11th International Conference on Recent Trends in Engineering, Science and Management
(IETE) Institution of Electronics and Telecommunication Engineers, Osmania University Campus, Hyderabad, India
22nd October 2017, www.conferenceworld.in (ICRTESM-17) ISBN: 978-93-86171-72-6

Prediction of User Topic Opinions based on Social Media Analysis

J.Sai Harshith Rao[1], K.Sriniketh Reddy [2] P.Rohith Reddy[3], S.G.Sanjeevi[4]

[1,2,3]Undergraduate Scholar ,[4] Professor, Dept. of CSE,

National Institute of Technology, Warangal (India)

ABSTRACT

Micro-blogs are common platforms for the people and organizations to create, share, or exchange information, experiences and ideas. Peoples conversations on social media portray their emotional status. Therefore, extensive research and analysis is performed on the data gathered from various sites, which includes user's social and contextual information. A Challenging and most essential task is to use the learned knowledge in the prediction of user's opinion towards specific topics which are not directly stated. Unlike previous works, we predict individual's topic opinions on specific topics which are not given directly, using obtained knowledge.

Keywords: Collaborative Filtering, Sentimental Analysis, ScTcMF Framework, SentiStrength, Theory of Homophily

I. INTRODUCTION

Various micro-blogging sites provides a common platform to discuss and share their everyday activities, in an informal and casual manner. In order to understand and analyse experiences of each individual, Researchers and practitioners can acquire large amount of absolute knowledge from these digital traces. This can be acquired with the help of machine learning analysis.

Twitter, Facebook, and You-Tube are the famous microblogging sites, where people from diverse nationalities, dialects share their experiences, feelings and can acquire knowledge of live activities. User topic opinions on social media after undergoing coherent tracing and logical analysis carries great value and can be used in different scenarios. Thus, its a prime necessity to detect and analyze user topic opinions.

In real-life applications, a very detailed analysis is required in order to acquire accurate information. Both public and private organizations are using this approach to fetch user data and analyze the corresponding user opinion on their products and services. The problem with direct opinion analysis is that every user will not be blogging about each and every topic. Therefore, the particular users opinion on a specific topic cannot be identified with this conventional method of analysis. Thus, this leads to study the theory of homophily. Therefore his Social relationship status provides base for the prediction of user topic opinion.

1.1 Homophily

The homophily theory states that relationships in social media are more probably created among users with identical characteristics [1]. In some former work [1],[2], validation of the homophilic effect in social networks was considered as a vital step. Bollen et al. proved the general happiness among Twitter users is assortative across the Twitter social network. Based on theory of homophily, an assumption that users and their followers on Twitter may coincide in their topical perspectives.

1.2 ScTcMF Framework

ScTcMF stands for Social relations context and topics context MF. In this users-topics prediction of opinions problem is defined as collaborative filtering task, and then finally the social relations context and topics context incorporating the factorization of matrix method is proposed which in turn include them as the regularizing constrains in order to achieve the goal.

The proposed frame-work is a general one and it can be applied without any difficulty for any other social relations network related environments or it can also be extended for incorporating the other regularizing constraints. We shall be defining the social relations network related information as social relations context and analyze the same for users-topics prediction of opinion problems.

A users-topics prediction of opinions problem has been proposed, for which, we understand and analyze social context which provides social network information. The data required is extracted from twitter and then sentimental analysis is done on the data in order to identify whether the user is having a positive or negative or neutral opinion on that topic.

II. DATA PROCESSING

As a part of data pre-processing , we need to extract the data from twitter and then it need to be translated into English as all tweets may not be in English and then finally the data is to be sent for sentimental analysis.

2.1 Data Collection

There are many tools and methods used to extract data from Twitter. We have used Twitter API to extract the data because Twitter provides data publicly for every developer. Analyzing the data and formulating the algorithm using machine learning analysis is the main step in our project. We went through various methods like extracting Tweets using R and tweepy python.

The problem with extraction of tweets using R is that this method can be employed only while we try to extract tweets based on certain hash tags. But in our project we need to extract tweets based on user-name and have to maintain the track of all the followers of the user profiles for which the tweets are extracted. Hence we moved on to tweepy python for the extraction of user profiles and then the tweets of respective users.

2.1.1 Extracting Tweets using Tweepy Python

It is very demanding and Challenging task to collect user data from various Social media. We have collected user data of 2,000 profiles from Twitter using Tweepy Python. The process of acquiring data was analytic. We have constructed a tree structure for each user profile. In the tree structure of a particular user, child nodes are his followers and the structure is iterated until a threshold is reached.

Twitter provides an API, which in turn provides access to entire Twitter restful API methods. We can provide one or more and can retrieve the data accordingly. The results from the called API methods will be returned as a class defined in Tweepy python which formats the JSON Format data into the data of our desired format.

After all the list of followers are collected, then the tweets of the respective users are to be collected. Furthermore, twitter API will provides a way to query for each user and the results are returned in JSON format and it can be parsed with a python script.

Some tips for writing cron job tasks that we found helpful during data collection are - building scripts in such a way so that it iterates through API keys to stay within the rate limit indicate and writing catch statement to handle exceptions that may occur while accessing Twitters API. This provision helps our program to run even when there is an error and can start execution once the prescribed limit is over.

2.2 Data Translation

Though we selected those users who set the language option in their profiles as en, we found there were a part of users posting tweets which are not in English in the data set. Therefore, we made use of Microsoft Translator Platform to filter the non-English content.

For the purpose of our convenience, we converted every tweet into English using Microsoft Azure Marketplace and Memsource. Memsource is a Translation Platform, a tool to convert a document in source language to Target language i.e. in our case the source language is the actual language in which the tweet is extracted and the target language is the English language.

2.3 Data Analysis

After translation of data, we have to analyze the kind of data and to what extent the words in the sentence lead to positive, negative or neutral opinions. For this, we need to do Sentimental Analysis of the Twitter data that is extracted. Sentiment Analysis is the process of determining whether a particular script or tweet is positive, negative or neutral. It is also known as the opinion extraction of the particular users opinion on. It is needed to estimate the strength of positive and negative sentiment in short texts, even for informal language. For this purpose, we used SentiStrength, a jar file with a few tweaks. Getting the collected tweets as input to the application. Building a java application using SentiStrength for the analysis. Feeding the input and collect the emotional strengths of tweets in an output file.

This is done by having a text file Emotion Look-up, in which there are numerical values assigned to the word with a range [-5,5] depending upon the strength of emotion. The application takes the collected tweets as an input file and maps them a value -1, 0 or 1 based on the sentiment of the tweet, to an output file. Here is a code snippet of getting the opinion strength through a java application.

From the results of sentimental analysis we can infer that if the results is -1 - if the negative strength is greater than positive strength of the statement 0 - if the negative strength is equal to positive strength of the statement 1 - if the positive strength is greater than negative strength of the statement

III. INCORPORATING SOCIAL AND TOPICAL CONTEXT FOR PREDICTING USER-TOPIC OPINION

Considering the predictive accuracy, scalability and flexibility for incorporating additional information, matrix factorization methods are extensively utilized in the state-of-the-art collaborative filtering tasks [3],[4],[5],[6].

3.1 Incorporating Social Context

Following mechanism provided by Twitter enables a user to follow other users and a social relation is established. A user who follows other users is called as their follower. A user who is followed by other user is called as the followers friend, no matter whether they follow back that particular user or not. Updated data of a user will appear in Home tab of the follower, who subscribes to the user tweets on Twitter. Unlike the case of celebrity following, creating a following relationship implies that the follower and the friend may have similar taste, so with higher probability, they would hold more similar opinions towards the same topic[7].

Now, an undirected weighted graph is constructed with a symmetric adjacency matrix S(denoting the social context)which defines the social context of Twitter users. Next, the social context hypothesis on Twitter will be described with high probability, the social friends hold more identical opinions on the topics than the non friends, which will be validated experimentally in forthcoming sections.It is because one tend to follow a person only if he suits his ideas and opinions.

As stated earlier, the value of S(i,j) implies the prior opinion similarity between social friends ui and uj. Firstly, the cosine similarity between the two corresponding row vectors O(i;:) and O(j;:) of the user-topic opinion matrix O is calculated, to record the divergence of social friends opinions towards different topics, and define it as User Opinion Similarity (UOS), thus eq(1):

Here the cosine similarity function is applied on the original opinion matrix to obtain the UOS(User Opinion Similarity) matrix: And the following equation is implemented:

$$\text{UOS(Ui,Uj)}= \frac{(\sum_{k=1}^{n} Oik.Ojk)}{(\sqrt{\sum_{k=1}^{n} Ojk.Ojk})(\sqrt{\sum_{k=1}^{n} Oik.Oik})}. \quad (1)$$

3.2 Incorporating Topical Context

Definite links are not established between topics in topical context unlike the social follower relationships between users in social context. However, research study on information retrieval and text mining explored content-based topic correlations and necessary refinements are made.

Thus, to establish these links we try to identify the topics which best represents a particular hash tag.Then with all the available topic features we try to construct a topical feature vectors.with thus obtained topical features vectors we apply cosine similarity and then the topical context similarity matrix is created[8].

$$\text{TCS(Ui,Uj)}= \frac{(\sum_{k=1}^{n} TFik.TFjk)}{(\sqrt{\sum_{k=1}^{n} TFjk.TFjk})(\sqrt{\sum_{k=1}^{n} TFik.TFik})}. \quad (2)$$

where tfi and tfj denote the term frequency vectors of ti and tj respectively, and N is the number of features in the vectors. In this definition, the similarity values range from 0 to 1, since the term frequencies cannot be negative.

IV. SCTCMF: PROPOSED FRAMEWORK WITH SOCIAL RELATIONS AND TOPICAL CONTEXT

The hypothesis is formulated as the social relations context and the topics context and then finally modelled the regularizing constraints along with them respectively. Also in this chapter, we shall also try to propose a most general framework in order to incorporated the both the social and also the topics context which is named as the ScTcMF.

$$Ł_U f(U_t,H_t) = -2(Y\,O_t)H_t + 2Y\,(U_t H_t^T)H_(t) + 2_1 U_t + 2L_s U_t. \quad (3)$$

$$Ł_H f(U_t,H_t) = -2(Y\,O_t)H_t + 2Y\,(U_t H_t^T)H_(t) + 2_1 U_t + 2L_s U_t. \quad (4)$$

Above functions U and H are the gradients in the t+1 step, dened as partial derivatives to U and H respectively.

Algorithm[17]: ScTcMF: The proposed framework with social and topical context.

Input: social context GS , Topical context Gt the set of user topic opinion labels OL

Output: The predicted user-topic opinion matrix$_O$

1: Initialise Uo randomly

2: Initialise Ho randomly

3: Construct the indicator matrix Y and Laplacian matrice LS and LT

4: While not convergent do:

5: Compute $Ł_U f(U_t,H_t)$

6: Compute $Ł_H f(U_t,H_t)$

7: Set $U_(t+1) = U_t - {}_U f(U_t,H_t)$

8: Set $H_(t+1) = H - {}_H f(U_t,H_t)$

9: end While

10: Set U = Ut+1

11: Set H = Ht+1

12: Compute $^O = UHT$

The complete procedure for ScTcMF is shown in above Algorithm. The lines 1 to 3, we initialize the required matrices for the algorithm. And then the lines 4 to 9, we shall update H and U up to the achievement of convergence along the direction of negative gradient to result in a matrix which includes the closest approximation of the opinions predicted.

V. IMPLEMENTATION AND RESULTS

The main contribution to our paper comes from the extraction of topic features by adding weights to the topics which best denotes the hash-tags i.e. topics.

1: Eliminate all the stop words from the tweets

2: Hash all the words individually

3: If there is any re-occurrence of a word already hashed then increment the weight of the word exponentially (2 in our case).

4: Continue similarly until all the tweets are hashed

5: Now fix a threshold for the topics and consider all topics whose word count is greater than the threshold

6: Include all those word in the topic feature vector of that particular topic

7: Repeat the procedure for all topics.

After extracting the topics we need to calculate the topic feature values for all topics respectively. That is we need to calculate how many times a particular topic is representing (present) among all the tweets relating to a particular hashtag. We can achieve this with the following algorithm:

Algorithm for topic feature selection:

1: Consider each hash-tag one by one and for each hash-tag do:

2: for all the tweets containing the particular hash-tag parse each tweet individually

3: for each word in the tweet do:

4: if that word is in the topic feature vector list then increment the count of that particular topic feature.

5: continue until the entire tweet is parsed

6: end if all the tweets of a particular hash-tag are over

7: continue until all the hash-tags are used

8: return the resultant vectors of all the topic features for each hash-tag individually.

The final result as measured with RMSE and the methods of Accuracy are displayed below:

Training Set	ScTcMF of Previous Algorithm	ScTcMF of our Algorithm
10%	0.9771	0.9654
20%	0.9645	0.9551
50%	0.9561	0.9678
80%	0.9513	0.9518
90%	0.9488	0.9513

Table: Comparison based on RMSE by Using the Different data

VI. CONCLUSION AND FUTURE WORK

6.1 Conclusion

The main idea of this project is to successfully identify the opinion of a particular user on a defined topic which he didn't give his opinion directly. The following are our contributions:

Differing from all related works previously i.e. recognizing of the sentiments or emotions on the topics but neglecting who is having that opinion, we try to identify that who has what opinion well in prior before they actually express any opinion directly which can be used in case of recommendations for that user.

In order to propose the solution, we shall try considering the homophily between the social relations (may be friends) in our case from twitter on opinion similarity topics and shall try to formulate mathematically both the social relations and topics contexts.

By making use of the trained emotional data sets from extracted tweets also from the social relations and topics contexts we proposed framework based on ScTcMF to successfully predict the indirect users-topics opinion. Finally, the implementation part id done in order to verify the framework which is proposed based on ScTcMF and the results shown that the social relations and topics contexts will helps us to improve performances for indirect users-topics predictions. Though we encountered few limitations we set them as our future work

6.2 Future Work

Though the predicted models are giving satisfactory results no model is absolutely accurate i.e. provides 100 percent accuracy hence there is always scope for advancement in this area. Some of the areas are:

We have considered relations between the social friends in our case twitter in order to model the social relations context. But the relationship strength may not be decided only by the topics similarity among the users, also the additional features like the vicinity of the users also when the users are being on-line etc.

We also try to explore in the other direction in identifying the target topics of the user with more accuracy. Sometime the hash-tags which are labelled by users will completely be under their assumption they can user their own methodology in labelling the topics i.e. one can give an abbreviation for the topic and the other can give the full name as there is no fixed standard in hash-tag labelling.

To successfully predict the multiple emotion/opinion states that the user is having on the topics also is an very interesting idea in order to analyze, in the coming future.

REFERENCES

[1] M. McPherson, L. Smith-Lovin, and J.M. Cook, Birds of a Feather: Homophily in Social Networks, Annual Rev. of Sociology, vol. 27, pp. 415-444, 2001.

[2] J. Tang, H. Gao, X. Hu, and H. Liu, Exploiting Homophily Effect for Trust Prediction, Proc. Sixth ACM Intl Conf. Web Search and Data Mining, 2013.

[3] H. Ma, H. Yang, M. Lyu, and I. King, SoRec: Social Recommendation Using Probabilistic Matrix Factorization, Proc. 17th ACM Conf. Information and Knowledge Management, pp. 931-940, 2008.

[4] Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181–184. IEEE Press, New York (2001)

[5] Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)

[6] Y. Koren, Factorization Meets the Neighborhood: A Multifac- eted Collaborative Filtering Model, Proc. 14th ACM SIGKDD Intl Conf. Knowledge Discovery and Data Mining, pp. 426-434, 2008.

[7] J. Tang, Y. Zhang, J. Sun, J. Rao, W. Yu, Y. Chen, and A. Fong, Quantitative Study of Individual Emotional States in Social Networks, IEEE Trans. Affective Computing, vol. 3, no. 2, pp. 132- 144, April-June 2012.