International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp. 13884-13890
© Research India Publications. http://www.ripublication.com

Pairwise Similarity Analysis and Quality Estimation on Classical Chinese Poetry of Ancient Korea in 15th Century

Shohrukh Bekmirzaev[1] and Tae-Hyong Kim[2,*]

[1] Department of Computer Engineering, Kumoh National Institute of Technology,
61 Daehak-ro, Gumi, Gyeongbuk 39177, Republic of Korea.
*(*Corresponding author), Orcid Id: 0000-0003-3806-2517*

Byoung-Chan Lee[3]

[2] Sol Liberal Arts School, Woosong University,
171 Dongdaejeon-ro, Dong-gu, Daejeon 34606, Republic of Korea.

Abstract

Big data analytics can assist studies on Classical literature by enabling extensive and in-depth analysis of massive data. This paper represents a preliminary study on Classical Chinese poetry in the middle period of Joseon with text mining techniques. We tried to evaluate the relations between poems with several similarity analysis methods adjusted for the characteristics of Classical Chinese poetry. The quality of poems was also tried to estimate based on their vocabulary with similar analysis methods. Nine poem books in the 15th century have been selected for analysis to validate the effectiveness of this approach. Analysis results show a high potential of this approach which correspond to the existing intuitive evaluation on the relations among the target poems and poets.

Keywords: Classical Chinese poetry, text mining, relation analysis, quality estimation, document similarity**.**

INTRODUCTION

Current high interest and success stories on big data analytics owes massive amount of digital data generated every day, extensive computing power with clustering technology, competent machine learning algorithms including deep learning, and so on. As, among various kinds of data, text data are still main targets for analytics such as online documents and social messages, various machine learning techniques have been developed to extract valuable information from text data. Historical data are emerging targets on data analytics which has been digitized rapidly. There is huge amount of classical literature written in various languages. Digitization and opening to the public by websites of classical literature can accelerate its studies owing to convenient user interface such as easy access and fast search. This environment may be able to change the styles of studies. Ordinary studies are usually microscopic. For example, they often try to clarify the meanings of specific parts or to read between lines. Computer-based text mining can expedite extensive analyses by handling massive data in detail such as trends and relations analysis.

Classical Chinese poetry is ancient poetry written in Classical Chinese, mainly in China and nearby east Asian cultural sphere using Classical Chinese characters, such as Korea, Japan, and Vietnam [1]. Classical Chinese poetry usually has fixed verse forms such as strict rhyme and pitch rules and is classified according to the number of characters in a line. A line has mostly five or seven characters, sometimes four or six, and a verse has often 4 or 8 lines. A title is usually put on a poem which is constructed from several verses. Each verse in a poem is usually independent of the others and sings the same theme differently. Like other literature, a poem is highly evaluated if it shows creative images and new understandings on the theme. Poets in Classical Chinese poetry also try to use their own unique words, and to avoid repeating same words in their poems.

Major topics on studies of Classical Chinese poetry are interpretation of poems, analysis of poets' views, trends on themes and types of poems in a certain period, and relations among poems and poets. While interpretation of poems and analysis of poets' views require deep understanding those poets and poems, analysis on trends and relations among poems and poets usually require thorough investigation of the target poems and poets, which can be assisted by computer-based text mining.

There have been very limited number of studies on analysis of poetry with text mining approach. Von-Wun Soo *et al.* examined methods of semantic retrieval and automated ontology acquisition from semantically annotated poems based on a Chinese thesaurus [2]. They also defined a scoring scheme to assess semantic similarity for semantic retrieval. David M. Kaplan and D. M. Blei proposed a quantitative method to assess the style of American poems by developing metrics that analyze orthographic, syntactic and phonemic features [3]. They mapped a poem text to a multi-dimensional vector representing its place in stylistic space. Shinji Kikuchi *et al.* presented a method to estimate the artistic quality of Haiku, Japanese style short poem, text using neural networks [4]. They constructed word-based and syllable-based vector models and then Haiku quality estimation function with a convolutional neural network. 'Likes' information given from viewers in a

Haiku community site was used as target data for learning. John Lee and Mengqi Luo explored the phenomenon of parallelism in Classical Chinese poetry with a graph-based clustering method [5]. For clustering to detect parallel lines in a poem, they used similarity scores between two words using frequency statistics and word embedding. No relation or quality analysis on Classical Chinese poems has been made with text mining or machine learning approach yet.

In this paper, we try to evaluate the relations between poems with several document similarity analysis methods, and also to estimate the quality of poems based on styles of using words. This work is a evaluation study on big data analysis of Classical Chinese poetry, especially in the middle period of Joseon, an old Korean dynastic kingdom. In the 15th century in Joseon, there appeared a lot of poets and was a change of characteristics in Classical Chinese poetry, which can be a good target for analysis on trends and relations among poems and poets. There are 1,260 poem books in the 15th century, among which we selected 9 to validate the effectiveness of this study. In order to overcome the limited number of data, we slightly change the existing similarity measure. We also try various methods to derive similarity and quality information by considering the structure of Classical Chinese poetry. The next section shows preliminary work which includes information on the target poem books and their relations and basic notations used in the poetic data models. Then relation analysis and quality estimation methods of Classical Chinese poetry are explained in detail. Finally, the results of this analysis are shown and discussed and conclusion is derived.

PRELIMINARIES

Before representing details of analysis methods, we introduce the target data sets for better understanding the goal, and also notations of data models to be used in later equations.

Target Dataset

In this study, we selected 9 poem books in the middle period of Joseon to validate our text mining-based poetry analysis as shown in Table 1. Those poem books were written by different poets who have some relations each other. All those poem data can be downloaded at the website of Korean classical literature integrated database that has been built by Institute for the translation of Korean classics [6].

Table 1: Target poem books

No	Title	Poet	Volume
1	Cheong- eum-jip	Kim, Sang-Heon (1570~1652)	756 poems, 50,674 words
2	Gok-un-jip	Kim, Su-Jeung (1624~1701)	97 poems, 20,107 words
3	Nong-am -jip	Kim, Chang-Hyeop (1651~1708)	573 poems, 47,986 words
4	Hyeon-ju -jip	Yun, Shin-Ji (1582~1657)	320 poems, 21,860 words
5	Hyeon-ju -jip	Lee, So-Han (1598~1645)	488 poems, 22,815 words
6	Go-mun -jin-bo	Various (13th century)	241 poems, 26,405 words
7	Man-chwi -jip	Oh, Eok-Ryeong (1552~1618)	152 poems, 10,429 words
8	Baek-sa-jip	Lee, Hang-Bok (1556~1618)	251 poems, 12,756 words
9	Eo-u-zip	Yu, Mong-In (1559~1623)	402 poems, 29,362 words

The relations among the poets of the target poem books are depicted in Figure 1. The main relation analysis target are the poets of poem books 1, 2, and 3. Let us denote poet *n* as the writer of poem book *n* for simplicity. While poet 1 is the father of poet 2, and the grandfather of poet 3, poet 2 is an uncle of poet 3. Therefore, there may be possibly some similarity among their poems. Poet 5 may be another hub; the father of poet 5 had some friendship with poets 1, 8, and 9, and poet 5 is also a friend of poet 4. Poets 7, 8, and 9 are contemporary poets but there were little interactions among them. They were randomly chosen as a control group for comparison. Poem book 6 is a collection of poems composed by famous Chinese poets such as Li Bai and Du Fu which was published in the 13th century and introduced to Joseon in the 14th century. As it had been used as a textbook by students studying Classical Chinese poetry, it will be interesting to check the relations between this poem book and the others.

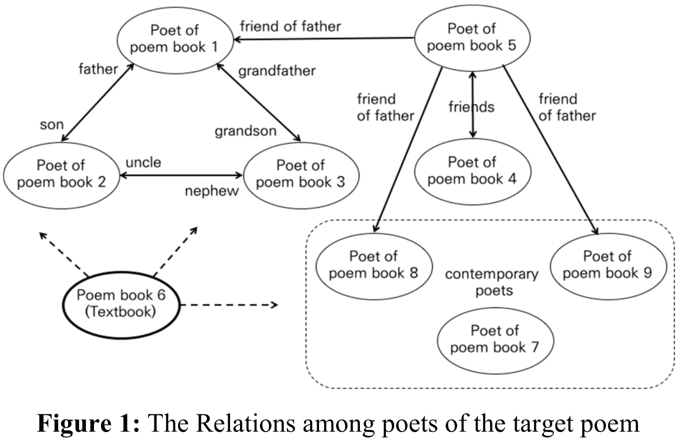

Figure 1: The Relations among poets of the target poem books

Data Model

In text mining analysis, understanding data plays a crucial role. Hence before performing pairwise relation analysis and quality estimation, we ought to define a model of poetic structure with some notations, which will be used to explain our analysis methods.

Let D be the set of poem books in the dataset $D = \{D_1, D_2, D_n\}$, where n is the number of poem books. We assume that each poem book is written by a different poet, so there are also n poets. Let the i-th poem book $D_i(1 \leq i \leq n)$ include poems $P^i = \{P_1^i, P_2^i, \dots, P_m^i\}$, where m is the number of poems in D_i. Let poem P_t^i have the title T_t^i, where $1 \leq i \leq n$ and $1 \leq t \leq m$. Let D_i be composed of a series of strings $L_1^i, L_2^i, \ldots, L_{l_i}^i$, where l_i is the number of lines in D_i. Let D_i be also composed of a series of characters $c^i_{(1,1)}, \ldots, c^i_{(1,z(1))}, c^i_{(2,1)}, \ldots, c^i_{(2,z(2))}, \ldots, c^i_{(l_i,1)}, \ldots, c^i_{(l_i,z(l_i))}$, where $z(k)$ is the number of characters in the k-th line in $D_i(1 \leq k \leq l_i)$. Therefore $c^i_{(j,k)}$ indicates the k-th character in the j-th line of the i-th poem book.

PAIRWISE RELATION ANALYSIS

We analyze the relations between two poem books based on their contents, titles, and extracted topic words.

Contents-based Analysis

Basically, the contents of poems are main targets for relation analysis. Contents are composed of each line without titles in a poem. When dealing with Classical Chinese characters, it is impossible to directly compute the term frequency-inverse document frequency (TD-IDF) as they are not based on alphabetic writing system [7]. The TD-IDF has been thus adjusted to the logographic writing system on which Classical Chinese characters are based. Since one character usually makes a word and two or more characters sometimes construct a word, basically characters are tokenized and sliding character windows are also used for tokenization. With that adjustment, poem data can be analyzed by traditional text mining methods. In order to overcome the limited number of documents, poem books in our case, we also consider the line frequency (LF), how many times a character appears in each line in the entire poem data, in addition to the document frequency (DF). The inverse line frequency (ILF) is computed as follows:

$$ILF = \log(1 + \frac{N}{LF}) \qquad (1)$$

where N is the total number of lines in the whole poem dataset. The ILF helps to compute commonness of characters with a small size of poem data and is used to compute the TF-ILF.

The TF-IDF or TF-ILF are used to put weights in calculating similarity between two different poem books. We use the well-known Jaccard and Cosine similarity algorithms [8, 9]. They are adjusted to the weighted vector zone scoring model defined by the TF-IDF or TF-ILF weights. Binary vectors are obtained from the bag-of-words model which validates a character as 0 or 1 according to the occurrence of characters [10]. Then a vector is generated for each line with binary values constructed from the bag-of-words model. For computation of accurate pairwise relations, binary vectors should be re-initialized to weighted vectors by multiplying the TF-IDF or TF-ILF index. This makes distinctive characters have higher similarity scores than common ones.

The comparison can be made by line-by-line ($L^i_{k_1}$, $L^j_{k_2}$) or poem-by-poem ($P^i_{t_1}, P^j_{t_2}$) between the pair of two poems (D_i, D_j) for $\forall\, k_1$, k_2 in $1 \leq k_1$, $k_2 \leq l_i$, and $\forall\, t_1$, t_2 in $1 \leq k_1$, $k_2 \leq m$. We use the line-by-line comparison due to the following two reasons. First, the lengths of lines in poems are mostly very similar while the lengths of poems are various. Second, one unique character might have different meanings according to its locations. The average line-based similarity between two poem books D_i and D_j is computed as follows:

$$Sim_L(D_i, D_j) = \frac{\sum_{k_1=1}^{l_i} \sum_{k_2=1}^{l_j} Jac_L|Cos_L(L^i_{k_1}, L^j_{k_2})}{l_i \times l_j} \qquad (2)$$

where $Jac_L\,|\,Cos_L\,(\cdot,\cdot)$ is the line-based Jaccard or Cosine similarity of two lines.

Title-based Analysis

The contents of a poem are naturally related to the title and sometimes the title can represent a bird's-eye view of the poem. We therefore add the title-based analysis to find another level of similarity between poems. The title-based relation analysis can be performed similarly to the content-based one as each poem has its title and a poem book can be reduced to a set of titles. Titles are considered a labeled corpus input for similarity algorithms.

In the title-by-title analysis, title $T^i_{t_1}$ in D_i is compared with other title $T^j_{t_2}$ in D_j for every pair of t_1 and t_2 in $1 \leq t_1 \leq m_i$ and $1 \leq t_2 \leq m_j$ where m_i and m_j are the numbers of poems in D_i and D_j respectively. The average title-based similarity between two poem books D_i and D_j is computed as the following two methods:

$$Sim_{T1}(D_i, D_j) = \frac{\sum_{t_1=1}^{m_i} \sum_{t_2=1}^{m_j} Jac_T|Cos_T(T^i_{t_1}, T^j_{t_2})}{m_i \times m_j} \qquad (3)$$

$$Sim_{T2}(D_i, D_j) = Jac_D|Cos_D(T^i, T^j) \qquad (4)$$

where $Jac_T|Cos_T(\cdot,\cdot)$ and $Jac_D|Cos_D(\cdot,\cdot)$ are the title-based and the document-based Jaccard or Cosine similarities of two titles and documents respectively, and T^i is the document constructed from a series of titles, $T^i_1, T^i_2, \cdots, T^i_{m_i}$ in D_i.

Topic Words-based Analysis

As poem titles are put by poets, titles can be considered abstracted keywords which represents the characteristics of the poem. Titles can be, however, composed of words which are not appeared in the poem itself. In this aspect, a title may not be an abstract from the poem contents. That is why we choose the topic words-based analysis. Topic words are set of words which often appear in a given document, but rarely occur in the other documents. With the topic words generated from our dataset, we could measure pairwise relations between poem books. The TF-IDF method is also used to extract topic words from a poem book. Due to the properties of our dataset, characters are used instead of words.

In the poem book *D*, the score s_c of each particular character *c* is calculated as s_c = *TF*(*c*, *D*)×*IDF*(*c*, *D*), where *TF*(*c*, *D*) and *IDF*(*c*, *D*) are the term frequency and the inverse document frequency of *c* in *D*. We can generate *N* topic words by choosing *N* highest-scored characters and the topic words of D_i can be denoted by $K^i = K_1^i, K_2^i, \cdots, K_N^i$, where K_k^i is the *k*-th highest-scored topic word in D_i. Then, the average topic words-based similarity between two poem books D_i and D_j , $Sim_K(D_i, D_j)$, is computed as follows:

$$Sim_K(D_i, D_j) = Jac_K|Cos_K(K^i, K^j) \qquad (5)$$

where $Jac_K|Cos_K(\cdot)$ is the topic words-based Jaccard or Cosine similarity of topic characters of two documents.

QUALITY ESTIMATION

The quality of a poem book will be computationally estimated based on its self-similarity and similarity to a reference textbook.

Self-Similarity-based Estimation

In terms of self-similarity-based estimation, high level of correlations inside a poem book might represent the shortage of poets' vocabulary or poetic creativity. As mentioned in the introduction, poets generally try to use their own unique words and to avoid using the same words in their poems to show their creativity. According to this consideration, we define the self-similarity based inverse quality of a poem book D_i, $IQ_L(D_i)$ as follows:

$$IQ_L(D_i) = \frac{\sum_{j=1}^{l_i}\sum_{k=1}^{l_i} Jac_L|Cos_L(L_j^i, L_k^i)}{l_i \times l_i} \qquad (6)$$

Reference-Similarity-based Estimation

It is possible to guess vocabulary-based quality of poems by comparing with reference textbooks. Students usually study literature by reading classic reference writings. They usually begin their own writing by imitating the styles of the references they like. But after they are experienced sufficiently, they can create their own writing styles. Considerable amount of correlations of a poem with the references may thus degrade the quality of the poem. According to this consideration, we define the reference-similarity-based inverse quality of a poem book D_i, $IQ_R(D_i)$ as follows:

$$IQ_R(D_i) = \frac{\sum_{j=1}^{l_i}\sum_{k=1}^{l_R} Jac_L|Cos_L(L_j^i, L_k^R)}{l_i \times l_R} \qquad (7)$$

where L_k^R is the string of the *k*-th line in the reference poem book D_R and l_R is the number of line in D_R.

RESULTS AND DISCUSSION

Pairwise Relations

We have already presented the target dataset, 9 poem books in the middle period of Joseon as shown in Table 1. The main targets of relation analysis are poem books 1 to 5. Poem book 6 is used as a reference, and poem books 7 to 9 are used as randomly chosen control data. Their poets seem to be related each other as shown in Figure 1.

Table 2 shows the contents-based relations of the target poem books with several similarity indexes. An underlined value of target poem books pair (D_i, D_j) indicates a relatively high similarity with respect to the average similarity between D_i and the others. A similarity value in bold indicates an absolute high similarity with respect to the average of all similarity values. A high similarity means at least 2% higher value than the average similarity. The pair of poem books indices (A, C) and (A, A) denote the average similarity between every pair of the control poem books and the target poem books, and between every pair of all the poem books except the reference poem book D_6, respectively

Table 2: The results of contents-based analysis

Tar-get	TF-IDF Based		TF-ILF Based	Tar-get	TF-IDF Based		TF-ILF Based
	Jaccard	Cosine	Cosine		Jaccard	Cosine	Cosine
1-2	0.0940	0.1721	0.2732	1-4	0.0919	0.1682	0.2685
1-3	0.0940	0.1714	0.2714	1-5	0.0922	0.1687	0.2740
2-1	0.0940	0.1721	0.2732	2-4	0.0942	0.1734	0.2656
2-3	**0.0982**	**0.1799**	**0.2730**	2-5	0.0942	0.1735	0.2704
3-1	0.0940	0.1714	0.2714	3-4	**0.0954**	**0.1751**	**0.2751**
3-2	**0.0982**	**0.1799**	0.2730	3-5	**0.0955**	**0.1751**	**0.2794**

4-1	0.0919	0.1682	0.2685	4-3	**0.0954**	**0.1751**	**0.2751**
4-2	0.0942	0.1734	0.2656	4-5	0.0925	0.1704	**0.2781**
5-1	0.0922	0.1687	0.2740	5-3	**0.0955**	**0.1751**	**0.2794**
5-2	0.0942	0.1735	0.2704	5-4	0.0925	0.1704	*0.2781*
A-C	0.0922	0.1696	0.2667	A-A	0.0929	0.1707	0.2689

The results in Table 2 show that the similarities to the control poem books are lower than those between target poem books. As control poem books have been chosen randomly, this seems to be reasonable. Among the target poem books, D_3 has high similarities to other poem books overall. More specifically, poem books pair (D_2, D_3) shows an obvious high similarity and poem book pairs (D_1, D_2), (D_1, D_3), and (D_1, D_5) have slightly high similarity. The results of the TF-IDF based similarity are similar to those of the TF-ILF based similarity, and Jaccard and Cosine similarity results are nearly the same. We need to investigate why there are differences between the TF-IDF based and the TF-ILF based similarities of (D_3, D_2) and (D_4, D_5) as a further study.

Table 3: The results of title-based analysis

Target Poem Books	Title-by-Title Analysis		Corpus Analysis with All Titles	
	Jaccard	Cosine	Jaccard	Cosine
1-2	0.1098	0.2019	0.0981	0.2133
1-3	0.1093	0.2014	**0.2674**	**0.2513**
1-4	0.1068	0.1961	**0.2664**	**0.3334**
1-5	0.1143	0.2095	**0.3187**	**0.2914**
2-1	0.1098	0.2019	0.0981	0.2133
2-3	0.1193	0.2164	**0.1361**	0.1819
2-4	0.1217	0.2183	0.1027	0.1693
2-5	0.1242	**0.2236**	0.1125	0.1610
3-1	0.1093	0.2014	**0.2674**	**0.2513**
3-2	0.1193	0.2164	0.1361	0.1819
3-4	0.1180	0.2142	**0.2130**	0.2520
3-5	0.1220	0.2206	**0.2443**	0.2672
4-1	0.1068	0.1961	**0.2664**	**0.3334**
4-2	0.1217	0.2183	0.1027	0.1693
4-3	0.1180	0.2142	**0.2130**	0.2520
4-5	0.1210	0.2191	**0.2686**	0.2378
5-1	0.1143	0.2095	**0.3187**	**0.2914**
5-2	0.1242	0.2236	0.1125	0.1610
5-3	0.1220	0.2206	**0.2443**	0.2672
5-4	0.1210	0.2191	**0.2686**	0.2378
A-C	0.1231	0.2235	0.1833	0.2258
A-A	0.1208	0.2194	0.1903	0.2294

Table 3 shows the similarity results based on the titles of poems. These results are somewhat different from those based on the contents. D_1 appears to have very high similarities to other poem books especially in the corpus analysis with all the titles. The results of title-by-title analysis results are more similar to those of contents-based line-by-line analysis except that the similarities to the control poem books are higher than those between the target poem books on the contrary. By the way, the corpus analysis with all the titles newly shows high similarity among D_3, D_4, and D_5.

The topic words based similarities shown in Table 4 indicate another different result. Poem books group (D_1, D_3, D_4, and D_5) shows high pairwise similarity, which is similar to the title-based similarity with corpus analysis. Group (D_2, D_3, and D_5) also shows high pairwise similarity, somewhat similar to the contents-based line-by-line similarity.

Table 4: The results of topic-words-based analysis

Target Poem Books	Topic-Words based Similarity		Target Poem Books	Topic-Words based Similarity	
	Jaccard	Cosine		Jaccard	Cosine
1-2	0.1942	0.3720	1-4	**0.2117**	**0.4069**
1-3	**0.3371**	**0.5043**	1-5	**0.2723**	**0.4798**
2-1	0.1942	0.3720	2-4	0.1829	0.3095
2-3	**0.1985**	**0.3804**	2-5	**0.2209**	**0.3622**
3-1	**0.3371**	**0.5043**	3-4	0.1528	0.3101
3-2	**0.1985**	**0.3804**	3-5	0.1827	0.3477
4-1	**0.2117**	**0.4069**	4-3	0.1528	0.3101
4-2	0.1829	0.3095	4-5	**0.2808**	**0.4398**
5-1	**0.2723**	**0.4798**	5-3	0.1827	0.3477
5-2	**0.2209**	**0.3622**	5-4	**0.2808**	**0.4398**
A-C	0.1551	0.3135	A-A	0.1795	0.3413

By combining the results in Tables 2 to 4, the overall pairwise similarity scores are derived as shown in Table 5. The relations with underlined values, in bold, and underlined in bold are given 1, 2, and 3 points respectively. Table 5 shows the total similarity points of all pairs of the target poem books. The bold and underlined similarity indexes indicate 20~40% and more than 40% higher similarity than the average similarity respectively. Figure 2 depicts the derived relations among the target poem books. Solid and dotted lines indicate 'high' and 'rather high' relations respectively. Compared by Figure 1, the relation analysis captured the similarity between poem books D_2 and D_3, D_1 and D_3, and D_1 and D_5 successfully as expected. However, the close relations between poem books D_1 and D_4, and D_3 and D_5 have newly come out, which needs further investigation with the result details in the future.

Table 5: The overall results of pairwise similarity

Target Poem Books	Total Similarity Pts	Target Poem Books	Total Similarity Pts
1-2	5	2-4	1
1-3	25	2-5	15
1-4	24	3-4	21
1-5	**28**	3-5	24
2-3	**30**	4-5	22
Average	19.5		

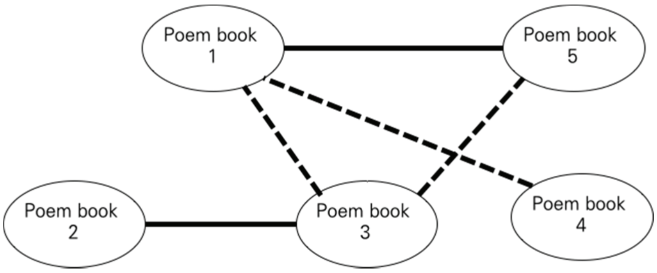

Figure 2: Derived pairwise relations between the target poem books

Quality Estimation

Now we estimate vocabulary-based quality of the poem books. Tables 6 and 7 show the inverse quality values computed by the equations (6) and (7) respectively. Self-similarity based inverse quality shown in Table 6 indicates that poem books D_2 and D_3 have high self-similarity. According to the TF-ILF based cosine similarity, poem book D_5 shows high self-similarity, and D_4, D_6, and D_9 also show somewhat high self-similarity. The pair of poem books indexes (S, S) denote the average self-similarity of all the poem books.

Table 6: The results of self-similarity analysis

Target Poem Books	TF-IDF Based		TF-ILF Based
	Jaccard	Cosine	Cosine
1-1	0.0916	0.1677	0.2682
2-2	**0.0988**	**0.1800**	0.2730
3-3	**0.0993**	**0.1810**	**0.2887**
4-4	0.0931	0.1710	0.2781
5-5	0.0943	0.1725	**0.2894**
6-6	0.0909	0.1674	0.2768
7-7	0.0920	0.1687	0.2721
8-8	0.0916	0.1681	0.2629
9-9	0.0911	0.1706	0.2743
S-S	0.0936	0.1719	0.2759
A-A	0.0929	0.1707	0.2689

Table 7: The results of reference-similarity analysis

Target Books	Contents-based		Title-based		Topic-words	
	Jaccard	Cosine	Jaccard	Cosine	Jaccard	Cosine
1-6	0.0904	0.1657	0.0951	0.1812	0.1696	0.2196
2-6	0.0928	0.1711	**0.1242**	0.2236	0.1274	0.2374
3-6	0.0941	0.1725	0.1121	0.2085	**0.1983**	0.2483
4-6	0.0906	0.1672	0.1112	0.2066	0.1318	0.2218
5-6	0.0907	0.1674	0.1088	0.2036	0.1221	0.2105
7-6	0.0898	0.1674	0.1050	0.1976	0.0359	0.1536
8-6	0.0899	0.1659	0.1156	0.2136	0.0961	0.2369
9-6	0.0910	0.1674	0.1191	0.2197	**0.1869**	0.2489
A-6	0.0911	0.1680	0.1114	0.2068	0.1335	0.2221
A-A	0.0929	0.1707	0.1208	0.2194	0.1795	0.3413

Table 7 shows the result of reference-similarity based inverse quality of the poem books, where the pair of poem books indexes (A, 6) denote the average reference-similarity of all the poem books. Poem books D_2, D_3, D_8, and D_9 appear to have relatively high similarity with the reference poem book D_6. Overall, we can say that the quality of poem books D_2 and D_3 might be lower than the others in the sense that they used somewhat limited vocabulary. Note that this is just an estimation from a simple character-based analysis, which is not related to their artistic quality. There should be more cooperative studies with researchers in the area of Classical Chinese poetry for further refined estimation of poetic quality.

CONCLUDING REMARKS

This study presented a text-mining based analysis method of Classical Chinese poetry by focusing on pairwise similarity between poems and quality estimation of a poem. The results have shown a high potential of this approach because they considerably correspond to the existing intuitive evaluation on the relations among the target poems and poets.

We are extending our study as follows. We are developing various measure for further relation analysis such as sentiment, theme, metaphor, color, citation, and the characters. We also try to analyze semantic similarity with precise word embedding such as word2vec [11]. In addition, deep neural networks are considered for extracting analytic features or characteristic styles of poems and poets. Ultimately, we aim to analyze all the poem books in the 15th century in Joseon to examine the whole relation between poets in that time.

ACKNOWLEDGMENT

This paper was supported by Research Fund, Kumoh National Institute of Technology.

REFERENCES

[1] Classical Chinese poetry, *Wikipedia*, Refer to https://en.wikipedia.org/wiki/Classical_Chinese_poetry.

[2] Von-Wun Soo, Shih-Yao Yang, Shu-Lei Chen and Yi-Ting Fu, "Ontology acquisition and semantic retrieval from semantic annotated Chinese poetry," *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries*, Tucson, AZ, USA, 2004, pp. 345-346.

[3] D. M. Kaplan and D. M. Blei, "A Computational Approach to Style in American Poetry," *Seventh IEEE International Conference on Data Mining* (ICDM 2007), Omaha, NE (2007), pp. 553-558.

[4] Shinji Kikuchi, Keizo Kato, Junya Saito, Seiji Okura, Kentaro Murase, Takaya Yamamoto and Akira Nakagawa, "Quality estimation for Japanese Haiku poems using Neural Network," *2016 IEEE Symposium Series on Computational Intelligence* (SSCI), Athens, (2016), pp. 1-8.

[5] J. Lee and M. Luo, "Word clustering for parallelism in Classical Chinese poems," *2016 International Conference on Asian Language Processing* (IALP), Tainan, 2016, pp. 49-52.

[6] Institute for the translation of Korean classics, Korean classical literature integrated database, Refer to http://db.itkc.or.kr/.

[7] Td-idf, Wikipedia, Refer to https://en.wikipedia.org/wiki/Td-idf.

[8] Jaccard index, *Wikipedia*, Refer to https://en.wikipedia.org/wiki/Jaccard_index.

[9] Cosine similarity, *Wikipedia*, Refer to https://en.wikipedia.org/wiki/Cosine_similarity.

[10] Bag-of-words model, *Wikipedia*, Refer to https://en.wikipedia.org/wiki/Bag-of-words_ model.

[11] C. H. Liu, Q. Liu and C. H. Lee, "Valence-arousal ratings prediction of Chinese words using similarity measures based on Word2Vec," *2016 International Conference on Asian Language Processing* (IALP), Tainan, (2016), pp. 317-319.