Text Visualization and LDA Model Based on R Language

Hongjie Li[1, a], Peng Cheng[2, b] and Huiyang Xie[1, c, *]

[1]College of Science, Beijing Forestry University, Beijing 100083, China

[2]Institute of Economics, Jinan University, Guangzhou 510632, China

[a]lhjlhj1991@163.com, [b]512496900@qq.com, [c]xhyang@bjfu.edu.cn

Keywords: R Language, Word cloud, community mining, LDA model.

Abstract. On the Internet, text is the mainly form of information generated by users, analyzing the text can get a lot of important information. Therefore, the text analysis has become an important means of dealing with text data. In this study, R is an open-source software, could be used to analyze users of sina Weibo and their comments. In order to find hot topics and dig the internal links in the comments, then constitute the network structure, this study used a variety of R language function package to visualize categorized word and word cloud. The LDA theme model is used to analyze the potential relationships among the entries, and provides a solution for analyzing the users' behaviors and habits in the social network and tracking hot topics.

1. Introduction

Nowadays, the rapid development of information network technology has brought a lot of text and image data. How to deal with and understand these data effectively and mining the important information under these data quickly has become an important topic. Text is the mainly form of information generated by users on the Internet, analyzing the text can get a lot of important information. For example, analyzing messages issued by users of Weibo to find out which topic that the users concerned most. This type of analysis is called text analysis. Analysis software such as SAS / EM, Insightful Miner, IBM IM and SPSS, and so on, they have completed functions and perfect performance, but expansion and high cost are their weakness. R integrates data manipulation, statistics and visualization capabilities, overcome the shortcomings of commercial data mining tools. This paper introduces the text visualization, community mining and LDA model based on R language and set an example based on a reality show named *The Amazing Race*.

2. Word segmentation and Word cloud

While using R to analyze text, the first thing is cut whole segment into terms with minimal meaning, this called word segmentation. Different word segment methods can cause different results and have different effectiveness. The "dictionary" provides the results, and the "disable dictionary" deletes invalid vocabulary. In this paper, word segment method is “jiebaR”, because it can support four types of word segmentation: the Maximum Mrobability, Hidden Markov model, Query Segment and Mix Segment.

>> cutter<- worker(bylines = T, user="user_dict.txt", top_word = "stopword.txt")

First of all, use gsub () to pre-processing the data, remove the topic, URL, symbols and other invalid content of Weibo. Then establish cutter word processor to process word segmentation.

Select the comments of one set of The Amazing Race issued on Weibo, all comments were pre-processed and word-segmented. Using unlist () function to compute the segmented word information could get the following word frequency distribution.

Table 2.1 Word frequency distribution table

Words	节目	张哲瀚	镜头	加油	极速前进	婷婷	喜欢	金星	晶刚	…
Frequency	240	215	214	178	177	169	168	149	135	…

Word cloud is a technique to analyze word frequency of large text data and generate visual images by using language analysis technology. In order to help people to understand and analyze the subject content of the micro-blogs and comments, we use visual tools to express the frequency distribution. This article provides two methods of visual word cloud, one is the word cloud package in R language. This package provides a good way to display the word frequency, including a variety of settings parameters to meet different requirements of the word cloud. As shown in Figure 2.1a.

wordcloud(words,freq,scale=c(4,5),min.freq=3,max.words=Inf,random.order=TRUE,random.color=FALSE,rot.per=.1,colors="black",ordered.colors=FALSE,use.r.layout=FALSE,...)

Figure 2.1 Word cloud diagram (a), Tagxedo diagram (b)

Another word cloud diagram method is relatively simple, and there is lots of online software or web pages also support the production of words cloud, users only need to upload terms and the corresponding frequency to make a word cloud diagram. Tagxedo is a high efficiency and exquisitely online software and provide a variety of colors and patterns. It can support Chinese word collocation, and the diagram is more accurate and beautiful than other software. As shown in Figure 2.1b.

3. The text content of the community mining

In the social network, each user corresponds to each point, the relationship between users constitute the entire network structure, in such a network, some users are connected closely, and some are long-distance. In this network, the more closely connected parts can be regarded as a community, the internal nodes have relatively close connections, and the relative connections between the two communities are relatively sparse, which is called Community structure. Many entries often appear in Text, and some entries often appear together. This means that some entries or text are related and linked, this requires people to find their connection and to identify the characteristics. When it is reflected in Weibo, some words appear together in many times, there is a link between them. If those words are names indicate that those people have some social relationship. For a program, it means the discussion on the evaluation of the program and the guests.

Modularity, also called modularity measurement, is a commonly method for measure the strength of the network community structure, it first proposed by Mark NewMan [1]. Modularity is defined as:

The size of the module value depends on the community distribution of nodes in the network, C, the community division of the network, which can be used to measure the quality of network community partitioning. The closer the value is to 1, the strength of the community structure is stronger, which means the quality of the division is better. So it is possible to obtain the optimal network community partitioning by maximizing the degree of modularity Q.

The modularity maximization problem is a classical optimization problem. Mark Newman proposed a greedy algorithm FN which maximizes the modularity based on the greedy idea. In order to reduce the time complexity of the algorithm, Vincent Blondel proposed another hierarchical greedy algorithm [2].

R integration of a variety of community clustering function [3], such as cluster_fast_greedy (test), walktrap.community (test), multilevel.community (test) and so on. In this paper, after the calculation of term-term matrix of relation, set the number of occurrences of the two entries as a weight, so that we can construct a weighted relationship diagram. Then calculate the clusters with cluster_fast_greedy (test) could get the corresponding clustering results. Modularity () can query the value of the module, it is generally considered that when the value is bigger than 0.3 we can get a good classification result. Membership () can display the classified category of each entry after classification. Finally, plot function is used to draw the final classification results. As shown in Figure 3.1.

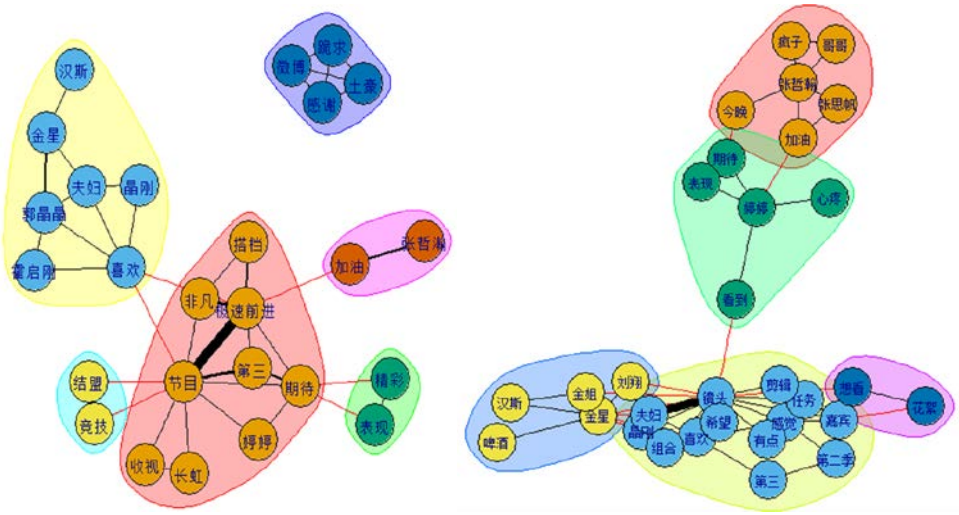

Figure 3.1 Association mining cluster results

>>plot(wc,test,layout=layout.fruchterman.reingold,edge.width=E(test)$weight/60,vertex.size=18,vertex.label.cex=0.7) 4. Conclusions

4. LDA theme model

The first time addressed the LDA (Latent Dirichlet Allocation) theme model [4] is by David M. Blei, Andrew Y. Ng and Michael I. Jordan in 2002.

In recent years, with the development of social media, text data become more and more important analytical data. Massive textual data sets new demands on the analytical capacity of social science researchers. Because the LDA theme model (Topic Model) is a probability model which could extract themes from a large number of texts, so it is increasingly applied to social science research including theme discovery and document markup [5-6].

Use LDAvis and lda packages in R to read the word and word frequency data. Then use LDA model to set its parameters, when enlarging the value of α could find that the result of each document close to a same topic, while β reflected in each topic is concentrated in a few words, or each word is transferred to a topic with the highest possible probability. Then use the following code to achieve visualization.

```
fit<- lda.collapsed.gibbs.sampler(documents = documents, K = K, vocab = vocab, num.iterations = G, alpha = alpha, eta = eta, initial = NULL, burnin = 0, compute.log.likelihood = TRUE)
theta <- t(apply(fit$document_sums + alpha, 2, function(x) x/sum(x)))  # Document - Topic Distribution Matrix
phi <- t(apply(t(fit$topics) + eta, 2, function(x) x/sum(x)))  # Theme - word distribution matrix
term.frequency <- as.integer(term.table) #Word frequency
json <- createJSON(phi = phi, theta = theta, doc.length = doc.length, vocab = vocab, term.frequency = term.frequency)
serVis(json,out.dir ="d:/LDAvis",open.browser= FALSE)
```

The results of web-based visualization analysis are generated in the corresponding folder. This paper opens the web page with firefox, you can get the interactive pages of LDA model. As shown in Figure 4.1.

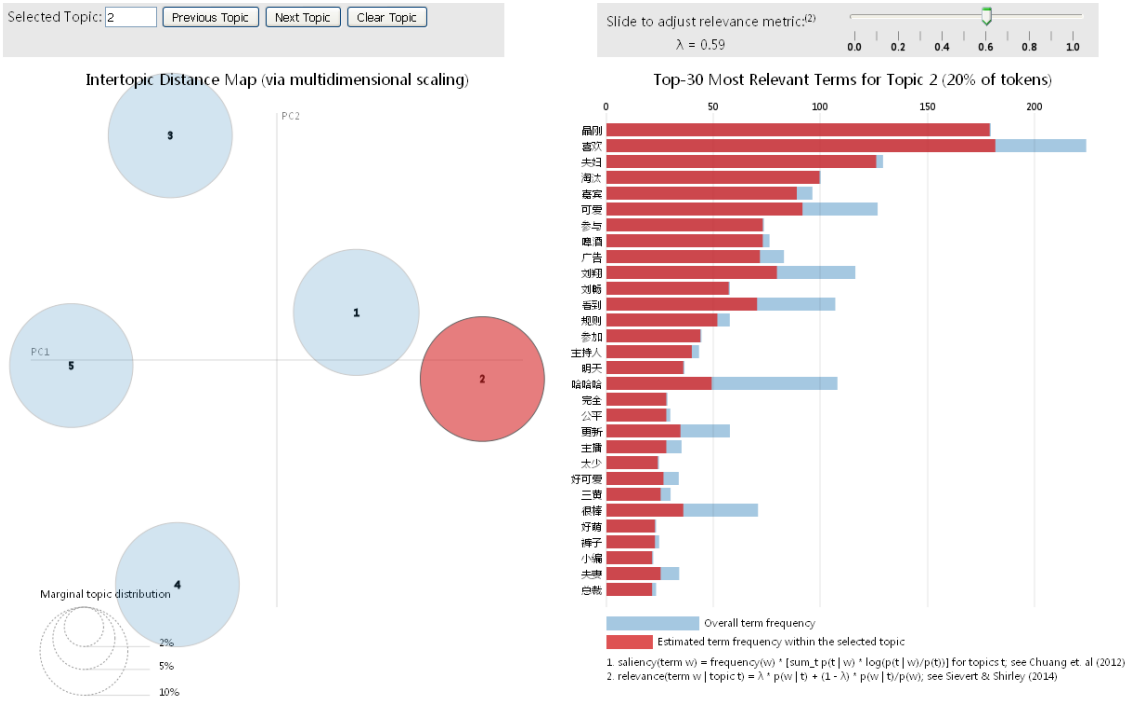

Figure 4.1 LDA theme model

5. Conclusion

By analyzing the comments data based on Weibo, this paper provides some methods of text data mining with R. The word segmentation system can support the Chinese text of the discriminant analysis and word cloud display. Using community mining to find the relationship the information, especially the visualization of information dissemination platform and the important functions of LDA analysis model in the environment of Big Data, provides scientific and operational solutions to analyze data and tracking social hot issues.

Acknowledgments

This paper supported by the National Natural Science Foundation of China under Grant No.61370193.

References

[1] Newman MEJ. Fast algorithm for detecting community structure in networks. *Physical review E Statistical Nonlinear & Soft Matter Physics*. 69 (6): 066133, 2004

[2] Blondel V. D, Guillaume J. L, Lambiotte R, and Lefebvre E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics Theory & Experiment.* 10: 155-168,2008.

[3] Clauset, A. Finding local community structure in networks.*Physical Review E Statistical Nonlinear & Soft Matter Physics.* 72(2), 254-271,2005.

[4] David M. Blei, AndrewY. Ng, and Michael I. Jordan, Latent Dirichlet Allocation, *Journal of Machine Learning Research*. 3: 993-1022, 2003.

[5] Airoldi E. M, David M. Blei, Fienberg Stephen E, And Xing Eric P. Mixed Mem-bership Stochastic Blockmodels. *Journal of Machine Learning Research*. 9(5): 1981-2014, 2008.

[6] David M. Blei And McAuliffe J. D. Supervised topic models. *Advances in Neural Information Processing Systems*. 3:327-332, 2008.