and 10,000.

Very similar findings were showed in (Kwak, et al., 2010) but the difference between followers and followings was not considered.

Recommendations should consider tweeters that tweet according to P.I, the larger the P.I and followers the more tweets should be expected. I should avoid recommendations with very low activity regardless the value of P.I.

6.1.5. Interesting cases

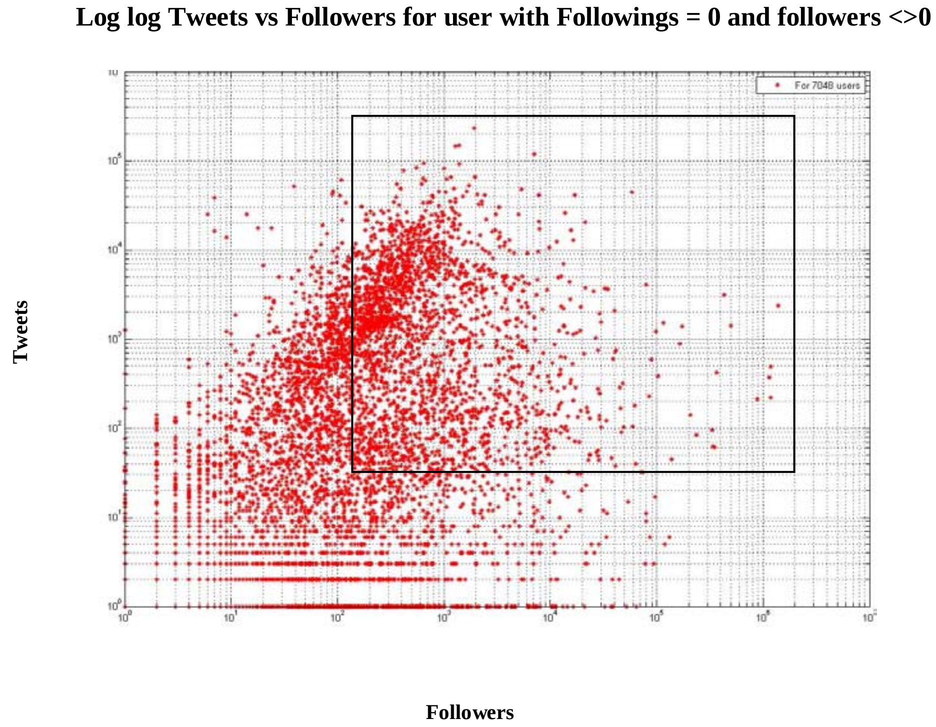


Figure 15 : Tweets vs Followers for user with followings = o and followers <> 0 for 7048 users

The previous figures did not consider cases where followers = 0 but included cases where followings =0. The reason is justified in

Figure 15, there are users that do not have followings but that still can be part of good recommendations. Users within the square are active users with considerable number of followers but that have chosen not to follow anybody. Should these users be considered for recommendations as well? If a user has a considerable amount of followers with a considerable amount of tweets and without the need to reciprocate at all then these users are certainly "interesting users" for recommendations.

6.1.6. TimeZone- Location

Another aspect that certainly is important at the moment of providing recommendations of people is their location. Certain people will prefer to connect with users who provide local information. For the location analysis I considered all the the data of the t *USER PROFILE* having 456,107 rows.

This dataset provides information about the "time zone" of the twitters in the dataset. The time zone, although not very accurate, provides a sense of location. Figure 16 shows a histogram with the corresponding distribution.

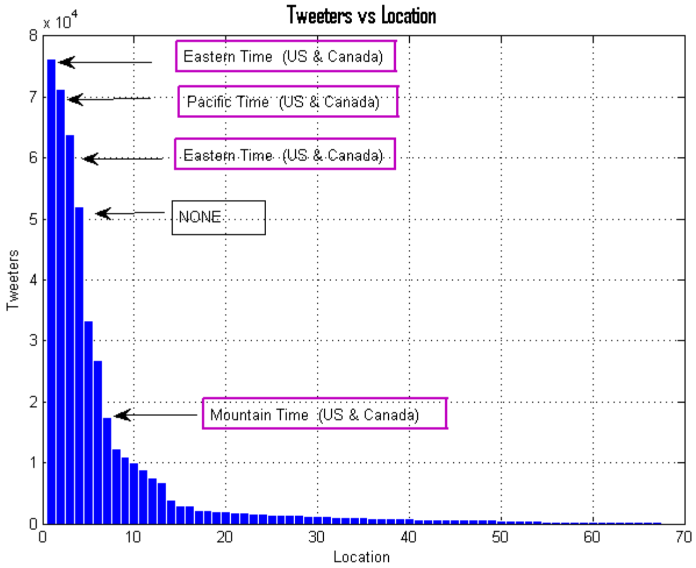


Figure 16 : Histogram Tweeters vs Location for all users

In the following graphs, I will see that the majority of people is found in the (US&Canada) region. I can classify location by time-zone.

Figure 17 shows that 50.17% of all tweeters belong to the United States of America and Canada Time Zone excluding Hawaii and Alaska

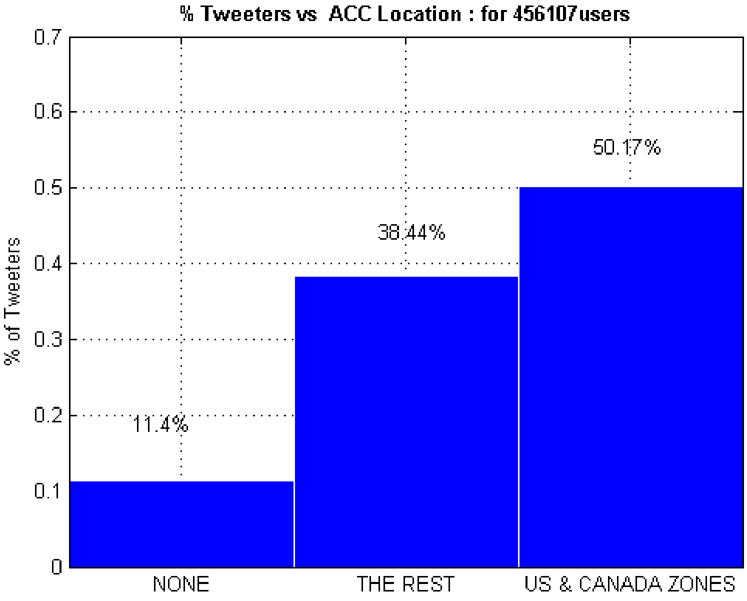


Figure 17: % of Tweeters vs Location (Three groups of location) for all users

The same classification can be done for the number of tweets.

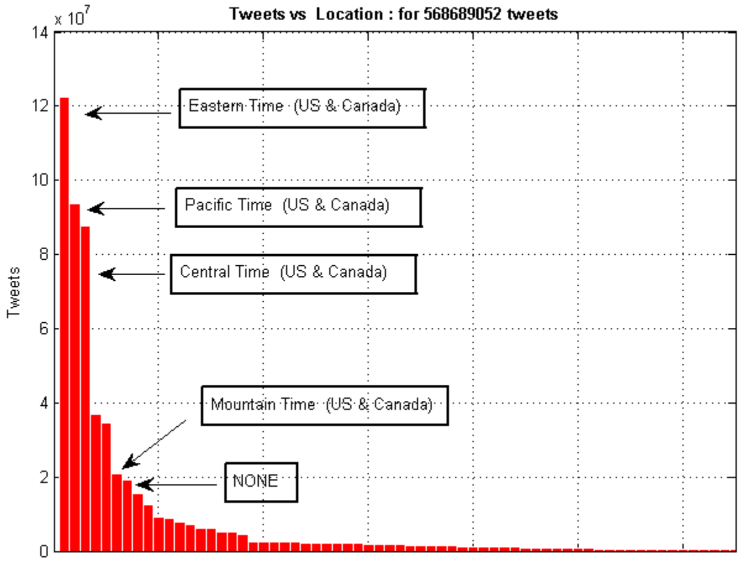
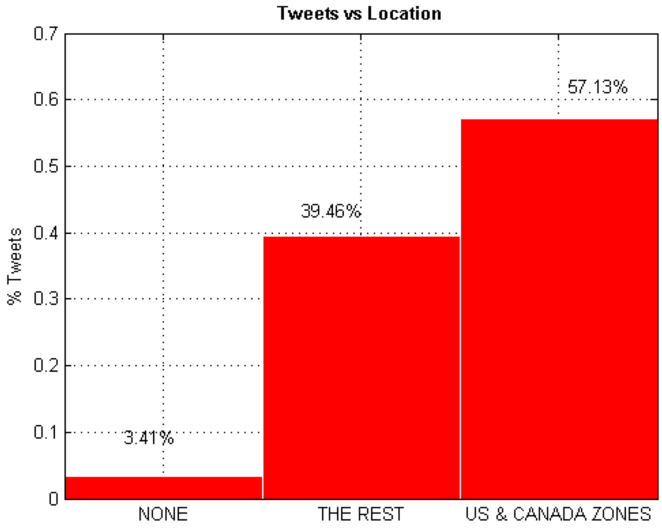


Figure 18: Tweets vs Location for all users

Grouping them by location, the majority of tweets also belong to the United States and Canada time zone accounting 57.13% of the total tweets.



6.2. SOCIALGRAPH : RECIPROCAL CONNECTIONS

6.2.1. GENERAL

The COMPLETE*FRIEND pairs* consider the *reciprocal connections* of a sample of users of user profile, considering the fixes specified in Section 5.3.2. For this table 2,476 users were considered and for each one, their reciprocal connections were listed.

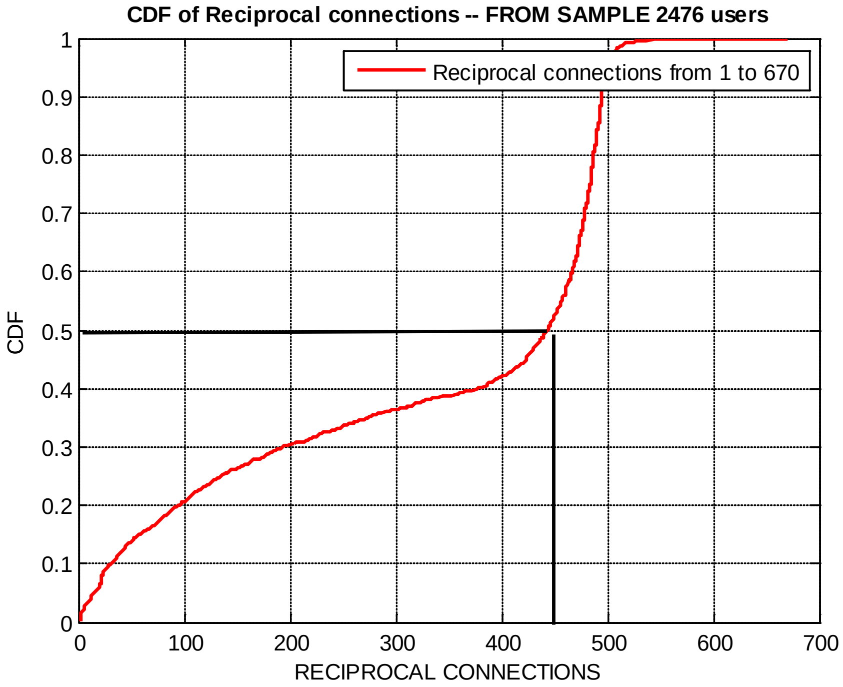


Figure 19 : CDF of reciprocal connections for simple of users in *CompleteFriend*

The maximum number of possible friends in the sample is 670. And more than 50% of users considered in *COMPLETEFRIEND pairs* have more than 450 friends (verified).

6.2.2. location of reciprocal connections

From this graph, I will find users that:

- Did not define a time zone
- Have cero friends in the same time zone
- Have all their friends in the same time zone
- Have at least 1 friend in the same time zone

From this classification, I will consider Central Time (US &Canada), Eastern Time (US & Canada), Mountain Time (US & Canada), Pacific Time (US & Canada) as on common "location." Even if I would consider these zones as separate regions, the results will not be very different (results were obtained for comparison).

The Users with location *NONE* are considered separately and not used for comparisons. The total amount of users in ***CompleteFriend*** with location = NONE is around 10%, that is, 237 tweeters. From the users having specified their location, I have that 2135 users have at least 1 friend in the same location and only 2476 – 237 – 2135 = 81 users have none of their friends in the same location.

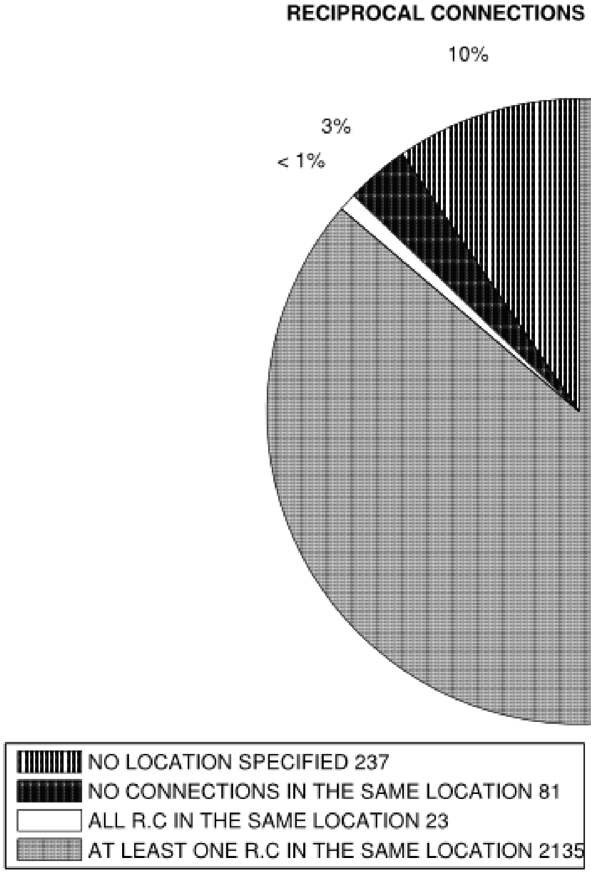


Figure 20 : Pie chart of the locations corresponding to the reciprocal pairs

I can see that the majority of users that specified their timezones will have reciprocal connections in and out of their timezones.

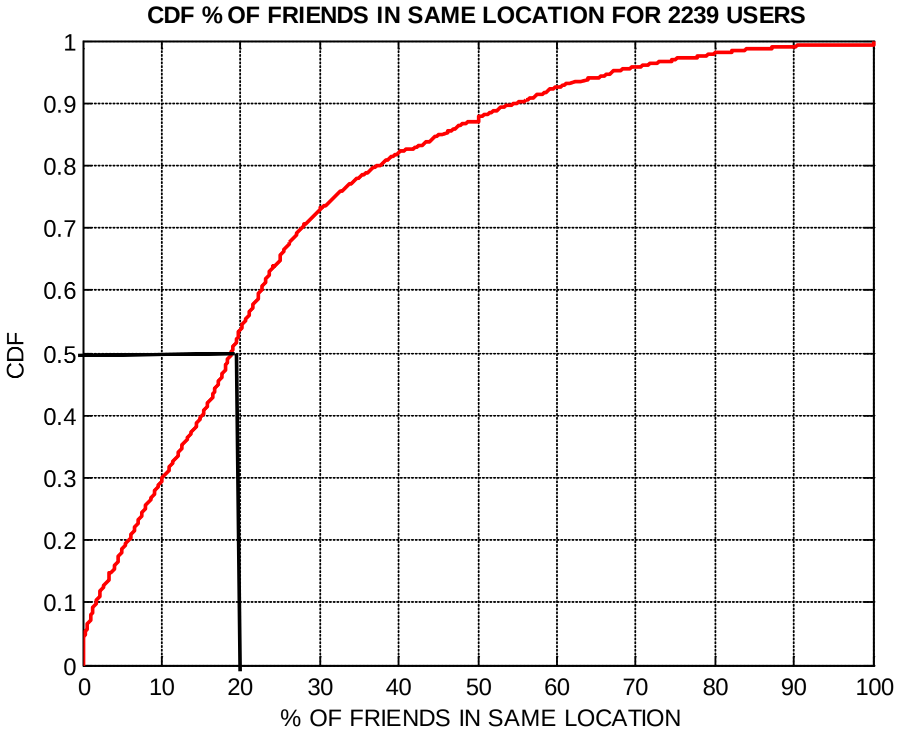


Figure 21: CDF % of friends in Same location – 2239 users analyzed

Figure 20 shows that 86% of users have at least one friend in his/her same location but Figure 21 shows that 50% of users have less than 20% of their friends in the same location and 90% of users have less than 50% of their friends in the same location. This implies that the majority of reciprocal connections of a tweeter are not in the same location. These findings make sense since twitter is not considered an OSN like facebook but rather an information network where public profiles share contents and news.

These findings go in accordance to (Scellato, et al., June 2010). In their dataset they found that Twitter pairs have an average separation length of 5,117 Km and that more than 80% of pairs are separated by distances longer than 1,000 km. They concluded that in fact Twitter users are engaged with a global audience of followers, even though there are also short-range social connections."

6.2.3. Popularity index of Reciprocal Connections

On subsection 6.1.2, the Popularity Index was analyzed for the user profile, showing that there are more people with P.I < 0. In this section, I will analyze the cumulative distribution frequency of the percentages that reciprocal connections with P.I <0 and P.I> 0 represent in the total number of connections for every user in ***completefriend pairs set***.

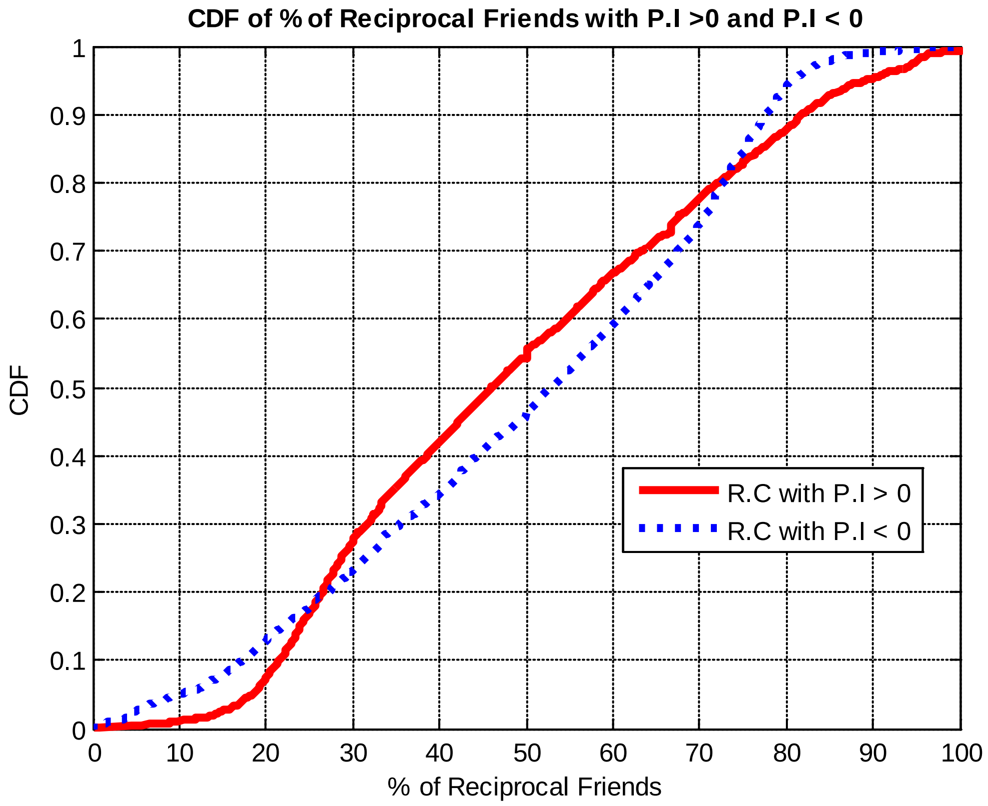


Figure 22 : CDF of % of R.C for cases with P.I > and P.I <0

Figure 22 shows that in the majority of the cases (middle part) there will be a bigger amount of R.C with P.I <0 among the reciprocal connections of every user. For higuer and lower values, R.C with P.I > 0 seem to take the lead. I can conclude for example that there will be more people having 80% or higuer of their R.C with P.I>0 but more people having 50% or higuer of their R.C with P.I<0.

7. EVALUATION METHOD

In order to simplify calculations and avoid consuming too much time in processing information, I sampled the table *Completefriend* using snow balling sampling method arriving to 61 users and more than 20,000 reciprocal links. These data was used for testing the classifiers. For more details about the data see Section 5.3.2 and Section 5.3.3.

The evaluation method chosen is Repeated Random Sampling with a 80/20 split because the amount of data for training and testing is limited. The training set corresponds to the 80% of the sample and the testing set corresponds to the remaining 20%. I tested how well the algorithms predicted the actual reciprocal friends of users in the testing set.

In order to make the training and the testing set representative, I divided the sample into the three regions identified in Figure 17 (None, the rest and USA& Canada) and 80% and 20% of pairs were added to the training set and testing set respectively from each one of the location groups. No repeated pairs are found between the testing and the training set. Moreover, I also repeated the whole process, training and testing, several times with different random samples in order to mitigate any bias caused by the particular sample.

After separating the testing set from the training set, I randomly added the same number of pairs already in the testing set from the table NoUserAfriend to the testing set in order to test the efficiency of the classifiers not only by how well they predict reciprocal connections but also how many mistakes they make by predicting no real reciprocal connections.

The results are reflected by Recall-precision curves. Precision shows how many of the recommendations are actually real connections and recall shows the percentage of actual friends left out from recommendations. A standard deviation was also calculated in order to show if resulting predictions were close to each other when choosing different samples for the training and testing set.

All classifiers were tested for different samples and the results stored in the table *Results*

8. CLASSIFIERS FOR RECOMMENDATIONS

At first, this study aimed to build classifiers based on

Popularity Index

Activity –Tweets

Location

Social graph

Content

For each pair of users in the testing set [userA, userB], the classifiers have to learn the preferences of userA from the training set. The recommendations are ranked according to different parameters depending on the classifier but all are based on what has been learned from the training set.

This approach may not be ideal in a real recommender system since the recommendations are not based on the "best options available" but actually on the preferences made by the user although they may not be part of the best options. As it was mentioned before, in this study we follow the same as in De Choudhury et al. [23] where reciprocal connections are "robust to spam" since "a normal user is less likely to follow a spam-like account."

Just as in Gayo-Avello,[24] the classifiers suggested in this study are not aimed to be directly applied to users when providing recommendations but, instead, as a weight within a given algorithm such as PageRank.

Unfortunately due to deadlines, the combination with famous algorithms as well the testing of the classifiers related to the tweets, location, socialgraph and content are left for further research. I only implemented the first classifier "Popularity-Index (a)." However, all classifiers are specified theoretically in the next sub-sections.

$$RS(A_i) = \ All\ the\ recommendations\ provided\ for\ the\ Ai\ in\ the\ testing\ set$$
$$TS(A_i)$$
$$= \ All\ the\ connections\ of\ user\ Ai\ in\ the\ testing\ set\ (actual\ reciprocal\ conenctions\ or\ not)$$
$$TR(A_i) = \ All\ reciprocal\ connections\ of\ user\ Ai\ in\ the\ traininsting\ set$$

8.1. RECOMMENDATIONS BASED ON POPULARITY INDEX

The number of followers and followings determine the popularity of a tweeter and therefore should be taken into account for recommendations. Of course, factors such as activity, locations and content must also be taken into consideration but those factors will be analyzed in other sections/subsections. For this test, I will only focus on recommendations based on followers and followings.

In order to determine the role of P.I in recommendations, I have tried to identify a method that calculates a P.I value from the training set to be used in the judgment of pairs in the testing set.

Parameter of recommendation

After obtaining this P.I, the recommendation follows that for every pair [User A, User B] in the testing set, the pair is considered to be a reciprocal connection if the P.I of User B is >= than the P.I found from the training set for User A.

[23] (De Choudhury, et al., 2010)

[24] (Gayo-Avello, 2010)

$\mathbf{P.I}(\boldsymbol{B_j})\ \boldsymbol{\epsilon}\ \boldsymbol{RS}(\boldsymbol{A_i})\ \boldsymbol{if}\ \mathbf{P.I}(\boldsymbol{B_j}) \geq \boldsymbol{P.I_r}(\boldsymbol{A_i})\ \boldsymbol{where}\ \boldsymbol{B_j}\ \boldsymbol{\epsilon}\ \boldsymbol{TS}(\boldsymbol{A_i})$ (Formula 5)

$P.I(B_j) = The\ user\ Bj\ with\ a\ P.I$
$P.I_r(A_i)\ =\ P.I\ used\ \ to\ judge\ connections\ of\ a\ certain\ User\ A\ in\ the\ testing\ set$

8.1.1. Calculation of the Popularity Index (P.I)

Several methods were analyzed to determine the P.I that could make more accurate predictions in the testing set. The best method was chosen based on its precision, recall and f-measure. At first, I analyzed a similar method considered in Daniel Gayo-Avello[25] , which corresponded to the ratio followers/followings in order to rank recommendations but the major problem found was in cases such as 10/1 and 1000/100: the ratio will be the same for both cases although it is more likely that the second case is a considerably more interesting user than the first one. For this reason, this "paradoxical_discounted" ratio was not used and it was decided to preserve the difference followers-followings since it reflects the actual *weight* of popularity, the bigger the difference the better the popularity. Of course, the activity and other factors should also be considered but as mentioned earlier this section only focuses on the popularity-Index.

After deciding to use ***followers-followings*** as the preferred method to *rank* recommendations, I faced the problem of deciding how to use this method in recommendations. I considered several methods and obtained the best results in the median, the average and a threshold controlling method.

In this regard, the top three methods are the following (in order):

a) MEDIAN OF P.I :
Median (P.I (Bj)) for $\boldsymbol{A_i}$ where $\boldsymbol{B_j}\ \boldsymbol{\epsilon}\ \boldsymbol{TR}(\boldsymbol{A_i})$ (Formula 6)

b) AVERAGE OF P.I :
Average (P.I (Bj)) for $\boldsymbol{A_i}$ where $\boldsymbol{B_j}\ \boldsymbol{\epsilon}\ \boldsymbol{TR}(\boldsymbol{A_i})$ (Formula 7)

c) CASE WITH THRESHOLDS

Choosing μ =0, 50, 100, 600 and 2000, analyze the percentage of reciprocal connections which P.I ≥ μ

a) If % (P.I (Bj) ≥ μ)≥ 50% $\boldsymbol{P.I_r}(\boldsymbol{A_i}) = \left(\frac{\boldsymbol{MIN}(\mathbf{P.I\ (Bj)} \geq \boldsymbol{\mu}) + \mu}{\mathbf{2}}\right)$ (Formula 8)

b) If 0>% (P.I (Bj) ≥ μ) < 50% $\boldsymbol{P.I_r}(\boldsymbol{A_i})$ = μ (Formula 9)

c) If % (P.I (Bj) ≥ μ) = 0 $\boldsymbol{P.I_r}(\boldsymbol{A_i}) = (\boldsymbol{MAX}(\mathbf{P.I\ (Bj)}))$ (Formula 10)

[25] (Gayo-Avello, 2010)

8.1.2. Results for the Popularity Index (P.I) Classifier

Formula 10 (*median)* reported fewer errors when making predictions. The average of the P.I indexes (Formula 11) was ranked second best method and the last method based on thresholds was ranked third. Thresholds are based on analysis made on subsection 6.1.2 (thresholds of P.I = 0, P.I = 50, P.I =100, P.I =600, P.I =2000).

For case c) and several other methods involving thresholds, I did not see significant changes in precision, recall and f-measure when changing μ. This led us to conclude that an ideal $P.I_r(A_i)$ for predictions is independent of μ for this evaluation method.

The Figure 23 shows the results obtained for recall and precision for a) b) and c) (for third case, different values of μ was used).

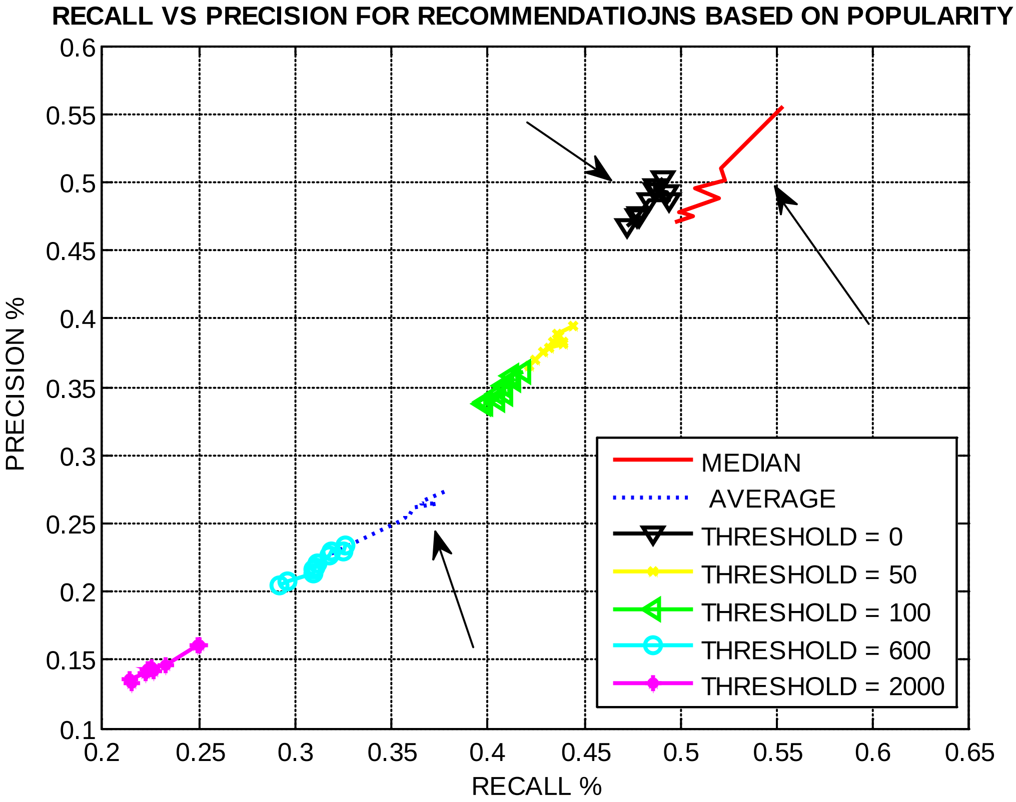
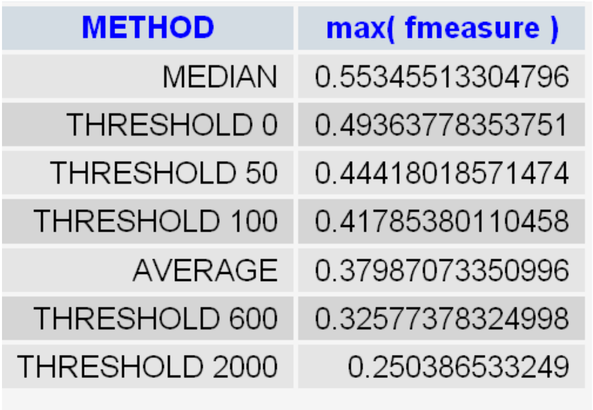


Figure 23 : Recall vs. Precision for different methods adopted in Popularity Classifier

The average presents more precision but does not perform as good with the recall. The median presented a better combination of precision and recall and it can be clearly seen in Figure 24.



METHOD	max(fmeasure)
MEDIAN	0.55345513304796
THRESHOLD 0	0.49363778353751
THRESHOLD 50	0.44418018571474
THRESHOLD 100	0.41785380110458
AVERAGE	0.37987073350996
THRESHOLD 600	0.32577378324998
THRESHOLD 2000	0.250386533249

Figure 24: Maximum F-measure performance on different tests for different methods adopted in Popularity Classifier

Although the median performed better than the rest of classifiers, in real life recommendations it is advisable to define thresholds regardless the preferences of users. Thresholds of minimum values are important so that we do not recommend spammers to spammers or heavily advertisers to similar users. A good recommender system will instead recommend people that are not likely to be spammers of heavily advertisers regardless the profile and preferences of the user.

Likewise, sometimes normal users are connected to friends that do not provide any valuable information and with very low P.I values, good recommender systems should try to recommend potential friends (since these users prefer be linked with friends) that have good P.I values.

8.2. RECOMMENDATIONS BASED ON ACTIVITY (TWEETS)

The level of activity should also be taken into account for recommendations. The Tweets value is the status count (number of tweets) posted since the profile was created. Active responsible users that are not spammers or advertisers actually read tweets coming from their followings and vice versa. These users will likely control the level of tweets that they want to receive from their followings and therefore their social graph will show the reciprocal connections that better suit this level of activity.
In order to determine the role of tweet number in recommendations, I will have to identify a method that calculates a maximum and minimum TW value (Number of Tweets) from the training set to be used in the judgment of pairs in the testing set.

Parameter of recommendation

After obtaining this TW value, the recommendation follows that for every pair [User A, User B] in the testing set, the pair is considered to be a reciprocal connection if the TW value of User B is within the MaxTW and MinTW found from the training set for User A.

$$\mathbf{TW}(\boldsymbol{B_j})\ \boldsymbol{\epsilon}\ \boldsymbol{RS}(\boldsymbol{A_i})\ \ \boldsymbol{if}\ \boldsymbol{MinTW_r}(\boldsymbol{A_i}) \leq \mathbf{TW}(\boldsymbol{B_j}) \leq \boldsymbol{MaxTW_r}(\boldsymbol{A_i})\ \ \boldsymbol{where}\ \boldsymbol{B_j}\ \boldsymbol{\epsilon}\ \boldsymbol{TS}(\boldsymbol{A_i})$$

(Formula 12)

$$TW\left(B_j\right) = The\ user\ Bj\ with\ a\ TW$$

$$\boldsymbol{MinTW_r(A_i)} = \textit{Minimum TW used to judge connections of a certain User A in the testing set}$$
$$\boldsymbol{MaxTW_r(A_i)} = \textit{Maximum TW used to judge connections of a certain User A in the testing set}$$

8.2.1. Calculation of the Activity (Tweets)

Three classifiers should be tested:

a) Max and Min of Reciprocal Friends :

$\boldsymbol{MaxTW_r(A_i)} = \boldsymbol{Max}\left(\mathbf{TW(B_j)}\right)$ $\boldsymbol{and\ MinTW_r(A_i)} = \boldsymbol{Min}\left(\mathbf{TW(B_j)}\right)$ where $\boldsymbol{B_j \epsilon\ TR(A_i)}$ and $\boldsymbol{Max}\left(\mathbf{TW(B_j)}\right) \neq \boldsymbol{Min}\left(\mathbf{TW(B_j)}\right)$ (Formula 13)

b) Max and Min of Reciprocal Friends with threshold for Max :

$\boldsymbol{MaxTW_r(A_i)} = \boldsymbol{Max}\left(\mathbf{TW(B_j)}\right)$ for $\boldsymbol{A_i}$ $\boldsymbol{and\ MinTW_r(A_i)} = \boldsymbol{Min}\left(\mathbf{TW(B_j)}\right)$ where $\boldsymbol{B_j \epsilon\ TR(A_i)}$ and $\boldsymbol{Max}\left(\mathbf{TW(B_j)}\right) \neq \boldsymbol{Min}\left(\mathbf{TW(B_j)}\right)$ $\boldsymbol{and}$ $\boldsymbol{Max}\left(\mathbf{TW(B_j)}\right) \geq \mu$ if not $\boldsymbol{Max}\left(\mathbf{TW(B_j)}\right) = \mu$

(Formula 14)

c) Max and Min of Reciprocal Friends with threshold for Min :

$\boldsymbol{MaxTW_r(A_i)} = \boldsymbol{Max}\left(\mathbf{TW(B_j)}\right)$ for $\boldsymbol{A_i}$ $\boldsymbol{and\ MinTW_r(A_i)} = \boldsymbol{Min}\left(\mathbf{TW(B_j)}\right)$ where $\boldsymbol{B_j \epsilon\ TR(A_i)}$ and $\boldsymbol{Max}\left(\mathbf{TW(B_j)}\right) \neq \boldsymbol{Min}\left(\mathbf{TW(B_j)}\right)$ $\boldsymbol{and}$ $\boldsymbol{Min}\left(\mathbf{TW(B_j)}\right) \geq \mu$ if not $\boldsymbol{Min}\left(\mathbf{TW(B_j)}\right) = \mu$

(Formula 15)

8.3. RECOMMENDATIONS BASED ON LOCATION (TWEETS)

The location should also be taken into account for recommendations specially when users are interested in trending topics happening locally or/and in reading tweets during certain time of the day. As it was said before, just like Kwak and lee.[26] in Twitter "it is hard to parse location due to its free form." The location is considered as the timezone of a user as an approximate indicator for the location of a user. Having the majority of reciprocal connections in the same location is an indicator of preference for locally tweeters

<u>Parameter of recommendation</u>

After obtaining the preferred location, the recommendation follows that for every pair [User A, User B] in the testing set, the pair is considered to be a reciprocal connection if :

[26] (Kwak, et al., 2010)

$\mathbf{L}(\boldsymbol{B_j})\ \boldsymbol{\epsilon}\ \boldsymbol{RS}(\boldsymbol{A_i})\ \ \boldsymbol{if}\ \ \boldsymbol{L}(\boldsymbol{B_j}) = \boldsymbol{L_r}(\boldsymbol{A_i})\ \ \boldsymbol{where}\ \boldsymbol{B_j}\ \boldsymbol{\epsilon}\ \ \boldsymbol{TS}(\boldsymbol{A_i})$ (Formula 16)

$L(B_j) = The\ user\ Bj\ with\ a\ Location$
$\boldsymbol{L_r}(\boldsymbol{A_i})\ \ =\ \ Preferred\ location\ used\ to\ judge\ connections\ for\ User\ A\ in\ the\ testing\ set$

8.3.1. Calculation of Recommendations based on Location

For this classifier, we find two cases all userA with no location specified should be ignored (only 10%), recommendations will be given according to the preferred location found on the training set. If there is no preferred location, the recommendation will be based on the location of UserA.
One classifier should be tested

Three classifiers should be tested:

a) Preferred location only :
$\boldsymbol{L_r}(\boldsymbol{A_i})\ \boldsymbol{if}\ \ \%\left(\mathbf{L}(\boldsymbol{B_j}) = \boldsymbol{L_r}(\boldsymbol{A_i})\right) \geq \mathbf{50\%}\ \boldsymbol{where}\ \boldsymbol{B_j}\ \boldsymbol{\epsilon}\ \ \boldsymbol{TR}(\boldsymbol{A_i})\ \boldsymbol{else\ ignore\ Ai}$ (Formula 17)

b) Preferred location if not Location of Ai :
$\boldsymbol{L_r}(\boldsymbol{A_i})\ \boldsymbol{if}\ \ \%\left(\mathbf{L}(\boldsymbol{B_j}) = \boldsymbol{L_r}(\boldsymbol{A_i})\right) \geq \mathbf{50\%}\ \boldsymbol{if\ not}\ \boldsymbol{L_r}(\boldsymbol{A_i})\ = \mathbf{L}(\mathbf{A_i})\ \ \boldsymbol{where}\ \ \boldsymbol{B_j}\ \boldsymbol{\epsilon}\ \ \boldsymbol{TR}(\boldsymbol{A_i})$
(Formula 18)

c) Location of Ai :
$\boldsymbol{L_r}(\boldsymbol{A_i})\ = \mathbf{L}(\mathbf{A_i})$
(Formula 19)

8.4. RECOMMENDATIONS BASED ON THE SOCIALGRAPH

The social graph should also be taken into account for recommendations specially when users are interested in following solely or majorly friends. Having a preference for friends implicates that recommendations should also be based on friends. In this case recommendations will be based in one-degree separation pairs.

<u>Parameter of recommendation and calculation</u>

In this classifier, the recommendation follows that for every pair [User A, User B] in the testing set, the pair is considered to be a reciprocal connection if User B is linked to any of the reciprocal connections of User A in the training set :

$\boldsymbol{SG}(\boldsymbol{C_k})\ \boldsymbol{\epsilon}\ \boldsymbol{RS}(\boldsymbol{A_i})\ \ \boldsymbol{if}\ \boldsymbol{C_k}\ \boldsymbol{\epsilon}\ \boldsymbol{R.C}(\boldsymbol{B_j})\ \boldsymbol{where}\ \boldsymbol{B_j}\ \boldsymbol{\epsilon}\ \ \boldsymbol{TR}(\boldsymbol{A_i})\ \textbf{and}\ \boldsymbol{C_k}\ \boldsymbol{\epsilon}\ \ \boldsymbol{TS}(\boldsymbol{A_i}\ )$ (Formula 20)

$SG(C_k)\ = Set\ of\ recommendations\ with\ one\ degree\ separation\ from\ \ A_i\ (Social\ Graph)$

$R.C(B_j) \;=\; The\ reciprocal\ connections\ set\ of\ B_j$

8.5. RECOMMENDATIONS BASED ON CONTENT

One of the main roles of Twitter is to be a source of information. Active responsible users tend to follow users who share common interests. For this reasons, recommendations should also be based on people tweeting topics of interest.

Parameter of recommendation

In this classifier, the recommendation follows that for every pair [User A, User B] in the testing set, the pair is considered to be a reciprocal connection if User B tweets topics that are interesting to User A:

$$\mathbf{T}(\boldsymbol{B_j})\ \boldsymbol{\epsilon}\ \boldsymbol{RS}(\boldsymbol{A_i})\ \boldsymbol{if}\ \boldsymbol{T}(\boldsymbol{B_j})\ \boldsymbol{\epsilon}\ \ \boldsymbol{T_r}(\boldsymbol{A_i})\ \boldsymbol{where}\ \boldsymbol{B_j}\ \boldsymbol{\epsilon}\ \ \boldsymbol{TS}(\boldsymbol{A_i}) \qquad \text{(Formula 21)}$$

$\mathrm{T}(B_j) \;= Main\ topics\ posted\ by\ \ A_i$
$T_r(A_i) \;= \ Topics\ of\ interest\ to\ A_i$

8.5.1. Calculation of Recommendations based on Content

The topic of interest for user A can be determined by finding the most common terms among their Reciprocal connections. There are several tools and algorithms to obtain these results.

Two classifiers will be considered:

a) Topics of interests among reciprocal connections

$$\mathrm{T}(B_j)\ \epsilon\ RS(A_i)\ \ if\ \ T(B_j)\ \epsilon\ \boldsymbol{T_r}(\boldsymbol{A_i}) = \ T(C_m)\ \ \ where\ B_j\ \epsilon\ \ TS(A_i) and\ C_m \epsilon\ \ TR(A_i) \qquad \text{(Formula 22)}$$

b) Topics of interests according to user A

$$\mathrm{T}(B_j)\ \epsilon\ RS(A_i)\ \ if\ \ T(B_j)\ \epsilon\ \ T_r(A_i) = \ \ T(A_i)\ \ \ where\ B_j\ \epsilon\ \ TS(A_i) \qquad \text{(Formula 23)}$$

c) Topics of interests of most influential reciprocal connections

$$\mathrm{T}(B_j)\ \epsilon\ RS(A_i)\ \ if\ \ T(B_j)\ \epsilon\ \ T_r(A_i) = \ \ T(INF(A_i))\ \ \ where\ B_j\ \epsilon\ \ TS(A_i) \qquad \text{(Formula 24)}$$

where

$INF(A_i) \;= Influential\ Reciprocal\ Connections\ of\ A_i$

Most influential users will be ranked by the number of retweets and mentions from A_i. At least three most influential users should be chosen

$$B_j \epsilon \; INF(\boldsymbol{A_i}) \; if \; B_j \; \epsilon \; TOP\,RT(D_m) \;\; and \;\; 0 < m \geq 3 \qquad \text{(Formula 25)}$$

where
$TOP\ RT(D_m) = users\ with\ more\ retweets\ or\ mentions\ from\ \ A_i$

8.6. WEIGHTING RECOMMENDATIONS

The weight of every classifier should be measured by finding the intersection of recommendations between them:

a) $\mathrm{P.I}(B_j) \cap \mathrm{TW}(B_j) \cap \mathrm{L}(B_j) \cap SG(C_k) \cap \mathrm{T}(B_j)$
b) $\mathrm{P.I}(B_j) \cap \mathrm{TW}(B_j) \cap SG(C_k) \cap \mathrm{T}(B_j)$
c) $\mathrm{P.I}(B_j) \cap SG(C_k) \cap \mathrm{T}(B_j)$
d) $\mathrm{P.I}(B_j) \cap \mathrm{T}(B_j)$

The Popularity Index is a very important classifier because it determines the real value of the popularity preferred. Contents, status count and locations can be misleading but certainly users with high P.I are interesting otherwise why a high number of people will chose to follow those users without asking for reciprocation?.

9. CONCLUSIONS

As it has been mentioned several of the results found from the analysis of the dataset used in this research is very similar to recent studies on Twitter.

On the other hand, the data analysis showed that the user profile provides information about: Popularity Index, number of followers, tweets and location. Different aspects were analyzed and interesting conclusions extracted.

a) There are more users with a negative P.I than positive P.I
b) The dataset presents users having a total number of followers from 0 to 3,909,704 and followings from 0 to 751,908.
c) Around 50% of users have more than 400 followers and followings but 50% of users with a P.I>0 have more than 400 followers and 200 followings.
d) For users with P.I>0 the 50% have P.I>100 whereas for users with P.I <0 50% have P.I > -150
e) The more followers you have with a P.I >0 the greater and uniform the P.I index
f) Around 50% of all users have posted more than 200 tweets

g) For users with a P.I > 0, around 50% have posted more than 400 tweets with a maximum of 496,338 tweets from 2006-2009. Users with P.I <0 have a lower maximum of tweets (212,926) and 50% posting more than 130 tweets.

h) People with more followers tend to tweet more but there are some cases where this rule does not apply.

i) Interesting case of followings =0 and large amount of followers

j) The majority of the users from this dataset are located in USA&CANADA timezone (taking into consideration four zones)

There are more unpopular people than popular people.

People with P.I >0 tweet more. The majority of users and activity is found in USA and Canada. Users tend to have more unpopular users among their reciprocal connections. The Popularity of reciprocal connections seem to be a promising predictor of connections and recommendations if considering the median. Other classifiers need and should be tested specially those proposed and analyzed or reviewed in this paper: activity, location, socialgraph and content.

Bibliography

Adams Jason The Mendicant Bug : Wanderings into computational linguistics, science, social media and life [Online] = What is your TunkRank?. - March 5, 2009. - 06 26, 2010. - http://mendicantbug.com/2009/03/05/what-is-your-tunkrank/.

Asur Sitaram and Huberman Bernardo A. Predicting the Future with Social Media [Online] / prod. Labs HP. - Hewlett-Packard , 2010. - http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf.

Cha Meeyoung [et al.] Measuring User Influence in twitter: The Million Follower Fallacy [Online]. - In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM), May 2010. - June 26, 2010. - http://an.kaist.ac.kr/~mycha/docs/icwsm2010_cha.pdf.

De Choudhury Munmun Munmun De Choudhury [Online]. - 2006 - 2010. - January 2010. - http://www.public.asu.edu/~mdechoud/datasets.html. - Social Datasets by Munmun De Choudhury is licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License..

De Choudhury Munmun [et al.] How Does the Data Sampling Strategy Impact the. Discovery of Information Diffusion in Social Media? [Online] / ed. Media In Proceedings of the 4th International AAAI Conference on Weblogs and Social. - 05 12, 2010. - http://www.public.asu.edu/~mdechoud/pubs/icwsm_10.pdf.

Gayo-Avello Daniel Nepotistic Relationships in Twitter and their Impact on Rank Prestige Algorithms [Online]. - 2010.

Kwak Haewoon and Lee Changhyun What is Twitter, a Social Network or a News Media? [Online]. - 2010.

Mccam Universal “Power to the people - Social Media Tracker Wave 4” “Next Thing Now” [Online]. - Oct , 2009.

Mccam Universal Power to the People - Social Media Tracker Wave 4 - Next Thing Now" [Online]. - July 2009.

Scellato Salvadore [et al.] Distance Matters: Geo-Social Metrics for Online Social Networks [Conference] // In Proceedings of the 3rd Workshop on Online Social Networks (WOSN'10). Colocated with USENIX'10. .. - Boston, MA, USA : [s.n.], June 2010. - p. 9. - http://www.cl.cam.ac.uk/~ss824/papers/wosn10_scellato.pdf.

Sobek Markus PageRank, A Survey of Google's PageRank [Online] / ed. eFactory. - 2002/2003. - 06 25, 2010. - http://pr.efactory.de/.

Tunkelang Daniel A Twitter Analog to PageRank (Authors' Blog) [Online]. - January 13, 2009 . - 06 24, 2010. - http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank/.

Twitter The Twitter Glossary [Online]. - May 04, 2010. - June 04, 2010. - http://help.twitter.com/entries/166337-the-twitter-glossary.

Weng Jianshu [et al.] TwitterRank: Finding Topic-sensitive Influential Twitterers [Online]. - 2010. - 04 26, 2010. - www.mysmu.edu/staff/jsweng/papers/TwitterRank_WSDM.pdf.

Witten Ian H. and Frank Eibe Data Mining Practical Machine Learning Tools and Techniques [Book]. - San Francisco : Elsevier In., 2005. - 2nd .