www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 11 November, 2014 Page No. 9211-9214

An approach to Re-evaluate CERS method for News Mining

Swati Vashisht[1], Dr. V. Bibhu[2], Tushar sharma[3], Smratika Sharma[4]

[1]Assistant Professor

Amity University Gr Noida

svashisht@gn.amity.edu

[2]Assistant Professor

Amity University Gr Noida

drvimal@gn.amity.edu

[3]Scholar B. Tech.(CSE)

Amity University Gr Noida

tusharsharma276@gmail.com

[4]Scholar B. Tech.(CSE)

Amity University Gr Noida

smrattika.sharma@gmail.com

Abstract: ***With the current growth rate of URLs we are at the age of online information overload and for many other domains such as web services data analysis. Text mining has been a key research topic for online information retrieval and information extraction. From online news and blog articles a human can often deduce information and knowledge for the prediction of market movements and other interesting activities occurring all around the world. However this recognition and comprehension process is very complex and requires some context knowledge about the domain in which trends are to detect.***

The analysis of news source represents an important challenge of our times. News not only reflects the different processes happening in the world but also influences the economic, political and social situation. A news source contains an enormous amount of information which can be compiled together and analyzed.

In this paper we proposed an approach that applies clustering methods on news articles and then CERS (Cross Entropy Reduction Sampling) technique to make a news article more effective to search and less cumbersome to get exact knowledge.

Keywords: News mining, information extraction, document clustering, news articles, CERS method.

1. Introduction

Text mining or text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, and concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling.

Text mining involves analyzing a large collection of documents to discover previously unknown information. The information might be relationships or patterns that are buried in the document collection and which would otherwise be extremely difficult, if not impossible, to discover. Text mining can be used to analyze natural language documents about any subject, although much of the interest at present is coming from the biological sciences [1]. Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining.

1.1 Information Extraction

Information Extraction (IE) is an important process in the field of Natural Language Processing (NLP) in which factual structured data is obtained from an unstructured natural language document. Often this involves defining the general form of the information that we are interested in as one or more templates, which are then used to guide the further extraction process. IE systems rely heavily on the data generated by NLP systems. Tasks that IE systems can perform include:

Term analysis: This identifies one or more words called terms, appearing in the documents. This can be helpful in extracting information from the large documents like research papers which contain complex multi –word terms.

Named-entity recognition: This identifies the textual information in a document relating the names of people, places, organizations, products and so on.

Fact extraction: This identifies and extracts complex facts from documents .Such facts could be relationships between entities or events.

1.2 Information retrieval

The classic approach of information retrieval based on keyword search makes it cumbersome for the users to look up for the exact and precise information from the search results. It is up to the user to go through each document to extract the relevant and necessary information from those search results. This is an impractical task. Text mining can be a better solution for this as it links all the extracted information together, pushes all the irrelevant information aside, and keep the relevant ones based on the question of interest.[3]

1.3 Clustering

Clustering is defined as a technique used to place data elements into related groups. There are various methods for clustering but K-means is one of the most efficient methods. From the given set of n data, k different clusters; each cluster characterized with a unique centroid (mean) is partitioned using the K-means algorithm. The elements belonging to one cluster are close to the centroid of that particular cluster and dissimilar to the elements belonging to the other cluster. Here we are discussing about K-mean clustering method. The letter “k” in the K-means algorithm refers to the number of groups we want to assign in the given dataset. If “n” objects have to be grouped into “k” clusters, k clusters centers have to be initialized. Each object is then assigned to its closest cluster center and the center of the cluster is updated until the state of no change in each cluster center is reached.

1.4 CERS method

It introduces a compression based approach for identifying a subset of representative and diverse documents (typically large) from a document collection (representative sampling of documents).

As input parameters, the algorithm accepts the sequence dataset D, the number m of representative sequences which should be extracted from D, as well as parameters required to compute the cross entropy reduction ΔH (x,D,S). The latter include the maximum order d of tree source models included in the mixture computation, the zero-order estimator function pe (.), and the CTW α parameter, which controls the prior over context tree structures. The result of the algorithm is a sequence set S, S ε D, |S|=m, containing representative sequences from the input dataset D.

```
CERS Algorithm

Input: Sequence dataset D={x; x ε ∑*}

Input: Desired sample size m>0

Input: Parameters required to compute ΔH: the model order d, zero order estimator routine pe(.) , and the CTW weighting parameter α.

Output: Representative set S of size m.

S←Ø;
/*Initialization*/

While |S|<m do

    S← arg $\max_{x\varepsilon D}$[ΔH(x,D,S)];
    /*Select representative point s*/

    S←S U{s};
    /Add s to S*/

    D←D\{s};
    /*Remove S from D*/

End
```

The objective function for selecting representative examples could also be defined as

$$\Delta H(x, D,S) = H(D,S) – H(D\backslash \{x\}, SU\{x\}),$$

i.e. the reduction in the average code length for D after x is moved to S. However this formulation is prone to selecting outliers, since moving outliers from D may improve compression of D only due to de-noising of D, and not due to improvements in the modeling of other data points in D.

2. Problem Formulation

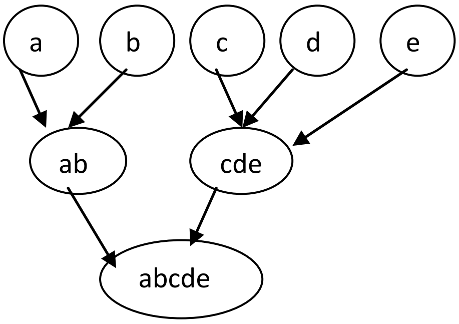
As we have studied about text mining with compression techniques, different approaches have been used to compress text according to the requirements of a particular application. In our research, we have chosen a specific application of text mining i.e. news mining for further assessment and analysis.

The CERS method was designed to extract a sample of representative and diverse data points from a large dataset. Such representative points should lie in dense areas which correspond to cluster centers. Data points selected by CERS method is a good set of centroids of clustering methods.

3. Proposed Work

In this paper we proposed an approach to re-evaluate CERS (Cross Entropy Reduction sampling) method. This method uses a compression based approach to find the subsets of the given sample data from the large database. If we implement clustering before CERS, then it would become easier to sample homogeneous dataset and to select representative data from a large dataset. K-mean clustering is applied before sampling method but the initial choice of seeds affects the quality of k mean clustering especially in the case of document clustering. A number of techniques are used in order to improve the quality of initial seeds which are picked for the initial process. Agglomerative clustering can be used to decide the initial seeds for clustering. We are using it only for initial clusters because the complexity of agglomerative clustering is O (n^3) which makes them too slow for large datasets.

Each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy. Here in this example we have taken five different datasets and based on their attributes and features, we have constructed two clusters i.e. political issues and social issues. These two can be initial seeds for further group of similar data.



After these two initial clusters, K-mean algorithm and CERS method is being applied for further processing.

4. Conclusion

The main advantage of compression algorithm is: It can improve detection function because compressed data will take less time for matching a pattern. K-mean clustering would be more efficient in determining similarity between different but related words in text. Response time will increase but it would be quite effective to extract knowledge.

The major advantage of clustering is fast processing time because it is dependent only on the number of cells in each dimension in the quantized space. Compressed data can become a little complex as compared to original data.

5. Future Scope

Area of future research includes experimental evaluation of CERS technique with clustering methods to calculate its response time, different compression methods for a large data set that will help to decrease response time, different mining methods that will help to extract sufficient and appropriate information from a huge collection of data. Future scope also includes analysis of performance of different clustering and classification methods.

6. References

[1] Bratko Andrej,"Text mining using data compression models"

[2] Sayood Khalid,"Introduction to data compression"

[3] Prabin Lama,"Clustering system based on text mining using the K-means algorithm"

[4] Inna Novalija,"Ontology extension using text mining for news analysis"

[5] Olga streibel, "Mining trends in texts on the web"

[6] Apirak Hoonlor, "Sequential patterns and temporal patterns for text mining"

[7] Shaldah Jusoh and Hejab M. Alfawareh, "Techniques, applications and challenging issues in text mining "

[8] Daggumalli sailaja and K.John Paul, "Text mining Detection in Effective model"

[9] V. Barnett and T. Lewis. Outliers in Statistical John Wiley & Sons, 1994.

[10] Zengyou He , XiaofeiXu, Shengchun Deng "A Fast Greedy Algorithm for Outlier Mining"

7. Author Profile

1. Swati Vashisht has received B.tech. degree in Information Technology & M.Tech. degree in Computer Science & Engineering. Her area of interest is data mining, data compression & digital circuits.

2. Dr. Vimal Bibhu has received his Ph.D. degree in Information Technology & B.tech. in Computer Science & Engineering. He has published many research papers in various national and international journals.

3. Tushar Sharma is pursuing B.Tech. in Computer Science & Engineering from UPTU. His area of interest is data mining & data strucures.

4. Smratika Sharma is pursuing B.Tech. in Computer Science & Engineering from UPTU. Her area of interest is data mining & data compression.