Volume 4, Issue 7, July 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Approach for Assigning of Proposals using Text Mining Techniques

Gurpreet Kaur*
Mtech,Research scholar
CSE,RIMT-IET, Mandigobindgarh,
India

Nidhi Bhatla
Assistant Professor
CSE,RIMT-IET, Mandigobindgarh,
India

Abstract— The management and extraction of useful information from large amount of data is very big challenge for the various work organisations etc. To find useful information from large database, data mining provides number of techniques. Text mining which is application of data mining provides clustering and classification technique for managing and extracting of important information. This paper presents a clustering and classification or decision tree algorithm for assigning of the proposal to the team according to the technology or area. The main aim of this paper is to provide an effective way of assigning proposal to team using data mining techniques and make it easy for the work organisations.

Keywords— Clustering, Classification, Decision Tree, Fuzzy-C means, CART algorithm.

I. INTRODUCTION

In today's World discovering patterns and trends from large databases becomes challenging issues as the amount of stored information has been enormously increasing day by day. The management or storage and extraction of useful information from unstructured data becomes a problem for many areas such as business, universities, research institutes, government funding agencies, and technology intensive companies. Data mining provides a solution for this problem.

Data mining emerged in 1980 for creating the useful information. Data mining is used to extract useful patterns and previously unknown trends from the large databases. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining has various techniques such as classification, clustering, decision trees, neural networks etc. The applications of data mining are text mining and web mining.

Text mining is the process of extracting important knowledge or patterns from the unstructured text that are from no. of sources. Text mining is a new and exciting research area that tries to take the challenge and produce the intelligence tool. The tool is a text mining system which has the capability to analyse large quantities of natural language text and detects lexical and linguistic usage patterns in an attempt to extract meaningful and useful information.

A. Clustering

Clustering is a unsupervised learning which provides a important role in a business environment. Clustering means that grouping of similar types of objects into one cluster. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Cluster Analysis is a process of grouping the objects, the objects can be physical like a customer or can be an abstract such as behaviour of a student, handwriting, hobbies. This process generates a group of objects called as a cluster/s. This cluster consists of the objects that are similar with each other in one cluster and dissimilar to the objects in other cluster. Thus the objective of clustering is exploratory in a nature to find a structure in dataset. The overall process of clustering is[1]:

1) Feature Selection or Extraction: It is the process of identifying the effective subsets of the original features so that they can be used in clustering, whereas feature extraction is the process of transforming one or more input features to produce new feature. Clustering process is mostly dependent on this step. If the feature selection is improper then it may increase the complexity and results into irrelevant clusters.

2) Clustering Algorithm Design or Selection: By applying domain knowledge, it is important to select an appropriate algorithm. Most of the algorithms are based on the input parameters, like number of clusters, termination condition, optimization/construction criterion, proximity measure etc. This different parameters and criteria are selected as a prerequisite of this step.

3) Cluster Validation: For clustering there is no appropriate algorithm because when we apply the different algorithms on the same dataset then they produces the different results and vice-versa. Therefore, it also becomes necessary to validate or evaluate the result produce by the clustering method. The evaluation criteria are categorized as:

- Internal indices: It evaluates the clusters by comparing it with the data only.
- External indices: By using the prior knowledge, it evaluates the clustering results, e.g. class labels.
- Relative indices: This criterion compares the results with various other results that are produced by the different algorithms.

4) Results Interpretation: This step deals with the representation of the clusters. Providing users with meaningful information from the original data is the main goal of clustering, so that, they can be effectively analysed and solved the problems.

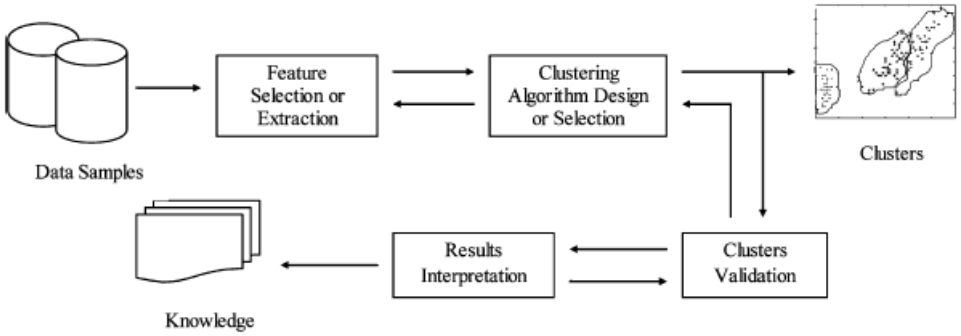

Fig.1 Process of clustering

B. Classification

In data mining, classification is supervised learning techniques in which classification rules are built by the classifier that is provided with the training data. Later if test data, is given to classifier, it will predict the values for unknown classes[3]. There is a widespread problem in the selection of the best classification algorithm for a given dataset. In this sense it requires to make several methodological choices.. The classification process groups the data into the classes on the basis of their differences. Some of the classification techniques or classifiers are the Neural Network Classifier , Decision Tree Classifier and Naïve Bayes Classifier and so on. Each of these techniques use the learning algorithm that generates the model that best fits the relationship between the predictors (attributes for prediction) and the prediction (class). The main motive of each of these techniques is to provide a model that accurately predicts the class of the unknown tuples records or tuples [16]. The steps given below represent the basic principle of working for each of these classifiers which is same [6]:

- The training set is provided that contains the training records along with their associated class label.
- By applying the learning algorithm the Classification model is built which is used in respective technique.
- On test set model built is applied that contains the tuples that do not have the associated class label.

Decision Tree algorithm is very useful and well known for the classification. It has an advantage of easy to understand the process of creating and displaying the results [4]. Decision tree is a predictive model that can be viewed as a tree where each branch of the tree is a classification question and leaves represent the partition of the data set with their classification. According to Investopedia, a Decision Tree is a schematic tree-shaped diagram used to determine a course of action or show a statistical probability [2]. Decision trees can be viewed from the business perspective as creating a segmentation of the original data set. Given a data set of attributes together with its classes, a decision tree produces sequences of rules that can be used to recognize the classes for decision making. The Decision tree method has gained popularity due to its high accuracy of classifying the data set [4]. Decision trees are trees that classify instances by sorting them based on feature values. In a decision tree each node represents a feature in an instance which is classified, and the value that the node can assumed is also represented by each branch. Classification of instances is starting from the root node and sorted on the basis of their feature values [5][17].

II. RELATED WORK

In Existing work, the text mining techniques clustering and classification are used with the concept of ontology. The tree like structure named ontology is constructed. It is created using the dataset of research proposals and reviewers. In previous work the classification and clustering of the proposals according to their area is done. The assignment of proposal to the reviewer is done manually as well as automatically. But the assignment of the proposals to the team is not done on the basis of the experience of reviewers and according to the proposals quotations like budget, time limit etc.

In [10], Preet Kaur et al, their work is for classification and clustering of research proposals and reviewers. Their main focus is on assigning the appropriate proposal group to the appropriate reviewer. The C4.5 decision tree algorithm is used for classification of research proposals and reviewers, and for clustering, k-means algorithm is used. In [11], N.Arunachalam et al, presents a framework on ontology based text mining to cluster external reviewers and research proposal on the basis of their research area and to assign concerned research proposals to reviewers systematically. Their paper has presented an framework on ontology based text mining for grouping research proposals and assigning the

grouped proposal to reviewers systematically. Research ontology is constructed to categorize the concept terms in different discipline areas and to form relationships between them. This facilitates text-mining and optimization techniques to cluster research proposals based on their similarities and then to assign them to reviewer according to their concerned area of research. The proposals in this are assigned to reviewer with the help of knowledge based agent. In [12], Prakash Pandey et al, presents a complete automatic ontology-based text-mining approach where one put paper and year of submission, then it automatically cluster research proposals based on their similarities in research areas. Hmway Hmway Tar et al [13], presents the concept of weight for text clustering system developed based on a k-means algorithm in accordance with the principles of ontology so that the important of words of a cluster can be identified by the weight values. Soumi Ghosh et al **[**14], their research work compares the two algorithms the k-means and Fuzzy C-mean algorithms on basis of efficiency of the output of clustering. Subhagata Chattopadhyay et al [9], their research also compares the two algorithms fuzzy c mean and entropy based fuzzy clustering (EFC) using four data sets IRIS, WINES, OLITOS, and psychosis. V. Sureka et al [15], compares the performance of enhanced ontological algorithms based on K-Means and DBScan clustering. The results showed that the efficiency of clustering and the performance of ontology-based DBScan algorithm is better than the ontology-based K-Means algorithm.

III. PROPOSED WORK

A. Objectives

- To provide an efficient and effective way for the assigning of research project proposals with the increasing number of research proposals and reviewers.
- To achieve appropriate assignment of proposal to the reviewer by using fuzzy C-means algorithm with decision tree algorithm CART on the basis of experience of team members and proposal quotation.
- To achieve the appropriate size of the team according to the proposals quotations.
- To analyze the results on the basis of parameters like accuracy, execution time, precision and recall.

B. Outline of Algorithm

The following algorithms will be used:

Fuzzy C-mean Algorithm

This algorithm is same as k-means algorithm because in this the value of C (number of cluster) has to be defined by the user. Fuzzy C-mean is a technique in which clustering is done by grouping datasets into n clusters. In this every data point belongs to every cluster with a high degree of belonging (connection) to that cluster and other which have low degree of belonging to that cluster lies far away from the centre of a cluster [7].

It is an approach, where the data points have their membership values with the cluster centers, which will be updated iteratively. The FCM algorithm consists of the following steps [9]:

Step 1: Let us suppose that M-dimensional N data points represented by xi (i = 1, 2, . . . ,N), are to be clustered.

Step 2: Assume the number of clusters that are to be made, that is, C, where $2 \leq C \leq N$.

Step 3: Select an appropriate level of cluster fuzziness $f > 1$.

Step 4: Initialize the $N \times C \times M$ sized membership matrix U, at random, such that $Uijm \in [0, 1]$ and $PCj=1\ Uijm = 1.0$, for each i and a fixed value of m.

Step 5: Determine the cluster centers CCjm, for jth cluster and its mth dimension by using the expression given below:

$$CCjm = \frac{\sum^{N}_{i=1} U^{f}_{ijm}\, x_{im}}{\sum^{N}_{i=1} U^{f}_{ijm}}$$

Step 6: Calculate the Euclidean distance between ith data point and jth cluster center with respect to, say mth dimension like the following:

$$Dijm = \| (xim - CCjm) \|$$

Step 7: Update fuzzy membership matrix U according to Dijm. If $Dijm > 0$, then

$$Uijm = \frac{1}{\sum^{C}_{c=1} \left(\frac{D_{ijm}}{D_{icm}} \right)^{\frac{2}{f-1}}}$$

If $Dijm = 0$, then the data point coincides with the corresponding data point of jth cluster center CCjm and it has the full membership value, that is, $Uijm = 1.0$.

Step 8: Step 5 to Step 7 repeat until the changes in $U \leq \varrho$, where ϱ is a pre-specified termination criterion.

Cart Algorithm

The CART addresses the classification and regression tree. In this, we build a binary decision tree on the basis of some splitting rule based on the predictor variables. We partitioned the space of predictor variables recursively according to the binary fashion. The partitioning is repeated until a node is reached where no further splitting is possible. A tree T has a root node whose descendant nodes, called daughters, further divided into two nodes i.e. terminal nodes and split nodes. The decision trees are built by collection of rules in the modelling data set on the basis of variables as[8]:

- Rules are selected on the basis of variable values to get the best split for differentiate observations on the basis of dependent variables.

- After selecting the rule the node is splits into two, and the same process is applied to each child node.
- Stop splitting when Cart detects that there can be no further gain made.

Each branch of the tree ends into a terminal node. Every observation comes under only one terminal node and every terminal node is uniquely defined by a set of rules.

IV. METHODOLOGY

The research methodology consists of data mining techniques such classification and clustering. In the research methodology firstly the database will be created that will be queried by user. For the classification of the proposals and team according to the technology the CART algorithm will be used. The clustering of proposal will be done by using the Fuzzy-C mean clustering algorithm. With the help of decision tree algorithm naming CART, the assignment of the proposal to the team according to the proposal quotations will be done. In last the analysis of the result will be done on the basis of different parameters.

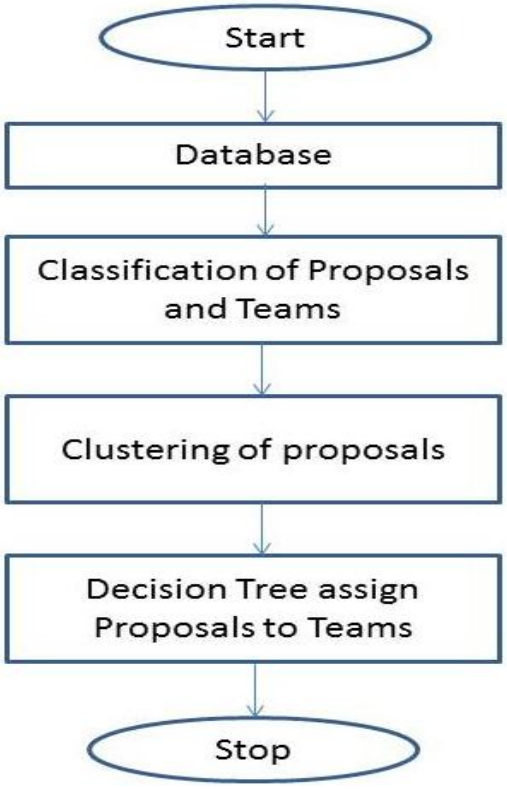

Fig. 2 Flow Diagram

V. CONCLUSIONS

In this paper the study of text mining techniques is done. The text mining, which is application of data mining, provides different methods such as classification, clustering etc. for extracting the important information from unstructured text documents or data. On the basis of study, concluded that for assigning of proposals to their relative teams will be done by using text mining techniques. In this paper methodology is proposed that will use fuzzy c means algorithm for clustering and CART algorithm for classification and decision tree. So, to assign proposals to their respective teams this will be an effective way by using text mining techniques and provide better results for various organizations.

ACKNOWLEDGMENT

The author would like to thank the RIMT Institutes, Mandi Gobindgarh-147301, Fatehgarh Sahib, Punjab, India. Author also extremely grateful and remain indebted to all the people who have given their intellectual support throughout the course of this work. And a special acknowledgement to the authors of various research papers and books which help me a lot.

REFERENCES

[1] Prof. Neha Soni, and Prof. Amit Ganatra, "*Categorization of Several Clustering Algorithms from Different Perspective: A Review*", IJARCSSE: International Journal of Advanced Research in Computer Science and Software Engineering (ISSN: 2277 128X), vol. 2, Issue 8, August 2012.

[2] Madhuri V. Joseph, Lipsa Sadath, Vanaja Rajan, "*Data Mining: A Comparative Study on Various Techniques and Methods*", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), ISSN: 2277 128X, Volume 3, Issue 2, February 2013, pp. 106-113.

[3] Mohd. Mahmood Ali, Mohd. S. Qaseem, Lakshmi Rajamani, A. Govardhan, "*Extracting Useful Rules Through Improved Decision Tree Induction Using Information Entropy*", International Journal of Information Sciences and Techniques (IJIST) Vol.3, No.1, January 2013.

[4] T.Miranda Lakshmi, A.Martin, R.Mumtaj Begum, Dr.V.Prasanna Venkatesan , "*An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data*", I.J.Modern Education and Computer Science, 2013, 5, 18-27 Published Online June 2013 in MECS.

[5] Thair Nu Phyu, "*Survey of Classification Techniques in Data Mining*", Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I,IMECS 2009, March 18 - 20, 2009, Hong Kong.

[6] Shabia Shabir Khan, Mushtaq Ahmed Peer, "*Evaluation of Knowledge Extraction Using Various Classification Data Mining Techniques*", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 3, Issue 6, June 2013.

[7] http://en.m.wikipedia.org/wiki/fuzzy_clustering.

[8] David G.T. Dension, Bani K. Mallick, Adrian F.M. Smith, "*A Bayesian CART Algorithm*", Biometrika, Vol. 85, No. 2 (Jun., 1998), 363-377.

[9] Subhagata Chattopadhyay, Dilip Kumar Pratihar and Sanjib Chandra De Sarkar, "*A Comparative Study Of Fuzzy C-Means Algorithm and Entropy Based Fuzzy Clustering Algorithms*", Computing and Informatics , Vol. 30, 2011, 701–720.

[10] Preet Kaur, Richa Sapra "Ontology Based Classification and Clustering of Research Proposal and External Research Reviewers" published in International Journal of Computers & Technology, Volume 5, No. 1, May - June, 2013, ISSN 2277-3061.

[11] N.Arunachalam, E.Sathya ,S.Hismath Begum and M.Uma Makeswari "An Ontology Based Text Mining Framework for R&D Project Selection" published in International Journal of Computer Science & Information Technology (IJCSIT) Vol 5, No 1, February 2013.

[12] Jay Prakash Pandey, Shrikant Lade, Manish Kumar Suman "Automatic Ontology Creation for Research paper classification" published in International Journal of Engineering Research and Science & Technology, Vol. 2, No. 4, November 2013.

[13] Hmway Hmway Tar, Thi Thi Soe Nyunt "Ontology-Based Concept Weighting for Text Documents", International Conference on Information Communication and Management IPCSIT vol.16 2011 IACSIT Press, Singapore.

[14] Soumi Ghosh and Sanjay Kumar Dubey "*Comparative Analysis of K-means and Fuzzy C-Means Algorithms*", International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.

[15] V.Sureka, S.C.Punitha **"**Approaches to Ontology Based Algorithms for Clustering Text Documents **"** published in Int.J.Computer Technology & Applications, Vol 3 (5), 1813-1817 .

[16] Prof. Saurabh Tandel, Prof. Vimal Vaghela, Dr. Nilesh Modi , Dr. Kalpesh Vandra , "*Multi Relational Data Mining Classification Processions – A Survey",* Int.J.Comp.Tech.Appl,Vol 2 (6), 3097-4002, ISSN:2229-6093.

[17] Suban Ravichandran, Vijay Bhanu Srinivasan and Chandrasekaran Ramasamy, "*Comparative Study on Decision Tree Techniques for Mobile Call Detail Record",* Journal of Communication and Computer 9 (2012) 1331-1335.

