Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 15(3): 170-174, Sep. 2017 **Regular paper**

Text Mining and Visualization of Papers Reviews Using R Language

Jiapei Li[1], Seong Yoon Shin[2], and Hyun Chang Lee[3*], *Member, KIICE*

[1]Department of Library Information Consulting, Hebei Geology University, Shijiazhuang 050031, China
[2]School of Computer Information & Communication Engineering, Kunsan National University, Gunsan 54150, Korea
[3]Department of Digital Contents Engineering, Wonkwang University, Iksan 54538, Korea

Abstract

Nowadays, people share and discuss scientific papers on social media such as the Web 2.0, big data, online forums, blogs, Twitter, Facebook and scholar community, etc. In addition to a variety of metrics such as numbers of citation, download, recommendation, etc., paper review text is also one of the effective resources for the study of scientific impact. The social media tools improve the research process: recording a series online scholarly behaviors. This paper aims to research the huge amount of paper reviews which have generated in the social media platforms to explore the implicit information about research papers. We implemented and shown the result of text mining on review texts using R language. And we found that Zika virus was the research hotspot and association research methods were widely used in 2016. We also mined the news review about one paper and derived the public opinion.

Index Terms: R language, Text mining, Visualization, Word cloud

I. INTRODUCTION

With the advent of the Web 2.0 and the big data, online forums, blogs, Twitter, Facebook and other social media services have developed rapidly. Researchers begin to conduct their work flow on social media tools. Scholarly literature is shared and discussed on Twitter and Facebook, organized in social reference managers like Mendeley and ReadCube, commented in blogs and micro blogs, reported in news, peer-reviewed after publication in Faculty of 1000. While the social media tools improve the research process and scholar communication efficiently, they have another powerful advantage: recording a series of online scholarly behaviors. The series of online scholarly behaviors are kinds of digital traces [1]. In “altmetrics: a manifesto”, Priem et al. [2] define altmetrics as follows: This diverse group of activities (that reflect and transmit scholarly impact on social media) forms a composite trace of impact far richer than any available before. We call the elements of this trace altmetrics (http://altmetrics.org/manifesto/). According to altmetric.com, altmetrics are metrics and qualitative data that are complementary to traditional, citation-based metrics. They can include (but are not limited to) peer reviews on Faculty of 1,000, citations on Wikipedia and in public policy documents, discussions on research blogs, mainstream media coverage, bookmarks on reference managers like Mendeley, and mentions on social networks such as Twitter. Compared with traditional bibliometrics and webmetrics, altmetrics are superior in that they provide rapid, real-time, public and transparent reports on scientific impact, and

Received 07 August 2017, **Revised** 14 August 2017, **Accepted** 20 September 2017
***Corresponding Author** Hyun Chang Lee (E-mail: hclglory@wku.ac.kr, Tel: +82-63-850-6260)
Department of Digital Contents Engineering, Wonkwang University, 460, Iksan-daero, Iksan 54538, Korea.

Open Access **https://doi.org/10.6109/jicce.2017.15.3.170** print ISSN: 2234-8255 online ISSN: 2234-8883

cover an extensive non-academic audience and diversified research findings and sources [3].

Social media platforms contain a lot of comment texts about scientific articles. We should better analyze them through statistical analysis, sentiment analysis, text classification and clustering, and machine learning to obtain implicit, unknown useful information from them, and thus better support scientific research and discovery. In this paper, we conducted text mining on the reviews of articles on social media, in an attempt to trace the focus of review and the direction of public opinion reflected in news reports.

II. RELATIVE WORKS AND DATASETS

Text mining encompasses a vast field of theoretical approaches and methods with one thing in common: text as input information. This allows various definitions, ranging from an extension of classical data mining to texts to more sophisticated formulations like “the use of large online text collections to discover new facts and trends about the world itself” [4]. In general, text mining is an interdisciplinary field of activity amongst data mining, linguistics, computational statistics, and computer science. Standard techniques are text classification, text clustering, ontology and taxonomy creation, document summarization and latent corpus analysis. In addition a lot of techniques from related fields like information retrieval are commonly used.

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: text has always been the default way of storing information for hundreds of years, and mainly time, personal and cost constraint prohibit us from bringing texts into well-structured formats (like data frames or tables).

The issue of text mining is of importance to publishers who hold large databases of information needing indexing for retrieval. This is especially true in scientific disciplines, in which highly specific information is often contained within written text. Therefore, initiatives have been taken such as Nature's proposal for an Open Text Mining Interface (OTMI) and the National Institutes of Health's common Journal Publishing Document Type Definition (DTD) that would provide semantic cues to machines to answer specific queries contained within text without removing publisher barriers to public access.

The automatic analysis of vast textual corpora has created the possibility for scholars to analysis millions of documents in multiple languages with very limited manual intervention. Key enabling technologies have been parsing, machine translation, topic categorization, and machine learning.

The automatic parsing of textual corpora has enabled the extraction of actors and their relational networks on a vast scale, turning textual data into network data. The resulting networks, which can contain thousands of nodes, are then analyzed by using tools from network theory to identify the key actors, the key communities or parties, and general properties such as robustness or structural stability of the overall network, or centrality of certain nodes [5]. This automates the approach introduced by quantitative narrative analysis [6], whereby subject-verb-object triplets are identified with pairs of actors linked by an action, or pairs formed by actor-object [7].

Content analysis has been a traditional part of social sciences and media studies for a long time. The automation of content analysis has allowed a “big data” revolution to take place in that field, with studies in social media and newspaper content that include millions of news items. Gender bias, readability, content similarity, reader preferences, and even mood have been analyzed based on text mining methods over millions of documents [8-11]. The analysis of readability, gender bias and topic bias was demonstrated in Flaounas et al. [12] showing how different topics have different gender biases and levels of readability; the possibility to detect mood shifts in a vast population by analyzing Twitter content was demonstrated as well [13].

In this paper, we chose the 100 highest-score articles in 2016 on Altmetrics.com, downloaded the datasets (December 7, 2016) via the link (https://figshare.com/collections/Altmetric_Top_100_2016/3590951).

III. METHODS

First we produced a plain text file “Top100.txt” which includes the summaries of all the 100 articles. Then we selected the highest-score article “United States Health Care Reform: Progress to Date and Next Steps” in 2016 and produced a text file based on mainstream media comments on it provided by Altmertics.com. Accordingly, we prepared two plain text files (one for the whole, and one for parts) for later text mining.

We used the RStudio version 3.3.3, including its statistical environment and the following packages: tm, dplyr, wordcloud2, etc. we implemented textual analysis of comment texts by studying the whole first and then narrowing the analysis scope to focus on some of them to obtain visualized word clouds and derived the idea of comments.

IV. RESULTS AND ANALYSIS

In continuous dissemination on social media, scientific articles not only leave digital records but also attract a host of comment texts on news outlets, blog and Twitter, etc.

These texts are important, rare source of strong support for evaluating the impact of scientific articles. We conducted a textual analysis based on the summary file of the 100 articles contained in the datasets and the news report file of one particular article among them. First, we entered the texts and the summary file of the 100 articles into the system. Second, we pre-processed the texts, such as deleting spaces, converting them into lowercase, deleting punctuation marks and words that are no longer in use. Third, we calculated the word frequency. Finally, we exported the visualized word clouds according to the word frequency. We used R language to program and the R script as follows:

```
1 library(wordcloud2)
2 library(dplyr)#data getting and cleaning
3 library(tm)
4 ##data cleaning, delete the blanks and punctuations
5 filePath<- "D:/R/top100wordcloud.txt"
6 text = readLines(filePath)
7 txt = text[text!=""]
8 txt = tolower(txt)
9 txt <- removeWords(txt,stopwords('english'))
10 txtList = lapply(txt, strsplit," ")
11 txtChar = unlist(txtList)
12 txtChar = gsub("\\.|,|\\!|:|;|\\?","",txtChar)
13 txtChar = txtChar[txtChar!=""]
14 data = as.data.frame(table(txtChar))
15 colnames(data) = c("Word","freq")
16 ordFreq = data[order(data$freq,decreasing=T),]
17 wordcloud2(ordFreq, size = 0.5,shape = 'star')
```

Thus, from the datasets we extracted 1,447 words and the seven most frequently used words are listed in Table. 1.

The words in the data set were displayed as word cloud according to word frequency. From Fig. 1 we can see that in 2016, people were more interested in the studies of human beings, in particular in the studies of cancers and the Zika virus that swept across Africa. From the frequently used word “association”, we discovered that most of the research was interdisciplinary, indicating the overlapping and fusion of scientific research. Besides, the research is “New”, meaning

Fig. 1. Visualized word cloud of comments on Top 100 articles.

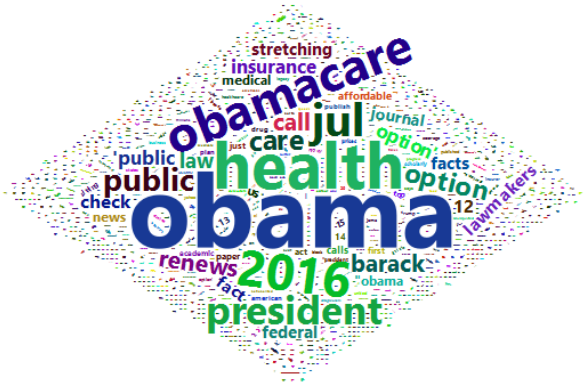

Fig. 2. Visualized word cloud of news review bout one paper.

that researchers adopt new methods, new perspectives and new approaches for pioneering research.

In addition, one paper in the datasets “United States Health Care Reform: Progress to Date and Next Steps” has received continuous media attention since its publication. We crawled a total of 31 titles of news reports on it and developed the visualized word cloud by using the same method. Fig. 2 gives that the common theme of these news reports shows that “former US president Obama rolled out Obama care in July 2016”.

Table 1. High frequency words

Words	Frequency (%)
Human	17
Cancer	13
Virus	12
Zika	12
Association	10
New	10
Life	9

V. CONCLUSIONS AND OUTLOOKS

Bormmann [14] considered that future research should focus more on the measurement of the extensive impact of the research, not on the comparison of altmetrics and traditional metrics. According to Davis et al. [15], text mining technology should be applied to track indirect citations of textual contents of research findings, particularly in blogs, news reports and government documents. We conducted text mining on the article summary file of the

datasets and found the focus of attention in scientific research from the public perspective and a new approach to the universal cooperation in scientific research in 2016. Text mining was also performed on titles of news reports on one particular article. Media comments about the article were visualized by word cloud. Deceptively simple, text mining tells us what the numbers recorded by altmetrics cannot tell. The visualized word cloud also makes the result more straightforward and easy to understand.

Altmetrics give us a unique social perspective to analyze the impact of academic research findings and trace academic communication among readers. There is a host of datasets to support the studies in academic social networking behaviors and even in the interaction between different metrics [16]. On top of that, visualization of academic exchange and community found at the social media level is another major research subject [17].

Social media platforms contain a lot of comment texts about scientific articles. We should better analyze them through statistical analysis, sentiment analysis, text classification and clustering, and machine learning to obtain implicit, unknown useful information from them, and thus better support scientific research and discovery.

ACKNOWLEDGMENTS

This paper was supported by Wonkwang University in 2017.

REFERENCES

[1] K. Weller, "Social media and altmetrics: an overview of current alternative approaches to measuring scholarly impact," in *Incentives and Performance*. Cham: Springer International Publishing, 2015.

[2] J. Priem, T. Taraaborelli, P. Groth, and Neylon, "Altmetrics: a manifesto," 2010 [Internet], Available: http://altmetrics.org/manifesto/.

[3] P. Wouters and R. Costas, "Users, narcissism and control: tracking the impact of scholarly publications in the 21st century," 2012 [Internet], Available: http://apo.org.au/node/28603.

[4] M. A. Hearst, "Untangling text data mining," in *Proceeding of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL)*, College Park, MD, pp. 3–10, 1999.

[5] S. Sudhahar, G. De Fazio, R. Franzosi, N. Cristianini, "Network analysis of narrative content in large corpora," *Natural Language Engineering*, vol. 21, no. 1, pp. 81-112, 2015.

[6] R. Franzosi, "Quantitative narrative analysis," *Journal of Bacteriology*, vol. 191, no. 7, pp. 2388-2391, 2016.

[7] S. Sudhahar, GA. Veltri, and N. Cristianini, "Automated analysis of the US presidential elections using big data and network analysis," *Big Data & Society*, vol. 2, no. 1, pp. 1-28, 2015.

[8] I. Flaounas, M. Turchi, O. Ali, N. Fyson, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini, "The structure of EU Mediasphere," *PLoS ONE*, vol. 5, no. 12, pp. e14243, 2010.

[9] V. Lampos and N. Cristianini, "Nowcasting events from the social web with statistical learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 4, pp. 1-22, 2012.

[10] I. Flaounas, O. Ali, M. Turchi, T. Snowsill, F. Nicart, and T. De Bie, "NOAM: news outlets analysis and monitoring system," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, Athens, Greece, pp. 1275-1277, 2011.

[11] N. Cristianini, "Automatic discovery of patterns in media content," in *Combinatorial Pattern Matching*. Cham: Springer International Publishing, pp. 2-13, 2011.

[12] I. Flaounas, O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini, "Research methods in the age of digital journalism," *Digital Journalism*, vol. 1, no. 1, pp. 102-116, 2013.

[13] T. Lansdall-Welfare, V. Lampos, and N. Cristianini, "Effects of the recession on public mood in the UK," in *Proceedings of International Conference on World Wide Web*, Lyon, France, pp. 1221-1226, 2012.

[14] L. Bornmann, "Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics," *Journal of Informetrics*, vol. 8, no. 4, pp. 895-903, 2014.

[15] B. Davis, I. Hulpuş, M. Taylor, and C. Hayes, "Challenges and opportunities for detecting and measuring diffusion of scientific impact across heterogeneous altmetric sources," 2015 [Internet], Available: http://altmetrics.org/wp-content/uploads/2015/09/altmetrics15_ paper_21.pdf.

[16] M. Taylor, "Exploring the boundaries: how altmetrics can expand our vision of scholarly communication and social impact," *Information Standards Quarterly*, vol. 25, no. 2, pp. 27-32, 2013.

[17] C. P. Hoffmann, C. Lutz, and M. Meckel, "A relational altmetric? Network centrality on ResearchGate as an indicator of scientific impact," *Journal of the Association for Information Science and Technology*, vol. 67, no. 4, pp. 765-775, 2015.

Jiapei Li
received her M.S. degree from information department in Tianjin normal university in China. From 2008 to the present, she has been an assistant professor in the Library of Hebei geology university in China. Her research interests include data science and text mining.

Seong Yoon Shin
received his M.S. and Ph.D. degrees from the Dept. of Computer Information Engineering of Kunsan National University, Gunsan, Korea, in 1997 and 2003, respectively. From 2006 to the present, he has been a professor in the same department. His research interests include image processing, computer vision, and virtual reality.

Hyun Chang Lee
Dean of Digital Content Engineering Department at Wonkwang University in Korea. He obtained his Ph.D. and M.S. from the University of Hongik, Seoul in Korea, in 2001 and 1996, respectively. Since 2008, he has been a professor in Wonkwang University. His major interests include business intelligence, IoT, AR&VR and data science fields.