International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Effective Text Clustering Method Based on Huffman Encoding Algorithm

Nikhil Pawar

Department of Computer Engineering, Savitribai Phule Pune University, India

Abstract: ***Huffman 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. 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.***

Keywords: accuracy, cluster analysis, huffman encoding, machine learning, text mining.

1. Introduction

Text Mining has become an important research area. Text Mining is the discovery of unknown information from different data resources. Text mining is a process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a field which draws on information retrieval systems, data mining systems, computational linguistics, machine learning. As over 80% information is stored as text, text mining has a high commercial potential value. There are many sources of information from which knowledge may be discovered; yet, unstructured text is the source of knowledge. The problem of extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques is faced in Knowledge Discovery from Text (KDT). 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 plays an important role in emerging applications, such as Text Understanding. Most of Text Mining operations are as follows: Feature Extraction, Text-based navigation, Search and Retrieval, Categorization (Supervised Classification), Clustering (Unsupervised Classification), Summarization.

Primary objective of the Feature Extraction operation is to identify facts and relations in text. Most of the times this includes distinguishing which noun phrase is a person, place, organization or other distinct object. Feature Extraction algorithms may use dictionaries to identify some terms and linguistic patterns to detect others Text mining is very important in many applications such as Publishing and media, Telecommunications, energy and other services industries, Information technology sector and Internet, Banks, insurance and financial markets, Political institutions, legal documents, Pharmaceutical and research companies and healthcare.

Data mining is about looking for patterns in data. In same way text mining is about finding patterns in text, it is the process of analyzing text to extract information that is useful for particular purposes. Aside from the raw analysis step, it involves interestingness metrics, data pre-processing, database and data management aspects, model and inference considerations, complexity considerations, post-processing structures, implementing visualization, and online updating.

2. Text Clustering Techniques

Clustering techniques apply when there is no class to be predicted but the instances are to be divided into natural groups. These clusters give signs of some mechanism that is at work in the domain from which instances are outlined, an mechanism that causes some instances to take a stronger likeness to each other than they do to the still in the same way examples. 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. 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 center. 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. 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. 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-down (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. 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.

3. Importance of Topic

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 Huffman 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.

4. Huffman Encoding

In this method, Huffman encoding algorithm is used to encode text data. In computer science and information theory, Huffman coding is an entropy encoding algorithm used for lossless data compression. The term refers to the use of a variable length code table for encoding a source symbol (such as a character in a file) where the variable length code table has been derived in a particular way based on the estimated probability of occurrence for each possible value of the source symbol. Huffman coding is based on frequency of occurrence of a data item. The principle is 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 this 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. 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. 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.

Phenomenon to design cluster merging criteria. We extended the semantic smoothing model into text data streams context firstly. After finding the instances codes, Attributes of dataset are declared as of numeric type and hierarchical clustering method is applied in order to discover the categories of data. As far as we know, the Huffman algorithm was so far improved by clustering methods in order to obtain more accurate data compression, but wasn't used as a part of a clustering algorithm in order to improve the data mining results. It is known that the Hierarchical cluster algorithm works better with numeric attributes, than using string attributes. This is the reason they encoded the string values using Huffman encoding algorithm, and declared these attributes as integer type in the cluster evaluation phase.

5. Effectiveness for this Method

Here Text Mining is performed on two datasets named SMS Spam Collection dataset and the Reuters Grain dataset; these datasets are obtained from the online UCI Machine Learning Repository. UCI is the Center 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.

In the first step of Clustering experiment on both datasets, we can observe that 19% instances of SMS Spam Collection dataset are clustered in class 0, and the remaining 81% instances are clustered in class 1, in Reuters Grain dataset cluster distribution class 0 consist of 20% instances and class 1 consist of 80% instances. In the second step, Huffman encoding algorithm applied on the considered dataset to generate integer code of each instances. In the third experiment Weka Machine Learning Tool is used to implement the clustering experiments. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. We chose to analyze the experiments with the Hierarchical clustering algorithm because this is the most developed clustering algorithm that works with string data.

After this experiment we can conclude that, hierarchical clustering is fail to distinguish between dataset categories, result shows that hierarchical clustering not make even distribution of instances over clusters , that is only cluster 0 consist of all instances and cluster 1 consist none. And after applying Huffman codes Hierarchical clustering is improved and it is able to distinguish between dataset categories.

It is clear that Huffman encoding method is add-on method to clustering text data, so additional time is require to perform encoding and decoding the data, but this time is gained in the clustering evaluation phase, because the learning method performed faster on numeric data, than it initially performed on string data.

6. 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, Huffman 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 these time is covered in clustering phase. And this method still has room for improvement.

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/