Knowledge Enhancement from Unstructured Text Descriptions

Nithya. A
Dept of Computer Science &Engg.-R&D Centre
East Point College of Engineering & Technology,
Bangalore, India.
nithya656@gmail.com

AnirbanBasu,
Dept of Computer Science &Engg,- R&D Centre
East Point College of Engineering & Technology,
Bangalore, India
abasu@anirbanbasu.in

Abstract: Unstructured textual data is being constantly generated through emails, web, blogs, tweets, customer reviews, etc. While the amount of textual data is increasing rapidly, ability to summarize, understand, and analyze such data is becoming challenging. This paper presents a method for text mining for discovering important knowledge from unstructured text descriptions.We make use of entropy analysis[1]to extract an A-term list, a list of terms that are important to characterize the documents of interest, a vector space model to represent features of important documents, and a constraint based k-means clustering algorithm to generate high purity clusters for use in detecting relevant documents.

This paper illustrates with an example on health care data. The search time on health care data is measured on Hadoop platform and results are discussed.

Key words:

Big Data, unstructured data, text mining, Hadoop, Clustering, important document detection.

I. INTRODUCTION

Big Data[2,3] applies to information with three primary characteristics of Volume, Variety and Velocity. Such data can be structured, semi-structured or unstructured. Big Data cannot be processed or analysed using traditional methods as a large percentage of Big Data is unstructured. Methods for analysis of such unstructured data are still in its infancy.

Huge amounts of textual data are generated daily in thousands of organizations. Text data are mostly in the unstructured form but contains important knowledge that can be used to make smart business decisions, improve processes etc. This paper proposes a method for discovering knowledge in textual data. The method has lot of applications in health care where medical diagnosis can be facilitated by clustering of unstructured data.

The performance of the proposed method has been measured on Hadoop [5, 7] platform.

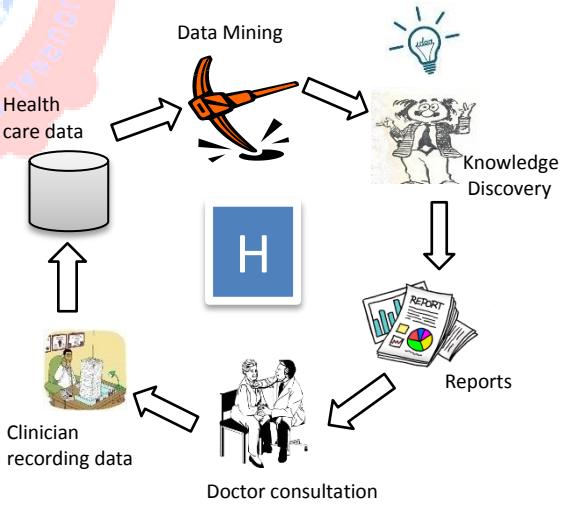


Figure 1: Typical Process in Health Care.

Figure 1 shows a typical process in health care. Healthcare information systems collect massive amounts of textual and numeric information about patients, visits, prescriptions, physician notes etc. Analysis of the information encapsulated within electronic clinical records can lead to improved medical diagnosis and treatment, promotion of clinical and research initiatives, fewer medical

errors and lower costs. However, the documents that comprise the health record vary in complexity, length and use of technical vocabulary. This makes knowledge discovery complex.

Typically data generated in Healthcare comprise of: Structured data, Semi-structured data, Unstructured data.

Structured Data: Data that resides in a fixed field within a record or file is called structured data. Data contained in relational databases and spreadsheets are examples of structured data. This data can be processed by traditional query languages.

Semi-structured Data: Is a form of data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables. Eg: XML

Unstructured data: Unstructured Data[8] refers to information which does not have a pre-defined structure i.e. information which is not stored in database in a row column format or is not organized in a pre-defined manner. Eg: Free text on web, audio, videos, pdf file, text document etc.

Text Mining[4]**:** also referred to as text data mining[12], analyses text to derive high quality information or patterns in text. Text mining normally requires a pre-processing phase such as spell checking, sentence splitting, word sense disambiguation, and more and also simple pattern matching, machine learning which are used to extract important concepts or detect hidden relationships in large "free-text" data.

Text mining is of great value in healthcare data as it supports the discovery of unknown disease correlations and the identification of previously unknown drug side effects.

II. RELATED WORK

There have been substantial researches in the area of Text Mining in unstructured data.

R. Kosala et al. [13] summarized the research work done for unstructured data and semi-structured data from information retrieval view. It shows that most of the researches use bag of words, which is based on the statistics about single words in isolation, to represent unstructured text and take single word found in the training corpus as features.

O. Zaiane et al. [14] proposed the idea of how to implement the OLAP technique on the Web mining. Their work on multimedia data also provided a valuable solution for content mining.

Dai et al. [15] proposed a Co-clustering based approach for this problem. In this method, they identified the word clusters among the source and target domains, via which the class information and knowledge propagated from source domain to target domain.

Leonid Churilov, AdylBagirov, Daniel Schwartz, Kate Smith and Michael Dally[16] had already studied about combined use of self-organizing maps and non-smooth, non-convex optimization techniques in order to produce a working case of a data driven risk classification system. The optimization approach strengthens the validity of self-organizing map results. This study is applied to cancer patients.

Different approaches to solve the problem of clustering analysis are mainly based on statistical, neural network, machine learning techniques. Bagirov et al. [17] propose the global optimization approach to clustering and demonstrate how the supervised data classification problem can be solved via clustering. Due to a large number of text data and the complexity function, general purpose global optimization techniques fail to solve such problem. It is very important therefore, to develop optimization algorithm which can handle huge amount of complex data efficiently.

Zhuang et al. [18] formulated a joint optimization framework of the two matrix tri-factorizations for the source and target domain data respectively, in which the associations between word clusters and document classes are shared between them for knowledge transfer. Although the basic assumption of this method is similar to our method, it lacks the probabilistic explanation of the model and is not easy to be extended to handle the tasks with multiple source and target domains.

The volume of published text mining research, and therefore the underlying knowledge base, is expanding at an increasing rate. Among the tools that help researchers in dealing with this information overload are text mining and knowledge extraction.

III. PRESENT METHOD

There are many challenges in automatic mining of knowledge from unstructured text descriptions. For example, when a patient visits the hospital

complaining about cough and cold, the doctor checks the patient and writes a report by examining the patient. Here cough and cold might be either common cold or might be related to serious health issues, each symptom has different treatment and prescription of drugs. The most challenging of all is that these text descriptions are often ill-structured, do not follow the English grammar, contain a lot of self-invented acronyms, and shorthand descriptions.

Clustering is the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are "similar" to each other but are "dissimilar" to the objects belonging to other clusters.

The Clustering methods[10] are:

Partitioned clustering.
Hierarchical methods.
Density based clustering.
Centroid-based clustering.
Distribution-based clustering.

K-means clustering is a "Centroid based Clustering" algorithm and aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

In text mining, documents are often represented in a form of vector space models for computational analysis.

Vector space model[9] is an algebraic model for representing text documents as vectors of identifiers e.g. index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings.

Entropy is defined as a measure of the expected information content or uncertainty of a probability distribution. Entropy is used to measure the evenness of the distribution of each word.

IV. WORK DONE

In this paper we focus on detecting important text documents through machine learning. Detecting important documents has many important applications. Different applications may have their own definition of "importance". The search problem can be modelled as extracting important terms.

In text mining, clustering[10] is widely employed for automatically structuring large document collections and enabling cluster based information browsing, retrieval and classification. The k-means clustering[11] technique is often used to partition documents into k clusters in which each document belongs to the cluster with the nearest distance to its mean.

Figure 2 shows the major machine learning algorithms developed for building an intelligent classification system that can accurately detect important documents. The algorithm, generating a term List, derives a list of terms and phrases that are important for describing important documents based on word entropy analysis. Then we model Important documents using a vector space that consists of the term list and a list of local and global weights.

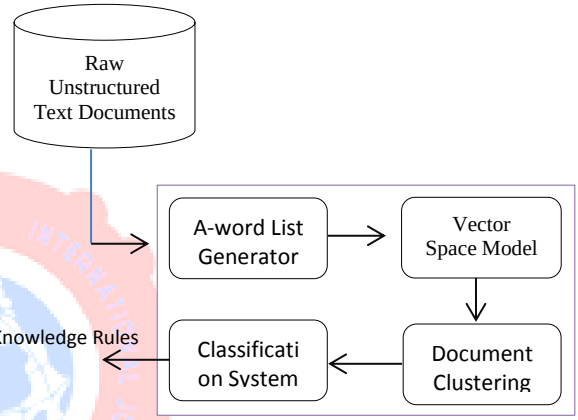


Figure 2. An intelligent system for detecting important documents

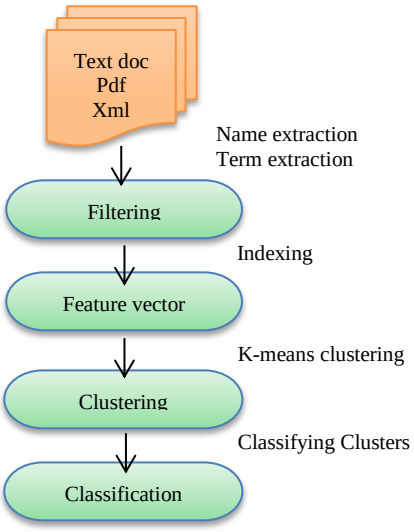


Figure 3.Work Flow Diagram

Figure 3 shows the work flow diagram. Initially text document which is in unstructured firm is provided as input.

- First step is to extract an A-term list, a list of terms that are important to characterize

the documents of interest, this A-term list is obtained by removing the stop words(Eg:- is, the, as etc.) present in the text document.

- In second step a vector space model is used to represent features of important documents as vectors of identifiers. Eg: index terms. It is used in information filtering, information retrieval, indexing.
- Finally a constraint based k-means clustering algorithm is used to generate high purity clusters for use in detecting relevant documents.

Eg:- In Health care data patients with similar symptoms of a particular disease can be grouped into a single cluster.

A. Extracting important term list

Our objective is to develop an intelligent system to accurately discriminate the important terms from unimportant. The following algorithm was developed to extract words that are important

- Extract all words from *the text documents* using tokenization, and then remove stop words to form an initial word list.
- For each term on the *word_list*, calculate the frequency rate of *Word* occurring in the document.

We use entropy to measure the evenness of the distribution of each word.

B. Vector Space Model for representing important documents

In text mining, documents are often represented in a form of vector space models for computational analysis. The word list described above is used as term list in the vector space model[9]. Here for each term we construct vector space model.

C. Detecting important documents through clustering

Clustering adds to the value of existing databases by revealing hidden relationships in the data, which are useful for understanding trends, making predictions of future events from historical data, or synthesising data records into meaningful clusters.

Here we use a constraint based k-means clustering algorithm[11] to produce clusters in an unstructured text document.

To improve the performance of the existing clustering methods, we use Hadoop platform and the performance of the proposed method is measured. We created search screen where documents of interest can be searched.

We proposed two different kinds of search

- Search by Document.
- Search by Content.

In search by document information displayed are:

- The cluster to which the document belongs.
- Inverted Index of the document.

In search by content information displayed are

- Matching documents for the term.
- Matching cluster for the document.

The k-means clustering algorithm was written in Java and its performance measured by running on Hadoop platform.

In Health care data patients with similar symptoms of a particular disease can be grouped into a single group.

A clinical syndrome is a set or a cluster of concurrent symptoms which indicate the presence and nature of a disease. Therefore, looking for concurrent symptoms is therefore one of the main tasks in medical diagnosis.

A classification system can be built based on these cluster centres and their assigned pattern class labels. Here we use constraint based k-means clustering algorithm to cluster the important documents.

V. IMPLEMENTATION DETAILS

The proposed system is implemented on Hadoop platform.
Hadoop is installed on Linux OS Ubuntu. The proposed method is run on Linux OS, next the entire method is eventually run on Hadoop using commands.

A. APACHE HADOOP

Apache Hadoop[5,7] is a Java-based software framework that enables data-intensive application in a distributed environment. Hadoop enables applications to work with thousands of nodes and terabytes of data, without concerning the user with too much detail on the allocation and distribution of data and calculation. Hadoop is a framework of tools used for running applications on Bigdata. Hadoop is open source and distributed under Apache license. The main components of Hadoop are: Map Reduce and HDFS.

• MAPREDUCE

Map Reduce[6,7] is a programming model for distributed computing.
Map Reduce works by breaking the processing into two phases: the map phase and the reduce phase. Each phase has key value pairs as input and output, the types of which may be chosen by the programmer. In MapReduce, the Map function processes the input in the form of key/value pairs to generate intermediate key/value pairs, and the Reduce function processes all intermediate values associated with the same intermediate key generated by the Map function. Map Reduce program performs sequence of transformations on key value pairs. The map function perform independent record transformations .The map function receives a key value pair of generic type and outputs zero or more key value pairs as of other types. The input type and output type can be same or different. The reduce function aggregates the results from map phase for every unique key. Reducer function receives list of values and outputs zero or more key value pairs. Map Reduce programs interact with framework to set up large tasks. The framework takes care of scheduling task, rerunning failed task, splitting input to feed map, moving map output to reducer and receiving output from reducer making it accessible in file system.

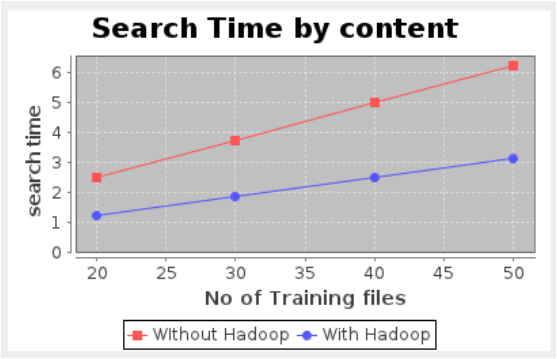
• HDFS

Hadoop distributed file system[7] is designed for storing very large files. Large files in this context means files that are hundreds of gigabytes and terabytes of size. HDFS is built around the idea that the most efficient data processing pattern is a write-once, read-many-times pattern. HDFS is designed to run on clusters of commodity hardware.

Map Reduce tasks use HDFS to read and write data. HDFS deployment includes a single Name Node and multiple DataNodes. When a user wants to read a file it uses a Hadoop HDFS client that contacts the NameNode. The NameNode then fetches the block locations and return the locations to the client, forcing the client to do the reading and merging of blocks from the DataNodes.

VI. PERFORMANCE ANALYSIS

The performance of the proposed method is measured on Hadoop platform and presented as a performance Graph.



Performance is measured on the basis of search time. Search time here is the time taken to search a keyword or content of interest.

When a keyword is entered the result displays all the documents containing the keyword, in addition the related clusters to which the document belongs are also displayed.

For example consider a typical health care data which contains files about the patients having common symptoms to a particular illness. When a keyword(symptoms) is entered the result displayed will contain the documents related to the entered keyword and also displays the cluster to which that document belongs to.

Training Files are the no:of source files present. The keyword entered is matched with each of these files. Therefore as the no:of source files are increased the search time for searching a particular keyword in all these files eventually increases.

The above graph shows the search time without using Hadoop and with Hadoop. The Search time increases as the training data increases without Hadoop. With Hadoop the search time is less even when the size of training dataset is increased.

VII. CONCLUSION.

This paper presents mining of unstructured textual descriptions using clustering algorithm. It presents two types of searches named as search by content and search by document. The performance is measured on the basis of search time when the system is run on Hadoop platform which shows a increase in search time with increase in training dataset without using Hadoop and better search time with increase in training dataset when using Hadoop.

VIII. REFERENCES

[1] T.Schürmann and P. Grassberger, Entropy Estimation of Symbol Sequences, CHAOS, Vol. 6, No. 3 (1996) 414–427.
[2] Laney, Douglas. "The Importance of 'Big Data': A Definition.
[3] Magoulas, Roger; Lorica, Ben (February 2009). "Introduction to Big Data".
[4] LiPing Huang and Yi L. Murphey, "Text Mining with Application to Engineering Diagnostics," The 19th International conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems, Annecy, France, June, 2006.
[5] Apache Software Foundation. Hdfs user guide. http://hadoop.apache.org/hdfs/docs/current/hdfs user guide.html, 2011.
[6] MapReduce tutorial-Apache Hadoop.
[7] Hadoop: The Definitive Guide 2009, by Tom White.
[8] Miner, G., Elder, J., Hill. T, Nisbet, R., Delen, D. and Fast, A. (2012). Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Elsevier Academic Press.
[9] G. Salton; A. Wong; C. S. Yang, A vector space model for automatic indexing, Communications of the ACM, v.18 n.11, p.613-620, Nov. 1975.
[10] Kaufman, L.; Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis
[11] Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; Wu, A. Y. (2002). "An efficient k-means clustering algorithm: Analysis and implementation".
[12] Fayyad, Usama; PiatetskyShapiro, Gregory; Smyth, Padhraic(1996). "From Data Mining To Knowledge Discovery In Databases".
[13]R. Kosala, H. Blockeel. Web mining Research: A Survey.
[14] O. Zaiane, M. Xin, J. Han. Discovering Web Access Patterns and Trends by applyingOLAP and Data Mining Technology on Web Logs. In Advances in Digital Libraries, pages 19-29, Santa Barbara, CA, 1998
[15]X. Liu, Y. Gong, W. Xu, and S. Zhu, "Document Clustering with Cluster Refinement and Model Selection Capabilities," Proc. 25thAnn. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '02), p. 191-198, 2002.
[16] Leonid Churilov. AdylBagirov, Daniel Schwarta, Kate Smith, Michael Dally, Journal of management information system : 2005, Data mining with combined use of optimization techniques and self-organizing maps for improving risk grouping rules: application to prostate cancer patients.
[17]Shakil Ahmed, FransCoenen, Paul Leng, Knowledge Information System : 2006, Tree based partitioning of data for association rule mining.
[18] Zhuang, Fuzhen; Luo, Ping; Shen, Zhiyong; He, Qing; Xiong, Yuhong; Shi, Zhongzhi; Xiong, Hui, "Mining Distinction and Commonality across Multiple Domains Using Generative Model for Text Classification," IEEE Transactions on Knowledge and Data Engineering, Volume: 24 , Issue: 11, p. 2025 – 2039, 2012.