IOP Conf. Series: Materials Science and Engineering **263** (2017) 042071 doi:10.1088/1757-899X/263/4/042071

Finding user personal interests by tweet-mining using advanced machine learning algorithm in R

Krithika L B, RoyP and Asha Jerlin M

School of Information Technology and Engineering, VIT University, Vellore-632014, Tamil Nadu, India.

E-mail**:** krithika.lb@vit.ac.in

AbstractThe social-media plays a key role in every individual's life by anyone's personal views about their liking-ness/disliking-ness. This methodology is a sharp departure from the traditional techniques of inferring interests of a user from the tweets that he/she posts or receives. It is showed that the topics of interest inferred by the proposed methodology are far superior than the topics extracted by state-of-the-art techniques such as using topic models (Labelled LDA) on tweets. Based upon the proposed methodology, a system has been built, "Who is interested in what", which can infer the interests of millions of Twitter users.

A novel mechanism is proposed to infer topics of interest of individual users in the twitter social network. It has been observed that in twitter, a user generally follows experts on various topics of his/her interest in order to acquire information on those topics. A methodology based on social annotations is used to first deduce the topical expertise of popular twitter users and then transitively infer the interests of the users who follow them.

1. Introduction

Today e-commerce market is the most competitive market in the business world; in order to sustain this capacity to be a leading company in the market various kind of latest technologies and the development environments are used. To lead in the market has become very necessary to perform application and system performance monitoring's for better customer based targeting to attract more clients and have more revenue. A system is required that will automatically get user interests. Based on that, it can predict what he /she can buy online and with that data our recommendation engine sends notifications to each user about new and updated products which can be of that person interest. These are completely automatic machine learning program which will self-sustain and self-replicate over time period to be more perfect in generating what a specific user wants to buy.

2. Related Work

Using the twitter tagging [1] and labels concept a list of categories or technically lists are built. Those lists acts like key to datasets which falls under same category and that will be used to map to user intrigues. Using list names and portrayals to find the topical aptitude of prevalent clients in twitter is the important aspect of that. The methodology in this paper suggests to distinguish themes of aptitude of a client , let's assume one test subject, v (whom he has subscribed to), they gather the lists which have v as a part, and concentrate the most regular terms that show up in the names and depictions of

the lists. Then recognize v as a specialist on a theme t if and as it were on the off chance that v has been recorded not no less than 10 times, i.e., if the term t shows up no less than 10 times in the names or depictions of the Records containing v. 1 Similar to [4, 10], this function has considered as themes just unigrams (single words, for example, 'legislative issues', 'music') furthermore, bigrams (two words which as often as possible happen together, e.g., 'online networking', 'computer game', 'inlet zone') .

In [2] Al Assad has provided the details about a small implementation about using the twitter API to get all the public data available in the twitter sphere which contains the hash tag of #PrayForMH370 to know the vibe of people about the Malaysian aircraft accident which was never found. In this post, they have done text data mining on Twitter tweets containing #PrayForMH370 from March 8 to March 20, 2014 using twitter API. First, they made an authentication on the twitter API, to obtain the data. In [3] the researchers have talked about geographical data rendering using tweets. The initial step is to limit down the subject - "digging tweets for geographic information" is an entirely expansive definition. By and large, however, just gather tweets with one of the three APIs - search, REST or streaming - and concentrate the geographic information from them. They can then plot the information on a guide or do some more perplexing investigation. It's entirely direct to separate geo-tagging in most scripting dialects.

In paper [4], it's been discussed that the customer wished to utilize its online networking knowledge stage to run investigations on particular tweets. These tweets needed to have an arrangement of catch phrases or expressions and twitter was to be checked for the same. What's more, tweets were to be gathered just from particular nations in Europe. Since base contribution was high and customer's center competency was not slithering huge datasets, they needed a facilitator for every one of this information.

While in the article [5] some social locales like 'Facebook' and 'LinkedIn' require the regular affirmation of a relationship between customers (which generally deduces a certifiable affiliation or something to that effect), Twitter's relationship model grants you to stay mindful of the latest happenings of some other customer, in spite of the way that that other customer may not tail someone back or even understand that that person exist.

From the survey on various twitter based systems and implementations related to the problem definition it has defined how the solution is different than others in terms of methodology and functionality. And in terms of speed how it is optimized at its level. While legacy platforms like Python and Java are there for implementing the proposed machine learning system, it has preferred the state of the art programming platform R because of its vast usage on data and the compatibility with large set of data and processing them efficiently.

3. Proposed Work

3.1. Various Kinds of Data-sets from Twitter Mining

Twitter is an online interpersonal interaction benefit that empowers clients to send and read short 140-character messages called "tweets". Enrolled clients can read and post tweets, yet the individuals who are unregistered can just read them. Clients access Twitter through the site interface, SMS or cell phone application [11]. Twitter Inc. is situated in San Francisco and has more than 25 workplaces around the globe. Twitter is a vast public data universe, various types of data sets are available there which can be used to get into some logical data driven conclusion.

Twitter may be portrayed as a continuous, exceptionally social micro blogging administration that permits clients to post short notices, called tweets that show up on courses of events. Tweets may incorporate one or more elements in their 140 characters of substance and reference one or more

places that guide to areas in this present reality. A comprehension of clients, tweets and courses of events is especially crucial to powerful utilization of Twitter's API, so a brief prologue to these essential Twitter Platform articles is all together before it is communicated with the API to bring a few information.

3.2 Tweet Mining Process

For a given Twitter client ‘u’ (whose hobbies are to be gathered), proposed technique comprises of the accompanying two stages. Initially, the system checks which different clients ‘u’ is taking after, i.e., clients from whom u is occupied with accepting data. Second, it distinguishes the points of ability of those clients (whom u is taking after) to induce u's hobbies, i.e., the subject on which u is occupied with getting data. Inferring topical skill utilizing Twitter Lists: Lists are an authoritative element, by which clients can gather specialists on subjects that intrigue them [1]. To make a list, a client determines a name and a discretionary portrayal, and after that includes different clients as individuals from the List; for case, a client can make a List named "Music and artists", and include records, for example, Lady Gaga, Katy Perry, and Yahoo Music. In earlier work [4, 10], a system proposed for using list names and portrayals to find the topical aptitude of well-known clients in Twitter.

To distinguish subjects of mastery of a client ‘v’ (whom u has subscribed to), the algorithm gathers the Lists which have ‘v’ as a part, and concentrate the most widely recognized terms that show up in the names and portrayals of the Lists. Consider ‘v’ as a specialist on a subject t if and only if v has been recorded on ‘t’ no less than 10 times, i.e., if the term t shows up to not less than 10 times in the names or depictions of the Lists containing v. Similar to [10], it has been considered as subjects just unigrams (single words, for example, 'legislative issues', 'music') and bigrams (two words which much of the time happen together, e.g., 'online networking', 'computer game', 'sound region') which are recognized as things or modifiers by a standard grammatical feature tagger.

The earlier work has demonstrated that this procedure precisely induces the themes of mastery of a huge number of well-known clients in Twitter. For example, a few points of skill of the client account ‘@BarackObama’, as interpreted by the above strategy, are 'legislative issues', 'celebs', 'government'. Correspondingly, a few themes surmised for the client account ‘@linuxfoundation’ are 'tech', 'Linux', 'programming', 'PC', and "designer. Understanding client intrigues: For the given client u (whose hobbies are to be inferred we utilize the above list-based philosophy to distinguish the subjects of aptitude of those to whom ‘u’ has subscribed. Naturally, if a client subscribes to tweets from a few specialists on a specific point, then the client is prone to be occupied with that theme. Machine considers u to be keen on point t if and just if u subscribes to no less than 3 specialists on theme t. In this manner, the process acquires an interest vector for u, which is a positioned rundown of themes, positioned on the premise.

3.2.1 Extracting the Data from Twitter

It just needs OAuth2 validation which is more straightforward to execute than OAuth1. When system have become it’s OAuth2 keys and ACCESS_TOKEN, the predefined application made in simple twitter base makes it simple to recover a rundown of supporters or tweets from a timetable. There is a coordinated correspondence between the strategy and the twitter API call. For every adherent being stored, the (maximum) 200 latest tweets, the dialect of the record, the quantity of tweets acquired and the screen name of the record. The R code to extract the information from twitter is accessible here total tweets into records, from where one record is made for each devotee by conglomerating his/her tweets. Every adherent then has a special archive which serves as the premise for our point investigation. Having archives in a few dialects will add demand to the point extraction and need to sift through timetables that are not 100% in English. Sifting clients by the "Lang"

parameter of their twitter record is not 100% solid. A few clients tweet in a few dialects in spite of the fact that their record dialect is proclaimed as "en" while others have not characterized the dialect of their record. (Lang = "und" for not defined).

In spite of the fact that it's sufficiently simple to skim the two arrangements of patterns and search for shared trait, R's set information structure should be utilized to consequently figure this for the system, since that is precisely the sort of thing that sets lend themselves to doing. In this example, a set alludes to the numerical thought of an information structure that stores an unordered accumulation of special things and can be registered upon with different arrangements of things and set wise operations. For instance, a set wise crossing point registers normal things between sets, a set wise union consolidates the majority of the things from sets, and the set wise contrast among sets acts kind of like a subtraction operation in which things from one set are expelled from another.

3.2.2 Cleaning the Mined data

Then the system clean up and prepare the documents for LDA like,

a) Remove URLs.
b) Remove documents with less than 100 words.
c) Tokenize: breaking the documents into words. This also removes punctuation.
d) Remove stop words, words that only occur once, digits and words composed of only 1 or 2 characters.

3.2.3 Applying LDA on the Clean Data-Set

Finding the right parameters for LDA is 'a workmanship'. Three fundamental parameters should be advanced:

a) K: the quantity of themes.
b) Alpha: which manages what number of themes a record possibly has. The lower alpha, the lower the quantity of points per records.
c) Beta: This directs the quantity of word per report.

Since the system is managing tweets, every devotee would have a set number of points to tweet about and along these lines set alpha to a low esteem 0.001. It is cleared out beta to its default setting.

3.3 Architectural Design

The proposed Twitter Mining Process as shown in Figure 1 has different phases namely: Calling Twitter API,mining tweets,process raw data, TDM creation and TF-IDF,stemming and lemmatizing,generating user interests.

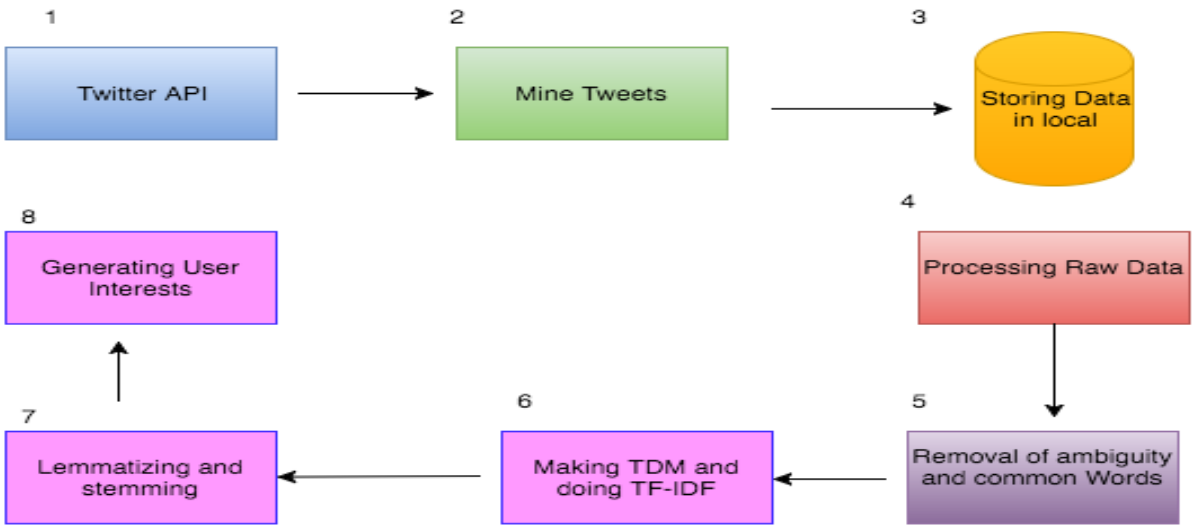

Figure 1.System Architecture of the proposed Twitter Mining Process.

3.4 Module Description

The proposed system is an application module, being a separate entity it results outcomes which will be used by further programs. The idea of tweet mining is some person's tweets and generate their user interests is very scientific as most of the people share their views about their interests in social media. This whole process has been divided into 6 modules and below all of them has been described briefly:

●Step 1: Calling Twitter API
Using the Twitter API via an authorized twitter app (which has been created for development purpose) and by specifying the user-name of a person, a request for getting all his tweets Twitter has a restriction of providing maximum 3200 tweets but here only totally 1000 tweets are used.

●Step 2: Mining Tweets
After a successful request tweet responds with a list of tweets along with person and those data are mined into local storage.

●Step 3:Process Raw Data
Now the whole data has been taken in consideration and raw data is cleaned by using clearance of random texts, symbols and unnatural strings using TM library.

●Step 4:TDM creation and TF-IDF
A term document matrix table is created which has all the frequency of words which occurs in the cleaned data and they are ordered according to their frequency. Also 'TF-IDF' concept is used to determine which use less word that has higher priority.

●Step 5:Stemming & Lemmatizing
These processes involves around making the corpus more subtle and consistent in terms of grammar and raw forms of each words.

●Step 6:Generating User Interests

Now using LDA topic modelling technique the most frequent interests of the user will be generated.

4. Results and Discussion

4.1. Results of the User Interests extraction

The figure 2, 3 and 4 are the results achieved for different user interest. For the test-purpose, the algorithm has been tried on some people's personal twitter handle with their consent via the application and their public twitter data has been mined and used to got the results.

4.1.1. About the Result

This is basically a word cloud having all the Interest Maps that the person has with most interested fields as bigger fonts are then smaller as the least interest. In all the images, the terms have been highlighted using colours and font sizes. This indicates certain significance, like being the bigger text tells that topic which has higher priority than other smaller ones. So from Table1it can be concluded that the first guy is most interested in EmberJS, WEB, Software, JavaScript and so on. While the second person is more like worldwide person and likes to follow business and development, third one is a movie buff and totally into movies and Bollywood.

Figure 2. User Interests for the handle @jaydevgajera

Figure 3. User interests for the handle @billgates

Figure 4. User interests for the handle @narendramodi

4.2 Comparison of the Original and the Resultant Interests

The real and actual interests of the people and the system generated interests are shown in table1 which comprises of twitter handle, original interest and resultant interest.

Twitter Handle	Original Interests	Resultant Interests
@jaydevgajera	Web development, JavaScript , HTML, CSS, UI , ember-Js ,User Interface, Code , engineer	Web , JavaScript , Development ,code, UI , Engineer
@nayanjadeja	Politics , Coding, Research , Leadership, Entrepreneur , Build , Deployment	Politics, Leadership, Build
@roypartha97	Coding , Java , JQuery , HTML ,WEB , Developer , Implementation , Movies , Bollywood	Java , Bollywood , HTML , Movies
@mb91	Singing , Dance , Music , News	Songs , Dance , Art
@billgates	Computer , Literacy , Philosopher , windows , development	Windows , development

Table 1. The real and actual interests of people and the system generated interests

5. Conclusion

Amongstthe assessment of utilizing human input, a few evaluators remarked that however the top taken themes precisely caught their expansive advantages; the points were infrequently excessively broad. They liked to see more particular premiums, for example, 'machine learning' or 'huge information' rather than "science" or 'innovation'. It has been watched that the more particular hobbies are without a doubt derived by the proposed technique; in any case, these particular hobbies were not getting incorporated into the main 20 themes (positioned by number of topical specialists that a client takes after) that were at first appeared to the evaluators. To consider this input, subjects were ordered into two classes in view of their all-inclusive statement, which was gauged by the worldwide number of clients who are keen on a theme. Out of the 36 thousand particular points gathered (as expressed above), the main five percentile of themes are considered as "general" subjects (on which there are a huge number of intrigued clients), and whatever remains of the less well known points as "corner" themes.

References

[1] Bhattacharya P 2002 Inferring User Interests in the Twitter Social Network *Proceedings of SPIE* **4666** 149–160
[2] Chen J, Nairn R, Nelson L, Bernstein M and Chi E H 2010 Short and tweet: experiments on recommending content from information streams *Proceedings of the 28th International Conference on Human Factors in Computing Systems DOI: 10.1145/1753326.1753503*
[3] Hong L and DavisonB D 2010 Empirical Study of Topic Modeling inTwitter *1st Workshop on Social Media Analytics* 80-88
[4] Smith J 2012 Mining Twitter with Python *ACM SIGCHI*
[5] Upadhyay A,Mao L and Krishna M G 2005 Mining data from twitter *Proceedings of ICIP* **2** 1110-1113
[6] http://stackoverflow.com/questions/4308554/simplest-way-to-read-json-from-a-url-in-java

[7] https://blog.twitter.com/2012/studying-rapidly-evolving-userinterestshttp://nlp.stanford.edu/IR book/html/htmledition/stemming-and-lemmatization.
[8] http://chimera.labs.oreilly.com/books/1234000001583/ch01.html
[9] http://alstatr.blogspot.in/2014/03/r-text-mining-on-twitter-prayformh370.html
[10] https://twittercommunity.com/t/mining-tweet-data/2680
[11] http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/
[12] http://www.cs.columbia.edu/~blei/topicmodeling.html
[13] https://github.com/minghui/Twitter-LDA
[14] http://blog.mathandpencil.com/using-latent-dirichlet-allocation-to-categorize-my-twitter-feed/