Linguistic Based and Location-Based Recommending Posts Using Various Clustering Methods

Prof. Pradhnya Mehta
Assistant Professor, Department of Computer Science Engineering, Marathwada Mitra Mandal's College of Engineering, Karvenagar, Pune, India

Mr. Ashish Saroj, Mr. Siddharth Sanas, Mr. Rajat Doijad,
B. E Students, Department of Computer Science Engineering, Marathwada Mitra Mandal's College of Engineering, Karvenagar, Pune, India

Abstract- Online social networks have produced bunches of online social groups where individuals can collaborate and shuffle their thoughts. In spite of, the real problems that conflicts with the user security and convenience are confidentiality break, groups without inception, confusion created out of various groups in which a user is a member of and difficulty in managing group regulations. This can be moderate to an extent by an automated filtering method required to categorizing members within a group based on their pattern of response. This paper proposes, the posts within a group are clustered based on stylistic, thematic, emotional, sentimental and psycholinguistic methods. After group members are categorized based on their response to the posts belonging to different clustering methods. The categorization provides security like the conflict associated with irrelevant notifications received from multiple groups, by recommending the users, posts that are likely to be of interest to them. It also helps to identify the group members intended towards spreading posts that violate group policies. The categorization posts shows increased performance in case of large number of candidate members in a famous group by performing clustering based on linguistic features. The contribution work is to implement location-aware personalized posts recommendation using both the users' personal interests and their geographical contexts. The system has been tested on Facebook group data where it offers a significant solution to an unaddressed problem associated with social networking groups.

Keywords- Emotion analysis, Multi-level clustering, Psycholinguistics, Sentiment analysis, Stylistics Clustering.

I. INTRODUCTION

The most popular social networking site Facebook has groups with over 100K members in it. Thus, it becomes difficult for the admin to track the members violating group policies. This shows that there exists a need for a measure to categorize the posts made by members in it based on acceptability and group behavior. A community is a fraternity that seeks a platform to discuss subjects that relate to the common cause that motivated the establishment of the group. The members within it enjoy discussion with regard to the purpose of the group. This is especially applicable in the case of academic and interest groups. Thus posting articles irrelevant to these groups can create unnecessary clutter that affects the comfort of the members within the group. Unnecessary advertisements and marketing matters targeting a different audience should be avoided from a group. There is no existing method to deal with this.

A method to control the influx of irrelevant messages within a community is very much essential for the smooth functioning of a social networking group. The existing policies of popular social networking site Facebook enable the administrator of a group to delete and monitor posts made by the members of a community. To screen all the messages posted by the members of a heavily populated group is very difficult. The existing settings provide no option for automated notification for group admin with regard to the members involved in posting articles that do not match the general interest of the group. Our proposed method aims at providing an automated notification to the group moderators regarding suspicious members who frequently post articles that appear to gather negative response from the members within the group. This novel approach has been experimented to account for the hypothesis that, though members may be socially connected; there might be disparity in their likings, thoughts, sentimental orientation and thematic inclination.

Motivation

To provide an automated notification to the group moderators regarding suspicious members who frequently post articles that appear to gather negative response from the members within the group.

II. RELATED WORK

The paper [1] proposes a new content based method for personalized tweet recommendation, based on conceptual relations between users' topics of interest. The Concept Graph is a way to exploit logical relations between topics of interest in order to provide interesting and efficient tweet recommendations. Advantages are: Provides a social media user with a new timeline that contains messages that strongly match ones interests and that are not necessarily posted by ones followings. This model is effective and efficient to recommend interesting tweets to users. Disadvantages are: The recommender still recommends some tweets that were not retweeted by the user.

The paper [2] proposes a TWIMER framework using language models as a basis for analyzing strategies and techniques for tweet recommendation based on user interest profiles. TWIMER consists of several components, including tweet retrieval (query formation and relevance model), tweet relevance verification, and final relevance ranking. Advantages are: Automatic query expansion. Higher performance in the language model retrieval entities.

In [3] paper, collaborative topic Poisson factorization (CTPF) can be used to build recommender systems by learning from reader histories and content to recommend personalized articles of interest. CTPF models both reader behavior and article texts with Poisson distributions, connecting the latent topics that represent the texts with the latent preferences that represent the readers. Advantages are: CTPF performs well in the face of massive, sparse, and long-tailed data. CTPF provides a natural mechanism to solve the "cold start" problem. CTPF scales more easily and provides significantly better recommendations than CTR.

In [4] paper, represents the problem of simultaneously predicting user decisions and modeling users' interests in social media by analyzing rich information gathered from Twitter. Proposes Co-Factorization Machines (CoFM), which deal with two (multiple) aspects of the dataset where each aspect is a separate FM. This type of model can easily predict user decisions while modeling user interests through content at the same time. Advantages are: CoFM can easily predict user decisions while modeling user interests through content at the same time. Factorization Machines to text data with constraints can mimic state-of-the art topic models and yet benefit from the efficiency of a simpler form of modeling. Disadvantages are: The services are only interacting with the user's interest not on user's behaviors.

The paper [5] focuses on recommending useful tweets that users are really interested in personally to reduce the users' effort to find useful information. The topic level latent factors of tweets to capture users' common interests over tweet content, which helps us to solve the problem of information sparsity in users' retweet actions. This allows us to adjust the collaborative filtering technique to solve the recommendation problem. Advantages are: A collaborative ranking method is better than collaborative filtering for different optimization criterion. The proposed CTR method greatly improves the recommendation performance. The CTR method is generic; it is easy to incorporate more information by adding extra features. Disadvantages are: It only works on user's interests over time not on user's history and tags of the tweet.

III. OPEN ISSUES

A method to control the influx of irrelevant messages within a community is very much essential for the smooth functioning of a social networking group. The existing policies of popular social networking site Facebook enable the administrator of a group to delete and monitor posts made by the members of a community. To screen all the messages posted by the members of a heavily populated group is very difficult. The existing settings provide no option for automated notification for group admin with regard to the members involved in posting articles that do not match the general interest of the group.

Disadvantages are:

1. Does not solve the clutter issues and group policy management hassles associated with social networking groups.
2. None of the methods deal with the issue of providing customized suggestions to posts in a community to reduce clutter associated with notifications generated by large groups.

3. Does not provide automated notification for group admin with regard to the members involved in posting articles.
4. Need to categorize the posts made by members in it based on acceptability and group behavior.

IV. SYSTEM OVERVIEW

The proposed method puts forward a filtering mechanism that enables a member of a social network community to get notified about posts that can be interesting to them. The method groups the members into different clusters based on their aspect-based characterization. The members belonging to the same clusters are those who exhibit resemblances with respect to sentimental, thematic, emotional, writing style and concept-related interests.

There are 5 types of clustering methods:

1) Sentiment Clustering

The posts in a network group vary based on the sentimental status associated with the contents. The overall sentiment associated with a post can be positive, negative or neutral. The assessment of the sentiment is made at the keyword level. This enables to classify the posts based on the attitude towards the trends. The response of a member to a post shows his/her attitude towards the entities. This can be illustrated by an example below:

- User A says "I love ice cream. I Love to have it every time"
- While User B says "I hate ice cream. I don't want it at all"

Here the keyword is ice-cream and the sentiment of the keyword with respect to user A is positive while that of user B is negative. Thus, analyze the sentiments associated with prominent keywords associated with a group.

2) Theme Clustering

The next type of clustering performed on the posts is the theme based clustering. Here the posts are grouped based on the thematic similarities. After consider concepts, entities and topics for determining the theme. Use the semantically rich common sense knowledge base ConceptNet for extracting contextual and conceptual information of a post. The ConceptNet toolkit provides numerous assertions related to a word [1]. Utilize them to arrive at a general concept of a post. Initially the posts are chunked to obtain keywords and key phrases. These phrases are fed to the ConceptNet to obtain generalized concepts. After exploit the analogy making and topic gisting features of ConceptNet to arrive at a conclusion regarding the important concepts of a post.

3) Emotional Clustering

Emotion based aspects can be investigated to gain insight into a personś emotional attitude. The similarity in these aspects can be exploited to predict the liking and sharing probability of members within a community. Thus, the users within a community are grouped together based on the emotional aspects. The clustering based on emotion is performed by taking into account the score of the entire post with respect to the emotional categories namely anger, sadness, fear, disgust and joy.

4) Stylistic Clustering

Stylistics refers to the writing-style followed in a document. Each person has a unique writing style of his or her own. The writing-style factor has been considered because the writing style followed determines the popularity of an article. There are phrases and usages peculiar to authors that can be of interest to the audience. This aspect has been exploited to find the like-minded audience of a particular style of writing. The stylistic features such as words and character n-grams can capture the writing style effectively. These features are clustered by using K-means clustering.

5) Psycholinguistic Clustering

The posts have been subjected to the analysis of psycholinguistic orientation of the posts. The psycholinguistic differences contribute to the nature of the posts and hence there will be difference in the audience based on the psycholinguistic perspective of members within a group. Psycholinguistic aspects throw light on various factors that vary based on a person’s personality, hobbies, passions, intellects, perception and context of references. Thus it reflects a person’s way of responding to a post.

The contribution of the proposed method can be summed up as follows:

1. The method proposes a two-level clustering mechanism that aims at categorizing the members of a group based on interest.
2. The method is extended to provide features for group policy management.
3. The contribution work is to implement location-aware personalized posts recommendation using both the users' personal interests and their geographical contexts.
4. The experiments have been conducted on limited content available for processing which makes the experiment standout from the state-of-the-art techniques that rely on lengthy text for processing.
5. The method is scalable in huge data-set.

A. Architecture:

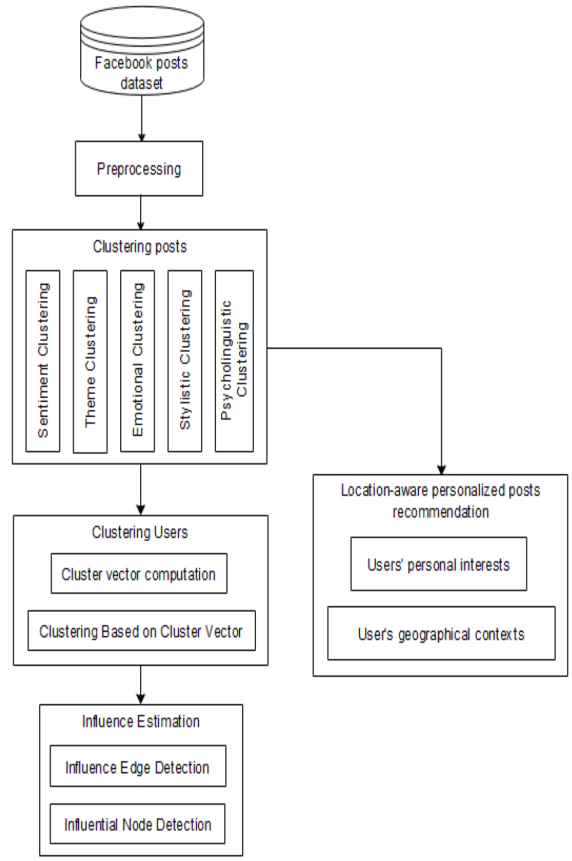

Fig. 1 Proposed System Architecture

Advantages are:

1. The approach is independent of the pre-existing social relations for members within a group and the approach considers behavioral similarity between the users in terms of their response pattern towards different aspects.
2. The approach proves to be beneficial in tackling “cold start problem”, i.e. the cases where a new post has not been rated yet.
3. Like minded members within a group are obtained by considering similarities in various dimensions.
4. The first work based on recommending posts relevant to a user based on their interest

expressed by them to the content of the posts, within a social network group and also one that helps in identifying members aimed at spreading unpopular posts within a group.

B. Mathematical Model

1. Clustering Based On Cluster Vector:

To check the validity of the similar users formed from initial clusters, the members within a cluster are made to undergo similarity check. This is to check the confidence values of the users. It is measured using Pearson´s correlation coefficient, which can be computed as follows:

$$sim_{m,n} = \frac{\sum_{p\in P}(r_{mp} - \overline{r_m})(\gamma_{np} - \overline{r_n})}{\sqrt{\sum_{p\in P}(r_{mp} - \overline{r_m})^2}\sqrt{\sum_{p\in P}(r_{np} - \overline{r_n})^2}} \quad (1)$$

Where m and n are users whose similarity is to be identified. P refers to the set of posts which are liked or shared by at least one of them. r_{mp} refers to the posts liked by m and r_m refers to the average rating of user m.

2. Group Policy Management:

Acceptability:

This term is defined in association with the posts. A post is checked for its acceptability with respect to how the different aspects (sentimental, psycholinguistic, stylistic, emotional and thematic) of the post are liked by the members of the community. It shows an aspect based score.

$$Acceptability(p) = \sum_{M} \sum_{i\in clus} val(C_{mi}) \quad (2)$$

Where, clus is a list that stores the cluster of the post p with respect to different aspects. $val(C_{mi})$ gives the vector values (0 or 1) associated with the cluster vector of member m with respect to the aspect i.

Popularity: The term popularity is defined for both posts and members within a community.

$$\begin{aligned} popularity(p) &= Acceptability(p) + share(p) \\ &+ (com_{pos} + likes_{pos}) \\ &- (com_{neg}(p) + likes_{neg}) \end{aligned} \quad (3)$$

Where $share(p)$ is the number of shares for the post p, com_{pos} is the number of positive comments for p, $likes_{pos}$ is the like obtained for positive comments, com_{neg} is the number of negative comments for p and $likes_{neg}$ is the negative comments obtained for p.

C. Algorithms

1. Sentiment Analysis using Sentiwordnet Dictionary

polarizedTokensList ← newList()
while tokenizedTicket.hasNext() **do**
token←tokenizedTicket.next()
lemma←token.lemma
polarityScore←null
if DomainDictionary.contains(lemma,pos) **then**
if SentiWordNet.contains(lemma,pos) **and**
SentiWordNet.getPolarity(lemma,pos) != 0) **then**
polarityScore ← SentiWordNet.getPolarity(lemma, pos)

else

domainDicToken←DomainDictionary.getToken(lemma, pos)

if domainDicToken.PolarityOrientation == ”POSITIVE” **then**

polarityScore ← DefaultPolarity.positive

else

polarityScore ← DefaultPolarity.negative

end if

end if

polarizedTokensList.add(token, polarityScore)

end if

end while

return polarizedTokensList

2. Latent Dirichlet Allocation (LDA) Algorithm:

First and major, LDA offers a generative model that describes how the files in a dataset were created. In this context, a dataset is a group of D files. Document is a set of phrases. So our generative version describes how each document obtains its phrases. Initially, permit’s anticipate that recognize K subject matter distributions for our dataset, meaning K multinomials containing V elements each, where V is the wide variety of terms in our corpus. Let βi represent the multinomial for the ith topic, where the size of βi is V: |βi|=V. Given these distributions, the LDA generative method is as follows:

Steps:

1. For every document:

(a) Randomly choose a distribution over subjects (a multinomial of length K)

(b) for each word within the document:

(i) Probabilistically draw one of the K subjects from the distribution over topics obtained in (a), say topic βj

(ii) Probabilistically draw one of the V words from βj

3. Cluster Vector Formulation of Users Algorithm

1: $M \leftarrow$ {members in a group}

2: $A \leftarrow$ {sentiment, theme, emotion, stylistics, psycholinguistic}

4: $M\_like[m] \leftarrow$ posts liked by member m

5: $Cluster[p][a] \leftarrow$ cluster of post p for aspect a

6: for $p \in M$ do

7: for $a \in A$ do

8: $x_i = 0$

9: for $p \in M\_like[m]$ do

10: $c \leftarrow cluster[p][a]$

11: $m\_cluster\_a[x_c] += 1$

12: $cv[a] \leftarrow argmax_i(m\_cluster\_a[x_i])$

13: for $i := 1\ to\ n(m\_cluster\_a)$ do

14: if $i = cv[a]$ then

15: $x_i = 1$

16: else

17: $x_i = 0$

18: $merge(m\_cluster\_a)$

V. RESULT AND DISCUSSIONS

Experiments are done by a personal computer with a configuration: Intel (R) Core (TM) i3-2120 CPU @ 3.30GHz, 4GB memory, Windows 7, MySQL 5.1 backend database and jdk 1.8. The application is web application used tool for design code in Eclipse and execute on Tomcat server. The real time tweet posts collection for dataset of this application using Twitter API with the help of Twitter4j-core and Twitter4j-stream jars. Some functions used in the algorithm are provided by list of jars like standfordcore-nlp jar for POS tagging etc.

Proposed work is expected to implement posts recommendation system which collects input dataset of list of posts from Twitter API. Apply all the clustering methods like, sentiment clustering, theme clustering, emotional clustering, stylistic clustering and psycholinguistic clustering posts to provide clutter free group environment. Expected outcome of this project is providing clutter free group environment to Twitter users. With the help of various aspects based clustering and the location-based clustering the posts with the help of personalized notifications. Finally, formulating groups of individuals within a group sharing similar interests. Fig. 2 represents number of clusters generated for each aspect. The Fig. 3 shows performance on combination of features for number of clusters effectively gives notifications to the users.

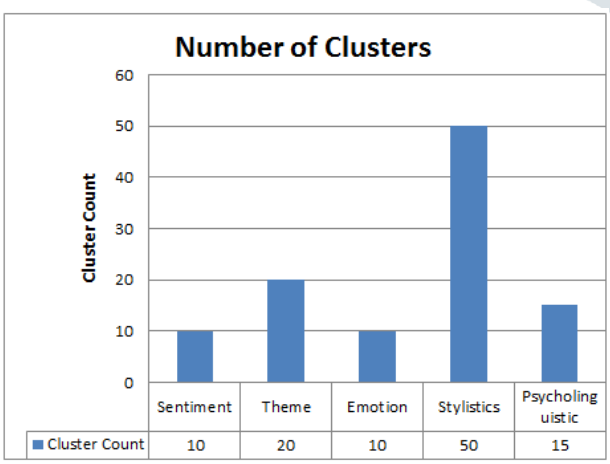

Fig. 2 Performance of Twitter posts of number of clusters for each aspect

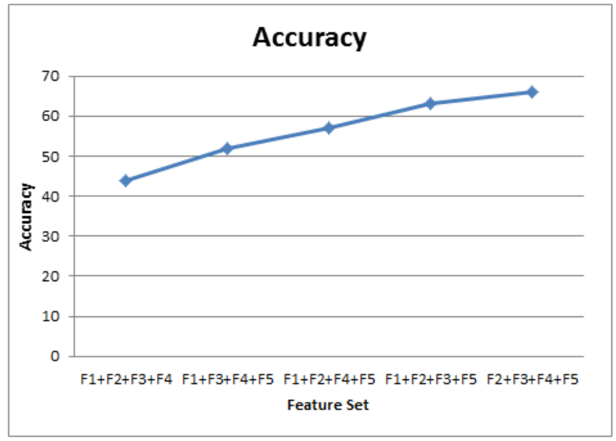

Fig. 3 Accuracy associated with clustering users by combination of four features

VI. CONCLUSION

The paper discusses a novel solution for the least addressed problem of clutter created out of group messages in online social networking sites. The method follows a different approach by considering linguistic features of data that are readily available from a social networking group. It does not depend on any of the pre-existing social relations between users unlike customary methods. The method shows a considerable degree of accuracy in predicting the response of a member to a post. The methodology proves to be an efficient means for managing group policies by providing a trusty environment. It offers a means to provide notification to group admin regarding activities of members within a community whose posting patterns do not suit the group principle.

REFERENCES

[1] Athira, U. and Sabu,M Thampi, “Linguistic Feature Based Filtering Mechanism for Recommending Posts in a Social Networking Group”, IEEE Access, pp:4470-4484 (2018) (SCI Impact Factor 3.244)

[2] D. P. Karidi, Y. Stavrakas, Y. Vassiliou, ”A Personalized Tweet Recommendation Approach Based on Concept Graphs”, In Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016,pp. 253–260

[3] R. Makki, A. J. Soto, S. Brooks, E. E. Milios, ”Twitter Message Recommendation Based on User Interest Profiles”, In Advances in Social Networks Analysis and Mining (ASONAM),IEEE/ACM International Conference, 2016,pp. 406–410

[4] P. K. Gopalan, L. Charlin, D. Blei, ”Content-based recommendations with Poisson factorization”, In Advances in Neural Information Processing Systems, 2014, pp. 3176–3184

[5] L. Hong, A. S. Doumith, B. D. Davison, ”Co-factorization machines: modeling user interests and predicting individual decisions in twitter” ,In Proceedings of the sixth ACM international conference on Web search and data mining, 2013, pp. 557–566

[6] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, Y. Yu ,”Collaborative Personalized Tweet Recommendation”, In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM, 2012, pp. 661–670.