Mining Current Topics Based On Client Elucidation

[1]Bavithra.N [2]Rajesh.P
[1]Student [2]Assistant Professor,
Kingston Engineering College, Vellore, India

***Abstract* - The paper provides the overview of essential analyses and methods, helpful for enterprise architecture improvement and based on social network approach. Social network is a place where people exchange and share information related to the current events all over the world .This behavior of users made us focus on this logic that processing these contents may lead us to the extraction the current topic of interest between the users. Applying data clustering technique like post Text-Frequency-based approach over these content may leads us up to the mark but there will be some chance of false positives. We propose a probability model that can capture both normal mentioning behavior of a user and also the frequency of users occurring in their mentions. It also works good even the contents of the messages are non-textual information like images etc .The proposed mention-anomaly based approaches can detect new topics at least as early as text-anomaly based approaches, and in some cases much earlier when the topic is poorly identified by the textual contents in the posts**.

***Keywords* - change point detection, anomaly scores, mentions**

I. INTRODUCTION

As in this internet world every one used to engage in social media is very familiar now days. Social media acts fast with the contents than any other media. Lot of contents in many format been scattered in the database were we can look forward to utilize those contents to build an automated news event. Since the information exchanged over social networks is not only texts but also URLs, images, and videos, they are challenging for the study of data mining. The interest is in the problem of detecting emerging topics from social streams. This can be used to create automated "breaking news", or discover hidden market needs or underground political movements. Compared to other media (news FM etc.) social media are able to capture the earliest, unedited voice of ordinary people. Problem is the challenge is to detect the emergence of a topic as early as possible at a moderate number of false positives.

Nowadays, in the times of strong competition, business organizations constantly look for tools and techniques to beat market opponents and become leaders among other companies. This paper focuses on corporate social network analysis as a possible way to improve enterprise architecture leading to the above mentioned goals. However, the proposed methods are suitable for all kind of organizations with the stable organizational structure, not only the commercial ones.

The interest in detecting emerging topics from social network streams based on monitoring the mentioning behavior of users (annotation like). Our basic assumption is that a new (emerging) topic is something people feel like discussing, commenting, or forwarding the information further to their friends. Conventional approaches for topic detection have mainly been concerned with the frequencies of (textual) words. A term-frequency-based approach could suffer from the ambiguity caused by synonyms or homonyms. It may also require complicated preprocessing (e.g., segmentation) depending on the target language. Moreover, it cannot be applied when the contents of the messages are mostly non textual information. On the other hand, the "words" formed by mentions are unique, require little preprocessing to obtain (the information is often separated from the contents), and are available regardless of the nature of the contents. Probability model that can capture the normal mentioning behavior of a user, which consists of both the number of mentions per post and the frequency of users occurring in the mentions. This model is used to measure the anomaly of future user behavior. Using the proposed probability model, we can quantitatively measure the novelty or possible impact of a post reflected in the mentioning behavior of the user. A term-frequency-based approach mainly depends upon the frequencies of (textual) words occurring in the social posts. This removes the verbal and adjective like words and considers only the nonverbal parts of the post. Word frequency is calculated for each word which will be taken mainly for extraction of the topic. The limitation is that a term-frequency-based approach could suffer from the ambiguity caused by synonyms or homonyms (plurals).It cannot be applied when the contents of the messages are mostly non-textual information. For e.g. "good life depends on liver", where liver may be organ or living person, so there will be an ambiguity problem. We cannot apply the technique when the content is non-textual information.

II.IMPLEMENTATION DETAILS

2.1 PROBABILITY DISTRIBUTION

Relevancy Probability Model

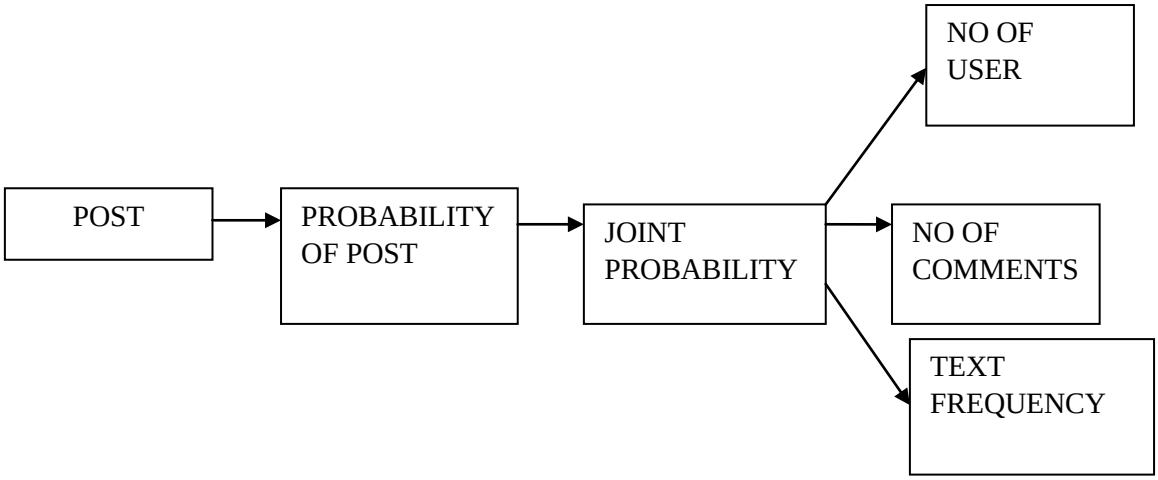

Fig.2.1 Relevancy probability model

We characterize a post in a social network by the number of comments it contains, and the set of users who are mentioned in the post. We also include the document frequency into our probability model which will enhance the detection process. In place of joint distribution we find conditional joint probability which consists of two parts: the probability of the number of comment/ mentions and the probability of number of mentioned, both based on term (word) frequency. The smaller the term frequency value smaller will be the probability of mentions and mentionees.

1: Find probability of number of mentions p (l|θ)

2: p (l|θ) = (1-θ) ^{s} (1)

3: Joint probability distribution of number of mentions, user and text frequency

P(s, v|θ, {π_v}) =p (l|θ) $\prod_{v \in V} \pi_v$, (2)

4: predictive distribution by using training set T= {(L_1, V1), (L_n, Vn)} **P (L, V|T) =p (L|T)** $\prod_{v \in V} P(V|T)$

2.2 TEXT FREQUENCY

Text frequency is used to find the similarity or relationship between each comments/mentions. This will also help us to identify the deviation of the comment in accurate manner. We create a text frequency function that generates a score based on comment relevance. Text frequency initial set up need a fixed dictionary word for which the frequency need to be calculated. After that the function search for the word in each comment in the post and assigns weights based on their occurrence in comment. These weights are used to generate the frequency score and it will be added with the anomaly score before emerging topic classification.

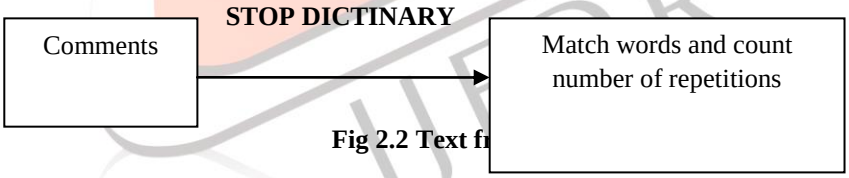

Fig 2.2 Text fr

Input: Set of Mention Content ($Cpost_{)}$ belong to Post P_i
Dictionary Word (D)
Output: Mention_Vectpr ($V_{post)}$
Start
String finalContent C_F
GetDictionaryWord ( )
Foreach word Wpost in C_{post}
Foreach W_{dic} in D
// where W_{dic} is element of Dictionary Word
If equals ($W_{dic,}$ Wpost) Then
concat($C_{F,}$ Wpost)

end If
end foreach
end foreach
arraylist Mention_vector V_{post}
foreach W_F in C_F
// where W_F is element of finalContent C_F
If Mention_vector.Contains(W_F) then

Mention_vector[W_F] = Mention_vector[W_F]+1
End if
Else then
Mention_vector[W_F] = 1
End else
End foreach
Return Mention_vector End

2.3 DERIVE ANOMALY SCORE

We calculate the link anomaly score for each post independently. Anomaly score is defined as the user's deviation from the post. If they pass the comments which is unrelated to the post called as the anomaly. The comments are either good or bad whether related to the post are determined by using link anomaly score. Accordingly, the link-anomaly score is defined by the following diagram.

1: Compute anomaly score of a new post x=(t,u,l,v)
L-mention-user-user-time
2: Find s(x)

$$s(x) = -\log(p(l|T_u^{(t)}) \prod_{v \in V} P(v|T_u^{(t)})$$

$$= -\log(p(l|T_u^{(t)}) \sum_{v \in V} \log P(V|T_u^{(t)}) \quad (3)$$

3:By using training set which consist of both number of user and mention compute anomaly score.
4: Finally we aggregate the anomaly score obtained for the post.
By using the joint probability distribution (conditional probability) both the text frequency as well as anomaly scores should be considered so that frequency of the user expectation about the topic can be easily mined and extracted from the comments.

2.4. CHECK-POINT RECOGNITION

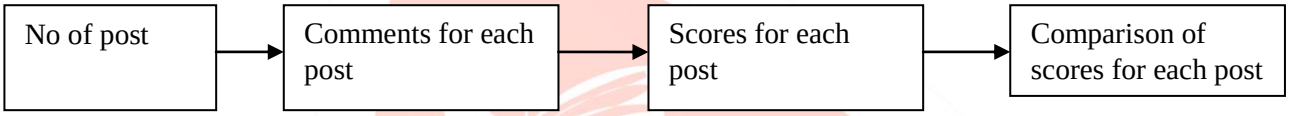

Fig 2.3 Check point detection

Change point act as the median based on the score value obtained. It is used to finds a change in the numerical trust construction of a time series by monitoring the compressibility of a innovative piece of data. It uses a sequential version of normalized maximum-likelihood (NML) coding called SDNML coding. A change point is detected through two layers of scoring processes. The first layer finds anomaly and the second layer finds change-points.

The issues of anomaly detection and change point detection from a data stream. In the area of data mining, there has been increased interest in these issues since the former is related to fraud detection, rare event discovery, etc., while the latter is related to event/trend change detection, activity monitoring, etc. Specifically, it is important to consider the situation where the data source is non-stationary, since the nature of data source may change over time in real application. The change point detection mainly states that if comments are passed only by few friends, i.e if the number of users is less, but the comments passed are more then we can say that it is a discussion but we cannot extract current emerging topic. So we can easily detect the outliers and find the change points.

2.5. (DTO) DYNAMIC THRESHOLD OPTIMIZATION

Finally we need to convert the change-point scores into binary alarms by thresholding as t.Maximum threshold value is 1.Change point detection act as a median depending on the score value for each and individual post. Binary alarm means a binary demonstration of true and false statement for the emerging topic. Since the distribution of change-point scores may change over time, we need to dynamically adjust the threshold to analyze a sequence over a long period of time. If the comments get added then the change point score gets varied. So Based on the generated score of each topic binary alarm differentiate the emerging topics.

III. SYSTEM ARCHITECTURE

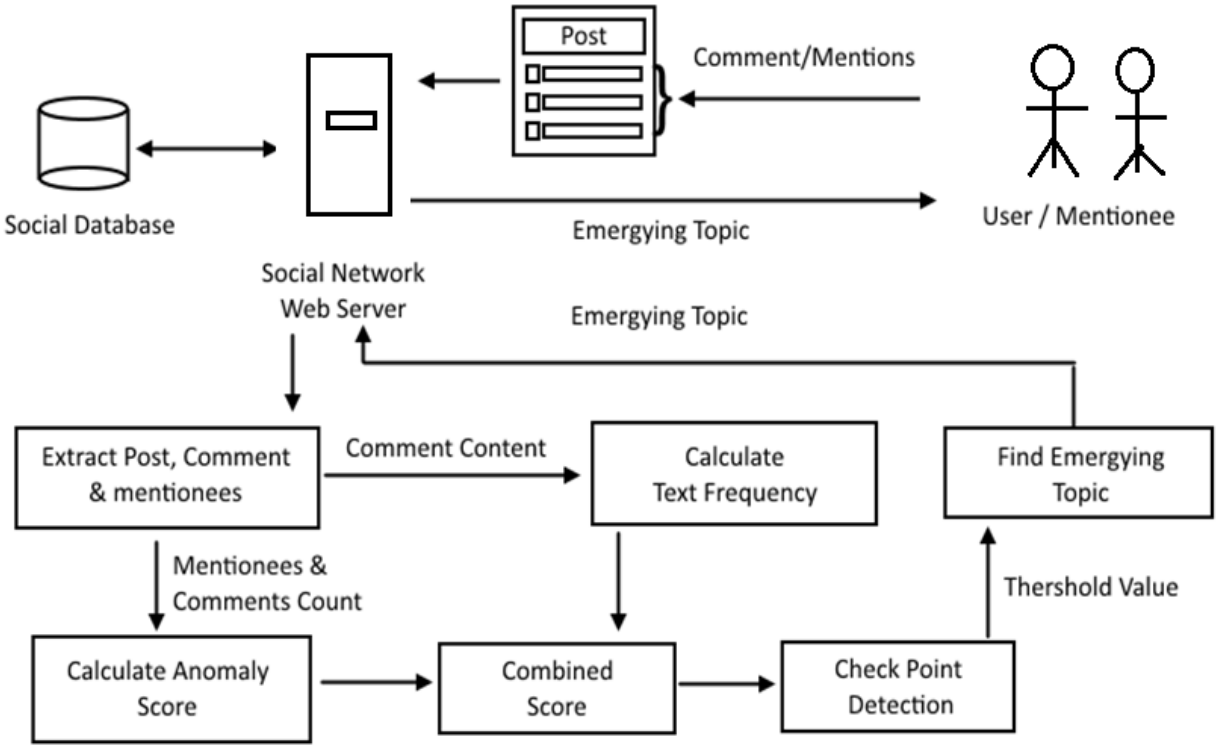

Fig3.1 Architecture diagram

The system architecture diagram explains that initially user tags the different post and friends pass comments for the post. Social network web server extracts the post, comments and mentionees from the social database. Then process the comments by calculating the anomaly score and then the text frequency for each and every post. Anomaly score determines the users deviations from the post, For calculating text frequency initially we need to setup the dictionary word, and we need to match the dictionary word with the comments, if the word gets repeated then we need to calculate text frequency for each and every word, then the anomaly score and the text frequency are combined, check point detection act as a median for the score obtained, and from the score we can find emerging topic from the different post. Social network web server will extract emerging topic to the user.

IV. RESULTS AND DISCUSSIONS

As discussed earlier, we can extract the emerging topic based on the user mention and by using the text frequency. The score (anomaly score) generated for both the text frequency as well as the user mention we can extract the correct emerging topic. Change point detection act as a median for different post. We can practically implement this project in multiple systems by using IIS (Internet information service) which can change the system into server which can be accessed by the client system.

V. CONCLUSION

We are interested in detecting emerging topics from social network stream based on monitoring the mentioning behavior of users. A term-frequency-based approach could suffer from the ambiguity caused by synonyms or homonyms. It may also require complicated preprocessing (e.g., segmentation) depending on the target language. The planned probability model determines both number of mentions per post and the frequency of the mentionee and this approach is used to find the emergence of topics in a social network stream .We have put forward a probability model that captures both the number of mentions per post and frequency of mentioning . The text frequency based methods used to determine how many times the text gets repeated and from that the repeated words are considered We combined the proposed mentioned model with the SDNML change point detection algorithm to pin point the emergence topic, the link anomaly based approach have detected emergence of the topic even earlier than the keyword based approach that use hand chosen keywords. It will be more effective when combining both text anomalies based and link anomaly based approach. We can also execute the process in multiple systems by using Internet Information services which is used to change the system into server.

REFERENCES

[1] Amandeepkaurmann,Navneenkaur'Survey paper on clustering techniques', international journal of science ,engineering and technology research ,volume2,issue 4 (2013).

[2] Anoopkumarjain, SatyamMaheswari'Survey of recent clustering techniques in data mining 'International journal of computer science and management research ISSN 2278-733x volume 1 issue -1 (2012).

[3] Adrian Gepps , J. Holton Wilson ,Kildee Kumar , Squanto Bhattacharya 'A Comparative Analysis of Decision Trees Vis-_a-vies Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection 'Journal of Data Science, volume 2,issue 4 (2012).

[4] A.DineshKumar, Dr.V.Radhika'A Survey on Predicting Student Performance' International Journal of Computer Science and Information Technologies, Volume 5 (5), 6147-6149, (2014).

[5] Dr.M.Hemalatha,N.Nagasaranya'A recent survey on knowledge discovery in spatial data mining'IJCSI International journal of computer scienceissues,Volumme 8,Issue 3,No.2 (2011).
[6] Edgar Moyotl-Hern_andez, H_ectorJim_enez-Salazar'An Analysis on Frequency of Terms for Text Categorization'international journal of computer science application,volume2 issue5(2002).
[7] Irena PletikosaCvijikj, Florien Michahelles' Monitoring Trends on Facebook 'Ninth IEEE International Conference on Dependable, Autonomic volume2 issue 5 (2011).
[8] JuhaVesanto, EsaAlhoniemi 'Clustering of self organizing map', IEEE transaction on neural networks, volume11,n0.3 (2000).
[9] K.V.Nagendra, C. Rajendra 'Customer behavior Analysis using CBA', national conference on research trends in computer science and technology (2012).
[10] Minghuang Chen, Seiji Yamada, YasufumiTakama'Investigating User Behavior in Document Similarity Judgment for Interactive Clustering-based Search engines' journal of emerging technologies in web intelligence, volume. 3, no. 1 (2011).

[11] PrzemysławKazienko, RadosławMichalski, Sebastian 'Social Network Analysis as a Tool for Improving Enterprise Architecture'Palus, proceedings of 5th international conference on agent, multiagent system (2011).
[12] Rajangupta, Nasibsingh gill's Data Mining Framework for Prevention and Detection of Financial Statement Fraud 'International Journal of Computer Applications (0975 – 8887) Volume 50 – No.8, July (2012).
[13] RajeshV.Argiddi,S.S.Apte'Future trend predication of Indian IT stock market using association rule mining of transaction data' International journal of computer science and management research 0975-8887,volume-39-No-10 (2012).