LAKEHEAD UNIVERSITY

# Local and Global Influence on Twitter

Thesis for the Degree of Master of Science in Electrical and Computer Engineering

**Shan ZONG**

**2016/8/22**

# Abstract

The analysis of influence in social network is drawing more and more attention. It can be applied in different areas such as political campaigns and marketing. In this work, the analysis of influence in Twitter, based on users' profile statistics in a real-time scale, was studied and discussed. Two methods of identifying influential users by given keyword in real-time are introduced.

To understand the relationship between users' influence features and social states in real life, two influence measures were presented: Local Influence which the user has on his/her immediate set of contacts and global Influence which the user has on the entire social network. This study describes in details these two metrics and shows their implementation for a real social network. Our case study, using Twitter, showed that the proposed model can create clusters of users in 2D space corresponding to their social standing, and can further be used to classify previously-unseen users into the correct classes with an f-measure of 0.82 which is significantly higher than benchmark algorithms. F-measure is often used for measuring the accuracy of the test for classification.

Keywords: Influence Analysis, Social Network Service, 2D classification

# Acknowledgement

I would like to thank Dr. Richard Khoury and Dr. Rachid Benlamri for kindly advising on my research work and supervising my study during the two years in Lakehead University. Thanks to my committee members: Dr. Samuel Pichardo, Dr. Sabah Mohammed and Dr. Carlos Christoffersen for reviewing my works and giving good suggestions on my works.

I would like to thank my friends in the lab for providing a good research environment in the lab. A lot of thanks to my friends for taking care of my catering while I was writing the thesis. Thanks to the friends in Thunder Bay for giving me so many joy and memories here. Many thanks to my parents for support me to come to study in Canada.

Shan ZONG (2016)

# Table of Contents

# List of Figures

# List of Tables

# 1. Introduction

## 1.1 Chapter Overview

This chapter introduces the motivations for the research work described in this thesis. We first introduce the concept of social network service with special focus on the usage of Twitter in the context of this research. Finally, we describe the objectives of the thesis and potential applications of the research.

## 1.2 Motivation

As the rate of Internet users is increasing, social network services including Facebook and Twitter are getting more popularity all around the world. The statistics of global social network users show that the number of users is rising every year. In 2016, there are approximately 3.17 billion Internet users, 68.3% of them are social network service users, especially within North America where 59% of the population uses social media outlets. By using social network services, each user has to create a personal profile, after which they are able to establish a connection with other users. To some extent, a user's online social media behavior could reflect the user's real social life. It provides us with a good foundation to do research on online social network services to understand how a user's thoughts or beliefs are propagated and may influence other users. Influence is the capacity to have an effect on others. In the case of influence within a social network, it can be described as an ability of the original node to induce reactions of other nodes within the network.

The analysis of such influences could be applied in areas such as political campaigns, marketing, and advertisement. For example, Yaron Singer proposed to identify a set of influential users, and designed a framework to encourage the influential user to broadcast the advertisement in the social network [1]. In this way, companies can make their product reach to more potential customers without spending large amount of money on TV commercials. Using social network service for political campaign is quite common, as it is convenient and fast to publish the candidates' news and collect the supporters 'opinions. And finding the influential user in the community could help the candidate winning more support.

## 1.3 Social Network Analysis

Social network analysis (SNA) is the process of investigating social structures through the use of network and graph theories. These graphs illustrate networked structures in terms of nodes (individual actors, people, or things within the network) and the connections between these nodes (relationships or interactions) [2].The research of social network analysis includes the structure of the network, the interaction between the users in the network and the user influence analysis.

In 1948, the concept of centrality was introduced for the analyzing people's communication in the social network by Bavelas. After that, centrality has been used as the indicators to identify the most importance nodes in a graph. There are different measures of centrality. These are given below:

- Degree Centrality: the number of ties that a node has in the graph. In a directed network, we usually define two kinds of measures: indegree (number of links directed to the node) and outdegree (number of links that the node directed to others).
- Closeness Centrality: the average distances of the shortest path between the node and all other nodes in the graph.
- Betweeness Centrality: the number of times a node connects two other nodes through the shortest path.

Previous social network analysis mainly focused on the relationship between the users in the network, from a stranger to a close friend. The relationship is quite important because it not only determines the interactions between users, but also affects the information flow, such as

which user shares the message with other users. The relationship is based on graph theory in the previous work, but in this work, we can find a method to measure the influence between users without using the graph theory.

## 1.4 Twitter and Usage Statistics

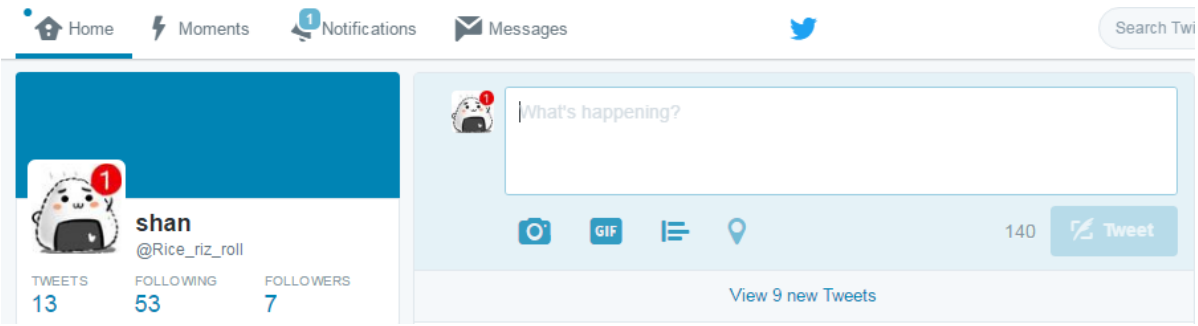


**Figure1. 1 Twitter User Account**

Twitter is a free social networking and micro blogging service based in the USA and it is one of the most popular social media platforms in the world. Users can easily access Twitter by visiting its website through their computer or by using an App through their smart phone or smart device. As shown in Figure 1.1, twitter users can publish messages which are referred to as tweets (messages no more than 140 characters). The user of twitter plays two kinds of roles: reader and author. As a reader, the user can read the tweets published by users they follow, after which the user may mark it as a favorite or comment. As an author, the user can publish a tweet or transfer the tweet. The limited information of the user including; personal information, tweets, and network information can be extracted by Twitter Application Programming Interface (API).

With over 300 million active users around the world, 500 million tweets are sent every day. Twitter has become a popular median of expression for politicians, companies, entertainers, and the public. It has likewise become an instructive data source for scientific study.

## 1.5 Thesis Objectives

As Twitter is a free but widely used social media platform, a message can travel all around the social network. This makes Twitter a perfect choice for many research institutions and companies to study the dynamics of users' information exchange, perception and behavior. The first goal of this thesis is to find how influence is defined in the Twitter world. Secondly, by using our definition of influence, this study will investigate the most influential users using specific keywords efficiently on Twitter. Thirdly, we introduce a new approach using the global influence and local influence measures defined in the thesis to classify users into different categories.

# 2. Background and Literature Review

## 2.1 Chapter Overview

In this chapter, the background of analysis of influence in Twitter is presented in details. The literature review covers three facets of the research in influence analysis in Twitter: the application of the analysis, the metrics to measure the influence and the measuring tools. First, we show how the utilization of the analysis is applied in political research and marketing areas. Then, we follow with the existing approaches to measure the influence, which includes both the offline and real-time analysis methods. Finally, some popular Twitter influence analytical tools are introduced.

## 2.2 Analysis of Influence in Twitter

Twitter is a typical application of microblog service. Users can easily express their ideas by simply publishing a tweet. The tweet can be simply some words, picture, video, or URL link. Tweet's variety has the advantage of attracting users' attention because it can provide all kinds of information from one simple platform.

The speed of the tweet's expansion and popularity in users from the entire world is astonishingly high. In the first quarter of 2010, there were around 30 million monthly active users; however, in the first quarter of 2016, the number reaches around 300 million[1], So more researchers are putting their efforts in studying Twitter [3].Consequently, studies on identifying influential users are drawing more and more attention [4], especially the studies aimed at using it in viral marketing [5] and political research.

### 2.2.1 Application in Political Research

As the Twittersphere in North America is quite large, a lot of meaningful analysis in the political area has been conducted by researchers. For example, Rodrigo Sandoval-Almazan uses the data from Twitter to predict the president candidate in a Mexico campaign [6];Ruan, Lotus Yang analyzes the Canadians' public sentiment on China [7].

[1] http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

**The Multiple Facets of Influence: Identifying Political Influentials and Opinion Leaders on Twitter [8]**

Many scholars pointing out that digital technology brings the opportunity of understanding political system to normal citizens, and made significant change in political mechanism [9,10,11]. A significant question is: Among a variety of metrics in Twitter, what is the most important facet to focus on to identify the influential user in the political world? In the paper, the authors compare metrics from four different facets to identify influential players in two of the most important Canadian political parties in the Twitter community: the Conservative Party of Canada and the New Democratic Party of Canada. The comparison of the metrics lead to some meaningful results.

#### 2.2.1.1 Metrics

The four facets are centrality, interaction, knowledge and local social embeddedness. The metrics they used are described in detail in the following.

1. Indegree: the number of incoming connections to a node. This metric belongs to the centrality facet.
2. Eigenvector: a measure of how connected a node's connections are. For a node i, we assume its eigenvector centrality can proportional to the summation of the neighbours'centralities, which can be defined as follows:

$$\lambda C_e = A^T C_e \text{ [12]} \qquad (2\text{-}1)$$

$C_e$denotes the eigenvector of adjacency matrix AT, and λ is the corresponding eigenvalue. This metric belongs to the centrality facet as well.

3. Clustering coefficient: a measure of how embedded within the network a node is, it defines the probability that two randomly selected friends of the node i are friends of each other. This metric belongs to the local social embeddedness facet.
4. Knowledge: the ratio of tweets featuring context-specific terms at that node. This metric belongs to the knowledge facet.
5. Interaction: the number of mentions of that node by other users, this metric belongs to the interaction facet.

#### 2.2.1.2 Data Collection

In the research, they established datasets for the two parties. First, they collected tweets with hash tag CPC and NPC, and then selected interesting users who publish the tweets. With more than 3,000 users chosen for each party, a social network graph was generated, and up to 200 recent tweets of each user were retrieved as well.

#### 2.2.1.3 Methodology and Results

The author used two different methods to analyze the metrics. First, is to rank the user with different metrics and compare the ranking result. In order to see the relationship among these metrics, the Kendall's $\tau$ is used in the measurement of each of the two metrics. Kendall's $\tau$ is used to measure the degree of correspondence between two rankings. The results show that indegree and eigenvector centrality have a high $\tau$, which means they both indicate how central a node is within a network. Regarding the influence, indegree and eigenvector focus on having a following. The knowledge and clustering coefficient have a tau approaching zero, which means these two are independent. Also, the interaction has a minor agreement with four other metrics.

The second method focuses on the profile content analysis. By analyzing the profile of the top 20 users in the rank, the researcher found that the indegree and eigenvector centrality metrics could identify the traditional political elite such as media outlet. The knowledge and interaction metrics could identify political commenters and bloggers, and the local clustering coefficient helps to identify opinion leader influence on their own local network.

In conclusion, different metrics suit different facets, and finding influential users can not be simply determined by a single metric. We should also consider the target group of the user, and combine the condition with the metric to identify the influential users accurately.

### 2.2.2 Application in Marketing

Besides regular private users, we could find a lot of company/ business accounts in Twitter as well. Twitter, as a free website, is a good platform for the enterprise to promote their products and services. The propagation of the message within Twitter is a hot topic for scientists to study on [13]. Qualitative analysis used in finding crucial users in Twitter is very meaningful for doing effective advertising for the company [14].

**Diffusion of Messages from an Electronic Cigarette Brand to Potential Users through Twitter [15]**

Twitter can be a good social media platform for product advertisements because the followers receive the tweets from the user from time to time. If the message can be sent effectively, the outcome can be competitive, as for example, being put on TV commercials. However, compared to the cost of TV commercials, the cost for tweeting is extremely low.

#### 2.2.2.1 Methodology

In this case, Chu and Unger presents how the tweets travel from an electronic cigarette brand to potential users. The research object is an e-cigarette brand called Blu. To figure out how the message flows, the social network is modeled into three layers. The researchers select Blu as the layer 0, define the followers of the brand as layer 1 and the followers of Blu follower as layer 2. They collect the tweets and the retweets of the target for 2 months to see how the size of the social network changes over time. By analyzing the profile of the users in layer 1 and layer 2, they classify the users in the following categories:

1. Person-Supporter: The users express a supportive view toward the e-cigarette by mentioning the related words in their profiles
2. Person-Basic Profile: The users who doesn't mention about the e-cigarette in their profiles
3. Researcher: The users who mentions they are doing the related research in their profiles
4. Nonperson: The users who mentions that they are groups or enterprise but not related with e-cigarette in their profiles
5. Industry-Retailer/Manufacturer: The users who mentions they are e-cigarette retailers or manufacturers in their profiles
6. Industry-Other: The users who mentions they are in the related industry of e-cigarette in their profiles
7. Unclassified: The users with blank or meaningless symbols in their profiles

#### 2.2.2.2 Discussion

The results show that in the retweet network, the users from layer 1 are mainly from person- supporter, however in the layer 2 the majority is the person-basic users. And the topic of the retweets changes at the level of followers as well. Layer 0 focuses on the social and entertainment events, layer 1 retweets more frequently about the news and laws, while layer 2 users' messages more focuses on ecig, which means they may not identify with certain messages released by layer 1 users.

In conclusion, the focus of tweets is shifting from product advertisement to social behavior. This work shows: first, how the message travels from the retweet network. Second, the main constitution of each layer is different. Third, the person-supporter is a good group for advertising. Finally, the message with social behavior travels further than the product advertisement.

## 2.3 Existing Approach to Measure Influence

There exists simple metrics which measures the popularity, for example, the Follower Rank [16]. Follower Rank is the result of the user's total follower number divided by the sum of user's follower number and followee number. Analysis of influence in Twitter can also be performed by classifying and collecting the data of the actions in Twitter. In the paper [17], the researchers proposed that four most important actions in Twitter are reply, retweet, mention and attribution. These four actions are four different ways for users to communicate.

Reply: user can reply to another user's comments.

Retweet: Transfer other user's tweet, and this tweet can be published on the user's page so the user's follower could see it.

Mention: When a user gets mentioned, he or she will get a notification. And it is shown as the symbol @ with the username in the tweet.

Attribution: This is used in transferring a tweet by marking the source, usually shown as "via"+ message source.

Reply should be considered related to the influence because it happens when another user is influenced by the content of the tweet, and so does retweet. Mention also should be related to influence because it is similar to the reply, and the attribution is similar to retweet as well. Just attribution is used to provide the citation of the previous published content. For the measurement performed by the researchers, the mentions also encapsulate attributions.

- **Methodology**

At the first step, two methods of measurement of the influence in Twitter have been examined. The first method consists of counting the total number of the user's follower; the second however calculates the ratio of the followers to followees of each user. The flaw of these two methods is apparent because they don't take the previous actions into account. Then the researchers collected relevant data from 12 users from celebrities, news outlets and social media analysts for 10 days.

The researcher [17] divided the actions into two kinds of categories: conversation-related (reply and mention) and content-related (retweet and attribution). To find the average

influence the user caused for each follower, they use the equation (2-2) and (2-3) for conversation-related and content-related influence.

Conversation-related= (# of reply + # of mention) / # of follower (2-2)

Content-related= (# of retweet + # of attribution) / # of follower (2-3)

To illustrate the average influence that the user caused by each tweet, they utilize the following equations 2-4 and 2-5:

Conversation-related= (# of reply + # of mention) / # of tweet (2-4)

Content-related= (# of retweet + # of attribution) / # of tweet (2-5)

- Discussion

The results show that equation (2-4) and (2-5) approach the accurate estimation most, however, this method still has shortcomings because it doesn't consider the follower's network of the user, and the results can vary considerably if considering the real situation. From the user group side, it seems celebrities with large number of followers foster more conversation than provide retweet content while news outlets influence followers to retweet their content to other users no matter the number of the followers.

### 2.3.1 Existing Offline Analysis Method

Many influences calculating algorithms are inspired from PageRank [18,19], which is an algorithm for measuring the importance of the website page. The similarity between social network and the network between webpages allow this algorithm to be performed on measuring

the importance of the user (node) in his or her social network. The PageRank is a good algorithm for calculating the user's eigenvector centrality.

Another quite important algorithm is HITS (Hyperlink-Induced Topic Search), also known as hubs and authorities. It is originally used for analyzing links to rate the web page [20]. It emphasizes the idea of hub and authority: hub user could lead a lot of other users to link to the authority; authority could then link to a lot of hub users.

The difference between HITS and the PageRank is the execution period. The PageRank is executed during indexing while HITS is executed during query. The main flaw of the static analysis method is the time, because the algorithm is iterative, it is not able to meet the need of the real-time analysis.

### 2.3.2 Measure Influence with given keywords

In the real marketing case, it's more important to get the influential user of the certain field because it's more effective to do the advertisement by focusing on the target group. However, how to find the target user is a meaningful question to be discussed. To answer this question, some researchers worked on the user influence ranking with given keywords [21].

#### 2.3.2.1 Methodology

The researchers in this study use twitter keyword search to get tweets matching the keyword, then they extract the user information and tweet relation. A reference graph could be built based on the retrieved information. They defined the TURKEYS score of each user as follows:

$$\mathrm{TURKEYS(u)} = \mathrm{TC}(u)^{w} \mathrm{x}\ \mathrm{UI}(u)^{1-w} \tag{2-9}$$

Where w is the weight ranges from 0 to 1, it could start from 0.5.

TC(u) is the tweet count score which count both user's original tweets and retweets, which can be represented as follows.

$$\mathrm{TC(u)} = \frac{|\{t|t\in T0 \cap t.user.id=u.id\}|}{\max u\prime\in Uall|\{t|t\in T0 \cap t.user.id=u\prime.id\}|} \quad (2\text{-}10)$$

Where t.user.id indicates the poster's ID of the tweet t, u.id indicates ID of the user u, and T0 is the set of 1000 recent tweets by querying with the keywords.

UI is the user influence score, which could be calculated with the reference graph consisting of user node which is based on the PageRank [22]. The first step is to create a user reference graph to represent the retweet, reply and mention relationship. Au is used to indicate the adjacency matrix of the graph.

$$\mathrm{Au}\ (\ \mathrm{ui, uj}\ )\ =\ \mathrm{retweet}\ (\mathrm{ui, uj})\ +\mathrm{mention}\ (\mathrm{ui, uj}) \quad (2\text{-}11)$$

Retweet (ui, uj) means the total number of times ui retweet uj's tweet and mention (ui, uj) means the total number of times ui mention uj. So the matrix can be transformed as follows.

$$\mathrm{Bu(ui, uj)} = \begin{cases} \frac{\mathrm{Au(ui,uj)}}{\sum_k \mathrm{Au(ui,uk)}}(1-\mathrm{d}) + \frac{\mathrm{d}}{|\mathrm{Uall}|} & \text{if } \sum_k \mathrm{Au(ui,uk)} \neq 0 \\ \frac{1}{|\mathrm{Uall}|} & \text{otherwise} \end{cases} \quad (2\text{-}12)$$

And the user influence score can be computed by the algorithm below. Line 1 gives the equation to show how to calculate the user influence score, where **u** stands for all users' column vector of the user influence score. Line 2 gives the initial value, and the line 3 set k as 1, then line 5 shows doing the kth iteration to calculate the user influence score, until the error between

the user influence score of kth and k-1th iteration coverage to ε, which shows in line 7. Then the final result is in line 9.

1 $\mathbf{u}=B_u^T\mathbf{u}$

2 $\mathbf{u}0=\left(\frac{1}{|Uall|},\frac{1}{|Uall|},\ldots\frac{1}{|Uall|}\right);$

3 k=1;

4 Repeat

5 $\mathbf{u}_k=B_u^T\mathbf{u}_{k-1;}$

6 k=k+1;

7 until $|\mathbf{u}_k-\mathbf{u}_{k-1}|<\varepsilon;$

8 return $\mathbf{u}_{k.}$

9 UI(uj) = $\frac{u(j)}{\max k\ u(k)}$

**Figure 2. 1 User Influence Calculation**

It could be seen in this step that the method is iterative, and therefore, it can't be finished in few seconds. Furthermore, when we apply this step, we need the graph of the users' social network. The more complete this graph is, the more accurate the result will be. Also, mining the users' information and creating a complete graph for each user can't be done in real time.

#### 2.3.2.2 Discussion

This method ranks good users to follow because it could exclude users who post similar tweets like advertisements (user with high tweet count score) and users who post few but are retweeted and replied to a lot (user with high user influence score). On the other hand, this

method does not take the text and the user profile into account, which could lead to a loss for the case that the user doesn't have many followers but are followed by the users interested in particular topic.

### 2.3.3 Real-time Algorithm: IARank [23]

The traditional ranking algorithms are mainly iterative, such as PageRank, which means they are time consuming. In the real world of twitter, world-wide event happens so fast, and the offline analysis may not be appropriate for real time ranking.

#### 2.3.3.1 Methodology

In this paper, researchers proposed a ranking method which could work in real-time based on their information amplification potential. The ranking scheme can be measured by two factors: Buzz and Structural Advantage.

Buzz measures the attention the user receives from other users in the network, which shows in the following equation.

$$\text{Buzz} = \frac{\#\text{Mention}}{\#\text{Event Activity}} \tag{2-13}$$

#Event Activity is number of times a user actively participated in the event， which is the times for posting tweets, retweeting and mentioning or replying to other user.

Structural Advantage measures whether the local network structure around a user is better suited to provide information to the network or seek information from the network, which can be defined in the following equation.

$$\text{Structural Advantage} = \frac{\#\text{Followers}}{(\#\text{followers}+\#\text{followees})} \quad (2\text{-}14)$$

We can get the formula to calculate the cumulative influence in the following equation by using the above equations.

$$\text{W}(\text{User(i)}) = \sum_{j=1}^{n} (Buzz + Structural\ Advantage)_{User(j)} \quad (2\text{-}15)$$

W(User(i)) is the cumulative influence achieved by the User(i), and n is the total number of the edges connecting to the User(i).

#### 2.3.3.2 Discussion

As the PageRank has been selected as benchmark, the comparison between benchmark and experiment results proves the quality of the method. This method has a good performance for real time ranking of the top 4 or 5 influential users, but the main flaw is the trade-off of the time when increasing the group size.

.

## 2.4 Twitter Influence Analytical Tool

As the importance of influence in social network service is drawing more and more attention, a variety of influence analytic tools have been developed to provide guidance to the user. These tools usually have their own methodology of calculating the user's influence, and some of them are made into product available to users with tailored suggestions to improve user's influence, such as Klout and PeerIndex. Some other tools are mainly for research purpose, such as tweetStimuli.

### 2.4.1 Klout

The most successful commercial case among the measurement tool could be Klout. Klout use the Klout score to represent the influence potential for the user, which varies from 1 to 100, higher score means higher influential power. The score is calculated by their own algorithm according to the data collected from 9 different social networks [24] (Facebook, Twitter, Wikipedia, Google+, etc.). Klout uses over 3600 features which captures the user's online social network actions to conduct the influence analysis. After training the model to decide the weight for the feature, the Klout score is determined. The user need to authorize Klout to connect to their social network service account, then Klout gives an evaluation to the user, as shown in Figure 1.2. Due to this limitation, it is not easy for researcher to access abundant data to evaluate the method. Klout not only provides how to increase influence service to individual user, but also provide company with the solution to combine influencer with the product, this part of service is provided by Klout and Lithium. Lithium is a company who provides social customer experience management software for the enterprise.

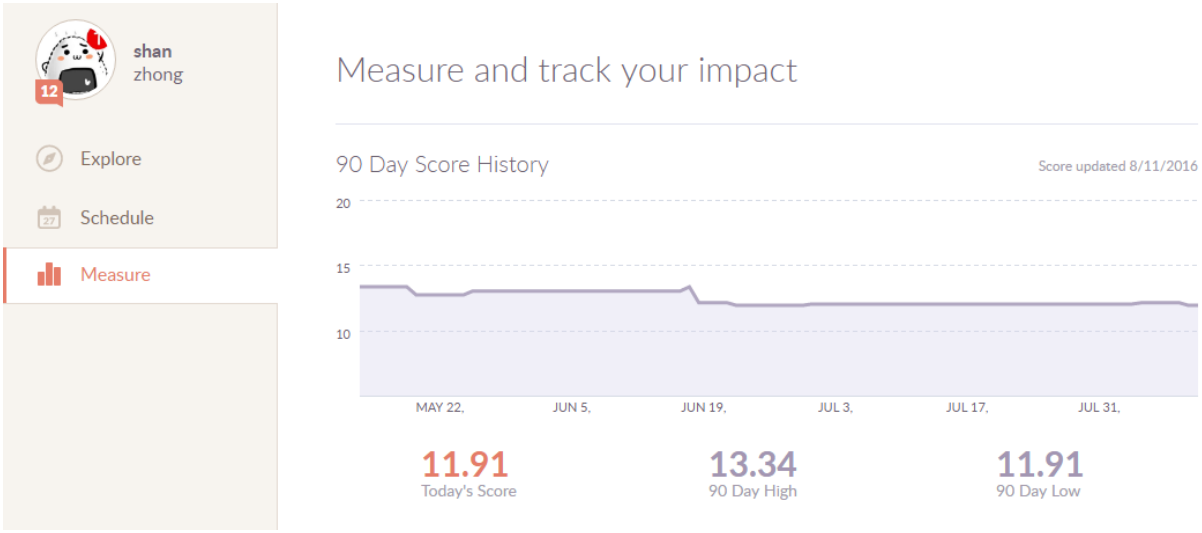


**Figure 2. 2 Klout Score Evaluate User's Influence**

### 2.4.2 PeerIndex

PeerIndex is similar to Klout. It provides social media analytics based on measuring influence from gathering the information from Facebook, LinkedIn, Quora and Twitter [25].Unlike Klout, PeerIndex focuses on helping companies working on their brand. PeerIndex was purchased by Brandwatch, a social media monitoring company, with a capital of about 15.7 million dollars.

PeerIndex collects the brands' information from social network and ranks them. The user could check the brand's influence in the sector the brand belongs to. The sectors provided are automotive, consumer technology, food and beverage, healthcare, luxury fashion, MLB, NBA, nonprofit, public sector, retail, telecommunication and TV network.

| BRAND | SOCIAL VISIBILITY | GENERAL VISIBILITY | NET SENTIMENT | REACH GROWTH | ENGAGEMENT & CONTENT | TOTAL |
|---|---|---|---|---|---|---|
| Lexus | 100 | 53 | 69 | 58 | 100 | 381 |
| Tesla | 73 | 74 | 72 | 64 | 92 | 376 |
| Mercedes-Benz | 90 | 61 | 86 | 58 | 48 | 342 |
| Audi | 62 | 71 | 100 | 59 | 48 | 340 |
| BMW | 76 | 80 | 85 | 60 | 38 | 339 |
| Ferrari | 77 | 69 | 85 | 64 | 42 | 337 |
| Porsche | 73 | 66 | 79 | 60 | 57 | 335 |
| Suzuki | 27 | 62 | 100 | 43 | 99 | 331 |
| Dodge | 79 | 74 | 61 | 55 | 58 | 327 |
| Ford | 65 | 100 | 66 | 57 | 36 | 324 |

**Figure 2. 3 Top 10 Automotive Brands in PeerIndex**

Figure 2.3 illustrates the top 10 automotive brands in PeerIndex, besides the total score, it also gives evaluation from five different facets: Social visibility, General visibility, Net sentiment, Reach growth and Engagement content. As PeerIndex does not provide rating for individuals, we could not use it as benchmark in this study. However, if we extend our study to measure the influence for company and organization account, PeerIndex can be a good choice as comparison benchmark.

### 2.4.3 tweetStimuli

Besides, there also exits some other analytic tools not for commercial use but for research purpose, such as tweetStimuli. TweetStimuli is specialized in discovering social structures of local influence [26].The definition of local influence is given by Bakshy and Eytan in [27]:given a target user A, it refers to who has been influenced by A focusing on diffusion cascades of depth 1 which are the most informative. TweetStimuli can help users investigate their social graphs and rankings about who influenced them and who has been influenced by them based on analyzing the last 100 retweets and favorites of the target user.

# 3. Finding Influential Users by Keywords in Real-time

## 3.1 Chapter Overview

This chapter mainly discusses the metrics and experimental results of finding influential users by keywords in real-time. In this chapter, the introduction of Twitter API and its limitation is presented. We then introduce several metrics, which are mainly developed based on those described in Chapter 2 with further improvements. We then show how the ranking is conducted. Finally, some ranking results are shown and conclusions are drawn.

## 3.2 Twitter API

Twitter API is officially provided by Twitter to programmers for accessing the data. There are two types of API: Streaming APIs and REST APIs. Streaming APIs provide the stream data to the developer hardly with any delay or overhead. It is usually used for monitoring the user's tweets in real-time.

Unlike Streaming style APIs, REST APIs are used for getting the data from Twitter, such as user's home timeline by given username. They allow users to establish a connection with Twitter and perform requests. The only authorized way to access Twitter REST API is OAuth (Open Authorization). OAuth is an open standard that allows the third party to apply the users' resource without touching the users' account information, such as username, password, so the security level of OAuth is high.

However, there is a limitation for developers using Twitter API. For REST type API, Twitter API defines 15 minutes as an interval and it only allows 180 requests in one interval. If the user passes the limit, an HTTP 429 type error will be sent to the user, which means too many requests were submitted.

In this thesis, we used Twitter REST API V1.1. Some parts of the code are attached in Appendices.

## 3.3 Metrics

### 3.3.1 Follower Rank

In Chapter 2, the concept of Follower Rank was defined by the following:

$$\text{Follower Rank} = \frac{\text{Follower Number}}{\text{Follower Number+Followee Number}} \quad (3\text{-}1)$$

It illustrates the connections of a user, and in the IARank method, the researcher uses the same expression for the concept of Structural Advantage which measures the potential to broadcast the user's influence.

The range of the Follower Rank is from 0 to 1. However, if the value is smaller than 0.2, it indicates the user follows many user accounts but doesn't get many users following back; on the other hand, if the value equals 1.0, it means the user doesn't follow other users, which is not a good influential user sign as well.

### 3.3.2 Iterative Follower Rank

The Iterative Follower Rank is a metric developed from Follower Rank. Follower Rank only takes account of the user's follower and followee number. However, Iterative Follower Rank also considers user's social network. Here we assume the user as layer 0, and the user's follower as layer 1.

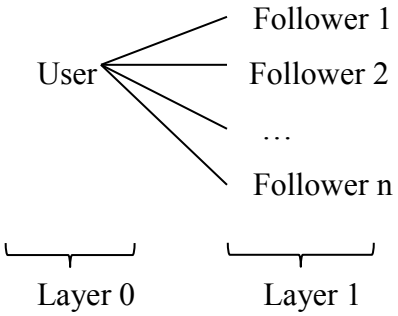


**Figure 3. 1 Principe of the Iterative Follower Rank**

The Iterative Follower Rank can be expressed as follows:

$$Iterative\ Follower\ Rank = a * Follower\ Rank_{Layer\ 0} + b * Average\ Follower\ Rank_{Layer\ 1} \quad (3\text{-}2)$$

Parameter *a* is the weight variable for the layer 0, and *b* is the weight variable for layer 1. In this thesis, we use 0.5 for *a* and 0.5 for *b*. In this way, layer 0 and layer 1 have both been taken into account when measuring the user's influence potential.

### 3.3.3 Compromised IArank

The IARank method was described in chapter 2. This is formed of Structural Advantage and Buzz. However, due to the limitation of the Twitter API, we can no longer get the number of times a single user is mentioned in real-time to determine Buzz. Therefore, we need to use the Retweet times of the user's most recent 100 tweets to replace the mention times in Buzz. The new Buzz can be expressed as following:

$$New\ Buzz = \frac{\sum Retweet\_count}{\text{Total } number\ of\ tweet} \quad (3\text{-}3)$$

As the Compromised IARank is composed of two parts, so we define $\alpha$, $\beta$ as the weight for Structual Adavantage and Buzz separately. It can be defined as follows:

$$Compromised\ IARank = \alpha * Structual\ Advantage + \beta * New\ Buzz \quad (3\text{-}4)$$

Here α, β both take 1.0.

## 3.4 Working Flow

A simulation of the above metrics was developed on a laptop with regular Internet connection. The keyword applied is RIO2016, and the results were obtained in real-time. The top 10 influential user rankings by Follower Rank, Iterative Follower Rank, Buzz and Compromised IARank are shown in the following tables separately.

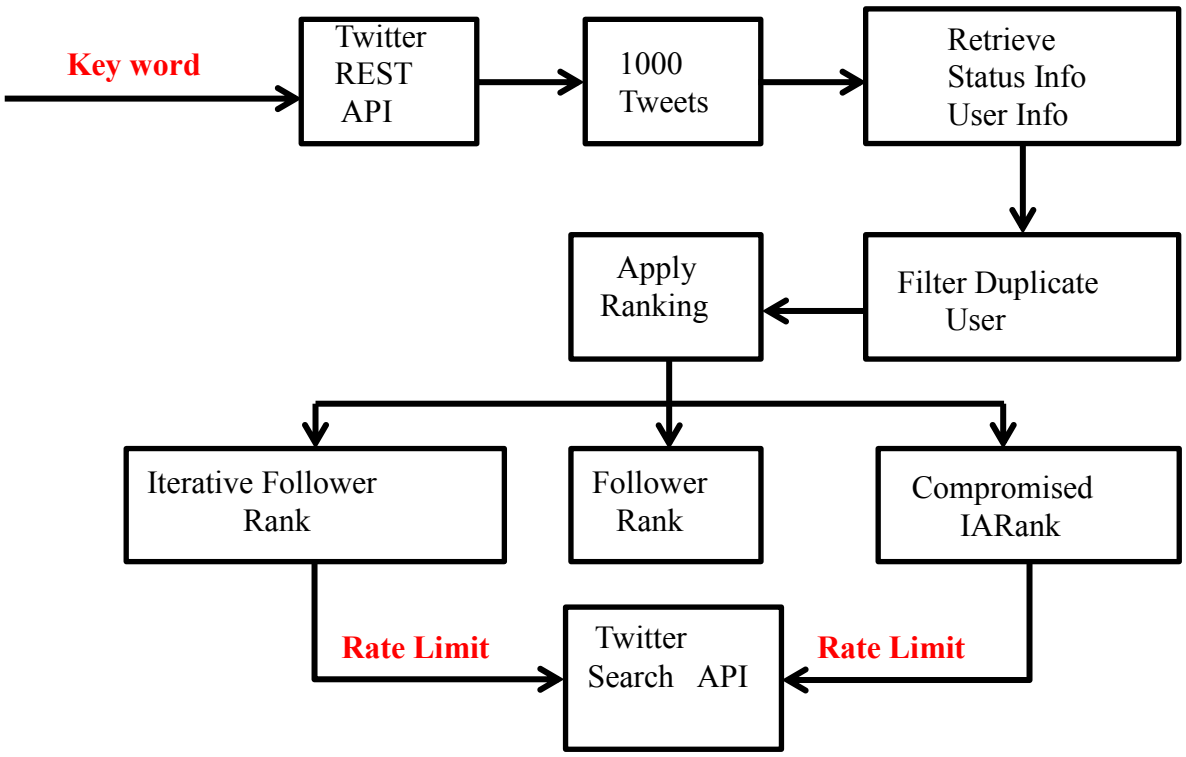


**Figure 3. 2 working flow diagram**

## 3.5 Results and Discussion

### 3.5.1 Experimental Results

The experiment is running on a laptop with regular Internet connection, the key word applied is RIO2016, and the results can be obtained in a real-time scale. The top 10 influential user rankings by Follower Rank, Iterative Follower Rank, Buzz and Compromised IARank are shown in the following tables separately.

**Table 3. 1 Top 10 user ranking by Follower Rank**

| User Name | Follower Rank |
|---|---|
| Ashleymckenzi12 | 0.987 |
| Loxollg | 0.961 |
| PaulGarciaPS | 0.914 |
| ____SadBoy | 0.913 |
| Theywillbefree | 0.912 |
| zinfo67 | 0.899 |
| Quantocksram | 0.89 |
| Margauxdelys | 0.887 |
| Javsxo | 0.883 |
| PaulaHorann | 0.873 |

From the Follower Rank result, the top 10 most influential users all get a score higher than 0.87, which means they all have more followers than followees. The first place is a judo player from Team GB who has 36k followers and only 479 followees. Third place is held by an NBA writer who has 6.9k followers and 650 followees. By using the Follower Rank scheme,

high score users are the group of users who have much more followers than followees, no matter their occupation. Using this scheme, user social networks cannot be known. It is also unknown if their tweets can be transferred by their followers.

However, the Iterative Follower Rank takes the user's followers' network into account. The first place is taken by an active user 'javsxo' who has 1.4k followers and 199 followees. This user is also in the top 10 Follower Rank list but not in the top 5.

**Table 3. 2 Top 10 user ranking by Iterative Follower Rank**

| UserName | Iterative Follower Rank |
|---|---|
| javsxo | 0.513 |
| zinfo67 | 0.497 |
| ericasara | 0.467 |
| ____SadBoy | 0.463 |
| syiramxhd | 0.461 |
| bradsdabomb | 0.459 |
| paleblueeyes24 | 0.451 |
| FMMtweets | 0.451 |
| pprettywindy | 0.449 |
| PaulaHorann | 0.448 |

The Buzz score is mainly designed to measure the potential of the user to attract attention from others. The first place user is the same user as Iterative Follower Rank, and the second place is taken by a fiction writer. Compared to the top 2 users, the other eight users' Buzz score is quite small.

**Table 3. 3 Top 10 user ranking by Buzz Score**

| User Name | Buzz |
|---|---|
| Javsxo | 0.288 |
| Peteryeoh | 0.233 |
| ohhthatslil_ | 0.07 |
| ____SadBoy | 0.068 |
| Laurinhabn | 0.045 |
| baby_Barnson | 0.026 |
| Unitedjohnsonn | 0.025 |
| Margauxdelys | 0.013 |
| Inthyo | 0.011 |
| InesGraveto | 0.009 |

The Compromised IARank is a combination of Follower Rank and Buzz score. Because of the large advantage of Buzz score, the top 2 users in Buzz still rank top 2 in Compromised IARank. The third place is taken by the first place user in Follower Rank.

**Table 3. 4 Top 10 user ranking by Compromised IARank**

| User Name | Compromised IARank |
|---|---|
| javsxo | 1.171 |
| peteryeoh | 0.988 |
| Ashleymckenzi12 | 0.987 |
| ____SadBoy | 0.981 |
| loxollg | 0.961 |
| PaulGarciaPS | 0.914 |
| theywillbefree | 0.912 |
| zinfo67 | 0.903 |
| margauxdelys | 0.9 |
| Quantocksram | 0.891 |

### 3.5.2 Discussion and Conclusion

From the simulation, we can find that the results from different ranking schemes are different. However, there is correspondence between them if we do a pairwise comparison.

#### 1. Follower Rank and Iterative Follower Rank

From Table 3.5, it can be shown that four of the same users appear in the top 10 list of the two ranking schemes. These four users are all active users who use Twitter a lot as a social network service. From their profile, they are not celebrities, journalists, or anyone who holds another special occupation; just ordinary people but active in Twitter. In the Follower Rank scheme, two kinds of users get high scores: users who have a lot of followers and users who seldom follow other user. Considering the Iterative Follower Rank scheme, it can help to filter the user whose followers' Follower Rank score is low.

**Table 3. 5 Comparison of Follower Rank and Iterative Follower Rank**

| Follower Rank | Iterative Follower Rank |
|---|---|
| Ashleymckenzi12 | javsxo |
| loxollg | zinfo67 |
| PaulGarciaPS | ericasara |
| ____SadBoy | ____SadBoy |
| theywillbefree | syiramxhd |
| zinfo67 | bradsdabomb |
| Quantocksram | paleblueeyes24 |
| margauxdelys | FMMtweets |
| javsxo | pprettywindy |
| PaulaHorann | PaulaHorann |

### 2. Buzz and Compromised IARank

From Table 3.6, it can be shown that there are four of the same users in the Buzz and Compromised IARank. Moreover, the top two users are exactly the same. Buzz score is used to measure the ability of the user to get attention from others. The Compromised IARank considers both Buzz score and the Follower Rank score, which means high score user maintain both high potential to receive attention from others and high potential to broadcast their influence.

**Table 3. 6 Comparison of Buzz and Compromised IARank**

| Buzz | Compromised IARank |
|---|---|
| javsxo | Javsxo |
| peteryeoh | Peteryeoh |
| ohhthatslil_ | Ashleymckenzi12 |
| ____SadBoy | ____SadBoy |
| laurinhabn | Loxollg |
| baby_Barnson | PaulGarciaPS |
| unitedjohnsonn | theywillbefree |
| margauxdelys | zinfo67 |
| Inthyo | margauxdelys |
| InesGraveto | Quantocksram |

### 3. Follower Rank and Compromised IARank

From Table 3.7, it can be shown that there are nine of the same users in Follower Rank and Compromised IARank. That's to say, these two ranking schemes have quite high interdependency. The one in Compromised IARank but not in Follower Rank, is a fiction writer who doesn't have too many followers. That proves again the flaw of Follower Rank which is the

inability to find users who don't have many followers but are followed by users interested in a particular topic.

**Table 3. 7 Comparison of Follower Rank and Compromised IARank**

| Follower Rank | Compromised IARank |
|---|---|
| Ashleymckenzi12 | Javsxo |
| loxollg | Peteryeoh |
| PaulGarciaPS | Ashleymckenzi12 |
| ____SadBoy | ____SadBoy |
| theywillbefree | Loxollg |
| zinfo67 | PaulGarciaPS |
| Quantocksram | theywillbefree |
| margauxdelys | zinfo67 |
| javsxo | margauxdelys |
| PaulaHorann | Quantocksram |

### 4. Iterative Follower Rank and Compromised IARank

From Table 3.8, it can be shown that there are only three of the same users in both Iterative Follower Rank and Compromised IARank. These two ranking schemes are the least correlative, compared with other combinations. The mutual users are those users who have both high quality followers and high potential to broadcast their influence.

**Table 3. 8 Comparison of Iterative Follower Rank and Compromised IARank**

| Iterative Follower Rank | Compromised IARank |
|---|---|
| javsxo | javsxo |
| zinfo67 | peteryeoh |
| ericasara | Ashleymckenzi12 |
| ____SadBoy | ____SadBoy |
| syiramxhd | loxollg |
| bradsdabomb | PaulGarciaPS |
| paleblueeyes24 | theywillbefree |
| FMMtweets | zinfo67 |
| pprettywindy | margauxdelys |
| PaulaHorann | Quantocksram |

### 5. Discussion

Follower Rank is good at finding the user who has celebrity quality, which means they have a lot of followers. The weak point is it can't find the user who has the specialist quality, which means the user doesn't have many followers but are followed by users interested in a special area. Compromised IARank make up this kind of loss by combining the Follower Rank and Buzz together. This scheme can find the user that holds both high potential to influence others and the ability to spread the influence.

## 3.6 Conclusion

In this chapter, we discussed the realization of using the metrics described in the literature review to measure the influence of users, such as IARank. I chose a basic ranking scheme (Follower Rank) as the starting point, and successfully developed it into Iterative Follower Rank. The advantage of the Iterative Follower Rank is to filter the user who doesn't

have good quality followers. Because of the limitation of Twitter API, we evolved the IARank to Compromised IARank. Then I compared the results with Follower Rank and Iterative Follower Rank. After looking deeply into the user's Follower Rank score and Buzz score, I found some interesting relationships between the score and the occupation of the user. This interesting observation makes me want to investigate the relationship between the influence and the user's occupation. A new method of classification of the user will be presented in detail in the next Chapter.

# 4. 2D User Classification by Twitter Profile Statistics

## 4.1 Chapter Overview

In this Chapter, we introduce a 2D classification of the user based on the user's basic profile statistics. Section 4.2 introduces a detailed description of the problem, and section 4.3 presents the potential key features to be used in the method. As the method contains two basic elements: local influence and global influence, the forth part presents the local influence, including the definition, the factors and the significance. Section 4.5 focuses on the global influence, and section 4.6 provides classification of users by influence features, followed by the conclusions.

## 4.2 Problem Statement

The discussions of the metrics in Chapter 3 lead us to another interesting finding about certain groups of users share similar influence score. H. Schoen, D. Gayo-Avello, P. Takis Metaxas, E. Mustafaraj, M. Strohmaier, and P. Gloor are working on the method of classifying users in Twitter, for a variety of applications such as, political campaign, prediction models [28] and product marketing. For example, Pennacchiotti, Marco, and Ana-Maria Popescu used the user behavior, network structure and semantic analysis of the user's tweet as key features to predict the user's political preference, ethnicity and attitude towards Starbucks [29].

The previous method to classify users heavily relied on the semantic analysis of the user's tweets which has obvious flaws. The tweets can be about any topic, thus, it is difficult to deal with various subjects of the tweets in a precise manner. Besides, the negation of the sentence, irony, the utilization of symbols and emote icons also make it hard to analyze the tweet correctly.

Here, we present a method based on the user's observable profile statistics as candidate features to achieve the classification of the users in different groups including, politicians, businessman, entertainment star and general public.

## 4.3 Key Features

In Chapter 3, we discussed the limitations in Twitter APIs when retrieving user's data. However, there are still some useful user profile statistics that can be used. Figure 4.1 is an example of truncated status retrieved from the user's home timeline. The status stands for the

tweet object in the Twitter API and contains not only the tweet information but also the author's information.

```
1 Status(
2 contributors=None,
3 retweet_count=1,
4 text=u'Good weather in Tbay!!!',
5  is_quote_status=False,
6  in_reply_to_status_id=None,
7  id=722515643644407808L,
8  favorite_count=0,
9  _api=<tweepy.api.API object at 0x021F1690>,
10 author=User(follow_request_sent=False,
11  has_extended_profile=False,
12  profile_use_background_image=True,
13 _json={
14  u'follow_request_sent': False,
15 u'has_extended_profile': False,
16 u'profile_use_background_image': True,
17  u'default_profile_image': False,
18  u'id': 2865156349L,
19 u'profile_background_image_url_https':
20 u'https://abs.twimg.com/images/themes/theme1/bg.png',
21 u'verified': False,
22 u'profile_text_color': u'333333',
23 u'profile_sidebar_fill_color': u'DDEEF6',
24 u'entities': {u'description': {u'urls': []}},
25 u'followers_count': 4,
26 u'profile_sidebar_border_color': u'C0DEED',
```

```
      27 u'id_str': u'2865156349',
      28  u'profile_background_color': u'C0DEED',
      29 u'listed_count': 0,
      30 u'is_translation_enabled': False,
      31 u'utc_offset': None,
      32 u'statuses_count': 12,
      33 u'description': u'',
      34 u'friends_count': 22,
      35 u'location': u'',
      36 u'following': False,
      37 u'geo_enabled': False,
      38 u'profile_background_image_url':
u'http://abs.twimg.com/images/themes/theme1/bg.png',
      39 u'screen_name': u'Rice_riz_roll',
      40 u'lang': u'zh-Hans',
      41 u'profile_background_tile': False,
      42 u'favourites_count': 0,
      43u'name': u'shan',
      44 u'notifications': False,
      45u'url': None,
      46 u'created_at': u'Sun Oct 19 14:13:17 +0000 2014',
      47u'time_zone': None,
      48u'protected': False,
      49 u'default_profile': True,
      50 u'is_translator': False},
      51 u'created_at': u'Tue Apr 19 20:02:07 +0000 2016',
```

**Figure 4. 1 Example of Twitter Status Object**

In Figure 4.1, the potential candidate features that indicate the influence of the user and the tweet that the user composed are demonstrated.

- Retweet count: The number of times the tweet has been retweeted can be found in line 3.

- Tweet ID: Each tweet has a unique ID that can be easily accessed for tweet details and content in line 7.
- Favorite count: The number of times the tweets indicated favorite by the user is shown in line 8.
- Created time: The time and date information of the tweet's creation time is found in line 51.
- User ID: The author's ID is uniquely assigned by Twitter which cannot be changed. This is highlighted in line 18 and in line 27, the user id is indicated in string format.
- Follower count: The number of the user's followers is an important element to show the user's social network, as highlighted in line 25.
- Friends count: The number of the user's 'followees' indicate the scope of the user's social network, as reflected in line 34.
- User created time: The time when the user started to use this Twitter account is reflected in line 46.

## 4.4 Local Influence

### 4.4.1 Motivation

From the observation of the Twitter user's key social features, it can be observed that a user can make strong influence on his or her surroundings such as, friends, and family members. This kind of ability of influencing is defined as an important local actor. This local actor is independent of how his or her influence reaches across the entire global social network. The

other motivation of my work is the limitation of the variables retrieved from Twitter APIs. For example, the lifelong number of retweets or the number of mentions used in the past research mentioned in the background review, are no longer publicly accessible through the Twitter APIs. Thus, we have designed our metrics to use more general values.

### 4.4.2 Definition

The user's local influence is the impact that the user has on their immediate surroundings. This is measured in two parts. In the first part, the attention that the user's messages receive from their contacts is considered. The second part however focuses on the frequency the user writes a tweet.

Social networks allow users to note interesting messages in a variety of ways, such as marking them as liked, sharing them, or following them. On Twitter, two of these methods are available—retweeting and marking as favourite. However, researchers have found them to be strongly correlated [30]. The focus is on only one of these two mechanisms namely retweets. Moreover, due to the limitation of Twitter API which only allows developers to retrieve a user's most recent 200 tweets and their retweet count, it is decided to get the user's recent 100 tweets, the average number of retweets. However, such a measure would be strongly biased in favour of users that have more friends or followers. To account for this, we make an average of the number of recent retweets by the number of followers the user has (also available from the Twitter API). Finally, as can be seen in equation (4-1), we use log functions in order to account for the size of the values involved, and the final result is named Local 1.

$$Local\ 1 = \frac{Log(\#Retweets)}{Log(\#Followers \times 100)} \tag{4-1}$$

Local 1 can be seen as the rate that the user's follower retweet their tweet. This value indicates the user's power of influencing their followers. In most cases, Local 1 is inferior to 1, and the phenomenon that user's Local 1 is rising in a short time that usually indicates the hot topic. For example, the Leicester's city mayor's Local 1 value increases because the Leicester's city football team unexpectedly won the championship of the Premier League for the first time, and the mayor's tweet which related to this event has been retweeted for many times.

The second part of the local score focuses on the frequency of the user writing a tweet, as indicated in equation (4-2). If a user composes relatively few messages, he or she will have a smaller impact on the followers. However, if a user who posts too many messages, it does not necessarily have a greater influence, since not all messages will be seen or read, or the messages may be part of an ongoing conversation. Going back to the Twitter APIs, the date of the first and last message in the set of 100 most recent tweets of a user can be accessed, and the number of days between those dates is computed. The log of that value (adding 1 to avoid negative numbers in the case of users with less than one tweet per day), and the final result is named Local 2.

$$Local\ 2 = Log\left(\frac{100}{\#days} + 1\right) \tag{4-2}$$

Local 2 reflects the user's ability of producing tweet, so the news feed accounts usually dominate the top of the ranking list. Some promoting accounts can also be productive. Therefore, there is a chance that high Local 2 score account can be a spamming account. On the other hand, low Local 2 score can be a sign of inactive account. However, there existing influential user does not tweet frequently but each tweet has good quality, and that is why Local 1 score and Local 2 score are combined to measure the user's local influence.

The total local score is the product of these two values, as shown in equation (4-3).

$$Local = \min(1, Local\ 1 \times Local\ 2) \quad (4\text{-}3)$$

Local 1 is the rate that the follower retweets the user's tweet. Local 2 is the frequency of the users' composing a tweet. Therefore, the product of Local 1 and Local 2 gives the impact power of a user for influencing their surrounding contacts. Usually, both Local 1 and Local 2 are below 1, thus, the Local influence score is usually below 1. In rare cases, for instance if the user has made a recent flurry of activity, or has published a long message as a series of tweets, the local score can be inflated above 1. However, since such peaks are artefacts of the Twitter setup and the need to break a message as 140 characters, we can cap the maximum at 1 to reduce the effect of such outliner values. Judging from the result, the user with Local Influence value varies from 0.5 to 1 has a high local influence on their direct surroundings. However, the user with the exact value of 1 does not mean that he or she is locally most influential person.

## 4.5 Global Influence

### 4.5.1 Motivation

Chapter 3 highlighted that the user could have impact on the entire social network. More followers mean the user have the potential to spread their influence. However, the Follower Rank which is mentioned in chapter 3 is not able to indicate the user's influence power due to lack of accuracy. The Follower Rank score is the result of the follower number divided by the total number of followers and followees. For example, the USA President Obama's Follower Rank score is similar to the Colorado Governor Hickenlooper's, but it does mean that they have similar global impact, as President Obama has 1,100 times more followers than Governor

Hickenlooper. In this section, a new mechanism to measure the global influence which is taking into account both the entire social network and the user's personal profile is introduced.

### 4.5.2 Definition

Global influence is the impact the user has globally on the entire social network. Taking into account of the user's follower number, and the research metric of the user's lifetime activity is considered. The research metric is composed of two parts. The first part is the proportion of the network that is paying attention to the user's messages. On Twitter, this is the ratio of the user's number of followers to the total number of users, or 310 million, and considering the size of the number, we use log function, as shown in equation (4-4), and is named Global 1.

$$Global\ 1 = \frac{Log(\#Followers)}{Log(\#\text{Total } Users)} \tag{4-4}$$

Global 1 focuses on the user's ability to broadcast the message to the entire global network. The range varies from 0 to 1. For example, pop singer Kate Perry who has the most followers, has the highest Global 1 value, near to 0.936.

The second part is the user's activity compared to the number of users following him. The assumption here is that a user that posts scarcely but is followed by a large crowd must have a greater impact on the network than either one who posts more but is followed by a smaller group or one that posts as much but is follower by fewer people. We compute this metric as the ratio of the user's total number of tweets to his total number of followers, taking 1 minus that value, and setting a minimum score of zero, as shown in equation (4-5). This value, named Global 2 is computed as follows:

$$Global\ 2 = max\left(0{,}1 - \frac{\#Tweets}{\#Followers}\right) \quad (4\text{-}5)$$

Accounting the user's productivity of composing tweet, Global 2 can be described as the power of influencing global network by each tweet. Suppose that user A and user B, both have 1000 follower. However, user A has published only100 tweets and user B published 10,000 tweets in total. In this case, user A has high global influence than user B. In other words, if two users have similar number of followers, the less tweet the user compose, the larger power each piece of tweet has, and the greater global influence the user has. The value of Global 2 varies from 0 to 1. Also this metric helps to filter some advertising account with fake followers and popping irrelevant messages.

Finally, the total global score is the product of Global 1 and Global 2, as shown in equation (4-6).

$$Global = Global\ 1 \times Global\ 2 \quad (4\text{-}6)$$

In this way, we have the new metric to measure the user's global influence. Recalling the example mentioned in the motivation section, and using the equation 4-6 for global influence, President Obama gets a global influence value as 0.9259, and the Governor Hickenlooper's global influence value is 0.3108. Compared to the Follower Rank score, the metric shows a great difference between these two users on the ability of influencing the global social network.

## 4.6 Classification of Users by Influence Features

With the two key influence features—Local Influence and Global Influence, a variety of users' data as the training set are collected. When we plot the influence measures from the training set, users having similar influence features cluster together, and they can be defined as the center for each cluster of users. When a new user data is introduced, the Local Influence and Global Influence can be calculated and the distance between the new user and the cluster center can be computed. The new user is then classified into a cluster that is closest to it.

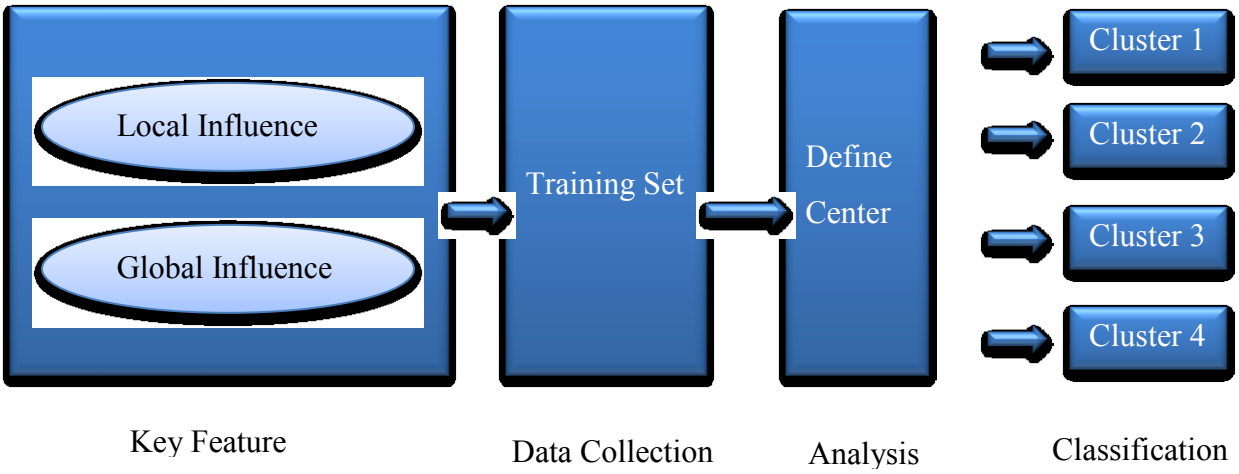


**Figure 4. 2 Classification of Users by Local Influence and Global Influence**

## 4.7 Conclusion

In this Chapter, we presented a new method of classifying Twitter users. We began by presenting important user features of social network. To achieve the classification of the users, two new designed metrics to measure user's influence: Local Influence score and Global Influence score were introduced. Regarding the Local Influence score, that measure user's power of impacting their direct connection in the social network, the account rate that the follower

retweet the user's tweet is taken into consideration, but also the user's ability of producing tweet is considered. For the Global Influence score which measure user's global impact on the entire social network, the user's ability to broadcast the message and the power of influencing global network by each tweet together to balance user's impact globally is combined.

For both of the metrics, we enhance the performance of making a distinction between two users who share similar social network statistical numbers, such as follower number.

In conclusion, the modeling method on the use of the influence features to classify users in Twitter is presented. In the following chapter, the entire experiment and results are discussed in details.

# 5. Experimental Results

## 5.1 Chapter Overview

This Chapter aims to present the experiments and results that prove the efficiency and accuracy of the new classification method mentioned in Chapter 4. Then, follows the presentation of the data collection used in this method. The test data sample can be classified into four categories according to their occupation. As well, the test data samples can be put into four clusters according to the features of the local influence and global influence. Each category and each cluster is discussed in details. The fourth part of this chapter shows all the results, and the fifth part gives the benchmarks. The results obtained from the new method are compared with the results using IARank and Klout system separately.

## 5.2 Data Collection

To capture their features, we collected users' social network information to compute the Local Influence score and Global Influence score. We took information of 234 users as the training database, including 107 politicians (from federal, provincial and municipal levels), 33 celebrities (movie stars, pop singers, TV hosts, sport stars), 56 businessmen (CEO, management from large international companies, middle size companies to small start-up companies), and 38 general public users. The test set had 50 users which included 13 politicians, 11 celebrities, 14 businessmen and 12 general public users, as shown in Table 5.1.

**Table 5. 1 Organization of the Training set and the Test set**

| | Politician | Celebrity | Businessman | General Public | Total |
|---|---|---|---|---|---|
| Training | 107 | 33 | 56 | 38 | 234 |
| Test | 13 | 11 | 14 | 12 | 50 |

Due to the popularity of using Twitter in North America, the politicians are mainly picked from USA and Canada and include federal, provincial and municipal. Besides, I also collected different level politicians from other countries such as UK, France and Australia. The general public users are picked randomly for the training set and test set by finding users who are related to the keyword 'good day', which is a general topic.

## 5.3 Four Categories and Four Clusters

As mentioned above, we collected user information from four different categories: politician, celebrity, businessman and general public user. The users' local and global scores

were computed to see if there was any correspondence between the category and the score. Besides this approach, we also used the IBM SPSS statistical software suite[2] to find the clusters which could gather the users together.

### 5.3.1 Four Categories

To find the social network features of different groups, we collected their information through Twitter API. In this part, we analyzed the user's social information features for each category in details separately.

#### 5.3.1.1 Politician

Regarding the politician group, we collected the information of politicians from different levels: Federal, Provincial and Municipal. We also collected the politicians' information from different regions and countries, and due to the rate of Twitter usage in North America, a high proportion of politicians from USA and Canada are selected in this training set. The center of the politician category's influence feature is Local Influence (0.3366) and Global Influence (0.3693).

[2]https://www.ibm.com/support/knowledgecenter/SSLVMB_21.0.0/com.ibm.spss.statistics.help/idh_quic.htm

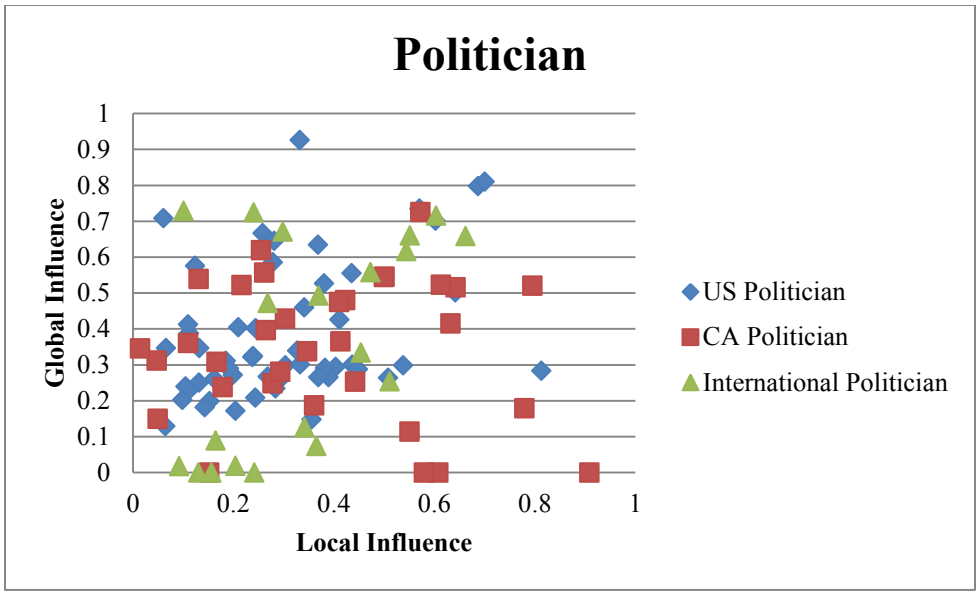


**Figure 5. 1 World Politicians' Influence Features**

Figure 5.1 illustrates the politicians' influence features, including 54 American politicians, 33 Canadian politicians and 20 international politicians from other countries (UK, France, and Australia). From the results, it can be seen that American politicians have a higher average Global Influence (0.3858) than the politicians from other countries. It can also be seen that Canadian politicians have a higher average Local Influence (0.3880) than the others. The Global Influence and the Local Influence are the influence measures computed according to the equation (4-6) and (4-3). These values indicate American politicians have stronger influence on the global social network (higher Global Influence than the average of politician category) while Canadian politicians have stronger impact on the direct surroundings (higher Local Influence than the average of the politician category). It should be noted here that these results cannot be revealed by the existing influence measures.

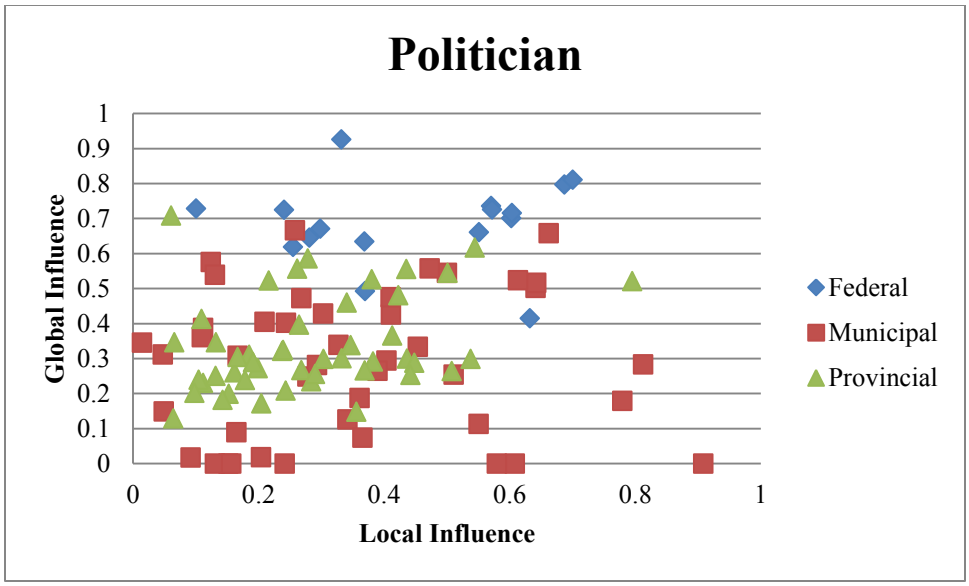


**Figure 5. 2 Influence Features of Politicians from Different Levels**

Figure 5.2 shows the influence features of politicians from different levels: Federal, Provincial and Municipal. There are 16 federal politicians including presidents and party leaders from different countries. They have a significantly higher average of Global Influence (0.6874) than politicians from the other two levels, and President Obama has the highest Global Influence (0.9259). There are 44 city mayors from USA, Canada, England, France and Australia, and 47 provincial politicians including governors and premiers. The importance of introducing Global vs Local influence measures is clearly demonstrated in this experiment. The results show that provincial politicians have a higher average of Global Influence (0.3378) than municipal level politicians, but municipal politicians have a higher average of Local Influence (0.3523).

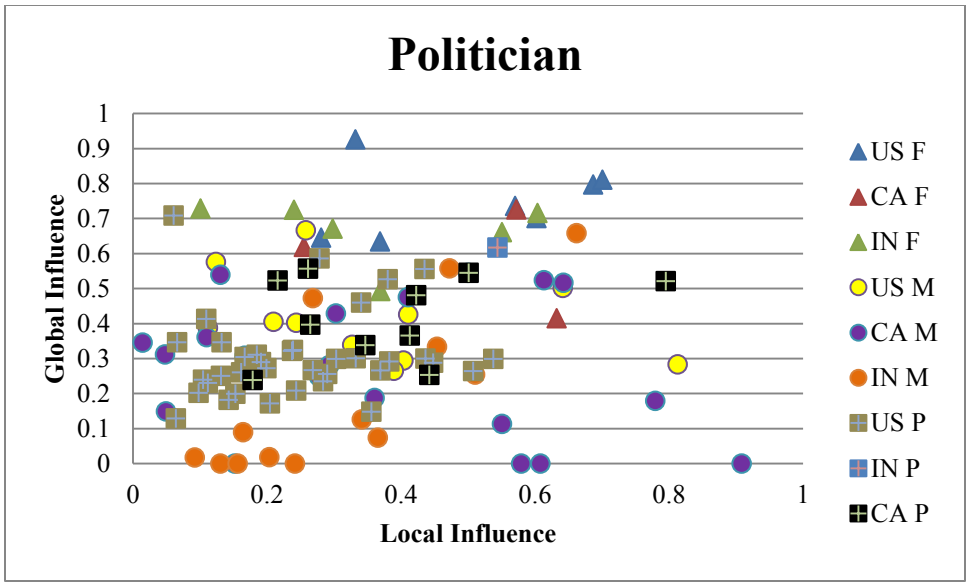


**Figure 5. 3 Influence Features of Politicians**

The influence features of different levels from different countries are marked in different colors and shapes in Figure 5.3. The federal level politicians from USA have the highest Global Influence (average: 0.7497). However, unlike President Obama (Local Influence: 0.3321), the four Presidential candidates: Hilary Clinton (0.6872), Donald Trump (0.7001), Bernie Sanders (0.5706) and Ted Cruz (0.6026) have a quite high Local Influence. President Obama's high Global Influence is due to his follower numbers (higher value in equation 4-4); his lower Local Influence is mainly because of the lower rate he writes tweets (lower value in equation 4-2). And this can be seen as the difference of a world leader and a candidate in the campaign. Overall, provincial level politicians' Local Influence is not as high as municipal level politicians. However, Alberta Premier Rachel Notley showed a very high Local Influence (0.7955), mainly because the Fort McMurray wild fire makes her tweets get more attention from the local people. Regarding the Global Influence, the mayors of large international cities also have a high impact on the entire global network. For example, Paris Mayor's Global Influence is 0.6580; New York

Mayor's Global Influence is 0.6660; both due to the high follower numbers (high value in equation 4-4).

#### 5.3.1.2 Celebrity

Regarding the celebrity category, we collected data of pop singers, movie stars, TV entertainers and sports stars. The center of the celebrity category's influence feature is Local Influence (0.2068) and Global Influence (0.7394). From the results in Figure 5.4, we can observe that most samples in this category have a quite high average of Global Influence (0.7394). Compared to the Global Influence, their average Local Influence is low, which is only 0.2068. However, we do find a TV entertainer (Kim Kardashian) who has both high Global Influence (0.9020) and Local Influence (0.6216). She is really active in producing tweets, so she received a high value in equation (4-2); meanwhile, she also has many followers retweeting her tweets causing high values in both equation (4-4) and (4-5). The results above show the power of the proposed measures in providing accurate picture of users' influence.

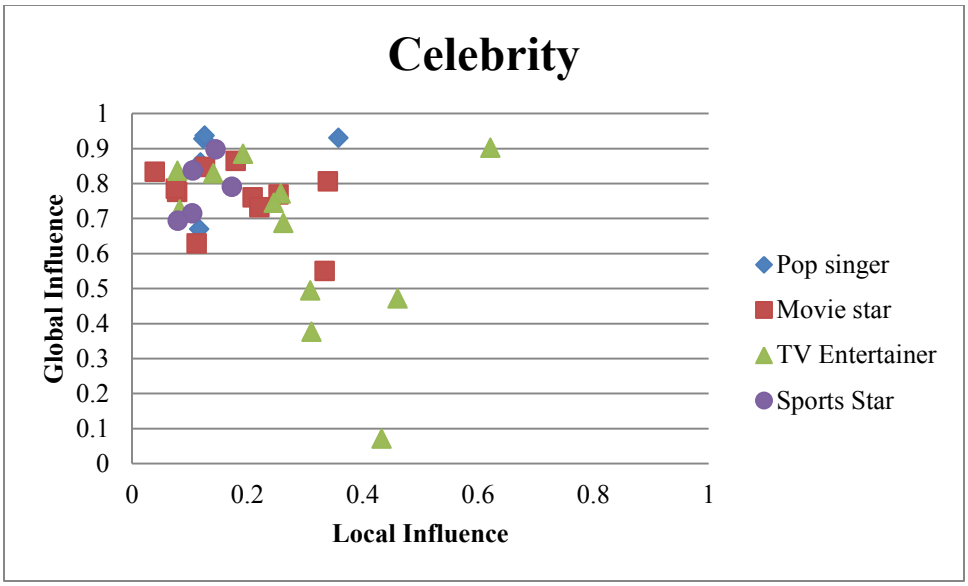


**Figure 5. 4 Influence Features of Celebrity Category**

#### 5.3.1.3 Businessman

Regarding the businessman category, the social network information of the CEOs of different size companies from different industries has been collected. In this work, we define the size of company by the number of their employees as shown in table 5.2. These were retrieved from LinkedIn.

**Table 5. 2 Definition of the company size**

| Size | Number of Employees |
|---|---|
| Large | 10,000+ |
| Medium | 200-10,000 |
| Small | 1-200 |

The center of the businessman category's influence feature is Local Influence (0.1694) and Global Influence (0.4519). And the businessmen's influence feature is illustrated in Figure 5.5.

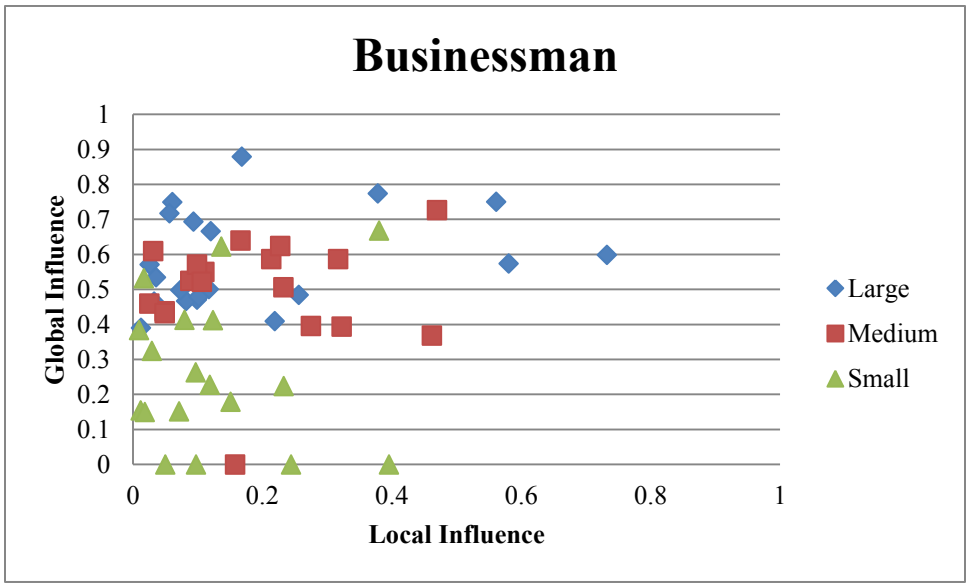


**Figure 5. 5 Influence Feature of Businessman**

In general, the businessman category's Local Influence is not high, except for one sample: Marc Benoiff, the CEO of salesforce.com. Marc's Local Influence is 0.7325, and he is not only a successful Internet entrepreneur, but also an active philanthropist that attracts a lot of public attention. Another thing to notice is there are some famous CEOs who have the Global Influence at the celebrity category's level, such as Bill Gates (Global Influence: 0.8788) and Elon Musk (Global Influence: 0.7740). From the results, it can be seen that large size company CEOs have higher Global (0.5835) and Local Influence (0.1905) compared to CEOs from medium and small sized companies.

#### 5.3.1.4 General Public User

In the training set, we also collected some random ordinary users' information in order to capture the influence feature of different categories. The sample of the general public user was picked at different times and days by searching subjects such as 'good day'. The influence feature of this category is illustrated in Figure 5.6. The center of the general public user category's influence feature is Local Influence (0.5546) and Global Influence (0.0159). Most samples in this category have lower Global Influence, and only two samples scored relative high Global Influence. After checking the two sample's profile, it was observed that one of them was a comedian (Global Influence: 0.3394), and the other was a chef of a restaurant in Times Square (Global Influence: 0.1955), thus explaining the obtained results.

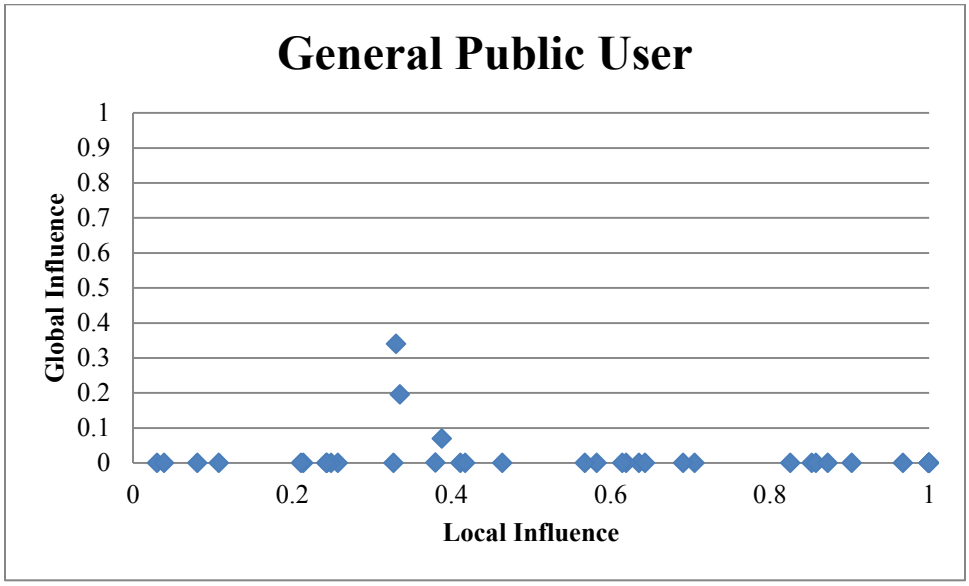


**Figure 5. 6 Influence Feature of General Public User**

#### 5.3.1.5 Conclusion

Figure 5.7 shows the regions by different categories of users. First, we can observe most of the ordinary users are at the bottom of the global axis, with an average Global Influence of 0.0159. Their Local Influence, however, varies the most of any class of users, going from almost 0 for a user with almost no followers or retweets to 1 for users who are very popular in their immediate circle of friends, leading to large variations in the result of equation (4-1) for that category.

Secondly, the celebrity category has the highest score among all the other categories in Global Influence (average 0. 7394) because of their very high number of followers in equation (4-4). Meanwhile, their Local Influence scores are mainly clustering in a low Local Influence region (average 0.2068), mainly because many of them simply do not tweet much, leading to a low value of equation (4-2).

Thirdly, the business CEO category seems to be a variation of the celebrities' category, with both fewer followers and fewer tweets, leading to lower global and local scores (average 0.4519 and 0. 1694 respectively).

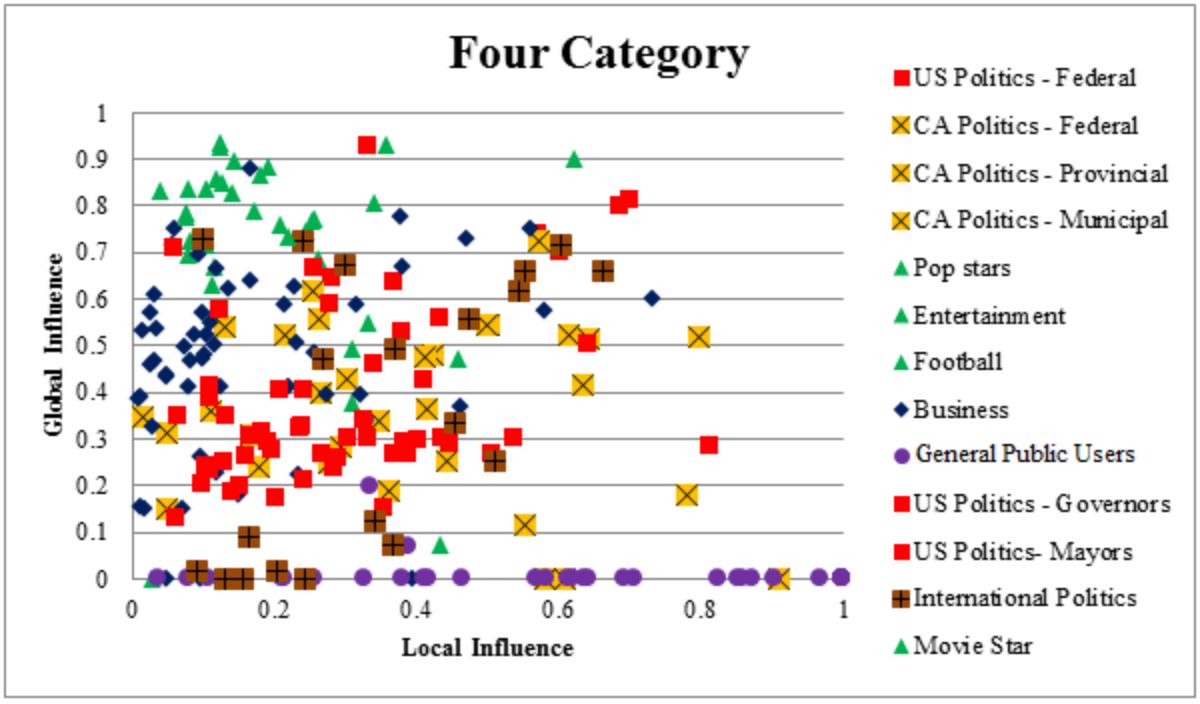


**Figure 5. 7 Local and Global Influence of Different Categories of Users**

Finally, the results for the political category show that they are distributed much more irregularly. Unlike the other three categories, they show no dominant cluster, and politicians seem to be found over the entire graph area. Moreover, there is important overlap between the scatter of US politicians (average Global: 0.3853, Local: 0.3043), Canadian politicians (average Global: 0.3480 Local: 0.3880), and international politicians (average Global: 0.3604 Local: 0.3387). The distinction between national, regional, and municipal politicians is a bit more notable. The cluster of national politicians (average Global: 0.6874, Local: 0.4479) shows a higher global average influence than that of regional politicians (average Global: 0.3378, Local: 0.2840), which in turn has a higher global average than that of municipal politicians (average Global: 0.2875, Local: 0.3523). However, the relationship clearly does not hold for Local Influence. The center of each category is shown in Table 5.3.

**Table 5. 3 Center of Four Categories**

| Category | Global Influence | Local Influence |
|---|---|---|
| Politician | 0.3693 | 0.3366 |
| Celebrity | 0.7394 | 0.2068 |
| Businessman | 0.4519 | 0.1694 |
| General Public User | 0.0159 | 0.5546 |

### 5.3.2 Influence Analysis Using Four Clusters

After analyzing the user influence from each category's point of view, we found that some categories have clear regions while some do not. So, we used the statistical software SPSS to apply the k-means algorithm to classify the users into different clusters. In the initial study, we have tried from two clusters to twenty clusters, and judging from the results, 4 clusters gave the best result. So, this section, we only show the results of the four best clusters which are given in figure 5.8.

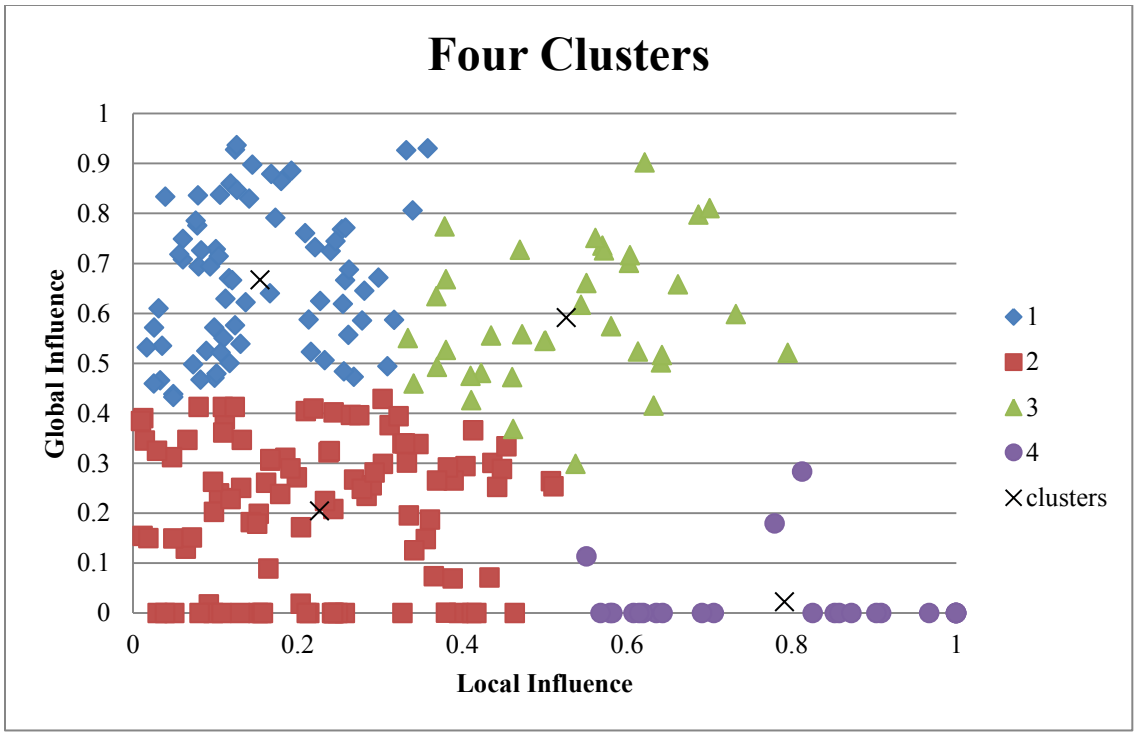


**Figure 5. 8 K-means Generation of Four Clusters**

From the figure above, we can see the disparity among the four clusters. Cluster 1 located in the top-left, has high Global Influence but low Local Influence. Cluster 2 located in the bottom-left of the chart, has relative low Global Influence and relative low Local Influence. Cluster 3 located in the top-right of the chart, has high Global Influence and relative high Local Influence. Cluster 4 located in the bottom-right of the chart, has high Local Influence and low Global Influence. The center of each cluster is marked with symbol X, and the exact value can be found in Table 5.4.

**Table 5. 4 Center of Four Clusters**

| Cluster | Global Influence | Local Influence |
|---|---|---|
| 1 | 0.6668 | 0.1544 |
| 2 | 0.2043 | 0.2267 |
| 3 | 0.5908 | 0.5265 |
| 4 | 0.0221 | 0.7916 |

#### 5.3.2.1 Cluster 1

Among 234 points in the set, there are 71 points in Cluster 1. As this cluster has relative high Global Influence and relative low Local Influence, we define it as the celebrity cluster. The cluster is composed of three kinds of users: celebrities, globally influential politicians, and large sized company CEOs. Out of 33 celebrities in the training set, 28 are located in this cluster. In addition, 14 of the most internationally influential politicians (including 4 federal level leaders, 2 party leaders, 4 international large city mayors and 4 provincial leaders of large region) and 29 large sized companies CEOs (such as Apple CEO Tim Cook) are included in this cluster.

#### 5.3.2.2 Cluster 2

Out of 234 points in total, 101 are in Cluster 2. In total, 107 politicians were included in the training set. Out of that number, 61 politicians are grouped in Cluster 2, including 35 provincial level politicians and 26 municipal level politicians. 20 businessmen appear in this cluster as well; they are mainly CEOs of small/ medium companies (such as Jamie Cheng, the founder of Klei Entertainment) or branch companies. Another source of this cluster is ordinary users. We define it as the regular politician cluster.

#### 5.3.2.3 Cluster 3

There are 36 points included in Cluster 3. 26 of them are politicians. These politicians are neither world-famous (Cluster 1) nor average (Cluster 2) but temporarily famous, such as those actively campaigning (US presidential nominees Donald Trump and Hilary Clinton are in this cluster, for instance). In addition, some businessmen who are recognized as philanthropists or involved with the public are in this cluster as well. As the users in this cluster are quite active with the public, we define this cluster as 'social activist'.

#### 5.3.2.4 Cluster 4

There are 26 points in Cluster 4 and it is mainly composed of ordinary users. The users in this cluster have the lowest Global Influence score and relative high Local Influence score, as a result of small numbers of followers (low value in equation 4-4 and high value in equation 4-1). Cluster 4 is defined as ordinary user cluster.

### 5.3.3 Category vs Cluster

To illustrate the relationship between the category and clusters, each category will be projected into the four clusters separately in this section.

#### 5.3.3.1 Politician vs Clusters

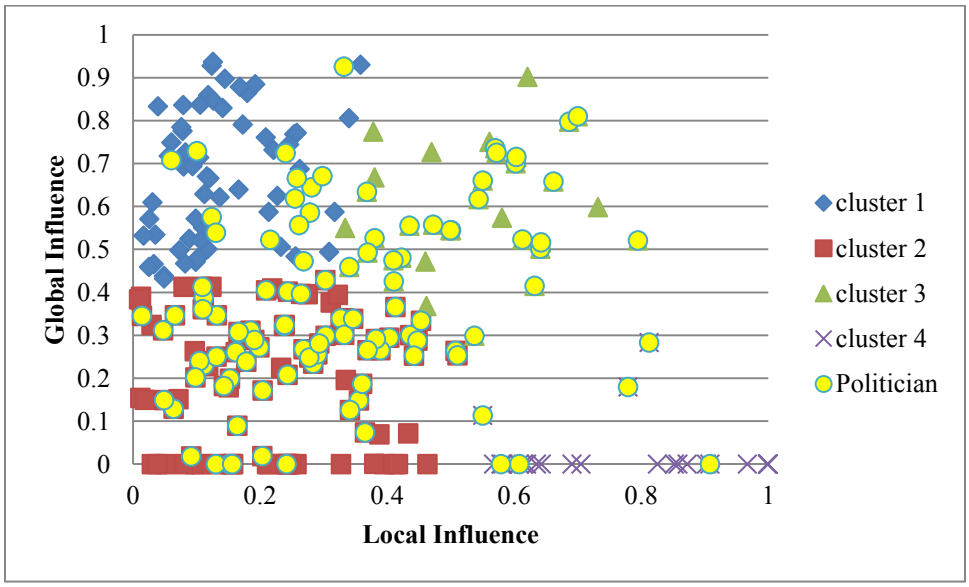


**Figure 5. 9 Politicians with 4 Clusters**

From the above figure, we can find that the politicians are mainly located in Cluster 1, 2, and 3. Among the 107 politicians, 14 world influential politicians are in the 'celebrity' cluster (Cluster 1), 61 of them are in the 'average politician' cluster (Cluster 2), 26 of them are in the 'social activist' cluster (Cluster 3), and 4 of them are in the 'ordinary user' cluster (Cluster 4).

#### 5.3.3.2 Celebrity vs Clusters

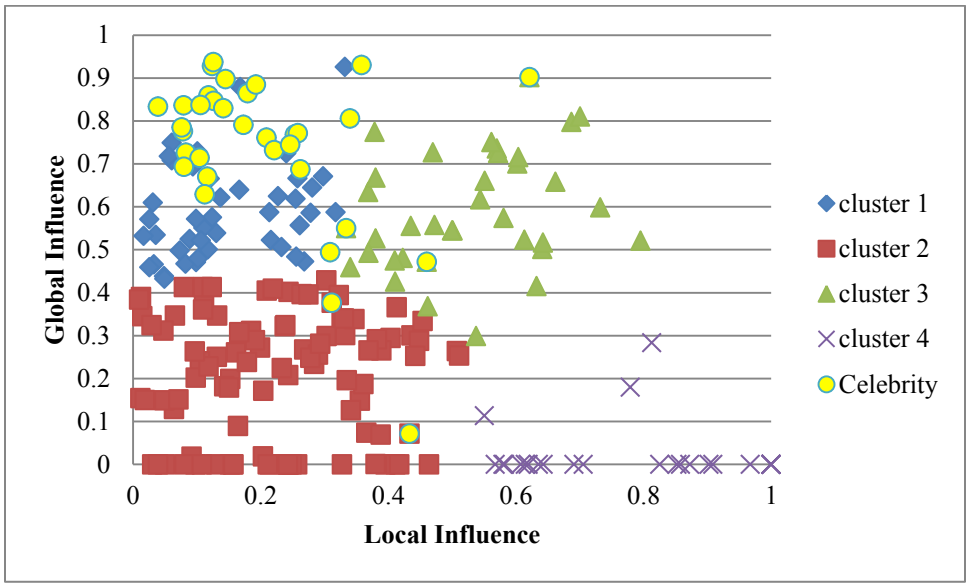


**Figure 5. 10 Celebrities with 4 Clusters**

Figure 5.10 illustrate the distribution of celebrities in the four clusters. Out of total 33 celebrities that are included in the training set, 28 points appear for Cluster 1. Three points are located near the junction of Cluster 1, 2, and 3. One outlier is Kim Kardashian, who is active in composing tweets and also has many followers retweeting her tweets (high value in both Local Influence and Global Influence). As a result, she is placed under Cluster 3. Another outlier is comedian Donovan Goliath, who has high Local Influence but low Global Influence due to a low number of tweets (low value in equation 4-5) and high number of recent activity (high value in equation 4-2). Thus, he is placed at the junction of Cluster 2 and Cluster 4.

### 5.3.3.3 Businessman vs Clusters

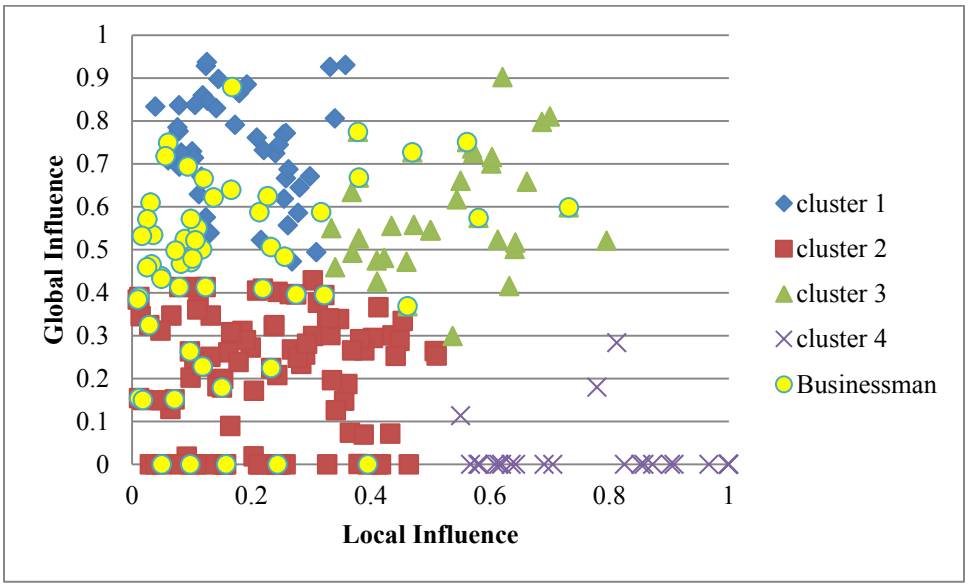


**Figure 5. 11 Businessman with 4 Clusters**

Figure 5.11 illustrate that 29 of 56 businessmen are found in Cluster 1, who are mainly CEOs of large companies. 20 CEOs of small and medium companies are in Cluster 2. 7 CEOs who are recognized as philanthropist and social activist are placed in Cluster 3.

### 5.3.3.4 General public users vs Clusters

Figure 5.12 demonstrate how general public users are distributed in the four clusters. Most of them are found at the bottom of the axis of Global Influence, but their Local Influence scores are not in a certain range. Thus, 18 out of 38 general public users are found at the bottom of Cluster 2 and the remaining 20 are found at the bottom of Cluster 4.

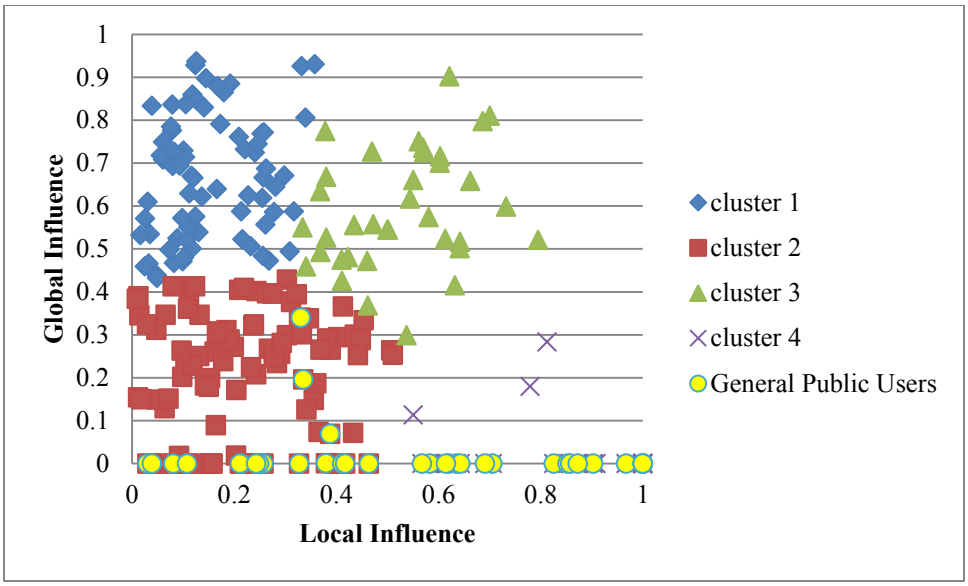


**Figure 5. 12 General Public Users with 4 Clusters**

#### 5.3.3.5 Conclusion

Figure 5.13 illustrates some striking similarities and differences. In the figure, the centers of four clusters are marked with X. The celebrity cluster center in the top-left side of figure lines up well with the Cluster 1 center. Likewise, the general public user cluster center in the bottom-right side of figure is found near the location of the Cluster 4 center. These two categories were those whose users were mostly found in single clusters. However, the business and politician cluster centers are not located near the centers of Clusters 2 and 3. These two categories of users are likewise the ones divided over multiple clusters as shown in Figure 5.8.

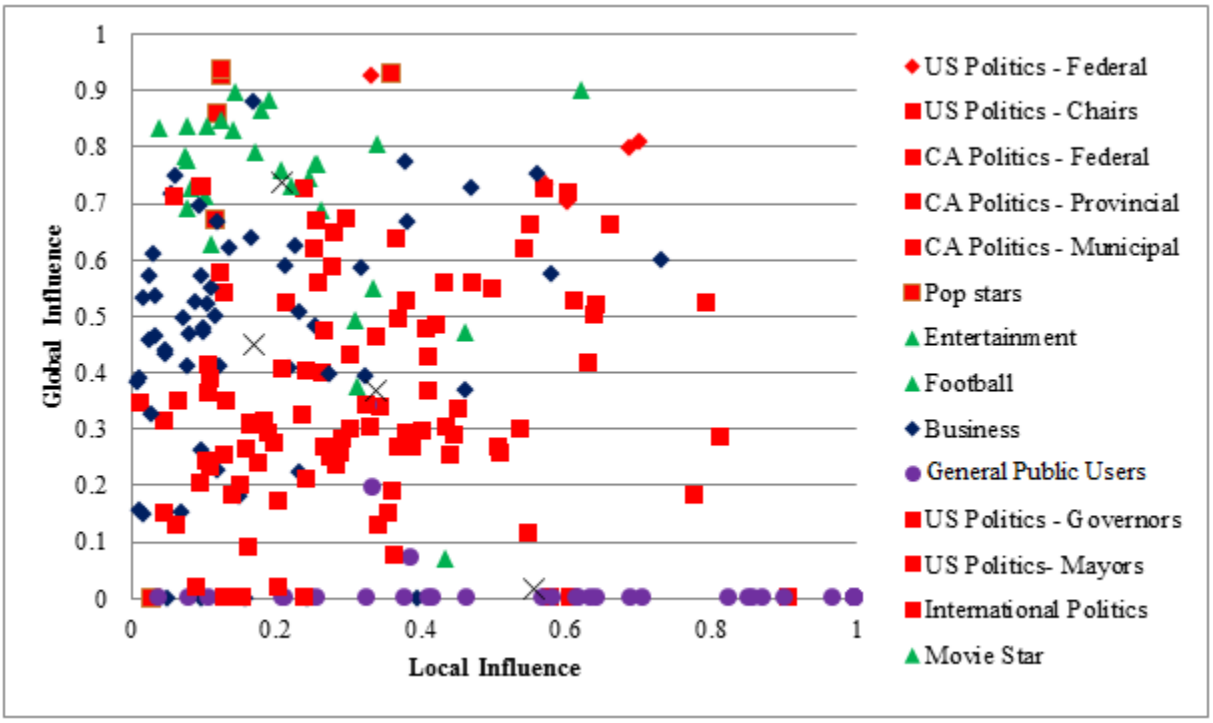


**Figure 5. 13 Category VS Cluster**

## 5.4 Evaluation of the Method

Modelling users based on their Global Influence and Local Influence was presented in the previous section. In this section, we show how to predict a user's class based on their statistics. To achieve this goal, we conducted two tests, "categorization"test, where the test users were matched with the category clusters, and "clustering" test, where the test users were matched into the four k-means clusters.

To evaluate our method, 50 users are identified for the test set which include 13 politicians, 11 celebrities, 14 businessmen and 12 General public users, as mentioned in the data collection section. We first collected users' information, and then, computed their Local

Influence and Global Influence by using equation (4-3) and (4-6) respectively. Predictions were then made on the categories and clusters they belong to. The next step consisted of computing the Euclidean distance and classifying each user to the nearest cluster center. Finally, results were compared and predictions were made.

The confusion matrix of categorization and clustering is shown in Table 5.5 and 5.6, respectively.

**Table 5. 5 Confusion Matrix of Categorization**

| Cat-Confusion Matrix | | Prediction | | | |
|---|---|---|---|---|---|
| | | Politician | Celebrity | Businessman | General Public User |
| Real Result | Politician | 5 | 6 | 2 | 0 |
| | Celebrity | 0 | 11 | 0 | 0 |
| | Businessman | 3 | 3 | 4 | 4 |
| | General Public User | 0 | 0 | 0 | 12 |

**Table 5. 6 Confusion Matrix of Clustering**

| Cluster Confusion Matrix | | Prediction | | | |
|---|---|---|---|---|---|
| | | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
| Real Result | Cluster 1 | 14 | 3 | 3 | 0 |
| | Cluster 2 | 0 | 5 | 0 | 0 |
| | Cluster 3 | 0 | 2 | 8 | 0 |
| | Cluster 4 | 0 | 3 | 0 | 12 |

To evaluate the performance of the classification method, here we introduce precision, recall and f-measure. Precision is the fraction of the retrieved instances that are relevant, and

recall is the fraction of the relevant instances that are retrieved. F-measure considers precision and recall, and is used for measuring the accuracy of the test.

The precision and recall computation for each cluster *i* (equations (5-1) and (5-2) respectively) is undertaken. The average over all classes (equations (5-3) and (5-4) respectively) were found, followed by the computation of the F-measure of the classification (equation (5-5)). The results are presented in Table 5.5.

$$P_i = \frac{TruePositive_i}{TruePositive_i + FalsePositive_i} \quad (5\text{-}1)$$

$$R_i = \frac{TruePositive_i}{TruePositive_i + FalseNegative_i} \quad (5\text{-}2)$$

$$P = \frac{1}{4}\sum_{i=0}^{3} P_i \quad (5\text{-}3)$$

$$R = \frac{1}{4}\sum_{i=0}^{3} R_i \quad (5\text{-}4)$$

$$F = \frac{2PR}{P+R} \quad (5\text{-}5)$$

**Table 5. 7 Experiment Results**

| Experiment | Precision | Recall | F-measure |
|---|---|---|---|
| Categorization | 0.62 | 0.65 | 0.64 |
| Clustering | 0.80 | 0.84 | 0.82 |

Table 5.7 shows that Clustering has better performance than Categorization. The precision, recall and f-measure value of Clustering are all above 0.8, which indicates that Clustering is able to classify the user in the correct class.

## 5.5 Benchmarks

To evaluate the proposed method, two systems for benchmark comparison are selected, the IARank system and the Klout system. These two systems provide a score which measure the users' influence, as explained in Chapter 4. The benchmark experiment generated the IAScores. The Klout score was retrieved for all users who were tracked using the same 234 users information to train a classifier and a clustering algorithm and the same 50 for testing.

### 5.5.1 IARank

IARank system is presented in Chapter 2. This system has the advantage of providing two-dimensional influence metrics of buzz and structural advantage, like the metrics presented in this section. However, its equations use the number of mentions a tweet receives, which is no longer the public information that can be accessed through the Twitter API. It is replaced with the number of retweets a tweet receives, which is still publicly available and is highly correlated to the number of mentions (correlation coefficient of 0.972 [21]). The result is shown in the Table 5.6.

**Table 5. 8 IARank Experiment Results**

| Experiment | Precision | Recall | F-measure |
|---|---|---|---|
| IARank-Categorization | 0.32 | 0.43 | 0.37 |
| IARank-Clustering | 0.57 | 0.59 | 0.58 |

The results show that clustering has better performance than categorization on precision, recall and F-measure. However, our method has better performance than the IARank on every single option.

### 5.5.2 Klout Score

The second benchmark is the Klout system [16] that analyses 3600 features of 750 million users of nine social networks to assign them influence scores. Unlike our system, the metric used in the Klout system provides a single value and not a two-dimensional value. Moreover, we are limited to looking up values on the Klout website. Scores are only available for 150 out of 284 users in the training set and test set. They are notably missing for almost all members of our "ordinary users" class, which was subsequently excluded from the benchmark test of that system. The center of each category's Klout score is computed for the users in the training set, we find the klout score for the user in test set, do the prediction. Then match the user with the nearest category center. However, as most of the ordinary user's klout score can not be accessible, we run the SPSS to get 3 clusters on the training set and computed the center of each cluster. We do the prediction and match the user in the nearest clustering center. The experimental results are shown in Table 5.9.

**Table 5. 9 Klout Experiment Results**

| Experiment | Precision | Recall | F-measure |
|---|---|---|---|
| Klout-Categorization | 0.28 | 0.28 | 0.28 |
| Klout-Clustering | 0.54 | 0.52 | 0.53 |

The Klout results are consistent with the results using the new metric presented in this work. The clustering precision, recall, and F-measure scores are 20% to 25% higher than the categorization scores. However, it is clear that the new metric performs significantly better than the benchmarks in all measures in these experiments.

## 5.6 Conclusion

In this Chapter, the results of the experiments using the Local Influence and Global Influence are presented. To better analyze the relationship between users' influence score and their social state in real life, a training set of 234 users' data are collected. The center of each category is defined with the influence scores. Each category's characteristics are described in detail. Four clusters are identified using the k-means classification algorithm. And the organization of each cluster is defined. The discussion on the relationship between category and cluster enables understanding the connection between users' influence score and their social state in real life.

It is clear from the above experiments and by comparison our results with two benchmark systems that the new method has better performance than the other two approaches. With an f-measure score at 0.82, the new method is able to put the user into the correct class.

# 6. Conclusion

In this thesis, we discussed the users' influence analysis based on users' social network information in Twitter. Unlike most existing methods, the influence analysis, proposed in this work, is carried out in real-time thus capturing the constant information changes on Twitter. This is one of the main features of the proposed research work.

## 6.1 Achievements of this work

In the early study of the influence analysis, two new ranking schemes for finding influential users by given keyword in real-time were developed based on the metrics mentioned in the background review.

- Iterative Follower Rank: Evolved from the follower rank, works better because it takes into account the user's followers in the social network.
- Compromised IARank: Evolved from the IARank, using a new Buzz score due to the limitation of the Twitter API.

After trying out these new schemes, the relationship between the influence score and user's social states in was discovered. More information and user's social network data were collected, and a new approach to modelling a user's influence on social networks was introduced in this thesis. The proposed model consists of two metrics: Local Influence which the user has on their immediate set of contacts and the Global Influence which the user has on the entire social network. General definitions of these two metrics and the implementation of the metrics in a real

social network (Twitter) are presented. The case study using Twitter showed that the new model can create clusters of users in 2D space corresponding to their social standing, and can further be used to classify previously-unseen users into the correct classes with an f-measure of 0.82, which is significantly higher than used benchmark algorithms.

The previous work of classification of users in Twitter was mainly based on the semantic analysis of the users' profile [31,32,33]. The semantic analysis requires a training process and the subjects in Twitter change really fast. One flaw of the method using semantic analysis is time consuming. Moreover, different languages require different training materials to complete the semantic analysis, so the scalability of using it in multi-language environment is not good. However, the method proposed in this thesis can be used solely based on the user's profile statistics instead of doing semantic analysis. Consequently, our method is fast, works without offline computation and is robust. This is because the user's profile statistics do not change frequently compared to the frequency of topics (or subjects) changes in Twitter. Moreover, the accuracy of putting users in the correct class is acceptable.

## 6.2 Potential Improvements

There are several aspects of this research work that can be improved. Considering the training set, more users from celebrity, businessman and general public users can be introduced. Also, further analysis on the politicians could be conducted to figure out if any sub-clusters exist. For example, the population and the Internet usage of the city of interest could be taken into account while analysing municipal politicians' Local Influence.

Another potential application area of this study is to focus on the celebrity category. More celebrities can be introduced to figure out if there are any sub-clusters of interest, such as singers, movie stars, fashion stars and sport stars.

The users in the training set are all personal accounts. We could also introduce some public accounts such as city accounts, news outlet accounts and company accounts. Comparing the influence scores of the public account and personal account from the same category and capturing these features can also be meaningful.

## 6.3 Future Work

The influence analysis can be used in various areas to help us understand the crucial element of influencing people in social networks. For example, the analysis on politicians can be used for political campaigns; the analysis on celebrity can be used for promoting new songs or movies; and the analysis on businessman can be used for promoting a company or product.

Another research direction can focus on the fashion industry. Many social network service users are concerned about this subject and there are many popular accounts aimed at teaching people to do makeup or stylish dressing. Finding influence features of the popular fashion accounts (including fashion news outlets, fashion icons, brand and popular bloggers) and characterizing their influence measures can be a good direction to apply the analysis.

The research in this thesis is on Twitter. However, the profile statistics of social network services are similar. The method could be also applied on Facebook, Instagram and LinkedIn by making slight changes: total followee number, total posts number, and created time of post in Instagram can be kept and the retweet number can be replaced by the total comments number.

# Appendices

# Appendix A

```
# Get user list
def get_ulist(T0):
        tlist=[]
        ulist_temp=[]
        ulist=[]
        for tweet in T0:
                #create empty dictionary
                twinfo={}
                users={}
                twinfo['tweet_id']=tweet.id
                twinfo['in_reply_to_status_id']=tweet.in_reply_to_status_id
                twinfo['in_reply_to_user_id']=tweet.in_reply_to_user_id
                twinfo['user_id']=tweet.user.id
                twinfo['retweet_count']=tweet.retweet_count
                tlist.append(twinfo)
                if twinfo['retweet_count']>100:
                        users['user_id']=tweet.user.id
                        users['screen_name']=tweet.user.screen_name

        users['FS']=float('%0.2f'%(tweet.user.followers_count))/float('%0.2f'%(tweet.user.statuse
s_count))
                        users['RAT']=1
                        if (tweet.user.followers_count + tweet.user.friends_count) > 0:
                                ## modified if users['FR'] > 0:
                                a = float('%0.2f'%(tweet.user.followers_count) )/
float('%0.2f'%(tweet.user.followers_count + tweet.user.friends_count))
                                users['score']= float('%0.3f'%a)
```

```
                                        users['R_score'] = float('%0.3f'%(users['score']-(8-
len(str(tweet.user.followers_count)))*0.100))
                                        if (users['R_score'] < 1.0 and users['R_score'] > 0.2):
                                                ulist_temp.append(users)
                ind={}
                for user in ulist_temp:
                        if user in ulist:
                                num=ulist.index(user)
                                if ind.has_key(num):
                                        ind[num] += 1
                                else:
                                        ind[num]=2
                        else:
                                ulist.append(user)

                for (k,v) in ind.items():
                        ulist[k]['RAT']=v
                return ulist
```

# Appendix B

```
# calculate the user's L1 follower 's  FR_score by given user id
def get_FR_score(api1,api2,id):
        #get L1 followers
      followers=get_friends_followers_ids(api2,
                                user_id=id,
                                #friends_limit=10,
                                followers_limit=10)
        # L1 follower list
        l1info=[]
        for follower in followers:
                l1={}
                l1userinfo=api1.get_user(id=follower)
                l1['follower']=l1userinfo.followers_count
                l1['follwee']=l1userinfo.friends_count
                l1['FR']=l1userinfo.followers_count+l1userinfo.friends_count

        # filter user whose FR =0
                if l1['FR'] > 0:
                        l1['score']= float('%0.3f'%(l1['follower']) )/ float('%0.3f'%(l1['FR']))
                         l1info.append(l1)
        # normalize the FR score
        # get an average FR score
        score=0
        for element in l1info:
                score =+ element['score']
        avg_score=float('%0.3f'%(score/len(l1info)))

        return avg_score
```

# Appendix C

```
#calcualte the buzz for each user


def get_buzz_list(api_twitter,ulist,max_num):
    ulist_buzz = []
    for user in ulist:
        tweets = harvest_user_timeline(api_twitter,user_id= user['user_id'],
max_results=100)
        sum_retweet = 0
        for tweet in tweets:
            retweet = tweet[u'retweet_count']
            sum_retweet += retweet
        user['buzz'] =float( '%0.2f'%sum_retweet)/float('%0.2f'%len(tweets))
        ulist_buzz.append(user)
    return ulist_buzz
```

# Bibliography


[1] Y. Singer, “How to win friends and influence people, truthfully: influence maximization mechanisms for social networks”, *Proceedings of the fifth ACM international conference on Web search and data mining*, ACM, 2012, pp. 733-742.

[2] E. Otte, and R. Rousseau, “Social network analysis: a powerful strategy, also for the information sciences”, *Journal of information Science*, 28(6), December 2002, pp. 441-453.

[3] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a social network or a news media?”, *Proceedings of the 19th international conference on World wide web*, ACM, 2010, pp. 591-600.

[4] A. Sameh, “A Twitter analytic tool to measure opinion, influence and trust”, *Journal of Industrial and Intelligent Information*, 1(1), 2013.

[5] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network”, *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2003, pp. 137-146.

[6] R. Sandoval-Almazan, "Using twitter in political campaigns: The case of the pri candidate in mexico", *International Journal of E-Politics (IJEP)*, 6(1), 2015, pp. 1-15.

[7] L. Ruan, and L. Yang, "Behind Canadians’ Public Sentiments on China: A Preliminary Social Network and Sentiment Analysis of the Canadian Twittersphere".

[8] E. Dubois, and D. Gaffney, "The multiple facets of influence identifying political influentials and opinion leaders on twitter", *American Behavioral Scientist*, 58(10), 2014, pp. 1260-1277.

[9] P. E. Agre, “Real-time politics: The Internet and the political process”, *The information society*, 18(5), 2002, pp. 311-331.

[10] E. Dubois, and W.H. Dutton, “The fifth estate in Internet governance: Collective accountability of a Canadian policy initiative”, *Revue française d’études américaines*, (4), 2012, pp. 81-97.

[11] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, “Predicting the political alignment of twitter users”, *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, IEEE, 2011, pp. 192-199.

[12] R. Zafarani, M. A. Abbasi, and H. Liu, *Social Media Mining: An Introduction*, Cambridge University Press, 2014.

[13] D.M. Romero, W. Galuba,S. Asur, and B. A. Huberman, “Influence and passivity in social media”, *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer Berlin Heidelberg, 2011, pp. 18-33.

[14] Z.Zhu, “Discovering the influential users oriented to viral marketing based on online social networks”, *Physica A: Statistical Mechanics and its Applications*, 392(16), 2013, pp. 3459-3469.

[15] K. H. Chu, J. B. Unger, J. P. Allem, M. Pattarroyo, D. Soto, T. B. Cruz, and C. C. Yang, "Diffusion of Messages from an Electronic Cigarette Brand to Potential Users through Twitter", *PloS* one, 10(12), December 18, 2015.

[16] R. Nagmoti, A. Teredesai, and M. De Cock, “Ranking approaches for microblog search”, *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, IEEE, Vol. 1, 2010, pp. 153-157.

[17] A. Leavitt, E. Burchard, D. Fisher, and S. Gilbert, "The Influentials: New Approaches forAnalyzing Influence on Twitter", *Web Ecology Project*, 4(2), September 2, 2009, pp. 1-18.

[18] A. Pal, and S. Counts. "Identifying topical authorities in microblogs", *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM, 2011, pp. 45-54.

[19] J. Weng, E. P. Lim, J. Jiang,and Q. He, “Twitterrank: finding topic-sensitive influential twitterers”, *Proceedings of the third ACM international conference on Web search and data mining*, ACM, 2010, pp. 261-270.

[20] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment", *Journal of the ACM (JACM)*, 46(5), 1999, pp. 604-632.

[21] T. Noro, F. Ru, F. Xiao and T. Tokuda, "Twitter user rank using keyword search", *Information Modelling and Knowledge Bases XXIV. Frontiers in Artificial Intelligence and Applications*, 251, 2013, pp. 31-48.

[22] L. Page, S. Brin, R. Motwani and T. Winograd, "The PageRank citation ranking: bringing order to the web", 1999.

[23] R. Cappelletti, and N. Sastry, "Iarank: Ranking users on twitter in near real-time, based on their information amplification potential", *Social Informatics (SocialInformatics)*, *2012 International Conference on*, IEEE, 2012, pp. 70-77.

[24] A. Rao, N. Spasojevic, Z. Li, and T. Dsouza, "Klout score: Measuring influence across multiple social networks", *Big Data (Big Data), 2015 IEEE International Conference on*, IEEE, 2015, pp. 2282-2289.

[25] J. del Campo-Ávila, N. Moreno-Vergara, and M. Trella-Lopez, "Bridging the gap between the least and the most influential Twitter users", *Procedia Computer Science*, 19, 2013, pp. 437-444.

[26] A. Tejeda-Gómez, M. Sànchez-Marrè, and J. M. Pujol, "Discovering social structures of local influence by using tweetStimuli", *International Journal of Computer Mathematics*, 91(2), 2014, pp. 291-303.

[27] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. "Everyone's an influencer: quantifying influence on twitter", *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM, 2011, pp. 65-74.

[28] H. Schoen, D. Gayo-Avello, P. Takis Metaxas, E. Mustafaraj, M. Strohmaier, and P. Gloor. "The power of prediction with social media", *Internet Research*, 23(5), 2013, pp. 528-543.

[29] M. Pennacchiotti, and A. M. Popescu, "A Machine Learning Approach to Twitter User Classification", *ICWSM*, 11(1), 2011, pp. 281-288.

[30] Y. Mei, Y. Zhong, J. Yang, "Finding and Analyzing Principal Features for Measuring User Influence on Twitter", *IEEE First International Conference on Big Data Computing Service and Applications*, 2015, pp. 478-486.

[31] L. De Silva, and E. Riloff , "User type classification of tweets with implications for event recognition", *ACL 2014*, 2014, p.98.

[32] S. Bergsma, and B. Van Durme, "Using Conceptual Class Attributes to Characterize Social Media Users", *ACL* (1), 2013, pp. 710-720.

[33] D. Preoţiuc-Pietro, V. Lampos, and N. Aletras, "An analysis of the user occupational class through Twitter content", The Association for Computational Linguistics, 2015.