The Text Mining Application of "Intelligent Government Affairs"

Huang Runchen, Wang Benhui, Yang Wenjue

College of Applied Mathematics, Beijing Normal University Zhuhai, Zhuhai, Guangdong, 519087

ABSTRACT
In recent years, with WeChat, microblog, mayor's mailbox and other network political platforms continue to become an important channel for the government to understand public opinion, gather people's wisdom, and condense people's morale. The number of texts of various data has increased sharply. The traditional manual data processing methods have appeared problems such as low efficiency and tedious work. Therefore, it is of great significance to use text analysis and data mining methods to deal with mass messages and improve the efficiency of government. In this paper, based on the records of people's political questions, through some algorithms and some mathematical tools, this paper analyzes the hot areas and hot issues, and gives a set of evaluation scheme for the quality of the reply from the perspective of relevance, integrity and interpretability. Finally, the reliability of the score is verified by various formulas, which shows that the model has high accuracy.
Keywords: *TF-IDF, F-score, correlation coefficient, significance level*

1. INTRODUCTION

In recent years, with WeChat, microblog, mayor's mailbox and other network political platforms continue to become an important channel for the government to understand public opinion, gather people's wisdom, and condense people's morale. The number of texts of various data has increased sharply. The traditional manual data processing methods have appeared problems such as low efficiency and tedious work. Therefore, it is of great significance to use text analysis and data mining methods to deal with mass messages and improve the efficiency of government.

In this paper, based on the records of people's political questions, through some algorithms and some mathematical tools, this paper analyzes the hot areas and hot issues, and gives a set of evaluation scheme for the quality of the reply from the perspective of relevance, integrity and interpretability.

2. RELATED CONCEPTS

2.1. weight

It refers to the importance of a factor or indicator relative to a certain thing, which is different from the general proportion. It not only reflects the percentage of a certain factor or indicator, but also emphasizes the relative importance of the factor or indicator, which tends to contribute or importance.

2.2. tf-idf algorithm

TF-IDF is a common weighting technology for information retrieval and data mining. TF is word frequency, IDF is inverse text frequency index.

2.3. accuracy and recall

The accuracy rate is based on our prediction results, which indicates how many samples with positive prediction are real positive samples. Recall rate is for our original sample, it shows how many positive examples in the sample are predicted correctly.

2.4. F-score

F-score is a kind of statistics, F-score is also called F-measure, F-measure is precision and recall weighted harmonic average, is a common evaluation standard in IR (information retrieval) field, which is often used to evaluate the quality of classification model.

3. WORD FREQUENCY STATISTICS FREQUENCY STATISTICS BASED ON TF-IDF

3.1. classification process

Now there are a number of messages, the operator can put the message one by one corresponding to its due first level label; or you can first group the same type of messages

together, and finally return them to their proper tags. Obviously, the latter is more efficient. If we want to put the same types of words together, they must have the same characteristics. Here we choose word frequency as their characteristics. The same type of messages will generally have the same words with higher frequency. For example, according to experience, it can be inferred that the words with the most frequent occurrence are "fight", "robbery" and other words related to the meaning of crime Vocabulary.

If you want to filter the keywords in the message accurately from the words that appear more frequently in the message, the operator needs to use TF-IDF algorithm. The specific principle of TF-IDF algorithm is as follows:

The first step is to calculate the word frequency

Word frequency (Ni) = the number of times a word appears in the text, which can be expressed as:

$$\sum_{i=1}^{n} N_i$$

Considering that the length of each message is inconsistent, the operator can standardize the word frequency, and set the total number of words in the P text as NQ:

Word frequency (TF) = (the number of times a word appears in the text) / (the total number of words in the text), which can be expressed as:

$$TF = \frac{\sum_{i=1}^{n} N_i}{nq}$$

3.2. result analysis

After calculating the word frequency of each keyword in each message, the operator can select the top 15 words with the largest frequency from each message as a word set to replace the message. There are 14 first-class tags in total, and there are a number of words which are replaced by word frequency. In this case, we can first collect the word set with the highest coincidence degree to get the set with higher repetition rate in the fourteen sets. Then, TF-IDF method is used to calculate the top ten words with the highest coincidence rate in each set. Finally, by artificial judgment, all the message sets replaced by the word set can be collected to their corresponding first level tags.

In order to know whether the above method is appropriate, we should use the F-score formula as follows:

$$F - score = \frac{1}{n}\sum_{i=1}^{n} * \quad (1+\beta)^2 * \frac{P_i * R_i}{R_i + P_i * \beta^2}$$

Where p is Precision and R is Recall. Although from the calculation formula, the two are not related, but in large-scale data sets, the two are often not mutually compatible but restrictive. In general, the higher the P, the lower the R, and the higher the R, the lower the P. β represents the weight in this formula. When β>1, the operator thinks that the recall rate is more important, when β<1, the operator thinks that the accuracy rate is higher, and when the operator thinks that the accuracy rate and the recall rate are equally important, β=1, So sometimes it is necessary to make trade-offs. In this question, the operator believes that both P and R are very important, so the operator assigns β to 1 so the above formula becomes:

$$F - score = \frac{1}{n} * \sum_{i=1}^{n} \frac{2 * Pi * Ri}{Pi + Ri}$$

Therefore, the better the above method, the higher the F-score value, and the worse the method, the lower the F-score value.

4. SCREEING OF HOT MESSAGES

The message in the source data can be divided into several parts by the standard of a certain group of people or a certain region in a certain period of time, so as to obtain several sets of data.

4.1. Data sorting

After obtaining several sets of data, suppose there are k sets of data, each set of data can be expressed as θj (0≤j≤k). Each set of data contains a number of original messages in Annex III, and each message in Annex III has a number of objections and likes. Suppose the data in the θj group contains a message; suppose the i (0≤i≤a) message has bi antilogs and ci likes; let di be the weighted sum of likes and antis, Let the weight of the likes be r, the weight of the antilog be s, and the weighted score is expressed as:

$$di = rc_i + sb_i$$

The weight of the number of messages in each group of data is u, so the weighted total score in the θ_j group of data is W_j;

$$w_j = ua + \sum_{i=0}^{n} d_i$$

The heat evaluation index can also be represented by W. The formula can modify the weight according to the needs. If the operator feels that the number of comments, likes or oppositions in a group of data groups is better than the other two In response to the heat of the problem, the weight can be set higher than the other two weights; if the operator feels that the number of comments or likes or objections in a group of data sets has a side effect on the hotspot of the problem, it can also be The weight is set to a negative value.

4.2. Data group arrangement

After k are obtained, the top five data sets with W values can be obtained by comparing with each other, and the problem IDs are ranked as 1, 2, 3, 4, and 5 from high to low.

When specifically studying the problem, the operator sets s to 0, that is, does not consider the influence of the antilog on the problem, and regards the number of messages in each data group as five points of praise: s=0, u=5, r=1; through the screening of Annex III, operators can get the top five hot issues are "car loan fraud", "noise problem caused by high-speed rail", "the road has public transportation, infrastructure, campus and other problems", "High-voltage

pole line problem" and "Community night supper smoke noise disturbs the people, there are also problems such as housing quality, housing purchase policy, and difficulty in enrolling in the school." By calculating their heat index is 1587, 726, 225, 208 and 165. Arrange them in turn to get the results as shown in the table.

Table 1 Problem heat results

Heat Ranking	Ques-tion ID	Heat Index	Time Range	Location/Crowd
1	1	1587	2019/1/11-2019/7/8	58 car loan cases in city A
2	2	726	2019/1/30-2019/9/6	A City Greenland Bund Community
3	3	225	2019/1/14-2019/11/21	Songya Lake in A7 County
4	4	208	2019/3/26-2019/4/12	Moon Island Road, Area A6
5	5	165	2019/1/8-2019/12/4	A City Charming City Community

5. SCORING MODEL FOR REPLY OPINIONS

5.1. Definition of scoring model

For the evaluation of relevance, the TF-IDF algorithm described in the text is still used to select keywords for reply comments at the same

Table 2 Comment keywords heat analysis

Question ID	Question description	Number of comments	Number of likes
1	Car loan fraud	6	1557
2	Noise problems caused by high-speed rail	7	691
3	This section has public transportation, infrastructure, campus and other issues	22	115
4	High voltage rod line problem	7	173
5	The nighttime night fume noise in the community disturbs the people, and there are still problems such as the quality of the house, the policy of purchasing the house, and the difficulty of entering school	29	20

time, and the scores of the keywords of the message and the reply comments are discriminated and scored. The relevance should be positive with the number of keyword coincidences Related.

The interpretability of the reply should actually be influenced by factors such as the logic of the reply itself and whether there are corresponding laws and regulations. It is difficult to judge the logic of language by machine scoring, so the relevant words such as ‘according to law’, ‘according to regulations','regulations','legal provisions"government issued' and so on in the reply are extracted, which has the credibility 'The frequency of occurrence of characteristic words is scored. Interpretability is positively related to the frequency of occurrence of such words.

The score of the completeness of the reply can integrate the evaluation of both relevance and interpretability. It is easy to understand that the more relevant the answer is to the question, the higher the interpretability, and the more complete the message comments. Therefore, completeness should be positively correlated with relevance and interpretability.

Due to the lag and untimeliness of the reply to the message,

the actual situation of the problem described in the message may gradually increase entropy with time; that is, the later the reply, the specific situation of the problem described in the message may have changed, and even relevant The impact of force majeure, such as changes and revisions of the law Therefore, it is reasonable to think that the time interval between the time of leaving a message and the time of answering is an important indicator of scoring. To this end, a new evaluation indicator 'lag' is introduced for this situation, which should be negatively correlated with the scoring.

5.2. Calculation of weighted average points

Suppose the operator scored from four perspectives (correlation, completeness, interpretability, lag). The correlation score is A, the integrity score is B, the interpretability score is C, and the hysteresis score is J. Next, the operator needs to separately weight the completeness and interpretability of the correlation, and set their weights to e, o, g, and v, respectively. From this, the operator can get the weighted average of the reply of the i-th division to Di, the formula can be expressed as

$$Di = \frac{e * A_i + o * B_i + g * C_i + v * J_i}{100 * e + 100 * o + 100 * g + 100 * J}$$

5.3. Feasibility test

The operator scored each answer by scoring with a machine, but in order to check whether the machine is properly tested, we need to review it.

The operator first selects the score of the F reply comments as the sample X from the reply opinions that have been fully scored by the machine, and the score of the i reply opinion is xi.

The calculation gives the mathematical expectation of machine scoring for sample X:：

$$E(X) = \frac{\sum_1^F x_i}{F}$$

Through calculation, the variance of sample X is:

$$D(X) = \frac{1}{F-1} * \sum_1^F [x_i - E(x)] * [x_i - E(X)]$$

After finding the expected variance, the operator will manually score these F answers, using the obtained score as the sample Y, and the answer opinion score of the i-th answer is yi.

The mathematical expectation of sample Y can be obtained by calculation:

$$E(Y) = \frac{\sum_1^F y_i}{F}$$

Through calculation, the variance of sample Y is:

$$D(Y) = \frac{1}{F-1} * \sum_1^F [y_i - E(Y)] * [y_i - E(Y)]$$

To determine whether the scores of the two are more consistent, the operator can start with their expected variance correlation coefficient.

Assuming that the degree of expected error that the operator can accept is h, when

$$h \leq |E(X) - E(Y)|$$

The operator considers that it does not meet the demand, when

$$h > |E(X) - E(Y)|$$

The operator is deemed to meet the conditions;

Similarly, assuming that the operator can accept the variance error degree as l, when

$$l \leq |D(X) - D(Y)|$$

The operator considers that it does not meet the demand, when

$$l > |D(X) - D(Y)|$$

The operator is deemed to satisfy the condition.

Similarly, assume that the minimum correlation coefficient that the operator can accept is m (m>0), when

$$\frac{Cov(X,Y)}{\sqrt{D(X)} * \sqrt{D(Y)}} \leq m$$

The operator is deemed to satisfy the conditions when

$$\frac{Cov(X,Y)}{\sqrt{D(X)} * \sqrt{D(Y)}} > m$$

The operator regards it as unsatisfactory.

In summary, when (1) (2) (3) is satisfied at the same time, the operator regards the machine scoring as equivalent to manual scoring. If one of the items is not satisfied, the operator regards the machine scoring as invalid.

6. CONCLUSION

This model uses Python to filter data to eliminate irrelevant conditions and reduce workload. A simple model is used to calculate complex models, which greatly reduces the amount of calculation. The hypothesis test model can be used to accurately determine whether there is a significant difference between manual scoring and machine scoring.

However, replacing the message with a word set that appears more frequently will still result in larger errors and missing important information in the message. Moreover, the model mentions the use of correlation coefficients for testing, but the correlation coefficients can only test whether they have linear correlation. There may be cases where two sets of data have correlations but not linear correlations but nonlinear correlations. Therefore, the model still has a certain degree of plasticity for further work.

REFERENCES

[1] Sheng Su. Probability Theory and Mathematical Statistics [Fourth Edition]. Zhejiang University

[2] Jia Junping. Statistics [Seventh Edition]. Renmin University of China

[3] Xue Wei. Statistical Analysis and Application of SPSS [Fifth Edition]. Renmin University of China

[4] Sun Haifeng, Zheng Zhongshu, Yang Wuyue's analysis and mining of online recruitment information. Beijing Forestry University. 2017

[5] Liang Changming, Sun Dongqiang. Empirical research on the evaluation index system of Weibo based on Sina's popular platform. Shandong Normal University. 2015