Visual Analysis of Large Data Text —— An Empirical Study Based on Sohu News Data

Shi Chen[1], **Zheng Li** [2], **Yijing Li** [3] and **Xiaosong Wu***
*Corresponding author email id: WXSZWD@YNUFE.EDU.CN

Date of publication (dd/mm/yyyy): 19/05/2017

***Abstract* — The arrival of the era of big data, making it very important to extract value from the massive data. Data mining on the basis of the new requirements. Open source R software integrates a variety of data analysis and visualization methods, with powerful data analysis capabilities and good scalability, suitable for data mining. In this paper, the application of R software in the main stage of text visualization is given based on the case of Sohu news data.**

***Keywords* — Big Data, R Language, Word Cloud**

I. Introduction

With the new information on blogs, social networks, location-based services LBS as the representative of the release of emerging, and the rise of cloud computing, networking and other technology, data is a hitherto unknown speed in constant growth and accumulation, the era of big data has come to [1]. Big data contains a huge value, has important strategic significance on the social, economic, scientific research, provide rich information for the hitherto unknown people better perception, understanding and control of the physical world.

The R language is a language environment and software system which has the function of statistical analysis as well as the powerful drawing function. It was founded by Ross Ihaka and Robert Gentleman of the Department of statistics, Oakland University, new zealand. R language in the GNU protocol issued free of charge, the source code can be downloaded free to use, there are compiled executable version can be downloaded. R language can be run on a variety of platforms, including UNIX (including FreeBSD and Linux), Windows and MacOS. R language development and maintenance of the core group (R Development Core Team) is responsible for the development of the R, the team members are mostly from the University of statistics and related departments. In addition to these developers, the R language also has a large number of contributors, they write code for the R language, modify program defects and document writing.

R language is the most attractive place is that it comes with a variety of statistical and digital analysis capabilities, and can be installed through the package to enhance the new features. So far, the number of packages on the official website of the R language has more than 4000, covering a wide range of industries and fields of data analysis applications. At the same time, the corresponding computer programs of all kinds of advanced statistical methods will be implemented in the form of R software package. Although the R language is mainly used for statistical analysis or the development of statistical software, but it was also used for matrix calculation. The analysis speed can be comparable to the free software and commercial software for matrix computation. Another advantage of the R language is the drawing function, which has the quality of printing and the addition of mathematical symbols.

With the explosive growth of Internet data, the traditional technology architecture is increasingly difficult to meet the needs of massive data processing. In order to solve the problem of data storage and data query, a lot of new technologies and tools have been developed. R software [2], [3] is an integrated open source software for data manipulation, statistics and visualization, which effectively overcomes the shortcomings of commercial data mining tools. A major advantage of R software is that analysts can use a simple R programming language to describe the processing process to build a powerful analysis function. In this paper, the Sohu news data analysis as an example, through a specific case, to explore the application of R software in a major stage of data mining [4]. First, using R software data mining technology on news data information was extracted; then the information obtained by using the R software package to function word segmentation, word segmentation, preprocessing, integration, classification, statistical information, form the corpus for text analysis. The finished data using R software visualization tools, visualization of the final analysis of the word cloud; R text processing software [5] data visualization and its feasibility and advantages, interpretation of the news in the big data environment importance.

II. Data Sources and Processing Methods

A. Data Sources

The data were collected in the laboratory (http://www.sogou.com/labs), Sogou and the development of the site for research for the Chinese Internet network information and network user behavior of Sogou company, to provide free data set. The use of the data set from Sohu news June 2012 -7 during the domestic, international, sports, social, entertainment and other 18 channels of news data, providing URL and text information.

B. Data Processing Methods

a) Word Cloud

Word cloud [6] is the use of language analysis technology, large data text for word frequency analysis, and the generation of visual image technology. From the computer department of Tsinghua University natural language text analysis laboratory developed Chinese "word cloud is just like an open door key data - one hundred thousand words," it only takes a few seconds to read, can quickly generate "trend, visualchart. In a more open and transparent, the public and the media have synchronous access to large data capacity of the era, in the map reading, shallow reading age, the report is the value of those looks out of order data screening, analysis, interpretation, readers "or" together to find the truth behind the data and readers. This requires human vision, but also the need for intelligent technology, the word cloud is one of the leading.

b) Key Technical Route

This paper gets the data set, the data sets will be converted into the corresponding format; and then imported into the R environment, extracted from the text content of the XPath language will be required for operation, generating a data frame; then loading the Rwordseg package to a text file for word processing, the use of TM components to calculate frequency. In the word frequency file to extract valuable vocabulary; finally, through the word cloud rendering word cloud.

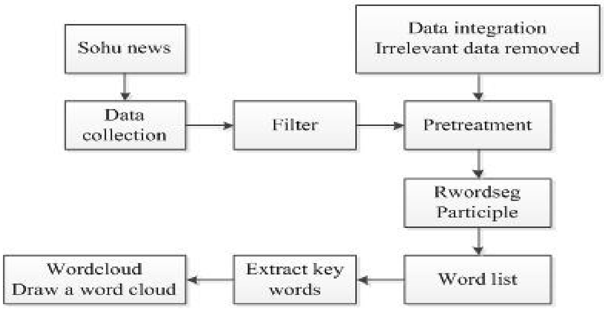

Fig.1 key technical route

c) Data Preprocessing

The existence of incomplete, noisy and inconsistent data is a common feature of a large database or data warehouse in the real world [7]. Therefore, the data must be pre processed before executing the data mining algorithm for the original data. Data cleaning is the most commonly used data preprocessing technology, through the data cleaning can add missing values, remove noise data, identify or eliminate outliers to ensure that the data：

```
parsed_views<- xmlParse(file = "views.xml")
views_content<- xpathS Apply(doc = parsed_views, path
= "//content", fun=xmlValue)
write.table(views_content.df, file =
"C:\\Users\\cs\\Desktop\\SOHU\\views_content.txt")
# output segmentation
segmentCN("C:\\Users\\cs\\Desktop\\SOHU\\views_cont
ent.txt", returnType="tm")
# word frequency statistics
views_content = views_content[views_content!=" "]
words = unlist(lapply(X = views_content, FUN =
segmentCN))
word = lapply(X = words, FUN = strsplit, " ")
v = table(unlist(word))
v = sort(v, decreasing = T)
write.table(v, "C:\\Users\\cs\\Desktop\\SOHU\\v.txt",
col.name = c("Vocabulary", "word frequency"),
row.name = F,sep="")
# picture word cloud
mydata<-
read.table("C:\\Users\\cs\\Desktop\\SOHU\\v1.txt",head=
TRUE)
mycolors<- brewer.pal(8,"Dark2")
wordcloud(mydata$ Vocabulary, mydata$ word
frequency,random.order=FALSE,random.color=FALSE,c
olors=mycolors,family="myFont3")
```

III. News Text Visualization Results Show

In this paper, we use the R software to analyze the data of the news text. The visual display is shown in Figure 2 below.

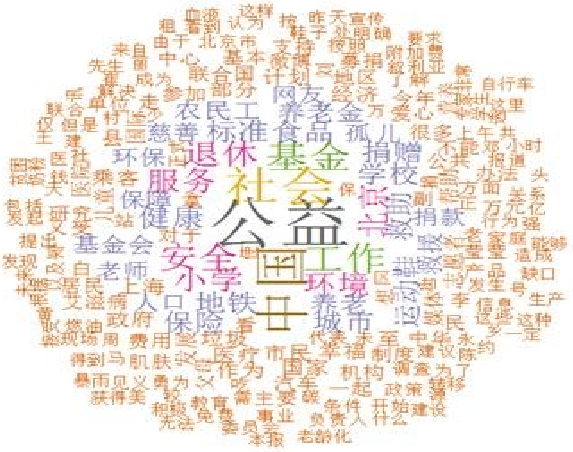

Fig. 2 word cloud[1]

The news text data preprocessing, the word cloud can see: the news media reported on the public are also included, China, social safety, environment, meeting and other related words. At the same time, the media for primary school, pension, relief, donations and other related reports, indicating that the media's attention to vulnerable groups. In addition, through the observation of the word cloud, we can find that the relationship between society and public welfare is higher.

IV. Conclusion and Prospect

A practical data mining system, on the one hand, need to have a good mining function, on the other hand, need to have a good user interface, R software has a strong practical data mining system to build a variety of conditions. In this paper, the application of R software in

[1] **Note:** Because the data used is from the Sohu web site, the rendering of the simulated word cloud is a Chinese form of atlas.

text data mining is given based on the case of data mining. Especially in the big data environment, the visual display of news interpretation provides a scientific and feasible solution for the timely analysis of the news content and tracking the hot issues of the society. When the amount of data needed to explore more than PB, and even reached ZB level when the text of the next step, how to use the R and Hadoop combination, can break the limitation of data calculated by Hadoop distributed Map Reduce, and can use the extended many excellent packages in the R, the rapid implementation of data processing and analysis required, is a to deal with the problem of big data is worth pondering.

REFERENCES

[1] Meng Xiaofeng, Ci Xiang. Big Data Management: Concepts, Techniques and Challenges [J]. Journal of Computer Research and Development, 2013, 50 (1): 146-169.

[2] Yi Xue, Liping Chen. Statistical modeling and R software [M]. Beijing: Tsinghua University press, 2007.

[3] Robert, I., Kabacoff, Gao Tao, Xiao Nan, Chen Gang. R, in, Action: Data, Analysis [M]. Beijing: Posts & Telecommunications Press, 2013.

[4] Wenwei Chen, et al. Data mining technology [M]. Beijing: Beijing University of Technology press, 2002.

[5] Paul, Teetor, Zhu, Li Hongcheng, Zhu Wenjia, Shen Yicheng. R Cookbook [M]. Beijing: China Machine Press, 2013.

[6] Clement， T，Plaisant， C，& Vuillemot， R. The story of one：Humanity scholarship with visualization and text analysis Tech Report HCIL-2008-33［Z. College Park，MD University of Maryland，Human-Computer Interaction Lab 2008.

[7] JIAWEI Han，MICHELIN Kamber．Ming Fan, Meng Xiaofeng. Datamining : concepts and techniques [M]. Beijing: China Machine Press, 2008, 6:30 - 31.

AUTHOR'S PROFILE

Shi Chen, was born in Hubei Province in 1992-03, Yunnan University of Finance and Economics graduate student 2015, main research direction, management information system, e-mail:chenshi3520@163.com;

Zheng Li， was born in Jiangsu Province in 1992-10, Yunnan University of Finance and Economics graduate student 2015, main research direction, management information system, e-mail:840884107@qq.com;

Yijing Li: female, 1963-11 was born in Anhui Province, Ph. D., Professor of Kunming University of Science and Technology, the main research direction of knowledge management and statistical data analysis, e-mail:lyjwxs@163.com;

Xiaosong Wu: male, 1962-12 was born in the city of Kunming, the Chinese Communist Party members, Professor of management, Ph.D., Vice Dean of information School of Yunnan University of Finance and Economics, teaching and research in large data and information resources in the field of science,
e-mail: wxszwd@ynufe.edu.cn