TEXT DATA MINING : CLUSTERING APPROACH

D. Saravanan[1] and K. Chonkanathan[2]

[1]*Asst.Prof., Sathyabama UniversitY, Sa_roin@yhaoo.com*
[2]*Asst.Prof., Sarswathi Velu Coll.of Eng., kchokkanathan@rediffmail.com*

ABSTRACT

Text mining usually involves the process of structuring the input text usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database, 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 interestingnes text mining has emerged as a different stream in data mining because of the unstructured nature associated with free text. Many algorithms have been developed to assist in text mining. This paper presents the use of text mining based on a novel high dimensional clustering algorithm that leads to the exploratory data mining on data associated with the text. Experimental results of analyzing a real-world text data set and associated data are also presented.

Keywords : Data Mining, Clustering, Organizing Map, Text Analysis.

1. INTRODUCTION

Databases have been associated with organizations since the beginning of wide usage of computers to manipulate and record day-to-day activities. The ability of computers to store and retrieve data efficiently and the ability of storing enormous volumes of data in a very small area have been the main reasons for the popularity of the use of computers in organizations. It has been a necessity for many organizations as traditional methods of retrieving data such as reports, queries, etc. do not allow the extraction of hidden information in the data. In some instances the data is in formats that are hard to process and retrieve using these traditional retrieval methods. Lengthy text documents and descriptions have been associated with many databases, but are seldom used in the data mining applications mainly because of the lack of structure in these textual information. Text mining has also been developed as a field of its own in data mining because of this complexities associated. Many text mining algorithms are currently being used in the industry and clustering algorithms such as feature map algorithms based on the Self Organizing Map (SOM) [1] dominate in this area. The unsupervised nature of these algorithms leads to the grouping of text based on their similarities. These algorithms require the data to be encoded into numbers using techniques such as converting the text into phrases or words and then encoding them using techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) [2]. Text is preprocessed in order to eliminate spelling mistakes and to form stem forms of the words to reduce complexity of the processing.

This paper uses a novel growing high dimensional feature map algorithm called the High Dimensional Growing Self Organizing Map with Randomness (HDGSOM*r*) [3, 4] for processing text data efficiently. The algorithm is capable of producing good clusters from very large collections of text iterating only for 50-60 rounds.

2. APPLICATIONS

Recently, text mining has received attention in many areas.

2.1 Security applications

Many text mining software packages are marketed towards security applications, particularly analysis of plain text sources such as Internet news.It also involves in the study of text encryption.

2.2 Biomedical applications

A range of text mining applications in the biomedical literature has been described. One example is PubGene that combines biomedical text mining with network visualization as an Internet service.

2.3 Software and applications

Research and development departments of major companies, including IBM and Microsoft, are researching text mining techniques and developing programs to further automate the mining and analysis processes.

2.4 Online Media applications

Text mining is being used by large media companies, such as the Tribune Company, to disambiguate information and to provide readers with greater search experiences, which in turn increases site "stickiness" and revenue.

2.5 Marketing applications

Text mining is starting to be used in marketing as well, more specifically in analytical Customer relationship management. Coussement and Van den Poel (2008)[6] apply it to improve predictive analytics models for customer churn (customer attrition).

2.6 Academic applications

The issue of text mining is of importance to publishers who hold large databases of information requiring indexing for retrieval. This is particularly true in scientific disciplines, in which highly specific information is often contained within written text. Therefore, initiatives have been taken such as Nature's proposal for an Open Text Mining Interface (OTMI) and the National Institutes of Health's common Journal Publishing Document Type Definition (DTD) that would provide semantic cues to machines to answer specific queries contained within text without removing publisher barriers to public access.

3. TEXT ANALYSIS PROCESSES

Subtasks — components of a larger text-analytics effort — typically include:

Information Retrieval or identification of a corpus is a preparatory step: collecting or identifying a set textual materials, on the Web or held in a file system, database, or content management system, for analysis.

Recognition of Pattern Identified Entities: Features such as telephone numbers, e-mail addresses, quantities (with units) can be discerned via regular expression or other pattern matches.

Coreference: identification of noun phrases and other terms that refer to the same object. For example, anaphora is a type of coreference.

Relationship, Fact, and Event Extraction: **I**dentification of associations among entities and other information in text Sentiment Analysis involves discerning subjective material and extracting various forms of attitudinal information: sentiment, opinion, mood, and emotion. Text analytics techniques are helpful in analyzing sentiment at the entity, concept, or topic level and in distinguishing opinion holder and opinion object

4. OVERLAYING DATA ONTO TEXT CLUSTERS

The problems faced with association rule mining is that the data needs to be formatted as separate fields for association rule mining to extract useful information. Text mining has the advantage of finding associations in data that is This paper highlights on the advantage of overlaying other associated data onto text records[5]. The technique presented in this paper is illustrated in Figure 1. The process begins by pre-processing the text records to be presented to the HDGSOMr algorithm described in the next section. The heart of the technique is on the HDGSOMr algorithm that efficiently clusters the text data records and produces a map that has clusters organized in such a way that related clusters are placed close to each other.

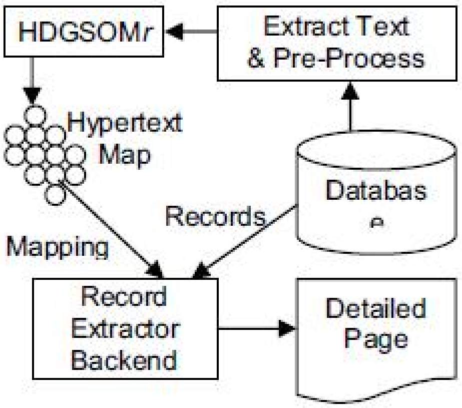


Figure 1. Block diagram of the proposed method.

A hypertext map is produced from the HDGSOMr algorithm, which is then used as the user interface for the technique. By clicking on the hyperlinks in the map, the data analyst can navigate to the records associated with the group. The records are extracted and a summary of the other fields associated with the records and the full texts are presented in the detail page.

4.1 The High Dimensional Growing Self Organizing Map with Randomness (HDGSOM*r*)

The High Dimensional Growing Self Organizing Map with Randomness (HDGSOMr) [3, 4] is an extension of the Growing Self Organizing Map (GSOM) . The GSOM itself is a growing variant of the popular Self Organizing Map (SOM) [1]. It was developed as a solution to the problem of identifying the correct height and width of the SOM depending on the natural distributions of the data. The HDGSOM the predecessor to HDGSOMr was developed as a high dimensional version of the GSOM that eliminated some problems encountered by the GSOM when applied to very large dimensional datap[6].

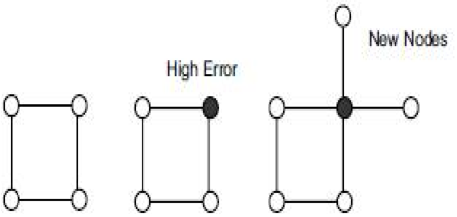


Figure 2. New node growth in the HDGSOMr algorithm.

Introduction of randomness into the HDGSOM algorithm enhanced the performance by converging the algorithm much quickly. The HDGSOMr algorithm has a growing phase that reduces a growth threshold value in several stages interleaved with several smoothing phases that lead to a smoother growing. The growing phase is then followed by more smoothing phases that resemble the operations of the SOM resulting in a better feature map than the original GSOM algorithm. The HDGSOMr algorithm in brief is as follows.

4.1.1. Initialization phase

The HDGSOM is created with 4 nodes connected to each other in a rectangular shape as in Figure 2 and initialized with random weight vectors. Since many high dimensional datasets such as those in text are very sparse, the growth thresholds used in the growing phase are calculated using the non-zero dimension distribution of the dataset. This involves calculating the average (µ) and standard deviation (□) of the non zero dimensions of all the inputs. A non zero dimension of an input is an attribute in the input that has a value greater than zero.

Two growth threshold values, (GT_1 and GT_2) are calculated to be used in the growing phase of the algorithm: (GT_2) –(µ+2□)Xln(SF)X 50 and GT_2=-ln(µ+2□)Xln(SF)X50 where *SF* is a parameter ranging between 0 and 1 called the *SpreadFactor* that allows the data analyst to control the spread of the map. Higher values of *SF* generate larger maps giving detailed clusters while smaller *SF* values produce denser clusters. The *SF* is usually set at 0.1.

4.1.2 Growing phase

During the growing phase of the HDGSOM, inputs are presented to the nodes as in the SOM algorithm and the winner node is identified. The error between the input and the weight vector is used to update the weights of the winner node and its neighbors similar to the SOM. If the accumulated error in a node exceeds a threshold called the growth threshold (GT explained next) and the node is a boundary node, then new nodes will be grown in all the vacant neighboring positions as illustrated in Figure 2.The growth threshold GT used to decide on when to grow new nodes varies from *GT1* to *GT2* in equal steps after each growing epoch..

4.1.4. Smoothing Phases

The growing phase is followed by two smoothing phases that are almost similar to the smoothing phases of the SOM, but have smaller numbers of epochs. These two phases have diminishing learning rates (α) that smooth out the map to produce crispier clusters.

4.1.5 Self Organization With Randomness

The performance of the algorithm is improved over the traditional self organizing process by introducing randomness into the weight adaptation process. This is achieved by modifying the weight adaptation process as follows:

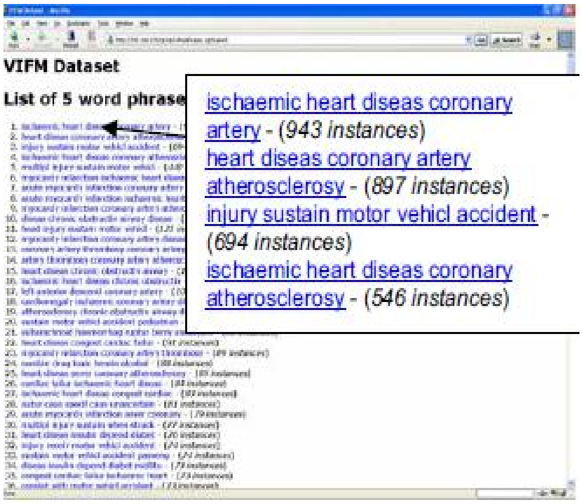
$$W_i^{new} = w_i^{old} + [a+(r-0.5)xax2]xnx(x_i - W_i^{old})$$

where is the updated weight of the i^{th} component of the weight vector, is the weight of the i^{th} component of the weight vector before updation, is the value of the i^{th} component of the input, α is the learning rate and *r* is a random number in the range of 0 and 1

5. EXPERIMENTS AND RESULTS

5.1 Searching the map

A search tool was developed. This allowed the user to give several words as input and find the nodes that had the most dominant occurrences of those words. Although searching from words enhanced the navigation of the map, it was a bit difficult to decide on which words to put in the search. It was evident that some guidance in the selection of the words was useful. An additional set of links based on the most occurring 2, 3, 4 and 5 word phrases were listed for the user as illustrated in Figure 3.



VIFM Dataset

List of 5 word phrase

ischaemic heart diseas coronary artery - (943 instances)
heart diseas coronary artery atherosclerosy - (897 instances)
injury sustain motor vehicl accident - (694 instances)
ischaemic heart diseas coronary atherosclerosy - (546 instances)

Figure 3. List of 5 word phrases in the dataset.

This gave the user a head start in navigating the map as most of these phrases are common phrases used by pathologists in the autopsy reports. For a non technical user such as a pathologist in this case, would be very familiar with phrases and can quickly start navigating the map through these leads.

5.2. Interesting findings Histogram

In statistics, a **histogram** is a graphical representation, showing a visual impression of the distribution of experimental dataThe total area of the histogram is equal to the number of data. A histogram may also be normalized displaying relative frequencies. It then shows the proportion of cases that fall into each of several categories The statistical analysis in this case was carried out by extracting the relevant records from the database and exporting them to a statistical tool. The evident advantage of combining the two have encouraged the authors to include most of these statistical capabilities to the visual exploration tool itself to enable the easy exploration of data by a non-technical user. The promising results from the VIFM dataset has lead to the exploration on other datasets.

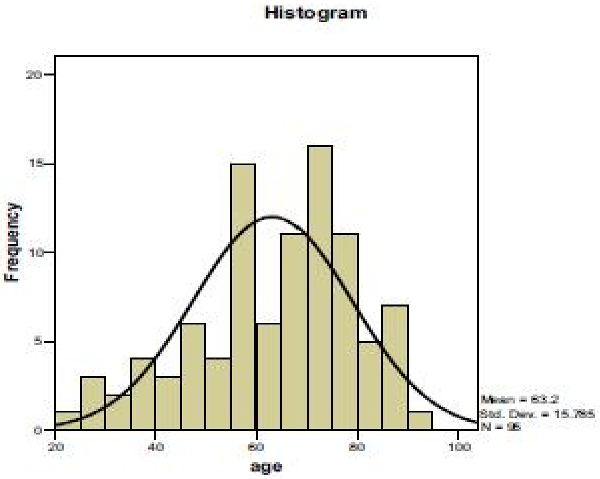


Figure 4. Age distribution for Node 91.

6. CONCLUSIONS AND FUTURE WORK

Traditional text mining algorithms have been mainly used on processing and finding similar documents in large collections of documents. In this paper we presented that traditional text mining can also be applied on text fields associated with databases which are harder to process using traditional query languages. The architecture presented in this paper uses a novel high dimensional growing self organizing map called the HDGSOMr that is capable of producing clusters from large collections of text and other high dimensional data much more efficiently than traditional SOM based feature maps..

7. REFERENCES

[1] T. Kohonen, "Self Organizing Maps", Springer, 2001.

[2] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", Information Processing & Management, vol. 24, 1988.pp 513-523

[3] R. Amarasiri, D. Alahakoon, M. Premaratne, and K.Smith, "Enhancing Clustering Performance of Feature Maps Using Randomness", presented at Workshop on Self Organizing Maps (WSOM) 2005, France, 2005.pp 463-470

[4] R. Amarasiri, D. Alahakoon, M. Premaratne, and K.Smith, "HDGSOMr: A High Dimensional Growing Self Organizing Map Using Randomness for Efficient Web and Text Mining", presented at IEEE/ACM/WIC Conference on Web Intelligence (WI) 2005, Paris, France, 2005.pp

[5] P. Srinivasan and A. K. Sehgal, "Mining MEDLINE for Similar Genes and Similar Drugs", Department of Computer Science, The University of Iowa TR# 03-02, July 2003.

[6] M. A. Hearst, "Untangling Text Data Mining", presented at 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, USA, 1999.pp

[9] A.J. McGrail, E. Gulski et al.. "Data Mining Techniques to Asses the Condition of High Voltage Electrical Plant". *CIGRE WG 15.11 of Study Committee 15*.

[10] R. Baker. "Is Gaming the System State-or-Trait? Educational Data Mining Through the Multi-Contextual Application of a Validated Behavioral Model". *Workshop on Data Mining for User Modeling 2007*.

[11] J.F. Superby, J-P. Vandamme, N. Meskens. "Determination of factors influencing the achievement of the first-year university students using data mining methods". *Workshop on Educational Data Mining 2006*.

[12] Xingquan Zhu, Ian Davidson (2007). *Knowledge Discovery and Data Mining: Challenges and Realities*. Hershey, New York. pp. 163–189. ISBN 978-159904252-7.

[13] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", presented at 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C.,1993.pp 207-216

[14] R. Amarasiri, D. Alahakoon, and K. Smith, "HDGSOM: A Modified Growing Self-Organizing Map for High Dimensional Data Clustering", presented at Hybrid Intelligent Systems 2004, Japan, 2004.pp 216 – 221

[15] Norén GN, Bate A, Hopstadius J, Star K, Edwards IR. Temporal Pattern Discovery for Trends and Transient Effects: Its Application to Patient Records. *Proceedings of the Fourteenth International Conference on Knowledge Discovery and Data Mining SIGKDD 2008*, pages 963–971. Las Vegas NV, 2008.

[16] Healey, R., 1991, Database Management Systems. In Maguire, D., Goodchild, M.F., and Rhind, D., (eds.), Geographic Information Systems: Principles and Applications (London: Longman).