Using Keyword Features to Automatically Classify Genre of Song Ci Poem

Yong Mu(✉)

Department of Chinese Language and Literature, School of Humanities, Tsinghua University, Beijing 100084, China
muyong2004@126.com

Abstract. The category of subjective aesthetic was distinguished into automatic text categorization of natural language processing problem. The Song dynasty poems were collected and randomly divided into training set and testing set based on keyword features. Three classification methods, which are K-nearest neighbor, naive Bayes classifier and support vector machine, were used to classify the poems' genres. Results showed that support vector machine can classify best and achieved above 95% accuracy.

Keywords: Text classification · Keyness · Genre of song-Ci poem · Natural language processing

1 Introduction

The Song-Ci (宋词) poem classification was first introduced by Zhang Yan (张綖) in the Ming Dynasty. He classified Song-Ci poems in one of his writings named "Shiyu Tupu • Fanli (诗余图谱·凡例). " There are two different styles of Song-Ci poetry: the Wan-Yue (婉约, delicate restraint), and the Hao-Fang (豪放, heroic abandon). Wan-Yue has a graceful and restrained characteristic, similar to Qin Shao-you (秦少游) 's works, whereas Hao-Fang is more like the works of Su Zi-zhan (苏子瞻) that makes readers feel bold and unconstrained. This classification was made based on personal knowledge, experience and perception, which effectively grasped the connotation and artistic conception. However, the process of classification necessitates a lot of time and energy, and the result could be biased by subjective judgment of personal preference.

Since the 1980's, Chinese scholars began to combine features of ancient poetry and natural language processing (NLP) technology to conduct computer-aided research on poetry understanding, such as "Computer supportive in Ancient poems research" laboratory, which was a cooperation between the Institute of Computational Linguistics of Peking University and Institute of Ancient Literature in Yuanze University. However, there is still not much research in this field. Zhou [10][6] introduced the concept of "Computational Poetics" and carried out a series of related research work. Li [2] proposed natural language processing technology based on word connection and applied it to understand the meaning of poetry.

Q. Lu and H.H. Gao (Eds.): CLSW 2015, LNAI 9332, pp. 478–485, 2015.
DOI: 10.1007/978-3-319-27194-1_48

Training computers to learn to "understand" and "categorize" the human language is still a challenge. Word, phrase and grammatical level have already been widely studied in NLP, but the semantic and emotional aspects are still narrow, particularly in style and genre, which is classified as the embodiment of personality in literary language.

Generally, the original document must be converted into a data structure that a computer can understand through the process called "pretreatment" before the computer is trained. This pretreatment process consists of many steps, such as word segmentation, tagging, stop word filtering, and so on. Because there is no space as the boundary between one word to another in Chinese language, one of the most important steps in the pretreatment process of Chinese language is text representation.

At present, there are two studies that researched how to classify the Song Ci genre in different text representation policies. The first was done by Yi in 2005 [8]. The researcher used the word segmentation policy in the text representation step, and applied a genetic algorithm with the three classification methods, which are Naïve Bayes, decision tree, and support vector machine, to categorize Song-Ci genre. The result found that the support vector machine gave the best solution set with 85% of average accuracy. In 2008, Wu introduced the frequent words co-occurrence policy to improve the result [9]. In the experiment, a vector space model with three classification methods, which are K-nearest neighbor, naive Bayes classifier, and support vector machine, was applied. The best result achieved 91.45% of average accuracy. Recently, Li [1] published an interesting piece that compared the effect of classifying the artistic conception of Tang Poetry based on single-word segmentation policy and word segmentation policy. The result found that the solution of word segmentation policy is slightly lower in accuracy.

Based on the characteristics of the language used in Chinese classical poems, Song-Ci is very concise and exquisite, normally using single characters to express complicated mood and emotion. Therefore, using the word segmentation system of modern Chinese to classify Song-Ci, as done in Yi [8] and Wu [9], proved to be a less efficient system. The significance of this research shows that text representation based on single-word segmentation policy and building keyness based on big data of Song-Ci [7] is more effective. Three classification methods, which are K-nearest neighbor, naive Bayes classifier and support vector machine, were trained to classify the poems' genre of 151 poems in Song Dynasty, in order to building the Song-Ci classification machine and classifying the big data of Song-Ci. The results were compared with the works of Yi and Wu [8][9].

2 Methodology

2.1 Experimental Design

The two categories of Wan-Yue and Hao-Fang from the Song-Ci genres were translated as text representation in this research. Evaluating which poem belonged to which genre can be classified as a text categorization problem. In the study, the preprocessed corpus was divided into two categories: Wan-Yue and Hao-Fang. Automatic text classifying was composed of three steps, which are the representation step, algorithm classifying step and result evaluation step.

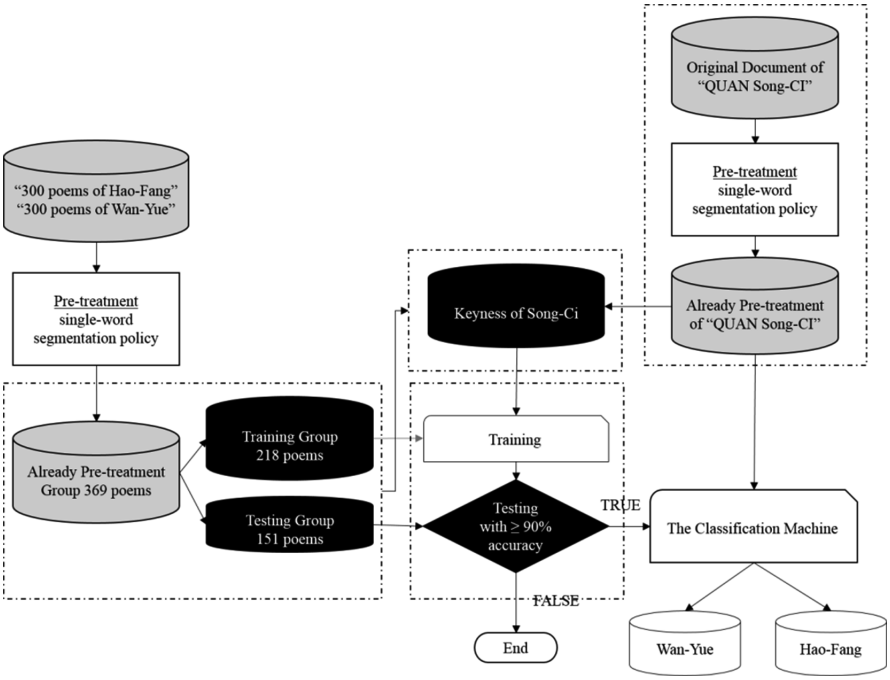

Fig. 1. Experimental flow chart

The representation step extracts features from documents by feature-selection strategy, and then transforms the document into a vector-space model for features. The algorithm classifying step chooses the most suitable algorithm for this type of document. Lastly, the result evaluation step chooses some classic evaluation indexes in the field of information retrieval, such as accuracy, recall rate, equilibrium value. The experiment procedure can be descripted as shown in figure 1 and figure 2.

Step 1. Selecting artificial classified works from Quan Song Ci (全宋词, The Whole Collection of Song Poetry) as the training set and testing set.
Step 2. Feature extraction from the training set and test set.
Step 3. Numeralization of the training set and test set by the features; each work is transformed into a vector.
Step 4. Modeling by using classification algorithm to train the data of training set, creating a classifier.
Step 5. Classifying the test set using the trained classifier and evaluating the results, selecting the highest score of the test model as the Song Ci genre classifier.
Step 6. Distinguishing the genre of Song Ci automatically by this classifier

Fig. 2. Experimental Step

2.2 Data Generating

The details of experimental data were shown in table 1 and table 2. Table 1 shows the number of samples in the training group and testing group. Testing groups are divided according to author's names and shown in table 2. This found that there is an author named Lu You who wrote both types of Song-Ci genre poetry.

Table 1. The number of sample in training group and testing group

Group	Song-Ci genre		Total
	Hao-Fang	Wan-Yue	
Training group	110	108	218
Testing group	66	85	151
Total	176	193	369

Table 2. Testing group divided by authors

Song-Ci genre			
Hao-Fang		Wan-Yue	
Author	Quantity	Author	Quantity
Ouyang Xiu	4	Zhang Xian	5
Huang Tingjian	4	Yan Shu	5
He Zhu	5	Liu Yong	13
Shu Shi	14	Yan Jidao	5
Ye Mengde	4	Zhou Bangyan	8
Zhang Yuangan	4	Qin Guan	6
Lu You	9	Fan Chengda	4
Zhang Xiaoxiang	6	**Lu You**	8
Liu Guo	4	Jiang Kui	10
Liu Kezhuang	8	Shi Dazhu	7
Wen Tianxiang	4	Wu Wenying	14

To generate the experimental data, two books of "Three Hundred poems of Hao-Fang" [4] and "Three Hundred poems of Wan-Yue" [5] were used as the reference for experimental data. According to these books that consist of poems from the Tang, Song, Ming and Qing Dynasty, it is necessary to classify poetry according to the time period before starting the experiment. After the classifying, 231 poems in Tang, Ming and Qing Dynasty were filtered out. The example group had 369 poetry from the Song dynasty, which can be divided into 176 poems of Hao-Fang and 193 poems of Wan-Yue. After that 369 poetry in example group were randomly divided into two groups, which were the training group and testing group by 3:2 ratio. The details are shown in table 1.

2.3 Solution Procedures

2.3.1 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a widely used method of classification that is based on learning the training data examples to classify new data and make predictions. Given

a test document, the system finds the nearest K "neighbors" in it, and gives grades to the candidates according to the classification of these neighbors. Taking the similarity of neighbor texts and test texts as the weight of neighbor classification, if part of the texts of the K "neighbors" belong to the same category, sum the weight of each neighbors in this category for the similarity of the document which will be classified and test. By sorting the scores of the candidate classification then giving a threshold, we can determine the classification of test documents. The decision rules can written as follows:

$$y(x, c_i) = \sum_{d_i \epsilon kNN} sim(x, d_i) y(d_i, c_j) - b_j$$

The value of $y\left(d_i, c_j\right)$ is 0 or 1, 1 said that the document d_i belongs to the classification C_j, 0 said that the document d_i does not belong to the classification C_j; $sim(x, d_i)$ is the similarity of test text x and training text d_i; b_j is the binary decision threshold , which is generally taken by including the angle cosine of two vectors.

2.3.2 Naive Bayes Classifier

Naive Bayes Classifier (NBC) is a method of classification that analyzes the probability of class variables and data features set. The hypothesis of Naive Bayes is that the data features and specific categories are independent of each other. When the text vector component is a Boolean value, 0 indicates that the corresponding keyword does not appear in the document while 1 indicates it appears. The probability of the document Doc belongs to the class C is:

$$P\left(Doc\left(F_j\right)|C\right) = \frac{1 + N\left(Doc\left(F_j\right)|C\right)}{2 + |D_c|}$$

$P\left(Doc\left(F_j\right)|C\right)$ is the Plath estimate of the conditional probability of feature F_j which appears in the class C, $N\left(Doc\left(F_j\right)|C\right)$ is the number of documents that features F_j which appears in the class C, $|D_c|$ is the number of documents in class C.

2.3.3 Support Vector Machine

The original Support Vector Machine (SVM) algorithm was invented by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963, but the current standard incarnation was proposed by Corinna Cortes and Vapnik in 1993 and published in 1995 [1]. The SVM showed good effect to solve nonlinear, small sample problems. The idea of SVM, in simple terms, is to find a decision surface in an vector space, and the plane can segment two class data points best [11].

2.4 Keyness

Keyness is a term used in linguistics to describe the likelihood a word or phrase has of being "key" in its context. Compare this with collocation, the quality linking two words or phrases usually assumed to be within a given span of each other. Keyness is a textual feature, not a language feature (a word has keyness in a certain textual context but may well not have keyness in other contexts), so it is very suitable for feature genre classification.

Table 3. The 45 keywords of Hao-Fang

Keyword	江	山	百	万	兴	雄	旗	北
Pinyin	*jiang1*	*shan1*	*bai3*	*wan4*	*xing1*	*xiong2*	*qi2*	*bei3*
Meaning	*river*	*mountain*	*hundred*	*ten hundred*	*prosper*	*hero*	*flag*	*north*
Keyword	经	事	关	战	剑	朝	英	军
Pinyin	*jing1*	*shi4*	*guan1*	*zhan4*	*jian4*	*chao2*	*ying1*	*jun1*
Meaning	*pass through*	*affair*	*barrier*	*war*	*sword*	*court*	*hero*	*army*
Keyword	吾	涛	可	问	胡	冲	帆	南
Pinyin	*wu1*	*tao1*	*ke3*	*wen4*	*hu2*	*chong1*	*fan1*	*nan2*
Meaning	*I*	*billow*	*can*	*ask*	*barbarian*	*rush*	*sail*	*south*
Keyword	壮	血	投	矣	要	之	白	神
Pinyin	*zhuang4*	*xue3*	*tou2*	*yi1*	*yao4*	*zhi1*	*bai2*	*shen2*
Meaning	*strong*	*blood*	*cast*	*interjection*	*need*	*this*	*white*	*deity*
Keyword	物	古	封	汉	潇	安	虎	诸
Pinyin	*wu4*	*gu3*	*feng1*	*han4*	*xiao1*	*an1*	*hu3*	*zhu1*
Meaning	*thing*	*ancient*	*envelop*	*Chinese*	*sound*	*safe*	*tiger*	*Every*
Keyword	磨	铁	发	名	闻			
Pinyin	*mo2*	*tie3*	*fa1*	*ming2*	*wen2*			
Meaning	*grind*	*iron*	*send out*	*name*	*hear*			

Table 4. The 49 keywords of Wan-Yue

Keyword	华	春	情	帘	莺	愁	娇	暖
Pinyin	*hua2*	*chun1*	*qing2*	*lian2*	*ying1*	*chou2*	*jiao1*	*nuan3*
Meaning	*flower*	*spring*	*feeling*	*curtain*	*warbler*	*anxious*	*tender*	*warm*
Keyword	转	香	怕	寄	宝	阑	小	薄
Pinyin	*zhuan3*	*xiang1*	*pa4*	*ji4*	*bao3*	*lan2*	*xiao2*	*bo2*
Meaning	*turn*	*fragrant*	*fear*	*send*	*treasure*	*separate*	*small*	*thin*
Keyword	思	甚	院	琴	禁	留	水	肠
Pinyin	*si1*	*shen4*	*yuan4*	*qin2*	*jin1*	*liu2*	*shui3*	*chang2*
Meaning	*miss*	*very*	*yard*	*lyre*	*bear*	*stay*	*water*	*bowel*
Keyword	芳	窗	梅	寒	碎	棠	团	拈
Pinyin	*fang4*	*chuang1*	*mei2*	*han2*	*sui4*	*tang2*	*tuan2*	*nian1*
Meaning	*fragrant*	*window*	*winter sweet*	*cold*	*fragmentary*	*begonia*	*group*	*pick up*
Keyword	红	半	又	燕	新	悄	眉	绿
Pinyin	*hong2*	*ban4*	*you4*	*yan4*	*xin1*	*qiao1*	*mei2*	*lv4*
Meaning	*red*	*half*	*again*	*swallow*	*new*	*quiet*	*eyebrow*	*green*
Keyword	魂	涯	晴	雁	还	夜	梳	腮
Pinyin	*hun2*	*ya2*	*qing2*	*yan4*	*huan2*	*ye4*	*shu1*	*sai1*
Meaning	*spirit*	*margin*	*fine*	*wild goose*	*back*	*night*	*comb*	*cheek*
Keyword	玉							
Pinyin	*yu4*							
Meaning	*jade*							

In this research, the concept of "keyness" was used to extract keywords from the training sets of Wan-Yue and Hao- Fang. The feature sets according to a certain threshold number were established. This study used Antconc to extract keywords from sample groups of Wan-Yue and Hao-Fang with the keyness⩾3. The results after extracting found that there were 49 and 45 keywords in Wan-Yue and Hao-fang, as shown in table 3 and table 4 respectively. These 94 keywords were used as the feature items (each keyword is a feature) to calculate the frequency of each feature in different texts in both the training set and testing set.

3 Results

To compare the performance of the classification methods, Precision (P) and Recall (R) were used——the most common evaluation standard index in the field of information retrieval. The precision indicates what proportion is correct among the selected document; the recall rate indicates how much was selected by the classifier in all the correct results. Table 4 shows the experimental results of the three classification methods compared with the work of Wu in 2008.

Table 5. The Comparison Performance of classification method (Unit :%)

	P%		R%	
	Hao-Fang	Wan-Yue	Hao-Fang	Wan-Yue
KNN w-fwc (Wu, 2008)	83.3	77.3	75.0	85.0
KKK w/o-fwc (Wu, 2008)	84.6	98.3	93.1	88.5
KNN	**90.9**	72.7	76.9	88.9
NBC	63.6	**100.0**	**100.0**	73.3
SVM	**90.9**	**100.0**	**100.0**	**91.7**

From table 5 we can see, KNN classification algorithm is more accurate on the Hao-Fang set than on the Wan-Yue set. NBC algorithm is very accurate on the Wan-Yue set, but undesirable on Wan-Yue set. For SVM algorithm, it shows 100% precision and a high recall rate on Wan-Yue set, 90.9% precision and perfect recall rate on Hao-Fang set. SVM with single-word segmentation had the best performance out of the three classification methods. When comparing with previous research, the performance of SVM with single-word segmentation is better than Wu 2008 [9].

4 Conclusion

In view of the human errors and limitations, this research considers using modern computer technology to support the classification of the large number of Ci poems. This does not only significantly reduce manpower and time, but also improve the efficiency. Moreover, by the quantitative method, the machine learning can deal with poems that are considered as "only to be sensed, not explained," and also provide some new information and reference, which the traditional research method in the literature cannot easily find. In this paper, we extract keywords where the keyness is more than 3. If the threshold is relaxed, with Keyness$\geqslant$2 as selection criteria, the feature vector space will be constructed by more dimensions. This may further improve the classification accuracy, but of course this remains to be tested.

References

1. Qi, L.: Research and Implement of Classical Poetry Artistic Conception Classification. The master degree thesis of Donghua University (2014). (in Chinese)
2. Ly, L.: A study on term connection oriented NLP technique and its applications [Ph.D. Thesis]. Chongqing: Chongqing University (2004). (in Chinese)
3. Ying, L.: Statistical Linguistics. Tsinghua University Press, Beijing (2014)
4. Zhuqin, L.: 300 Hao Fang Ci. San Qin Press, Xi'an (2003). (in Chinese)
5. Yin, L.: 300 Wan Yue Ci. San Qin Press, Xi'an (2003). (in Chinese)
6. Jingsong, S., Changle, Z., Yihong, L.: The Establishment of the Annotated Corpus of Song Dynasty PoetryBased on the Statistical Word Extraction and Rules and Forms. Journal of Chinese Information Processing **3** (2007). (in Chinese)
7. Guizhang, T.: Quan Song Ci. Zhonghua Book Company, Beijing (1965). (in Chinese)
8. Yong, Y.: A Study on Style Identification and Chinese Couplet Responses Oriented Computer Aided Poetry Composing. Ph.D. Dissertation of Chongqing University (2005). (in Chinese)
9. Chunlong, W.: The Research of Computer Assistant Analysis on Chinese Song Poems' Style. The master degree thesis of Xiamen University (2008). (in Chinese)
10. Zhou, C.L.: An Introduction to Computation of Mind and Brain. Tsinghua University Press, Beijing (2003). (in Chinese)
11. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273 (1995)