Determine the Close Friends for Twitter Users by Analyzing Users

Public Data Name

Haydar Hassan Safi
Computer Department
Ozyegin University
Istanbul – Turkye
hayderaljebory76@gmail.com

Khalid Moh. Alzaidi
Computer Department
Ozyegin University
Istanbul – Turkye
kha2005ms@yahoo.com

Zena Fawzi Kadem
Applied Sciences Department
University of Technology
Iraq _ Baghdad
zenakadem@gmail.com

Abstract

In this paper a new method to extract the closed friends to a certain twitter user is proposed. This algorithm can be used in many applications such as security and marketing. In the proposed algorithm four relation indicators are used. The shared twitter users, the common friends, the common followers and number of "retweeted from" are used to determine the closed friends of a certain twitter user. The great data downloading and mining tool KNIME is used with some R coding. The proposed model could ranks the shared twitter users according to their closeness to a certain twitter user using the four used relation indicator.

1. Introduction

Social media networks become the most important and biggest data source in the world. Users write everything about their life, experience, relations and even their feelings. A social network is a website on the Internet that brings people together in a central location to talk, share ideas and interests, or make new friends. Twitter is the second most important social network website in the world after Facebook. It is an online social networking service that enables users to send and read short 140-character messages called "tweets". After you become a registered user in Twitter, you could read and post tweets with your followers or friends. Unregistered users can only read posts from other twitter users and the other features are restricted. Users access Twitter through the website interface, SMS, or mobile device app. Twitter Inc. is based in San Francisco and has more than 25 offices around the world.[7].

Not like Facebook, twitter users could not determine their friends because there is no tool to describe this relation in the Twitter website. Instead of this there is the followers list which determines the twitter users that you want to follow. But of course this does not mean that the followers are the friends of the Twitter user. The main goal of this paper is to find the closest people (we means from other twitter users) to a certain twitter user. Determine the closest circle of a certain user in any social media is very important and can be used in many applications. For example, security authorities always need to determine the close circle of a certain user to determine the personality, trends and plans of that user. Furthermore this information is important for advertising companies including twitter itself to direct the advertisements more precisely and to now the preferences of their users.

This paper is organized as follow. In section two a brief review about research in social networks will be viewed and we will concentrate on Twitter research papers. Section three will explain the tools used in this paper to download the twitter data and analyze the downloaded data. The proposed algorithm to determine the friends or closed twitter users will be proposed in section four and we will also analyze and discuss the features of the proposed algorithm in this section. In last section a conclusion and future work will be discussed.

2. Previous Work

After the great using of social network sites, it becomes a rich research domain. There are many research papers in the literal propose algorithms to analyze the social networks data and tries to extract new information from it or analyze the performance of the topology of the users networks. In this section we will give some brief overview of this research and we will concentrate on Twitter website.

One of the frequently using of twitter data in last years is the election prediction Murphy Choy, Michelle Cheong, Ma Nang Laik and Koo Ping Shung [1] propose a model to predict USA Presidential Election 2012. In that time it was a very tight race between the two key candidates. They used the candidates' campaigns which continued several months and so they expect that the effects of these campaigns can be predicted from the internet and twitter. In their paper, the model described in Choy et. al. [2] was used. Before this paper there are many papers tried to do some election prediction in different ways using twitter data and another social network websites [3-5].

Because of the importance of stock markets there are a lot of research papers try to predict information about markets from social network public data. In 2010 [6] two researchers tried to answer this question, is the public mood of social networks correlated of economic indicators? They investigated whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. They analyzed the text content of daily Twitter feeds by two mood tracking tools and their results indicated that the accuracy of DJIA predictions can be improved by predicting the public mood from twitter.

There are many of other research new topics in social networks relating the friendship. In 2010 Teng-Sheng Moh and Alexander J. Murmann [7] publish a paper titled as “Can You Judge a Man by His Friends? Enhancing Spammer Detection on the Twitter Microblogging Platform Using Friends and Followers”. In their paper they proposed a new method to distinguish between spam users and real users using the information of their friends. In 2011 [8] a paper name “Where are my followers? Understanding the Locality Effect in Twitter” was proposed and the researchers of this paper demonstrated that language and cultural characteristics of twitter users can be used to determine the level of Locality expected for different countries.

3. Tools

In order to analyze the twitter data and determine the close people or close twitter accounts to a certain twitter user, we must first download the data of some twitter accounts using twitter API. Selecting the data is very important because it affects the efficiency of results. To download and analyze the data we use three popular tools which used frequently in data mining and social network analyzing. Next paragraphs explain these tools in details:

3.1. Twitter API

Twitter allows the users to interact with its data, tweets and several important attributes about tweets and hash tags using Twitter APIs. To use these API need to write some code in any language like Python, Perl, R and any other language to make requests to twitter API and receive the results which will be in JSON format that can be easily read by your program [9]. Twitter API includes many functions that return twitter data in different styles and quantities. It is preferable to use another package to simplify the data receiving and communication with Twitter API. There are many new tools were built to simplify the communication with Twitter API. In this work we will use two of the simplest and widely used tools which are KNIME and R language.

3.2. KNIME

In order to make our communication to the twitter API easy and simplify the processing of data in powerful way we use a very new and widely used tool which is KNIME [10]. KNIME, the Konstanz Information Miner, is an open source data analytics, reporting and integration platform. KNIME is a brilliant data mining platform with a wide variety of pre-built data preparation and analytics operators (nodes) that save huge amounts of time in the analytics development cycle. KNIME integrates various components for machine learning and data mining through its modular data pipelining concept. It includes a graphical user interface allows assembly of nodes for data downloading, preprocessing, processing and analyzing. There are many researchers use KNIME environment to download and analyze the data [11]. In this paper we will use KNIME to download the data and do some analyzing using different nodes as will explained later.

3.3. R language and TwitterR package

R is a language and environment for statistical computing and graphics. It is a GNU project which is implemented to be similar and based on the S language which was developed at Bell Laboratories. R language can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R [12]. R is available as Free Software to download and use. There are many R packages on the internet to simplify connecting to other applications. One of those packages is TwitterR. It is implemented to Provide an interface to the Twitter web API. This package is very new published in 2015 and powerful [13].

4. Proposed algorithm

Determining the closed friend or the most closed twitter users to a certain user is not an easy task. There are many public data available that can be exploits to determine the list of closed friends. In this paper we will use four relation indications to measure how a certain twitter user is closed to another twitter user. The following paragraphs summarize and explain these indicators:

4.1 Shared twitter users indicator

Let U1 be a twitter user with N1 followers and M1 friends, let U2 be another twitter user with N2 followers and M2 friends. If U1 follows U2 and U2 follows U1 we said that there is a shared relation between U1 and U2. This indicator is very important in measuring the relation between any two twitter users. It is reasonable to use this relation because if there is a twitter user that has a good relation to another one, they must follow each other. To determine the list of twitter users that have this shared relation we must find the twitter users that founded in the followers list and friends lists together for a certain twitter user. Figure one shows how compute this relation using KNIME.

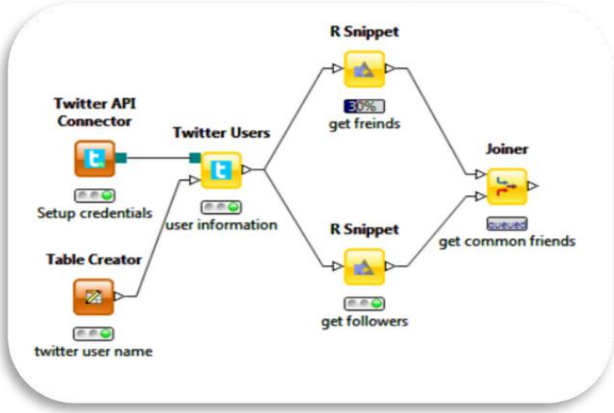

Fig. 1: **Compute the shared twitter users list**

First using (Twitter API connector) node our security information is entered to make the KNIME capable to connect correctly to the twitter API. After that we use the node (Twitter user node) to download all other information of twitter user such that name, number of retweets, number of followers and many other information. The two R Snippet nodes are used to write R code to download the friends and the followers of each twitter user. After that we use a Joiner node to determine the twitter users that listed in the two tables. As we see the KNIME simplify our work very much and make it possible to understand every step in our work. This step reduces the number of candidates that may be in the list of the closed friends very well. The experimental results show that this step approximately reduces the percent of candidate twitter users to 20% of the summation of the followers and friends lists.

4.2 Number of shared followers indicator

This indicator determines the number of twitter users that are following two different twitter users together. So if we have U1and U2 two different twitter users, if we said this indicator equals five, this means that there are five different twitter users following U1 and U2 together. This indicator is also important and can measures the relation between two twitter users because if this indicator has a big value between two twitter users then this means that there is a high number of users followed these two users together which mean that it must be a relation between these two users. Figure 2 shows how we compute this indicator for each friend of a certain twitter user.

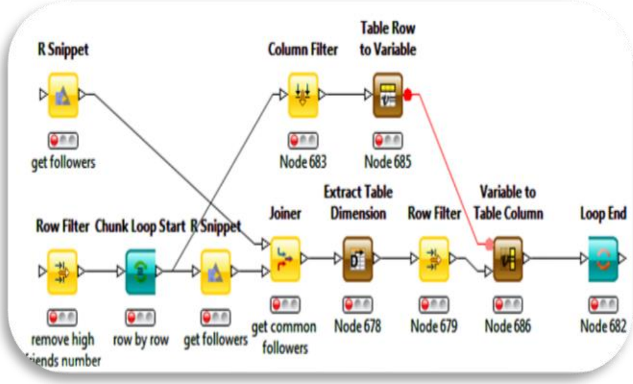

Fig. 2: Compute the number of shared followers for each friend of a twitter user

As we see in Figure 2 we get the friends of a certain twitter user one by one using (Chunk Loop) node, and for each one we get all the followers again using the R Snippet nodes by writing R code to download the followers. After that we use a Joiner node to determine the twitter users that listed in the two followers tables. To compute the number of common followers (Extract Table Dimension) node is used and after that we collect the results of all the friends of a twitter user. Note that the result of this step is a table contains two columns one for friend user name and the second column is a number to indicate the number of shared followers between this friend and the origin twitter user.

4.3 Number of shared friends indicator

This indicator is the same as previous indicator but it determines the number of twitter users that are friends for two different twitter users together. So if we have U1and U2 are two different twitter users, if we said this indicator equals five, this means that there are five different twitter users are friends to U1 and U2 together. This indicator is also important and can measures the relation between two twitter users because if this indicator has a big value between two twitter users then this means that there are a high number of common friends between these two users which means that it must be a relation between these two users. Note that this indicator is the same as the shared friend's property of Facebook. Figure 3 shows how we compute this indicator for each friend of a certain twitter user.

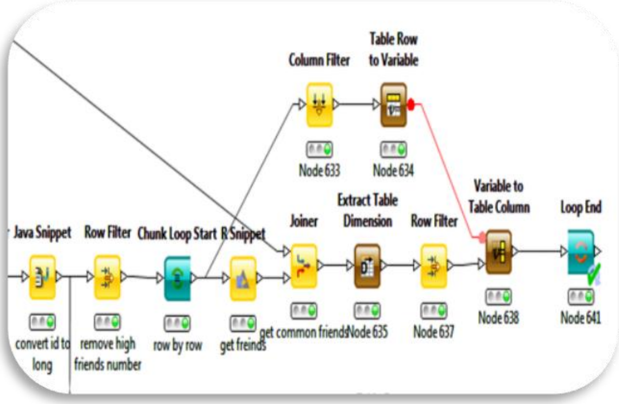

Fig. 3: Compute the number of shared friends for each friend of a twitter user

As we shown in Figure 3 we get the friends of a certain twitter user one by one using (Chunk Loop) node, and for each one we get all the friends again using the R Snippet nodes. After that we use a Joiner node to determine the twitter users that listed in the two friends tables. To compute the number of common friends (Extract Table Dimension) node is used and after that we collect the results of all the friends of a twitter user. Note that the result of this step is a table contains two columns one for friend user name and the second column is a number to indicate the number of shared friends between this friend and the origin twitter user.

4.4 “Retweeted from” indicator

This indicator is used to determine the number of retweeted from other twitter users. Of course if this value is high between two twitter users this means that there is a good relation predicted between them. To compute this value, the (Twitter Timeline) node is used after manage a secure connection to a twitter user using (Twitter API Connector) node. To compute the number of retweeted from each twitter user a powerful node (Statistics) is used as shown in Figure 4.

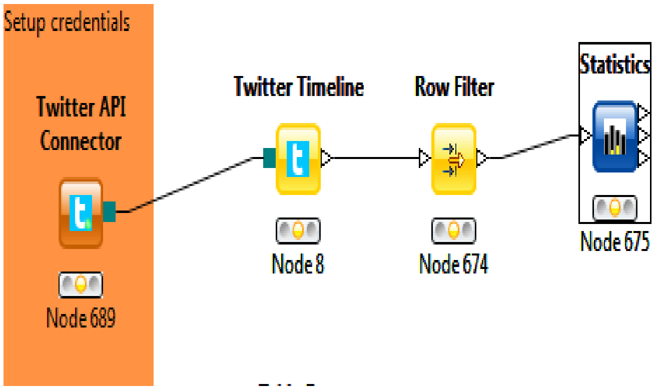

Fig. 4: Compute “Retweeted from” indicator for a certain twitter user

4.5 Combine the indicators

After we compute the four indicators explained in previous subsections, we must combine them in one algorithm to extract the closed friends of a certain twitter user. The reader maybe noted that we did not use training data in this algorithm because it is difficult to determine which twitter user is close to other one in order to use these data in training. The other important thing is that the degree of closeness change from person to person depending on the requirements and objectives. Therefore, we will propose general algorithm with three parameters that can control the degree of closeness as suited to the application which we will use this algorithm in. The following algorithm explains our proposed ***algorithm in details:-***

Let the twitter user that we want to compute the close friends for him be U.

Let the friends of U be as $fr_1, fr_2, fr_3 \dots . fr_n$ Where n is the number of U friends.

Let the followers of U be as $fo_1, fo_2, fo_3 \dots . fo_m$ where m is the number of U followers.

Step 1: compute the shared twitter users for U as explained in subsection 4.1. Assume that the resulted list will be $fs_1, fs_2, fs_3 \dots . fs_p$ where p < m and p < n. note that this list can be computed mathematically as follow: let m < n

```
For all $fo_i$ where i = 1,2… m
For all $fr_j$ where j = 1,2… n
    If $fo_i = fr_j$ add $fo_i$ to shared twitter users list
```

Step 2: for each shared twitter users $fs_1, fs_2, fs_3 \dots . fs_p$ compute the Number of shared follower’s indicator as described in subsection 4.2. Let these values be $Sfo_1, Sfo_2, \dots . Sfo_p$

Step 3: for each shared twitter users $fs_1, fs_2, fs_3 \dots . fs_p$ compute the Number of shared friend’s indicator as described in subsection 4.3. Let these values be $Sfr_1, Sfr_2, \dots . Sfr_p$

Step 4: compute the number of “retweeted from” for each twitter users $fs_1, fs_2, fs_3 \dots . fs_p$. Let these values be $rf_1, rf_2, \dots . rf_p$

Step 5: compute the shared friends' $fs_1, fs_2, fs_3 \ldots . fs_p$ closeness value as described in the ***following equation*** :

$$S_i = \alpha * Sfo_i + \beta * Sfr_i + \gamma * rf_i \qquad where\ \alpha + \beta + \gamma = 1$$

$$and\ \alpha, \beta, \gamma\ are\ weights\ to\ control\ the\ strength\ of\ the\ Sfo_i, Sfr_i and\ rf_i$$

friends' of U and the friend with highest value will be the most closed friend to U. We can choose suitable values for $\alpha, \beta\ and\ \gamma$ to get the wanted results in any application. These values can be determined easily using training data if it is available with any machine learning algorithm.

5. Conclusion and future work

In this paper a new method to extract the closed friends to a certain twitter user was proposed. This algorithm can be used in many applications such as security and marketing. In the proposed algorithm four relation indicators are used which are the shared twitter users, the common friends, the common followers and number of "retweeted from". The great data downloading and mining tool KNIME was used with some R coding. The proposed model computes a value to indicate the relation between twitter user and each one of his shared friends. Those values could rank the shared twitter users according to their closeness to the twitter. In future we will work to include other indicators to increase the accuracy of the algorithm such as the number of replies and making favorites to tweets. We also try to collect some training data to test our algorithm.

References

[1] Murphy Choy, Michelle Cheong, Ma Nang Laik, Koo Ping Shung, "US Presidential Election 2012 Prediction using Census Corrected Twitter Model" (Submitted on 5 Nov 2012, last revised 11 Nov 2012.

[2] Murphy Choy, Michelle Cheong, Ma Nang Laik, Koo Ping Shung, 2011. A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction, Arxiv.

[3] Andranik Tumasjan, T O Sprenger, P G Sandner, I M Welpe, 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment, International AAAI Conference on Weblogs and Social Media Washington DC (2010)

[4] Armstrong, J. S. & A. Graefe, 2011. Predicting elections from biographical information about candidates, Journal of Business Research, 64, 699-706.

[5] Böhringer, M., and Richter, A. 2009. Adopting Social Software to the Intranet: A Case Study on Enterprise Microblogging. In Proceedings of the 9th Mensch & Computer Conference, 293-302. Berlin

[6] Huina Maoa, Xiaojun Zeng "Twitter mood predicts the stock market Johan Bollena" Journal of Computational Science 2 (2011) 1–8

[7] Teng-Sheng Moh, Alexander J. Murmann "Can You Judge a Man by His Friends? - Enhancing Spammer Detection on the Twitter Microblogging Platform Using Friends and Followers" Communications in Computer and Information Science Volume 54, 2010, pp 210-220

[8] Roberto Gonzalez, Ruben Cuevas, Carmen Guerrero and Angel Cuevas "Where are my followers? Understanding the Locality Effect in Twitter" Institute Telecom, Telecom SudParis

[9] https://dev.twitter.com/overview/api

[10] http://www.knime.org/

[11] Tiwaria, Abhishek; Sekhar, Arvind K.T. (October 2007). "Workflow based framework for life science informatics". *Computational Biology and Chemistry*31 (5-6): 305–319.

[12] http://www.r-project.org/about.html

[13] Jeff Gentry, "Package twitteR" February 20, 2015. Title R Based Twitter Client Description Provides an interface to the Twitter web API. Version 1.1.8