J Inf Process Syst, Vol.19, No.4, pp.474~482, August 2023
https://doi.org/10.3745/JIPS.04.0281

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Data Mining Research on Maehwado Painting Poetry in the Early Joseon Dynasty

Haeyoung Park[1] and Younghoon An[2,*]

Abstract

Data mining is a technique for extracting valuable information from vast amounts of data by analyzing statistical and mathematical operations, rules, and relationships. In this study, we employed data mining technology to analyze the data concerning the painting poetry of Maehwado (plum blossom paintings) from the early Joseon Dynasty. The data was extracted from the Hanguk Munjip Chonggan (Korean Literary Collections in Classical Chinese) in the Hanguk Gojeon Jonghap database (Korea Classics DB). Using computer information processing techniques, we carried out web scraping and classification of the painting poetry from the Hanguk Munjip Chonggan. Subsequently, we narrowed down our focus to the painting poetry specifically related to Maehwado in the early Joseon Dynasty. Based on this, refined dataset, we conducted an in-depth analysis and interpretation of the text data at the syllable corpus level. As a result, we found a direct correlation between the corpus statistics for each syllable in Maehwado painting poetry and the symbolic meaning of plum blossoms.

Keywords

Corpus, Data Mining, Maehwado, Painting Poetry, Plum Blossoms, Symbolic Meaning

1. Introduction

This study analyzed the painting poetry data of Maehwado (plum blossom paintings) from the early Joseon Dynasty extracted from the Hanguk Munjip Chonggan (Korean Literary Collections in Classical Chinese) in the Hanguk Gojeon Jonghap database (Korea Classics Database) applying computer information processing technology. The purpose is to objectively clarify the expression methods and meanings of Hansi (poems in the Chinese style) by converting them into quantitative values based on a computer program.

It is one of the primary studies on Hansi that categorizes poems by their structure. Traditional studies on Hansi have emphasized qualitative evaluation, relying primarily on the researcher's judgment. Analyzing Hansi using data mining is a more objective, and quantitative method of evaluation. While existing digital humanities research has predominantly focused on the narrative genre [1,2], research on the poetry genre within digital humanities is exceedingly rare. Lee [3,4] is the sole scholar who extensively studied Hansi using computer information processing, but there is a limitation in terms of

Manuscript received February 2, 2023; first revision April 3, 2023; accepted April 25, 2023.

* **Corresponding Author:** Younghoon An (yhnahn@khu.ac.kr)
[1] Humanitas College, Kyung Hee University, Seoul, Korea (hy000p@khu.ac.kr)
[2] Dept. of Korean Language and Literature, Kyung Hee University, Seoul, Korea (yhnahn@khu.ac.kr)

scope. Therefore, we aim to classify various types of poetry and provide a more detailed interpretation of the language employed in the poems. This approach can potentially be expanded to encompass the entirety of Korean classical literature, facilitating deeper and more extensive research within each genre.

2. Research Methodology

2.1 Research Materials

The central focus of this study revolves around painting poetry. Painting poetry [5] constitutes a form of literary expression in which a poet encapsulates emotions or reflections concerning paintings within a condensed poetic structure. Painting poetry serves as a conduit for translating visual symbols, such as "paintings," into the realm of social symbolic tools, namely "language." In essence, it signifies the symbol itself, thereby culminating in the language intrinsic to painting poetry. Consequently, painting poetry facilitates our ability to apprehend and decipher the symbolic significance underlying the elements portrayed through artworks.

The principal source material underpinning the exploration of painting poetry emanates from the Hanguk Munjip Chonggan, housed within the Database of Hanguk Gojeon Beonyeokwon (Institute for the Translation of Korean Classics), colloquially referred to as the "Hanguk Gojeon Jonghap DB." The Hanguk Gojeon Beonyeokwon spearheads a comprehensive initiative aimed at translating, digitizing, and rendering Korean classics accessible to the general populace. This data is structured in XML format, supplemented by an open application programming interface (API). The dataset is meticulously categorized by attributes such as title, authorship, stylistic composition, publication year, and more. Moreover, it boasts the capability to selectively extract pertinent information by configuring search criteria through designated keywords. Building upon this framework, the painting poetry data from the early Joseon Dynasty has been meticulously extracted and subjected to analysis.

2.2 Data Mining Method

Data mining is a technique used to extract valuable information from large datasets by analyzing statistical and mathematical operations, rules, and relationships [6]. Today, this technology is employed across various major fields such as computing [7], business administration [8], and statistics [9,10]. We applied this technology to the realm of Korean classical literature studies. Initially, we selected painting poetry from the Hanguk Gojeon Jonghap DB and classified it by type. This marks the first instance of data mining applied to painting poetry. Subsequently, we narrowed our focus to the painting poetry of Maehwado in the early Joseon Dynasty. We then proceeded to analyze and interpret the textual data of Maehwado's painting poetry from the early Joseon Dynasty. This signifies the secondary data mining of phase painting poetry.

3. Result

3.1 First Data Mining: Scraping and Type Classification

In the first data mining phase, we initiated the process by sorting out Hansi from the Hanguk Munjip Chonggan within the Hanguk Gojeon Jonghap DB. We meticulously extracted and categorized poetry with specific keywords such as “hwa” (paint or painting), “do” (picture), “sa” (draw), “hoe” (draw), “cheop” (album), “chuk” (scroll), “jok” (hanging scroll), “byeong” (folding screen), and “muk” (ink) found in their titles (Fig. 1). Subsequently, we verified the titles and content of the selected poetry to ascertain if they were indeed inspired by paintings. This process constitutes the scraping method.

Originally, web scraping is a computer program that automatically extracts data from web pages (Fig. 2). However, the Hanguk Gojeon Jonghap DB (https://db.itkc.or.kr) already provides data in an XML format, categorized by item. Hence, researchers can extract painting poetry data using the search function as described above. We verified that there are 842 painting poetry pieces from the early Joseon Dynasty available on the Hanguk Munjip Chonggan [11].

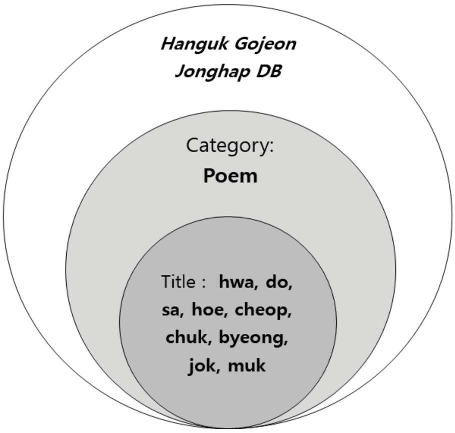


Fig. 1. Extracting of painting poetry of the early Joseon Dynasty.

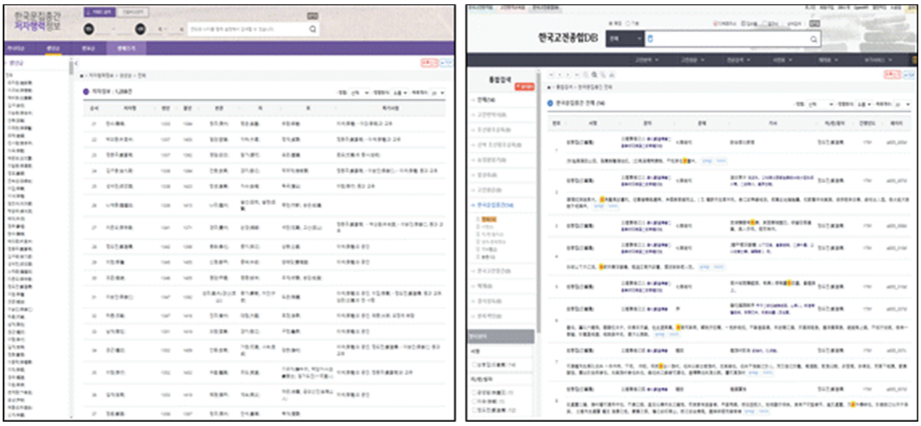


Fig. 2. Scraping of painting poetry in the early Joseon Dynasty on the Hanguk Gojeon Jonghap database.

A type-classification of painting poetry is based on the material of the paintings. This categorization is possible because the material of a painting is typically denoted by a distinct name, preceded and followed by terms such as “hwa” and “do,” which specifically refer to the painting. For instance, many titles of Maehwado painting poetry include expressions like “Hwamae (drawing a plum blossom)” or “MukMae (Inked a plum blossom).” The types of painting poetry can be further divided into higher concepts and subitems based on the material of the paintings. To illustrate, we categorized the painting poetry of

Maehwado from the early Joseon Dynasty by grouping it under higher concepts, such as “Flower painting poetry,” and “Birds Beasts Plants Trees painting poetry.” Within the category of painting poetry, we narrowed our focus to the subject of research, specifically “Birds Beasts Plants Trees → Flower → Maehwa (plum blossom).” Among the painting poetry in the early Joseon Dynasty, there are a total of 97 pieces classified as Maehwado types. Fig. 3 illustrates the result of categorizing Maehwado painting poetry from the early Joseon Dynasty.

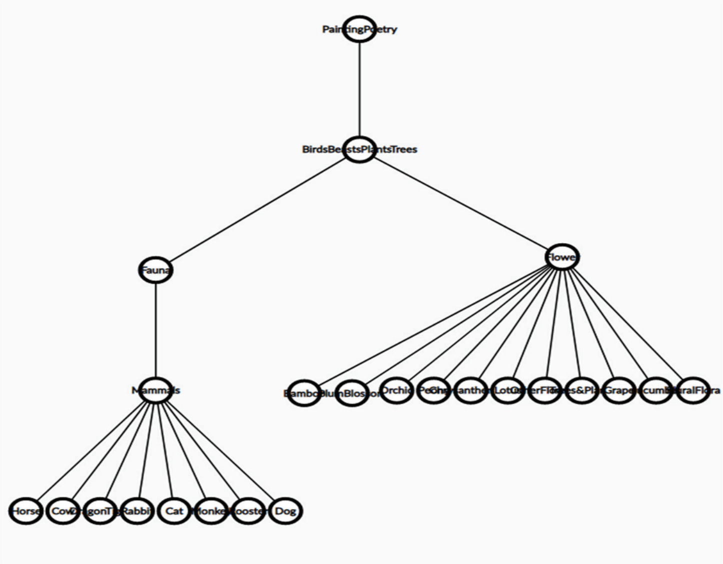


Fig. 3. A type-classification of Maehwado painting poetry.

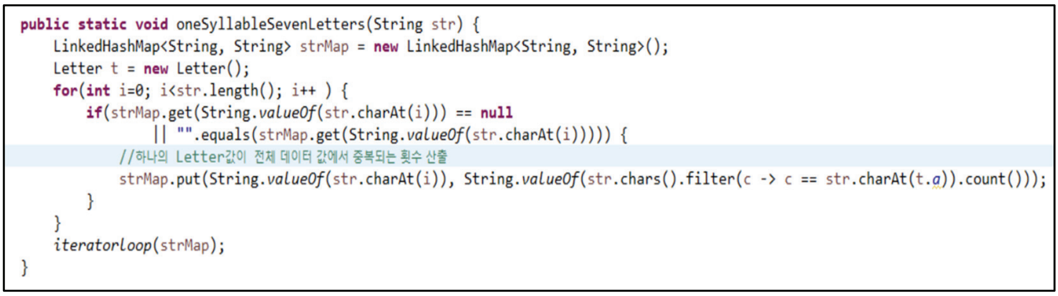
3.2 Second Data Mining: Statistics and Analysis of Syllable Corpus

Afterward, we proceeded to analyze the text data of Maehwado painting poetry from the early Joseon Dynasty, which served as the focal point of our study. The majority of Maehwado painting poetry in the early Joseon Dynasty conforms to a classical chinese verse structure, taking the form of Jeol-gu (four lines) and Bae-ryul (eight lines). Consequently, there exist four types of Hansi based on the number of words: five-word Jeol-gu, seven-word Jeol-gu, five-word Bae-ryul, and seven-word Bae-ryul. Each verse comprises a specific number of characters, forming a corpus consisting of 20 (5×4), 28 (7×4), 40 (5×8), or 56 (7×8) letters, constituting a complete poem.

To interpret a poem [12] effectively, it is essential to break it down into morphemes, which represent the smallest units of meaning. Since Hansi consists of ideograms, morphemes can be further disassembled into single-syllable units, referred to as corpus. Hansi essentially consists of five or seven words per line, and a line represents the fundamental unit of Hansi. In the text of painting poetry, o’eon (five-word) is divided into five types of corpus, ranging from one to five syllables , and chil’eon (seven-word) is divided into seven types of corpus, ranging from one to seven syllables. In essence, for each line, o’eon generates five corpus and chil’eon generates seven corpus.

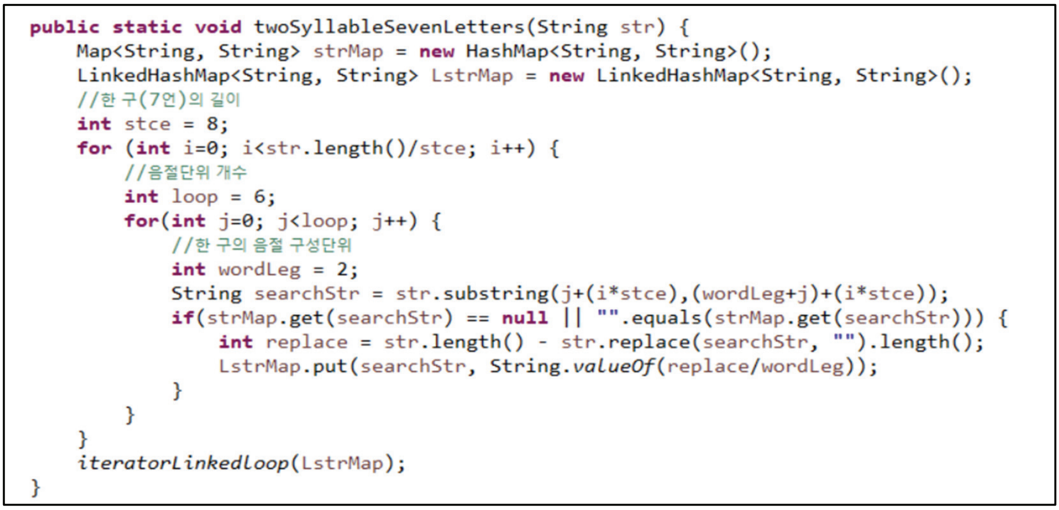
At this juncture, when dealing with corpora containing more than two syllables, the order in which they are combined becomes crucial. During the interpretation of poems, the usually approach is to sequentially interpret from the front to the back, but occasionally, due to the structure of the poem, interpretation may occur from back to front. In some instances, there might be instances of empty words within the poem, wherein these words do not contribute to the overall meaning. While translating, it is essential to take note of issues like postpositional modifications or the insertion of empty. However, such factors do not significantly impact the process of creating a corpus by segmenting the lines of a poem into syllable units. Since the corpus has already been divided into one-syllable segments, counting the frequency of letters with distinct semantic values remains unaffected even if the order of combination differs. In essence, the repeated word combinations found in the upper segment unit have already been segmented into one semantic value in the lower segment unit. If there are instances where the meaning varies based on the order, these should be treated as individual words, and their frequency should be measured separately.

In the early Joseon Dynasty, the total number of Maehwado painting poetry amounts to 97, comprising a total of 2,686 Chinese characters. These characters are organized in corpus units. Figs. 4 and 5 are a set of Java program instructions designed to calculate the frequency by segmenting the original painting poetry data based on syllables. A function diagram has been implemented to determine the number of repeated corpus units for each syllable within the original painting poetry text.



```
public static void oneSyllableSevenLetters(String str) {
    LinkedHashMap<String, String> strMap = new LinkedHashMap<String, String>();
    Letter t = new Letter();
    for(int i=0; i<str.length(); i++ ) {
        if(strMap.get(String.valueOf(str.charAt(i))) == null
                || "".equals(strMap.get(String.valueOf(str.charAt(i))))) {
            //하나의 Letter값이 전체 데이터 값에서 중복되는 횟수 산출
            strMap.put(String.valueOf(str.charAt(i)), String.valueOf(str.chars().filter(c -> c == str.charAt(t.a)).count()));
        }
    }
    iteratorloop(strMap);
}
```

Fig. 4. Chil'eon 1-syllable (7) [1/2/3/4/5/6/7] corpus unit analysis function.



```
public static void twoSyllableSevenLetters(String str) {
    Map<String, String> strMap = new HashMap<String, String>();
    LinkedHashMap<String, String> LstrMap = new LinkedHashMap<String, String>();
    //한 구(7언)의 길이
    int stce = 8;
    for (int i=0; i<str.length()/stce; i++) {
        //음절단위 개수
        int loop = 6;
        for(int j=0; j<loop; j++) {
            //한 구의 음절 구성단위
            int wordLeg = 2;
            String searchStr = str.substring(j+(i*stce),(wordLeg+j)+(i*stce));
            if(strMap.get(searchStr) == null || "".equals(strMap.get(searchStr))) {
                int replace = str.length() - str.replace(searchStr, "").length();
                LstrMap.put(searchStr, String.valueOf(replace/wordLeg));
            }
        }
    }
    iteratorLinkedLoop(LstrMap);
}
```

Fig. 5. Chil'eon 2-syllable (6) [12/23/34/45/56/67] corpus unit analysis function .

Figs. 6 and 7 present results obtained by dividing the text into syllables and calculating their frequency, using a Java program. The measurements aimed to determine the number of corpus units in the original painting poetry text. The 1–5-syllable corpus data represent the combined data for both o'eon and chil'eon painting poetry, while the 6–7-syllable corpus data solely comprise chil'eon painting poetry. Notably, only the 1–3-syllable corpus data was considered valid, as all 4–7-syllable corpus data contained unique word combinations that appeared only once, rendering them meaningless for analysis. Moreover, the 3-syllable corpus data was excluded from semantic analysis, due to infrequent occurrence (less than one to three times), resulting in limited meaning for the combination of 3-syllable units. Consequently, only the word usage frequency data from 1-syllable and 2-syllable corpus units is presented.

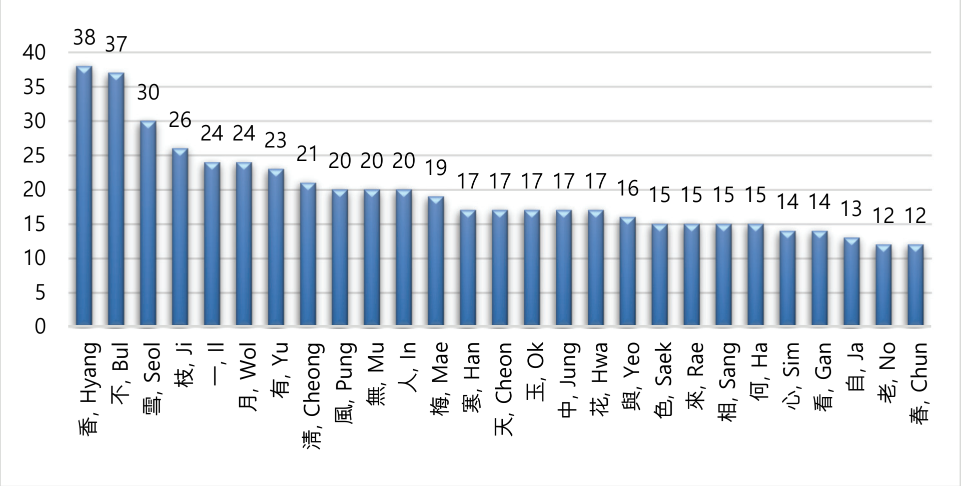


Fig. 6. One-syllable corpus statistics of Maehwado painting poetry.

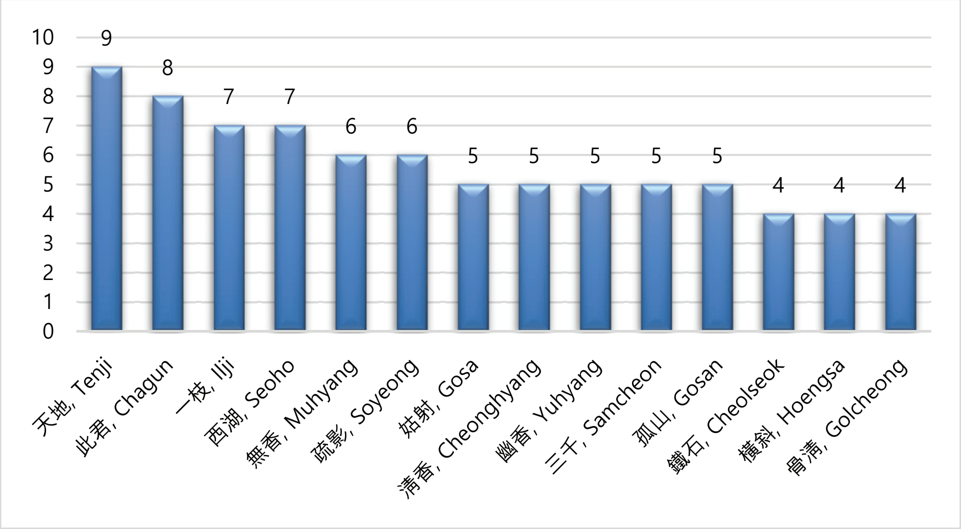


Fig. 7. Two-syllable corpus statistics of Maehwado painting poetry.

First, a significant data value in the corpus statistics for Maehwado painting poetry lies in the 2-syllable corpus. Notably, certain words frequently appear in the 2-syllable corpus, including "Cha-Gun" (this gunja: 8 times), "Seo-Ho" (West Lake: 7 times), "Go-San" (Mt. Gushan: 5 times), "Go-Ya" (Miaogusheshan:

5 times), “Cheol-Seok” (iron and stone: 4 times). These terms align with the symbolic representation of plum blossoms during the early Joseon Dynasty’s Maehwado. In a previous study of Maehwado painting poetry [13], it was discovered that plum blossoms in the early Joseon Dynasty’s paintings were depicted in the form of “Seonnyeo” (fairy), “Sinryong” (god dragon), and “Gunja” (man of virtue).

Conventionally, the beautiful fairy image utilizes the literary allusions to Miaogusheshan (Mt. Miaogushe) and Luofushan (Mt. Luofu) to enhance the sense of romance. Moreover, it gracefully incorporates poetic descriptions of ice, snow, moon, and twilight to further evoke a clean and clear imagery. The dragon, with its unique form, becomes particularly prominent when depicted as gomae (old plum), accentuating its novelty as the winding branches are likened to the intricate appearance of a dragon. The representation of plum blossoms, symbolizing a man of virtue at the right time, varies depending on the political perspective. In Maehwado, the portrayal of wise ministers like Buyeol takes the form of Choolsa (government service), connecting them with the attributes of plum blossoms and plum fruit. On the other hand, Limpo residing in Gushan (Mt. Gushan), is depicted as Maecheohakja, choosing not to serve the power, a notion expressed through the concept of Cheosa. Through this analysis, it has been confirmed that the meaning of plum blossoms in the painting poetry of the early Joseon Dynasty encompasses themes of “romance and pureness,” “mysteriousness and bizarrity,” and “government service and not serving the power.”

Therefore, “Go-Ya” is used to refer to a seonnyeo, “Cha-Gun” to a Gunja, “Cheol-Seok” to a Choolsa, and “Seo-ho” and “Go-San” to Cheosa. Additionally, when describing plum blossoms, the nouns “Hyang” (scent: 38 times), “Seol” (snow: 30 times), “Wol” (moon: 24 times), and “Pung” (wind: 20 times) are employed, reinforcing the meaning of plum blossoms with adjectives like “Cheong” (clear: 21 times), “Ok” (jade: 17 times), “No” (old: 12 times), “Jeong” (upright: 8 times), and “Hon” (spirit: 8 times). Other words with high frequency include “Ten-Ji” (world: 9 times), “Il-Ji” (a branch of a tree: 7 times), “Il” (one: 24 times), “Mu-Hyang” (unscented: 7 times) and “Bul” (no: 37 times), among others. These words are used to emphasize the existence of a single thing [Ten-Ji, Il-Ji, Il] when expressing plum blossoms or to further underscore their meaning through irony using negatives [Mu-Hyang, Bu]. As a result, we can confirm that the corpus statistics for each syllable in Maehwado painting poetry are closely linked to the symbolic meaning of plum blossoms.

4. Conclusion

The more a symbol is shared, the more impactful its meaning becomes. Through data mining, a computer information processing method, we have confirmed that a series of identical words is repeated in the same type of painting poetry. The meaning implied by this corpus represents the symbol of the painting, which, in turn, symbolizes the essence of the object portrayed in the painting. This process provides us with a convincing means to analyze and derive the symbols of objects more clearly. Data mining proves to be a valuable tool not only for revealing the meaning of Maehwado painting poetry but also for exploring other types of painting poetry. Furthermore, we aim to expand this study to gain a deeper understanding of communication and meaning within literature.

Acknowledgement

This paper is a revised and expanded version of the paper “A Case Study on Data Mining for Korean Painting Poetry,” which was presented at CSA 2022. A study in the same vein of painting poetry data analysis is “A study of Mukjukdo painting poetry using N-gram analysis: focusing on the early Joseon Dynasty's works (*The Studies of Korean Literature*, no. 79, pp. 57-82, 2023).” The data calculation function setting using the Java program was performed with the valuable assistance of Seokjune Kang, an expert in computer science software. We extend our sincere gratitude for his practical help with computer information processing.

References

[1] Y. H. An, C. H. Cha, and D. K. Kim, “The present condition and task on database of Hangeul materials of classical novel,” *The Research of Old Korean Novel*, vol. 42, pp. 47-81, 2016.

[2] K. Kwon, W. Choi, and D. Kim, “A lengthwise comparative study of different versions of Yadam: based on ‘Ok So-seon’,” *The Research of the Korean Classic*, no. 57, pp. 87-120, 2022.

[3] B. C. Lee, “A study of Korean Chinese poetry similarity analysis,” *Journal of Korean Literature in Chinese*, vol. 59, pp. 361-383, 2021.

[4] B. C. Lee, “A study on the construction and utilization of Korean Chinese poetry corpus,” *Journal of Korean Literature in Chinese*, vol. 53, pp. 153-177, 2019. http://doi.org/10.17260/jklc.2019.53..153

[5] H. Park, “A study on the painting poetry of the latter part of the Goryeo Dynasty,” Master's thesis, Kyung Hee University, Seoul, Korea, 2015.

[6] Wikipedia, “Data mining,” c2023 [Online]. Available: ko.wikipedia.org/wiki/Data Mining.

[7] A. G. Finogeev, D. S. Parygin, and A. A. Finogeev, “The convergence computing model for big sensor data mining and knowledge discovery,” *Human-centric Computing and Information Sciences*, vol. 7, article no. 11, 2017. https://doi.org/10.1186/s13673-017-0092-7

[8] K. Paul and C. M. Parra, “Corporate social responsibility in international business literature: results from text data mining of the Journal of International Business Studies,” *International Journal of Corporate Social Responsibility*, vol. 6, article no. 12, 2021. https://doi.org/10.1186/s40991-021-00066-6

[9] P. Vilakone, K. Xinchang, and D. S. Park, “Personalized movie recommendation system combining data mining with the k-clique method,” *Journal of Information Processing Systems*, vol. 15, no. 5, pp. 1141-1155, 2019. https://doi.org/10.3745/JIPS.04.0138

[10] V. E. Sathishkumar, M. Lee, J. Lim, Y. Kim, C. Shin, J. Park, and Y. Cho, “An energy consumption prediction model for smart factory using data mining algorithms,” *KIPS Transactions on Software and Data Engineering*, vol. 9, no. 5, pp. 153-160, 2020. https://doi.org/10.3745/KTSDE.2020.9.5.153

[11] H. Park, “A study on the painting poetry of the former part of the Joseon Dynasty,” Ph.D. dissertation, Kyung Hee University, Seoul, Korea, 2021.

[12] H. Park, “A study of Mukjukdo painting poetry using N-gram analysis: focusing on the early Joseon Dynasty's works,” *The Studies of Korean Literature*, no. 79, pp. 57-82, 2023. http://dx.doi.org/10.20864/skl.2023.7.79.57

[13] H. Park, “A study on the symbolic meaning of painting poetry about plum blossom paintings in the early Joseon Dynasty,” *The Study of the Eastern Classic*, no. 89, pp. 9-41, 2022. http://dx.doi.org/10.23904/SEC.2022.89.01.009

Haeyoung Park **https://orcid.org/0000-0002-0161-8465**

She received her M.S. and Ph.D. degrees in Korean language and literature from Kyung Hee University in 2015 and 2021, respectively. Since March 2018, she has been serving as a lecturer in the Humanitas College at Kyung Hee University, Seoul, Korea. Her current research interests encompass literary symbol communication and digital information processing.

Younghoon An **https://orcid.org/0000-0002-3012-9724**

He received his M.S. and Ph.D. degrees in Korean language and literature from Kyung Hee University in 1993 and 1998, respectively. He is currently a professor in the Department of Korean Language and Literature at Kyung Hee University, Seoul, Korea. His research interests primarily revolve around modeling Korean literature information.