A study on the subject classification of the NBA match reports

Zhe Wang[1,a], Baoan Li[2], Xueqiang Lv[3] and Zhian Dong[4]

[1]Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing, China
[2]Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing, China
[3]Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing, China
[4]Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing, China

Abstract. In this paper, we study the task of template building in automatically generate NBA match reports from NBA live text. As a preliminary study, we collect and process the historical reports compiled by the editors and get different kinds of sentences. Our innovative proposal is to divide the NBA match reports into 11 categories, which covering almost all cases. We use different machine learning methods to classify sentences. Each class finally constructs a template library to service the next automatic writing. By comparing different methods, we get a higher accuracy classification structure. The evaluation results show that our method does construct a template library.

1 Introduction

Basketball is more and more popular because of its athletics and appreciation. NBA match reporters occupy a large proportion in sports news as an important source of information for people to understand basketball match result. Therefore, the research of computer automatic writing for basketball is becoming a hot topic. The idea of automatic writing has a long history. With the development of big data, Natural Language Processing and other artificial intelligence technology, the exploration and practice of automatic generation of match reports has been gradually developed in recent years [1].

In this paper, we hope to construct templates for sentence classification in NBA match reports to apply to NBA match reports automatic writing. Through the analysis of the writing behavior of the existing NBA match reports, we summarize writing words, writing subjects, writing emphasis and writing methods. The subject of writing refers to the situation of the competition in a certain period based on the time change of the match. There are 3 main tasks in the experiment. First, we determine the subject classification by analyzing the sentences in the NBA match reports. Second, we mark the subject tagging of the sentences in the NBA match report. Third, we classify the sentences in the NBA match report. We first rank all candidate sentences according to a sentence scoring scheme and then select a few sentences according to certain criteria to form the final generated new [2]. The final match reports should have a real result and a strong readability.

The automated categorization of texts into predefined categories has witnessed a booming interest in the last ten years, due to the increased availability of documents in digital form and the ensuing need to organize them [3].

2 Subject categories

To better understand the editorial writing behavior, we need to define each subject. The content category mainly refers to the facts that are reflected in the competition process. It constantly changes with the time of the game and the performance of the players and the teams. They are basically fixed in 11 subjects.

- Definition 1: expand the score gap, Team A leads the team B S points at time T1. After the period T2-T1, the team A is not continuously chased by the team B, and at the time T2, the team A leads the team B score > S points, then call this team A expand the score in the period T2-T1.
- Definition 2: small climax, Team A leads the B team S points (or S points behind) at time T1. After the period T2-T1, team A scores consecutively and score => 8 points. If team B does not score or score <= 3 points, then called this team A played a small climax during the period T2-T1.
- Definition 3: stable the score gap, Team A leads the B team S points at time T1. After the period T2-T1, team A leads the team B with score < S points. After a pause or the player adjust himself, the team A leads the team B approximately equals S points in the

* Corresponding author:[a] Zhe Wang: 525322746@qq.com

period T3-T2, then call this team A stable score in the period T3-T2.

- Definition 4: maximum difference, Team A leads the team B S points at time T1. After T2-T1, T3-T2, ...Tn+1-Tn, team A leads team B with score <= S points, then call this team A leads score is the maximum difference at time T1 for this game.
- Definition 5: score shortage, In the period T2-T1, if team A does not score at least 3 consecutive rounds, or if there are only a few scores in multiple rounds, team A will be called in scoring drought during the period T2-T1.
- Definition 6: reduce the score gap, Team A behind the B team S points at time T1. After the period T2-T1, the team A behind the team B score < S points, then call this team A to reduce the difference in period T2-T1.
- Definition 7: same score, Team A behind team B at time T1. Team A and team B have the same score during the period T2-T1, team A is said to be equal to team B at time T2.
- Definition 8: both sides score, In the period T2-T1, team A and team B alternated to score, then team A and team B were said to face each other.
- Definition 9: Both sides score shortage, In the period T2-T1, neither A nor B scored, and the team A and the team B hit each other.
- Definition 10: overtake, Team A behinds the team B S points at time T1. After the period T2-T1, team A leads team B, then team A claimed that team A overtakes the score at time T2.
- Definition 11: alternate lead, Team A leads the team B S points at time T1. After the period T2-T1, team A behind team B. After the period T3-T2, Team A leads team B, then call this team A and team B call alternate lead during the period T3-T1.

Table 1. Classification example

category	Typical sentence
expand the score gap	Shortly after the start of the third quarter, Adams scored a 61-53 lead in his backhand layup.
small climax	In the 45 second time of the festival, Green scored two points, and the Golden State Warrior made a wave of 8-0, drawing the gap by 94-80.
stable the score gap	Hood scored a key three points, and jazz remained steady.
Maximum difference	The gap between the two sides has reached 19 points, and Jazz is hard to chase the score to 44-58.
score shortage	They narrowed the gap to 2 points but no score again in the last 5 minutes and 46 seconds.
reduce the score gap	Chandler scored three points, and Denver Nuggets followed 98-104 last 4 minutes of the fourth quarter.
same score	In the game with 1 minutes and 16 seconds, Cousins scored two points, Sacramento Kings scored 102-102.
both sides score	The two teams fought against each other in large numbers, and each scored more than five points in 2 minutes in the fourth quarter.
both sides score shortage	In 4 minutes, the Orlando Magic failed to score, and the Los Angeles Clippers score only once.
overtake	In the second 54.2 seconds of the game, Henson scored 3 points and the Bucks were 96-94.
alternate lead	The two teams returned to the same race line for more intense competition and several times to take the lead.

3 Corpus

The real-time data is annotated according to historical news, and the train set is obtained [4]. We used a multi-person cross-labeling method and annotated 600 NBA match report data. First, the machine automatically removed the background information, and then divided the sentence into a sentence-level. Finally, we assigned data to N individuals to annotate, and cross-validate the labeled results.

Since there is some background information when the NBA Newsletter was written, it has nothing to do with the facts of the game. The live text cannot be generated at all. It requires historical data and professional knowledge. We need to remove the background information. We divide the rest of the data by periods. Each line of data is represented by a sentence.

Experiments need to be marked with many NBA match reports corpus to study the writing characteristics of editors. We label data according to the following rules: First, we designate the labeling rules, complete the labeling according to the categories we have defined in advance, and conduct centralized communication with the corpus makers to adjust the rules. Secondly, we assign the N parts of the corpus to N individuals for annotation. We extract each sentence in the NBA match report, and label the data from the structure and content, and use the '\t' key to segment it.

In the experiment, cross check is adopted in the experiment. The data is checked by multi wheel, and the check annotation data are compared with the original. The data of the problems are unified and discussed. Finally, the results of the annotation are confirmed.

4 Feature extractions

Due to the large number of vocabularies in the text and the large scale of the dictionary, when constructing the vector space model, it will encounter dimension disasters, and construct a stop word vocabulary, use different methods for feature selection or feature extraction to reduce dimensions.

The feature selection is mainly selected as a feature word by selecting some keywords that are representative without Influence the classification effect. The usual practice is to select some algorithms and score the feature

of each dimension. The most important feature is to select the features of the score. In the experiment, the TF-IDF algorithm is more suitable for the selection of Chinese text feature words is used to assign specific weights to each word, and the Boolean weight method is used to select feature words for NBA match reports, then constructing a vector space model. Use SVM for classification. Before text classification, the corpus is preprocessed, feature sets are determined, and feature words are extracted.

This paper mainly uses the Mutual Information (MI), the Information Gain (IG), the Chi-squared Distribution (CHI), and the Weighted Log Likelihood Ratio (WLLR) to extract feature words. The results were tested by intersection and union.

The data is preprocessed during the experiment: word segmentation. We construct a new vocabulary list (NBA player name, team name, action specific nouns, etc.) and a stop word list (NBA player name, team name, punctuation, numbers, auxiliary words, etc.) to make entries more consistent with the rules of NBA match report and remove useless words.

Handle scores and convert different scores into corresponding words. Through the analysis of the corpus, we can see that the structure when the report form of the score is xx-xx, and we define it as number one and number two. The score in report is divided into:

1.Report on the team leading the score: the number one is bigger than the number two. In this case, it can be divided into leading the score, expand the score gap and overtake. As these situations are difficult to distinguish only by differences in scores, we unify the markup to “expand the score gap”. In addition, it also be divided into small climax. Usually, the number one is one to two digits, the number two is one digit, and there is a big gap between the two numbers, marking this situation as “small climax”.

2.Report on the team behind the score: the number one is less than or equal to the number two. In this case, according to the needs of the existing classification, it is divided into: when the number one is less than the number two, it is marked as “reduce the score gap”; the number one is equal to the number two, marking this situation: “same score”.

The comparison experiments were conducted from the pre-processed corpus at different stages, setting different thresholds, setting different dimensions, overall extraction and extraction in the same class to obtain the best classification results.

5 Subject classification of sentences

The experiment analyzes the factors that affect the classification of sentence structure, label categories, and the number of corpora. We select several classification methods suitable for the sentence of NBA match reports. include: Naive Bayes, KNN, SVM, Decision tree and Neural network [5].

Naive Bayes is a classification algorithm based on probability theory. Its advantages lie in its simple principle, implementation, high learning and prediction efficiency. KNN is a method for classifying majority voting by determining the category of the sample under test from the largest of the k-neighbor samples of the test sample. SVM maps all points to be classified to the "high-dimensional space", then finds a "super-plane" in the high-dimensional space that separates these points. The decision tree is based on the known probability of occurrence of various conditions, by calculating the influence of different factors on the classification result, various features are taken as the branch nodes of the tree. The final category is taken as the leaf node, which is determined by each decision. The neural network classifies by simulating the human brain's way of thinking and learning. The neural network can often play the best role when the data set is large. The best classification method was selected through the comparison of the results of the classification tests on different methods.

6 Experimental results and analysis

6.1 Experimental data

A total of 3033 sentences of NBA match reports were marked. After removing the stop words, some short sentences were removed because they were all useless words. After taking out, there were 3,024 sentences.

The verification method chooses k-fold cross-validation. The 1/k of the data set is used as the test set, and the rest is used as the train set. Each model is trained k times and tested k times. The accuracy rate is the average of k times. The value of k is set to 10, so the train set occupies 90% of the total data and the test set occupies 10% of the total data. Different methods are compared by the final test results. We choose the best method to classify.

6.2 Evaluation index

The precision, recall, and f1-score used in text classification evaluation and the accuracy rate were used for evaluation. For category C, the results of the classification can be divided into the following situations:

1) The original class C is divided into class C, the number is denoted as a

2) The original non-C class is divided into Class C, and the quantity is denoted as b

3) The original class C is divided into non-C class, the number is denoted as c

4) The original non-C class is divided into non-C class, the number is recorded as d

Precision: $$P = \frac{a}{a+b} \times 100\% \quad (1)$$

Recall: $$R = \frac{a}{a+c} \times 100\% \quad (2)$$

f1-score: $$F = \frac{2 \times P \times R}{P+R} \times 100\% \quad (3)$$

Accuracy: $$A = \frac{a+d}{a+b+c+d} \times 100\% \quad (4)$$

6.3 Experimental results and analysis

By using TF-IDF algorithm to construct feature vectors, using SVM to train and predict. Texts processed at different stages of the text are tested separately and the following results are obtained:

When using SVM based on TF-IDF algorithm for text classification, the train set has a good degree of fitting, the accuracy reached 88.84%, but there is a big gap for the test set, only up to 77.25%.

The emphasis of SVM based on Boolean weight method lies in the selection of feature words. Selecting good feature words has a great influence on the accuracy of the model. The following needs to use some text feature extraction algorithm to extract more feature words to improve the classification model accuracy.

At the beginning, we set a threshold and all words larger than the threshold are extracted as feature words. Since the threshold under each category is set higher, the dimension of the feature words obtained is lower. The accuracy of the model also shows different states, but the overall situation is poor. Because the MI and IG methods tend to get lower frequency words, the accuracy rate is very low when the Boolean eigenvector has a low dimension. Next, we reduce the threshold. So, the number of eigenvalues is increased, which increases the dimension of Boolean eigenvectors. The results are shown in the table 2.

Although the accuracy of the train set is lower than that of the TF-IDF algorithm, the accuracy of the test set has been greatly improved. Intersection and union are usually good, but union sets tend to be over fitting because of their high dimension, which makes the accuracy of the test set decrease.

In the experiment, we extract n feature words to test the effect of the number of feature words on the result, as shown in the table 3.

Table 2. SVM classification results of different feature words extraction method

	Threshold	Feature Count	Accuracy (train set)	Accuracy (test set)
MI	0.1	131	0.6848	0.6267
IG	0.03	141	0.8098	0.7646
CHI	3	111	0.8217	0.7881
WLLR	0.1	112	0.8261	0.7811
Intersection		152	0.8377	0.7785
Union		213	0.8610	0.7805

It can be seen from the results that with the increase in the number of terms, the degree of fitting of the train set becomes higher and higher, the accuracy becomes higher and higher and the test set fluctuates within a relatively small range. Excessive dimensions can cause overfitting, making the test set less effective.

In addition, classification tests are performed on the above-mentioned several eigenvectors by using different classification methods. Choose the best classification method. Here is a simple listing of the classification results under a 140-dimensional Boolean eigenvector in the table 4.

The classification accuracy of KNN and decision tree algorithms is poor, and it is not suitable for the classification of the sentences of NBA match reports. Naive Bayes is slightly less effective. The results of neural networks and SVM are similar, but SVM is faster. SVM is finally selected as a classification method for classification. The results of the various types of results are shown in the table 5 and table 6.

Table 3. SVM classification results of Boolean eigenvector with different dimensional

	Train set				Test set			
dimension	accuracy	precision	recall	f1-score	accuracy	precision	recall	f1-score
20	0.6658	0.70	0.67	0.66	0.6525	0.69	0.65	0.65
60	0.7810	0.79	0.78	0.78	0.7593	0.78	0.76	0.75
100	0.8109	0.82	0.81	0.81	0.7808	0.79	0.78	0.78
140	0.8265	0.83	0.83	0.82	0.7844	0.79	0.78	0.78
180	0.8399	0.84	0.84	0.84	0.7844	0.79	0.78	0.78
220	0.8460	0.85	0.85	0.84	0.7848	0.79	0.78	0.78
260	0.8492	0.85	0.85	0.85	0.7817	0.79	0.78	0.78
300	0.8603	0.86	0.86	0.86	0.7801	0.79	0.78	0.77
340	0.8678	0.87	0.87	0.87	0.7831	0.79	0.78	0.78

Table 4. CHI extracts classification results of different classification methods for 140-dimensional Boolean eigenvector

	Train set				Test set			
method	accuracy	precision	recall	f1-score	accuracy	precision	recall	f1-score
Naive Bayes	0.7876	0.79	0.79	0.78	0.7732	0.78	0.77	0.77
KNN	0.7169	0.75	0.72	0.69	0.6859	0.71	0.69	0.66
SVM	0.8265	0.83	0.83	0.82	0.7844	0.79	0.78	0.78
Decision Tree	0.8537	0.87	0.85	0.85	0.7232	0.74	0.72	0.78
Neural Network	0.8262	0.83	0.83	0.82	0.7814	0.79	0.78	0.78

Table 5. SVM classification results of CHI extract 140-dimensional Boolean eigenvector (train set)

category	precision	recall	f1-score	support
expand the score gap	0.78	0.87	0.82	769
small climax	0.88	0.84	0.86	474
stable the score gap	0.87	0.66	0.75	116
maximum difference	0.93	0.75	0.83	55
score shortage	0.84	0.78	0.81	274
reduce the score gap	0.81	0.90	0.85	525
same score	0.94	0.94	0.94	88
both sides score	0.83	0.50	0.62	90
both sides score shortage	1.00	0.72	0.84	32
overtake	0.84	0.81	0.83	220
alternate lead	0.91	0.68	0.78	78
avg/total	0.83	0.83	0.83	2721

Table 6. SVM classification results of CHI extract 140-dimensional Boolean eigenvector (test set)

category	precision	recall	f1-score	support
expand the score gap	0.68	0.86	0.76	73
small climax	0.90	0.77	0.83	61
stable the score gap	0.75	0.27	0.40	11
maximum difference	0.50	0.67	0.57	3
score shortage	0.93	0.68	0.79	41
reduce the score gap	0.72	0.85	0.78	55
same score	0.73	0.89	0.80	9
both sides score	0.75	0.55	0.63	11
both sides score shortage	1.00	0.50	0.67	2
overtake	0.76	0.79	0.78	24
alternate lead	1.00	0.85	0.92	13
avg/total	0.80	0.78	0.77	303

7 Conclusions

We analyze and compare the accuracy of each category, found that the categories appear more often in the text of the match report and have more than one explicit feature word can be obtained with high accuracy, and some cases with low frequency, even when multi-word analysis is required. The accuracy is low.

Analysis of test results: TF-IDF algorithm has a better recall compared to Boolean weight method using CHI as a feature word extraction, but overall the method chosen in the experiment is more accurate on the test set. The other two methods contain situations in which some sentences are too short to be classified correctly. Some sentences maybe contain multiple tags and error of manual annotation also lead to a decline in the accuracy. In addition, some sentences also need some logical judgments, and machines are difficult to classify.

If we have a larger train set, the effect of the neural network will be more prominent. In the later period, we can pre-mark corpus through SVM, and then perform manual verification checks, and finally apply it to RNN or other models for classification.

The paper mainly classifies the report data. Firstly, preprocessing the data and extracting the annotated data. Secondly, the extracted sentences are manually annotated and cross validated. Finally, the classification features are classified by SVM classification. The experimental results show that the proposed method is very effective for the sentence classification and provides service for subsequent domain word extraction and domain template library extraction.

Acknowledgements

This project was supported by the Funding Project for Natural Science Foundation of China (Grant No. 61671070) and the Opening Project of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (Grant No. ICDD201708).

References

1. Chen Y, Xueqiang L, Zhou J, et al. Research on Automatic Writing of NBA Sports News. *Acta Scientiarum Naturalium Universitatis Pekinensis*, **53(2)**, 211-218 (2017)
2. Zhang J, Yao J G, Wan X. Towards Constructing Sports News from Live Text Commentary. *Meeting of the Association for Computational Linguistics*, 1361-1371 (2016)
3. Manyika J, Chui M, Brown B, et al. Big Data: The Next Frontier For Innovation, Competition, And Productivity. *Analytics* (2011)
4. Wang W, Xueqiang L, Zhang K, et al. Research on Automatic Writing of Football Game News. *Acta Scientiarum Naturalium Universitatis Pekinensis* (2018)
5. Sebastiani F. Machine learning in automated text categorization. *Acm Computing Surveys*, **34(1)**, 1-47 (2002)