Social Media Analysis using Optimized K-Means Clustering

K. Madhuri¹, Mr. K. Srinivasa Rao²

¹Associate Professor, ²Final M.Tech Student
¹²Department of CSE, Sarada Institute of Science,
¹²Technology and Management (SISTAM), Srikakulam, Andhra Pradesh, India

ABSTRACT
Now a day’s social media place an important role for sharing human social behaviors and participation of multi users in the network. The social media will create opportunity for study human social behavior 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 behavior 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 overcoming exist problems. By implementing our approach will be very scalable and optimized for real time clustering of social media.

1. INTRODUCTION
Clustering is the process of partitioning or grouping a given set of patterns into disjoint clusters. This is done such that patterns in the same cluster are alike and patterns belonging to different clusters are different. Clustering has been a widely studied problem in a variety of application domains including neural networks, AI, and statistics. Data clustering is considered an interesting approach for finding similarities in data and putting similar data into groups. Clustering partitions a data set into several groups such that the similarity within a group is larger than that among groups. The idea of data grouping, or clustering, is simple in its nature and close to the human way of thinking; whenever we are presented with a large amount of data, we usually tend to summarize this huge number of data into a small number of groups or categories in order to further facilitate its analysis. Moreover, most of the data collected in many problems seem to have some inherent properties that lend themselves to natural groupings. Nevertheless, finding these groupings or trying to categorize the data is not a simple task for humans unless the data is of low dimensionality (two or three dimensions at maximum.) This is why some methods in soft computing have been proposed to solve this kind of problem. Those methods are called “Data Clustering Methods” and they are the subject of this paper. Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction. By finding similarities in data, one can represent similar data with fewer symbols for example. Also if we can find groups of data, we can build a model of the problem based on those groupings.

As mentioned earlier, data clustering is concerned with the partitioning of a data set into several groups such that the similarity within a group is larger than that among groups. This implies that the data set to be partitioned has to have an inherent grouping to some extent; otherwise if the data is uniformly distributed, trying to find clusters of data will fail, or will lead to artificially introduced partitions. Another problem that may arise is the overlapping of data groups.
EXISTING SYSTEM:
In the existing technique will take by performing clustering of social media information will face so many problems. For example take the k means clustering algorithm for performing clustering process will take more time and will not get efficient clustering result. By implementing the k means algorithm we can also face the problem of space complexity. In the k means algorithm we can also face the problems of random centroid selection. By choosing random centroids is poor generation of cluster groups. By overcome those problems we can implementing bisecting optimized cluster distance algorithm.

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 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. 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 cluster distance 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 (di) from dataset and calculate the Manhattan distance from data point to centroids’ (ci). Distance = |ci-di|
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. After completion of step 6 we can get group of clustered data.
7. The calculation of Manhattan distance we can also calculate each cluster sum squared error by using following equation.
   \[ SSE = \sum_{i=1}^{n} \text{dis}(c_i, d_i) \]

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

Optimized Fuzzy Means Clustering Algorithm:
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 di (1<=i<=n) and each cluster Center cj (1<=j<=k).
3. After completion of matched word we can find out sum squared error by using following formula.
   \[ SSE = \frac{1}{w^2} \]
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.
   \[ \text{Weight}(W_i) = \frac{1}{\text{dist}(d_i, c_j)} \]
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, c_j) and assign the data object di to the nearest cluster.
   Set cluster[i] = j;
   Set w[i] = d (di, c_j).
7. For each cluster center j (1<=j<=k), recalculate the centers;
8. Until the center is same.
9. Output the clustering result.

Operational Feasibility:
Proposed projects are beneficial only if they can be turned out into information system. That will meet the organization’s operating requirements. Operational feasibility aspects of the project are to be taken as an important part of the project implementation. Some of the important issues raised are to test the operational feasibility of a project includes the following: –
- Is there sufficient support for the management from the users?
- Will the system be used and work properly if it is being developed and implemented?
- Will there be any resistance from the user that will undermine the possible application benefits?

This system is targeted to be in accordance with the above-mentioned issues. Beforehand, the management issues and user requirements have been taken into consideration. So there is no question of resistance from the users that can undermine the possible application benefits.

The well-planned design would ensure the optimal utilization of the computer resources and would help in the improvement of performance status.

2. SYSTEM DESIGN
UML Diagrams:
Systems design is the process or art of defining the architecture, components, modules, interfaces, and data for a
system to satisfy specified requirements. One could see it as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture, and systems engineering. If the broader topic of product development “blends the perspective of marketing, design, and manufacturing into a single approach to product development,” then design is the act of taking the marketing information and creating the design of the product to be manufactured. Systems design is therefore the process of defining and developing systems to satisfy specified requirements of the user. Until the 1990s systems design had a crucial and respected role in the data processing industry. In the 1990s standardization of hardware and software resulted in the ability to build modular systems. The increasing importance of software running on generic platforms has enhanced the discipline of software engineering. Object-oriented analysis and design methods are becoming the most widely used methods for computer system design [citation needed]. The UML has become the standard language used in Object-oriented analysis and design [citation needed]. It is widely used for modeling software systems and is increasingly used for high designing non-software systems and organizations.

Use Case Relationships:
Three relationships among use cases are used often in practice:

Elements of a Collaboration Diagram:
A Collaboration diagram consists of the following elements:

<table>
<thead>
<tr>
<th>Element and its description</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Object</strong>: The objects interacting with each other in the system. Depicted by a rectangle with the name of the object in it, preceded by a colon and underlined.</td>
<td><img src="image" alt="Object Diagram" /></td>
</tr>
<tr>
<td><strong>Relation/Association</strong>: A link connecting the associated objects. Qualifiers can be placed on either end of the association to depict cardinality.</td>
<td><img src="image" alt="Relation Diagram" /></td>
</tr>
<tr>
<td><strong>Messages</strong>: An arrow pointing from the commoning object to the destination object shows the interaction between the objects. The number represents the order/sequence of this interaction.</td>
<td><img src="image" alt="Message Diagram" /></td>
</tr>
</tbody>
</table>

3. **LANGUAGE SPECIFICATIONS**

**Java Technology**
Java Architecture: Java’s architecture arises out of four distinct but interrelated technologies:
- The Java programming language
- The Java class file format
- The Java Application Programming Interface
- The Java virtual machine

When you write and run a Java program, you are tapping the power of these four technologies. You express the program in source files written in the Java programming language, compile the source to Java class files, and run the class files on a Java virtual machine. When you write your program, you access system resources (such as I/O, for example) by calling methods in the classes that implement the Java Application Programming Interface, or Java API. As your program runs, it fulfills your program’s Java API calls by invoking methods in class files that implement the Java API. You can see the relationship between these four parts.
Software Development Life Cycle

The Systems Development Life Cycle (SDLC), or Software Development Life Cycle in systems engineering, information systems and software engineering, is the process of creating or altering systems, and the models and methodologies that people use to develop these systems.

In software engineering the SDLC concept underpins many kinds of software development methodologies. These methodologies form the framework for planning and controlling the creation of an information system the software development process.

WHAT IS SDLC

A software cycle deals with various parts and phases from planning to testing and deploying software. All these activities are carried out in different ways, as per the needs. Each way is known as a Software Development Lifecycle MODEL (SDLC)[2]. A software life cycle model is either a descriptive or prescriptive characterization of how software is or should be developed. A descriptive model describes the history of how a particular software system was developed. Descriptive models may be used as the basis for understanding and improving software development processes or for building empirically grounded prescriptive models.

SDLC models:

- **The Linear model (Waterfall)** - Separate and distinct phases of specification and development. - All activities in linear fashion. - Next phase starts only when first one is complete.
- **Evolutionary development** - Specification and development are interleaved (Spiral, incremental, prototype based, Rapid Application development). - Incremental Model (Waterfall in iteration), - RAD (Rapid Application Development) - Focus is on developing quality product in less time, - **Spiral Model** - We start from smaller module and keeps on building it like a spiral. It is also called Component based development.
- **Formal systems development** - A mathematical system model is formally transformed to an implementation.
- **Agile Methods** - Inducing flexibility into development.
- **Reuse-based development** - The system is assembled from existing components.
5. CONCLUSION

This paper we are proposed an efficient clustering algorithm for reduce the time complexity and space complexity. This paper proposes optimized fuzzy means clustering 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.

6. REFERENCES

[8] KiriWagstaff and Claire Cardie Department of computer science, Cornell University, USA “Constrained k-means algorithm with background knowledge”.