Research on Clustering and Classification for the Green Edition of Southern Weekend

Sheqi Zhang*, Xiangmei Dou
Beijing National Day School, Beijing, China.

* Corresponding author. Tel.: 18310770963; email: sheqi0620@icloud.com
Manuscript submitted August 13, 2017; accepted November 21, 2017.
doi: 10.17706/jcp.13.9.1027-1036

Abstract: The corpus consists of 2099 reports of the Green Edition of Southern Weekend from October 8, 2009 to September 30, 2016. Word frequency and term frequency–inverse document frequency (TFIDF) are counted and sorted. Word clouds of each year are given according to words of high TFIDF. Key words and word scopes of each year are different from the word clouds. The feature words are generated scientifically and automatically. All reports are clustered into one category automatically using hierarchical clustering algorithm. We can find the clustering results intuitively and visually from the hierarchical clustering. The clustering results provide scientific and effective bases for automatic classifications of a large scale of short reports. 2099 reports are classified into 20 categories automatically and the classification is very reasonable.

Key words: Green edition of southern weekend, word cloud, hierarchical clustering, scree plot, automatic classification.

1. Introduction

In the recent ten years, people tend to pay attention to environmental protection instead of pursuing business interests blindly. Chinese environmental issues are becoming better and better with the help of the government and the society. The media also concerns about people's attitude to environment and plays a positive role in recording and supervising environmental protection. Southern Weekend, one of the most influential media in China, hopes to record and promote the green process of the country. The Green Edition of Southern Weekend (Short for the Green edition) was set up on October 8, 2009, which broke the traditional news classification model such as politics, society, economy and culture. The Green Edition covers all aspects of environmental protection including environmental protection, energy, low carbon economy, sustainable development and public safety. [1]

The Green Edition of Southern Weekend was selected as the corpus. Mathematical model and R were used, and the data were analyzed in a visual and intuitive way. Quantitative and qualitative method are combined to analyzing the Green Edition, which can describe the change of the Green Edition's focus and give readers a social and scientific depict for Chinese environmental protection.

Wang Min, Zhang Yan, Deng XiaoXuan, Wang Guanhui, Yuan Duanduan, Zhang Haifeng, Wang Jilong, Chen qingmei combine content analysis method and the statistical method of simple frequency and frequency distribution to analyze a small part of the Green Edition. They analyze one or more aspects of the theme,

[1] Born for Green, The Green Edition of Southern Weekend, October 8, 2009.

the ways, the source of news, the content, the type, the category of reports and the shortcomings of the Green Edition [1]-[8].

Compared with the previous researches on the Green Edition, the research focus on the following:

(1) A larger scale of corpus is used.

2099 reports of the Green Edition of Southern Weekend from October 8, 2009 to September 30, 2016 are used for the corpus of the paper, while fewer reporters are selected in the previous researches on the Green Edition.

(2) TFIDF and hierarchical clustering model are used.

The changing trends of theme words over the years can be determined by using complex mathematical model and algorithm. Mathematical model, computer science and the method of humanities are combined to analyze the Green Edition of Southern Weekend, which has not been found in the previous research on the Green Edition.

(3) Quantitative method and qualitative method are combined.

Qualitative method is mainly used for analyzing the Green Edition in the previous research. Quantitative method and qualitative method are combined to analyze the Green Edition in the paper.

(4) The statistical results are shown by using lots of figures and tables.

Tables, word clouds, hierarchical clustering tree, scree plot and automatic classification are applied for a large scale of corpus, and meaningful contents are mined from the corpus. The tables and the figures show the scale, theme words, scope, focuses and the classification result of the corpus in a clear, vivid and direct way.

It is difficult to cluster automatically for similar and a large scale of shorter texts. Feature words are selected by scientifically weighting method and 2099 similar reports are clustered into a tree.

The method is also used to identify plagiarism papers, identify the authorship of the literature works, filter rubbish emails, help to solve criminal cases, evaluate online products and predict emotion of microblogs, speeches and literary works, which provide valuable and scientific bases for the tasks.

2. Mathematical Model

2.1. TF-IDF

Term frequency–inverse document frequency (TF-IDF) is a numerical statistic that is intended to reflect how important a word is to a document in a corpus.

It is often used as a weighting factor in information retrieval, text mining, and user modeling.

TF-IDF is made up of Term Frequency (TF) and Inverse Document Frequency (IDF).

$$TF_{ik} = \frac{tf_{ik}}{\sum_{j=1}^{t} tf_{jk}} \tag{1}$$

Tf_{ik} is a term frequency of term i in document k. tf_{ik} is the number of times that term w_i occurs in $\sum_{j=1}^{t} tf_{jk}$ document k is the total number of term s in the document k.

$$IDF_{ik} = \log\left(\frac{N}{n_k + L}\right) \tag{2}$$

IDF_{ik} is inverse document frequency (a global parameter), N is the total number of documents in the corpus and N=2099 in the paper. n_k is the number of documents containing term w_i. L=1.

$$TFIDF_{ik} = TF_{ik} \times IDF_{ik} \tag{3}$$

The *tf-idf* value increases proportionally to the number of times a term appears in the document, but is offset by the frequency of the term in the corpus, which helps to adjust for the fact that some terms appear more frequently in general [9].

2.2. Euclidean Distance

Euclidean distance is to compute the distance between texts [9].

$$d_{ij}(2) = \| X_i - X_j \| = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{4}$$

(x_{i1}, x_{i2}, ……x_{ip}) is the vector representing text *i*, and (x_{j1}, x_{j2}, ……x_{jp}) is the vector representing text j. $d_{ij}(2)$ is Euclidean distance between text *i* and text *j*.

2.3. Standard Score

A standard score is also called Z-score, which indicates how many standard deviations an element is from the mean. A z-score can be calculated from the formula (5).

Suppose the matrix $[x_{ij}]$ $(i = 1,\ 2,\ 3,\ \ldots,\ n;\ j = 1,\ 2,\ 3,\ \ldots\ p)$

$$X_{xj}^{*} = \frac{X_{xj} - \overline{X}_j}{S_j} (i = 1,\ 2,\ 3,\ \ldots n;\ j = 1,\ 2,\ 3,\ \ldots,\ p) \tag{5}$$

$$\overline{X}_j = \frac{\sum_{i=1}^{n} X_{ij}}{n} \tag{6}$$

$$S_j = \sqrt{\frac{1}{n-1} \sum_{n-1}^{n} (X_{ij} - \overline{X}_j)^2} \tag{7}$$

$\overline{X}_j$ is the mean of x_{ij} (i=1, 2, …, n), and S_j is the standard deviation of x_{ij} (i=1, 2, …, n). X_{ij}^{*} is the standard score of X_{ij}.

2.4. Sum of Squares of Deviations (Ward's Method)

Sum of Squares of Deviations is put forward by Ward [9], which is called Ward's method.

The diameter D_k of category G_k is defined by the formula (8).

$$D_K = \sum_{i=1}^{m} (X_i - \overline{X}_K)^T (X_i - \overline{X}_K) \tag{8}$$

The category G_k and the category G_L are clustered into a new category G_{K+L}. The formula (9) defines the square of the distance between the category G_k and the category G_L.

$$D_{KL}^2 = D_{K+L} - D_K - D_L \tag{9}$$

D_L is the diameter of the category G_L, and D_{K+L} is the diameter of the category G_{K+L}. They are defined by the formula (8).

3. Corpus

All the reports of the Green Edition of Southern Weekend[2] from Oct. 8, 2009 to Sep. 30, 2016 are downloaded and comprised the corpus. The corpus is made up of 214 pages, 2099 reports or 2,573,846 words. The Table 1 gives the numbers of reports and words in each year for the Green Edition of Southern

[2] http://www.infzm.com/green.shtml

Weekend.

Table 1. The Numbers of Reports and Words in Each Year

year	No. of reports	No. of words	year	No. of reports	No. of words
2009	58	56155	2010	332	365245
2011	235	283445	2012	357	413097
2013	378	460062	2014	300	388574
2015	233	326844	2016	206	280424

From Table 1, the number of words and reports in 2009 is the least, because the Green Edition of Southern Weekend began publication in October of 2009. The number of words and reports in 2013 is the most.

4. The Processing Procedure

The process is as Fig. 1.

① Download the reports from the web page of the Green Edition of Southern Weekend from October 8, 2009 to September 30, 2016 using Octopus V6.2.1[3], and create the corpus containing 2099 reports.

② Segment the corpus using the package JiebaR of R[4]. Organization name, person name and some phrases(For example, "垃圾焚烧", "环境保护", "发展中国家") are segmented into a segmentation unit.

③ Count frequency and TFIDF of each word according to the formula (3).

④ Create the word lists of high frequency and high TFIDF respectively. We use stop words from Information Retrieval Laboratory of Harbin Institute of Technology and some added stop words(for example, "南方周末"). The stop words are filtered out before generating high frequency words and high TFIDF words. There are all together 1228 words, which are punctuations, conjunctions, auxiliary words and adverbs and so on.

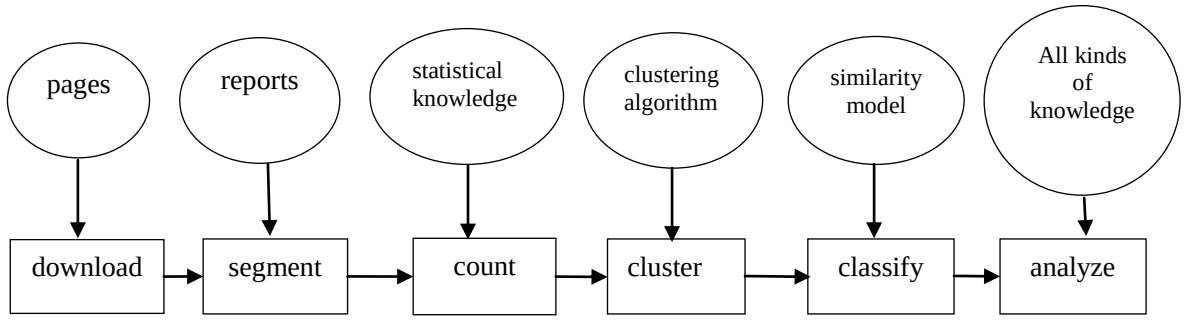

Fig. 1. The process of corpora processed, counted and analyzed.

⑤ Generate word clouds automatically according to higher TFIDF words.

⑥ Each report is represented as a feature vector. Cluster all the reports into a category based on hierarchical clustering principle. The hierarchical clustering result is a tree.

⑦ Classify all the reports based on the clustering tree.

⑧ Analyze the results based on statistical knowledge, linguistic knowledge, cultural background and commonsense knowledge.

R[5] is used for segmenting the corpus, counting word frequency and TFIDF, generating word clouds,

[3] http://www.bazhuyu.com
[4] http://www.r-project.org/
[5] http://www.r-project.org/

clustering the reports and classifying the reports.

5. Word Clouds

Transgenosis, food safety, air quality, new energy, climate change, bird flu, refuse burning, solar energy, developing country, developed country, air pollution, environmental protection, the World Expo, additive, renewable sources of energy, nuclear power station, and drinking water are the highest TFIDF words for the Green Edition of Southern Weekend. The words are indeed the theme words of the Green Edition of Southern Weekend. There are many high frequency numerals and ordinal numbers, which aren't the theme words of the Green Edition of Southern Weekend. Obviously, higher TFIDF words are more suitable for acting as the theme words than higher frequency words.

Word cloud of each year is automatically generated according to words with higher TFIDF. First, the threshold of each year is set to make the number of words of each year is about 100. The Table 2 gives the threshold of each year. Secondly. We select the words which TFIDF is greater than the threshold of each year. Thirdly, the word cloud is depicted using the function wordcloud () according to the word's TFIDF. The greater the word's TFIDF is, the greater the word is depicted.

Table 2. TFIDF Threshold of Each Year from 2009 to 2016

year	threshold	year	threshold	year	threshold	year	threshold
2009	0.5	2010	1.7	2011	1.3	2012	1.7
2013	1.8	2014	1.2	2015	1.2	2016	1.2

Fig. 2. Word cloud of 2009.

Fig. 3. Word cloud of 2010.

Fig. 4. Word cloud of 2011.

Fig. 5. Word cloud of 2012.

Fig. 6. Word cloud of 2013.

Fig. 7. Word cloud of 2014.

Fig. 8. Word cloud of 2015.

Fig. 9. Word cloud of 2016.

The greater the word is depicted, the greater important the word is. Therefore, we can find out the hot themes and scopes of each year visually from the word clouds. The hot themes vary from year to year. Copenhagen and climate change were most concerned in 2009, the World Expo, Copenhagen and climate change were most concerned in 2010, fifteen of the first month of lunar year, National Bureau of Oceanography and senior high school students were most concerned in 2011, transgenosis and new energy were most concerned in 2012, h7n9 and bird flu were most concerned in 2013, transgenosis and most h7n9 were most concerned in 2014, food safety and waste incineration were most concerned in 2015, and Aids and clinical test were most concerned in 2016.

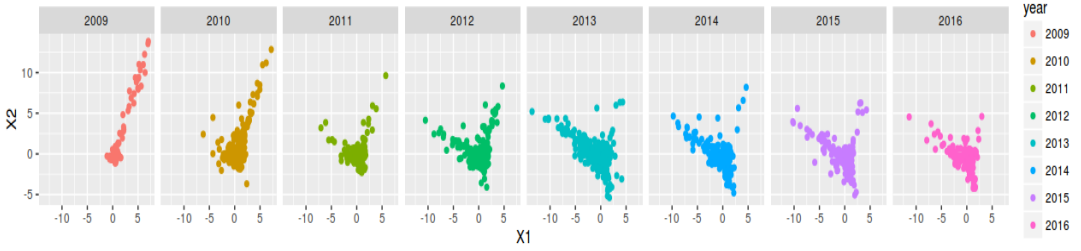

Fig. 10. The scatter plot of reports after reduction from multi-dimensional vector to two dimensions by year.

Words with higher TFIDF vary from 2009 to 2016, which explains the scope of themes vary from year to year for the Green Edition of Southern Weekend. In order to show the different themes of different year visually, we reduce multi- dimensional scale of each report to two- dimensional scale and keep the distance of any two reports unchanged. The scatter plot of reports is depicted in Fig. 10.

The scope of themes varies from year to year from Fig. 10.

6. Text Clustering

6.1. Feature Words

The words to categorize the reports based on content are feature words. We can find the themes of the report according to the feature words. Reports with similar themes tend to use semantically similar words, and semantically similar words tend to co-occur in reports with similar themes. The themes vary from year to year according to word clouds of each year. Words with higher TFIDF grouped by year are selected and considered as the feature words. The first ten words of each year according to TFIDF are selected, then 80 words are selected. Repeated words and number words are deleted from 80 words, and the remaining words are the feature words.

6.2. Hierarchical Clustering

Each report is represented by a feature vector. The feature vector is an 48 dimensional vector of numerical features. The features are words from Table 3, and the weight of each feature is the word's TFIDF in each report. So 2099 feature vectors consist of a matrix.

Hierarchical clustering is a process of grouping a set of reports into some groups so that the similarity of the reports in the same category is higher, while the similarity of the reports in different category is lower. Hierarchical clustering tries to build a hierarchy of clusters.

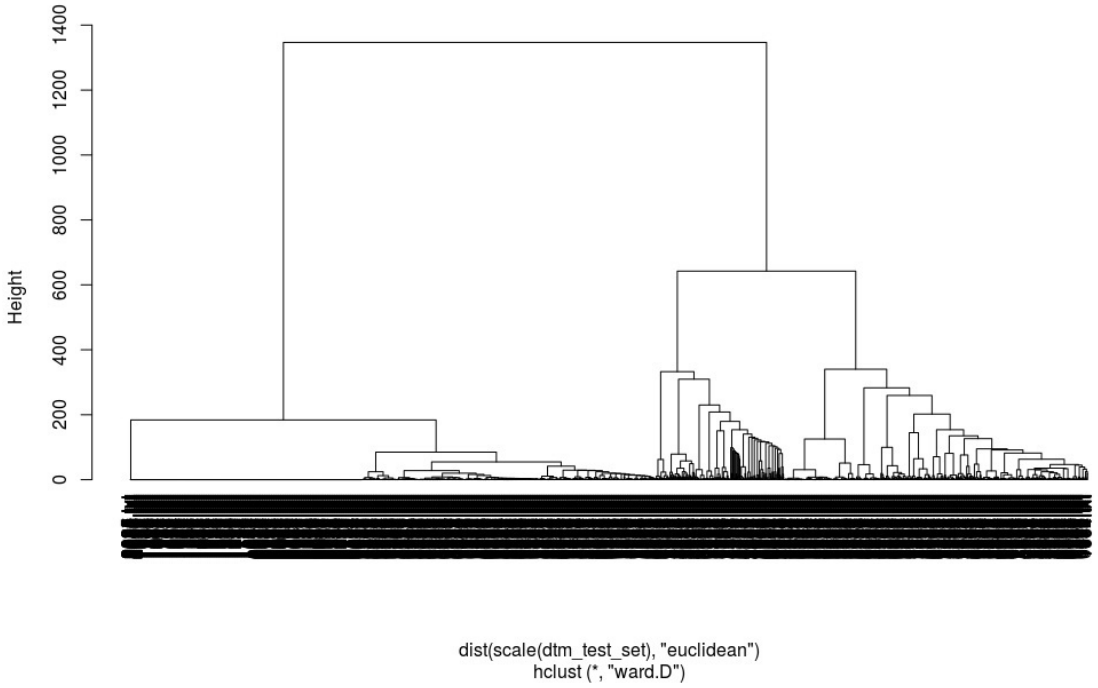

Fig. 11. The hierarchical clustering tree of 2099 reports.

The process of hierarchical clustering is as follows:

① Take every report as a category.

② Calculate the similarities between any two categories.

③ Find the greatest similarities between two categories, which are clustered as a new category.

④ Repeat ② and ③ until all the reports are clustered into a category.

Hierarchical clustering can not only measure the similarities between two reports, but also measure the similarities between two categories.

The TFIDF of each feature is calculated according to the formula (3) and standardized according to the formula (5). The similarity between two reports is measured according to Euclidean distance of the formula (4). The similarity between two categories is measured according to Ward's method of the formula (9). The hierarchical clustering result of 2099 reports is as the Fig. 11.

2099 reports are clustered into a tree. The bigger the distance between two categories is, the smaller the similarity between two categories is. The smaller the distance between two categories is, the bigger the similarity between two categories is.

7. Automatic Classification

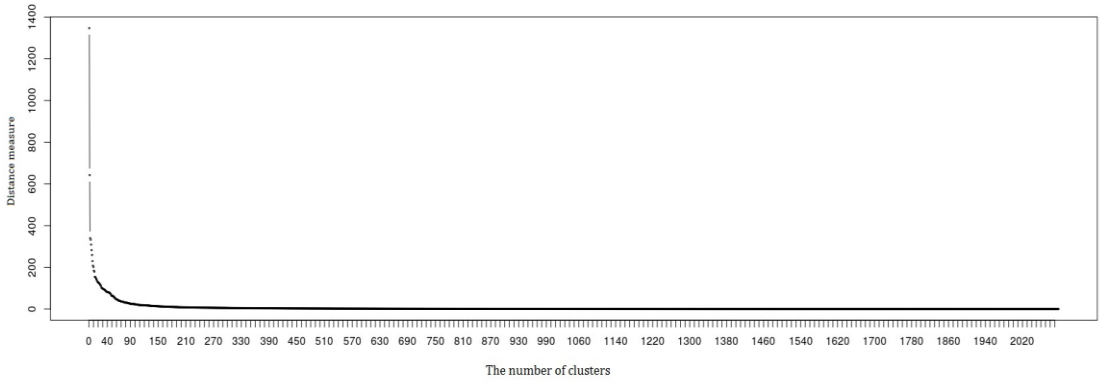

Fig. 12. The scree plot of hierarchical clustering.

A scree plot displays the distance associated with categories in descending order versus the number of categories. This scree plot shows that 20 of 2009 categories explain most of the variability because the line starts to straighten after 20 categories. The remaining categories explain a very small proportion of the variability and are unlikely unimportant.

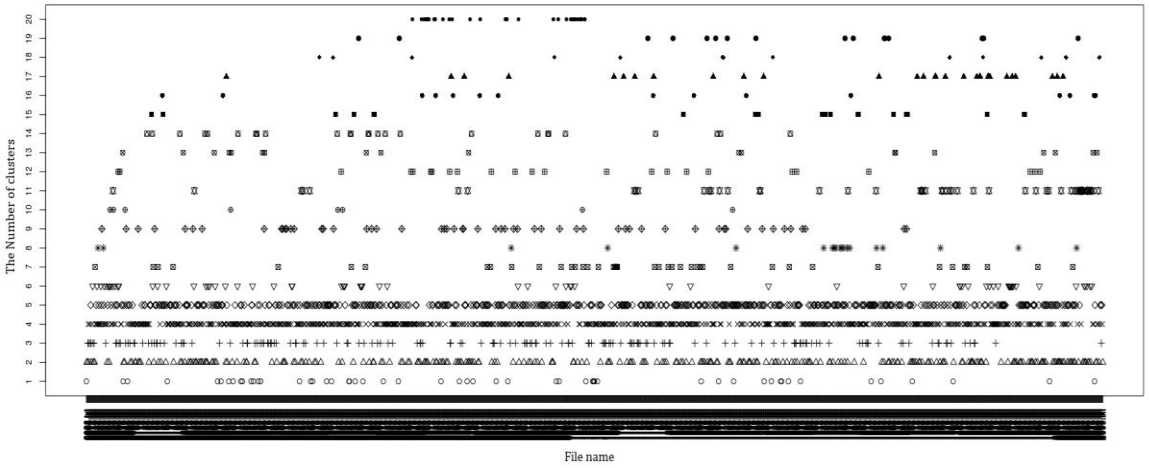

Fig. 13. 20 categories of hierarchical clustering.

Categories are automatically classified into 20 categories according to the clustering results. Fig. 13 shows 20 categories. Each row shows a category. For example, Fig. 14 shows the 20th category.

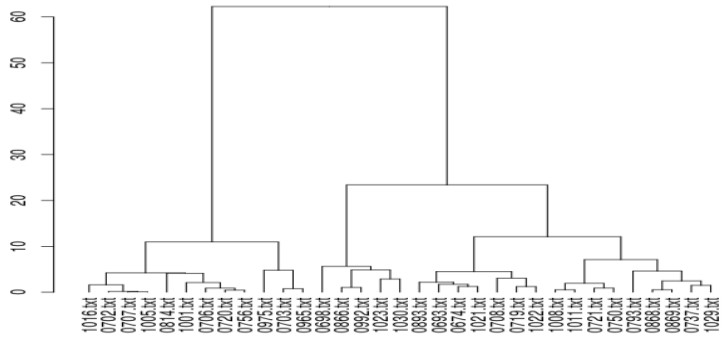

Fig. 14. Reports of the 20th category.

In Fig. 14, the serial number of reports are represented in horizontal axis, while the distance between two categories is represented in vertical axis. Reports of Fig. 14 are concerned with new bird flu. Fig. 14 are a subtree of Fig. 11. Semantically similar reports are indeed clustered into one small category using the selected feature words.

8. Conclusion

The themes and scopes of the Green Edition of Southern Weekend are given in a scientific and visual manner. 2099 reports are clustered into one category automatically. The key words and word scopes vary from year to year from the word clouds and the scatter plot. More similar the reports are, more earlier the reports are clustered. This proves the features selected by using TFIDF are very effective. The hierarchical clustering provides scientific, reasonable and effective bases for automatic classifications of a large scale of short reports. 2099 reports are classified into 20 categories according to the hierarchical clustering. The future work is to automatically predict emotions of every report.

Acknowledgement

I really appreciate the guide and help of Bin Wang and Chao Li.

References

[1] Wang, M. (2013). *Research on Environmental News of the Green Edition of Southern Weekend*. Unpublished master thesis, Nanjing Normal University, Nanjing.

[2] Zhang, Y. (2013). *Study of Reporting Features of Environmental News in the Green Edition of Southern Weekend*. Unpublished master thesis, Shanghai Jiao Tong University, Shanghai.

[3] Deng, X. X. (2013). *Study of Text Frames of News in the Green Edition of Southern Weekend*. Unpublished master thesis, Northwest University, Xi'an.

[4] Wang, G. H. (2012). *Study of Environmental News in the Green Edition of Southern Weekend*. Unpublished master thesis, Henan University, Kaifeng.

[5] Yuan, D. D. (2012). *The Innovation of the Concept and Methods of Green News in the Low Carbon Economy Background-Based on the Green Edition of Southern Weekend*. Unpublished master thesis, Jinan University, Jinan.

[6] Zhang, H. F. (2011). The content analysis of the green edition of southern weekend. *Journalism Lover, 6*, 93-95.

[7] Wang, J. L. (2013). Development, problems and causes of green journalistic publishing in China — A case study of southern weekend green. *Journal of Shanghai Jiao Tong University (Philosophy and Social Science), 21(5)*, 62-69.

[8] Chen, Q. M. (2011). New forms of environmental news reports in China — The green edition of

southern weekend. *Academic Research*, 349-350.
[9] Liu, Y. (2014). *Statistical Linguistics.* Beijing: Tsinghua University Press.

Sheqi Zhang is with Beijing National Day School, Beijing, China. Her research interest is computer application.

Xiangmei Dou is with Beijing National Day School, Beijing, China. She got the Ph.D from Chinese Academy of Sciences, Beijing, China. Her research interests are bioscience and computer application.