POWER USER DETECTION

AKIL ADESHWAR(2013103002)
BHAGYASHREE(2013103502)
ASWIN.M(2013103006)
College of Engneering Guindy.

April 21, 2016

1 Acknowledgement

We would like to thank Dr. Mahalakshmi.G.S for constantly motivating us and identifying the areas of improvement on the project. We would like to extend our immense gratitude to our project guide, for her perpetual support and able guidance which was instrumental in taking the project to its successful conclusion. Finally, we would like to thank the Head of the department, Dr. D. Manjula for providing a conducive environment and amenities to facilitate our project work.

Contents

1 Acknowledgement **2**

2 Keywords **4**

3 Introduction **4**

4 Related Works **5**

5 Problem Description **6**

6 Power User Definition **6**

7 Power User Detection-Methodology **6**

7.1 Dataset Collection 6
7.2 Processing the Dataset 8
7.2.1 Calculating User Mentions 9
7.2.2 Retweet History 9
7.2.3 Hashtag Analysis 10
7.2.4 Favorite Tweet Analysis 10
7.2.5 Score Calculation 10
7.2.6 Phase 1 Results Calculation 10
7.3 Power User-Phase 2 10
7.3.1 Collection Phase 10
7.3.2 Topic Modeling Phase 11
7.3.3 Similarity Calculation 12
7.3.4 Phase 2 Result Calculation 13
7.3.5 Power User Detection 13

8 Tools **13**

9 Assumptions and Limitations **13**

10 Future Work and Conclusion **14**

10.1 Topic Modelling using Photo Analysis 14
10.2 Follower/Following Matching 14
10.3 Topic Modelling using Song Analysis 14
10.4 Interest segregation based on Fashion 14

Abstract

An enormous number of tweets are generated everyday. This provides huge amounts of data to analyze, recognize patterns, construct models and predict user behavior. Tweet Analysis can help understand user behavior and help service providers improve their user experience. In this paper, we define a power user and propose a method to identify whether the base user is a power user based on their tweets, favorites, re-tweets, hash tags and mentions. Dataset is collected for a base user. This dataset is then processed and analyzed based on the Power User Detection method. The Power User Detection is a method which is used to detect whether a cycle of influence exists for a particular user. The power user detection method involves two phases. The results obtained using this method is dynamic as user interests vary with time.

2 Keywords

Power user, Twitter , Tweets ,influential user

3 Introduction

Twitter is an online social networking service that enables users to send and read short 140-character messages called "tweets". Registered users can read and post tweets, but those who are unregistered can only read them. Users access Twitter through the website interface, SMS or mobile device app.Twitter was created in March 2006 by Jack Dorsey, Evan Williams, Biz Stone, and Noah Glass and launched in July 2006. The service rapidly gained worldwide popularity, with more than 100 million users posting 340 million tweets a day in 2012. The service also handled 1.6 billion search queries per day.Users can tweet via the Twitter website, compatible external applications (such as for smartphones). Users may subscribe to other users tweets this is known as "following" and subscribers are known as "followers" or "tweeps". Individual tweets can be forwarded by other users to their own feed, a process known as a "retweet". Users can also "like" (favorite) individual tweets.Twitter allows users to update their profile via their mobile phone by apps released for certain smartphones and tablets.

Detection of the influential user have helped users with providing interested tweets to them. Our aim is to extend the idea of influential users, a user is said to be a power user if he/she gets influenced by users and he/she influences users. A power user hence, influences and gets influenced. Tweets of a user along with his favorites,re-tweets,mentions and hash tags are collected and given for further processing. After processing is done, a user who influences the base user is found this is end of Phase 1 which is, processing the dataset. The dataset processing is explained clearly in the section 6.2. In Phase 2(in section 6.3), topic modelling is done with the dataset collected and a matching percentage is found for each user against the base user. If the matching value is greater than the threshold value it means that the base user influences that particular user otherwise that user is just ignored.

4 Related Works

[1]In the micro blogging networks, there exist users or actors who attract different users of the network towards the documents posted by him or her. Under this attraction, different users utilize the different blog services and usage of these services becomes viral. They are referring these users as Influential Users. These users compel other users to actively use the different services provided by micro blogging networks upon their posted documents.

[2]Into the bloggers or a blog network, there are some users who cause a great influence over other users of the network. They refer these kinds of users as Influential Users (IU). IUs are those users that cause the other users to do some actions on the documents and contents published by him or her. The IU is being used by different organizations for viral marketing by using blogging sites. The organization wants to market a new product by using a small group of potential users to get profit. They focused on the various approaches that helps in determination of IUs, some of them are based on the topology of the social network and some are based on hyperlink and later we discuss the new approach to finding the influential user which is based on the activities that the users performs in social networks, utilizing their diffusion history.

[3]Discovering top-k influential users plays a central role in many social network applications.they study a challenging problem of discovering item-based top-k influential users in social networks. Specifically, they present a dynamic selection approach (referred to as Item-based top-K influential user Discovering Approach, IDA for short), to identify the top-k influential users for a given item based on real-world diffusion traces and on-line relationships. In particular, IDA first softly divides users involved in a diffusion trace into different communities by topic, and ranks users' influence degrees in these topic communities with activeness, follower-counts, and follower participation-rates (including forwards and comments). In doing so, the top-K influential users for a given item can be obtained w.r.t. different topic communities. Experimental results on real world data sets demonstrate the performance of our approach.

[4]On-line support forums are a common method for businesses to provide product support for customers. In addition to trouble-shooting and how-to guides, on-line forums also serve the important purpose of allowing customers to interact and discuss the business's products. These interaction are an important factor in influencing customer opinions, and subsequently the adoption and use of products and services. The identification of influential users on these forums would therefore enable businesses to more effectively disseminate information and communicate with customers. In this paper we develop a method for identifying influential users in support forums using topical expertise and social network analysis. One of the key challenges when analyzing influence in this context is that the users are generally less socially active than users on other social networks such as Twitter and Facebook. In order to address this issue we have taken a broader view of a social network and considered all of the users that a particular user has interacted with instead of just the subset of users for which there is an explicit relationship. The user's expertise in a

particular category is then used to determine the weight or influence of each individual interaction. Finally, the influence of the top influential users is then categorized as positive or negative based on sentiment analysis of their posts.

[5]Social Influence can be described as the ability to have an effect on the thoughts or actions of others. The objective of [5] is to investigate the use of language in detecting the influential users in a specific topic on Twitter. From a collection of tweets matching a specified query, we want to detect the influential users from the tweets' text. The study investigates the Arabic Egyptian dialect and if it can be used for detecting the author's influence. Using a Statistical Language Model, we found a correlation between the users' average Retweets counts and their tweets' perplexity, consolidating the hypothesis that SLM can be trained to detect the highly retweeted tweets. However, the use of the perplexity for identifying influential users resulted in low precision values. The simplistic approach carried out did not produce good results. There is still work to be done for the SLM to be used for identifying influential users.

5 Problem Description

When users login on Twitter, they see a stream of tweets sent by friends which composes their timeline. Many of these tweets are conversational tweets and/or are not of personal interest to the user.The goal of our model is to detect the cycle of influence for a particular user so that they can interact more with that influencer.

6 Power User Definition

A user is said to be a power user if he/she has an influential user and also influences other users. In this paper, we use the Power User Detection method to identify power users in a given dataset. Power users are common links between two sets of users in a given dataset. Power Users can be used to identify super influential users in a given dataset.

7 Power User Detection-Methodology

Power User Detection is a method used to detect the powers users in a given dataset. This method involves two phases. In phase one, the influential users with respect to the base user are identified. In phase two, the users to whom the base user is the influential user are identified. If for a given user, both phase one and two are successfully done, then the given base user is a power user. The block diagram in 1 depicts various stages in both the phases. The two phases are dicussed in detail below.

7.1 Dataset Collection

We used the Twitter API to gather information about a users social links and tweets. We launched our crawler for all user IDs ranging from 0 to 80 million. This API has a restriction of 15 requests per 15 minutes. We did not look

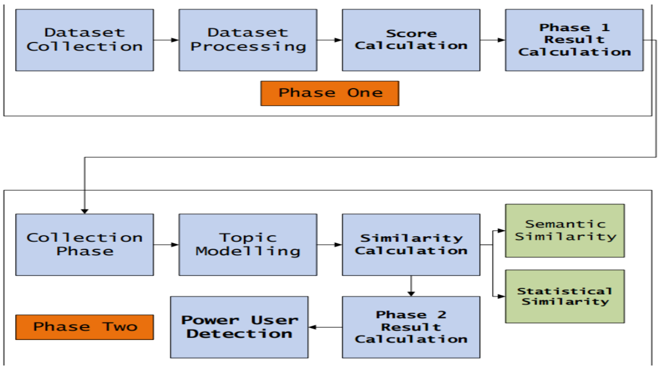

Figure 1: BLOCK DIAGRAM

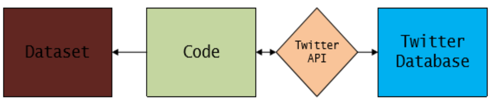

Figure 2: DATASET COLLECTION

beyond 80 million, because no single user in the collected data had a link to a user whose ID was greater than that value. Out of 80 million possible IDs, we found 54,981,152 in-use accounts, which were connected to each other by 1,963,263,821 social links. We gathered information about a users follow links and all tweets ever posted by each user since the early days of the service. In total, there were 1,755,925,520 tweets. Nearly 8% of all user accounts were set private, so that only their friends could view their tweets. We ignore these users in our analysis. The social link information is based on the final snapshot of the network topology at the time of crawling and we do not know when the links were formed. The network of Twitter users comprises a single disproportionately large connected component (containing 94.8% of users), singletons (5%), and smaller components (0.2%). The largest component contains 99% of all links and tweets. Our goal is to explore influence of users, hence we focus on the largest component of the network, which is conceptually a single interaction domain for users. Because it is hard to determine influence of users who have few tweets, we borrowed the concept of active users from the traditional media research and focused on those users with some minimum level of activity. We ignored users who had posted fewer than 10 tweets during their entire lifetime. We also ignored users for whom we did not have a valid screen name, because this information is crucial in identifying the number of times a user was mentioned or retweeted by others. After filtering, there were 1,048,636 users, whom we focus on in the remainder of this paper.
We have also collected the dataset based on some unique characteristics as 4 csv files separately for each user based on his/her screen name in twitter.

First file is screen_name_tweets.csv which has id, account created date, tweet, entities, retweet_count, favorites_count, in_reply_to_screen_name, language.

Second file is screen_name_retweets.csv which has id, account created date, tweet, entities, retweet_count, favorites_count, in_reply_to_screen_name, language.

Third file is screen_name_mentions_count.csv which has a screen_name which the user has mentioned and its count on the other column. It is used to find the maximum mentioned person for a particular user.

Fourth file is screen_name_hashtag_count.csv which has a @screen_name which the user has used and its count on the other column.

Majority of the dataset was collected using the Tweepy Python Module. This is a wrapper API for the Twitter API. Python was used to collect the dataset. Python was the primary programming language used to collect the dataset. Around 10 GB of dataset was collected to test the Power User Detection Method.

7.2 Processing the Dataset

Dataset Processing is the second stage in phase 1. Once the dataset has been collected, it has to be processed in order to make any inference. Dataset processing plays a major role in phase one as it segregates the dataset in to vital parts which can be used during the score calculation stage. Processing is done based on the tweet information contained in the dataset. Dataset processing

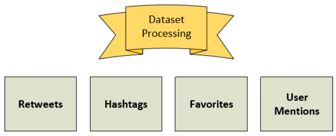

Figure 3: DATASET PROCESSING

	A	B	C
1	Name	Count	
2	CarterCen	3	
3	nickcollisc	1	
4	ndtv	4	
5	AJLifeline	1	
6	anildash	1	
7	jeremys	2	
8	KellyAyot	1	
9	TEDchris	6	
10	duolingo	1	
11	gatesfoun	59	
12	tomfriedn	1	
13	TED_TALK	1	
14	StateDept	2	
15	RyanSeacr	8	
16	RecordSet	1	
17	KevinSpac	1	
18	jeancase	1	
19	WuDunn	1	
20	FCBarcelo	2	
21	AdeAdepi	3	
22	globalfun	2	
23	iack	1	

Figure 4: Results of Mentions count

involves four major sub stages. These stages help model user behavior and provide information on user interests. The four sub stages are as follows:

7.2.1 Calculating User Mentions

Every tweet by the base user may contain mentions of other users.The number of user mentions for every user is calculated and stored in a separate file. The user mentions are obtained from tweets, retweets, favorites and hashtags. If a hashtag forms a substring of a user, the user mentions count of that user is incremented by one. User mentions is one of the important factors for score calculation as the base user directly mentions the target user in the tweets. User mentions from retweets are also added.

7.2.2 Retweet History

A Retweet is a tweet shared by the base user but created by another user. Retweets help in understanding what topics the user wants to share with others. In this paper, retweets is majorly used in topic modeling so obtain the topics which interest the user. For every retweet, the mentions count of the owner of the retweet is incremented by one.

7.2.3 Hashtag Analysis

Hashtag refers to a word that begins with the symbol "#". Hashtags generally refers to collection of words used by an user to describe the context of the tweet. Hashtags are used in topic modeling and user mentions count as mentioned above.

7.2.4 Favorite Tweet Analysis

Favorites refer to the tweets liked by an user. Favorites majorly define the interests of the base user. For every favorite, the mentions count of the owner of the favorite tweet is incremented by one. Favorite tweets can be used to model the favorite topics of the base user.

7.2.5 Score Calculation

The score calculation stage uses the files generated by the data processing stage. The scores are provided to each user in the files mentions above based on constant multiper value. The score calculation stage assigns each user with a certain score with respect to the base user.

Score Calculation Formula:

Final User Score: 1 * Tweet Mentions Count + 0.5 * Hashtags Mentions Count + 0.5 * Retweets Mentions Count + 1 * Favorites Mentions Count

The score calculation formula is used to calculate the score of each user with respect to the base user. The results of score calculation are stored in a separate file.

7.2.6 Phase 1 Results Calculation

Phase 1 result calculation is the final stage of phase one. This stage uses the file generated by score calculation stage. In this stage, the scores file is sorted in a non increasing order based on the scores of each user. The top ten users are obtained from the new sorted list. The top ten users are stored in a separate file. The file serves as the input to phase two. The top 10 users are the influential users with respect to the base user and the user with the highest score being the most influential among them. This completes phase one of the Power User Detection methodology.

7.3 Power User-Phase 2

7.3.1 Collection Phase

Collection stage is the first stage in Phase two. The output of the phase 1 results calculation stage is the input to the collection stage of phase 2. The collection stage involves collection users who may be influenced by the base user. The list

Name	Score
RealHugh.	67.2
Deborra_l	51.1
livelaughi	48.3
TheRiverF	25.2
MPTF	21
EddieEagl	15.4
jimmyfall	14.7
TheGPP	14.7
panmovie	13.3
TaronEger	11.9
MatthewT	9.8
Hughceva	9.8
GlblCtzn	9.8
WorldVisi	9.8
AdoptCha	9.1
iTunesMo	9.1
Wimbled(	9.1
russellcro	7.7
wbpicture	7
GusWorla	6.3

Figure 5: Phase 1 Results Calculation

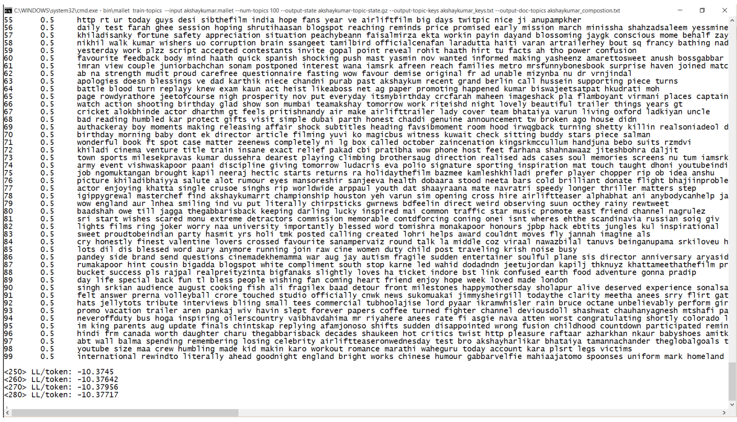

Figure 6: Results from mallet

of users collected during this stage are stored in a separate file and given as input to the topic modeling stage. Collection stage, only involves users who belong to the dataset.Collection stage collects all the information including retweets, hashtags, tweets, favorites and user mentions with respect to given user. All the above mentioned details are stored in a separate file for each user. This file is used for topic modeling.

7.3.2 Topic Modeling Phase

Topic Modeling plays a major in phase 2. Topic modeling is done separately for each collected user. We used Mallet as the primary tool for topic modeling. Mallet uses an optimized LDA algorithm at its base to generate topics. The results of topic modeling depicts the target users interested topics. The results of topic modeling are stored in a separate file for each user. The top hundred topics are generated for each user. Mallet trains itself multiple times to generate the final topics file.

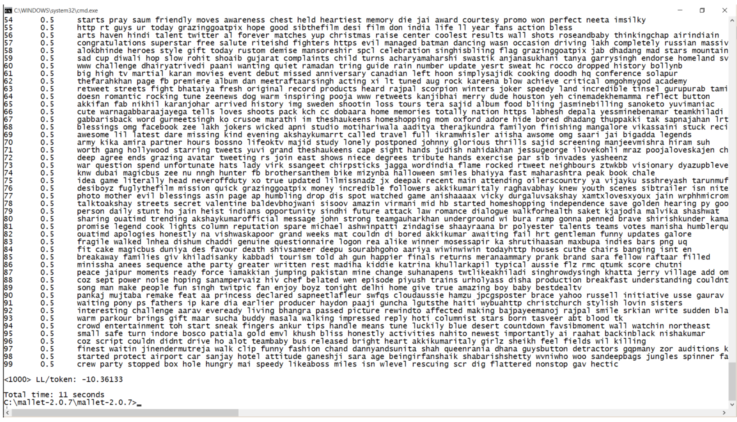

Figure 7: Results from Topic Modeling

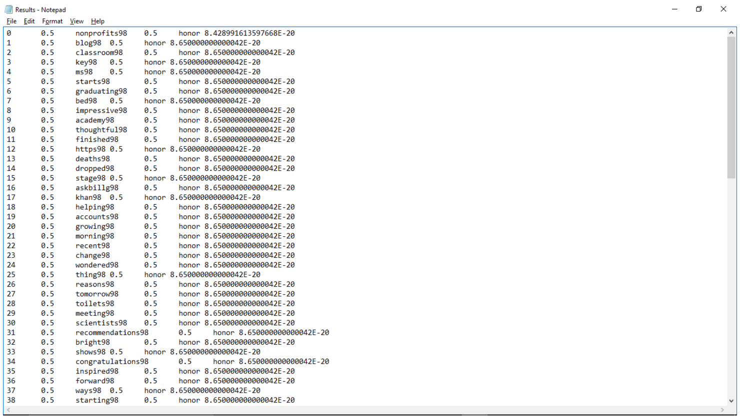

Figure 8: Results from Similarity calculation

7.3.3 Similarity Calculation

The output of topic modelling is used by the Similarity Calculation stage. In this stage, the topic modelling file of the base user is compared with the topic modelling files of the rest of the users. To compare the similarity of the two files, two types of similarity checking is used - Statistical and Semantic. Under statistical similarity,Kumar Hasan Brook is used to calculate similarity. Under semantic similarity, Wu and Palmer similarity is used. Similarity calculation helps identify helps identify the measure of common interest between two users. The similarity score is calculated as below:

Similarity Score = 0.75 * Semantic Similarity + 0.25 * Statistical Similarity

The similarity score is calculated for the collected users with respect to the base user. All the scores are stored in a separate file which is used by final result calculation stage.

7.3.4 Phase 2 Result Calculation

Phase 2 result calculation is the final calculation in Power User Detection method. This stage uses the file generated by similarity score calculation stage. In this stage, the scores file is sorted in a non increasing order based on the scores of each user. The number of users for whom the similarity score is more than 40% is calculated and stored in a separate file.

7.3.5 Power User Detection

Power User Detection stage is the final stage of this process. The results of the previous stage is used in order to the finalize the result. If the similarity score is more than 40% for at least one collected user, then base user becomes the influential user for that user based on topic modelling. If both the phases produce successfull results, the base user is both influencer and influenced, resulting in becoming the Power User, a strong link in the network. Thus, using the Power User Detection method, we have successfully identified the base user as the Power User.

8 Tools

Tweepy: Tweepy is a python package used as a wrapper for the Twitter API. It simplifies the use of Twitter APIs. Tweepy provides great functionality to easily access twitter data with OAuth requirements. Version: 3.3.0

Excel : All the data in used in paper was always stored in 'csv' file. Excel made it easier to organize and manipulate data.

Python Scripts: Scripts written in python were used to process the dataset. Version: 2.7

Mallet: Mallet was used for topic modelling. Mallet is written in Java. Mallet command line can also be used. Version: 2.0.7

Semantic Similarity: Semilar was used to calculate semantic similarity. Semilar is built on Java. Semilar uses Standford CoreNLP Apache OpenNLP, and WordNet for calculating semantic similarity. Version: 1.0.2

9 Assumptions and Limitations

The following contain the assumptions and limitation of this paper:
1.The dataset contains all the influential users with respect to a given base user.
2.Only a small subset of factors are considered in influential user detection. With improvement in technology, more new factors will rise.
3.Mallet is used for topic modelling. The accuracy of this paper depends on the accuracy of mallet as a topic modelling tool.
4.Similarity Calculation assumes both Wu-Palmer Similarity and Kumar Hasan

Brook always produce best possible results.
5.Only top 100 favorite topics are taken for a given user.

10 Future Work and Conclusion

10.1 Topic Modelling using Photo Analysis

A lot of pictures are posted as a tweet. Machine Learning algorithms can be used to detect topics in a given picture. This can also be done using IBM Watson API. Given a picture IBM Watson predicts different objects present in the picture to a great accuracy.

10.2 Follower/Following Matching

Measure of Follower/Following matching for the base user and target user can be considered as a factor in Influential User detection.

10.3 Topic Modelling using Song Analysis

Music sharing has never been so popular in the human history. User interests can be modeled by the genre of songs the user prefers. This can be an added factor to determine common interest

10.4 Interest segregation based on Fashion

Image processing can be used to figure out the fashion of the base user and the target user. The measure of fashion similarity also depicts measure of common interests.

In this paper, we have defined Power User and proposed a method, Power User Detection which can be used to detect Power Users in a given dataset. Power user play a vital role in a given dataset. They act as strong links between users.

References

[1] Yadav, Mahendra Kumar, and Manoj Kumar. "Determining influential users in blogosphere-A survey." Green Computing, Communication and Conservation of Energy (ICGCE), 2013 International Conference on. IEEE, 2013.

[2] Singh, Sushil, Nitesh Mishra, and Shantanu Sharma. "Survey of various techniques for determining influential users in social networks." Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), 2013 International Conference on. IEEE, 2013.

[3] Guo, Jing, et al. "Item-based top-k influential user discovery in social networks." Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on. IEEE, 2013.

[4] Munger, Tyler, and Jiabin Zhao. "Identifying Influential Users in On-line Support Forums using Topical Expertise and Social Network Analysis." Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. ACM, 2015.

[5] Shalaby, May, and Ahmed Rafea. "Identifying the Topic-Specific Influential Users Using SLM." 2015 First International Conference on Arabic Computational Linguistics (ACLing). IEEE, 2015.