Volume 7, Issue 5, May 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Mining the Attitude of Social Network Users using K-means Clustering

Vairaprakash Gurusamy
School of IT, Madurai
Kamaraj University,
Tamilnadu, India

S. Kannan
School of IT, Madurai
Kamaraj University,
Tamilnadu, India

J. Regan Prabhu
Department of CS,
Thiagarajar College,
Tamilnadu, India

DOI: 10.23956/ijarcsse/SV7I5/0231

Abstract: Social media is the collective of online communications channels dedicated to community-based input, interaction, content-sharing and collaboration. Social media has become a central point of a person's daily life for many people around the world with the ability to be connected to these sites through access to cellphones, tablets, and computers. The ease of sharing information has allowed people to keep in contact with friends and family and keep them updated on life changes, views of various subjects, collaborate on projects, and much more. It has also made it possible for groups or individuals who can unlike or retweet your posts. User's opinions may be any of forms such as Text, Image, Audio, and video. In this work, take Text format to mining the users' attitude for the social network. The user may tweet a comment using any of the social media to a particular topic from different place and time. K-Means Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Popular notions of clusters include groups with small distances among the Cluster members, dense areas of the data space, and intervals of particular statistical distributions. Using clustering techniques we group the similar and dissimilar of users' attitude.

Keywords: Mining, Rule induction, Social Network, Clustering.

I. INTRODUCTION

In Recent years micro-blogging platforms such as Twitter have seen a rapid growth, and the emergence of social media campaigns, where a massive number of people express strong opinions and provide support for social causes of public interest. Social media blogs are the free microblogging service that allows registered members to broadcast short posts called tweets or posts. Members from Social Media can broadcast tweets or posts and follow other users' posts by using multiple platforms and devices [1, 8]. Posts or Tweets have recently gained a lot of importance due to their ability to disseminate information rapidly. Popular search engines like Google and Bing have started including feeds from Facebook, Twitter and related social blogs in their search results [2]. Implementation of Data Mining in the Social media is the process of representing, analyzing, and extracting actionable patterns from social media data. Vast amounts of new information and data are generated every day through economic, academic and social activities, much with significant potential economic and societal value. Techniques such as text and data mining and analytics are required to exploit this potential. Some of the techniques used for data mining are Artificial Neural networks, Decision Trees, Rule Induction, Genetic Algorithms and Nearest Neighbor.

Artificial Neural networks are Non-linear predictive models that learn through training and resemble biological neural networks in structure. Decision Trees are Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. The extraction of useful if-then rules from data based on statistical significance is Rule Induction. Genetic Algorithms are Optimization techniques based on the concepts of genetic combination, mutation, and natural selection. A classification technique that classifies each record based on the records most similar to it in a historical database comes under Nearest Neighbor.

Clustering is a widely studied data mining problem in the text domains. The problem finds numerous applications in customer segmentation, classification, collaborative filtering, visualization, document organization, and indexing. Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters) so that the data in each subset (ideally) share some common trait often according to some defined distance measure. Clustering classified into many types such as Hierarchical Clustering, Partitioning, Relocation Clustering, Density-Based Partitioning, Constraint-based Clustering, Co-Clustering and Grid-Based Methods [6,7]

II. RELATED WORK

The social network is an important and positive change in our society. It becomes an important aspect of our day to day activities. It is obvious that they contribute a lot, and still contributing. Online networking is made up of clusters of people, bounding themselves together on the World Wide Web [4]. To be able to sort out the many different clusters we belong to we use online groups to helps us arrange and make sense of all our contacts. This sense-making is rooted within us, we sort and put people into compartments or sort by categories to make sense and try to understand our relationships

with the people around us. Online social networking groups, therefore, enable us to do the same thing online. So we are using the efficient clustering method for data mining. K-Means is one of the partition clustering Technique in which a cluster is represented by its centroid. It takes the iterative approach to minimize the sum of distances between data points and their respective nearest centroid [9, 10, 11]. Since we've moved a huge part of our social life to the internet online social networking groups has become very important for us to maintain a structure in our social life. Here, the Opinion Mining plays an important role.

Opinion Mining can be defined as a sub-discipline of computational linguistics that focuses on extracting opinion of persons from the web. Human beings are always keen to know what other peoples think. Whenever a decision has to be taken, the opinions of friends and relatives are always considered. But now a days where the Internet is used by everyone. It presents a theoretical work that consists in defining formally an opinion- oriented model. We have experimented by using it in order to rank forum messages from the most to the least interesting [2, 3].

Social network mining is a growing and exciting area of research that has in front of itself a long way to go, with the contribution of many research fields. The broad range of problems and challenges in mining social networks calls for powerful new methods such as Natural Language Processing and Density Based Clustering algorithms towards new frontiers in understanding the exciting phenomenon of social networks [8].

Social sites have undoubtedly best owed unimaginable privilege on their users to access readily available never-ending uncensored information. Twitter, permits its users to post events in real time way ahead of the broadcast of such events on traditional news media. Also, social network allows users to express their views. We extensively evaluate their relative performance in numerous scenarios [1]. A Large amount of new data created every minute on social networking sites. It is difficult to obtain and interpret. The data and to allow for further analysis. Twitter acts as a great source of rich information for millions of users on the internet and therefore is apt for applying data mining. The notion of community in this social networking world has caught lots of attention. Such algorithms are even harder to analyze users on twitter as it is an asymmetric microblogging service [1, 5].

Online Social Networking sites are getting much popular day by day. Well known sites such as Facebook, LinkedIn, Twitter, Orkut and Google+ have millions of users across the globe. Nowadays in this social network, text mining research is valuable for analyze the people mind and our personal opinions to create a social report. Therefore, our proposals are also quite efficient. K-means clustering is a well-known partitioning method. It is useful for undirected knowledge discovery and is relatively simple. K-means has found widespread usage in a lot of fields, and it has the biggest advantage of clustering large data sets and its performance increases as a number of clusters increases. But its use is limited to numeric values. However, the *K*-means method is the most efficient in terms of execution time [9, 10, 11]. Hence the performance of K- mean algorithm is better than Hierarchical Clustering Algorithm [12].

Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is therefore called are hierarchical agglomerative clustering or HAC. Top-down clustering requires a method for splitting a cluster. It proceeds by splitting clusters recursively until individual documents are reached [12].

Model-based clustering is hypothesized for each cluster and find the best fit of data to the given model. This method locates the clusters by clustering the density function. This reflects the spatial distribution of the data points. This method also serves a way of automatically determining a number of clusters based on standard statistics, taking outlier or noise into account. It, therefore, yields robust clustering methods.

1) Data Retrieval from Twitter Site

Text mining is roughly equivalent to text analytics that refers to the process of deriving high-quality information from twitter. Data retrieval involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including the link and association analysis, visualization, and predictive analytics.

Twitter has followed the procedure to retrieve user commands. We have to create an account in attitude analyzer app in the twitter app management. Now redirect twitter authenticate with the tool using to retrieve a data. OAuth tool has authenticated by twitter app account details. Attitude analyzer application is available in the twitter account to get the information. Now we have rights to retrieve the data so install the required packages in the tool. Finally mining a user's attitude about "WORLD CUP 2015" page. Observe the comments variables that are twitted in the page for the opinion of the peoples. Now the data are moved to the next level data preprocessing to remove the repeated words and unwanted data's.

2) Data Preprocessing

Text preprocessing aims to make the input documents more consistent to facilitate text representation, which is necessary for most text analytics tasks. Stop word removal eliminates words using a stop word list in which the words are considered more general and meaningless. Stemming algorithms differ in respect to performance and accuracy and how certain stemming obstacles are overcome. A simple stemmer looks up the inflected form in a lookup table. The advantages of this approach are simple, fast and easily handles exceptions. Tokenization is the task of chopping it up into pieces, called tokens, perhaps, at the same time throwing away certain characters, such as punctuation. This method is used to find out the root/stem of a word. For example, the words user, users, used, using all can be stemmed to the word

"USE". The purpose of this method is to remove various suffixes, to reduce the number of words, to have exactly matching stems, to save memory space and time.

3) Clustering Methods

K-Means Clustering is a partition method technique which finds mutual exclusive clusters of spherical shape. It generates a specific number of disjoint, flat (non-hierarchical) clusters K- Means algorithm organizes objects into k – partitions where each partition represents a cluster. We start out with an initial set of means and classify cases based on their distances to their centers.

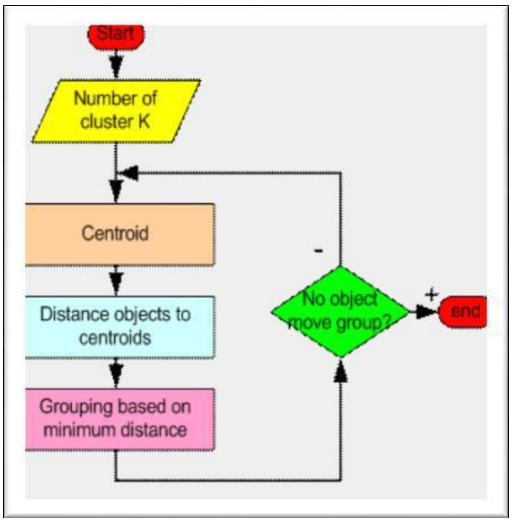

Fig.1

Step1: Begin with a decision on the value of K= number of clusters.
Step2: Put any initial partition that classifies the data into k clusters. You may assign the training samples randomly, or systematically as the following:

i. Take the first k training sample as single-element Clusters
ii. Assign each of the remaining (N-k) training sample to the cluster with the nearest centroid. After each assignment recomputed the centroid of the gaining cluster.

Step3: Take each sample in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.
Step4: Repeat step 3 until convergence is achieved, that is until a pass through the training sample causes no new assignments

III. PROPOSEDMETHOD

Cluster 1: In the first cluster, various names are grouped together, especially country names grouped together. Hence from this cluster, we can extract the user attitude more idea and knowledge about the countries.

Cluster 2: In the second cluster, it indicates that more verbs are grouped together, especially action verbs grouped together. Hence from this cluster, we can extract the user attitude, more idea and knowledge about the action verbs.

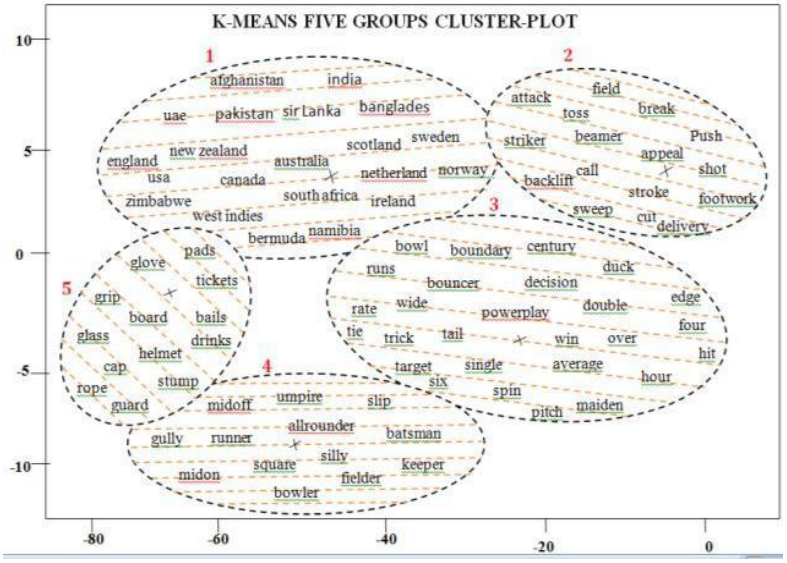

Fig.2

Cluster 3: In the third cluster, it indicates the player name and position in the ground are grouped together. Hence from this cluster, we can extract the user attitude, more idea and knowledge about players and their positions.

Cluster 4: In the fourth, cluster it indicates more objects are grouped together, especially the objects are used by the cricket players. Hence from this cluster, we can extract the user attitude, more idea and knowledge about objects.

Cluster 5: In the first cluster various names are grouped together, especially country names grouped together. Hence from this cluster, we can extract the user attitude more idea and knowledge about the countries.

K-Means Four Group Cluster Plot

In the four group cluster, verb and adverb words are grouped together, rest of them remains same in the same cluster group.

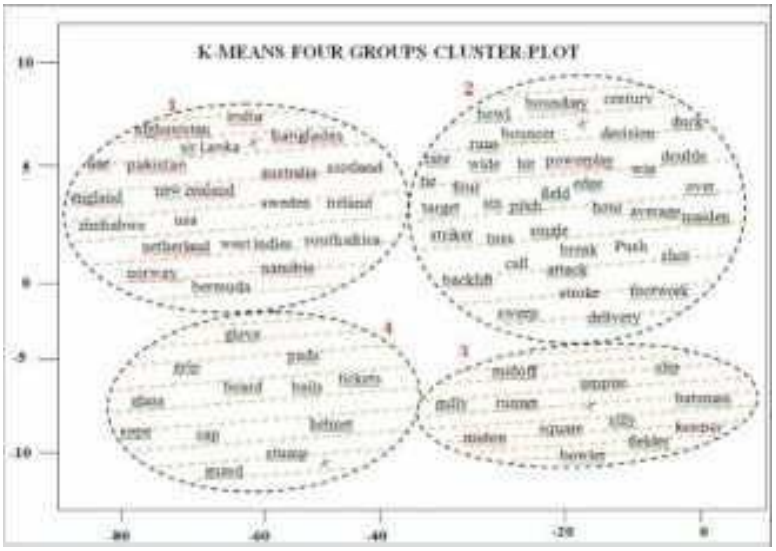

Fig.3

In the two group cluster county, names and object are one group, while verb and adverbs are another groups of the cluster. Rest of them remains same in the same cluster group.

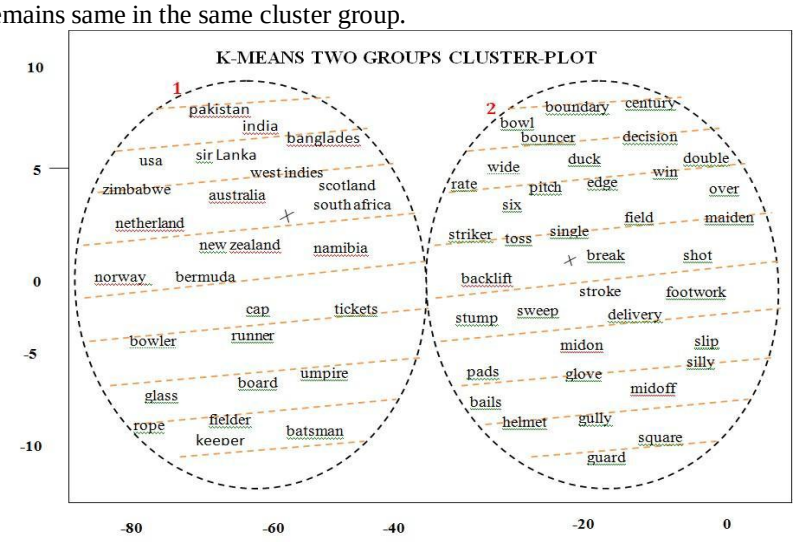

Fig.4

IV. CONCLUSION

As the number of document increase, the performance of Hierarchical will decrease and time for execution increased. K-mean algorithm also increases its time of execution but as compared to Hierarchical, its performance is better. The hierarchical algorithm shows more quality as compared to k-mean. In general conclusion, the k-mean algorithm is good for the large dataset and Hierarchical and Model-Based is good for small datasets.

REFERENCES

[1] Dugundji,E.R, Poorthuis,A and Van meeteren "MModeling:User Behavior in Adoption and Diffusion of Privacy, Twitter Clients". Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom),

[2] Kanavos A, Perikos I, Vikatos P, Hatzilygeroudis I, Makris and Tsakalidis A "Conversation Emotional Modeling in Social Networks" Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on.2014

[3] Kandias M, Stavrou V, Bozovic N, Mitrou L, and Gritzalis "Can We Trust This User? Predicting Insider's Attitude via YouTube Usage Ubiquitous Profiling" Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC), 2013

[4] Devmane M.A, Rana N.K "Usability trends and access of Online Social Network by Indian Nascent population and its analysis". Technologies in the Engineering Field (ICNTE), 2015 International Conference on.2015

[5] Ostrowski, D.A. "Feature Selection for Twitter Classification". Semantic Computing (ICSC), 2014 IEEE International Conference, 2014

[6] Fasheng Liu, Lu Xiong, "Survey on text clustering algorithm". Software Engineering and Service Science (ICSESS), 2011 IEEE 2nd International Conference, 2011

[7] Xiang-Wei Liu, Pi-Lian He, and Hui-Ying Wang. "The research of text clustering algorithms Machine based on frequent term sets" Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on (Volume:4) 2005

[8] Li Xinwu, "Research on Text Clustering Algorithm Based on Improved K-means Design and Applications" (ICCDA), 2010 International Conference (Volume: 4),2010.

[9] Yujun Lin, Ting Luo, Sheng Yao, Kaikai Mo, Tingting Xu and Caiming Zhong "An improved clustering-method based on k means", Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference, 2012

[10] Yufang Liu, Shibin Xiao, Xueqiang Lv, Shuicai Shi "Research on K-means Text Clustering Algorithm Based on Semantic", 2014

[11] Huiying Wang, Xiangwei Liu, "Study based on frequent term set" Eighth International Conference on Text Mining (Volume:2), 2011

