ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web

Jasmine Kanav[1], Jyoti Killedar[1], Rutuja Trimbake[1], Trupti Sapkale[1], Aparna Hajare[2]

Bachelor of Engineering, Department of Computer, Cummins College of Engineering for Women, Pune, Maharashtra, India[1]

Assistant Professor, Department of Computer, Cummins College of Engineering for Women, Pune, Maharashtra, India[2]

ABSTRACT: Twitter has millions of tweets of users on almost all topics under the sun. This data is extremely big and consists of huge amount of tweets, retweets, trends and much more to add to the context. We have observed that in twitter, a user generally follows another user based on his interests. It is challenging to find interest/likes of that person. The person might tweet about certain things more frequently, but it is a challenging task to analyze whether his tweets are in favor of the topic or against.Users' profiles from social media websites such as Facebook or Google Plus are used as a distant source of supervision for extraction of their attributes from user-generated text. In addition to traditional linguistic features used in distant supervision for information extraction, our approach also takes into account network information, a unique opportunity offered by social media. The project investigates semantic user modeling based on Twitter posts.Our system is analyzing the tweets of a user and then it is segregating them into two domains, namely music and sports. According tothe number of times a user tweets about particulardomain he/she is designated a level of expertise. This makes searching a topic expert easier.We have included the location feature, wherein a user can look for a topic expert in a particular locality.

KEYWORDS: Tweet; Hashtag; Classification; Stemming

I. INTRODUCTION

Social Networking sites have a huge amount of data. In order to know about the particular interest of a user we would have to go through a large amount of his social networking feed. If we consider the social networking platform Twitter, on a daily basis a user tweets about many things. Some tweets are about important events, such as a warning about a natural disaster, whereas most of the tweets are about daily activities of the user. Thus finding particular interest of the user becomes a mammoth task. Our system analyzes the tweets of a user and accordingly categorizes them into two domains music and sports. Then based on the number of times a user tweets about a particular domain he is assigned a level of expertise. Thus finding topic experts becomes simple. We have also added the location feature, where a user can look for a topic expert in a particular locality. We analyze the tweets of a user registered on our website and based on the number of times he/she tweets about a particular topic we categorize the use into various levels of expertise. The user can be a novice, an intermediate or an expert in a particular domain based on the number of times he has tweeted about that topic. Categorizing the users makes it easier for a third person to look for a domain expert. We also have integrated location services, wherein the exact location of a user will be displayed, which make area wise searching for an expert easier.

II. RELATED WORK

[1] This paper presents real time investigation of events such as earthquakes, using Twitter. For this purpose an algorithm is proposed to monitor tweets and to detect a target event. The algorithm mainly devises a classifier of tweets based on features such as the keywords in a tweet, the number of words and the context in it. Semantic analysis is applied to tweets to classify them into positive and negative class. Along with that, location estimation methods such as

Kalman filters and particle filtering are used to estimate the location of the events. It mainly uses the power of microblogging on twitter for a novel approach to notify people promptly of an earthquake event. The content in the paper and its features are relevant to the features of our proposed project as the same kind of processing will be required.**[2]** The authors of the paper advocate the use of real-time micro-blogging data. It mainly proposes a method to use the micro-blogging data stream to detect fresh URLs. Also the use of microblogging data is extended to compute novel and effective features for ranking fresh URLs. These methods ultimately are the effective ways to improve portal of web search engine for real time search.**[3]** In this paper the authors have crawled the entire twitter site and have obtained 41.7 million user profiles, 47 billion social relations, 4,262 trending topics, and 106 million tweets. They have analyzed the follower and following topology and have found a non-power follower distribution, a short effective diameter, and low reciprocity, which all mark a deviation from known characteristics of human social networks. For identifying influential on twitter, they have ranked users by the number of followers and by PageRank and found two things rankings similar. So again when the ranking was done based on tweets and retweets, it differed from the previous two rankings which indicated a gap between influence inferred from the number of followers and that from the popularity of one's tweets.[4]This paper addresses the knowledge management problem by going beyond the general perspective of information finding in twitter that asks "What is happening right now?" towards an individual user perspective, and asks "What is interesting to me right now?".A collaborative filtering is considered as an online ranking problem and present RMFO, a method that creates, in real-time, user-specific rankings for a set of tweets based on individual preferences that are inferred from the user's past system interactions. Experiments on the 476 million Twitter tweets dataset show that our online approach largely outperforms recommendations based on Twitter'sglobal trend and Weighted Regularized Matrix Factorization (WRMF), a highly competitive state-of-the-art Collaborative filtering technique demonstrating the efficacy of the given approach.**[5]** Finding topic experts on microblogging sites with millions of users, such as Twitter, are a hard and challenging problem. In this paper, author proposes and investigates a new methodology for discovering topic experts in the popular Twitter social network. This methodology relies on the wisdom of the Twitter crowds – it leverages Twitter Lists, which are often carefully curated by individual users to include experts on topics that interest them and whose meta-data (List namesand descriptions) provides valuable semantic cues to the experts' domain of expertise. They have mined List information to build Cognos, a system for finding topic experts in Twitter. Detailed experimental evaluation based on a real-world deployment shows that: (a) Cognos infers a user's expertise more accurately and comprehensively than state-of-the-art systems that rely on the user's bio or tweet content, (b) Cognos scales well due to built-in mechanisms to efficiently up- date its experts' database with new users, and (c) Despite relying only on a single feature, namely crowdsourced Lists, Cognos yields results comparable to, if not better than, those given by the official Twitter experts search engine for a wide range of queries in user tests.

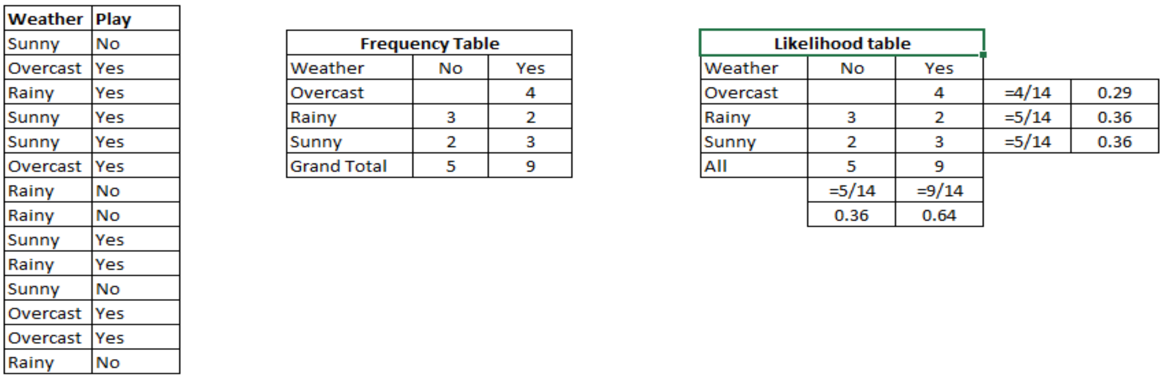
III. PROPOSED ALGORITHM

a. Description of algorithm used

1. Naive Bayes Algorithm

Step 1: Convert the data set into a frequency table

Step 2: Create Likelihood table by finding the probabilities like Overcast probability = 0.29 and probability of playing is 0.64.



Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Fig No 01Naive Bayes Algorithm Flow

Step 3: Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

2. Stemming Algorithm

```
1. For each document Di in the document set
      For each word Wj in the document
            If(not stop-word)
                  Stem Wj using Porter Stemmer
                  Increment TF[i,j]
2. Build the Document Frequency vector using the Term
Frequency matrix
DF[j]=Total values of I for which TF[I,j ] is non zero.
3. Build Inverse Document Frequency vector
  $IDF[j]=\log(D/DF[j])$,D is total number documents
4. Normalize the Term Frequency matrix
  $TF[I,j]=TF[i,j]/\max(TF[i,j]_{j=1\ldots N,}$ N is the total terms
5. Multiply Term Frequency & Document Frequency for each
   Term
            $TF\text{-}IDF[i,j]=TF[i,j]*IDF[j]$
```

IV. PROPOSED SYSTEM

Our system consists of the following modules-

1) User Interface:-

a. Collect the user’s tweets along with all the user details.
b. In this Module we use available dataset for analytics.
c. Here different users register first and then login then they tweet. Different users have different interests.
d. This data is stored in database according to the interests.

2) Tweet feature Extraction:-

a. In this module we are processing on the collected tweets.
b. Extracting tweets from tweet table by Using Feature Extraction Algorithm.
c. Here we extract the positive and negative features on the basis of profile

3)Tweet Analysis:-

d. Analyzing Extracted tweets.
e. Separating these tweets according to the type of the tweet

4)User Profile Generation:-

f. User specific contents are extracted from classified tweets.
g. Find the user profile through the tweet analysis module
h. Obtain the profile along with the location.

5)Tweet Storage:-

i. Storing classified tweets to generate tweet summarization.
j. The separated tweets are stored into database.
k. This tweet is used to identify the user

V. SYSTEM ARCHITECTURE

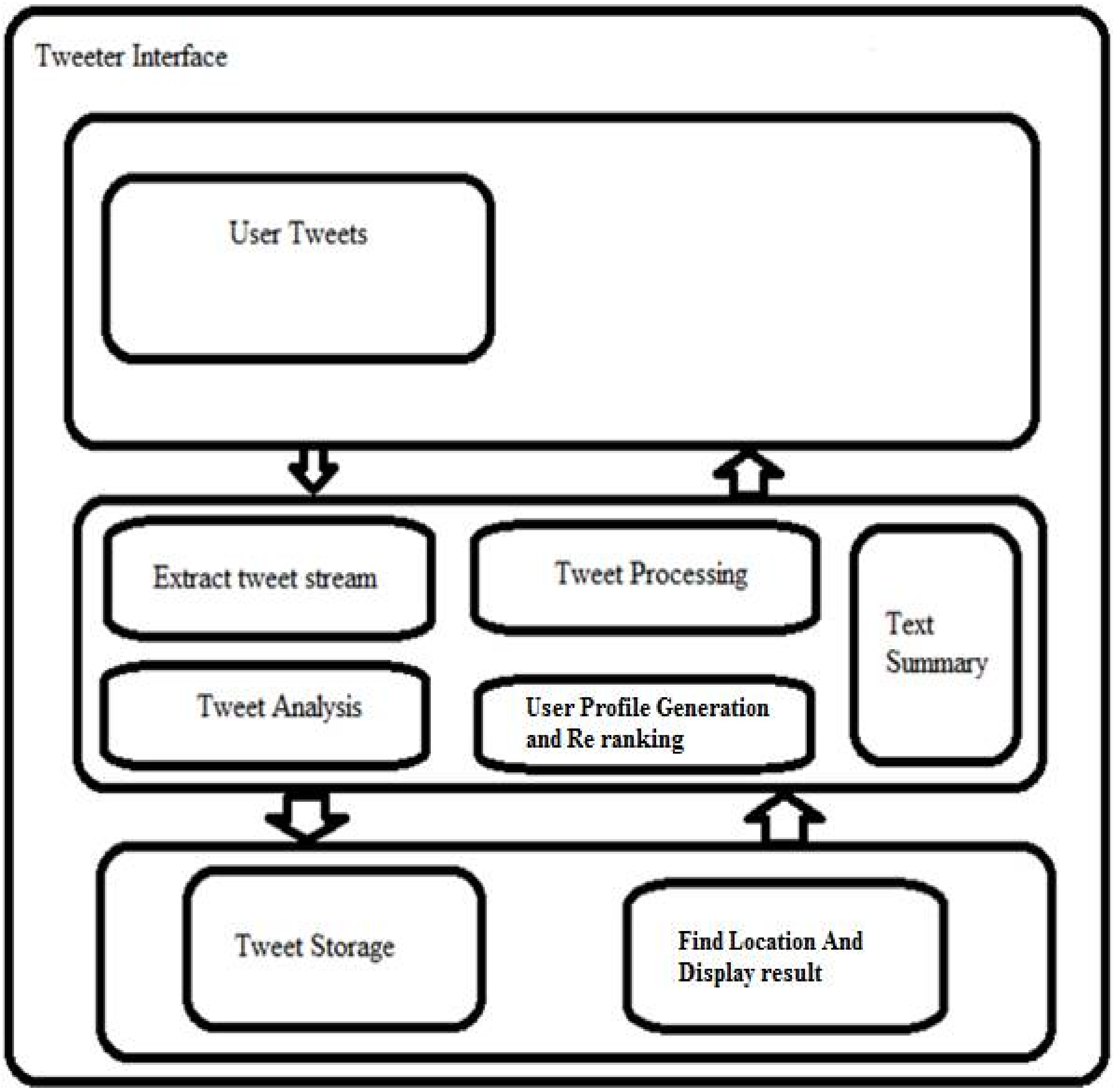


Fig No 02SYSTEM ARCHITECTURE

VI. SIMULATION RESULTS

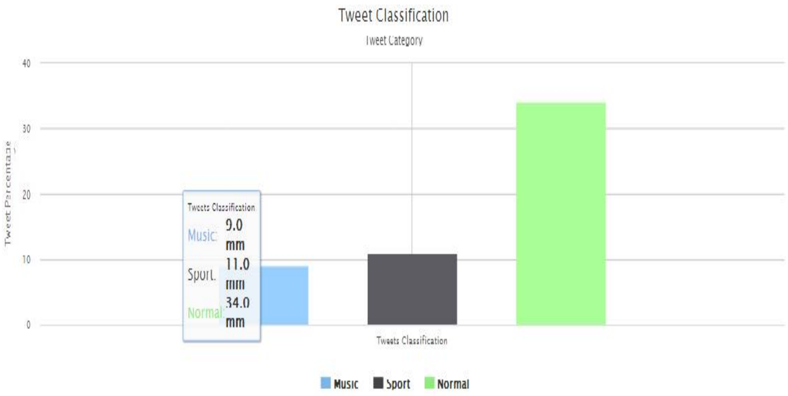


Fig.3.Tweet Classification

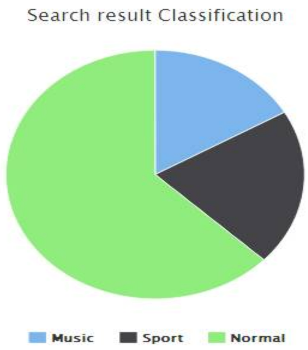
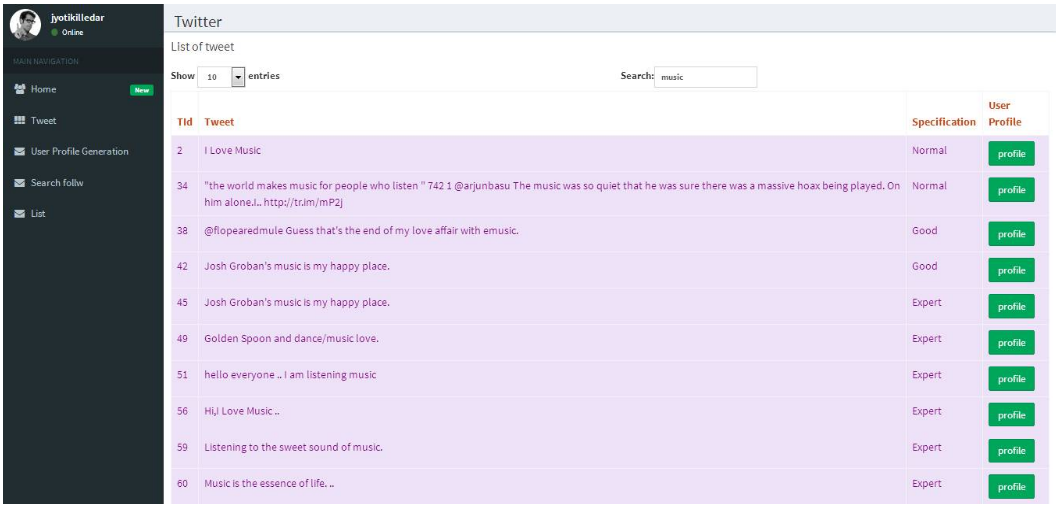


Fig.4. Search Result Classification



Twitter

List of tweet

Show 10 entries   Search: music

TId	Tweet	Specification	User Profile
2	I Love Music	Normal	profile
34	"the world makes music for people who listen " 742 1 @arjunbasu The music was so quiet that he was sure there was a massive hoax being played. On him alone.I.. http://tr.im/mP2j	Normal	profile
38	@flopearedmule Guess that's the end of my love affair with emusic.	Good	profile
42	Josh Groban's music is my happy place.	Good	profile
45	Josh Groban's music is my happy place.	Expert	profile
49	Golden Spoon and dance/music love.	Expert	profile
51	hello everyone .. I am listening music	Expert	profile
56	Hi,I Love Music ..	Expert	profile
59	Listening to the sweet sound of music.	Expert	profile
60	Music is the essence of life. ..	Expert	profile

Fig.5. Search Music Results

VII. CONCLUSION AND FUTURE WORK

In this project we have investigated semantic user modeling based on Twitter posts. We analyze method for segregating Twitter posts in order to contextualize Tweeting activities. Our system extracts tweets from a database and separates the user profiles based on their specific interests into different domains. The Domains are music and sports. Then we have ranked the user profiles and determined the experts of the domain based on those rankings. Ultimately we are finding experts according to geographical area or location.A large-scale evaluation validates the benefits of our approach and shows that our method relates tweets with high precision and coverage, enriching the semantics of tweets clearly. This helps the users to find experts specific to a domain based on location.

REFERENCES

[1] Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proc. of 19th Int. Conf. on World Wide Web, New York, NY, USA, ACM (2010)
[2] Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., Zheng, Z., Zha, H.: Time is of the essence: improving recency ranking using Twitter data. In: Proc. Of 19th Int. Conf. on World Wide Web. New York, NY, USA, ACM (2010)
[3] Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proc. of the 19th Int. Conf. on World Wide Web. New York, NY, USA, ACM (2010)
[4] E. Diaz-Aviles et al. What is Happening Right Now ... That Interests Me?: Online Topic Discovery and Recommendation in Twitter. In ACM CIKM, 2012.s
[5] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In ACM SIGIR, 2012.
[6] Passant, A., Laublet, P.: Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. In: Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW 2008), Beijing, China. (2008)
[7] Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and tweet: experiments on recommending content from information streams. In: Proc. of 28th Int. Conf. on Human Factors in Computing Systems. New York, NY, USA, ACM (2010)
[8] Jadhav, A., Purohit, H., Kapanipathi, P., Ananthram, P., Ranabahu, A., Nguyen, V., Mendes, P.N., Smith, A.G., Cooney, M., Sheth, A.: Twitris2.0 : Semantically empowered system for understanding perceptions from social data. In: Proc. Of the Int. Semantic Web Challenge. (2010)
[9] Mendes, P.N., Passant, A., Kapanipathi, P.: Twarql: tapping into the wisdom of the crowd. In: Proc. of the 6th International Conference on Semantic Systems. New York, NY, USA, ACM (2010)
[10] Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J.: Twitterstand: news in tweets. In: Proc. of 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York, NY, USA.

BIOGRAPHY

Jasmine Kanav, Jyoti Killedar, Rutuja Trimbake, Trupti Sapkale are students pursuing Bachelor of Computer Engineering from the Cummins College of Engineering for Women, Pune, Maharashtra, India.

Mrs. Aparna Hajare is an Assistant Professor in Computer Department at the Cummins college of engineering for women, Pune, Maharashtra ,India