International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 7 4797 - 4802

An Approach towards Data Clustering By Using NLP and Annotated Text Categorization

Amol V. Kale
Department of Computer Technology
Priyadarshni College of Engineering,
Nagpur, India
amol11022@gmail.com

Shaikh Phiroj Chhaware
Department of Computer Technology
Priyadarshini College of Engineering,
Nagpur, India
firoj466@yahoo.com

Animesh R. Tayal
Department of Computer Technology
Priyadarshini College of Engineering,
Nagpur, India
annu09in@gmail.com

***Abstract*—**Aim is to develop system for clustering of data into user defines clusters with the help of language processing. The main objective behind this research is to solve the problem of data classification into large dataset to get an efficient system which classifies data not only on basis of the dataset, but also on basis of the property of keyword and specified class. This provides the best optimization and segmentation, which incorporate a priori knowledge of existing dataset. This will help end user to choose the item from the particular data cluster from its previous parches or search from the dataset. This field leads to: event resolution, grammar annotation, information mining, knowledgebase, labeling, question/answer, redundancy reduction, similarity measure, summarization, word sense disambiguation, and word sense induction. Implementation of application of Apriory algorithm on the given data to classify the data into the categories. Bisecting K-Means algorithm and hierarchical clustering used categorizing all objects in single cluster. PDDP is the latest development of SVD-based partitioning techniques.

Keywords —*Dataset, Cluster, Hierarchical Clustering, NLP, Apriory algorithm, Bisecting K-means Algorithm.*

*****

I. INTRODUCTION

The Web is huge, diverse, and dynamic and thus raises the scalability, multimedia data, and temporal issues respectively. Due to those situations, we are currently drowning in information and facing information overload [5].

In present day human beings are used in the different technologies to adequate in the society. Each and every day the human beings are using the vast data and these data are in the different fields[5]. It may be in the form of documents perhaps graphical formats, may be the video, and may be record. As the data are available in the different formats so that the proper action to be taken. Not only to analyze these, data but also make good decision and maintain the data[1]. As and when the customer will require the data should be retrieved from the database and make the better decision. This technique is actually we called as a data mining or Knowledge Hub or simply KDD (Knowledge Discovery Process) [2][7]. The important reason that attracted a great deal of attention in information technology the discovery of useful information from large collections of data industry towards field of "Data mining" is due to the perception of "we are data rich but information poor"[3]. There is huge volume of data, but we hardly able to turn them in to useful information and knowledge for managerial decision making in business. To generate information it requires massive collection of data. It may be different formats like audio/video, numbers, text, figures, and Hypertext formats [13][5]. Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions.

Increasing no of database has become web accessible through web based search interface. The data units which are encoded to the machine processable, which required for many applications on the internet, that's why they extracted in useful manner [7][12]. The automatic annotations approach that segment the extracted text into the appropriate semantic cluster. Then each segment we represent as the aggregate annotation label for extracted text. The referred web data base or dataset has multiple records, each extracted word belongs to the multiple search result records from web database[1]. These extracted words belongs to the real world entity, it corresponds to the value of word under the attribute[8]. Employ the text mining algorithms that make the use of ontology to identify the necessary artifacts such as different parts and different predefined labels. Text driven development methodology for unstructured text data help to categorize the extracted text to the related cluster[11]. NLP algorithm proposed to automatically develop the structured text data to the predefined data labels or cluster[14]. This improves the following system capability,

- Text information retrieval from predefined dataset.
- Manages the extracted text in proper labeled format of predefined dataset.

Clustering is a very powerful data mining technique for text discovery from data set. Clustering report well for the extracted text data and provide minimized and maximized the optimization problem [8]. Patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster [3]. An example of clustering is depicted in Figure 1. The input patterns are shown in Figure 1(a), and

the desired clusters are shown in Figure 1(b). Here, points belonging to the same cluster are given the same label. The variety of techniques for representing data, measuring proximity (similarity) between data elements, and grouping data elements has produced a rich and often confusing assortment of clustering methods[8].

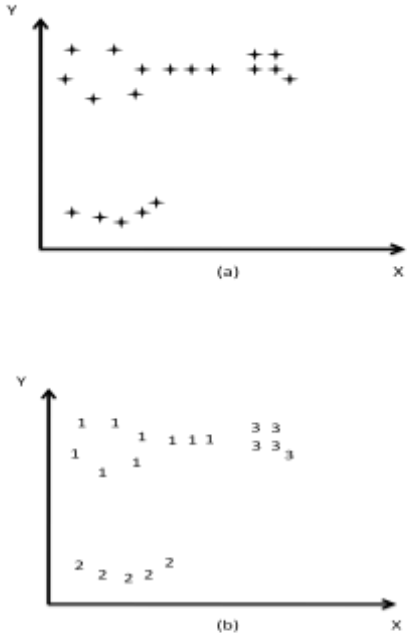


Fig 1: Data Clustering

Clustering is useful in several exploratory pattern-analysis, grouping, decision- making, and machine-learning situations; including data mining, document retrieval, image segmentation, and pattern classification.

The problem of data clustering is generally defined as follows:

Given a set of text data, we would like to partition them into a predetermined or an automatically derived number of clusters, such that the data fields assigned to each cluster are more similar to each other than the text data assigned to different clusters [2]. In other words, the text data in one cluster share the same topic, and the text data in different clusters represent different topics[1]. Depending on how many data units a text node may contain, we identify the following four types of relationships between data unit and text node.

1. One-to-One Relationship: Each text node contains exactly one data unit, i.e., node contains the value of a single attribute.
2. One-to-Many Relationship: In this type, multiple data units are encoded in one text node. One Data node can include many data node.
3. Many-to-One Relationship: Multiple text nodes together form a data unit. Data fields are valid if the sub node is extraction of the main node.
4. One-To-Nothing Relationship: The text nodes belonging to this category are not part of any data unit inside dataset. Ex. Text nodes like “Vehicle” and “Gear” are not data units, but due to data labels describing the meanings of the corresponding data units. We employ a some basic Data unit and Text node they are as follows.

Data Content: The data units with the often share certain keywords. The data units corresponding to the search field where the user enters a search condition usually contain the search keywords [6]. Sometime the user required data are put into label to get the exact data.

Presentation Style: This display a data unit presentation over webpage. It consists of different style features like font face, font size, font color, font weight, text decoration, bold, italic [11]. Data units with same concept in different text data are usually displayed in the same style.

Data Type: It contains some of predefined data types for the different attributes like Date, time, Currency, Integer, Decimal. If the result is same data field it belong to another data field also, like if resulted text is Monday then it follows the data label of weekdays [13].

Tag Path: This contains its own value or name that lead to path or data field in resulted database.

Adjacency (AD): Same set of data labels are in one set are immediate before and after the preceding and succeeding data [6].

Basic Annotators: In a result page containing multiple text data the data units corresponding to the same concept share special common features [6]. These associated with the same or different pattern, Based on this we define four basic annotators to label data units, with each of them considering a special type of patterns [6][7]. Four of these annotators are as follows:

1. Table annotator
2. Query-based annotator
3. In text prefix/suffix annotator
4. Common knowledge annotator

On the basis of these annotators different data labels are follows different data units.

II. Problem Statement

The problem could be defined in higher length for big dataset when we are splitting the data from it, sometime they are specified [7], sometimes they are specified, but not according to the input. Same problem associated with the sub clusters which get defines by the main cluster in table. Mapping of data into cluster according to the basic property with the help of text categorization [12]. Many time there are

restrictions on the clusters, at defining each level is a problem at data node, for this purpose the balanced and hierarchical clustering is used [1]. Part of speech, tagging and chunking are get best use by using the formation of NLP. This provides the good utilization of data words or data set to manage the clusters which are defined by the user[9][10].

III. Literature Review

Data annotation problem and proposed a multi-annotator approach to automatically merging of extracted text data to the predefined dataset. This approach can be naturally divided into two sub-problems:

1. The problem of choosing which cluster must be divided
2. The problem of splitting the selected cluster.

This approach consists in recursively splitting a cluster into two sub-clusters, starting from the main data-set. This is one of the more basic and common problems in fields like pattern analysis, data mining, document retrieval, image segmentation, decision making, the problem is how we are splitting the selected cluster [8]. The new method for the selection of the cluster to split is proposed. This provides the best compromise between computational complexities.

Clusters consist of smaller records, therefore, fewer pages from secondary memory are access to process transactions that retrieve or update only some attributes from the relation, instead of the entire record. This leads to better query performance. The advent of the Internet has made cluster computing a powerful and cost-effective way to share and process data [8]. Autoclust can take advantage of this computing paradigm to not only speed up its execution time, but also produce an efficient data storage scheme where the query response time of the resulting database is faster than that of Autoclust running on a single node [1]. In response to clustering, the computing world is relying more and more on automated self-managing systems capable of making intelligent decision on their own. The area of autonomic computing has been getting a lot of attention. For Autoclust to be useful it should be fully automated, which implies that no human intervention is needed during the clustering process [7].

Auto clustering algorithm that is based on data mining techniques. The idea is to form clusters of attributes that correspond to closed item sets of attributes found in the queries. Preliminary tests results indicate that this algorithm returns an excellent quality solution in record time [1][3]. Auto Clustering can be adapted to fit a cluster computing environment with or without data nodes. In the case of data nodes we showed that a different clustering scheme could be implement on each data node, thereby providing an optimal query response time for any query to run from the database. [8][2]

The advantages of Auto cluster architecture are many, Different queries can be run in parallel against different nodes. Each query has a chance to get routed towards the node that would execute it in the most efficient manner [8]. Another advantage of redundant data nodes, when it comes to clustering is that data node can be clustered or re-clustered, and the data node still have the operation on that another node [1].

IV. Methodology

Web data extraction and annotation has been an active research area in recent years [1]. Many systems rely on human users to mark the desired information on sample pages and label the marked data at the same time, and then the system can find a series of rules to extract the same set of information's on web data from the same source. Bisecting k-means is reported outperforming k-means as well as the agglomerative approach in terms of accuracy and efficiency. In bisecting k-means, initially the whole data set is treated as a cluster [8]. Based on a rule, it selects a cluster to split into two by using the basic k-means algorithm. These bisecting steps are repeated until the desired number of clusters is obtained.

The goal of this function is to optimize different aspects of intra-cluster similarity, inter-cluster dissimilarity, and their combinations.. The function can be used in the family of k-means algorithms to assign each document to a cluster with the most similar cluster in an effort to maximize the intra-cluster similarity. There are some limitations with the k-means algorithm:

Bisecting K-means

Bisecting k-Means is like a combination of k-Means and hierarchical clustering. It starts with all objects in a single cluster.

The pseudo code of the algorithm is displayed below:

Basic Bisecting K-means Algorithm for finding K clusters

1. Pick a cluster to split.
2. Find two sub-clusters using the basic k-Means algorithm.
3. Repeat step 2, the bisecting step, for ITER times and take the split that produces the clustering with the highest overall similarity.
4. Repeat steps 1, 2 and 3 until the desired number of clusters is reached.

V. System Architecture

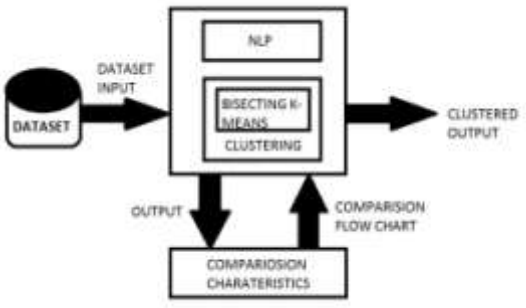


Fig. System Architecture

 ISSN: 2321-8169
4797 - 4802

Above architecture describe the complete scenario about dataset (text data) clustering. Basically dataset contains any file format like .doc, .docx, .txt, .csv, .rtf, .err, .sub, .log and many more. Input data is process for the NLP and annotated text categorization, and then get clustered according to the requirement. NLP formats text data which only utilized in meaningful way where unnecessary text get sorted out. K-means has limitations when clusters are of differing Sizes, Densities, Non-globular shapes, Problems with outliers, Empty clusters.

Bisecting K-MEANS apply better clustering to the formatted data due to some of advanced features over simple K-MEANS algorithm, and further process data for the characteristics comparison. Characteristics comparison compares output data to previous technology for data clustering, and provides the filtered characteristics with advancements in output.

The data units with the often share certain keywords. The data units corresponding to the search field where the user enters a search condition usually contain the search keywords. Sometime the user required data are put into label to get the exact data.

VI. IMPLEMENTATION OF CLUSTERING WITH NLP AND ANNOTATED TEXT CATEGORIZATION

(a) Screenshot For Annotation and Clustering.

Above screenshot contains procedure to input text for clustering. This phase contains four basic options which are Read Dataset, Apply NLP, Apply Clustering and Exit.

(b)Reading of Input Dataset For Clustering.

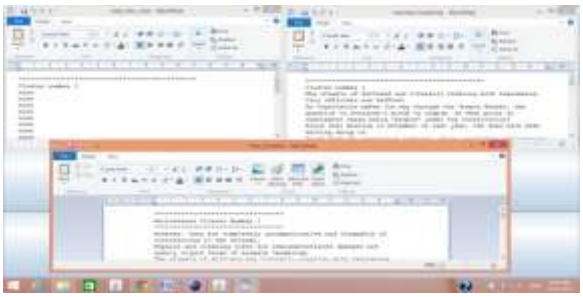
As per the first function (Read Dataset) used to provide input to the system. Input dataset contains data in text format. Different text files get supported like doc, docx, txt, log, err and many more.

(c)Applying NLP to dataset for language Processing.

NLP process text data into trash which has no effects on the original data, it also performs tagging, chunking, POS over input text data.

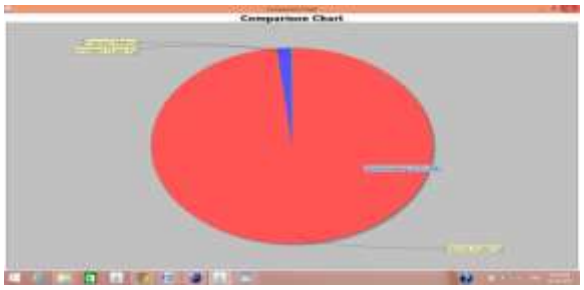
(d) Clustering the Dataset according to required cluster.

Number of cluster describe after completion of NLP process on data. User can provide number of cluster according to the requirement, where input text get clustered.



(e) Final Clustered document.

Resulted cluster file generated automatically after number of cluster given. It generates three output file every time each file provides more accuracy than previous one.



(f) Comparison Chart with previous Methodology.

Comparison chart provides performance report between normal clustering, proposed clustering, and annotated clustering. By using this chart it is get easy to understand which method is suitable for text clustering.

VII. CONCLUSION

We successfully implement the clustering algorithms used for the web mining to partition of the data. After studying all this methods, we create the better clustering format for the items in the large dataset. Autoclust concept can be useful in manner to manage the large dataset for better future use to user. Data clustering provides the better output for data in large dataset.

Current framework we proposed is used for the software requirement filed. In future this framework will also be used in any filed that used Natural languages.

VIII. POSSIBLE RESEARCH DIRECTION

By using the different methodology clustering of data can be improve in better manner to perform and process annotated text categorization. As NLP and annotation of text is vast in research field to take some of different advancements which improves data quality and reduce time consumption, detection of data discrepancy, deliberate error and data decay. Automated extraction of such information can be helpful in automated conceptual modeling of natural language software requirement specification.

ACKNOWLEDGEMENT

All the faculty members should be praised for contributing to the success of this survey in various ways. Also I want to thanks my guide to research of this topic with me on this survey and the references I have used throughout this project as well as the anonymous reviewers for their valuable comments.

IX. REFERENCES

[1] Semi-supervised Linear Discriminant Clustering, Chien-Liang Liu, Wen-Hoar, Chia-Hoang Lee, and Fu-Sheng Gou, IEEE transactions on cybernetics, vol. 44, no. 7, july 2014.

[2] An Onltology Based Text Mining Method To Develop D-Matrix From Unstructured Text,DnyaneshG. Rajpathak,andSatnam Singh, IEEE transactions on systems, man, and cybernetics: systems 2013.

[3] Data Mining With Big Data, Xindong Wu, Xingquan Zhu, Gong-Quing Wu, and Ding, IEEE transactions on Knowledge and Data engineering, vol. 26, no. 1, january 2014.

[4] Active Learning Constraints For Semi Supervised Clustering, Sicheng Xiong, Javad Azimi, and Xiaoli Z. Fern, IEEE transactions on knowledge and data engineering, vol. 26, no. 1, january 2014.

[5] On The Use Of Side Information for Mining Text Data, Charu C. Aggarwal, Yuchen Zhao, and Philip S. Yu, IEEE transactions on knowledge and data engineering, vol. 26, no. 6, june 2014.

[6] Annotating Search Result from Web Databases, Yiyao Lu, Hai He, Hongkun Zhao, WeiyiMeng, and Clement Yu, IEEE transactions on knowledge and data engineering, vol. 25, no. 3, march 2013.

[7] Data and Knowledge Engineering, CongnanLuo, YanjunLi,Soon M.Chung,Journal data k (jdatak), data & knowledge engineering 68 (2009)

[8] Using Cluster Computing to Support Automatic and Dynamic Database Clustering, Sylvian Guinepain, Le gruenwald, Third international Workshop on Automatic Performance Tuning IEEE 2008.

[9] Word Sense Disambiguation Based on Vicarious Words, Zhimao Lu,Dong, Mei Fan, Rubo Zhang, Fourth International Conference on Natural Computation, IEEE 2008.

[10] A Word Sense Disambiguation Approach for Converting Natural Language Text into a Common Semantic Description, Francisco Tacoa, Hiroshi Uchida, Mitsuru Ishizuka, 2010 IEEE Fourth International Conference on Semantic Computing.

[11] The Role of Apriori Algorithm for Finding the Association Rules in Data Mining, Jugendra Dongre, Gend Lal Prajapati, S.V. Tokekar, IEEE 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques.

[12] Big Data Application to the Vegetable Production and Distribution System, AlirezaAhrary, Dennis A. Ludena R, 2014 IEEE 10th International Colloquium on Signal Processing & its Applications.

[13] Analytical Study of Agent Based Distributed Data Mining and its Ontology, Bimalendu Pathak, Dr. Madhavi Sinha, 2014 International Conference on Computing for Sustainable Global Development.

[14] "Word Sense Disambiguation with Automatically Acquired Knowledge", Ping Chen, Wei Ding, Max Choly, Chris Bowes.

[15] Part-of-Speech Tagging by Latent Analogy, Jerome R. Bellegarda, IEEE journal of selected topics in signal processing, vol. 4, no. 6, December 2010.

BIOGRAPHIES

Prof. Shaikh Phiroj is a member of IAENG and IACSIT societies. He had completed his B.E. in Computer Technology and M. Tech in Computer Science & Engineering from Nagpur University, India. He has over 14 years teaching experience at undergraduate and post graduate level. His main research areas include Real time operating system, Embedded System, Machine Learning and Data Mining. Currently he is pursuing his PhD in Computer Science & Engineering from Nagpur University in the area of Web Data Mining.

Prof. Animesh R. Tayal received Bachelor of Engineering Degree in Computer technology from Nagpur University, and Master of Engineering Degree in Wireless Communication and Computing from G. H. Raisoni College of Engineering, Nagpur, India in 2002 and 2009 respectively. His research area is Natural Language Processing and Wireless Sensor Networks. He is having 10 years of teaching experience. Presently he is Assistant Professor in Priyadarshini College of Engineering, Nagpur. He is the author of fifteen research papers in International and National Journal, Conferences.

Mr. Amol V. Kale has received his B.E degree in Information Technology from Nagpur University, India. He is currently a postgraduate student of wireless communication and computing field with the department of Computer Technology, in Priyadarshini College of Engineering, from Nagpur University, Nagpur. His research interests include Wireless Communication, Data Mining, Natural Language Processing.