International Journal of Data Mining Techniques and Applications
Volume: 04 Issue: 02 December 2015, Page No.88-91
ISSN: 2278-2419

A New Arithmetic Encoding Algorithm Approach for Text Clustering

Nikhil Pawar[1], P.K Deshmukh[2]
[1]M.E Second Year Student, [2]Professor,
Department of Computer Engineering, JSPM's Rajarshi shahu College of Engineering Savitribai phule Pune University, India
Email: nikhilhpawar@gmail.com, pkdeshmukh9@gmail.com

Abstract- In this paper we propose a new method for improving the clustering accuracy of text data. Our method encodes the string values of a dataset using Arithmetic encoding algorithm, and declares these attributes as integer in the clustering phase. In the experimental part, we calculate the efficiency of proposed method, and we obtained a better clustering accuracy than the one found with traditional methods. This method is useful when the dataset to be clustered has only string attributes, because in this case, a traditional clustering method does not recognize, or recognize with a low accuracy, the category of instances.

Keywords - Accuracy, cluster analysis, Huffman encoding, Arithmetic encoding, machine learning, text mining.

I. INTRODUCTION

Text data mining is the process of retrieving information from text. Text data mining is roughly equivalent to text analytics which refers to High-quality information. It is typically derived through the patterns and trends through methods such as statistical pattern learning. Text Mining is the discovery of unknown information from different data resources which has become an important area in the researches. The process of extracting interesting and non-trivial information and knowledge from unstructured text is known as Text mining. It is a field which is derived from information retrieval systems, computational linguistics data mining systems, as well as machine learning. On an average, about 80% of information is stored as text and thus text mining has a high commercial potential value. From many of the information sources from which knowledge can be derived; still, unstructured text is the source of knowledge. There are many problems in Knowledge Discovery from Text (KDT) like extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. But it increases efficiency in large quantities of text data. KDT, is rooted in NLP, draws on methods from statistics, information extraction, knowledge management, machine learning, reasoning, and others for its discovery process.

KDT has an important role in emerging applications like Text Understanding. There are many Text Mining operations like: Text-based navigation, Summarization, Feature Extraction, Search and Retrieval, Categorization (Supervised Classification) and Clustering (Unsupervised Classification). Like Data mining is about searching for patterns in data, in same way text mining is about finding patterns in text. It is the process of analyzing text to retrieve information that is useful for particular purposes. Besides from analysis step, it involves data pre-processing, interestingness metrics, database and data management aspects, model and inference Considerations, complexity considerations, implementing visualization, post-processing structures, and online updating. Arithmetic coding is a form of entropy encoding used in lossless data compression [13]. Generally, the words are represented using a fixed number of bits per character, similar to the ASCII code. A string is converted to arithmetic encoding, then the frequently used characters will be stored with fewer bits and not-so-frequently occurring characters will be stored with more bits. Therefore it results in fewer bits used in total. Arithmetic coding is different from the Huffman coding in which the input is converted into component symbols and replaced by a code, whereas arithmetic coding encodes the entire input into a single number [13]. Texts need to be compressed with lossless compression algorithms because a loss in a text will change its originality. Even the repeated data is important in text compression which can be compressed to a high ratio. This is possible due to the compression algorithms which generally eliminate repeated data. The comparison between arithmetic and Huffman coding algorithms for different text files with different capacities has been performed. The compression ratio of the arithmetic coding for text files is better than Huffman coding [10].

This paper involves the following sections which are divided as further: Section II is about related work studied till now. Section III present Issues, section IV present some techniques used for clustering, Section V presents implementation details, algorithms used, mathematical model and experimental setup tended to by this paper. Section IV depicts results and discussion part. Section V draws conclusions and presents future work.

II. RELATED WORK

Previously, the Huffman encoding algorithm [12] was used to encode text data [5]. The Huffman coding is an encoding algorithm which is used for lossless data compression. It uses a variable length code table for encoding a character symbol in a file where the variable length code table has been derived in a specific manner based on the estimated probability of occurrence for each possible value of the source character symbol. Huffman coding is based on the principle to use a lower number of bits to encode the data that occurs more frequently. The average length of a Huffman code depends on the statistical frequency with which the source produces each symbol from its alphabet. A Huffman code dictionary, which associates each data symbol with a codeword, has the property that no codeword in the dictionary is a prefix of any other codeword in the dictionary. Huffman encoding is a way to assign binary codes to symbols that reduces the overall number of bits used to encode a typical string of those symbols. This Huffman encoding method is used in that method for encoding

string data. This is new method for improving the clustering accuracy of text data. This method encodes the string values of a dataset using Huffman encoding algorithm, and declares these attributes as integer in the cluster evaluation phase Demonstrated that this method clusters with a higher accuracy the instances of such a dataset [8]. The clustering methods have generally focused on the case of quantitative data, in which the attributes of the data are numeric. The problem has also been studied for the case of categorical data, in which the attributes may take on nominal values. There are also approaches on improving the clustering of text data streams.

In the paper, we present a method for massive-domain clustering of data streams. The results obtained that a sketch-based clustering method can provide similar results to an infinite space clustering algorithm with high probability. We focus on view points and measures in hierarchical clustering. The research is particularly focused in studying and making use of cluster overlapping.

III. ISSUSES

When the dataset to be clustered has only string attributes, because in this case, a traditional clustering method does not recognize, or recognize with a low accuracy, the category of instances. Our proposal is to improve the clustering accuracy of a text dataset by encoding the instances with Arithmetic encoding algorithm. After finding the instances codes, we declared the attributes of the dataset of numeric type and we applied a hierarchical clustering method in order to discover the categories of data.

IV. TEXT CLUSTERING TECHNIQUES

Clustering techniques apply when there is no class to be predicted but we want to divide instances in to natural groups. These clusters give signs of some mechanism that is at work in the domain from which instances are outlined, a mechanism that causes some instances to take a stronger likeness to each other than they do to the still in the same way examples. [1] Clustering naturally has need of different techniques to the classification and association learning methods. With the popularity of Internet and great-scale getting better in the level of undertaking information, the bursting substance growth of useable things, the research of text mining, information filtering and information search. So, the cluster technology is becoming the core of text information mining technologies. The main objective of clustering is to partition unlabelled patterns into homogeneous clusters. Clustering algorithm can be divided into the following categories: hierarchical clustering partitioned clustering, density-based algorithm, self organizing maps algorithm [2]. At the same time, the text clustering problem has its particularity.

On one hand, the text vector is a high-dimensional vector, usually thousands or even ten thousands; On the other hand, the text vector is usually sparse vector, so it is difficult for the choice of cluster centre. As an unsupervised machine learning method, because of not need to train the process and manual label document at category in advance, clustering has certain flexibility and high automation handling ability. It is become an important mean which pays attention for more and more researchers. The purpose of text clustering is large-scale text data sets which can be grouped into several categories, and made between the text information in the same class which has high similarity, rather than the difference of text between the different types. There are many clustering techniques used for clustering text such as: Hierarchical clustering, Partitioned clustering, Density-based algorithm, and Organizing Maps algorithm. In this paper we focus on Hierarchical clustering to improve clustering efficiency [3] [11].

Text clustering is a typical problem of unsupervised machine learning. Hierarchical clustering algorithm by combining the appropriate similarity measure similarity such as cosine similarity, Dice coefficient, Jaccard similarity coefficient, has become the mainstream technology on the document clustering. Hierarchical clustering is commonly text clustering method, which can generate hierarchical nested class. Hierarchical clustering method takes category as hierarchical, in other words, with the change of category hierarchical, object also corresponding change. This method allows classifying data at different granularity. In accordance with generation methods of the category tree, hierarchical clustering method can be divided into two categories, one kind is integration method (bottom-up method), and the other kind is to split methods (top-down method).Hierarchical clustering accuracy is relatively high, but when each class merges, it needs to compare all classes' similarity in the global and selecting the most similar of two classes, so it's relatively slow [7]. The defect of hierarchical clustering is that once a step (merge or split) completed, it cannot be revoked, so it can not correct the wrong decision. Hierarchical clustering methods are generally divided into bottom-up hierarchical clustering method and top-down hierarchical clustering method.

Bottom-up (merge) hierarchical clustering method starts from a single object, first takes an object as a separate category, and then repeatedly merges two or more appropriate categories, until meeting stop conditions. Top-d0own (splitting) hierarchical clustering method starts from the objects complete works, and gradually be divided into more categories. The typical approach is to construct a minimum spanning tree on similar graphs, and then at each step choosing a side which in the smallest similarity of the spanning tree (or in the farthest distance of the spanning tree) and removing it. If it deletes one side, it can create a new category [4]. When the smallest similarity achieves some threshold value the cluster may stop. In general, the amount of computation of top-down method is greater than the bottom-up method, and the applications of top-down method is inferior widespread than the latter.

V. PROBLEM STATEMENT

There should be a mechanism to improve clustering efficiency; to improve clustering accuracy of text data we can use a new text clustering method based on Arithmetic encoding algorithm. The main concept behind this study is, It is observed that when the dataset to be clustered has only string attributes, a traditional clustering method does not recognize, or recognize with a low accuracy, When we first convert it in to integer then clustering is perform well, the category of instances and it is Demonstrated that this method clusters with a higher accuracy the instances of such a dataset.

International Journal of Data Mining Techniques and Applications
Volume: 04 Issue: 02 December 2015, Page No.88-91
ISSN: 2278-2419

A. System Architecture :

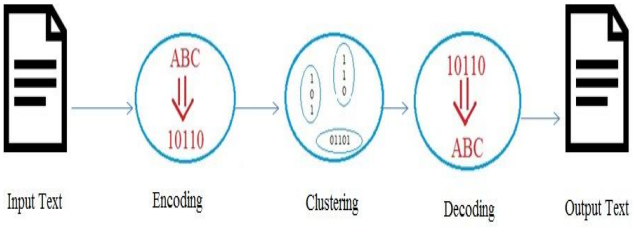

Figure 1: propose system Architecture

- Input : User Provide input as text instances
- Process Mechanism : Encoding text data in to integer then clustering is perform on integer attributes, to get text in original form decoding is done in next step
- Output: we have clusters to perform text mining.

B. Mathematical model:

Let the system S is represented as: S = {T,E,C,D}

1. Input data set Consider, T is a set for learning dataset

T = {*t1, t2, t3, . . . ,tn*} Where *t1,t2. . . tn* are the input data

2. Encoding Phase Let E be the set for Encoding phase E = {*e1, e2, e3, . . . ,en*} Where, *e1,e2,e3,. . . en* are the Encoded data.

3. Clustering Phase Let C is a set for Clustering C = {*c1, c2, c3,. . . .,cn*} Where, *c1,c2,c3. . . cn* are the clusters after Clustering step.

Max{d(x,y): x ϵ X, y ϵ Y}

Min{d(x,y): x ϵ X, y ϵ Y}

$$\frac{1}{|X|.|Y|}\sum_{x\in X}\sum_{y\in Y} d(x,y)$$

4. Decoding Phase Let D is set for decoding Phase D = {*d1,d2,d3,. . . dn*} Where, *d1,d2,. . . dn* are the Decoded data .

C. Algorithm :

1. Read and validate input
2. for (all the instances of the dataset)
{
For (all the attribute values of an instance)
{
Read 0an attribute value
If (value = string type)
{
Encode the value with Arithmetic encoding algorithm
Delete the value from the dataset and write the Arithmetic code obtained in the previous step
}
}
}
3. Replace the string type of the attributed with integer type in the attribute declaration
4. Evaluate the dataset
5. Label the instances with the cluster name
6. for (all the instances of the dataset)
{
For (all the attribute values of an instance)
{
Read an attribute value
If (value = integer type)
{
Decode the value with Arithmetic decoding algorithm
Delete the Arithmetic code from the dataset and write the string value obtained in the previous step
} } }

VI. RESULTS

A. Data set

Here Text Mining is performed on two datasets named Physics dataset and the Biology dataset; these datasets are obtained from the online UCI Machine Learning Repository. UCI is the Centre for Machine Learning and Intelligent Systems. It holds a repository of datasets which are used by practitioners and researchers in the fields of Artificial Intelligence, Pattern Recognition, Machine Learning, Neural Networks, Data Mining, Bio-informatics and others these are referred to as the UCI datasets. Both dataset consist of string attributes.

B. Results :

Table 1 presents the clustering results of the datasets with Physics and biology data sets. Our proposed method performs clustering with great efficiency, after observing these reading we can estimate clustering accuracy as 84.6%.

TABLE I: The Clustering Results

Data set	*Cluster 0*	*Cluster 1*
Physics	61	331
Biology	509	99

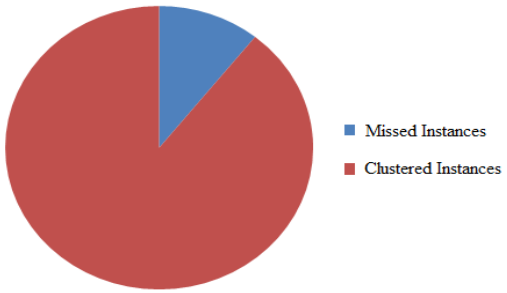

Figure 3: biology Cluster

International Journal of Data Mining Techniques and Applications
Volume: 04 Issue: 02 December 2015, Page No.88-91
ISSN: 2278-2419

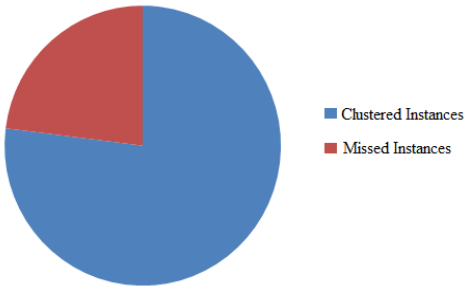

Figure 4: Physics Cluster

We take input of 1000 instances of physics and biology datasets; we performed experiment on these 1000 instances, from experiment, cluster 0 represents Biology instances and cluster 1 represent Physics instances. From experiment we get 509 instances in biology cluster and 331 instances in Physics cluster, so we get 84.6% accurate clustering with these datasets. We get 840 instances correctly clustered. Figure 3 shows biology cluster in which we get 509 instances clustered correctly and Figure 4 shows Physics Cluster in which we get 331 instances correctly clustered, We observe a huge improvement in the first experiment, where we used the initial dataset for the evaluation.

VII. CONCLUSION

This is very effective technique to improve clustering accuracy of text data, it has been observed that traditional clustering methods not perform well on string attributes to improve clustering accuracy, Compression ratio of Arithmetic encoding is better than Huffman encoding. So Arithmetic encoding algorithm is used to encode data instances in to integer, clustering performed on integer instances is much more effective than clustering performed on string instances, here additional time require in encoding and decoding phase but this time is covered in clustering phase.

ACKNOWLEDGEMENT

The authors would like to thank the Department of Computer Engineering of JSPM's Rajarshi Shahu College of Engineering, as well as researchers for making their resources available and teachers for their guidance. We are thankful to the authorities Board of Studies Computer Engineering of Savitribai Phule Pune University for their constant guidelines and support. We are also thankful to reviewer for their valuable suggestions. We also thank the college authorities for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

REFERENCES

[1] B., Zheng, J., Chen, S., Xia, Y., Jin, "Data Analysis of Vessel Traffic Flow Using Clustering Algorithms", 2008 International Conference on Intelligent Computation Technology and Automation, Changsha, Hunan, China, pp. 243 – 246, 2008.

[2]M., Moslem, A., Hosein, and M.-B., Behrouz, "Neural Network nsembles using Clustering Ensemble and Genetic Algorithm", Third 2008 International Conference on Convergence and Hybrid Information Techology, Busan, South Korea, pp. 1924-1929, 2008.

[3] N., RaghavaRao, K., Sravankumar, P., Madhu, "A Survey On Document Clustering With Hierarchical Methods And Similarity Measures", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 7, ISSN: 2278-0181, pp. 1-7, 2012.

[4] C., C., Aggarwal, C., X., Zhai, Mining Text Data, chapter: A Survey of Text Clustering Algorithms, Springer US Publisher, Print ISBN 978-1-4614-3222-7, Online ISBN 978-1-4614-3223-4, pp. 77-128, 2012.

[5] P., R., Suri, and M., Goel, "Ternary Tree and Clustering Based Huffman Coding Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, ISSN (Online): 1694-0814, 2010.

[6] http://www.cs.waikato.ac.nz/ml/weka/

[7] N., RaghavaRao, K., Sravankumar, P., Madhu, "A Survey On Document Clustering With Hierarchical Methods And Similarity Measures", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 7, ISSN: 2278-0181, pp. 1-7, 2012.

[8] Y., B., Liu, J., R., Cai, J., Yin, A., Wai-Chee Fu, "Clustering text data streams", Journal of Computer Science and Technology, 23(1), pp. 112-128, 2008.

[9] P., R., Suri, and M., Goel, "Ternary Tree and Clustering Based Huffman Coding Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, ISSN (Online): 1694-0814, 2010.

[10] J., Dvorsky, J., Martinovic, J., Platos, "Using The Clustering for Improve of WLZ77 Compression", First International Conference on the Applications of Digital Information and Web Technologies, ICADIWT 2008, Ostrava, E-ISBN: 978-1-4244-2624-9, Print ISBN: 978-1-4244-2623-2, pp. 308 – 313, 2008.

[11] C., C., Aggarwal, C., X., Zhai, Mining Text Data, chapter: A Survey of Text Clustering Algorithms, Springer US Publisher, Print ISBN 978-1-4614-3222-7, Online ISBN 978-1-4614-3223-4, pp. 77-128, 2012.

[12] http://rosettacode.org/wiki/Huffman_coding

[13] Rissanen, J. J. and Langdon, G. G.: 'Arithmetic Coding'. IBM Journal of Research and Development, 23(2):146–162, March 1979.