DATA MINING TECHNIQUE FOR OPINION RETRIEVAL IN HEALTHCARE SYSTEM

A.Ananda Shankar[1] and Dr.K.R.Ananda Kumar[2]

[1]Associate Professor in Dept. of CSE, Reva University Bangalore, India
[2]Professor and HOD Dept. of CSE, S.J.B.I.T, Bangalore, India

ABSTRACT

The aim of this paper is to use Text mining(TM) concepts in the field of Health care System. We no that now days decision making in health care involves number of opinions given by the group of medical experts for specific disease in the form of decisions which will be presented in medical database in the form of text. These decisions are then mined from database with the help of Data Mining techniques. Text document clustering is considered as tool for performing information based operations. For clustering normally K-means clustering technique is used. In this paper we use Bisecting K-means clustering technique and it is better compared to traditional K-means technique. The objective is to study the revealed groupings of similar opinion-types associated with the likelihood of physicians and medical experts.

KEYWORDS

Text Mining(TM), Opinion Mining (OM), Sentimental Analysis (SA), Data Mining (DM).

1. INTRODUCTION

Healthcare related data mining(DM) is one of the most rewarding and challenging areas of application in data mining and knowledge discovery. The challenges are due to the datasets which are large, complex, heterogeneous, hierarchical, time series and varying of quality. As the available healthcare datasets are fragmented and distributed in nature, thereby making the process of data integration is a highly challenging task.[1]

Classifying opinions into groups is a new phenomenon. In fact, it can be said that the idea is based on the notion of a search for a natural ordering of things, which is a basic characteristic of human beings [2]. Fairly recent additions to this concept, however, are 1) the wide-scale application of clustering and classification techniques to intra- and inter institutionally for determining medical resource utilization and 2) the growing importance being attached to the reliability and validity aspects of classification procedures and the resulting schemes in general.[3]

Certain critical decisions must be made in order to properly utilize cluster analysis. The goal of clustering and cluster analysis is to group and distinguish comparable units and to separate them

DOI : 10.5121/ijdkp.2015.5506

from differing units. Towards this end, cluster analysis encompasses a wide range of statistical techniques.[4] In cluster analysis, one attempts "to group large numbers of persons, jobs, or objects into smaller numbers of mutually exclusive classes in which the members have similar characteristics". The ultimate objective is to develop clusters whose configurations would be such that each entity in the analysis would be classified into only a single, unique cluster[5].

The product of this analysis is referred to by a variety of terms, including types, taxons, groups, classes, categories, classifications, or clusters. By extension, therefore, cluster analysis is also referred to as typological analysis, numerical taxonomy, pattern recognition, classification, or botryology. Although the foregoing implies that classification can be used synonymously with clustering, clustering is associated with the concept of forming classes, whereas classification has been used in the sense of identifying or assigning individual objects to predetermined classes based on specific criteria.[6]

Frequently, cluster analysis is used to determine groups so that subsequent assigning of new objects to these groups can be achieved Cluster analysis has been applied for such varied objectives as finding a true typology, model fitting, prediction based on groups, hypothesis generation, hypothesis testing, data exploration, data reduction, and grouping similar entities. into homogeneous classes [7] .

When reviewing methodologies used to build existing medical resource classification schemes, it becomes apparent that there is a standard set of decisions that must be made when doing medical clustering. These decisions have not received prominent attention in the proliferation of literature that has accompanied the wide-scale application of clustering techniques.

The focus of our paper is on that portion of the literature concerning conventional clustering procedures, which facilitates an understanding of how clustering methods may be used to develop opinions resource utilization and what decisions are required to use such methods.

2. RELATED WORK

Opinion mining(OM) (or sentiment analysis(SA)) is the computational study of people's opinions, appraisals, attitudes and emotions toward entities, individuals, issues, events, topics and their attributes. It has become a very active research area in the past few years due to challenging research problems and a wide arrange of applications. There are now at least 40 companies in abroad alone that provide some kinds of opinion mining services. Opinions are important because they are key influences on our behaviors. It is well known that our beliefs and perceptions of reality are to a considerable degree conditioned on how others see the world. For this reason, when we need to make a decision we often seek out the opinions of others. This is true not only for individuals, organizations and also for the doctors in medical field. With the explosive growth of social media (i.e., reviews, forum discussions, blogs and social networks, etc) in the past 10 years, individuals and organizations are increasingly using these media for their decision making [8].

For efficiency by reducing system wide costs linked to under treatment, over treatment, by reducing errors, cost and duplication in diagnosis. Nowadays, in medical domain doctors take decision for critical diagnosis with multiple opinion [op]. The literature of U.S healthcare and

other sectors tells that 61 % of public and doctors seek for multiple opinions before taking decisions for diagnosis of these diseases [9].

Now days in medical domain, it is difficult to make a decision for complex diseases henceforth doctors seek multiple opinions of different experts in order to achieve the accurate diagnosis process for the diseases [10].

The objective of successful diagnosis is by experts past experience or the knowledge gained from those experience. Experts can make prediction from previous observations (solved cases) and produce diagnosis for new cases. Using these experience experts can suggest good opinions about the diseases. Similarly different experts having different background knowledge (experience) can suggest different opinions, which leads for multiple opinions [11].

3. PROPOSED WORK

The proposed system uses Bisecting K-means algorithm for clustering the common types of opinion decisions given by a set of experts for particular case, where this algorithm will effectively cope up with outliers. In order to retrieve decisions the Best Position algorithm is used.

3.1 Advantages of the Proposed work:

- Bisecting K-means tends to produce clusters of relatively uniform size whereas K-means produce clusters of non-uniform size.
- If the number of clusters is large, then bisecting K-means is more efficient than the Regular K-means algorithm.
- The Bisecting K-means algorithm for document clustering which effectively cope up with outliers.
- Bisecting K-means produces uniform cluster irrespective of centroid selected.
- Best Position algorithm stops early than threshold algorithm and its execution cost Never higher than threshold algorithm.

The Figure (1) below shows flowchart for pre-processing and clustering of documents of opinions. First opinion decisions in the form of text documents are taken as input. In this paper three clusters are used for storing documents. If the number of documents less than number of cluster then the system not going to pre-process and cluster the documents. If the number of documents more than the number of clusters then, the system going to pre-process the documents.

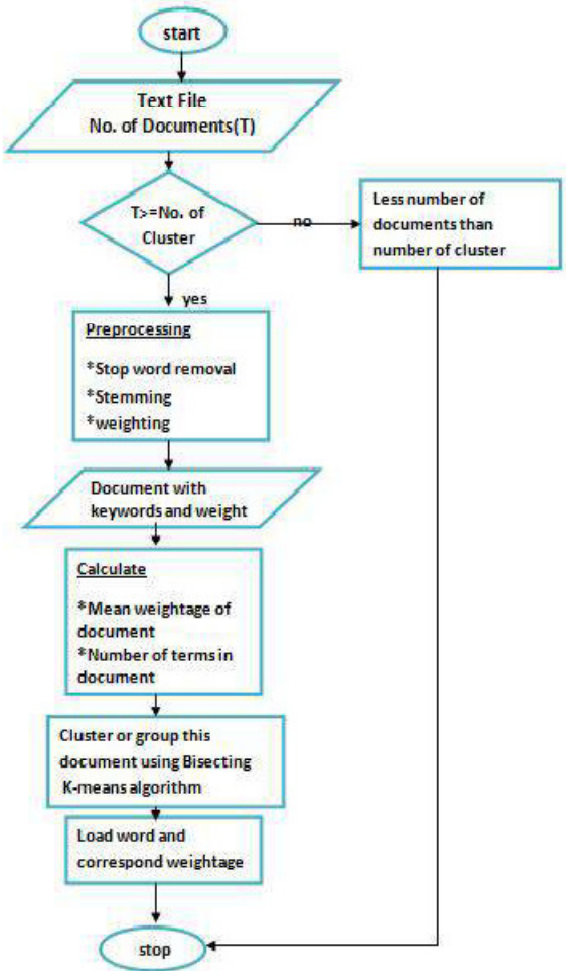

Figure (1)

4. IMPLEMENTATION

The algorithm is implemented in Java and proposed algorithm is based on the clustering of opinion decisions given by the experts for particular case, which is stored in the form of text documents.

4.1. List of Modules

1. Documents pre-processing module.
2. Documents clustering module.
3. Gateway module.
4. User interface module.

4.1.1. Documents pre-processing module:

Document pre-processing is the process of introducing a new document to the information retrieval system in which each document introduced is represented by a set of index terms. The goal of document pre-processing is to represent the documents in such a way that their storage in

the system and retrieval from the system are very efficient. Document pre-processing includes the following stages.

Stop word removal:

Stop words are very common words in the natural language. This stop words increase the time of searching particular phrases. Example the, and, who, what, are, was, then etc. Stop words are specific to particular language. It is very difficult to identifying all stop words in a language. We have to identify those stop words manually.

Steps involved in identifying and eliminating stop words:

1. Store the documents contents in the file stream.
2. Manually specify the stop words to remove.
3. Copy the content from file stream to another document.
4. If any stop word encounters don't copy those words to document.

Stemming:

Stemming is to reduce the variant form of words to the normal form. Example connection, connections, connectives, connected, connecting are variant form of the word connect.

Steps involved in identifying and eliminating stem words:

1. Store the documents contents in the file stream.
2. Manually specify the required stemming word.
3. Copy the content from file stream to another document.
4. If any stem word appears before the space character stems those words and copy remaining part to document.

Weighting:

Weighting for the documents done using term frequency inverse document frequency (TF-IDF) weighting function. Number of terms in entire document is document frequency. Number of times the particular word repeated is the term frequency. Weighting is calculated by dividing both measures. Example if document having 1000 words, then the document frequency (DF) is 1000. If a particular word repeated 10 times in document, then the term frequency (TF) is 10. Weight for the particular word is calculated using dividing

(TF) / (DF) that is 10/1000=0.01.

4.1.2. Documents clustering module:

After document pre-processed, all these documents are grouped in three clusters using Bisecting k-means algorithm with the use of two measures mean weight age of all document and mean number of words in all documents.

Bisecting K-means algorithm:

Bisecting K-means tends to produce clusters of relatively uniform size whereas K-means produce clusters of non-uniform size. If the number of clusters is large, then bisecting K-means is more efficient than the regular K-means algorithm. Bisecting K-means is an excellent algorithm for clustering a large number of documents. The Bisecting K-means algorithm for document clustering is effectively coping up with outliers. Bisecting K-means produces uniform cluster irrespective of centroid selected.

Steps followed to clustering documents using Bisecting K-means:

1. Initially put all the documents in a single cluster.
2. Calculate number of words in document and mean weight for documents for all the documents.
3. Calculate the initial centroid by mean number of words in all documents and mean weight age documents.
4. Select the random document and calculate the symmetric point for random documents by considering initial centroid as a midpoint.
5. Find 2 sub-clusters using the basic K-means algorithm.
6. Repeat step 2, the bisecting step, for a fixed number of times and take the split that Produces the clustering with the highest overall similarity. (For each cluster, it similarity is the average pair wise document similarity, and we seek to minimize that sum over all clusters.)
7. Repeat steps 3, 4 and 5 until the desired number of clusters is reached.

4.1.3 Gateway module:

After grouping the documents, Documents action words and weights are transfer to tables. Information extraction is done by gateway module by using the best position algorithm for stopping condition.

Gateway module responsible for the following steps:

1. Receiving the user request.
2. Finding the cluster regarding the request word.
3. Finding the files in cluster belong to the word based on weight age.

Best Position Algorithm (BPA):

Main idea:

Take into account the positions (and scores) of the seen items for stopping Condition and Enables BPA to stop much sooner than Threshold Algorithm (TA).

Best position:

The greatest seen position in a list such that any position before it is also seen. Thus, we are sure that all positions between 1 and best position have been seen.

Stopping condition:

Based on best positions overall score, i.e. the overall score computed based on the best positions in all lists

How the algorithm works:

- Do sorted access in parallel to each list Li, For each data item seen in Li, Do random access to the other lists to retrieve the item's score and position.
- Maintain the positions and scores of the seen data item, Compute best position in Li and Compute best positions overall score.
- Stop when there are at least k data items whose overall score ≥ best positions overall score

4.1.4 User interface module:

First user has to register and login with those registration details. The user has to enter keyword for retrieving documents. The requested keyword passes into Gateway module to retrieve the documents. If documents having the keyword the user searched then documents are retrieved based on highest weight age.

User responsible for the following action:

- User has to register and login with registered details.
- Input search word.
- View the Received file name.
- Open the file content and view.

5. TEST CASES :

5.1.Test Cases for Unit Testing:

Test Case-1	UTC-1
Name of Test	Pre-process button in User Interface Window
Items being tested	Pre-Process Button
Sample Input	Click on Pre-process Button
Expected Output	Files in the input folder should be weighted and stemmed.
Actual output	Files in the input folder are processed according to the expectation.
Remarks	Pass

5.2. Efficient opinion decisions retrieval Testing and validation:

Test Case-2	UTC-2
Name of Test	Bisecting K-means Button in User Interface window.
Items being tested	Bisecting K-Means Button
Sample Input	Click on Bisecting K-Means Button.
Expected Output	After clicking the button clustering operation should performed on those files.
Actual output	Clustering operation is happened after clicking the button.
Remarks	Pass

5.3. Efficient opinion decisions retrieval Testing and validation:

Test Case-3	UTC-3
Name of Test	User Home-Submit
Items being tested	One word in text field with submit button
Sample Input	Click on submit Button
Expected Output	It will check the files which are found with the worked what you are typing. If found then it should display the file downloaded message. In Gateway frame should display the selected files with highest weight.
Actual output	It is displaying.
Remarks	Pass

5.4.Test Cases for Integration Testing:

Test Case-4	IT
Name of Test	Clustering Documents
Description	It should cluster the documents and stores in respective cluster
Sample Input	Pre-Processed Documents.
Expected Output	Documents should be stored in respective cluster.
Actual output	Documents are stored in respective cluster based on Bisecting K-Means.
Remarks	Pass

5.5. Checking the overall working of system:

Test Case-5	ST
Name of Test	Checking the overall working of system
Description	It will retrieve the highest weighted documents.
Sample Input	Text documents.
Expected Output	Documents having requested keyword with highest weight.
Actual output	It is retrieving documents having requested keyword with highest weight.
Remarks	Pass

6. CONCLUSION

In this work, propose a novel approach that carefully uses text mining techniques in order to pre-process and clustering text documents. Normally K-means clustering technique used for text document clustering. In this work uses Bisecting K-means clustering technique and it is better compared to traditional K-means technique. This paper concentrates on finding Top opinions file retrieval based on keyword weight in document.

7. FUTURE ENHANCEMENT

The future work is pursuing in the following direction. The paper can be extended to deal with still more formats of documents and also to test on specific Cases.

REFERENCES

[1] Atul Kumar Pandey*, Prabhat Pandey**, K.L. Jaiswal***, Ashish Kumar Sen**** " DataMining Clustering Techniques in the Prediction of Heart Disease using Attribute election Method" ISSN: 2278 – 7798 International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 10, October 2013.

[2] Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley and Sons, 2001.

[3] Bay KS, Leatt P, Stinson SM. A patient classification system for long-term care. Med Care. 1982;20:468-11.

[4] Biljen EJ. Cluster Analysis: Survey and Evaluation of Techniques. Groningen, The Netherlands: Tilburg University Press, 2009.

[5] Ward JH, Hook ME. Application of an hierarchical grouping procedure to a problem of grouping profiles. Educ Psycho1 Meas. 196323:69-11.

[6] Bock HH (ed). Classification and Related Methods of Data Analysis. Amsterdam: The Netherlands; North Holland, 2008.

[7] Gordon AD. Classification. London, U.K.: Chapman and Hall, 2011.

[8] Yuefeng li, Ning Zhong, Raymond Y.K.lau " Topic Feature Discovery and Opinion Mining"10thIEEEinternational Conference on Data Mining(ICDM'10. Sydney,Australia.

[9] James Manyika,Michael Chui,Brad Brown,Jacques Bughin, Richard Dobbs, Charles Roxburgh Angela Hung Byers from McKinsey Global Institute " Big data: The next frontier for innovation, competition, and productivity " McKinsey & Company 2011.

[10] Mitja Lenic, Petra Povalej, Milan Zorman, Peter Kokol,faculty of ectrical engineering and computer science, university of Maribor, slovenia,Proceedings of the 17th IEEE symposium on computer-based medical systems(CBMS'04) 1063- 7125/04,2004 IEEE.

[11] Mitja Lenic, Petra Povalej, Milan Zorman, Peter Kokol,faculty of ectrical engineering and computer science, university of Maribor, slovenia,Proceedings of the 17th IEEE symposium on computer-based medical systems(CBMS'04) 1063- 7125/04,2004 IEEE.

AUTHORS

A.Ananda Shankar[1] is a M.Tech graduate in Computer Science and Engineering. Currently works as Associate Professor in the School of Computing and Information Technology at Reva University, Bangalore, Karnataka. Currently he is doing his research work in the area of Medical Data mining under Visvesvaraya Technological University, Belgaum. He has Two international conference publications and One international journal publications and Two national conference publication to his credit.

K. R. Ananda Kumar[2] holds a Doctoral Degree in Computer Science and Engineering. Currently works as Professor and Head of the Department in the Department of Computer Science and Engineering, S.J.B.I.T, Bangalore, Karnataka. He has a vast teaching experience of about 25 years. His research interest includes medical data mining; data stream mining, Artificial intelligence, intelligent agents and web mining. He is currently guiding five research scholars. He has Over 51 papers in International & National Journals and conferences to his credit.