한시(漢詩) 주제 자동 분류를 위한 사용자 피드백 적용 시스템

이정한[1] · 하민수[2] · 지영익[3] · 민경주[4,*]

충남대학교

User Feedback Integration System for Automatic Topic Classification of Ancient Sino-Korean Poetry

Jung-han Lee[1] · Min-su Ha[2] · Young-Ik Ji[3] · Kyoung-ju Min[4,**]

Chungnam National University

E-mail : {air_need, alstn8742, podo1234567}@naver.com / nadopro@gmail.com

요 약

한시(漢詩)는 표의 문자인 한자를 사용하여 제한된 음절로 구성된 정형화된 형식의 시문학으로, 의미가 함축적이고 비유적 표현이 많아 인공지능 번역이나 주제 분류에 어려움이 있으며 뚜렷한 성과를 이루지 못하고 있다. 인공지능 및 딥러닝을 활용한 기존의 연구들은 의미 맥락을 분석하고 주제를 분류하려 하였으나, 결과의 타당성을 검증하고 사용자 피드백을 반영하는 데 한계가 있었다. 본 연구에서는 한시의 주제를 먼저 분류한 후, 시어를 n-그램 분석하여 출현 빈도에 따라 가중치를 부여함으로써 주제를 분류하는 방법을 제안한다. 3음절 이하의 시어를 주제별로 선정한 후, 사용자가 주제별 시어를 추가, 갱신하며, 그에 따른 피드백을 반영해 주제 분류의 정확성을 개선할 수 있다. 이를 위해 데이터 분석의 병목 현상을 최소화하고, 사용자 피드백을 실시간으로 분석하여 시각화된 결과를 제공하는 방식을 채택하였다.

ABSTRACT

Ancient Sino-Korean Poetry (漢詩), written in Chinese characters, follows structured formats and contains implicit, metaphorical expressions that challenge AI-based translation and topic classification. Traditional AI and deep learning approaches often struggle to validate results and incorporate user feedback effectively. This study aims to improve topic classification by first categorizing themes, then using n-gram analysis to assign weights based on word frequency. Poetic words of three syllables or fewer are selected per theme, with user feedback used to update the vocabulary and refine accuracy. By minimizing database access, real-time analysis and visualization of feedback are enabled to enhance classification.

키워드

Topic Classification, n-gram analysis, User feedback, Ancient Sino-Korean Poetry, Data Visualization

Ⅰ. 서 론

표음 문자인 한글과 달리, 표의 문자인 한자는 각 글자마다 고유한 의미를 지닌다. 이러한 한자로 쓰인 한시(漢詩)는 대부분 20자, 28자, 40자, 또는 56자의 정형화된 형식으로 구성되며, 의미가 함축적이고 비유적인 표현이 많아 의미 맥락을 이해하기 어렵다. 이러한 이유로 인공지능(Artificial Intelligence, AI)이나 딥러닝(Deep Learning) 기술을 활용한 한시 번역과 주제 분류에 대한 다양한 시

* 교신저자

도가 이루어지고 있지만, 아직 뚜렷한 성과를 거두지 못하고 있다. 특히, BERT(Bidirectional Encoder Representations from Transformers)와 같은 자연어 처리(Natural Language Processing, NLP) 모델을 사용한 연구에서는 언어적 의미를 추출하여 그룹화된 결과를 확인할 수는 있지만, 사용자 피드백을 반영해 알고리즘을 개선하는 데에는 한계가 있다.

본 연구는 사용자 피드백을 반영하기 어려운 주제 분류의 한계를 극복하기 위하여, 한시의 주제를 확장 가능한 40여 개로 구분하고 각 주제별로 3음절 이하의 시어(詩語, Poetry Word)를 설정하였다. 이를 바탕으로 n-그램 분석을 통해 시어별 가중치를 부여하여 주제를 분류하고자 하였으며, 한자어의 사전화가 어려운 특성도 고려하였다. 분석 결과에 따라 주제별 시어를 추가, 갱신, 삭제하고 사용자 피드백을 반영하여 주제 분류의 정확도를 지속적으로 개선할 수 있다.

대규모 한시를 분석하는 데 있어 연산 속도는 매우 중요하며, 이를 위해 데이터베이스의 접근을 최소화하여 데이터 분석의 병목을 줄이고[1, 2], 사용자 피드백을 실시간 분석에 반영한다. 이러한 반영된 결과는 시각화된 형태로 제공되어 데이터 분석과 결과의 직관성을 높인다[3, 4]. 이는 학령 인구가 감소하는 등 인구 구조의 변화로 인한 인문학 연구에 디지털 분석을 통한 해결책을 제시하는 시도이기도 하다. 이를 위해 인문학 연구자가 사용하기 쉬운 웹 기반 도구를 개발하였으며, 데이터 시각화를 위해 D3.js(Data Driven Document JavaScript)를 활용하였다[5, 6]. 이 연구는 인구 구조 변화로 인해 어려움에 직면한 인문학 분야에 디지털 분석을 통해 새로운 돌파구를 제공하고, 연구 환경의 변화를 효과적으로 대응하는 방법을 제시하고자 한다.

II. 관련 연구 및 개발 환경

2.1 관련 연구

한문은 표의 문자이기 때문에 디지털 분석에서 표면적 분석과 의미적 분석으로 나눌 수 있다. 표면적 분석을 위해 n-그램 분석은 한문 자료의 기초 통계 분석에 활용되며, 의미적 분석에는 BERT와 같은 NLP 모델이 사용되고 있다. 최근에는 생성형 AI인 ChatGPT가 등장하면서 AI 기반 분석이 연구의 화두로 떠오르고 있다. 하지만 NLP 모델은 사용자의 피드백을 반영하기 어려워, 제시된 결과를 확인하는 용도로만 활용되고 있다.

디지털 인문학은 공학 기술의 영향을 많이 받지만, 인문학 연구자가 이를 이해하고 활용하기 어려운 면이 있다. 이를 해결하기 위해 데이터 분석 결과를 네트워크 다이어그램 등 시각화 기술과 접목하여 연구자가 쉽게 접근할 수 있도록 하는 시도가 이루어지고 있다.

2.2 개발 환경

본 연구를 위한 개발 환경은 다음 표 1과 같다.

표 1. 개발 환경

항목	항목
개발환경	xampp 8.0.3
OS	Windows 10 Enterprise Ed.
DB	MariaDB 10.4.22
시각화	d3.js ver.7
원문 자료	농암집, 청음집의 한시

III. 한시 주제 분류

3.1 주제별 시어

한시는 5언(글자), 7언으로 이루어진 절구(4줄)나 율시(8줄)로 구성되며, 짧으면서 함축적이고 비유적인 표현을 사용한다. 예를 들어, 강(江, River)은 자연물, 시간의 흐름, 영속성, 장소와 경계, 단절을 의미하는 이별의 공간을 의미한다. 백운(白雲, White Cloud)는 하늘의 흰 구름, 자유와 해탈, 한가로운 생활, 돌아가신 아버지를 상징하는 시어로도 사용된다. 이러한 인문학적 의미를 반영해 자연, 계절, 고독과 우울, 삶과 죽음, 여행, 자아성찰 등 40여 개의 주제를 미리 선정하고, 주제별 3음절 이하의 시어 100개씩을 선정하였다. 시어 선정 과정에서는 ChatGPT 4.o를 활용하여 초기 자료를 입력한 후, 연구자가 이를 보정하였다.

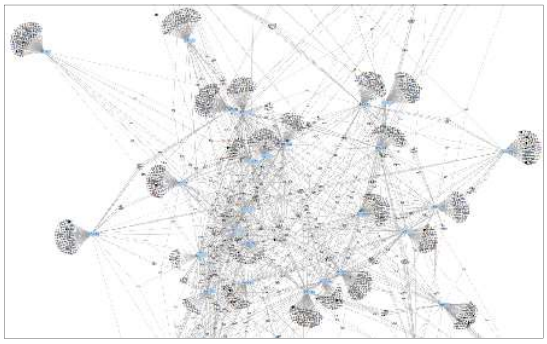

그림 1. 주제어별 2음절어의 분포

그림 1은 주제별 2음절 시어 분포의 일부를 보여주고 있다. 그림 1에서 여러 주제에 동시에 등장하는 시어와 특정 주제에만 분포하는 시어를 확인할 수 있다. 이는 기존의 주제 분류가 단일 주제에 초점을 맞추었던 것과 달리, 다수의 주제로 분류하는 방식이 인문학적 관점에서 더 바람직한지 검토하기 위한 시도이다.

3.2 한시 주제 자동 분류 알고리즘

주제를 설정한 후, 시어의 출현 빈도를 바탕으로 한 한시의 주제 자동 분류를 위한 전체 알고리즘

은 그림 2와 같다.

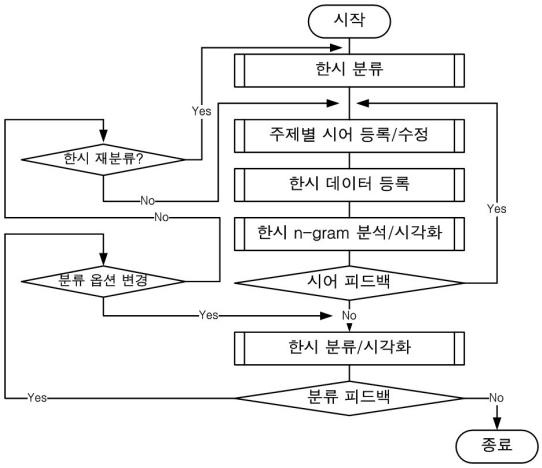

그림 2. 한시 주제 자동 분류 알고리즘

그림 2에서 보는 바와 같이, 한시의 분류, 주제별 시어, 한시 분류 결과에 따라 지속적인 사용자 피드백으로 최적의 답을 찾아가는 방식으로 동작한다.

3.3 한시 데이터 특성 이해를 위한 시각화

지금까지 수천 종의 문집이 인문 연구자들에 의해 디지털로 변환되고 번역 작업까지 이루어져왔다. 그중 농암 김창협(1651-1708)의 농암집에 수록된 902수의 시 전체를 n-그램 분석한 결과, 4회 이상 출현한 3음절 이하의 시어 분포는 그림 3과 같다.

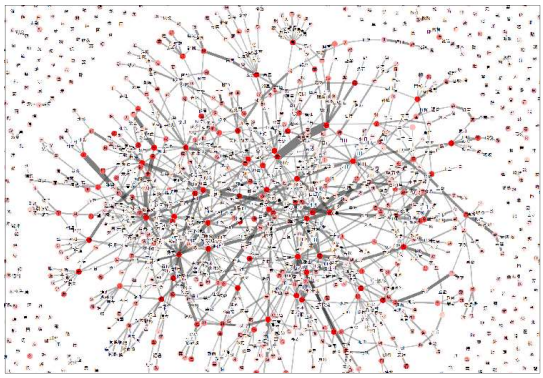

그림 3. 농암집의 4회 이상 출현어 시각화

그림 3에서, 노드의 농도와 링크의 두께로 구분된 결과를 통해 특정 시어가 반복적으로 사용되는 특성을 확인할 수 있다. 이는 본 연구에서의 주제 분류가 다중 주제로 분류될 가능성을 보여준다.

3.4 한시 주제 분류

한시의 주제, 분류 대상 문집인 농암집의 902수 자료, 주제별 시어는 모두 데이터베이스에 저장되어 있다. 데이터베이스는 특수한 파일시스템에서 동작하는 소프트웨어이기 때문에, 주제 분류를 위해 데이터베이스를 반복적으로 접근하면 성능 저하를 피할 수 없다. 이를 해결위해 그림 4와 같은 알고리즘을 사용하였다.

```
while($catData = fetchCategory()) {      // ①
    addCategoryArray($catData);
    $nodeArray = addNodeArray($catData);
}

while($bookData = fetch_book()) {      // ②
    ngramAnalysis();
    $nodeArray = addNodeArray($bookData);
    $linkArray = addLinkArray($bookData);
    calculateTopicPoint();
    sortTopicPoint();
    adjustTopicPoint();
    determinTopic();
}

makeJson($nodeArray, $linkArray);
showTopicGraph();
```

그림 4. 한시 분류 알고리즘

그림 4의 ①은 주제별 시어를 메모리에 배열 형태로 한 번 탑재하여 이후의 연산을 메모리에서 처리함으로써 성능을 높이는 과정을 보여준다. ②에서는 문집자료를 읽어오는 작업을 최소화하고, 접근시 메모리에 탑재하여 효율성을 극대화한다. adjustTopicPoint()는 시의 길이에 따라 값을 보정하는 기능으로, 긴 시와 짧은 시가 직접 비교될 때 발생할 수 있는 데이터 왜곡을 방지하는 역할을 한다.

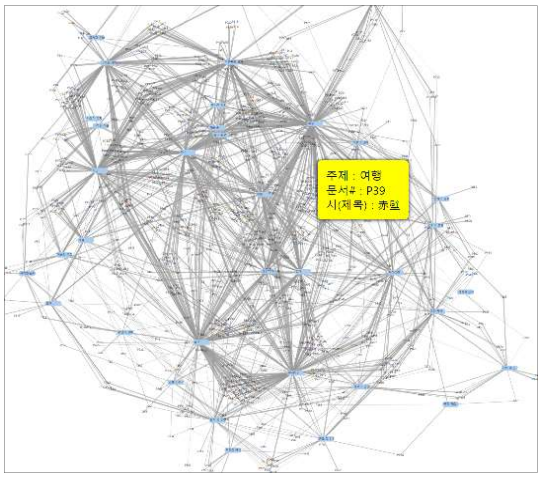

그림 5. 한시 분류 결과의 시각화

그림 5는 위 알고리즘을 적용하여 농암집의 902

수 시 자료를 n-그램 분석한 후, 출현 빈도에 따라 주제를 분류한 결과를 보여준다. 이 결과는 JSON(JavaScript Object Notation) 파일로 저장된 후, D3.js 네트워크 다이어그램을 통해 시각화되었다. 복잡한 결과 속에서도 사용자가 링크에 마우스를 올리면 주제, 문서 번호, 시의 제목을 확인할 수 있도록 하여, 자료 접근의 편의성을 높였다. 그림에서 시가 단일 주제로 분류되는 경우도 있지만, 다중 주제로 분류되는 것이 바람직한 경우를 확인할 수 있다.

주제어를 메모리에 탑재하는 데 걸린 시간은 0.00363초 (그림 4의 ① 과정), 902수 자료 분석까지는 3.038646초(그림 4의 ② 과정), 그리고 JSON 작업까지 포함한 전체 누적 시간 3.042886초로, 주제 분류에 대부분의 시간이 소요되었다.

Ⅴ. 결 론

본 연구는 인공지능 및 딥러닝 기술이 한시(漢詩)의 번역 및 주제 분류에 있어 한계를 가지는 문제를 해결하기 위해 진행되었다. 특히, 기존의 NLP 모델이나 BERT와 같은 딥러닝 기술은 사용자 피드백을 실시간으로 반영하고 그에 따라 알고리즘을 개선하는 데 어려움이 있다. 이를 해결하기 위해, 본 연구에서는 40여 개의 주제를 미리 선정하고, 각 주제별로 3음절 이하의 시어를 n-그램 분석을 통해 가중치를 부여하여 주제를 자동으로 분류하는 방식을 제안하였다.

본 연구의 핵심은 주제 분류의 과정을 사용자 피드백을 통해 지속적으로 개선할 수 있는 알고리즘을 제안했다는 점이다. 데이터를 메모리 기반으로 처리하여 데이터베이스 접근을 최소화함으로써, 연산 속도를 크게 향상시키고 실시간 피드백을 반영한 분석 결과를 빠르게 시각화 할 수 있었다. 이러한 방법은 인문학 연구자들이 디지털 분석 기술을 더 쉽게 사용할 수 있도록 돕는 동시에, 결과의 직관성과 투명성을 높이는데 기여한다.

또한, 본 연구는 인공지능과 딥러닝 기술이 피드백을 반영하기 어려운 문제를 보완하는 알고리즘을 제시함으로써, 향후 인문학 분야에서 디지털 분석의 효율성을 증대시키고 새로운 연구 패러다임을 제시할 가능성을 열었다. 특히, 본 연구에서 제안된 알고리즘은 다양한 주제에서 다중 분류가 가능하도록 개선될 수 있으며, 이를 통해 보다 정확하고 신뢰성 있는 주제 분류 결과를 제공할 수 있을 것이다.

Acknowledgement

이 논문은 인문사회융합인재양성사업단 [HUSS] 지원을 받아 작성됨.

References

[1] K. J. Min, J. Y. Cho, M. H. Jung and H. B. Lee, “Analysis of Impact Between Data Analysis Performance and Database,” *Journal of Information and Communication Convergence Engineering*, Vol. 21, No. 3, pp. 244-251, Sep. 2023.

[2] K. J. Min, J. Y. Cho, M. H. Jung and H. B. Lee, “Optimization for Large-Scale n-ary Family Tree Visualization,” *Journal of Information and Communication Convergence Engineering*, Vol. 21, No. 1, pp. 54-61, Mar. 2023.

[3] B. C. Lee and K. J. Min, “A Study on Visualization of the Analysis between the Collections of Korean Literature in Korea Class DB,” *Journal of Korean Classics*, Vol. 57, pp. 5-32, Mar. 2021.

[4] K. J. Min and B. C. Lee, “The Analysis of Chosun Dynasty Poetry Using 3D Data Visualization,” *Journal of the Korea Institute of Information and Communication Engineering*, Vol. 25, No. 7, pp. 861-868, Jul. 2021.

[5] K. J. Min, “Plan for the Improvement of the Interpersonal Relationship Network in the Korean Classics DB,” *Journal of Korean Classics,* Vol. 59, , pp. 197-241, Nov. 2021.

[6] K. J. Min, B. C. Jin and M. H. Jung, “Massive Graph Expression and Shortest path Search in Interpersonal Relationship Network,” *Journal of the Korea Institute of Information and Communication Engineering*, Vol. 26, No. 4, pp. 624-632, Apr. 2022.