DLINE JOURNALS

Journal of Digital Information Management

JDIM

ISSN 0972-7272

A Recognition Learning System Based on Poetry Database and Text Pattern Function

Wang Jinli[1,*], Zheng Xiang[2]
[*]Corresponding author
[1]Universitas Indonesia, West Java Province
Indonesia
[2]Universitas Islam Indonesia
A metropolitan suburb of Makassar in the province of South Sulawesi
Sfdsfsdwe3243@protonmail.com

ABSTRACT: *This article introduces a recognition learning system based on a poetry database and text pattern function. The system aims to help students better understand and remember poetry and improve the efficiency of Chinese language learning. The system's core components include a poetry database, text pattern recognition, and personalized learning modules. The poetry database contains a wealth of poetry works, providing students with rich learning resources. The text pattern recognition module can automatically recognize patterns in poetry, such as rhyme and contrast, to help students understand the structural characteristics of poetry. The personalized learning module provides suggestions and exercises based on students' learning progress and abilities to achieve precise teaching.*

Subject Categories and Descriptors: [H.3 INFORMATION STORAGE AND RETRIEVAL]; Linguistic processing: [H.2.4 Systems Textual databases] [I.2.7 Natural Language Processing]; Text analysis

General Terms: Learning Systems, Text Patterns, Poetry Databases

Keywords: Computer, Composing styles of Shi and Ci, Couplets antithesis

Received: 18 October 2024, Revised 22 December 2023, Accepted 30 December 2023

Review Metrics: Review Scale: 0/6; Review Score: 4.8; Inter-reviewer Consistency: 92.3%

DOI: https://10.6025/jdim/2024/22/1/1-7

1. Introduction

Chinese *Shi* and *Ci* poetry runs through all the time, becoming an influential constituent of traditional Chinese culture. The *Shi* and *Ci* culture in China possesses unique styles and a sense of beauty, especially in expressing emotion thoroughly. However, even modern society requires professionals who have fundamentally researched this cultural system to identify styles of *Shi* and *Ci* poetry. Currently, there is not yet a universally applicable

approach in this field. However, with the development of modern computer technology, there is a new research direction concerning the innovation of universal identification approaches to styles of *Shi* and *Ci* culture, scilicet applying computer data analysis technology to mine characteristics of *Shi* and *Ci* and identify correlative styles. This research direction has made certain breakthroughs in modern research progress. Still, it's not suitable for universal application for the time being, which won't influence this paper's research on it.

This paper mainly analyzes couplets in the *Shi* and *Ci* cultures, which were passed from ancient times and have already been incorporated into the everyday life of Chinese people, becoming a part of traditional customs. Consequently, couplets highly represent the cultural system of *Shi* and *Ci* poetry. This paper researches couplets by constructing an intelligent system based on computer-aided creation of *Shi*, *Ci*, *Qu* and couplets and studying couplets such as "a couple of orioles tweeting on the green willows, a line of egrets flying in the blue sky". The research identifies the object styles of *Shi* and *Ci* through the intelligent system. It evaluates artistic composing levels of *Shi*, *Ci*, *Qu* and couplets based on the aesthetic sense of literature, on which basis the validity of the system raised by this paper can be affirmed. In addition, to avoid shoddy couplets, *Shi* and *Ci*, such as simple piling up of words and mechanical enumeration of expressions, the intelligent system established by this paper will be equipped with the ability to compose *Shi* and *Ci* poetry and the attribute of identifying distinct styles of different writers. On that basis, shoddy composition can be prevented. To further verify the comprehensiveness of identification, the system's ability to identify *Shi* and *Ci* poetry can be completely illustrated through the intelligent system's judgment of *Shi* and *Ci* poetry composed by computers, which, from the perspective of future development, is advantageous for the literature and aesthetic values of poetry composed by computers.

2. Approaches

2.1. Definition Of Shi and Ci Poetry's Style

Essentially, there is no fixed definition of the style of *Shi* and *Ci* poetry, which is the definition and degree of emotion. Theoretically, the style of *Shi* and *Ci* poetry refers to the experience and feelings of readers when or after they read poetry. Thus, if the definition of the style of *Shi* and *Ci* poetry is necessary, it can be defined as the features readers feel. Currently, the styles of *Shi* and *Ci* poetry in China are mainly divided into two categories, namely the bold and unconstrained style that mainly aims to demonstrate masculine beauty and vent fevered and surging feelings, and the graceful and restrained style that is prone to the feminine beauty and gently voice the mild, vague and exquisite emotion. Further analysis is listed below to comprehend the two schools' styles fully.

The bold and unconstrained style. This style mainly demonstrates masculine beauty with the selection and combination of words that embody the sense of great momenta, such as the sentence that expresses grand, stirring and solemn emotion: the world abounds with magnificence and grandeur, inviting numerous heroes to their mighty feats. In addition, poetry in a bold and unconstrained style usually describes the landscapes of China, mainly including the grandiose territory, vast universe, and extensive land, to make readers feel the sublime mightiness.

The graceful and restrained style. This style is inclined to be feminine. Common poetry of this style usually depicts the sentiment of missing someone, birds' twittering, the fragrance of flowers and a sense of sorrow and melancholy; on the other hand, writers must select and combine words that can better embody the grief of parting, etc. The whole style is consistent but also delicate and dainty, able to induce soft emotion in readers.

The above analysis mainly involves the two major *Shi* and *Ci* poetry styles. However, it's hard to define it qualitatively due to the ambiguity. The formation of a certain style mainly results from a writer's memories, feelings, experiences and thoughts and includes the comprehensive characteristics of a writer. Consequently, the explanation and identification of a style shall not be limited to a single perspective. In addition, a relatively representative statement on the explanation of the style of *Shi* and *Ci* poetry comes from The Literary Mind and the Carving of Dragons: people may have ordinary or eminent talent, masculine or feminine temperament, shallow or profound knowledge and elegant or erratic habits, which are all decided by people's disposition and developed by a posteriori improvement, which, consequently, generates their diverse and various composition styles that the drastic change of the clouds and waves can't even match.

2.2. Construction of Shi and Ci Poetry Database

Poetry indexes are normally reflected as digital bibliographic indexes. [11]. The poetry database consists of structural units, grammatical features and sound devices, i.e. rhyme patterns, rhyme pairs, rhythm, alliterations and the main phonological features of words. [12] As the inheritance of traditional Chinese culture and exquisite artistic works, *Shi*, *Ci*, Qu and couplets are all facing the issue of rearranging and information mining in the information era. From the modern perspective, it's necessary to apply information processing tools to conduct deep analysis and intelligent simulation, especially when issues such as structuring storage of mass data and automatic processing of machines have been solved with the development of corpus technology and machine learning technology. Under the machine learning structure, there are initial conditions to explore and analyze problems of traditional *Shi* and *Ci* problems. Comparatively speaking, traditional methods of analyzing literature can be refreshed from the perspective of analyzing ideas and angles with the assistance of information technology. This makes the mass and complex analysis and study of classical literature more understandable, and its processing is more automatic.

Based on this paper's research requirements, the construction of the database of *Shi* and *Ci* poetry is divided into two parts: the structuring of *Shi* and *Ci* poetry from the Tang and Song dynasties and the establishment of the couplet corpus, which are expounded below in sequence.

Tang and Song dynasties were when the culture of *Shi* and *Ci* prevailed, making it relatively representative of the evolution of Chinese *Shi* and *Ci*. Consequently, this paper structures the database mainly based on *Shi* and *Ci* poetry from the two dynasties. This paper uses the Complete Collection of *Shi* Poetry from the Tang Dynasty and *Ci* Poetry from the Song Dynasty as the data source. On this basis, the data construction is conducted. The construction steps: firstly, this paper constructs the data structures of the title database of *Shi* and *Ci*, including the number of poetry (*poem_id*); titles (title); authors (author); sub-titles (*sub_title*); styles of *Shi* (style). Then, this paper constructs the data structure of the author database, including the names of authors (name); alternative names of authors (*sub_name*); eras of their birth and death (*birth_date0*); the number of their works (*poem_number*); the number of *Shi* poetry (*shi_number*); the number of *Ci* poetry (*ci_number*). In the end, this paper constructs data structures of the poetry database, including the number of the overall lines in a poem (*serial_number*), the number of poetry (*poem_id*), and the marking of sentences (*s_number*).

The couplet is a cultural convention of China that has passed from ancient eras and has become a noteworthy custom in this country. When constructing the couplet corpus, this paper mainly collects couplets according to Complete Works of Couplets and other correlative sources, after which it digitalizes couplets by scanning and identifying technology, ultimately acquiring 9,002 pairs of couplets. Concerning the structure of the couplet corpus, it is divided into the left roll structure and the right roll structure, which are manifested as follows: the corpus structure of the left roll includes the number of the couplet (*couplet_id*); the title (title); the author (author); the year (year); the left roll (*up_couplet*); the pronunciation tones (*pingze*); word segment results of the left roll (*up_segment*); note (note). The corpus structure of the right roll includes the number of the couplet (*couplet_id*); the number of the right roll (*up_couplet_id*); the title (title); the author (author); the year (year); the right roll (*down_couplet*); the pronunciation tone (*pingze*); word segment results of the right roll (*down_segment*); note (note).

2.3. Definition of Shi and Ci Poetry's Style

The database system of *Shi* and *Ci* constructed above shall be equipped with the function of text identification to identify the styles of *Shi* and *Ci*. The paper defines the identification function of text modes as a mapping process that can map unmarked and uncategorized texts into existing categories under the classification system based on the texts' characteristics. The mapping form is one-to-one mapping in the text identification function of this paper and might be one-to-multiple mapping in other research as well, both of which have their characteristics. This paper adopts the prior one mainly due to its high clarity, which is conducive to the research. The mathematical formula of one-to-one mapping is: $j : A \to b$. In the formula; $A = (D_1, D_2, D_n)\ B = (C_1, C_2, C_n)$; A , which can be infinite, is the set of all the uncategorized texts and B, which must be finite, is the set of all categories under the given classification system.

2.4. The Flowchart of Identifying Styles of Shi and Ci Poetry

In the system constructed by this paper, the flowchart of identifying styles of *Shi* and *Ci* poetry is shown in Figure 1:

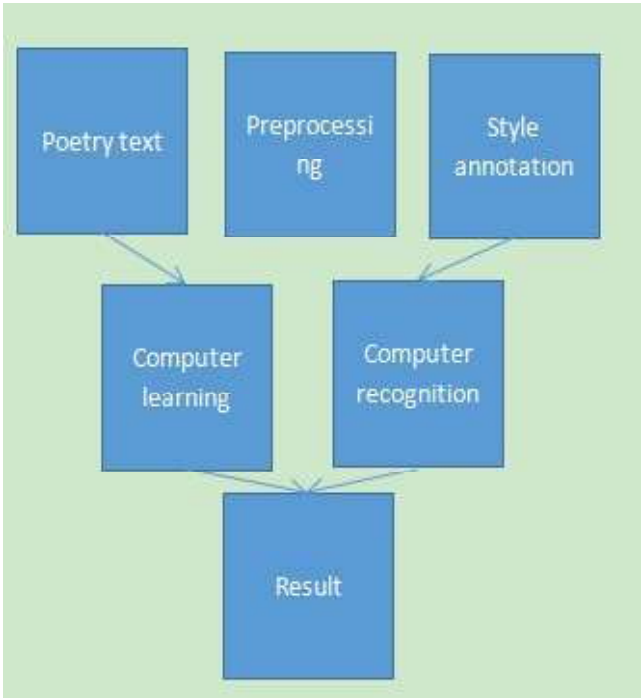

Figure 1. Flow chart of style pattern recognition

2.5. Identification Methods of Text Identification

This paper mainly adopts the support vector machine algorithm, which bases its theory on statistics to identify small-scale samples and nonlinear and high-dimensional patterns effectively. When the system in this paper applies the support vector machine algorithm for identification, the primary identification module is shown below:

As is manifested in Figure 2, A and B are defined as two training samples in the system. At the same time, H serves as the classification line that correctly separates A from B. H1 acts as the dot in one of the samples that are closest to the classification line, while H2 acts as the line parallel to the classification line; the distance between H1 and H2 is called the classification interval of the two samples. Consequently, the so-called optimal classification line is the line that correctly divides the two samples

to the largest classification interval. When applying the same theory to high-dimensional space, the optimal classification line will be the optimal interface.

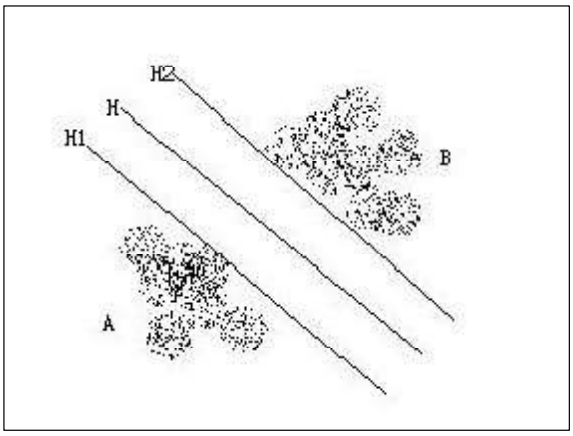

Figure 2. Support vector machine algorithm recognition model

3. Results and Analysis

3.1. Overview of Identification Experiment of Shi and Ci Authors Based on Word Appearance

The system in this paper identifies and judges the styles of all 866 poems of *Li Po* from the Complete Collection of *Shi* Poetry from the Tang Dynasty. During this process, this paper is programmed using advanced programming languages. This auxiliary experiment system has functions that include sequential storage rate, extraction of Chinese characters, and compiling word frequency module statistics. The paper marked the 866 poems of *Li Po* as 1b and identified them with classification modules such as Naïve Bayes.

3.2. Results of Identification Experiment of Shi and Ci Authors Based on Word Appearance

This paper first established the Naïve Bayes learning and identification module applied for identifying styles of *Shi* and *Ci* poetry (Shown in Table 1), then adopted the comparative approach to construct, based on the hill-climbing strategy of the heuristic search algorithm, a correlatively simple character subset algorithm (Shown in Table 1) that defined all the Chinese characters collected by this paper from *Li Po's* poems as *Shi* and *Ci* characteristics of system learning. Each Chinese character from the Chinese character set that hadn't been entered was tested in sequence. At this time, provided that the classification accuracy is improved, the system will enter the tested and unmatched Chinese characters into the database to improve the identification precision of the system classification. In circumstances where the system test results show that all Chinese characters have been entered and identification precision has reached its ceiling, the system learning will begin. In the end, this paper adopted the information gain approach to evaluate the relevance between the characteristics of each Chinese character and the styles of poetry writers. To offer more convenience to the research, this paper selected the 100 most relevant Chinese characters as the characteristics that were later identified by the system again, and the results showed the precision of the system to identify linguistic data from poetry authors has reached 98.3%, a 20% increase compared to prior results, which is highly satisfying.

The below text results prove that the Bayes module that adopts optimal information gain to choose characteristics can ensure identification accuracy after authors of *Shi* and *Ci* poetry are entered into and learned by the system, which demonstrates this module's high application value. Table 2 is the proportion comparison of approaches to identify poets under the ROC curve.

3.3. Analysis of Results of Identifying Characteristics of Shi and Ci Styles

The identification objects are the bold, unconstrained, graceful, and restrained styles, while the main identification method is the level clustering method. This paper applies the level clustering method to analyze training

Method serial number	Discriminant model	Model accuracy	Characteristic number
1	Naïve Bayes Model	0.735	4531
2	KNN Model	0.511	4531
3	Improvement of Bayes model by mountain climbing method	0.956	14 ~ 18
4	Bayes model forfeature selection using optimal information gain	0.983	100

Table 1. System Recognition Model and Algorithm

Method serial number	Discriminant model	Encircling area under ROC curve	Characteristic number
1	Naïve Bayes Model	0.760858	4531
2	KNN Model	0.7530137	4531
3	Improvement of Bayesmodel by mountain climbing method	0.963	14 ~ 18
4	Bayes model forfeature selection using optimal information gain	0.976	100

Table 2. Comparison of the Area Under the ROC Curve of the Discriminant Method of Poets

data of Shi and Ci texts, shown by vector space arithmetic. The comparison results of the two styles are shown in Table 3.

Based on the results above, this paper discovered the co-occurrence relationship among different words in *Shi* and *Ci*, and further concludes the characteristics of some characters in *Shi* and *Ci* of different styles:

This paper discovered that some closely related characters usually appear in the form of twocharacter combinations, such as *fu rong* (hibiscus), *qiao cui* (withered), fang cao (green grass), *ying xiong* (hero), *Huang hun* (dusk), *pi pa* (a four-stringed Chinese lute) and *ou lu* (gull and heron), which are also frequently combined and used in modern Chinese to the extent that they sometimes are referred to as words instead of character combinations. These characters often co-appear in *Shi* and *Ci* poetry.

The characters *gu* (bone) and ji (muscle) also frequently appear in sentences, and in most cases, right next to each other, which indicates it's of high certainty that the two characters co-appear in poetry of the graceful land restrained style.

Judging from the above results, *Ci* of the bold and unconstrained style usually contains content about wars and includes words, such as heroes of magnificent feats, bronze pipa and freezing iron plates, which can stimulate the mind-heaving feeling about the vague sense of belonging, demonstrating an open and extensive state of authors. On the other hand, *Ci* 's graceful and restrained style usually delivers temperament with words such as being pure and noble and being similar to green grass and *fu rong*, showing its meandering artistic conception. Thus, the system of this paper can effectively identify sentences from *Shi* and *Ci*'s poetry, obtaining satisfying results of style identification and comparison, further proving the validity of this system to identify poetry.

3.4. Results of Analyzing Couplets Antithesis

The origin of the couplet is related to Shi and Ci poetry to some extent because, from the perspective of its development history, it derived from symmetry sentences in Shi and Ci poetry. Analyzing couplets requires an understanding of correlative characteristics. This paper summarized the couplet's characteristics by referring to the literature on couplets, consequently obtaining the following results.

Style classification	Comparison of words with significantcontribution to hierarchical clustering
Bold and unrestrained	"Feather, arrow", "river, mountain", "carving,bow", "pipa, Pipa", "Ying, Xiong", "Ho," and so on.
Euphemism	"Bone, muscle", "Rong, Fu", "gaunt, haggard", "yellow and dusk", "Fang, grass", "flower, spring", and so on.

Table 3. Obvious Relationship Between Bold and Graceful Poetry and Comparative Poetry

The parts of speech of linguistic units from the left and the right rolls correspond, with scilicet characters in the same positions from the given sentence and the pairing sentence of identical parts of speech.

Words from the given and the pairing sentences are of identical semantic categories.

The couplet emphasizes the beauty of symmetry. Thus, once a couplet is made, the traits of words contained, such as the number of characters and parts of speech, must be symmetrical.

The characters at the beginning and end of each sentence must rhyme with their pairing character.

There are seldom conjunctions or modal particles. The semantic meanings of pairing sentences are identical.

In the process of the system identification of couplets, this paper discovered that some couplets lack clear subject-predicate-object structures due to the lack of conjunctions and modal particles, which are substituted by other characters or words, such as the expression "world with vitality, spring with feet". Consequently, in applying the system to conduct learning and identification, it's infeasible to adopt the grammar and other literature systems identical to *Shi* and *Ci* poetry's system; instead, new language processing approaches shall be applied.

To identify couplets, this paper mainly adopts the couplet corpus, including classic couplets of generations and existing Spring Festival couplets, which, after being digitalized, helped obtain the following results:

"A gad on the chest with millions of soldiers, a heart for the country of thousands of trees. Cock singing reminds warring in the south; horse neighing recalls fighting in the north. Green pines reach the blue sky in vitality, and withered leaves fall into the green river in lethargy. The wind blows the clouds, but the stars, and the water pushes the boats but the banks. Apricot blossoms fall in the drizzle; weeping willows swing in the breeze. Rays of smoke doddering in the green grass ferry, threads of rain falling in the apricot flower village. Slanting shadows of branches crossing in the shallow water, faint fragrance of flowers floating under the dizzy moon." The identification results are shown in Table 4.

Connection	In the east of the river, the wave of the ancientheroes
Generation of system recognition after system recognition increased	In the spring of the river, the mountains and rivers have in a hundred years
Original association	Xiao Yuan springs back; the warbler evokes a beautiful courtyard.

Table 4. Identifies the Results of a Linked Instance

Judging from Table 4, after identifying the original left roll, "The river runs to the east, washing away the feats and grandeur of generations of heroes", the system automatically created the right roll that's different from the left roll. Currently, the indexes adopted by the system to determine the validity of the right roll are from the characteristic system of couplets, including parts of speech and the number of characters, et cetera. In this manner, the characters of the right roll are identical to those of the left roll, and the result demonstrates good performances of the system concerning segmentation, parts of speech, sentence fluency, rhyme and semantic fluency, which further proves the validity of the system of this paper.

4. Conclusion

This paper mainly analyzes the composing styles of *Shi* and *Ci* poetry and couplet antithesis based on the assistance of computers. It first defines the style of *Shi* and *Ci* and, based on readers' feelings about *Shi* and Ci poetry, divides it into the bold and unconstrained style and the graceful and restrained style, which are briefly analyzed by this article. Then, this paper constructs an identification system with a database of *Shi* and *Ci* poetry and the function of identifying text modes, through which this paper identifies examples from the Complete Collection of *Shi* Poetry from the Tang Dynasty and the Complete Collection of *Ci* Poetry from the Song Dynasty with the support vector machine algorithm. Following this, this paper adopts the Naïve Bayes module to analyze identification results, showing the accuracy of the system's identification of authors based on the corpus has reached 98.3%, which is a satisfying result. In the end, the system raised by this paper identifies couplets with results showing the system's ability to create paring sentences that completely correspond to the given sentence, further proving this system's validity.

References

[1] Zhao, H., Ning, J., Mao, S. Y., Xu, W. F. (2014). Design and Implementation of an Automatic Hydrological Telemetry System. *Applied Mechanics & Materials*, 511-512, 752-756.

[2] Chen, L. H., Liao, F. Q. (2014). The Development and Application of Small Watershed Hydrological Telemetry System. *Advanced Materials Research*, 834-836 (4), 947-953.

[3] Fassnacht, S. R., Deitemeyer, D. C., Venable, N. B. H. (2014). Capitalizing on the daily time step of snow telemetry data to model the snowmelt components of the hydrograph for small watersheds. *Hydrological Processes*, 28 (16), 4654-4668.

[4] Chen, H. T., Wu, C. C. (2014). Design and Implementation of the Intelligent Mobile Phone Platform for English Learning. *Advanced Materials Research*, 926-930, 4485-4488.

[5] Lee, E. W., Jin, G. S., Yong, K. (2015). Design and Implementation of English Learning Application for Early Childhood. *Indian Journal of Science & Technology*, 8 (S7), 679.

[6] Liu, C., Liu, Z. (2016). A Creative Design and Implementation of Student-led Flipped Classroom Model in English Learning. *Theory & Practice in Language Studies*, 6 (10), 2036.

[7] Afrilyasanti, R., Cahyono, B. Y., Astuti, U. P. (2017). Indonesian EFL Students' Perceptions on the Implementation of Flipped Classroom Model Rida Afrilyasanti Sekolah Menengah Atas (Senior High School) Negeri 8 at Malang, Indonesia. *Journal of Language Teaching & Research*, 8 (3), 476-484.

[8] Trung, H. D., Hung, P. T., Khanh, N. D., Dung, H. V. (2013). Design and implementation of mobile vehicle monitoring system based on android smartphone. *Information & Communication Technologies*, 51-56.

[9] Jusik, S., Szoszkiewicz, K., Kupiec, J. M., Lewin, I., Samecka-Cymerman, A. (2015). Development of comprehensive river typology based on macrophytes in the mountain-lowland gradient of different Central European ecoregions. *Hydrobiologia*, 745 (1), 241-262.

[10] Stenger-Kovács, C., Tóth, L., Tóth, F., Hajnal, É., Padisák, J. (2014). Stream order-dependent diversity metrics of epilithic diatom assemblages. *Hydrobiologia*, 721 (1), 67-75.

[11] Houston, N. M. (2019). Modelling the Poem on the Page: Encoding the Database Schema for the Periodical Poetry Index. *Victorian Periodicals Review*, 52 (3), 626-635.

[12] Timári, G. P. (2022). ELTE Poetry Corpus: A Machine Annotated Database of Canonical Hungarian Poetry. *In Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3471–3478.