International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 12

ISSN: 2454-4248
370 – 374

A Novel Approach for Clustering of Heterogeneous Xml and HTML Data Using K-means

Meena Saini
Computer Science & Engineering
Sobhasaria Engineering College,
Sikar,Rajasthan,
Indiasainimeena44@gmail.com

Yashwant Soni
Computer Science & Engineering
Sobhasaria Engineering College,
Sikar,Rajasthan, India
yashalw80@gmail.com

Abstract:- Data mining is a phenomenon of extraction of knowledgeable information from large sets of data. Now a day's data will not found to be structured. However, there are different formats to store data either online or offline. So it added two other categories for types of data excluding structured which is semi structured and unstructured. Semi structured data includes XML etc. and unstructured data includes HTML and email, audio, video and web pages etc. In this paper data mining of heterogeneous data over Xml and HTML, implementation is based on extraction of data from text file and web pages by using the popular data mining techniques and final result will be after sentimental analysis of text, semi-structured documents that is XML files and unstructured data extraction of web page with HTML code, there will be an extraction of structure/semantic of code alone and also both structure and content.. Implementation of this paper is done using R is a programming language on Rstudio environment which commonly used in statistical computing, data analytics and scientific research. It is one of the most popular languages used by statisticians, data analysts, researchers and marketers to retrieve, clean, analyze, visualize, and present data.

Keywords:- *Unstructured Data, Semi-structured Data, Heterogeneous Data, Data Mining, Clustering-means.*

*****

1. INTRODUCTION

Many recent applications deal with heterogeneous data that consists of several parts, each part being of a different type of domain or modality, for example many Web data sets, network activity data, scientific data sets, and census data sets typically comprise several parts that are of different types: numerical, categorical, transactional, free text, ratings, social relationships, etc. Traditionally each of these different types of data has been best clustered with a different specialized clustering algorithm or with a specialized dissimilarity measure. A very common approach to cluster data with mixed types has been to either convert all data types to the same type (e.g: from categorical to numerical or vice-versa) and then cluster the data with a standard clustering algorithm that is suitable for that target domain; or to use a different dissimilarity measure for each domain, then combine them into one dissimilarity measure and cluster this dissimilarity matrix.

Web mining is a part of both information extraction and information retrieval. For improve the classification of text, web mining is used to support machine learning [4]. The main objective of web mining is to extract useful information from web. Web mining is integration of information that is gathered by traditional data mining techniques with information gathered over World Wide Web. Web mining is decomposed into following subtasks:

i. Resource Discovery: It helps in retrieving services and unfamiliar documents on web.

ii. Information selection and preprocessing: This step is used for automatically selection of the specific and required information, and then preprocesses this information from the web sources.

iii. Generalization: It uncovers general pattern at individual web sites as well as across multiple sites.

iv. Analysis: It validates and interprets the mined pattern.

v. Visualization: It presents the result in visual and easy to understand way.

Web mining is divided into three main categories depending on the type of data as web content mining, web structure mining and web usage mining [17].we are mainly focusing on web content mining then applying sentimental analysis through frequency count and WordCloud. The main purpose of web content mining is to gather, organize, categorize and provide the user with the best possible information that is available on World Wide Web . The detailed study and analysis of each web content mining technique have been done in this paper. The future scope of web content mining is to predict the user needs to improve the usability and scalability

Clustering is the process of grouping data items or element in a way that makes the elements in a given group similar to each other in some aspect. Clustering has many applications such as data mining, statistical data analysis and bioinformatics. It is also used for classifying large amount of data, which in turn is useful when analyzing data generated from search engine queries, articles and texts, images etc.

 ISSN: 2454-4248
370 – 374

Information Retrieval is the path toward masterminding data (for the most part scholarly data) and building algorithms so people can form inquiries to recover the data they require. Think of Google. Web pages are combination of text, links, and multimedia.

Information retrieval is an activity to satisfy the information need by finding material(usually documents) from large sets data(generally stored in computers) which is unstructured data nature(text data).Information retrieval is fast becoming the dominant form of the information access. Information retrieval is a Query answer-oriented discipline, concerning with the query for Effective and efficient answer of desired information between human generator and human user.

In other words:

• The indexing, ranking and retrieval of textual documents.
• Firstly, Concerned with retrieving relevant documents to a query.
• Secondly, Concerned with retrieving from large collection of documents efficiently.

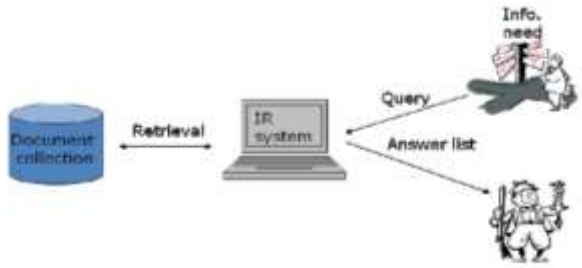


Fig.1. Process of Information Retrieval

The organization of this papers is as follows ; In Section we have discussed the related work to our work , In Section 3 we have proposed our new approach, Section 4 shows the implementation and result analysis on the basis of collected unstructured data using Rstudion. The final Section 5 concludes our paper.

2. RELATED WORK

Ming-Syan Chen et.al [1], Describes about the basic definition of Data mining. It is defined as a process of extracting or mining useful knowledge from huge amounts of data, or simply knowledge discovery in databases.

S. R. Dhamankaret. al [2], Describes about ontology and it states that more complex the application is, the larger the gap comes into existence between application and users . The data mining applications to illustrate the concepts and selection a better model to match business requirements to data mining categories to connect complex data mining concepts with business problems and assists users to choose the best data mining solution.

D.W. Jordan [19] describes the thoughts on Data Mining can generally be defined as automatically extracting useful knowledge from large amounts of data.

W. Himmel et.al [20],Text mining algorithms have been used in many applications such as summarizing and analyzing web content and managing scientific publications. Text mining generally starts with a text pre-processing step, where unstructured text is transformed into a structured form, which is then used for clustering or classification.

PrakashR.Andhale et.al [21], presents the characteristics of HACE theorem which provides the description of heterogeneous data and proposes a model for processing of heterogeneous data from the view of data mining. This information extraction model involves the information extraction, data analysis and provides the security and privacy mechanism to the data.

Hiroki Arimura et.al [22], proposed the algorithm and optimizes the performance for the text mining of semi structured data and unstructured data. A basic idea behind their method is to use an average set of texts as the control set used for canceling the occurrences of frequent and non-informative keywords.

Calvilloet.al [10], Describes about the text mining and about usage of data mining technique clustering. Automated text classification is the task of assigning a category to a document.

3. PROPOSED SYSTEM

Heterogeneous data is the combination of different semantics of data. This means that data stored in different formats such as HTML, XML, audio, video, PDF etc. different types of data is categorized by different authors which is semi-structured and unstructured data, In this paper we are working with data mining of semi structured data (XML) and unstructured data (website content in HTML format) with data analysis in the form of WordCloud and frequency count.

Fig. 2 Architecture of approach

 ISSN: 2454-4248
370 – 374

3.1 Loading text from website, xml file

It is a first activity of the paperframework, considering website www.nptel.com as the reference for this paper. Extracting the semantic contents from this website (with HTML code) and xml file will be the first step.

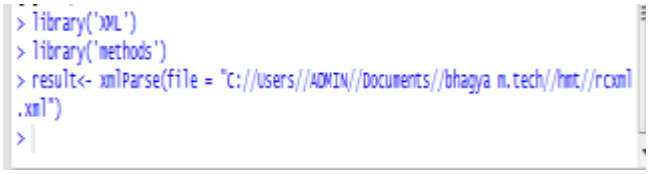


Fig. 3 Loading of semi structured xml file

Fig. 4 Loading of unstructured data from website (HTML)

3.2 Pre-processing of text

Pre-processing of text data activities are combination of1)Extract semantic of website data and xml data 2) Text extraction from structure 3) Loading content of website in csv file and 4) reading file content in R. this three steps are main steps from which further results will be occurred**.** Extraction of semantic of website means the page source code will extracted as it is as displayed on web in the Rstudio. Now the further step is to extract the relevant contents from that entire HTML programming structure, it is also called cleaning of data by removing html tags and all. Now the third step is to load the data extracted in Rstudio should be stored in some document file so commonly used file formats extensions in Rstudio is CSV file and XLSX file. Fourth step is to read the file content in Rstudio, by this step the data will appear in excel format with contents. Pre-processing step also include removing of numbers, special characters, converting the data into lowercase, remove punctuations etc.

3.3 Term document matrix

A term-document matrix represents the relationship between terms and documents, where each row stands for a term and each column for a document, and an entry is the number of occurrences of the term in the document. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms and organize the data according to their frequency and also removing the common terms, this will make 10% matrix empty. Term document matrix is shown in next figures.

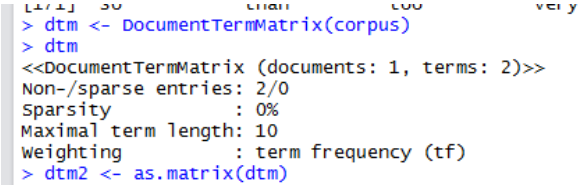


Fig. 5 XML file

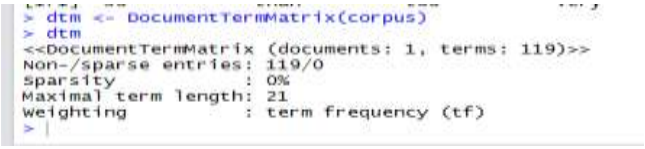


Fig. 6 HTML content

3.4 Calculate Frequent Word

In this we are going to find out the word that are most frequently occurred in the documents and we also adjust the frequency of word like 3, 4.Suppose we will adjust the minimum frequency 3 26 and then it will plot the word frequency graph as shown in Fig.5, which shows the word that occurs more than 3 or more times in documents. Note that the frequent terms are ordered alphabetically, instead of by frequency or popularity. To show the top frequent words visually, we make a barplot of them using ggplot2 package of Rstudio.

3.5 Relationship between Terms

Term Correlations - If you have a term in mind that you have found to be particularly meaningful to your analysis, then you may find it helpful to identify the words that most highly correlate with that term.

3.6 WordCloud

We can show the importance of words pictorially with a WordCloud. In the code below, we first convert the term-document matrix to a normal matrix, and then calculate word frequencies. After that we use WordCloud to make a pictorial. Humans are generally strong at visual analytics. That is part of the reason that these have become so popular. What follows are a variety of alternatives for constructing word clouds with your text. But first you will need to load the package that makes word clouds in R. We also form a word cloud of words which are frequently occurs in the documents. Similarly we can form a colored word cloud of words.

4. Implementation And Results

4.1 Extraction of xml data

 ISSN: 2454-4248
370 – 374

In this implementation, how to do data mining of semi structured data will be shown. Extracting the data from file name rcxml.xml. The data in this file is in the form of XML coding. Fig.7 shows the structure of the data in file which in the form of XML coding .and Fig.8 shows the result in which the content of the file is extracted using r programming in Rstudio.

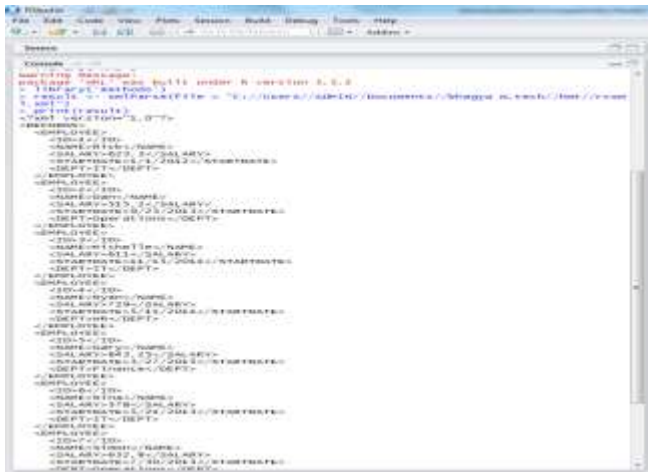
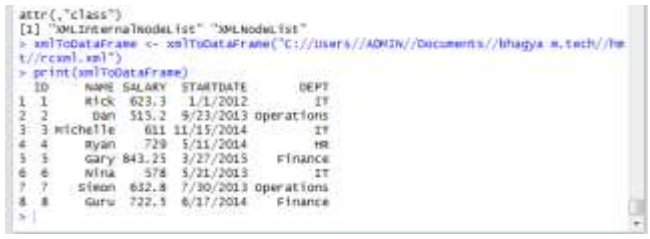


Fig.7 Semantic of Xml data in file



```
attr(,"class")
[1] "XMLInternalNodeList" "XMLNodeList"
> xmlToDataFrame <- xmlToDataFrame("C://Users//ADMIN//Documents//bhagya m.tech//he
t//rcxml.xml")
> print(xmlToDataFrame)
  ID     NAME SALARY  STARTDATE       DEPT
1  1     Rick  623.3   1/1/2012         IT
2  2      Dan  515.2  9/23/2013 operations
3  3 Michelle    611 11/15/2014         IT
4  4     Ryan    729  5/11/2014         HR
5  5     Gary 843.25  3/27/2015    Finance
6  6     Nina    578  5/21/2013         IT
7  7    Simon  632.8  7/30/2013 operations
8  8     Guru  722.5  6/17/2014    Finance
>
```

Fig.8 Content of Xml data in file

4.2 Extraction of webpage data/ html data

In this paper extraction of data from the website 'http://nptel.ac.in/courses/117105135/' which is demonstrated Step by step in the following sections. Extraction procedure will be displayed in the following sections here:

Fig..9 website page view

A) First estimating the distance between words and then cluster them according to similarity.

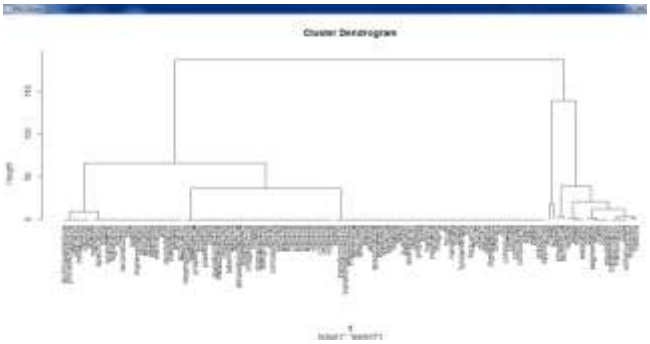


Fig.10 Dendogram of The Text

Helping to Read a Dendrogram- To get a better idea of where the groups are in the dendrogram, you can also ask R to help identify the clusters. Here, I have arbitrarily chosen to look at five clusters, as indicated by the red boxes. If you would like to highlight a different number of groups, then feel free to change the code accordingly. Clusters are divided in five different clusters as shown in the Fig. 11.

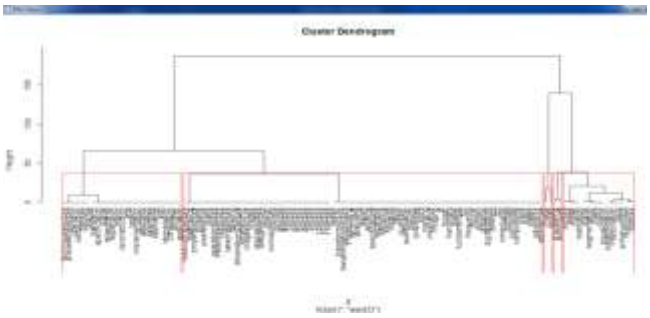


Fig.11 Specified clusters in red

B) K-means clustering

The k-means clustering method will attempt to cluster words into a specified number of groups (shown in Fig.12), such that the sum of squared distances between individual words and one of the group centers. You can change the number of groups you seek by changing the number specified within the kmeans() command.

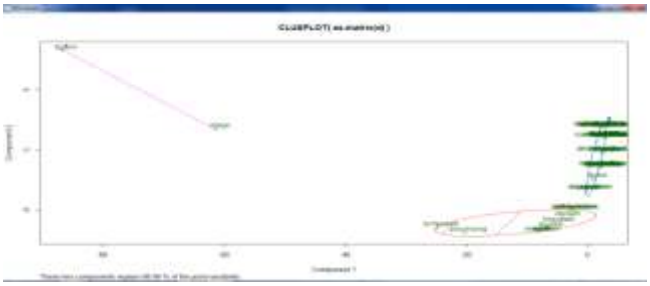


Fig.12 Graphplot of k-means clustering

5. Conclusion And Future Work

In this paper, the framework for web mining is implemented using data mining tool Rstudio. Most important aspect of this paper is to extract data from website which is obviously unstructured data. It found difficult to extract content from unstructured data source. Other aspects of this framework is

to identify the documents and the data they contained and evaluate the feasibility to apply text mining which may achieve good performance with high efficiency when dealing with thousands of documents, by separating the data contained by documents into bag of words. From our experiment we analyze, pre-processing does play an important role. Frequent words and associations are found from the matrix. A word cloud is used to present frequently occurring words in documents. Two main types of clustering techniques used(Hierarchical and k-means)applied on data set from that we can analyze the data.

The work presented in paper can be enhanced further by applying it to heterogeneous datasets, like Image, Audio, Video, Social Networking etc. we can also apply different tasks data mining such as classification, association, regression analysis and so on, also compare the work of these different tasks on the same data. Due to computer speed and memory limitations, data set was relatively small in this work. One of the future directions for this work is to perform a more detailed statistical analysis of heterogeneous data.

REFERENCES

[1] Ming-Syan Chen, Jiawei Han, and Philip S.Yu, "Data Mining – An Overview from Database Perspective",Knowledge and Data Engineering, IEEE Transactions on ,Volume 8 , No.6 , pp 866-883,Dec 1996.

[2] S. R. Dhamankar, Y. Lee, A. Doan, A. Halevy, and P. Domingos, "iMAP: Discovering Complex Semantic Matches between Database Schemas", International Conference on Management of Data,ACM SIGMOD, pp 383-394,2004.

[3] E. Rahm, P.A. Bernstein. "A Survey of Approaches to Automatic Schema Matching". VLDB Journal, Volume 10, No. 4, pp 334-350,2001.

[4] Piatetsky-Shapiro, Gregory, "The Data-Mining Industry Coming of Age" ,IEEE Intelligent Systems, Volume 6, pp 32-34,2000.

[5] Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi, "The survey of Data Mining Applications and Future Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[6] Nicholas J Belkin and W Bruce Croft,"Retrieval techniques", Annual Review of Information Science and Technology,Volume 22,pp 109-45,Information Today,1987.

[7] Romero, Cristobal, Sebastián Ventura, and Paul De Bra, "Knowledge discovery with genetic programming for providing feedback to courseware authors." Volume 14,Issue 5, pp 425-464, 2004.

[8] Ansari S., Kohavi R., Meason L., and Zheng Z., "Integrating E-Commerce and Data Mining: Architecture and Challenges",IEEE International Conference on Data Mining,pp 27-34,2001.

[9] Jadhav, S. R., and Kumbargoudar, P., "Multimedia Data Mining in Digital Libraries: Standards and Features READIT, pp.54-59,2007

[10] Calvillo, E. Alan, Alexandra Padilla, Jaime Munoz, Julio Ponce, and Jesualdo T. Fernandez, "Searching research papers using clustering and text mining." International conference on Electronics, Communications and Computing ,pp. 78-81, IEEE, 2013.

[11] Franklin, Michael, Alon Halevy, and David Maier. "From databases to dataspaces: a new abstraction for information management" ACM Sigmod Record,Volume 34,Issue 4,pp 27-33,2005

[12] Niranjan Lal, Samimul Qamar, "Comparison of Ranking Algorithm with Dataspace", International Conference On Advances in Computer Engineering and Application(ICACEA),pp 565-572, March 2015.

[13] Mark Hall, Eibe Frank, G. Holmes, B. Pfahringer, and P. Reutemann, " The WEKA data mining software: An update",ACM SIGKDD explorations newsletter, volume 11, Issue 1,pp 10-18, june 2009.

[14] Sunita B Aher, Mr. LOBO L.M.R.J., "Data Mining in Educational System using WEKA", International Conference on Emerging Technology Trends (ICETT),Volume 3,pp 20-25,2011.

[15] V. S. Jagadheeswaran, V. N. Saranya, "A Survey on Data Mining Application & Tools", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue5, pp. 477-480, May 2015.

[16] Vikas Gupta, Prof. Devanand, "A survey on Data Mining: Tools, Techniques,Applications, Trends and Issues",International Journal of Scientific & Engineering Research,Volume 4, Issue3, pp 20-33, March 2013.

[17] Bharati m. ramageri," Data mining techniques and applications", Indian journal of computer science and engineering, vol. 1 no. 4 301-305

[18] PrakashR.Andhale1 , S.M.Rokade2, "A Decisive Mining for Heterogeneous Data", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 12,pp. 43-437, December 2015.

[19] D.W. Jordan, "Re-thinking student written comments in course evaluations: text mining unstructured data for program and institutional assessment," Ph.D. dissertation, California State Univ., 2011.

[20] W. Himmel, U. Reincke, and H. Michelmann, "Text mining and Natural language Processing Approaches for automatic categorization of lay requests to web-based expert forums", Journal of Medical Internet Research, vol. 11, no. 3, pp. 25, 2009.

[21] Prakash R.Andhale, S.M.Rokade, "A Decisive Mining for Heterogeneous Data", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 12, pp. 43-437, December 2015.

[22] Asai, T. Arimura, H., Abe, K., Kawasoe, S. Arikawa, Online Algorithms for Mining Semi-structured Data Stream, In Proc. IEEE ICDM'02, 27–34, 2002.