데이터 기반 공모전 분석 프레임워크 설계: 텍스트마이닝을 활용한 공모전 기획 인사이트 도출

임채진[1], 한정수[2], 이현섭[2*]
[1]백석대학교 일반대학원 소프트웨어융합 전공 박사과정, [2]백석대학교 컴퓨터공학부 교수

Data-Based Contest Analysis Framework Design: Using Text Mining to Derive Insights for Contest Planning

Chae-Jin LIM[1], Jung-Soo HAN[2], Hyun-Seob LEE[2*]
[1]PhD candidate, Software Convergence, Baekseok University
[2]Professor, Division of Computer Engineering, Baekseok University

요 약 2014년 한국에서 열린 공모전 수는 약 4천 건 이였으며 2024년에는 약 1만여 건으로 증가하였다. 이처럼 공모전은 분야별로 꾸준히 증가하고 있다. 이에 따라 공모전 기획자들에게는 비용을 절감하며 최대한의 효과를 얻을 수 있는 효율적인 주제 선정이 중요해지고 있다. 자연히 이를 도울 수 있는 데이터 기반 분석 도구의 필요성도 대두된다. 본 연구에서는 2011년부터 2024년까지의 만화 공모전 데이터 7,587건(중복 제거 후 5,646건)을 텍스트마이닝 기법을 활용해 분석하는 프레임워크를 제안한다. 본 연구는 TF-IDF(Term Frequency-Inverse Document Frequency)와 Word2Vec(Word-To-Vector)을 결합한 조인 벡터(Join Vector) 기법을 통해 키워드의 출현 빈도와 의미의 연관성을 함께 반영하고, 시간 흐름에 따른 경향을 종합적으로 파악하여 공모전 기획자에게 실질적인 인사이트를 제공할 수 있는 방법을 제시한다. 이 프레임워크는 빈도 기반 중요도 점수와 의미의 복합 벡터를 통한 중요 키워드 순위를 만든다. 최종적으로는 연도별 변화율의 평균치를 통하여 전체 기간을 관통하는 추세 지표를 산출한다.

주제어 : 공모전, 만화, 텍스트마이닝, 빅데이터 분석, 콘텐츠-기술 간 융합

Abstract In 2014, the number of contests held in Korea was about 4,000, and it is expected to increase to about 10,000 by 2024. The number of contests is steadily increasing by sector. It's becoming increasingly important for contest organisers to select topics efficiently to maximise impact while reducing costs. This increases the need for data-driven analytical tools to help them do this. In this study, we propose a framework to analyse 7,587 (5,646 after deduplication) comic contest data from 2011 to 2024 using text mining techniques. Through an analysis methodology that combines TF-IDF and Word2Vec, we propose a methodology that reflects the frequency of keyword occurrence and semantic associations together, and comprehensively identifies trends over time to provide practical insights for contest planners. The framework creates a ranking of important keywords through a composite vector of frequency-based importance scores and semantics. Finally, it calculates a trend metric that spans the entire time period by averaging the percentage change from year to year.

Key Words : Contest, Comics, Text-mining, Big-Data Analysis, Converging content & technology

*이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-지역지능화혁신인재양성사업의 지원을 받아 수행된 연구임 (IITP-2025-RS-2024-00436765)
*교신저자 : 이현섭(hyunseob@bu.ac.kr)
접수일 2025년 04월 03일 수정일 2025년 04월 28일 심사완료일 2025년 05월 13일

1. 서론

공모전은 창작자의 독창적 아이디어와 스토리텔링이 반영되는 문화적 이벤트다. 이 이벤트에서 주최 측은 저렴한 비용으로 마케팅 효과와 인력수급을 해결하려고 하는 반면, 참가자들은 자기 아이디어로 경력과 성장의 기회를 잡으려고 한다. 2014년 한국에서 열린 공모전 수는 약 4천 건 이였으며 2024년에는 약 1만여 건으로 증가하였다. 이처럼 공모전은 분야별로 꾸준히 증가하고 있다.

한국에서 공모전은 2014년 한 해 4천 건에서 2024년 들어서는 10,384건이 열려, 10년 만에 약 260%에 달하는 양적 성장세를 보여주고 있다[1,2]. 공모전을 주최하는 주체는 주로 기업체와 기관, 지방자치단체, 사회·시민단체 등이다. 이들이 공모전을 여는 까닭은 ① 다양한 아이디어 획득 ② 사회공헌을 포함한 홍보 마케팅의 최적 수단 ③ 잠재 고객 유치 및 친(親) 기업 정서 형성 ④ 창의적 인재 발굴 [2,3] 등이다. 이상과 같은 개최 사유는 비용 대비 효과 면에서 주최 측에 매력적이다[4,5]. 하지만 비교적 젊은 연령대를 차지하고 있는 참가자들은 대부분 공모전을 수익보다는 경력 획득과 성장을 이루려는 기회로 활용하고 있다.

본 연구는 TF-IDF(Term Frequency - Inverse Document Frequency)와 Word2Vec (Word-To-Vector)을 결합한 분석 방법론을 통해 키워드의 출현 빈도와 의미의 연관성을 함께 반영하고, 시간 흐름에 따른 경향을 종합적으로 파악하여 공모전 기획자에게 실질적인 인사이트를 제공할 수 있는 방법을 제시한다.

2. 공모전 분석의 필요성

공모전의 공고문 본문은 해당 공모전에 대한 모든 정보를 압축하고 있다. 다만 건마다의 길이가 길지 않아 정보량이 제한적이고 안내를 위해 반복적으로 등장하는 고정 문구 등 가치 낮은 정보들이 많이 섞여 있다. 하지만 공모전 정보란 각각의 유효 정보량은 매우 적으나 시기별로 쌓인 양은 많은 데이터 집합이며, 중심이 되는 정보들에 핵심적으로 접근할 수 있다면 종합적인 판단을 내리기 위한 기초 자료를 산출할 수 있다.

텍스트마이닝은 비정형적인 텍스트 데이터에서 유의미한 정보를 뽑아내는 기술로, 주로 대량의 문서 세트 속에서 문서의 주제와 경향을 파악하는 데에 쓰인다. 수 만 건의 개최 건수가 쌓인 공모전의 정보를 분석하여 공모전의 패턴과 인사이트를 산출하는 데에는 텍스트마이닝이 유효한 도구가 될 수 있다. 텍스트 마이닝을 활용해 공고문의 중심이자 경향 해석의 핵심이 되는 키워드(Keyword)들을 뽑고 여기에 중요도를 산정해 순위를 매길 수 있다면, 그 자체가 곧 공모전에 대한 판단 기준을 돕는 분석 프레임워크의 기초가 된다.

본 연구는 다양한 공모전 주제 가운데에 '만화' 분야의 공모전을 특정한다. 만화 공모전은 아이디어와 스토리텔링이 융합된 결과물을 다방면에 활용하기 좋은 형태로 얻을 수 있으며, 영상 등에 비해 제작 인력이 적고 비용이 낮아 다수의 기관과 업체, 지방자치단체들이 앞다퉈 주최하고 있다. 즉, 만화 공모전은 공모전이라는 주제를 탐색하는 데에 적합한 소재라 할 수 있다.

본 연구는 만화 공모전이 요구하는 바를 확인하기 위해 국내 주요 공모전 정보 사이트 네 곳(씽굿, 콘테스트코리아, 위비티, 씽유)[6,7,8,9]에서 만화/웹툰/인스타툰/카툰 키워드로 검색된 총 7,587건의 데이터를 수집한 후 정제, 전처리를 진행해 중복분을 제거한 5,656건의 제목과 주제 등의 공모전 정보 중 1건뿐인 2011년과 1개월분 일부만 있는 2025년을 제외한 13년 치의 정보에 텍스트마이닝을 적용해 분석하였다.

3. 관련 연구

공모전 분석에 텍스트 마이닝을 접목한 선행 연구로는 건축 공모전의 키워드 변화 양상을 텍스트 마이닝으로 분석한 사광균(2024)의 연구를 들 수 있다. 사광균의 연구는 텍스트 마이닝을 통해 공모전 정보에서 "현시대의 건축적 주제를 파악"하려는 의도를 보여준다[10]. 그러나 이 연구는 코로나19를 전후한 시기의 공모전 주제를 비교 분석하는 방식이어서 특정 분야의 전체 경향을 세밀하게 조망하는 데에는 이르지 않아 본 연구의 방향과는 다소 차이가 있다.

이와 같은 문제에 대안점을 제시하는 연구가 박대서 외 1인의 연구[11]다. 이 연구는 키워드의 중요도를 측정하는 데에 널리 쓰이는 통계적 지표인 TF-IDF에 의미적 요소를 반영하기 위한 결합 벡터를 제안한다. 이 연구는 TF-IDF로 빈도 기반 벡터화를 진행한 후, 키워드 간 의미적 유사도를 제공하는 Word2Vec을 통해 의미적으로 벡터화를 진행해 둘을 결합하는 결합 벡터, 이른바 조인 벡터(Join Vector) 아이디어를 제공하고 있다.

박대서 외 1인의 연구에서 빈도와 의미 벡터 둘을 결

합하는 까닭은, 조회수 기반 뉴스 기사 추천 알고리즘이 보여주는 편향과 참여 의존의 한계를 극복하기 위한 방편으로 추천을 위한 기초 자료로서의 키워드를 유의미한 형태로 뽑아내기 위함이다. 그런데 본 연구가 대상으로 삼는 공모전 공고문의 특성 또한 TF-IDF의 한계점을 드러낸다. TF-IDF는 여러 문서 속 단어의 중요도를 문서 내 빈도와 전체 문서에서의 희소성을 바탕으로 계산하는데, 문서의 성격에 따라 빈도 기반 통계 지수의 특성상 변별력이 약화하는 현상이 나타난다. 공모전 공고문은 비교적 형식이 일정하고 길지 않으며 전형적인 안내 문구 등이 반복되는 특징이 있다. 최종적으로 키워드 간 순위를 매겨야 하는 것이 목적인 연구의 특성상 이는 근본적인 문제를 야기한다. 박대서의 연구가 인용한 이성익 외 1인의 연구(2009)는 TF-IDF의 변형식을 고안하여 의미 없는 키워드를 제거하고 있으나, 문서별이 아닌 전체 문서 집합에서 키워드를 추출하고 있다는 점에서 본 연구의 연도별 추세 확인이라는 목적과는 맞지 않았다.

결과적으로 본 연구는 박대서 외 1인의 조인 벡터 아이디어를 공모전 공고문 분석에 차용하여 개최 연도별 주요 키워드를 산출하되, 목적에 맞추어 분석 대상의 특징에 따라 정렬이 가능한 지수로 만들고 연도별 변화도를 측정하기 위한 방식을 개발하였다.

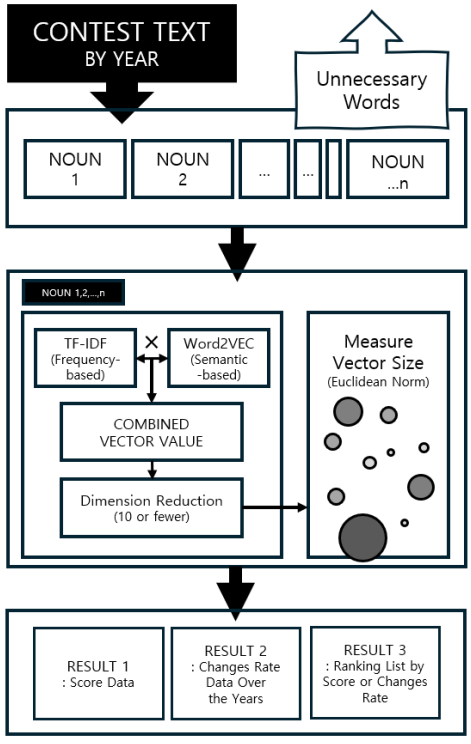
4. 공모전 분석 프레임워크

4.1 실험 환경

본 연구의 텍스트마이닝 기반 분석은 Python 3.10 환경에서 수행되었으며, 주요 라이브러리로는 scikit-learn 1.6.1, gensim 4.3.3, pandas 2.2.2, matplotlib 3.9.0 등이 사용되었다. 데이터 전처리 및 형태소 분석에는 KoNLPy와 Okt(Open Korean Text) 형태소 분석기가 활용되었고, Word2Vec 모델은 gensim의 Skip-gram 알고리즘으로 학습되었다. 실험은 Windows 10 운영체제를 사용하는 Intel i7-8700 CPU(3.20GHz), 32GB RAM의 사양을 갖춘 PC에서 수행되었다.

4.2 핵심 아이디어

본 연구에서 제안하는 공모전 분석 프레임워크(Contest Keyword Analysis Framework)는 ① 키워드 추출 ② 전처리 ③ 빈도/의미 벡터 분석 ④ 양 벡터의 결합 ⑤ 점수화 ⑥ 기타 정렬을 통한 순위화 및 해석 자료 제작의 단계로 구성된다. 그 중 핵심은 벡터의 결합(조인 벡터)이라는 아이디어를 통해 복합 지표를 단일 지표로 정량화하는 것이다. 여러 기준을 하나로 반영할 수 있다면 하나의 수치로 다양한 해석이 가능해지기 때문이다. 그 결과 이 프레임워크는 ① 키워드의 연도별 중요도를 나타내는 조인 벡터의 크기 ② 키워드의 연도별 점수의 변화 추이를 나타내는 변화도 ③ 전체 기간 혹은 특정 기간 내 점수와 변화율에 따른 순위 목록화를 산출한다.



[Fig. 1] Framework Structure Diagram

4.3 수집 및 전처리

별도 웹스크래핑을 통해 수집된 공모전 데이터를 전처리하고 키워드를 추출한다. 그 과정은 다음과 같다.

① 정규표현식을 통해 특수문자나 HTML 태그 등 쓸데 없는 정보를 제거한다.

② 형태소 분석을 통해 명사만 추출한다.

③ 약 700건으로 구축된 불용어 사전을 통해 불용어를 제거하고 한 글자 단어 등 의미 파악이 어려운 단어들을 지운다.

④ 키워드를 추출하기 위한 전 단계에서 전 기간의 모든 공모전 공고문을 한 덩어리로 묶지 않고 연도별

로 모아 묶는다. 이는 각 연도별로 중요한 이슈를 별도로 판단할 수 있게끔 하기 위함이다.

4.4 단어의 벡터화 / 차원축소 / 조인 벡터화

형태소 분석을 통해 추출된 문자열을 벡터화한다. 이 과정에서 빈도 기반인 TF-IDF와 의미론적 유사도 기반의 워드 임베딩 기법인 Word2Vec 알고리즘을 이용한다.

$$\mathrm{TFIDF}(t,d,D) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t,D)$$

[Eq. 1] TF-IDF formula

$$\mathrm{TF}(t,d) = \frac{f(t,d)}{\sum f(w,d)}$$

[Eq. 2] TF formula

$$\mathrm{IDF}(t) = \log\left(\frac{N}{1+|d \in D : t \in d|}\right)$$

[Eq. 3] IDF formula

TF-IDF는 특정 키워드가 문서 내에서 어느 정도 자주 나타나는지(TF, Term Fequency)와 해당 단어가 모든 문서 뭉치 안에서 얼마나 자주 나타나는지(DF, Document Frequency)의 역수(I : Inverse)를 조합한 수치이며 그 식은 [Eq. 1], [Eq. 2], [Eq. 3]과 같다. 식에서 t는 특정 단어(term), d는 특정 문서(document), D는 문서 뭉치를 뜻하며 N은 총 문서 수, $|d \in D : t \in d|$는 문서 d에 나타난 단어 t의 개수다[12]. TF-IDF는 특정 문서에서 자주 등장하지만 다른 문서에서는 보기 드문 단어를 더 중요하게 평가하고, 전체적으로 너무 흔하게 등장하는 관사나 조사 같은 단어들은 상대적으로 낮게 평가하는 경향을 보인다[13]. 즉 높은 TF-IDF는 통계적으로 해당 단어가 해당 문서에서 중요한 의미가 있다고 판단할 수 있다.

본 연구에서 TF-IDF와 Word2Vec으로 만들어진 의미 벡터를 결합하는 것은 단어 수준에서의 결합이며 그 식은 [Eq. 4]와 같다.

$$C_w = \mathrm{TFIDF}(w,d) \cdot V_w$$

[Eq. 4] Combined Vector Formula

결합된(Combined) 벡터 C_w를 얻기 위해서는 문서 d 내 단어 w의 TF-IDF 값을 Word2Vec에 학습시켜 나온 단어 w의 임베딩 벡터 V_w의 각 차원에 곱해준다. 즉 이 단계에서 의미 벡터의 차원 수는 유지된다. 이를 정량화하기 위해서는 PCA(Principal Component Analysis)로 차원을 축소하는 과정이 필요하다. 고차원인 벡터값을 파이썬의 scikit-learn 라이브러리 속 PCA 함수를 이용해 10차원까지 축소한 후, 이를 크기를 나타내는 숫자 하나로 만들었다. 여기에는 두 가지 방법이 있는데, 하나는 한 키워드에 해당하는 벡터값을 모두 일괄 합산하는 방식이고 다른 하나는 벡터의 거리를 재는 방식이다. 본 연구에서는 후자를 재는 방식으로 유클리드 노름(Euclidean Norm)이 선택되었다. 유클리드 노름은 벡터가 원점에서 얼마나 떨어져 있는지를 측정하는 방식이다.

$$\| V \|_2 = \sqrt{\sum_{i=1}^{n} v_i^2}$$

[Eq. 5] Euclidean Norm Formula

식 [Eq. 5]로 표시되는 유클리드 노름은 n차원 좌표평면(유클리드 공간)에서의 벡터 크기를 계산하여 붙은 이름이다. V가 벡터라 할 때, 벡터의 크기 또는 길이인 $\| V \|_2$를 벡터의 i번째 성분 v_i를 제곱한 값을 모두 더한 후 제곱근으로 만들어 표시한다. 이렇게 하여 만들어진 숫자가 원점에서 해당 벡터가 위치한 점까지의 직선거리라 할 수 있다[14]. 이 크기가 조인 벡터의 크기이자 빈도와 의미를 반영한 점수다.

5. 측정

5.1 평균절대편차를 통한 수치 변별력 비교

빈도와 의미를 반영한 키워드별 조인 벡터의 크기가 중요도 측정에 흔히 쓰이는 TF-IDF에 비해 나은 결과를 내는지를 확인하기 위해 양 값의 평균절대편차(Mean Absolute Deviation, MAD)를 확인한 표는 〈Table 1〉과 같다.

〈Table 1〉 MAD Value

Year	TF-IDF MAD	Join Vector MAD
2012	0.154289352	0.011689827
2013	0.10254205	0.009562694
2014	0.183720196	0.012045199
2015	0.417433023	0.125124646
2016	0.506777329	0.506397787
2017	0.579003922	0.738336696

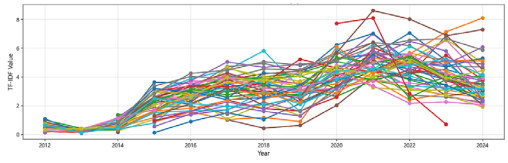
2018	0.60939997	1.155911563
2019	0.582085596	0.55798226
2020	0.723372049	1.967191448
2021	0.821545591	2.48815499
2022	0.8231687	2.271316023
2023	0.752781382	1.212537112
2024	0.667871571	0.702105035

편차는 각 데이터 값에서 해당 데이터의 평균을 뺀 값으로 데이터 값과 평균의 차이를 뜻하며, 절대편차는 편차의 절대값으로 데이터값이 평균에서 얼마나 떨어져 있는지를 양수로 나타내 평균에서의 거리만을 측정한다. 평균절대편차는 모든 데이터값의 절대편차를 평균 낸 값으로 그 식은 [Eq. 6]과 같다.

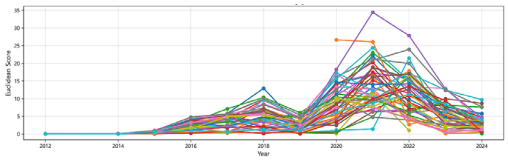
$$\mathrm{MAD}= \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

[Eq. 6] MAD formula

평균절대편차의 식에서 n은 데이터의 개수, i는 데이터 항목의 인덱스다. x_i는 i번째 데이터 값, $\bar{x}$은 데이터의 평균이다. 평균절대편차는 데이터 값들이 평균으로부터 얼마나 멀리 떨어져 있는지를 뜻하며, 따라서 숫자가 큰 쪽의 열은 각 데이터값 사이의 간격(흩어진 정도)가 더 크다고 할 수 있다. E.A.H.Elamir(2012)의 연구에 따르면 평균 절대 편차는 데이터 분포의 스케일만이 아니라 형태를 분석하는 데에 유용한 지표이며, 표준편차보다 이상치(outlier)에 덜 민감해 데이터의 변동성을 더 직관적으로 반영한다[15]. 이에 따르면 조인 벡터를 통해 측정한 연도별 공모전 데이터는 개최 건수가 낮은 일부 연도를 빼면 2020년부터는 TF-IDF를 압도하기 시작한다. 평균치의 합은 조인 벡터가 TF-IDF 단일 측정의 2.408배를 나타냈다. 즉 조인 벡터를 활용한 공모전 데이터 분석은 TF-IDF값만을 이용한 것보다 변별력이 좋은 결과를 얻을 수 있다고 볼 수 있다. 이는 연도별 TF-IDF과 조인 벡터의 상위 50위권으로 그린 그래프에서도 나타난다.



[Fig. 2] TF-IDF value by year(descending)



[Fig. 3] Join Vector value by year(descending)

[Fig. 2]와 [Fig. 3]의 X축은 연도, Y축은 값인데 [Fig. 3]의 Y축이 훨씬 더 많은 수를 보여주고 있음을 알 수 있다. 값 사이의 변별력이 낮으면 순위를 매기는 행위의 의미가 퇴색된다.

〈Table 2〉 Comparing Deduplication Rate

Division	Total	Remove	Remain
TF-IDF	16,801 (100%)	4,357 (25.94%)	12,444 (74.06%)
Join Vector	4,561 (100%)	0 (0%)	4,561 (100%)

〈Table 2〉는 TF-IDF와 조인 벡터의 전체 기간 평균치를 통해 겹치는 값이 몇이 나오는가를 정리한 표다. TF-IDF에서는 25% 가량이 중복값으로 판명되어 삭제된 반면, 최소 등장 횟수(5회)를 충족하여 벡터화된 이후의 키워드는 한 값도 중복되지 않아 순위를 원활하게 매길 수 있었다.

5.2 키워드별 평균 변화율 비교

$$ChangeRate_t = \frac{CurrentJV_t - PrevJV_t}{PrevJV_t}$$

[Eq. 7] Change Rate formula

[Eq. 7]은 변화율을 구하는 식이다. 키워드 t의 변화율 $ChangeRate_t$는 t의 현재 조인 벡터값 $CurrentJV_t$에서 t가 앞에서 출현했을 때의 조인 벡터값 $PrevJV_t$을 뺀 값을 다시 $PrevJV_t$로 나누어 구한다. 조인 벡터의 평균 변화율은 이렇게 산출한 시기별 조인 벡터의 변화율을 키워드별로 전체 기간 평균을 내 구한다.

〈Table 3〉 TF-IDF Change Rate
(20 Rank, sort by Weighted_Average)

rank	Keyword	AverageValue	Yearcount	WeightedAverage
1	Money	5.841408	5	10.4664
2	Galsses	6.297526	4	10.13548
3	Disconnected	5.754644	4	9.261743

4	Theater	6.289551	3	8.719169
5	force	4.040522	6	7.862492
6	Death Penalty	6.139066	2	6.744454
7	SEJIN	5.958127	2	6.545672
8	rest	5.935805	2	6.521149
9	deductions	4.608049	3	6.388112
10	electronics	5.560853	2	6.109222
11	security	2.601009	9	5.989043
12	EunPeyong-Gu	2.553606	8	5.610846
13	Hwa-bi	5.100429	2	5.603393
14	SooWon station	3.991326	3	5.533153
15	Alcohol	5.003343	2	5.496734
16	type	2.604946	7	5.416833
17	democratization	2.459375	8	5.403799
18	META	3.002084	5	5.379012
19	mirror	3.002084	5	5.379012
20	culrual center	2.4670006	7	5.129983

〈Table 4〉 Join Vector Change Rate
(20 Rank, sort by Weighted_Average)

rank	Keyword	AverageValue	Year count	WeightedAverage
1	GyeongGi	19.24975	12	49.37464
2	InCheon	19.13398	11	47.54616
3	tourism	15.46097	13	40.8024
4	forests	17.95412	7	37.33453
5	overcome	13.58801	11	33.76493
6	point	13.12216	12	33.65766
7	Sansam	28.91705	2	31.76862
8	DaeGoo	12.51386	11	31.09577
9	BooCheon	10.9456	12	28.07492
10	city	10.42648	12	26.74338
11	period	10.3128	12	26.4518
12	chairman	9.43644	11	23.44867
13	affiliation	8.730305	11	21.69399
14	fantasdy	8.883636	10	21.30203
15	rural district	8.1282	12	20.84842
16	value	7.802823	13	20.5921
17	equipment	8.071322	11	20.05648
18	nuclear fusion	16.64986	2	18.29174
19	tourist attraction	7.723089	9	17.78307
20	news paper	6.4436353	13	17.005123

2012년부터 13개년 전체에 걸친 중요 키워드의 TF-IDF와 조인 벡터 평균 변화율을 내림차순 정렬한 표는 각기 〈Tabel 3〉 〈Table 4〉와 같다. 이 표에는 각기 조인 벡터로 산출한 변화율의 평균값과 출현 연도 수에 따라 가중치를 부여한 가중 평균 변화율이 기재돼 있다.

$$Weighted\,Average = ChangeRateAverage \times log(YearCount + 1)$$

[Eq. 8] Change Rate's Weight Average Formula

[Eq. 8]은 [Eq. 7]에서 계산된 키워드의 변화율의 전체 평균 *ChangeRateAverage*에 출현 연도 수에 따른 가중치를 로그값으로 부여하여 가중 평균값인 *WeightedAverage*를 산출하는 식이다. 이 알고리즘은 연도 카운트가 클수록 더 높은 가중치를 부여하는 방식이다. 그 결과에 따르면 조인 벡터 랭킹이 TF-IDF 랭킹에 비해 수치 면에서 변별력이 있음을 알 수 있다. 특히 TF-IDF는 전체 20위권이라는 좁은 범위 안에서도 '메타'와 '거울'이 동일하게 나온다. 이상의 수치 데이터는 공모전의 전체 분석에는 조인 벡터의 활용성이 좋다는 것을 보여준다. 내용 면에서도 조인 벡터는 TF-IDF에 비해 지역 또는 산림 등 주최 측의 성격이 확연히 드러나는 키워드가 상위권에 넓게 분포되는 결과를 보여주며, 달리 해석될 여지가 없는 키워드들의 비중이 더 높음을 볼 수 있다. 변화율은 해당 시기의 키워드별 조인 벡터가 보여주는 변동폭을 통해 반응도의 추세를 판단하는 지표이므로, ① 계속해서 큰 변동이 없는 키워드 ② 큰 변동폭을 유지하며 상승세를 보이는 키워드 ③ 큰 변동폭을 보이며 하락세를 보이는 키워드를 파악할 수 있다. 이를 평균 낸 평균 변화율은 자주 등장하는 수에 구애받기보다 산출된 값들의 추세에 영향을 받으며, 이 값이 크고 작음으로 종합적인 견지에서의 추세를 판단할 수 있다. 즉 이 수치를 기준으로 오름차순 정렬하면 변화율이 전체적으로 꺾이고 있는 키워드를 확인할 수 있다. 공모전 기획자 입장에서 평균 변화율이 높은 키워드는 상승 추세를 보인 키워드로 채택 대상에 놓을 수 있지만 반대로 평균 변화율이 낮은 키워드는 하락세를 보인 키워드로 재고 대상에 놓을 수 있다.

6. 결론

본 연구는 연도별 추세 파악이 중요한 판단 기준이 되는 공모전을 소재로 유의미한 지표를 산출하는 분석 프레임워크의 구성을 목적으로 했다. 키워드 중요도 순위에 많이 쓰이는 빈도 기반의 TF-IDF 지표를 넘어 좀 더 정확한 결과를 산출하기 위한 정량화된 점수를 개발하기 위해, TF-IDF와 의미 기반인 Word2Vec을 결합해 조인 벡터를 산출하였으며, 이를 통한 전 기간 평균과 출현 연

도 수에 따른 가중치를 부여함으로써 기간별 중요도와 전체 기간 단위의 중요도를 모두 반영한 복합 지표를 만들어냈다. 이 복합 지표는 단순 TF-IDF 기반 순위와는 확연한 질적 차이를 보였다.

다만 본 연구는 데이터를 바탕으로 과거를 분석하는 데 유효한 지표의 개발까지를 진행하여, 예측 모델 알고리즘에 따른 복합 예측치를 산출하는 데에까지는 이르지 못하였다. 이후 연구를 통해 예측과 더불어 핵심 주제를 반영한 분류(Classification)와 상관관계 분석이 가능한 프레임워크 시스템을 개발하고자 한다.

REFERENCES

[1] Thinkgood, https://thinkcontest.com/thinkgood/user/lab/strategy-statistics.do?mode=view&articleNo=62503

[2] Nice Economy, https://www.niceeconomy.co.kr/news/articleView.html?idxno=82869

[3] HanKyeongRecruit, http://hkrecruit.co.kr/news/articleView.html?idxno=13357

[4] ChosunMedia A Better Future, https://futurechosun.com/archives/7502

[5] M.J.SHIN, "Experiences inside and outside of art school, and what it means to support art students in the public sphere", Art in the City×Trying 2019~2021 Urban Field Program for Young Artist Discourse Collection, Seoul Cultural Foundation, p97, 2021.

[6] Thinkgood, https://www.thinkcontest.com/

[7] Contestkorea, https://contestkorea.com/

[8] Wevity, https://www.wevity.com/

[9] Thinkyou, https://thinkyou.co.kr/

[10] G.G.SA, "A Study on the Keyword Change of Architecture Competition Using Text Mining", Journal of KICA, No.85, pp218-226, 2024.

[11] D.S.Park, H.J.Kim, "A Proposal of Join Vector for Semantic Factor Reflection in TF-IDF Based Keyword Extraction", Journal of KIIT. Vol. 16, No. 2, pp. 1-16, 2018.

[12] C.D.Manning, P.Raghavan, H.Schütze, "Introduction to Information Retrieval", Cambridge University Press, pp100-123, 2008.

[13] B.S.Desikan, "Natural Language Processing and Computational Linguistics", Acorn(Korean Translated), 2019.

[14] M.E.Celebi, F.Celiker, H.A.Kingravi, "On Euclidean Norm Approximations", Pattern Recognition, Vol.44, Issue 2, 2011.

[15] E.A.H.Elamir, "Mean Absolute Deviation about Median as a Tool of Explanatory Data Analysis", Proceedings of the World Congress on Engineering, Vol I, pp324-329, 2012.

임 채 진(Chae-Jin LIM) [정회원]

- 2004년 8월 : 백석대학교 정보통신학부 컴퓨터학 전공 (공학사)
- 2014년 2월 : 성공회대학교 문화대학원 미디어.문화연구전공 (문학석사)
- 2025년 3월 ~ : 백석대학교 일반대학원 소프트웨어융합 전공(박사과정생)

〈관심분야〉
콘텐트-정보통신 융합, 인공지능

한 정 수(Jung-Soo HAN) [정회원]

- 1992년 8월 : 경희대학교 컴퓨터공학부(공학석사)
- 2000년 8월 : 경희대학교 대학원 컴퓨터공학부(공학박사)
- 2001년 3월 ~ 현재 : 백석대학교 컴퓨터공학부 교수

〈관심분야〉
AI 교육, 자율주행, 데이터 분석, SW 모델링

이 현 섭(Hyun-Seob LEE) [종신회원]

- 2007년 2월 : 한양대학교 컴퓨터공학과 (공학 석사)
- 2013년 2월 : 한양대학교 컴퓨터공학과 (공학 박사)
- 2012년 3월 ~ 2021년 2월 : 삼성전자 책임연구원
- 2021년 3월 ~ 현재 : 백석대학교 컴퓨터공학부 조교수

〈관심분야〉
인공지능, 저장시스템, 임베디드 시스템