An Optimized Fuzzy Means Clustering Algorithm for Grouping of Social Media Data

Ronanki Umarao[1], BeharaVineela [2]

Final M.Tech Student[1], Asst.professor[2]

[1,2]Dept of CSE, *Sarada Institute of Science, Technology and Management (SISTAM), Srikakulam, Andhra Pradesh*

Abstract

Now a day's social media place an important role for sharing human social behaviours and participation of multi users in the network. The social media will create opportunity for study human social behaviour to analyze large amount of data streams. In this social media one of the interesting problems is users will introduce some issues and discuss those issues in the social media. So that those discuss will contain positive or negative attitudes of each user in the social network. By taking those problems we can consider formal interpretation social media logs and also take the sharing of information that can spread person to person in the social media. Once the social media of user information is parsed in the network and identified relationship of network can be applied group of different types of data mining techniques. However, the appropriate granularity of user communities and their behaviour is hardly captured by existing methods. In this paper we are proposed optimized fuzzy means clustering algorithm for grouping related information. By implementing this algorithm we can get best group result and also reduce time complexity for generating cluster groups. The main goal of our proposed framework is twofold for overcome existing problems. By implementing our approach will be very scalable and optimized for real time clustering of social media.

Keywords

Social Media data set, Data Mining, Clusters, Manhattan Distance Means Clustering Algorithm.

I. INTRODUCTION

Social media have grown quickly in popularity in a relatively short time. Social media serves as a ubiquitous public platform which remains accessible to users as a multiple group of internet applications. Within the applications, the user creates individual and unique expression for data exchange [2]. What remains valuable and fascinating is the level of social media data influence as the platform remains an intense portal of human interactions and behaviors. What remains dynamic about social media is the level of opportunity that influences individuals, groups and society. The study of this data by industry specialists seeking new and inventive methods to collect data for analysis remains important to the future of social media [3]. There are so many social media network sites, Twitter and Facebook remain the most well-known but other forms being used are blogs, wikis and platforms with unfiltered text and information. Industry researchers remain focused on social media application business, bioscience and social science. What has been found is that the social media is extremely valuable to statistical study of information technology, social behavior [4] within quantitative attributes for e-learning process and simulation design for further data mining [5].

The rate of digital interaction and exchange of data amongst users increases. These platforms like Facebook and Twitter and some more sites, user expression of opinions and the place many are getting his or her news [6]. These are platforms of free speech and protests [7], organization and sharing of common interests and ways to keep family and friends in the loop of everyday life. Still challenge to industry specialists that collect data from these platforms and social communities. Such data mining actions require the ability to interpret the massive amount of content continuously produced on online social media and manual labeling is infeasible on a large scale. Textual content equals a unit of information and this can also be coded to represent a particular trait of the individual posting the content. Each piece of content also represents a user's score point and this makes the process of assessing information from the standpoint of learning methods as these acts as individual means of identifying group behavior's.

A cluster is generally thought of as a group of items (objects, points) in which each item is closed .A simple cluster representation. To a central item of a cluster and that members of different clusters are "far away" from each other. Clusters can be viewed as "high density regions" of some multidimensional space. A method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. Cluster analysis is another but different from those in other groups. In marketing there is keen

interest among managers in developing products and strategies to target segments. The challenge with cluster analysis is that it involves both art and science, and it always produces an answer whether there really are clean and separable segments or whether consumers are positioned in a continuous cloud. Complicating matters further, there are numerous cluster analysis routines, which can lead to different results. Popular statistical tool for finding groups of respondents, objects, or cases that are similar to one There are given a set of n data points in dimensional space Rd and an integer k and the problem is to determine a set of k points in Rd, called centers, so as to minimize the mean squared distance from each data point to its nearest center.

A popular heuristic for k-means clustering is Lloyd's algorithm. In this paper, a simple and efficient implementation of Lloyd's k-means clustering algorithm is presented, which is called the filtering algorithm. This algorithm is easy to implement, requiring a k means as the only major data structure. On World Wide Web there is a computer program which allows such a statistical technique to be carried out in a very simple way. This paper also shows how this approach can be used with cross cultural data to extract similarities and differences between societies in a systematic fashion. Although the example used focuses on the economic systems of foragers, the methodology is also applicable to a wide variety of other cross-cultural research problems [8].In paper [9] there is introduced an improved k-mean algorithm which is based on background knowledge. Background knowledge is used as constraints to produce desired result and named as "Constrained k-means algorithm with background knowledge." This paper includes four parts. Second part is analysis of standard k-means algorithm and shows the shortcomings of the standard k-means algorithm [10] [11]. The third part introduces the optimized fuzzy means cluster distance algorithm and fourth part shows conclusions.

II. RELATED WORK

The amount of information shared on online social media has been growing during recent years [13]. Much can be learned about the retail and finance behaviors of users by studying social media analysis. It is nothing new that retail companies market via social networks to discover what consumers think about branding, customer relationship management, and other strategies including risk prevention. A good example is the found correlation of data on Twitter with industry market behavior and sentiment posted by users. Wolfram [14] used Twitter data to develop machine learning model using Support Vector Regression and predicted prices of individual stocks and found significant advantages of using social media data for forecasting future prices.

Social media analysis data can be used to track health issues like smoking and obesity for bio-scientific study like Penn State University found innovative systems and techniques to track the spread of infectious diseases because the data social media reflects about users within these groups [15]. Social science applications are concerned, it includes monitoring public responses to announcements, speeches and events with emphasis on political comments and initiatives. It also gives insights about community behavior, social media Polling within groups and early detection of emerging events like, For example by using the computational linguistics, the automatic prediction impact of news on the public perception of political candidates was implemented. Yessenov and Misailovic [16] use reviewing comments of movie. Karabulut found that Facebook also exhibits and captures major public events in its data.

Social network analysis has a well-defined relation and background in sociology [17]. With the rapid growth of the web forums and blogs, the user's participation on content creation led to a huge amount of dataset. Hence the advancement of data mining techniques is required. An overall discussion of one news forum called Slashdot, can be found in Social networks, it focus work like face pager. It is used to access data from social media like facebook by using this data to develop a clustering framework using optimized K-Means algorithm that is more accurate than existing methods. Clustering is used as an exploratory analysis tool that aims at categorizing objects into categories, so the association degree between the objects is maximal when belonging to the same categories. Clustering structures the data into a collection of objects that are similar or dissimilar and is considered an unsupervised learning. The application of our method is mainly on finding user groups based on activities and attitude features as suggested in the authority model.

III.PROPOSED SYSTEM

The amount of information shared on online social media has been growing during recent years. Much can be learned about the retail and finance behaviours of users by studying social media analysis. It is nothing new that retail companies market via social networks to discover what consumers think about branding, customer relationship management, and other strategies including risk prevention. A good example is the found correlation of data on Twitter with industry market behaviour and sentiment posted by users. Social network analysis has a well-defined relation and background in sociology. With the rapid growth of the web forums and blogs, the user's

participation on content creation led to a huge amount of dataset. Hence the advancement of data mining techniques is required. An overall discussion of one news forum called Slashdot, can be found in Social networks, it focus work like face pager. It is used to access data from social media like face book by using this data to develop a clustering framework using optimized fuzzy means clustering algorithm that is more accurate than existing methods. Clustering is used as an exploratory analysis tool that aims at categorizing objects into categories, so the association degree between the objects is maximal when belonging to the same categories. Clustering structures the data into a collection of objects that are similar or dissimilar and is considered an unsupervised learning. The application of our method is mainly on finding user groups based on activities and attitude features as suggested in the authority model.

The standard k-means algorithm takes extra time in calculating distance from each cluster's center in each iteration. The implementation process of k means algorithm is as follows.

1. Read the twitter data set from the twitter server.

2. Enter number of clusters to be performing and randomly choose the centroids from twitter dataset.

3. Take each data point (d_i) from dataset and calculate the Manhattan distance from data point to centroids' (c_i).
   Distance= (c_i -d_i)
4. If check the closet distance of each centroid from the data point and that data points will be put into those clusters.

5. The step 3 and 4 will be repeated until there is no change in the centroids.
6.
7. After completion of step 6 we can get group of clustered data.

8. The calculation of Manhattan distance we can also calculate each cluster sum squared error by using following equation.
   $$SSE=\sum_{i=1}^{n} dis\ (c_i, di)$$

By implementing this algorithm will take time complexity and space complexity. This extra time can be saved by adapting this method. The implementation process of optimized fuzzy means cluster distance algorithm is as follows:

IV. OPTIMIZED FUZZY MEANS CLUSTERING ALGORITHM

Input:
The number of desired clusters, k, and a dataset D= (d1, d2,……., dn) containing n data objects.

Output:

A set of k clusters.

Steps:

1) Randomly select k data objects from dataset D as initial clusters.

2) Calculate the matched words between each data object d_i (1<=i<=n) and each cluster center c_j (1<=j<=k).

3) After completion of matched word we can find out sum squared error by using following formula.
$$SSE=1/w^2$$
4) Calculate total number of words in a data point and centroid find out weight of each data points to centroid. The calculation of weight each twit is as follows.

$$Weight\ (W_i) = 1/dist\ (d_i,c_i)^2\ /\sum_{q=1}^{k} 1/dist(C_i,d_i)$$
5) After completion of weight of each data point to centroids check which data point is near by the centroids.

6) For every cluster center cj (1<=j<=k), it compute the weight of data points d (di, cj) and assign the data object di to the nearest cluster.
   Set cluster[i] = j;
   Set w[i] = d (di, cj).

7) For each cluster center j (1<=j<=k), recalculate the centers;

9) Until the center is same.

10) Output the clustering result.

The optimized fuzzy means cluster distance algorithm is used to reduce time complexity and also space complexity of data objects. This paper does not require calculating distance in each iteration. The time complexity of this algorithm is O (nk). If a data point remains in its initial cluster then the time complexity will be O (1) otherwise O (k). If half of the data points move from its initial cluster then the time complexity will be (nk/2). So the proposed algorithm effectively increases the speed of standard k-means algorithm. But this algorithm also requires the value of k in advance. If one wants the optimal solution then he must test for different values of k.

V. CONCLUSIONS

This paper we are proposed an efficient clustering algorithm for reduce the time complexity and space complexity. This paper proposes optimized fuzzy means cluster distance algorithm for getting better cluster result in data set. By implementing this process we can easily find out similar data object in data set by calculating weight of each data object to centroids. The calculation of weight of data object will repeat until the no changes occur in the centroids. By applying this process we can reduce number of iteration compared to existing algorithm of k means. So that each data point from each cluster center in each iteration due to which running time of algorithm is saved. By implementing proposed system we can efficiently improve speed of the clustering and accuracy by reducing the computational complexity of standard k-means algorithm.

REFERENCES

[1] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. Business horizons, 53(1):59–68, 2010.

[2] Bogdan Batrinca and Philip C Treleaven. Social media analytics: a survey of techniques, tools and platforms. AI & SOCIETY, 30(1):89– 116, 2015.

[3] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. Science (New York, NY), 323(5915):721, 2009.

[4] Claudio Cioffi-Revilla. Computational social science. Wiley Interdisciplinary Reviews: Computational Statistics, 2(3):259–271, 2010.

[5] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In Proceedings of the 19th international conference on World Wide Web, pages 591–600. ACM, 2010.

[6] Michael D Conover, Clayton Davis, Emilio Ferrara, Karissa McKelvey, Filippo Menczer, and Alessandro Flammini. The geospatial characteristics of a social movemen communication network. PloS one, 8(3):e55957, 2013.

[7] Bruce A. Maxwell, Frederic L. Pryor, Casey Smith, "Cluster analysis in cross-cultural research" World Cultures 13(1): 22-38, 2002.

[8] Kiri Wagstaff and Claire Cardie Department of computer science, Cornell University, USA "Constrained k- means algorithm with background knowledge".

[9] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest, Introduction to Algorithms, Prentice Hall, 1990.

[10] Anil K. Jain, M. N. Murty, P. J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, 31(3): 264-323 (1999).

[11] Bogdan Batrinca and Philip C Treleaven. Social media analytics: a survey of techniques, tools and platforms. AI & SOCIETY, 30(1):89– 116, 2015.

[12] Emilio Ferrara, Mohsen JafariAsbagh, Onur Varol, Vahed Qazvinian, Filippo Menczer, and Alessandro Flammini. Clustering memes in social media. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 548–555. IEEE, 2013.

[13] M Sebastian A Wolfram. Modelling the stock market using twitter. School of Informatics, page 74, 2010.

[14] Marcel Salathe, Linus Bengtsson, Todd J Bodnar, Devon D Brewer, John S Brownstein, Caroline Buckee, Ellsworth M Campbell, Ciro Cattuto, Shashank Khandelwal, Patricia L Mabry, et al. Digital epidemiology. PLoS Comput Biol, 8(7):e1002616, 2012.

[15] Kuat Yessenov and Saˇsa Misailovic. Sentiment analysis of movie review comments. Methodology, pages 1–17, 2009.

[16] John Scott. Social network analysis. SAGE Publications Ltd, 2013

[17] Vicenc¸ G´omez, Andreas Kaltenbrunner, and Vicente L´opez. Statistical analysis of the social network and discussion threads in slashdot. In Proceedings of the 17th international conference on World Wide Web, pages 645–654. ACM, 2008.