FRAMEWORK FOR CLUSTERING ON SOCIAL NETWORKING MESSAGE

Satish Babu M
M.Tech Student Computer Science,
Rajiv Gandhi Institute of Technology, Bangalore-32, India,
satish.babu20133@gmail.com

Geetha Pawar
Assistant Professor Computer Science,
Rajiv Gandhi Institute of Technology, Bangalore-32, India,
geethamnayak@gmail.com

Abstract - Normally, the messages collected from the social networking applications are highly unstructured and cluttered that pose a significant level of difficulty for applying conventional mining technique. Is such data are not filtered effectively; it will pose a greater deal of difficulty for the user if s/he wants to generate a customized level of query. This problem can be solved using clustering technique. Hence, we proposed a technique that is capable of visualizing the exact data representation of messages by applying a unique semantics operation followed by clustering. The study outcome of the technique is found to have better f1-score and accuracy in comparison to existing system of clustering social networking data.

***Keywords*: Social Networking message, Clustering, Accuracy, Recall, Semantics.**

1. Introduction

With the advancement of the mobile communication system, there is a significant revolution in the social networking applications too [1]. Various forms of social networking applications allows multiple people to form a specific group in order to share their views on a specific topic. Such views are usually in the form of text, which are sometimes highly cluttered one. Various social networking applications like Facebook, Twitter, and MySpace etc can be accessed from both desktop machine as well as smartphones. However, there are various other social networking applications which are equally gaining popularity e.g. whatsapp, viber, etc. Along with image, video, and audio, the higher percentage of data are mainly in the form of text. For the purpose of communication, the users make use of their own personalized dialects in English, which doesn't matches with lexical database many times [2]. Moreover, such data are sometime so much unstructured that it is not possible to perform datamining techniques on the top of it. In such application, usually, the communicated and transacted data are all gathered in one place which is a mixed result of discussion laid by multiple communicating user. Normally, 90% of the mixed information that gathered in such places are not much valuable of user, which posses a very big impediment when the user wants to place a customized query. Hence, it is required that user should only get to see the data which are of some value for them. This is a problem of clustering, which is still unsolved in the domain of social networking analysis. Although, there are multiple studies of clustering algorithms presented till date [3][4], but very few of them are found to be effective for dynamic messages collected from the social networking applications. Hence, there is a true need of identifying an effective clustering mechanism for social networking application.

Therefore, this paper presents a technique which is capable for identifying the amount of unstructured data and converts it in the form which allows datamining algorithm in future to be implemented. The technique also implements a unique logic of semantic that can perform better data perception followed by non-conventional mechanism of clustering and labeling to represent the extracted data. The proposed technique makes use of graph theory for solving the problem of better data representation in social networking applications. Section 2 discusses about the existing research techniques for addressing the problems of clustering the social networking data followed by brief discussion of problem identification in Section 3. The proposed system to address the unsolved problem of clustering is discussed in Section 4 followed by research methodology in Section 5. Section 6 discusses about the result analysis and finally summarization of the work is carried out in Section 7.

2. Related Work

This section discusses about the existing studies being carried out in the area of clustering techniques over social networking applications.

Ferrara et al. [5] have discussed a study towards analyzing the data of social networking application exclusively focusing on specific unit of information using supervised learning technique. The authors have also used pre-clustering technique along with features being extracted from the heterogeneous fields. Ajorlou et al. [6] have introduced a unique algorithm that can perform an efficient clustering over data collected from social networking sites. The techniques used by the authors are highly empirical in nature using quality threshold along with conventional utilization of k-means clustering algorithm to achieve faster convergence rate. The study outcome was analyzed with respect to histogram over the mutual spatial factors. Hajeer et al. [7] has adopted evolutionary techniques for performing clustering operation over social networking data. The study uses genetic algorithm as mean to perform knowledge extraction for the purpose of identifying the clusters for a given data. The study is implemented over smaller number of synthetic data sets (post, emails, comments, etc). Gao et al. [8] performed the investigation of parallel clustering focusing on high dimensional data gathered from social networking applications. The authors have used acyclic graph over cloud to perform better and effective synchronization of the unstructured and high-dimensional data using Apace Storm which is essentially deployed for processing the stream engine and supports better event-driven designing. The technique was implemented over Hadoop system using ActiveMQ broker. Reuter et al. [9] have focused on investigating on various social networking sites for performing clustering. The study uses the concept of event identification of the data captured from the online transactions where the outcomes are measured with respect to normalized mutual information as well as f1-measure.

Sharma and Gupta [10] have presented a clustering technique that uses planning of business system as the core agenda. The technique has the capability to perform clustering of the multiple classes of social networking data just on the basis of the links as well as relational identity that it shares with other classes. The technique is found to use less memory consumption. Lee et al. [11] have presented a technique of tracking multiple sub-graphs for a large network of dynamic types. The study also extracts the mathematical relationship that it bears with sub-components over fading time window. The study outcome was evaluated over precision and recall rate over various numbers of hash tags and unigrams. Wahl and Sheppard [12] have introduced a study of clustering using Fuzzy logic. Using Jaccard similarity, the prime network is decomposed into smaller data coefficient experimented over multiple forms of network e.g. Dolphin network, Alaska Campaign Finance, Zachary Karate network. Giuliani and Pietrobelli [13] have discussed a technique of evaluating the process of developing clusters over data of social networking sites. The study outcomes were evaluated with respect to the deterministic approach, stochastic approach, as well as probabilistic approach. Whang et al. [14] have addressed the problem of scalability in the clustering process as well as memory conservation. The authors have adopted clustering based on graphical concepts. The study also uses k-means clustering technique with slight amendment over the weighted kernel usage. Aamtrian [15] have presented an approach of mining based on highly customized suggestion system. The study has focused on the problem of the recommendation system existing at present and addressed the problem using datamining approach.

Li et al. [16] have presented a technique of extracting feature using unsupervised approach. The technique makes use of information available from the links in order to apply supervised streaming for extracting features. The study outcome was evaluated with respect to accuracy mainly. Zhao and Zhang [17] have introduced a graph-based technique that uses non-binary tree structure and unique feature in order to perform clustering. The technique has been evaluated and assessed with respect to multiple dataset to find the technique yields better classification techniques. Handcock and Raftery [18] have developed a technique using Bayesian method and Markov modelling in order to formulate the data clustering from the social networking applications. The numerical assessment of the study was carried out using R statistical language. Opsahl and Panzarasa [19] have presented a technique that can perform evaluation of the clustering coefficient for the purpose of retaining the data encoded over the nodes of social networking tree. The study outcome was tested with multiple dataset and the outcome shows better clustering analysis.

Hence, it can be seen that there are multiple techniques present in existing system that is meant for performing clustering of data collected from social networking application. All the above mentioned techniques definitely provide a better guideline for future research work as well as all the associated study is also shrouded with pitfalls that are briefly discussed in next section of problem identification.

3. Problem Identification

The major problem with the existing systems are i) lack of benchmarked outcomes, ii) less focus on unstructured data form present in social networking sites, iii) less efficient use of corpus management over graphs, and iv) usage of sophisticated semantics leads to inapplicability to understand the highly customized text in social networking transactional messages. Moreover in social networking sites, users interact with each other majorly using a text which is highly customized, highly-personalized, and sometimes codes. Hence, if all the data are collected than it eventually formulates a massive and highly complex unstructured data, which is a bigger obstruction towards mining social networking data. Hence, in order to perform efficient clustering, it is important that unstructured data to be converted to structured data using an efficient usage of semantics, lexical, and proper management of corpus to make the system more faster search-efficient with reduced response time from the query end. The next section discusses about the proposed system.

4. Proposed System

The prime purpose of the proposed system to develop a framework that can perform an effective clustering of the social networking messages. The prime objectives of the proposed system are as follows:

(1) To introduce a framework that can manage a massive amount of data in the form of clusters narrowing down to further meaningful and relevant clusters according to the user's requirement.

(2) To perform clustering and social networking messages by converting the unstructured text data to structured text data.

(3) To incorporate the concept of semantic and syntactic operation for better data perception of the unstructured transactional data of social networking applications.

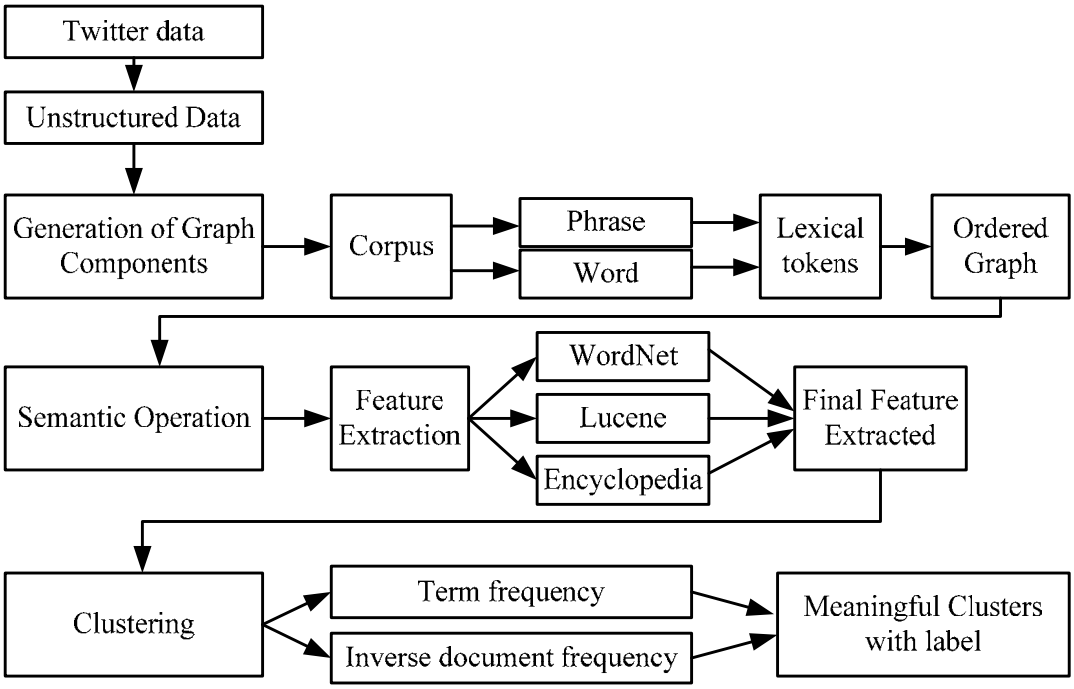

Figure 1. Schematic Architecture of Proposed System

The schematic architecture of the proposed system is highlighted in Fig.1. The proposed system takes the input of unstructured twitter data and applies graph theory in order to perform better data representation. The system performs corpus definition using the phrase and words, which after applying lexical tokens generates an ordered graph. This is finally carried over in next operation that applies semantics. The system using multiple tools e.g. WordNet, Lucene, and different encyclopedia sites to finally extract a feature. Finally, the system performs clustering using term frequency and inverse document frequency. The next section will further elaborate about the methodology adopted for implementing the proposed system with algorithm description.

5. Proposed Methodology

The proposed research work is carried out using analytical research methodology, where the prime idea is to perform clustering of the Social networking messages. The development of the proposed methodology is discussed with respect to the core modules involved in the proposed system as follows:

5.1 Generation of Graph Components

This module is responsible for taking the input of Social networking message and performs decomposition based on phrase and word factor, which finally converts the Social networking message compatible in the form of tree structure for better analysis. The message extracted from the Social networking is quite short and uses colloquial language which makes the mining task quite difficult. Hence, this module implements a graph theory in order to frame the Social networking message and classify the operation with respect to phrase level and word level. The proposed system emphasize on the phrase factor for the purpose of clustering. We use an ordered graph structure with presence of root for representing the string structure. The proposed system uses lexical tokens in order to construct a parse tree. In order to avoid the problem of high dimensional data, we don't analyze all the components of the graph, but perform selection of few of them. We also believe that noun and verb are the important grammar in majority of the text, so we use few combinations of nouns and verbs in order to extract more graph components. In order to resist the generation of redundancies from the graph components, the proposed system chooses only phase or word leaf nodes. The algorithm for the generation of the graph component is shown in Fig.2.

Algorithm for Generating Graph Components

Input: W (corpus), ω (lexical token), O_{grap} (ordered graph)

Output: generation of graph component

Start:

1. For ωεW
2. ω←W && O←O_{grah} (ω)
3. $\omega^{'} = \sum_{i=1}^{n} \omega_i$
4. G→Select graph component
5. G=G+1

End

Figure 2 Algorithm for Generating Graph Component

5.2. Semantic Operation:

This module is responsible for taking the corpus input of Social networking data with respect to phase factor and online encyclopedia in order to perform the generation of semantic concepts. The graph components from the syntactic decomposition phase are not sufficient for the representation of the message due to lack of semantics. This module is responsible for mapping the actual raw Social networking message to highly structured space of semantics. The advantage of this module is that it performs extracts semantics for every graph components give better edge of meaning of the unstructured Social networking message. The meaning extracted from syntactic operation is not enough and hence is further subjected to extraction of features from semantic operation. The first step is to extracts the graph component in order to encapsulate the sub-topics as well as information pertaining to structure in the Social networking message. The proposed system therefore uses semantic factors from phrase and word aspect in order to perform mapping. The proposed system also uses an encyclopedia sites in order to perform collateral information sharing process as reference. We also use WordNet for better semantic operation and extraction of lexical knowledge. The system also uses open source search tool Lucene that we use for indexing the extracted semantic feature with the encyclopedia sites. The algorithm is more focused on phrases in contrast to words. The query is formulated using logical operation of AND that calls for the response page also focusing on the key phrase of the provided message of Social networking. Therefore, the proposed algorithm performs dual processing of both phrase and words in order to provide the feature from the semantic operation that assists more in understanding and converging to relevant messages of the groups. All the problems of synonyms can be handled using semantic operation in proposed algorithm. The steps of the second algorithm is shown in Fig.3.

Algorithm for Semantic Operation

Input: G(graph component), P_F(phrase factor), q(query), ENC (encyclopedia), β (WordNet concept), f (Search tool Lucene), r_{page} (response page)

Output: Extracted feature of semantic (η)

Start

1. For GεG'
2. If GεP_F than
3. G.q←f (G, AND)
4. ENC←r_{page}(G.q)
5. η← η+ $ENC_{concept}$(ENC)
6. else
7. β← β(G)
8. η← η+ β
9. end if
10. end for

End

Figure 3 Algorithm for Semantic Operation

5.3 Clustering Formulation

A new space of feature is developed by integrating syntactic decomposition and semantic operation module in order to perform an efficient clustering processing. As it is a Social networking application, hence there is a possibility that multiple data (both relevant and irrelevant data) surfaces in the console application of blogging. Hence, in order to sort this problem, we proposed a clustering which is based on relevant information of an user which uses semantics for classifying the clusters. The system aggregates all the graph components from the previous module and specific number of semantic concepts from the base of knowledge of semantics. Hence, we compute the quantity of the semantic concepts by dividing product of graph components and interval with cumulative intervals. We choose binary logic as the interval factor corresponding to zero will mean extraction of feature is carried out from individual graph components while when the interval factor is one than it will mean extraction of features are done on the basis of semantic operation.

We extracts the quantity of the concepts that can be derived from semantic using following equation

$$\delta = \frac{\alpha_1 + \alpha_2}{\phi} \qquad (1)$$

In the above equation, α_1 and α_2 are term frequency and inverse term frequency, while the denominator ϕ is the coefficient factor for graph component and depth. We can represent it as a function of generated graph components and depth of graph components of the complete graph. Using the above mentioned concept, it becomes easier to investigate the different features and topics as per the user's choice. After the clustering is carried out, the next task is to perform cataloging the extracted outcomes (or features). The proposed system performs labeling the candidate text by multiplying the variable δ with α_1 and α_2. However, here δ can be said to be weight of original graph while other two variable α_1 and α_2. will represent the weight factor of entire candidate graphs. We look for the labels with maximum value and consider it to be the final output for performing labeling. Therefore, the proposed system gives better opportunity to convert the unstructured data to highly structured data so that it can be subjected to clustering, mining, sub-space exploration, and many other statistical operation for datamining of Social networking messages.

6. Result Analysis

The development of the proposed system is carried out on simple 32 bit machine with windows operating system. The programming is done using java. The study outcome was compared with the similar category of the work done by Tripathy et al. [20]. Following are the outcomes accomplished.

Table 1

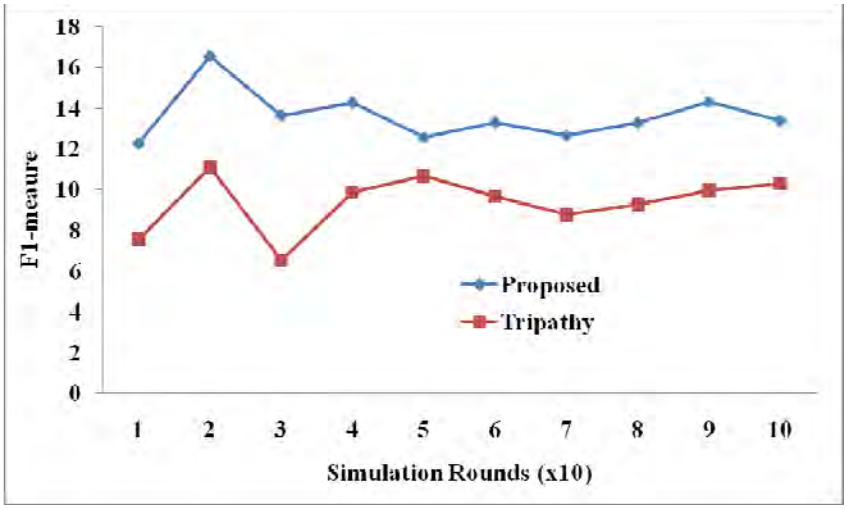

Figure 4 Analysis of F1 measure

Fig.4 shows that F1-measure of the proposed system as well as Tripathy et al. [20] system with respect to the improvement being done over the bag of words method. The improvement of F1 measure of proposed system is found to be 4.25% in contrast to Tripathy et al. [20]. The prime reason behind this is Tripathy et al. [20] has used mathemaical modelling using sparse matrix mainly using distance-based relationship of graphical components, which doesn't support scalability. It will mean that with increasing simulation (with increasing dataset size), the distance increases leading to maximum improvement in F1-score of 10%. Whereas proposed system performs syntactic decomposition technique which makes each unstructured word high structured and then only it occupies the nodal points of the graph followed by semantic operation and clustering, which significantly increases the F1-score value of proposed system.

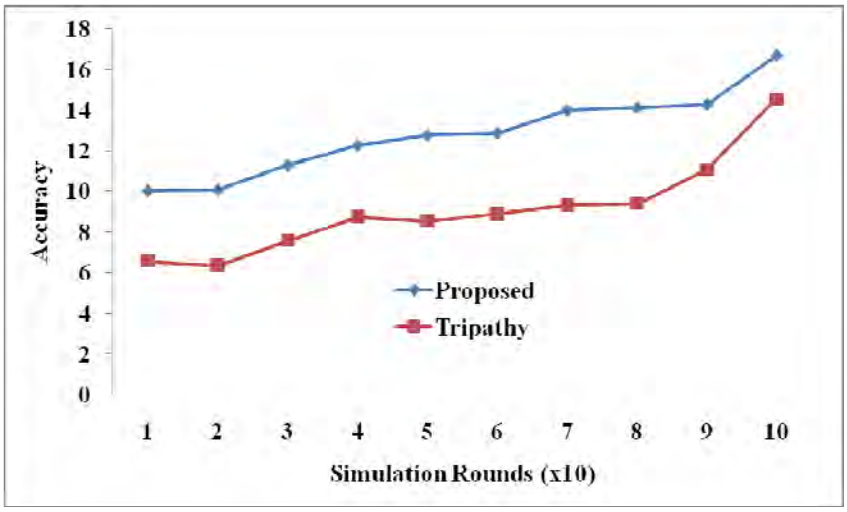

Figure 5 Analysis of Accuracy

Fig.5 highlights the analysis of the proposed system as well as Tripathy et al. [20] with respect to accuracy. The outcome shows that proposed system improves 3.74% of accuracy in contrast to the Tripathy et al. [20] work. The existing system of Tripathy et al. [20] uses bigram technique of Wikipedia using taxonomy graph, whereas there exists many types of encyclopedia sites. The next problem is usage of minimum distance that raises scalability issues in case of synonyms. Moreover, it cannot address the words which are unstructured that found more often in Social networking message. Therefore, the accuracy of the proposed system is high as it offers better comprehension to words by converting unstructured to structured-data with better data representation. It

also uses phrase and word-based syntactical data analysis that leads to better compatibility with semantic operation using WordNet and Lucene.

7. Conclusion

This paper presents a unique clustering mechanism that allows to segregate the collected messages of other users based on relevancy to the user's context. The mechanism applies graph theory where multiple messages are mapped over graphical nodes and are subjected to various forms of syntactic decomposition and semantic operation followed by clustering and labeling. Our future work will further focus on applying optimization algorithm for better convergence of search.

References

[1] Kurylo, A.; Dumova, T. (2016): Social Networking: Redefining Communication in the Digital Age, Rowman & Littlefield, Social Science, pp.214
[2] Vossen, P. (2013): WordNet: A multilingual database with lexical semantic networks, Springer Science & Business Media, Computers, pp.180
[3] Subhasish, D. (2013): Studies in Virtual Communities, Blogs, and Modern Social Networking: Measurements, Analysis, and Investigations: Measurements, Analysis, and Investigations, IGI Global, Technology & Engineering, pp.317
[4] Aggarwal, C.C., Reddy, C. K. (2016): Data Clustering: Algorithms and Applications, CRC Press, Business & Economics, pp. 652
[5] Ferrara1, E., Asbagh, M.J., Varol, O., Qazvinian, V., Menczer, F., Flammini, A. (2013): Clustering Memes in Social Media, arXiv, 2013
[6] Saeede Ajorlou, Issac Shams, Kai Yang, A Fast Clustering Algorithm for Mining Social Network Data, Proceedings of the World Congress on Engineering
[7] Hajeer, M.H., Singh, A., Dasgupta, D., Sanyal, S. (2013): Clustering online social network communities using genetic algorithms, arXiv preprint arXiv
[8] Gao, X., Ferrara, E., Qiu, J. (2015):Parallel Clustering of High-Dimensional Social Media Data Streams, IEEE/ACM International Symposium
[9] Reuter, T., Cimiano, P., Drumond, L., Buza, K. (2011): Scalable Event-based Clustering of Social Media via Record Linkage Techniques, In ICWSM
[10] Sharma, S., and Gupta, R. K. (2010): Improved BSP Clustering Algorithm for Social Network Analysis, International Journal of Grid and Distributed Computing, 3:3
[11] Lee, P., Laks V.S. Lakshmanan, Milios, E.E. (2014): Incremental Cluster Evolution Tracking from Highly Dynamic Network Data, International Conference
[12] Wahl, S., and Sheppard, J. (2015): Hierarchical Fuzzy Spectral Clustering in Social Networks Using Spectral Characterization, In FLAIRS Conference
[13] Giuliani, E., Pietrobelli, C. (2011): Social Network Analysis Methodologies for the Evaluation of Cluster Development Programs, Technical Notes
[14] Jiyoung, J., Sui, X., Dhillon, I.S. (2012): Scalable and Memory-Efficient Clustering of Large-Scale Social Networks, IEEE 12th International Conference
[15] Amatriain, X. (2013): Mining Large Streams of User Data for Personalized Recommendations, SIGKDD Explorations, 14, 2
[16] Liy, J., Huz, X., Tang, J., and Liuy, H. (2005): Unsupervised Streaming Feature Selection in Social Media, International on Conference on Information and Knowledge Management, pp. 1041-1050
[17] Zhao, P., Zhang, C-Q. (2011): A new clustering method and its application in social networks, Elsevier- Pattern Recognition Letters, 32, pp. 2109–2118
[18] Handcock, Mark S., and Raftery, A. E., and Tantrum, Jeremy M.(2007): Model-based clustering for social networks, Journal of the Royal Statistical Society: Series A (Statistics in Society)
[19] Opsah, T., Panzarasa, P. (2009): Clustering in weighted networks, Elsevier- Social Networks, 31, pp.155-163
[20] Tripathy, R M., Sharma, S., Joshi, S.(2014): Theme Based Clustering of Tweets, In Proceedings of the 1st IKDD Conference on Data Sciences, pp. 1-5