ISSN: 0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 9, Issue, 1(D), pp. 23121-23131, January, 2018

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

FINDING THE CO-OCCURRENCE OF TEXT ENTITIES FROM THE CORPUS

A.Muthusamy*

Department of Computer Technology, Dr.N.G.P Arts and Science College, Coimbatore, India

DOI: http://dx.doi.org/10.24327/ijrsr.2018.0901.1401

ARTICLE INFO

Article History:

Received 17th October, 2017
Received in revised form 21st November, 2017
Accepted 05th December, 2017
Published online 28th January, 2018

Key Words:

Information Extraction, Retrieval; Text mining, Text Analytics, Statistical Computing;

ABSTRACT

The web text documents are often structured, un-structured or semi-structured format are available on the internet. Multiple web text documents are downloaded and loaded into text mining framework that aims to extract co-occurrence of the entities from textual information. To achieve this process multiple text documents with text-mining scripts using R are presented in an efficient manner. In that, Term Document Matrix and the association between the terms is statistically computed. For statistical computing, R provides a class as term-document-matrices transported from a Corpus use bag-of-words mechanism, which implies that lists all occurrences of words within the corpus and this approach results in a matrix format.

INTRODUCTION

The development of information society in recent decades has enabled collecting, filtering and storing enormous amount of information. This knowledge should be further processed to achieve valuable information and knowledge. The scientific field managing with information extraction has evolved rapidly to manage with the extent and growth of knowledge sources. Then the primeval techniques for information extraction, retrieval, knowledge acquisition have to personalize for the dynamic, heterogeneous, and unstructured data on the WWW. In 1990s, no one truly determine what type of medium was emerging. The conception of hypertext coined by Tim Berners-Lee, and the underlying technological infrastructure, Internet, was not much spread beyond some university institutions. This was to change rapidly within the following decade in a breath-taking pace. The countless of Web servers began to host countless of all kinds of documents, and the Web's dimension doubled every six months.

It became clear that the new medium had a huge potential to exploit. Sergey Brin and Larry Page recognize the amazing possibilities of what was now called the World Wide Web and to make practical attempts to turn it into something more manageable. From 1996 to 1998, they designed and enforced Google, a search engine for the Web. They were aware that the Web had one particularity that standard information retrieval (IR) systems of that time failed to handle well. This feature was the presence of hyperlinks between Web documents. In accordance to analysis targets, web mining are often divided into three different categories. Web content mining is that the method of mining information from the contents of Web sites and Web documents frequently text, images, audio, and video files [1]. Techniques utilized in this discipline are heavily drawn from natural language processing (NLP). Web structure mining is the method of analyzing the nodes and link structure of a website through the exploit of graph theory. To learn from this the structure of a website in terms of however it has connected to different sites and structure of the document on however every page is connect on to the web itself. Web usage mining is the method of extracting patterns and knowledge from server logs to achieve insight on user activity including users location, counting the number of times user visited the web site. In [1] summarized the research works done for unstructured data or semi-structured data from information retrieval analysis.

The researches follow bag of words mechanism based on the statistics about single words in isolation, to signify the unstructured text and take single word found from corpus. Then, for the semi-structured data exploit the HTML structures inside the documents. In addition, the hyperlink structure between the documents for document version is also used. As for the database view, in order to have the better information

Corresponding author:* **A.Muthusamy
Department of Computer Technology, Dr.N.G.P Arts and Science College, Coimbatore, India

management and querying on the web, web mining tries to infer within the web site structure to convert a web page to become a database. This kind of mining uses the principles of data mining and knowledge discovery technique to screen the exact data. Web makes use as a source is unfortunately more complex than working with static databases. Since, it has dynamic nature and infinite number of documents, there is a need for clarification that is not depending on accessing the complete data on the outset [2]. Another essential aspect is the appearance of query results. Due to massive size, a web query can retrieve thousands of resulting webpage's. Thus significant methods for presenting these huge results are necessary to help a user to select the most interesting content.

Problem Identification

Text Mining is dissimilar from web findings [20]. Within, the web findings, the client usually seem to be for known things, which have written by others. The purpose of the optimizing the text is to discover out unknown information which no one yet knows. Otherwise, Intellectual Text Analytics consign to mine important information with knowledge as of un-ordered text. Optimizing the text and data both are same [22], but optimization tools on data are planned to hold ordered data as of datasets and optimization tools on text are effort on un-ordered or semi-ordered data as of datasets. Examples of un-ordered data are Emails, HTML documents, etc.., since the researcher said mining, knowledge extract from text [21]. The need of Text mining is data mining which is practical to textual information. Text is unstructured, vague and difficult to deal among the data. In general, it is the most frequent method for formal exchange of information. Whereas data mining belongs in the commercial globe as that's where mainly databases are, text mining assure to move machine learning tools not in the concern and keen on residence as an increasingly necessary Internet adjunct i.e., as "web data mining" offer a recent reconsider of web data extract tool kit. Text mining is nobody apart from non-conventional information retrieval scheme intends to shrink the attempt of essential users to obtain valuable information from huge automated text data sources. The conventional information retrieval scheme concurrently recalls both fewer and a lot of information from the textual data. The non- conventional scheme correspond to a useful system that must go beyond simple retrieval as shown in Fig. 1.

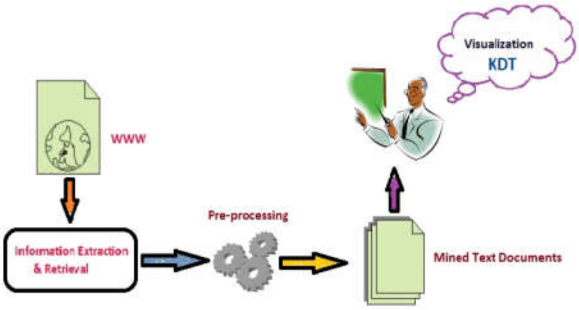

Figure 1 Flow of Text Mining Techniques for Knowledge Presentation

Why to choose R Language?

R is an free and open source software. It is an interpreted computer language and software environment for statistical computing and graphics. It is an implementation of S programming language integrated with lexical scoping semantics. S was created by John Chambers at Bell Laboratories (formerly AT&T, now Lucent Technologies). R was formed by Ross Ihaka and Robert Gentleman at the university of Auckland. R can be extended easily via packages. There are around 4000 packages available in the CRAN package repository for different tasks. Some tasks related to data mining are Machine Learning, statistical learning and graphical methods, and is highly extensible. R provides an Open Source way to contribution in that doings. R's force is the ease with which well-designed publication-quality design can be fashioned, it comprises mathematical symbols and formulae. R is available as a Free Software in which it compiles and runs on a extensive range of UNIX platforms and related systems, Windows and Mac OS. Text mining include enormous field of abstract and methods with one thing in common text as input information. It permit diverse characterization, series from an expansion of classical data mining to texts to more sophisticated exploit of large text collections to determine novel facts and fashion about the globe itself (Hearst 1999). In common, text mining is an interdisciplinary field of movement between data mining, linguistics, computational statistics, and computer science. The regular techniques are text classification, text clustering, ontology, and taxonomy creation, abstract, and underlying corpus study. In addition a lot of techniques from related fields like information retrieval are commonly used.

The modern development in document replace has conveyed valuable concepts for automatic handling of texts. The semantic web [24] propagates standardized formats for document replace to enable agents to perform semantic operations on them. It employ with metadata and by interpreting the text with HTML tags. One key format is RDF [25] efforts to handle this layout have previously been made in R with the Bio-conductor project [24, 25]. This development offers great flexibility in document exchange. But with the growing popularity of XML based formats (e.g., RDF/XML) tools required to handle XML documents and metadata. The advantage of text mining approach with the huge quantity of valuable information latent in texts which is not available in classical planned data layout for diverse causes text has always been the default way of accumulate information for hundreds of existence, and generally time, delicate and expenditure limitation prohibit us from bringing texts into well planned layout resembling data frames otherwise tables. According to a recent review on text mining products the capabilities and features include,

- Data preparation, importing, cleaning and general pre-processing,
- Association analysis means finding relationship for a given word found on counting co-occurrence frequencies.
- Clustering of related documents into the similar groups,
- Summarization is main concept in a text. Generally, these related to high frequency terms.
- Classification of texts into pre-defined categories,
- Availability of Application Programming Interfaces to expand the program with plug-ins.

Literature Review

In this part we depict the diverse techniques with various authors which are related to the Information extraction and retrieval from the web. Web Content mining [5] is completed by retrieving information from unstructured document such as open text and semi-structured document such as hypertext documents. In unstructured documents mining is done by word pose in the documents, text classification, event detection and tracking, finding extraction patterns in the text documents. The method used for semi-structured documents are hypertext classification and clustering, learning associations among web documents, learning pattern extraction or rules and finding patterns in semi-structured data [6]. The web content mining is used in various areas [7] mining online news sites. In past, for analyzing the news they observant on present people concern and measured the collective significance of ongoing assesses. Active crawler exploit for resource discovery intended for trend analysis, domain independent statistical analysis is used. The stages of active information analysis are resource recognition, pre-processing, generality and examine. In Resource recognition stage, active web crawler downloads web page from current URL and then filters the downloaded web pages and examine the recognized news information repeat the steps until the queue of URL's are empty. In pre-processing stage the news are converted into planned layout. Exciting fashion among new topics found in Generalization stage. In analysis phase user analyzes the pattern and the process is repeated until interesting news is found. Accordingly in this method, crawler downloaded 350 web pages every day and only 130 were chosen for further analysis.

An additional part where web content mining has proved extremely useful is a web content suggestion system for distance learning depicted in [9]. Two ways of plan are collaborative filtering and content based filtering. Collaborative offer clusters students into set with similar behaviour. Content based filtering offer web pages to the students having steering records. Web page navigation behaviour is stored in personal records. Students who are new to attend the course will be having less navigation record so they are asked to poll the interest. Content Suggestion system works with the facilitate of six modules such as Student Subordinate Agents, Student Recognition, Implication Generation, Suggestion Rescue, Data store. A new algorithm for Web Content mining succeeded in the information extraction from uncertainty interfaces and then goes with interrelated feature. First it mines the content of uncertainty interfaces and using clustering techniques information is extracted and they are placed in special domains. Uncertainties interfaces exist in the domain are equal with the system user query and finally uncertainty interface nearly similar to the user query chosen. Jaccard measure utilized by this algorithm to distinguish positive and negative correlated attribute [8].

The troubles oppose by Web Content mining like information extraction from heterogeneous environment, the redundancy, the associated web environment, the dynamic and noisy nature of the web were highlighted. Solutions for some of the above problems were also discussed [10, 11]. Web usage mining result can be improved by analyzing web content. The method combines web page clustering into record association mining and cluster labels are used as web page content indicators. The Web page clustering was done using K-means algorithm. The clusters obtained from the web log file and integrated data file were physically reviewed. Then the Apriori algorithm was applied. This system utilized Web content mining for web usage mining [12]. Integration of web content mining into web usage mining is also possible [13, 14]. In [13] the textual substances of the web pages are mined during frequent word sequence. Then they are combined with web server logs to study association rule of user's behaviour. The effect of the projected system helps in improved reference, personalization, creation and web user report. The relation among web content and web structure mining was examined in [15]. In this method the web page content is evaluated with the information definite formation of the web site. Every web page is explained with a group of keyword. The information combined with the relation formation produces perspective-based report. This assessment helps in discovered out web information and its neighborhood. Page Content algorithm was designed and the objective of the task was to fashion a improved algorithm than Page Rank algorithm. The significance of page determines the importance of term which the page contains. The term is computed based on a given uncertainty. For internal classification, page Content Rank utilize neural network [16]. A method was proposed it provides extraneous data along with the valuable data thus improving the effect of web content mining [18]. A review was done for web content mining and it clarifies how it can apply to business field helps to both the customer and the producer [17].

In [19] first, it defines how web mining research area focuses on mining research and retrieval research (i.e. data, information retrieval from web, optimizing the data, and text). Second, it categorizes the Web mining as content mining (i.e. information retrieval for texts, images and other contents), structure mining (i.e. fact discovery from the relation of web pages) and usage mining (i.e. Optimize the useful information from web). Web content primarily focuses on the formation of internal document whereas web structure mining aims is to find the linkage assembly of the hyperlinks at the internal document level. The usage mining includes three major phases i.e. pre-processing, pattern discovery and pattern analysis. The survey article [29] explains about the extraction and retrieval of personal name alias using various techniques from the web with the help of search engine. The presented technique help to improve the depth of knowledge relevant to alias extraction and retrieval process. It describes about how the alias name are ranked, then page counts on the web, word co-occurrence utilize anchor text and methods like term frequency (tf), inverse document frequency (idf), log likelihood ratio. Chi-squared tests etc.., are used for measuring the association and similarities between words. Rashmi Agrawal, Mridula Batra [3] proposed a exhaustive revise lying on text optimization methods in which conventional keyword findings rescue documents includes pre-definite keywords. Text mining extracts exact information found on further keywords, such as entities or perception, association, phrases, sentences and even statistical information in the environment.

- Optimized the text in a efficient, comprehensive and reproducible method, and business critical information can be confine usually.

- By using wildcard operator one can ask query without even having to know the keywords for which he seems for and still obtain high quality formation results.

Douglas E. Appelt proposed TextPro [4] is an information extraction system have broadly tested, and used in a practical software estimation. It proposes elevated presentation in terms of speediness and simplicity of domain precise growth, and is particularly permits fitted appropriate information extraction tasks such as name tagging. TextPro obtain to begin as classic guerrilla software [4] to enlarge this program for the complete excitement of Mac chopping, and constructing amazing [4] thought was both excitement and valuable. Its novel intention was to service as a test bed for the regular prototype design. In [30] has presented the need of graph mining algorithm, association orders among alias name are discussed. Searching the exact person names on the web [28] is a complicated activity when a distinct person name is shared by multiple persons. The summit people search in the web outcome for an individual name were addressed by two activities, a clustering task consists of grouping mutually web people pages consigning to the similar person, and an extraction activity, which consists of extracting relevant feature for each of the persons allocating the same name. Three data sets one for alias names, professions and personal names data sets are evaluated. The discovered results are commonly presented in a line of results often consigned to as search engine results pages (SERPs) listing of results in terms of patterns revisit by a people search engine in response to a keyword query. The person name is identified with the help of pattern extraction which simplifies the model. In addition, the description, assets, estimate method, outcome for grouping attributes extraction activity has presented.

METHODOLOGY

Web mining utilizes the data mining techniques for mining information from the web documents and services. Web Content Mining is the process of mining useful information from the contents of web pages and web documents which are chiefly encompasses of texts, pictures, audios and videos. These techniques applied in this discipline have been totally drawn from Natural Language Processing (NLP) and Information Retrieval (IR). Web Structure Mining is the process of analyzing structure (nodes and connection) of a website through the use of graph theory. To obtain the formation of a website, first, how it is linked with other web sites. Next, the document structure of the website itself as how each page is connected. Web Usage Mining is the process of mining patterns and information from server register to increase insight on user tasks. It includes where the users are from, how many clicks are to be done on the site and the kinds of performance being done on the site. Text mining is otherwise, referred as Text Data Mining. Text mining is a method of obtaining high feature information. It usually involves the process of structuring the entered text (parsing), originate patterns within the ordered data and finally evaluation and interpretation of the output as shown in Fig. 2. The flow of the text mining obtain the word documents as input, and then extracts words from the web document, and attached in the library or database. Text Mining has an ability to process the unstructured document typically the very large set of documents such as thousands or millions to deduce the implication and repeatedly recognize and extract their model as well as the association among the concept to directly respond query of their significance. Text Mining is different from web search. In web exploration, the user is frequently looking for recognized belongings and has written by others. The goal of the text mining is to discover indefinite information, which no one however knows. Text Mining is also known as Intelligent Text Analysis. It means that the essential information and knowledge can be had from the unstructured text. Text mining from the semantic web based Data Mining, IR (Information Retrieval), Machine Learning, Rule-Based Modeling of Natural Language and Statistics techniques. The majority of the information from the web (More than 80%) is stored as a text.

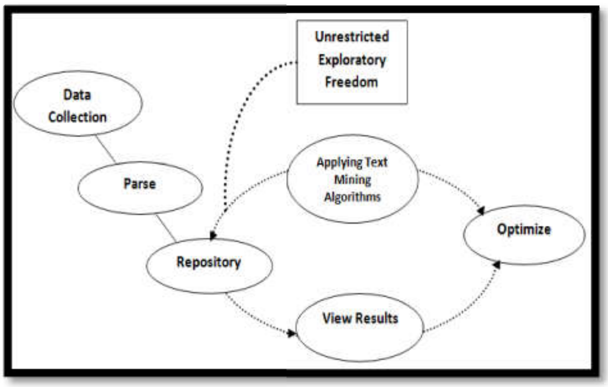

Figure 2 Text Mining Process

Text Mining and Data Mining are same, except that data mining tools planned to handle structural data from datasets and text mining tools are used to handle unstructured or semi-structured text from web, text documents and services. Examples of Unstructured data are Emails, HTML documents, etc., since the researchers assumed that text mining extracts knowledge from the web. Therefore, the problem of the KDT (Knowledge Discovery from Text) is to mine essential concepts and relating to use Natural Language Processing (NLP).

Text Mining Techniques

Text Mining Vs Web Mining: In text mining, the input is free unstructured text, but in web mining web sources are ordered.

Texts Mining Vs Data Mining: Through text mining process patterns are mined from natural language text except in data mining patterns are mined from databases.

Mining Plain Text : This section describes the major process ways in which text is extract while the input information is in simple natural language, rather than partially structured Web documents. We begin along with issues that engage mining information for individual use. Here are the various techniques, which mine the simple text similar to summarization of text, document and information retrieval, measure document similarity and text categorization.

Text Summarization and Document Retrieval

It produces a compressed representation of its input, which specifies human consumption. It too includes single documents or set of documents. Text Compression is a related area but the output of text summarization is precise to be individual legible. The outcome of text compression procedures are definitely not individual legible and besides not exploitable. It merely sustains decompression technique i.e.., automatic reconstruction of the original text. Summarization of text vary from several types of text mining in that there are people, namely professional abstractors, who are skilled in the technology of fabricating synopsis and performing the activity as part of their proficient existence. It is the task of identifying and returning the most relevant documents. Traditional libraries offer lists that permit the users to recognize documents based on resources which comprises metadata. Metadata is an extremely ordered document for review, and flourishing methodologies have been developed for manually mining metadata and intended for recognizing significant documents based on it, methodologies that are widely taught in library. Automatic mining of metadata for instance, topic, language, creator, and key-phrases is a prime application of text mining techniques. The proposal is to index all entity phrases in the document group. It specifies many effective and popular document retrieval techniques.

Information Retrieval and Accessing the Document Similarity

It is considered as an extension to document retrieval where the documents that are returned are processed to reduce or mine the exacting information required by the user. Thus, the document retrieval is followed through a summarization technique that focuses on the uncertainty cause by the client, or an information extraction stage. The documents standard might accustom at each entity part comprise a unit in its own right, in an endeavor to focus outcome on individual piece of information quite than prolonged documents. Numerous text mining issues involve assessing the similarity between different documents, for instance, conveying documents to be already defined type and set of documents into usual clusters. These are the fundamental issues in data mining besides, and have been a centre for research in text mining, perhaps, the success of diverse techniques might estimate and matched with using standard, objective and measures of success.

Text Categorization

Text categorization is the task of natural language processing the documents to be predefined categories with their text substances. The group of categories is frequently known as controlled terminology. The text categorization is an extensive repute usual technique for information retrieval in documents, where topic rival creator as the leading entry to library contents although they are future harder to consign objectively than originator. The systematized text categorization has numerous realistic purposes. It includes index of document retrieval, metadata extraction, word sense disambiguation with identifying the theme of a document conceal, organizing and maintaining huge lists of web resources. As in other areas of text mining, until the 1990's text categorization was conquered by adhoc method of data engineering that required obtaining categorization policy from individual professional and coding them into a system that could apply them automatically to new documents. Since then and particularly in the research community the dominant approach has been to utilize methods of machine learning to assume the categories systematically from training set that already defined in the categorize documents. Recently, text categorization is a talking subject in machine learning. The already defined categories are figurative design with no extra definition. While categorizing a document, excluding information are not used for the content of the documents. The several tasks limit the document to a distinct category, but for others all documents might have numerous categories. Sometimes category labeling is probabilistic rather than deterministic, or the objective is to rank the categories by their estimated relevance to a particular document. Sometimes documents are processed one by one, with a given set of classes; otherwise there might be a distinct class possibly a latest version that has been added to the group and the task was determined in which it contains the various documents. The numerous machine learning method have employed for text categorization.

Mining Structured Text

The Internet includes an unambiguous basic mark-up tag that usually varies from simple text content. Some mark-up is internal and indicates document structure or format; some is external and gives explicit hypertext links between documents. These information sources give added benefits for optimizing the web documents. The individual web page designer collects the sources of information that are very noisy as they engage random and unpredictable options. However, these demerits are offset by the total amount of existing data, which is fairly impartial since it is combined more numerous dissimilar information providers. Therefore, Web mining is rising as a recent subfield, related to text mining but catching the benefit of additional information obtainable in web documents, mainly hyperlinks and still realizing on the existence of topic directories in the web itself to develop better results. In brief re-examine three methods of extracting ordered text. In primarily, wrapper initiation utilizes inside mark-up information to raise the efficiency of text mining in marked-up documents. Next, document grouping and determining the ability of web documents, exploit on the exterior mark-up information that is present in hypertext in the form of explicit links to other documents.

Wrapper Induction

An Internet resource includes relational data. For instance, telephone information bank, product catalogue, etc., utilize formatting mark-up to clearly present the information to the users. In spite of, among basic HTML, it is relatively complex to mine data from such resources in a systematic way. The XML mark-up language is intended to defeat these issues by encouraging page authors to mark their content in a way that reflects document structure at a detailed level; but the user is not cleared to know how to share the XML document structure, and still if they perform enormous numbers of legacy web pages flourish. Many software systems use external online resources by hand-coding effortless parsing section, frequently known as wrappers, to analyze the web page formation and mine the essential information. This is a breed of text mining, but individual build upon the input having a permanent, determined formation in which the information might be mined algorithmically. Agreed that this hypothesis is fulfilled, the information extraction issue is fairly insignificant but this is unusually emerging. Web page formation is different; errors that are insignificant to human readers throw automatic extraction procedures completely through the dynamic web sites. There is a well-built wrapper for systematic induction to diminish the issues with small alteration happens, and to formulate it clear to construct recent sets of extraction policy when structures change completely.

Text Mining Frameworks

Text mining is a process that employs a set of algorithms for converting unstructured text into structured data objects and the quantitative methods used to analyze the data objects. The benefit of text mining comes with the large amount of valuable information latent in texts which is not available in classical structured data formats for various reasons: The text has always been the default way of storing information from the century, primarily period, delicate and expenditure limitation forbid us from bringing texts into well structured formats (like data frames or tables). The statistical contexts for text mining applications in research and business intelligence include business applications making use of unstructured texts. Recently, all most significant statistical computing invention propose text mining latent, and numerous familiar data mining inventions offer result for text mining activities. It includes the features as,

- By operating the pre-processing technique on textual data it incorporates the formation, significance and uncontaminated of data.
- The **r**elationship investigation on textual data support the mutual discovery of a known term based on counting the co-occurrence frequencies.
- Clustering method collects the linked textual documents formed into the similar position.
- Textual data summarization is an important technique to identify the most frequent terms.
- Categorization is the process of ordering the textual data interested in already defined class.
- Application Programming Interfaces aid towards to access the program with its peripheral.

In Table 1. Shows an overview of the most used commercial products intended for text mining with certain freeware tool and aspects. It seems that most commercial tools lack easy-to-use API integration and provide a relatively monolithic structure regarding extensibility since their source code is not freely available. Among well known open source data mining tools offering text mining functionality is Weka suite, a collection of machine learning algorithms for data mining tasks also offering classification and clustering techniques with extension projects for text mining like KEA for keyword extraction. It provides good API support and has a wide user support. Then there is GATE, an established text mining framework with architecture for language processing, information extraction, ontology management and machine learning algorithms. The former tools are Rapid Miner, a system for knowledge discovery and data mining, and Pimiento a basic Java framework for text mining. However, many existing open-source products tend to offer rather specialized solutions in the text mining context, such as Shogun, a tool kit for string kernels, or the Bow tool kit, a C library useful for statistical text analysis, language modeling and information retrieval.

Text Mining in R

The extension package which offer interface to existing text mining tool kit which integrate well with the tm package, and are also freely available at Comprehensive R Archive Network. The tm package suggests functionality for managing text documents, conceptual the process of document strategy and simplicity the usage of heterogeneous text design in R. The package has integrated database backend sustain to reduce memory insists. A highly developed Metadata management is employed for collections of text documents to improve the usage of huge and with metadata enhanced document sets. With the package distribute local sustain for conduct the Reuters 21578 dataset, Gmane Rss feeds, e-mails, and several classic file formats (e.g. plain text, CSV text and PDFs).

Table 1 Overview of Text Mining Products with Available Features Marked as Tick (✓) Sign

Products	Pre-process	Associate	Cluster	Summa rize	Categorize	API
			Commercial			
ClearForest	✓	✓	✓	✓		
Copernic Sum	✓			✓		
dtSearch	✓	✓		✓		
Insightful Infact	✓	✓	✓	✓	✓	✓
Inxight	✓	✓	✓	✓	✓	✓
SPSS Clementine	✓	✓	✓	✓	✓	
SAS Text Miner	✓	✓	✓	✓	✓	
TEMIS	✓	✓	✓	✓	✓	
WordStat	✓	✓	✓	✓	✓	
			Open Source			
GATE	✓	✓	✓	✓	✓	✓
RapidMiner	✓	✓	✓	✓	✓	✓
Weka/KEA	✓	✓	✓	✓	✓	✓
R/tm	✓	✓	✓	✓	✓	✓

The data structures and algorithms might be expanded to robust as custom demands, the tm package is designed in a modular way to facilitate effortless integration of new file formats, readers, transformations and filter operations. It (tm) provides easy access to pre-processing and manipulation mechanisms such as whitespace removal, stop words removal, and stemming. It supports exportation from document collections to term-document matrices simply build from web text documents as shown in Fig. 3. To recognize the related terms from corpus procedures should be followed systematically.

The contributions of the work summarized as follows,

- The essential Text mining functions tasks such as Pre-processing, Data Cleaning are perform using R.
- In-order to rank the terms, Term Document Matrix (TDM) or Document Term Matrix (DTM) has created.
- In addition, to recognize the strength of association among the terms has computed statistically.
- Word cloud and Histogram have graphically created to identify the most frequent terms from the corpus.

Recognition of Related Terms Extraction Algorithm

The recognition of related terms extraction algorithm is as follows,

Step 1: The downloaded text document from the web is loaded into an R environment using corpus() transformation is performed using tm_map() function to replace, special characters from the text.

Step 2: The tm_map() function is used to remove unnecessary white space, to convert the text to lower case, to remove common discontinue words like “the”, “we”, ”for” etc.., and punctuation mark is also removed from corpus then the extracted text is transformed.

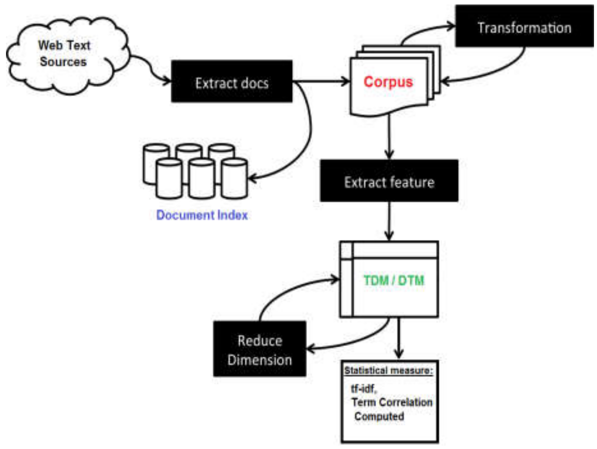

Figure 3 The Process of Text Mining Functions in R

Step 3: In-order to rank the terms again loads the corpus; build a Term Document Matrix to explore frequent terms and their association. Term Document Matrix contains the frequency of the words represented by column as words and row as document name generated.

Step 4: Load the processed text document to Open Natural Language Processing it supports the most common tasks, such as tokenization, sentence segmentation, part-of-speech tagging, chunking, parsing, and named entity extraction have been performed.

Step 5: For Graphical representation, a word cloud, and histogram have presented to exhibit the frequently occurring words in the corpus.

Implementation

To train and evaluate the proposed method, the information is extracted and retrieved from the web with the help of query as “dhoni * cricket” these will return URL of the web pages relevant to any person with the name as Mahendra Singh Dhoni. The text mining functions implemented are,

Install and Load required packages

Text mining and Wordcloud packages are required. They can be installed and loaded using the R code below,

```
# install
install.packages("tm") # for text mining
install.package("SnowballC") # for text stemming
install.packages("wordcloud") # word-cloud generator
install.packages("RColorBrewer") # color palettes
```

Initially load the tm package as this is not loaded by default. This is done using the library() function like,

```
library(tm)
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
```

The dependent packages are loaded automatically.

Load the Text

The collection of documents extracted from the web placed in a TextMining folder. In R, we identify it by using file path. It involves loading the multiple text documents availed in the TextMining folder into a Corpus object. The tm package offers the Corpus() function to do this task. There are different ways to build a Corpus. The option we will use

```
# Create Corpus - Change Path as Needed
docs1 <- Corpus(DirSource("D:/extraction"))
```

Few line that starts by # represent comment line, and the “<- “ tells R to assign the result to the right hand side variable or assignment operator. In this case, the Corpus object created is stored in variable called docs1. Equals sign (=) is another operator for assignment is permitted in R. The content of the document can be inspected in the R terminal by using the below R code as,

```
# Check detailsinspect(docs1)
```

If you choose to seem merely one of the documents load, then it can be specified which one using some extent as,

```
#To examine a particular document
inspect(docs1[1])
```

```
# Another way to represent
as,writeLines(as.character(docs1[[1]]))
```

Pre-processing

Transformation is performed using tm_map() function to replace, for example, special characters from the text.

```
# Replacing “/”, “@” and “|” with space
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))

docs1 <- tm_map(docs1, toSpace, "/")
docs1 <- tm_map(docs1, toSpace, "@")
docs1 <- tm_map(docs1, toSpace, "\\|")
docs1 <- tm_map(docs1, toSpace, ":")
docs1 <- tm_map(docs1, toSpace, """")
docs 1<- tm_map(docs1, toSpace, " -")
```

Cleaning the text

The purpose of **tm_map()** utilized to eliminate redundant white space, to convert the text to small letter, to eliminate common discontinue words like “the”, “we”. The information worth of “stop words” is near zero as they are so common in a language. Eliminating this type of words is useful before further analysis. The stop words, sustain languages are danish, dutch, english, finnish, French etc., Note that the language names are case sensitive. It can also possible to remove numbers and punctuation with **removeNumbers** and **removePunctuation** arguments. Another important pre-processing step is to formulate a **text stemming** which ease words to their origin form. In addition, this process eliminates suffixes from words to make it uncomplicated and to get the common origin. For instance, a stemming process ease the words “moving,” “moved” and “movement” to the root word as “move.”

```
    # Convert the text to lower case
    docs1 <- tm_map(docs1, content_transformer(tolower))
    # Remove numbers
    docs1 <- tm_map (docs1, removeNumbers)
    # Remove english common stopwords
    docs1<-tm_map(docs1,removeWords,
    stopwords("english"))
    # Remove your own stop word
    # specify your stopwords as a character vector
    docs1<-tm_map(docs1,removeWords,c("blabla1",
    "blabla2"))
    # Remove punctuations
    docs1 <- tm_map(docs1, removePunctuation)
    # Eliminate extra white spaces
    docs1 <- tm_map(docs1, stripWhitespace)
    # Text stemming
     docs1 <- tm_map(docs1, stemDocument)
```

Experiment results

Build term document matrix(tdm or dtm)

It is a table include the occurrence of the words. Column names are words and row names are documents. The purpose of TermDocumentMatrix() in the text mining wrap up utilized as follows,

```
# create document-term matrix
dtm1 <- DocumentTermMatrix(docs1)
```

These construct document term matrix on corpus and save the outcome in the variable of dtm1. We get summary information on the matrix by entering the variable name in the console and hitting revisit as,

```
dtm1

  <<DocumentTermMatrix (documents: 3, terms: 2125)>>
           Non-/sparse entries   : 2288/4087
           Sparsity              : 64%
           Maximal term length : 45
           Weighting             : term frequency (tf)


# View first 3 docs1 & first 9 terms
inspect(dtm1[1:3, 1000:1008])
```

```
<<DocumentTermMatrix (documents: 3, terms: 9)>>
Non-/sparse entries  : 10/17
Sparsity             : 63%
Maximal term length: 9
Weighting            : term frequency (tf)
Terms
Docs1         gautam gavaskar gave geet george gerry getting
ghosts gilchrist
 Fin1.txt       0        0      0    0     0     0       0
0       0
 test1.txt      2        4      0    2     4     0       0
0       0
 test2.txt      1        0      1    0     0     1       1
1       2
```

This command displays terms 1000 through 1008 in the first two rows of the DTM. Note that the results may differ.

Mining the Corpus

By constructing the TDM, we have converted a corpus of text into a mathematical object that can be analysed using quantitative techniques of matrix algebra. Therefore, that the TDM (or DTM) is the starting point for quantitative text analysis. To compute the frequency of occurrence of each word in the corpus, we simply sum over all rows to give column sums

```
freq <- colSums(as.matrix(dtm1))
```

Here initially we concealed the TDM into a exact matrix using the as.matrix() function. We have then summed above all rows to give us the total for each column. The result is stored in the (column matrix) variable freq. To check the dimension of frequency equals the number of terms as,

```
# length should be total number of terms
length(freq)
[1] 2125
```

Next, we sort frequency in descending order of term count.

```
# create sort order (descending)
ord1 <- order(freq,decreasing=TRUE)
```

Then list the majority and slightest frequently occurring terms:

```
freq[head(ord1)]

dhoni   india    odi   test   world cricket
  82     130      75    51     51      42
```

```
# inspect least frequently occurring terms
freq[tail(ord1)]
zealand   zealand,   zealand. zealandmclean  ziva. zone[35]
    1          1          1            1                 1
```

```
# List the majority of frequent terms. Lower bound specified as
second argument
findFreqTerms(dtmr,lowfreq=40)
[1] "cricket" "dhoni"  "india"  "odi"    "test"   "world"
```

The function findFreqTerms() returns all terms that arise greater than 40 times in the entire corpus. However, outcome is sorted alphabetically not by regularity. At present, we have most frequently occurring terms, we can check for correlations among few terms that occur in the corpus. In this context, correlation is a quantitative measure of the co-occurrence of words in multiple documents.

Relationship between Terms and Correlation

The tm package provides the findAssocs() function to specify the DTM, the term of interest and the correlation limit. If the final number lies between 0 and 1 it provides a lower bound for the strength of correlation between the search and outcome terms. For instance, if the correlation limit is 0.98, findAssocs() will return the words that forever co-occur with the discovered term. A correlation limit of 0.5 will return terms that have a search term co-occurrence of at least 50% and so on. Here the results of running findAssocs() on some of the frequently occurring terms at a correlation of 98%.

```
# State a correlation limit of 0.98
findAssocs(dtm1, c("name"), corlimit=0.98)
$name
```

mahi,	msd	thala
0.98	0.98	0.98
Cool,	dhoni	keeper,
0.98	0.98	0.98
Tournament	praveenkumar	
0.98	0.98	

Creation of Word cloud

The importance of terms can be illustrated in Fig.4

```
# wordcloud
library(wordcloud)
#setting the same seed each time ensures consistent look across clouds
set.seed(142)
# limit words by specifying min frequency
wordcloud(names(freq),freq, min.freq=25)
# add color
    wordcloud(names(freqr),freqr,minimum.frequency=100,
colors=brewer.pal(6,"Dark2"))
```

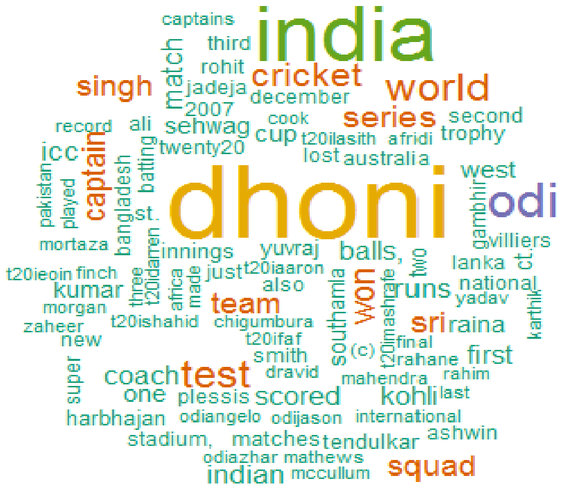

Figure 4 Cloud representation of Mahendra Singh Dhoni

Histogram

The initial code generates a data frame a listing columns of equal length. A data frame also contains the column name in this casing these are term frequency respectively. We then invoke ggplot(), telling it to consider plot only those terms that occur more than 30 times. The aes option in ggplot describes plot aesthetics in this case, we use it to specify the x and y axis labels. Then stat="identity" option in geom_bar() ensures that the height of each bar is proportional to the data value that is mapped to the y axis (i.e occurrences). The final line specifies that the x axis labels must be at a 45 degree and horizontal. It can be depicted in Fig. 5.

```
#histogram

wf1=data.frame(term=names(frequency), occurrences=
frequency)library(ggplot2)
p1 <- ggplot(subset(wf1, frequency > 30), aes(term,
occurrences))
p1 <- p1 + geom_bar(stat="identity",fill = "light blue",    colour
= "red")
p1 <- p1 + theme(axis.text.x=element_text(angle=45, hjust=1))
p1 <- p1 + theme(text = element_text(size=25,color = "black"))
p1
```

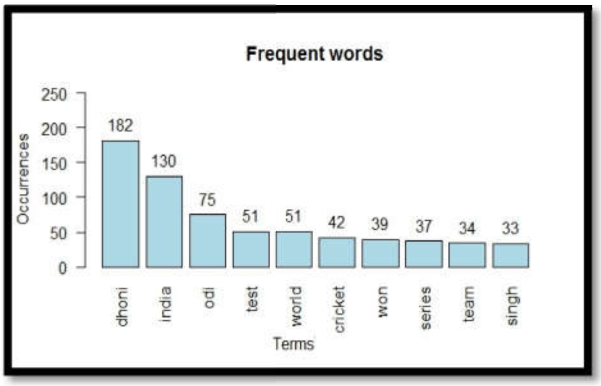

Figure 5 Term-occurrence histogram (freq>30)

Chi Square Test

Chi square analysis is valuable in verifying the statistical impact level of association or correlation. In addition, use to investigate whether distributions of categorical variables differ from one another. To do this test has observed that most frequent terms X_i through mining the corpus and NLP tasks represents several terms that occur when terms in X are present or absent in a dataset. There will be 2^k such terms since there are k terms in X that could either be present or absent in a dataset. The probabilities of these 2^k events can be measured from the data and the observed probabilities can be denoted as $O(X_i)$. These probabilities might estimated by multiplying the probabilities of occurrence or absence of items in X. The estimated probabilities are denoted as $E(X_i)$. The X^2 value of X formulated as follows,

$$X^{2(X)} = \sum (O(X_i) - E(X_i))^2 / E(X_i)$$

Data Set

In general, the training dataset has constructed with the use of observed values as frequency of 100 terms from mining the corpus. In addition, NLP tasks used to compute the expected values. So far, 300 text documents have analyzed from that it yields a significant value of each item has recorded in the Table 2. This test proves the statistical method assessing the goodness of fit between a set of observed values and those expected values.

Script

In Figure. 5 Shows how to evaluate and execute Chi-Square Test in R window. The functions and variables are described in the script as follows,

- The function chisq.test() carry out the chi square tests and goodness of fit tests.
- x variable represents a numeric vector or matrix.
- p is a vector of probabilities of the same length of x. An error was exhibit if any entry of p is negative.

```
x<-
read.csv("D:/extraction/PersonDataSet.csv",header=T,sep=",",d
ec=".")
out <- file("output_chisq.txt","w")
title <- names(x)
writeLines(paste(title[1],title[2],title[3],title[4],title[5],"Chisq P
Value", sep=","),    con=out,sep="\n")
```

```
xR <- nrow(x)
mat<- array(dim=c(2,2))
for (i in 1:xR)
{
mat [1,] <- c(x[i,2],x[i,3])
mat [2,] <- c(x[i,4],x[i,5])
pv<- chisq.test(mat)$p.value
writeLines(paste(x[i,1],x[i,2],x[i,3],x[i,4],x[i,5],pv,sep=","),con
=out,sep="\n")
}
close(out)
```

Table 2 Person Name Data Set Includes Most Frequent Terms with Observed and Expected Values

Terms	Observed Unique Terms O(X_i)	Expected Unique Terms E(Xi)	Duplicate Terms Observed O(Xi)	Duplicate Terms Expected E(Xi)
Dhoni	182	220	315	470
India	130	250	303	340
odi	75	220	400	465
test	51	225	125	160
world	51	230	100	156
Cricket	42	210	120	200
Mahi	40	150	90	120
series	37	60	54	80
Captain cool	31	100	30	45
Indian	28	70	60	50
MSD	25	110	70	50
Thala	21	60	20	40
WicketKeeper	20	15	6	30

The experimental results shown in Table 3. Comprises Unique Terms Expected as E(X_i) and Unique Term Observed as O(X_i). To do this test has observed the most frequent terms X_i through mining and NLP tasks in which terms in X might be present or absent. The term frequencies among those words have plotted separately. It clearly states that the variations among O(X_i) is low than the E(X_i) values. In addition, The experimental results shown in Table 3. Comprises Duplicate Terms Expected as E(X_i) and Duplicate Terms Observed as O(X_i). To do this test has examined the most frequent terms X_i through mining and NLP tasks in which terms in X might be present or absent. Here, the duplicated term frequencies among those words have plotted separately. It clearly states that the variations among E(X_i) is high than the O(X_i) values.

Table 3 Chi-Square Probability (P) Value

Terms	Observed Unique Terms O(X_i)	Expected Unique Terms E(Xi)	Duplicate Terms Observed O(Xi)	Duplicate Terms Expected E(Xi)	Chi square P Value
Dhoni	182	220	315	470	0.101
India	130	250	303	340	7.082
odi	75	220	400	465	5.270
test	51	225	125	160	1.700
world	51	230	100	156	1.238
Cricket	42	210	120	200	6.789
Mahi	40	150	90	120	5.555
series	37	60	54	80	0.845
Captain cool	31	100	30	45	0.020
Indian	28	70	60	50	0.002
MSD	25	110	70	50	1.243
Thala	21	60	20	40	0.441
WicketKeeper	20	15	6	30	0.009

$X^{2(X)} = \sum P_i$ / Total No. of Terms = 30.29 / 100 => **0.302**

The value 0.302 is close to zero, indicating that the observed negative correlation between the observed and the expected unique terms result is probably due to a random chance, than any actual relationship between them.

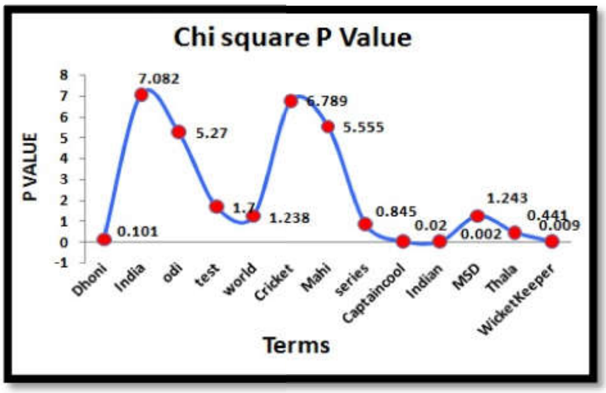

Figure 6 Probability (*P*) Value using Chi-Square Test

The Chi-square probability (*p*) values have computed statistically with the help of parameters as Unique Terms Expected, Unique Terms Observed, Duplicate Terms Expected, and Duplicate Terms Observed are denoted as O(X_i) and E(X_i). Whereas,

O(X_i) – Observed values
E(X_i) – Expected values

In Fig. 6. It clearly depicts that a parabola curve for the probability (P) value is achieved. It proves the statistical method assessing the goodness of fit among numerous frequent terms with set of observed values and those expected values.

CONCLUSION

In this paper text analytics for multiple text documents are performed. The several text mining functions are implemented using R language. In that, Term Document Matrix and the association between the terms were computed statistically which, yields a correlation limit of 98%. The Chi-Square tests yields a value of 0.302 is close to zero prove the statistical significance level of association or correlation between the terms. The word cloud and histogram are presented graphically to represent frequent terms and recognize the best term. In future, the machine learning algorithm is introduced with this model to produce further text classification and also for ranking the documents within corpus.

Acknoowledgement

I would like to thank the Management of Dr.N.G.P. Educational Institutions and Principal, Dr.N.G.P. Arts and Science College, Coimbatore for the precious support.

References

1. R. Kosala, H. Blockeel, “Web Mining Research: A Survey”, in SIGKDD Explorations 2(1), ACM, July 2000. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.
2. T. Berners-Lee, R. Cailliau, A. Loutonen, H. Nielsen, and A. Secret. The World-Wide Web. Communications of the ACM, 37(8):76 - 82, 1994.
3. Rashmi Agrawal, Mridula Batra, “A Detailed Study on Text Mining Techniques”, *International Journal of Soft Computing and Engineering (IJSCE)* ISSN: 2231-2307, Volume-2, Issue-6, January 2013 pp: 118-121.

4. Appelt, D.E. (1999) “Introduction to information extraction technology.” Tutorial, Int Joint Conf on Artificial Intelligence IJCAI’99. Morgan Kaufmann, San Mateo. Tutorial notes available at www.ai.sri.com/~appelt/ie-tutorial.
5. Faustina Johnson, Santosh Kumar Gupta,” Web Content Mining Techniques: A Survey”, *International Journal of Computer Applications* (0975 – 888) Volume 47– No.11, June 2012 pp:44-50.
6. Kosla, R. and Blockeel, H. 2000. Web Mining Research: A Survey. SIG KDD Explorations. Vol. 2, 1-15.
7. Torreblanca, A. M., Gomez, M. M. and Lopez, A. L. 2002. A Trend Discovery System for Dynamic Web Content Mining. Proceedings of the 11th International Conference on Computing.
8. Ajoudanian, S. and Jazi, M. D. 2009. Deep Web Content Mining. World Academy of Science*, Engineering and Technology* 49.
9. Yang, C. Y., Hsu, H. H. and Hung, J. C. 2006. A Web Content Suggestion System for Distance Learning. *Tamkang Journal of Science and Engineering*. Vol. 9, No. 3, 243-254.
10. Liu, B. and Chiang K. C. 2004. Editorial Special Issue on Web Content Mining. ACM. *Journal of Machine Learning Research* 4, 177-210.
11. Singh, B. and Singh, H. K. 2010. Web Data Mining Research: A Survey. Computational Intelligence and Computing Research (ICCIC). IEEE International Conference, 1-10.
12. Guo, J., Keselj, V. and Gao, Q. 2005. Integrating Web Content Clustering into Web Log Association Rule Mining. Springer Verlag. Vol. 3501 LNAI, 182-193.
13. Taherizadeh, S. and Moghadam, N. 2009. Integrating Web Content Mining into Web Usage Mining for Finding Patterns and redicting User’s Behaviors. *International Journal of Information Science and Management*. Vol. 7, Issue No. 1.
14. Kazienko, P. and Kiewra, M. 2003. Link Recommendation Method Based on Web Content and Usage Mining. New Trends in Intelligent Information Processing and Web Mining Proc. of the International IIS: IIPWM ’03 Conference. Advances in soft Computing, Springer Verlag. 529-534.
15. Gedov, V., Stolz, C., Neuneir, R., Skubacz, M. and Siepel, D. 2004. Matching Web Site Structure andContent. ACM. Proceedings of the 13th International World Wide Web Conference on Alternate track papers and posters.
16. Pokorny, J. and Smigansky, J. 2005. Page Content Rank: An Approach to the Web Content Mining. In proceedings of IADIS International Conference Applied Computing. Algarve, Portugal.
17. Ahmed, S. S., Halim, Z., Blaig, R. and Bashir, S. 2008. Web Content Mining: A Solution to Consumers Product Hunt. *International Journal of Social and Human Sciences* 2, 6-11.
18. Poonkuzhali, G., Thiagarajan, K., Sarukesi, K. and Uma G. V. 2009. Signed Approach for Mining Web Content Outliers. World Academy of Science, Engineering and Technology 56.
19. Karan Sukhija,” Web Content Mining equipped Natural Language Processing for handling web data”, *International Journal of Computer Applications Technology and Research* Volume 4– Issue 3, 209 - 213, 2015, ISSN:- 2319–8656.
20. Berry Michael W., (2004) -Automatic Discovery of Similar Words‖, in -Analysis Survey of Text Mining: Clustering, Classification and Retrieval”, Springer Verlag, New York, LLC, 24-43.
21. Vishal Gupta, Gurpreet S. Lehal, -A Survey of Text Mining Techniques and Applications”. *Journal of Emerging Technologies In Web Intelligence*, Vol. 1, No. 1, August 2009.
22. J.H. Kroeze, M.C. Matthee, -Differentiating between data-mining and text-mining terminology‖ , South *African Journal of Information Management*, Vol.6(4) December 2004.
23. D. Bollegala, Y. Matsuo and M. Ishizuka, “Automatic Discovery of Personal Name Aliases from the Web,” IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 6, (2011) June.
24. Tim Berners-Lee, James Hendler and Ora Lassila, "The Semantic Web", Scientific American, May 2001, p. 29-37.
25. Manola F, Miller E (2004). RDF Primer. World Wide Web Consortium. URL http://www.w3.org/TR/rdf-primer/.
26. Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dodoit, S., editors (2005). Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer - Verlag.
27. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80.1-16.
28. Mr.A.Muthusamy and Dr.A.Subramani “Lexical Pattern Extraction from Data Set Make Use of Personal Name Aliases”, in *International Journal of Advancements in Computing Technology* ISSN: 2005-8039(print), 2233-9337(online) Vol 7, No.3 May 2015,pp. 102-108.
29. Mr.A.Muthusamy and Dr.A.Subramani “ A Survey of Automatic Extraction of Personal Name Alias from the Web”, in *International Journal of Signal Processing, Image Processing and Pattern Recognition* ISSN: 2005-4254 Vol. 7, No. 6 (2014), pp. 75-84.
30. A.Muthusamy, Dr.A.Subramani, “Second or Higher Order Associations between a Name and Candidate Name Aliases”, *Journal of Computer Applications* (JCA) ISSN: 0974-1925, Volume V, Issue 3, 2012.

****