A Survey of Clustering Techniques in Topic Detection

Prakruti Parmar[1] Shafin Vahora[2]
[1,2]Department of Computer Engineering
[1,2]Ipcowala Institute of Engineering & Technology, Dharmaj, Anand, Gujarat, India- 388430

Abstract— The overall goal of text mining is to extract the hidden information from any unstructured data. In text mining, the meaningful patterns are extracted from the gathered data. Cluster analysis is the process in which the similar objects are gathered in one group and the dissimilar objects are considered as outliers. It plays a fundamental role in detecting the topics. Topic detection is the procedure of detecting the topics from the accessible articles. It detects the arrival of new events. In this survey paper, an outline is given about different clustering algorithms with their pros and cons which can be used in topic detection.

Key words: Text mining, Topic Detection, Clustering, Agglomerative Hierarchical Clustering, K-means

I. INTRODUCTION

Text mining is associated to data mining, apart from that data mining tools are considered to handle structured data, but text mining can work with unshaped or semi- structured data sets [1]. The text mining techniques starts with gathering of text documents, than a text mining tool for pre-processing is applied. This pre-processing technique cleans and formats the data; moreover it is responsible for extracting the significant characteristics from these documents. In next stage, the text mining techniques such as clustering algorithm is used to set up the documents. Figure 1 describes the whole process of text mining.

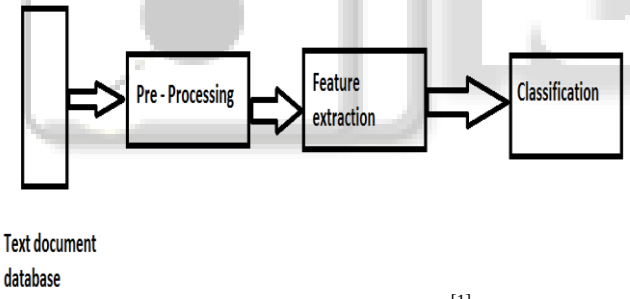

Fig. 1: Text mining Process [1]

Cluster analysis is the process in which the similar objects are gathered in one group and the dissimilar objects are considered as outliers. Cluster analysis divides data into significant or valuable clusters. If a significant cluster is the aim, then the resulting clusters should be same as the original data. However, in other cases, cluster analysis is only a valuable starting point for other purposes, for example, efficiently finding the nearest neighbours of objects in the data. Whether for understanding or convenience, cluster analysis is used in a broad diversity of fields: psychology, biology, pattern recognition, information retrieval, machine learning, and data mining.

Classification predicts the label of unlabeled data, whereas Clustering groups things into "natural" categories when there is no class label available. It is unsupervised learning. It requires that it automatically decides on the grouping structure.

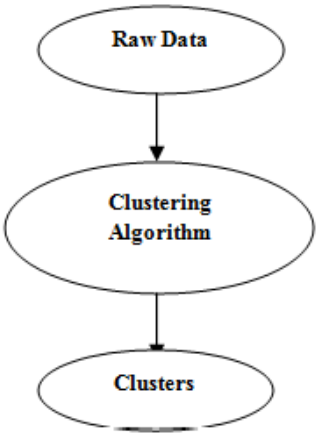

Fig. 2: Stages of Clustering [5]

II. TOPIC DETECTION

The internet has become the most vital source of information and a number one in the creation of events over the past few years [6]. It has the open civilization of publishing news and information and therefore made it an important implication for the pulse of the society. Social networks, now-a-days, have become a necessary source of information. Blogs, Twitter and Face book, have played a great part in near past and in current events over the globe. So, due to this, it was very significant to have a system that can mine this information without any human interference. Topic detection is the process of detecting the occurrence of a new event such as an earthquake event, a murder, any sports activity, or a political humiliation in a flow of news stories from various sources.

The topic detection evaluates technologies that determine novel, previously unidentified, topics. In the topic tracking, topics are defined by clustering the stories that consider the topic. However, topic detection systems are not given a priori information of the topic. Therefore, systems must indicate an understanding of what the topics say about . The task is multilingual and for that reason systems must make clusters that cover languages. The systems detect clusters of stories that discuss the same topic. The concept of clustering is easily applied to news stories, but the evaluation of performance is difficult because stories repeatedly discuss numerous topics. This fact not only means the topic clusters are dependent on previously processed stories, but also that degeneration of presentation into casual subsets is misleading.

Topic detection is an unsupervised assignment. Its input is a set of topic and the output is a definite clustering of topics or events. The type of clustering used determines whether a story can be assigned to several clusters or not. The clustering depends on the characteristics selected in the data vector. This task overlaps with First Story Detection.

All rights reserved by *www.ijsrd.com*

III. TOPIC DETECTION TASKS:

A. First Story Detection (FSD):

It is the problem of recognizing the arrival of new topic that had not been discussed previously. If the system is good FSD then it will detect the first news story like earthquake disaster, any political humiliation or a bomb's outburst. After the appearance of a new story, its feature is compared to all previously stories. If there is dissimilarity between them then it is marked as a first story otherwise it is not [2][3]. The output of the FSD systems says YES or No to the question that does the detected story has a new topic.

B. Story Segmentation:

The Story Segmentation task is to automatically detect the boundaries between stories. It is a pre-task for the remaining three tasks. That is, each of the other tasks is thought of at the story level, so it has stories as its input. Evaluation is measured by judging at each position the correctness of the segmentation:

- Segmentation is judged correct, if there is both a computed and reference boundary within the interval.
- Segmentation is judged correct, if within the interval, there are no computed or reference boundary.
- If within the interval which contains a reference boundary but do not contain computed boundary is said as a missed detection.
- A false alarm is considered if there is a computed boundary within the interval but no reference boundary [2][3].

C. Cluster Detection:

Here, the goal is not only to detect the arrival of new topics but also to cluster the stories which have the same topic into the bins. When a first story arrives that time a new bin is created, but the system has no knowledge about the number of bins previously. The problem in cluster detection is to find the appropriate evaluation [2]. The current evaluation says that no story can reside in more than one bin, but some stories discuss many topics.

D. Story Link Detection (SLD):

Here, the goal is not only to detect the arrival of new topics but also to cluster the stories which have the same topic into the bins. When a first story arrives that time a new bin is created, but the system has no knowledge about the number of bins previously. The problem in cluster detection is to find the suitable evaluation. The current evaluation says that no story can reside in more than one bin, but some stories discuss many topics.

Here, the system is given two new stories and asks it to determine whether or not they have same topic. These tasks compare similarity functions to determine which ones suit the best [2]. But it is not strongly adopted by the research community, as this task is not clear about how one would use this technique.

IV. METHODS FOR TOPIC DETECTION:

A. Distance-based Clustering Algorithms:

They are designed by using a similarity function which is used to measure the distance between the topics. The most popular similarity function is the cosine similarity function.

1) Agglomerative Hierarchical Clustering Algorithm (AHC):

AHC is a bottom-up method. Here, each objects are initially taken as a cluster and the most similar objects are clustered into one cluster [5]. This procedure continues until all objects get into one single cluster or till the termination condition. The resulting tree of clusters is known as a dendogram.

a) Pros and Cons of AHC:

It easily aggregate texts into a smaller and sufficient class. It helps us to produce the ordering of the objects which is useful for displaying the data. If the merger decision is wrongly made then errors are generated which leads to low quality clusters. In AHC, every pre-processing step could not undo and also objects cannot be exchanged between classes.

B. Distance-based Partitioning Algorithms:

1) K-medoid Clustering Algorithms:

In this algorithm, some set of points are used from the original data as medoids and around them clusters are built. Each document is assigned to its closest medoid. So, it develops a set of clusters from the dataset which are improved by the randomized process.

a) Pros and Cons of K-medoid Clustering Algorithm:

It is simple to understand and also implementing it. It is fast and convergent as it has only finite processing steps. This algorithm is very slow as it requires a large number of iterations to achieve convergence. Also, it doesn't work well for sparse data like text.

2) K-mean algorithm:

This algorithm also uses a set of k representatives due to which clusters are built. But these representatives are not obtained from the original data.

a) Steps in K-mean:

- Firstly, it selects any random k of the objects.
- Each object initially represents a cluster centre or mean.
- To the remaining objects, an object is assigned to the cluster.
- For each cluster, the new mean is computed.
- This procedure is repeated unless the function not converges.

b) Pros and Cons of K-mean algorithm:

Implementation of k-mean is easy. Also, it may be computed faster than hierarchical clustering with a big number of variables. It produces tighter clusters than AHC algorithm. But, if there are fixed number of clusters then it makes difficult to predict K that what it should be.

C. Density-based algorithm:

This algorithm plays a important role to find the non-linear shaped structure which is based upon the density. Here, clusters are considered by areas of higher density than the dataset remainder. The goal of this algorithm is reachability

All rights reserved by www.ijsrd.com

of density and connectivity of density. The most widely used density-based clustering algorithm is the Density-Based Spatial Clustering of Applications with Noise (DBSCAN). It depends on a density-based cluster notion. Also, it discovers arbitrary shape clusters with noise in spatial database [5].

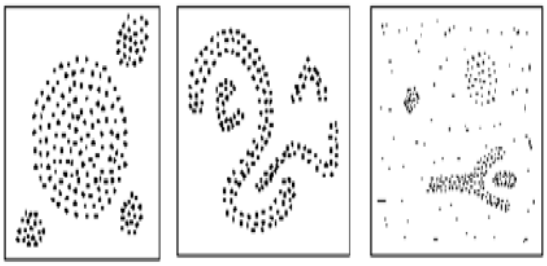

Fig. 3: Density Based Clustering[4]

1) Pros and Cons of Density-Based Algorithm:

The advantage of this algorithm is that it doesn't require Apriori specification. It identifies noisy data while clustering process. If there is a neck type dataset then it fails and it doesn't work properly in high dimensionality data.

V. Conclusion

In this paper, various clustering techniques for topic detection is surveyed. The topic detection evaluates technologies that determine newly and previously unidentified topics. AHC is used for topic detection which easily aggregates texts into a smaller and sufficient class. It helps us to produce the ordering of the objects which is useful for displaying the data.

References

[1] Mr. Rahul Patel, Mr. Gaurav Sharma, "A survey on text mining techniques", IJECS, 2014.
[2] J. Allan, "Introduction to topic detection and traking", Kluwer Academic Publishers, 2002.
[3] Jonathan G. Fiscus, George R. Doddington,"Topic Detection and Tracking Evaluation Overview.
[4] K. Kameshwaram, K. Malarvizhi, "Survey on Clustering Techniques in Data Mining", IJCST, 2014.
[5] Amandeep Kaur Mann, Navneet Kaur, "Review Paper on Clustering Techniques", GJCST, 2013.
[6] M. Vijayalakshmi, M. Renuka Devi, "A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets", IJARCSSE, 2012.
[7] http://www.cse.aucegypt_edu/~rafea/CSCE590/Spring11/Presemtations/TopicDetectionandTrackingwithinSocialNetworks.pdf

All rights reserved by *www.ijsrd.com*