International Journal of u- and e- Service, Science and Technology
Vol.9, No. 11 (2016), pp.71-80
http://dx.doi.org/10.14257/ijunesst.2016.9.11.07

The Text Mining Analysis of Macau's Participation in International Organizations

Wu Shianghau[1], Liu Chengkun[2], Shi Haoyu[3] and Pang Chuan[4]

[1,2,4]School of Business, Macau University of Science and Technology,
[3]Institute for Social and Cultural Research, Macau University of Science and Technology
[1]shwu@must.edu.mo, [2]ckliu@must.edu.mo;[3]about.s@hotmail.com,
[4]cpang@must.edu.mo

Abstract

In recent years, Macau has actively participated in international organizations according to the Macau Basic Law. After fifteen years of regaining sovereignty by China, Macau has faced some internal and external challenges. In order to comprehend the new policy directions regarding Macau's participation in international organizations, this paper made the depth interviews and applied the text mining analysis to the interviews responses. Through applying the random forests model and fruit fly optimization algorithm to keywords classification, the study concluded several points related to Macau's participation in international organizations.

Keywords*: text mining; wavelet transformation; random forests; fruit fly optimization algorithm*

1. Introduction

Macau is the special administrative region of China. Some scholars referred to Macau's international participation as the example of "paradiplomacy". Paradiplomacy is the emerging policy capacity of sub-state entities which can be enjoyed by both states and sub-state entities [1]. Some scholars think Macau's paradiplomacy as the "outreach of non-sovereign jurisdictions to actors beyond their borders and frontiers of their metropolitan relationships or claimant states. The power of the sub-states in international relations was decided by the sovereign state [2]. In recent years, Macau's international participation has become the research focus of international relations researchers. In order to comprehend the new policy directions regarding Macau's participation in international organizations, this paper made the depth interviews on two renowned specialists in politics and economy fields and applied the text mining analysis to the interviewees' responses.

2. Methodology

2.1. Text Mining

Text mining is one of the data mining methods, which learn from samples of past experience. In the text mining method, the text will be processes and transformed into a numerical representation.

According to the recent literature of text mining, the research focuses of text mining can be divided as the application of text mining to different areas and the new design of text mining algorithm and software.

As for the text mining application research, it is widely applied to information management on websites, biological data and customer relationship management. Other

ISSN: 2005-4246 IJUNESST

applications mainly relate to library management and literature critique. For example, Clement (2008) used the text mining method to analyze Gertrude Stein's "The Making of Americans" and has different conclusions from that of other literature critics [3]. And Trappey (2007) applied the text mining method to analyze patent brochures [4].

As for the new design of text mining algorithm and software research, Schmidt (2010) used the text mining method to analyze the marketing-oriented textual information[5]. Esuli and Sebastiani (2010) described an industrial-strength software system for automatically coding open-ended survey responses and compared the accuracy of the software against the accuracy of human coders [6]. Lee and Bradlow (2011) described a new text mining algorithm for analyzing online customer reviews to facilitate the analysis of market structure [7].

2.2. Depth Interviews

In order to explore more thorough results, the study made depth interviews on two of Macau's Legislative Assembly Members. The aim was to comprehend the Macau's problems and challenges on the international participation and conclude Macau's new policy directions. The interviews were taken place respectively on the first week of this January. Each interview took about forty-five minutes. The depth interview questions were made according to the news reports and research papers related to Macau's participation in international organizations.

The depth interview questions were listed as follows:

(1) What is the possible development for Macau to participate in international organizations?

(2) According to your professional background, what kinds of international organizations are suitable for Macau's participation?

(3) Macau joined the Asian Pacific anti- money laundry organization. Why does Macau still link to money laundry according to some international press reports?

(4) What are your viewpoints about the cooperation between Macau's civil organizations and its international counterparts?

(5) What are the related measures for Macau Special Administrative government to handle the international affairs?

(6) What are your views regarding Macau's roles in international organizations?

(7) Can you give some examples about Macau's participation in international organizations?

(8) Can you make the comparison with Macau's participation in international organizations before and after 1999?

(9) In your opinion, what kinds of international organizations or conventions are important for Macau's participation?

(10) What kinds of efforts should Macau make for more roles in international organizations?

(11) Do you think it is important to encourage Macau citizens to participate in international affairs?

(12) Do you think the central government's "regional cooperation" concept contradict to the Macau's internationalization?

The study assorted the depth interview response data in order to facilitate the application to the text mining method.

2.3. Research Design

The text mining method of the study is implemented by means of the text processing function of the R software and its application packages . The study uses the text mining application ("tm") package of R software to analyze the depth interview text data.

The text mining method of the study is implemented by means of the text mining package (tm) and "Rwordseg" package of the R software [8] The text mining package follows the procedures as follows[9] :

- Parsing: The text and the structure were extracted and represented in a data structure.
- N Chars Filter: The node filters terms consisting of less than a specified number. In the study, the specified number is set as four in order to filter out definite and indefinite articles (e.g., a, an, the).
- Number Filter: The node filters all terms consisting of numbers only.
- The Punctuation Erasure: The node removes all punctuation marks.
- Stop Words Filter: The node filters all stop words.

The study also calculated the term frequency (tf), and the inverse document frequency (*idf*) to choose the major keywords. The study got major keywords from the contents of two interview responses including: "Basic Law", "central government", "APEC", "service industry", "regularize", "financial industry", "productivity", and "diversification". The study got fifteen keywords frequency data and categorized the keywords data from the first interviewee as "Type 0" and keywords data from the second interviewee as "Type 1".

2.4. Wavelet Transformation

The wavelet transformation is a tool that separates data, functions, or operators into different frequency components and then explores each component with a resolution matched to its scale. Wavelet transformations were developed to express the frequency domain and the time locality of an input function. The fact that wavelets capture the temporal nature of the data is quite essential [10].

Many researches were devoted to the application of wavelet transformation to data mining. Hussain et. al. (2008) used the data mining method to extract the informative proteins in the discrete stationary wavelet transform domain[11]. Bassani and Nievola (2008) used an explanatory approach for data mining of EEG signal based on continuous wavelet transformation (CWT) and wavelet coherence (WC) statistical analysis [12].

The study used the "WaveThresh" package of R language [13] to make the wavelet transformation. The wavelet transformation enables us to find the independent coefficients using either Normalized Elliptic Fourier or Discrete Wavelet, which can facilitate further data handling. The study used the original keywords idf data to make the wavelet transformation. The R commands were as follows.

```
program WaveThresh
   {Let (y1,…yn) as the keywords idf data to become input.}
Begin:
   ywdS <- wd(y, filter.number=1, family="DaubExPhase",
   +type="station")
   accessD(ywd, level=2)
end
```

According to the R commands, the study got finest-scale coefficients. And the wavelet was the Daubechies wavelets as defaults [14].

2.5. Random Forests

After making the wavelet transformation, the study also applied the random forests classification analysis for the transformed data to explore the relationship of Type 0 and Type 1 data and the importance of keywords. The random forests classification included the following steps [15-16].

Step (1): Draw the ntree bootstrap samples from the original data.

Step (2): For each of the bootstrap samples grow an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample mtry of the predictors and choose the best split from among those variables.

Step (3): Predict new data by aggregating the predictions of the ntree trees (i.e., majority votes for classification, average for regression).

The study categorized the keywords idf data and made the keywords data from the response context of the first interviewee as "Type 0", the keywords data from the response context of the second interviewee as "Type 1". The number of trees was set as 500, and the number of variables tried at each split was set as 2. The "rattle" package of the R software used 8 keywords data as the test data (6 Type 0 data, 2 Type 1 data) to build the model. The overall error of the random forests model for the data was 0. The study also used the ROC (Receiver Operating Characteristic) curve to determine whether the model is the suitable model. The ROC curve plots the true positive rate against the false positive rate. The method is to consider the square measures of areas under the ROC curves. If the square measure approaches to 0.5, it would be the less corresponding model. If the square measure equals to 1, it would bet the model with perfect accuracy. According to the calculation, the square measure of the area under the ROC curve was 1. The ROC curve of the random forests model in the study is shown on Fig.1.

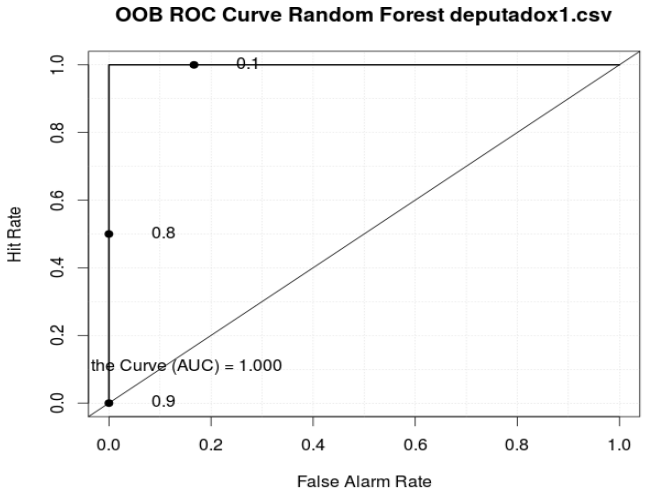

Figure 1. ROC Curve of the Random Forests Model

The random forests model calculated the variable importance, mean decrease accuracy and mean decrease gini of keywords which were listed as Figure 2, and Table 1. The error matrix of the random forests model was shown on Table 2.

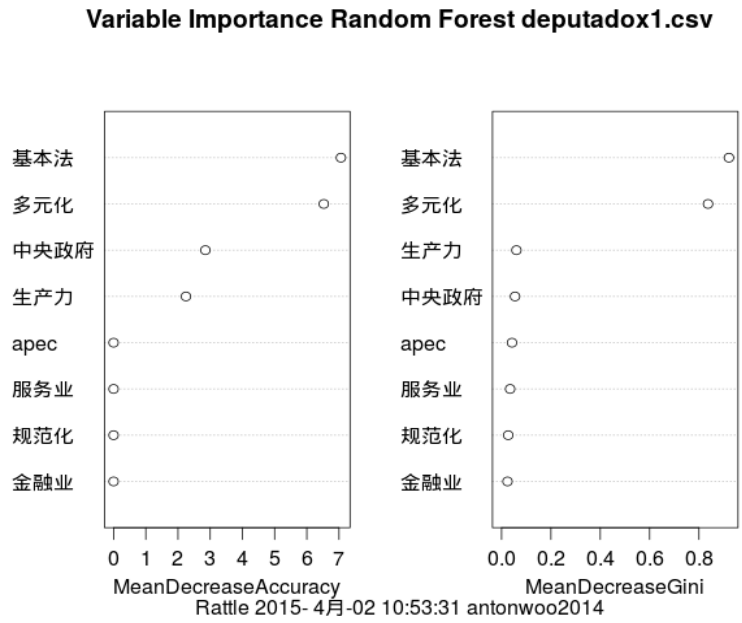

Figure 2. The Variable Importance, Mean Decrease Accuracy and Mean Decrease Gini of Keywords from Random Forests Model

Table 1. Valuable Importance, Mean Decrease Accuracy and Mean Decrease Gini in Random Forests Model

keywords	Valuable Importance (Type 0 Data)	Valuable Importance (Type 1 Data)	Mean Decrease Accuracy	Mean Decrease Gini
Basic Law	6.59	7.20	7.06	0.92
diversificati on	6.32	6.59	6.53	0.84
Central government	0.00	2.85	2.85	0.05
productivity	0.00	2.25	2.25	0.06
APEC	0.00	0.00	0.00	0.04
Service industry	0.00	0.00	0.00	0.03
regularizati on	0.00	0.00	0.00	0.03
Financial industry	0.00	0.00	0.00	0.02

Table 2. Error Matrix for the Random Forests Model

Observed	**Predicted**		
	Type No.0	***Type No.1***	***Percentage Correct***
Type No. 0	10	0	100.00
Type No. 1	1	3	75.00
Overall Error Percentage	7.14%		

2.6. Fruit Fly Optimization Algorithm

The Fruit Fly Optimization Algorithm (FOA) is a new optimization method developed by Pan (2011) [17]. The FOA was based on the sniffing behavior of fruit flies. Pan (2011) stipulated the FOA can be applied to gray system, data mining and neural networks. The FOA program assumed the fruit fires group located in a random space and found the food's location by sniffing [18]. When the fruit fries reached the food's location, the swarm of fruit flies will fly to the location directly by sense vision (sniffing) and the location would be confirmed. The basic step for FOA calculation was as follows [19],

Step1. The method set the fruit fly group's location by random (X-axis, Y-axis).

Step2. The method endowed individual fruit fly wih random value for locating food location(X,Y).

X= X-axis+ Random Value;

Y= Y-axis+ Random Value.

Step3. The method calculated the personal fruit fly's distance from the original point (0,0), and found the S value of density, this value equals to the inverse value of distance.

$$Dist = \sqrt{x^2 + y^2}; S = 1/Dist \quad (1)$$

Step 4. The method induced the smell density Value S into the density judge function, and find out the smell density value of personal fruit fly at confirmed location. Smell=Function(S).

Step 5. The method repeated the step two to step four, calculating the smell density of all the fruit flies in the swarm, and found out the fruit fly with largest and lowest smell density value.

Step 6. The method recorded the smell density value and location (X,Y) of the best fruit fly, and the flying routine of the group fly to that location.

Step 7. The method made the iterative optimization, and repeated step two to step five, and checked whether the smell density value was better than the previous one, if yes, carried on the step six, or carry on the step seven, until match the max iterations, then the calculating finished.

The study attempted to comprehend the relationship of each keyword in each analyzed document. In order to achieve the goal, the study stipulated the Keywords Frequency Composite Function (KFCF), which intended to describe the relationship of each keyword. The study took the Composite Keywords Frequency Index as Z, which means the category of each keyword data. The inverse document frequency of each keyword in each document was set as Xi, the coefficient of each keyword as a_i and the number of keywords as n. The KFCF was set as follows,

$$Z = \sum_{i=1}^{n} a_i Xi$$

The study used the Fruit Fly Optimization Algorithm to get the coefficients of each keyword by the optimization process.

The study attempted to fit the Keywords Frequency Composite Function by FOA optimization. The iterations of FOA were set as 100, and the size of fruit flies was set as 8, which meant each fly was designated for each keyword. The iterative optimization process was depicted as Figure 3.

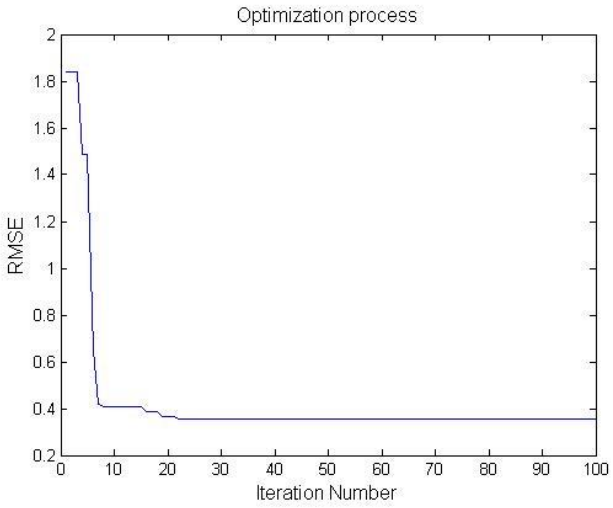

Figure 3. Iterative Optimization Process

According to Figure 2, the forecasting error between the real and forecasting coefficients of each keyword in the KFCF was kept on 0.3528 after 27 iterations.

The searching route of fruit flies was depicted as Figure 4.

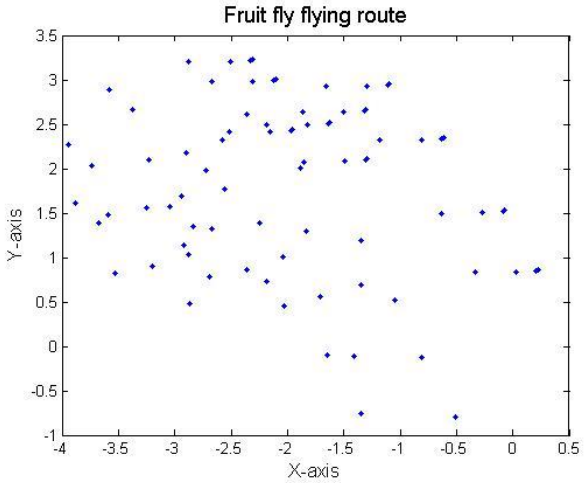

Figure 4. The Searching Route for Fruit Flies of KFCF

As the optimization process was converged after 27 iterations, the coefficient of each keyword in the KFCF can be ascertained. The KFCF was as follows.

Z = 0.2787*APEC + 0.2880*service industry + 0.2475*regularization + 0.2882*financial industry + 0.2593 *productivity + 0.2330 *central government + 0.2354*diversification + 0.3106* Basic Law

The ROC curve of the Fruit Fly Optimization Algorithm was on Fig.5.

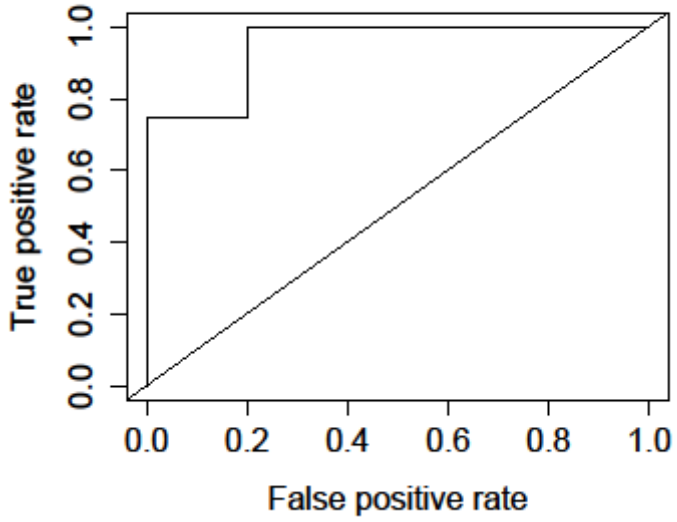

Figure 5. ROC Curve of FOA Model

According to the calculation, the square measure under the ROC curve was 0.95. The error matrix of the FOA model was shown on Table 3.

Table 3. Error Matrix for the FOA Model

Observed	Predicted		
	Type No.0	***Type No.1***	***Percentage Correct***
Type No. 0	10	0	100.00
Type No. 1	1	3	75.00
Overall Error Percentage	7.14%		

3. Discussion

The study applied the text mining to analyze the depth interview responses to get the inverse document frequency (*idf*). After making wavelet transform, the study used the random forests algorithm to make the classification and got top four major keywords as "Basic Law", "diversification", "central government" and "productivity".

In order to explore the relationship among the eight major keywords, the study used the fruit fly optimization algorithm to fit the keywords composite function and found the keyword "Basic Law" ,"financial industry" and "service industry" and "APEC" were major keywords. As for the comparison between the two models, the study found the random forests model had the larger square measure of the area under the ROC curve, while it had the same classification error percentage with the FOA model. We can conclude that the random forests model is the suitable model for the keywords classification.

The study concluded the interview responses as the following points according to the above-mentioned results:

(1)Macau's participation in international organizations relied on the regulations of Basic Law and central government.

(2)The related concepts of Macau's future participation included "industrial diversification" and "financial industry".

(3) Macau's service industry had played the important role in the future participation in international organizations.

4. Conclusions

The study attempted to explore the important points of Macau's future road map of international participation by means of depth interview and text mining analysis. The study found the key points of two interviewees related to the Basic Law, central government, industrial diversification and financial industry according to the random forests model and FOA model fitting results. It can be inferred that Macau's international participation should be under the framework of Basic Law and add on the concepts of industrial diversification and the participation of financial industry according to the text mining and statistical analysis results.

Acknowledgments

The authors would like to show their gratitude for assistance of the Institute for Social and Cultural Research at Macau University of Science and Technology and the Macau Foundation Research Fund.

References

[1] S. Wolff, "Paradiplomacy: Scope, Opportunities and Challenges, The Bologna Center Journal of International Affairs, vol. 10, no. 1, **(2007)**, pp. 141–150.
[2] J. C. Matias, "Macao, China and Portuguese Speaking Countries China's Macao Transformed: Challenge and Development in the 21st Century", China's Macao Transformed: Challenge and Development in the 21st Century, **(2014)**.
[3] T. E. Clement, "A Thing not Beginning and not Ending: Using Digital Tools to Distant-Read Gertrude Stein's The Making of Americans", Literary and Linguistic Computing, vol. 23, no. 3, **(2008)**, pp. 361–380.
[4] A. J. C. Trappey and C. V. Trappey, "An R & D knowledge Management Method for Patent Document Summarization", Industrial Management Data Systems, vol. 108, no. 2, **(2007)**, pp. 245–257.
[5] M. Schmidt, "Quantification of Transcripts from Depth Interviews, Open-Ended Responses and Focus Groups", International Journal of Market Research, vol. 52, no. 4, **(2010)**, pp. 483–509.
[6] A. Esuli and F. Sebastiani, "Machines that Learn How to Code Open-ended Survey Data", International Journal of Market Research, vol. 52, no. 6, **(2010)**, pp. 775–800.
[7] T. Lee and E. BradLow, "Automated Marketing Research Using Online Customer Reviews", Journal of Marketing Research, vol. 48, **(2011)**, pp.881–894.
[8] J. Li, "Rwordseg Usage", **(2013)**, https://r-forge.r-project.org/R/?groupid = 1054
[9] I. Feinerer, K.Hornik and D.Meyer, "Text Mining Infrastructure in R", Journal of Statistical Software, vol. 25, no. 5, **(2008)**.
[10] K. Swati and S. Kumar, "A Comparative Study of Various Data Transformation Techniques in Data Mining", International Journal of Scientific Engineering and Technology, vol. 4, no. 3, **(2015)**, pp.146–148.
[11] M. K. Hussain, M. H. Miran-Baygi and M. H. Moradi, "A data-mining approach to biomarker identification from protein profiles using discrete stationary wavelet transform", Journal of Zhejiang University Science B, vol. 9, no.11, **(2008)**, pp. 863-870.
[12] T. Bassani and J. C. Nievola, "Brain-computer interface using wavelet transformation and naive bayes classifier", Brain Inspired Cognitive Systems 2008, Springer New York, **(2010)**, pp. 147-165.
[13] L. A. Libungan and S. Pálsson, "ShapeR: An R Package to Study Otolith Shape Variation among Fish Populations", PloS one, vol. 10, **(2015)**, e0121102.
[14] G. Nason, "Wavelet methods in statistics with R", Springer Science & Business Media, New York, **(2010)**.
[15] A. Liaw and M. Wiener, "Classification and Regression by Random Forest", The R Journal, vol. 2, no. 3, **(2002)**, pp. 18-22.
[16] L. Breiman, "Random Forests", Machine Learning, vol. 45, **(2001)**, pp. 5-32.
[17] W. Pan, "Fruit Fly Optimization Algorithm", Taipei: The Sea Press, **(2011)**, pp. 10-12.

[18] W. Pan, "Using Fruit Fly Optimization Algorithm Optimized General Regression Neural Network to Construct the Operating Performance of Enterprises Model", Journal of Taiyuan University of Technology (Social Sciences Edition), vol. 39, no. 4, **(2011)**, pp. 1-4.
[19] F. Xu, and Y. Tao, "The Improvement of Fruit Fly Optimization Algorithm: Using Bivariable Function as Example", Proceedings of the 2012 2nd International Conference on Computer and Information Application (ICCIA 2012), Atlantis Press, Paris, France, **(2012)**, pp. 1516-1520.

Authors

Wu Shianghau, Associate Professor, School of Business, Macau University of Science and Technology.

Liu Chengkun, Associate Professor, School of Business, Macau University of Science and Technology.

Shi Haoyu, Doctoral Candidate, Institute for Social and Cultural Research, Macau University of Science and Technology

Pang Chuan, Professor, School of Business, Macau University of Science and Technology

