Journal of KIIT. Vol. 19, No. 11, pp. 115-124, Nov. 30, 2021. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2021.19.11.115

대학 평판도 분석을 위한 인스타그램 데이터 타당성 연구

정민수[1]*, 김장원*[2], 온병원*[3], 정동원*[4]

Instagram Data Feasibility Study for Analysis of University Reputation

Minsu Jung*[1], Jangwon Gim*[2], Byung-Won On*[3], and Dongwon Jeong*[4]

이 연구는 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임
(NRF-2019R1I1A3A01060826)

요 약

기존의 대학 평판도 평가 연구는 온라인 뉴스 기사를 활용하였다. 그러나 온라인 뉴스 기사는 주관적인 의견이 반영되지 않고 특정 대학의 기사가 다양한 대학을 언급하는 한계가 존재한다. 이러한 문제점을 해결하기 위해 인스타그램의 데이터 수집과 타당성 검증이 필요하다. 본 논문에서는 대학 평판도 분석을 위해 수집한 인스타그램 데이터의 타당성을 분석한다. 기존 감성 사전을 확장하고 감성 어휘를 고려한 페이지랭크 방법을 이용하여 허브 노드를 추출하여 허브 노드를 기점으로 수집한 데이터를 분석하였다. 분석결과에서, 인스타그램은 낮은 감성 해시태그 비율을 보였다. 대학 평판도 분석을 위한 수집 매체로 적합하지 않음을 확인하였다. 이 연구는 인스타그램을 활용할 연구 분야에서 사전 자료로 활용될 수 있다.

Abstract

Existing university's reputation assessment study utilized online news articles. However, online news articles has limits that do not reflect subjective opinions and articles of certain universities refer to various universities. To solve these problems, data collection and validation of Instagram are needed. In this article, we analyze the feasibility of Instagram data collected for university reputation analysis. This work extended the existing emotional dictionary and analyzed the data collected from the hub node using the pagerank method considering the emotional vocabulary. The analysis result shows that, Instagram is not proper data collection media for university's reputation analysis. It showed a low rate of emotional hashtags. This study can be used as basic data in the field of research to utilize Instagram.

Keywords
instagram hashtag, emotional vocabulary, university reputation

* 군산대학교 소프트웨어융합공학과(*[3,4] 교신저자)
- ORCID[1]: https://orcid.org/0000-0003-4704-6279
- ORCID[2]: https://orcid.org/0000-0002-4480-7944
- ORCID[3]: https://orcid.org/0000-0001-6929-3188
- ORCID[4]: https://orcid.org/0000-0001-9881-5336

· Received: Jun. 30, 2021, Revised: Oct. 06, 2021, Accepted: Oct. 09, 2021
· Corresponding Authors: Byung-Won On and Dongwon Jeong
Dept. of Software Convergence Engineering, Kunsan National University,
558, Deahak-ro, Gunsan, Jeollabuk-do, Korea,
Tel.: +82-63-469-8911, Email: {bwon, djeong}@kunsan.ac.kr

Ⅰ. 서 론

대학 평판도는 기업이나 기관에서 대학과 관련된 요인을 분석하여 평가한 결과이다. 입시를 준비하는 수험생 및 보호자와 취업을 준비하는 졸업예정자는 대학 평판도에 민감한 반응을 보인다. 대학 신입생들은 대학선택 요인으로 대학의 명성, 취업률, 경제적 지원을 고려하였다. 이중 가장 큰 요인은 대학의 명성이었으며 대학의 홍보정책이나 교육프로그램에 대해서 주목하지 않았다. 또한, 대학진학의 이유로 취업의 중간 과정으로만 생각하여 대학의 이름과 성장 가능성만을 고려했다[1]. 따라서 대학 평판도는 단순히 대학 평가를 넘어 이를 참고하여 활용함으로써 신뢰성 있는 분석결과가 필요하다.

대학 평판도 분석을 위해 사용하는 방법은 설문조사를 활용한 평가 방법과 자연어처리를 통한 방법으로 분류할 수 있다. 설문조사를 활용한 대학 평판도 평가는 영국의 평가기관 QS(Quacquarelli Symonds)에서 진행하는 QS 세계 대학 랭킹과 중앙일보의 대학 평판도 평가가 존재한다[2][3]. QS는 대학 평판도 평가요인으로 연구성과 등의 정량지표를 참고하지만, 설문조사를 활용한 학교평가, 기업으로부터의 평판에 대한 지표를 50%로 설정하였다[2]. 중앙일보 대학 평판도 평가는 매년 다양한 지표에 대한 설문조사를 시행하여 이를 기반으로 대학평판도 평가를 시행하였다[3]. 그러나 설문조사를 활용한 평가 방법은 응답자의 통제가 어려우므로 설문 내용에 불성실하게 응할 가능성이 존재하며 표본의 크기가 제한적이다. 따라서 설문조사 결과의 신뢰성에 문제를 제기할 수 있다.

설문조사를 활용한 대학 평판도 평가 방법의 문제해결을 위해 자연어처리를 활용하여 대학 평판도를 분석하는 연구가 수행되었다[4][5]. [4]는 수집한 온라인 뉴스 데이터를 기반으로 AR-KNU(Academic Reputation-KNU) 감성사전을 구축하였다. 군집화 알고리즘을 이용하여 대학에 관한 주제를 추출하고 군집명을 부여하였다. 군집명을 기반으로 문단을 학습하여 대학의 긍정·부정 비율을 통해 평판도를 분석하였다. 그러나 적은 데이터 크기와 대학별 데이터 편차로 인해 긍정·부정의 편향이 발생하였다.

[5]는 [4]에서 수집한 데이터의 수집 기간 확장을 통해 더 많은 감성 어휘를 수집하였으며, AR-KNU에 감성 어휘를 추가하여 AR-KNU+를 구축하였다[6]. 구축한 감성사전을 기반으로 문단을 학습하여 대학의 긍정·부정 비율을 통해 평판도를 분석하였다. 데이터로 사용한 온라인 뉴스 기사는 객관성을 전제로 보도하므로 높은 신뢰성을 가진다[7]. 그러나 특정 대학만 언급하는 것이 아니라 다양한 대학이 언급되는 문제점이 존재한다. 또한, 객관적인 보도를 위해 주관적인 감정을 최대한 배제하여 개인의 의견이 반영되지 않는 한계점이 존재한다. 따라서 해시태그를 통해 특정 단어를 강조할 수 있으며 주관적인 게시물을 공유하는 온라인 소셜 네트워크 서비스(SNS, Soical Network Service)의 데이터 수집이 필요하다.

SNS 중 인스타그램(Instagram)은 대표적인 소셜 네트워크로서 20, 30대가 가장 많이 이용하는 매체이다[8]. SNS 사용자는 해시태그로 다양한 문구를 작성하여 게시물을 설명하고 감정을 표현한다. 해시태그를 이용하여 사용자의 감성 분석 연구가 진행되고 있다[9][10]. SNS에서 감성 분석을 극대화하기 위해서는 네트워크 구조에서 가장 파급력 있는 노드인 허브 노드를 추출할 필요가 있다.

허브 노드를 추출하는 기법의 하나인 페이지랭크(Pagerank)를 활용한 연구가 진행되고 있다[11][12]. 페이지랭크는 하이퍼링크 구조를 가지는 문서에 상대적인 중요도에 따라 가중치를 부여하여 페이지의 우선순위를 계산한다[13]. 따라서 네트워크 분석 시 한 노드가 네트워크에 임의로 중요도 조작을 할 수 없는 장점을 갖고 있다. 그러나 하이퍼링크만을 고려하기 때문에 감성 어휘를 포함한 게시물 수집 시 감성 어휘가 적은 페이지가 높은 우선순위를 차지할 수 있는 문제점이 발생한다. 이를 해결하기 위해 감성 어휘를 고려한 페이지랭크 기법이 필요하다.

본 논문에서는 대학 평판도 분석을 위한 데이터 수집 매체로 인스타그램을 선정하고 데이터 타당성을 분석한다. 수집한 데이터의 해시태그를 감성 분석하여 기존 감성사전인 AR-KNU+를 확장한다. 확장한 감성사전을 기반으로 감성 게시물 분류 후 페이지랭크를 위한 그래프를 구축한다. 구축한 그래프를 기반으로 허브 노드를 추출하여 허브 노드를 기점으로 한 데이터 수집을 진행한다.

이 논문의 구성은 다음과 같다. 제2장은 관련 연구를 소개하고, 제3장은 제안하는 방법을 설명하고 제4장은 실험을 수행하고, 결과를 서술한다. 제5장은 결론을 서술한다.

II. 관련 연구

이 장에서는 대학평판도 분석에 관한 연구에 관해 기술하고 기존 연구의 문제점을 통해 SNS 데이터 수집의 필요성을 제시한다.

2.1 뉴스 데이터 기반의 대학 평판도 평가

[4]는 대학 평판도 평가를 위해 온라인 뉴스 데이터를 활용했다. 수집한 온라인 뉴스 데이터를 기반으로 AR-KNU 감성사전을 구축하였다. 군집화 알고리즘을 이용하여 대학에 관한 주제를 추출하고 군집명을 부여하였다. 군집명을 기반으로 문단을 학습하여 대학의 긍정·부정 비율을 통해 평판도를 분석하였다. 그러나 대학별 데이터 크기 차이로 인해 긍정·부정의 편향이 발생하였으며 특정 대학의 기사에서 다양한 대학이 언급되는 문제점이 존재하였다.

[5]는 대학 평판도 평가를 위해 온라인 뉴스 데이터를 활용했다. 데이터 수집 기간의 확장을 통해 더 많은 감성 어휘를 수집하였으며, AR-KNU에 감성 어휘를 추가하여 AR-KNU+를 구축하였다. 감성 사전을 기반으로 문단을 학습하여 대학의 긍정·부정 비율을 통해 평판도를 분석하였다. 그러나 온라인 기사는 특정 대학만 언급하는 것이 아니라 다양한 대학이 언급되는 문제점이 존재하였다.

[14]는 대학 평판도 평가를 위해 대학알리미에서 제공하는 대학별 신입생 충원율, 재학생 충원율, 취업률과 온라인 뉴스 데이터를 수집하여 연구를 수행하였다. 수집한 기사로 형태소 분석을 수행하였고 빈도수가 높은 단어를 추출하여 감성 사전을 이용한 감성 분석을 수행하였다. 그 결과, 정량지표를 통해 신뢰성을 높일 수 있었지만, 특정 대학의 기사에서 다양한 대학이 언급되는 문제점이 존재하였다.

뉴스 데이터를 통한 대학 평판도 평가는 적은 데이터 크기로 인한 편향과 특정 대학의 기사가 아닌 다양한 대학을 언급하는 문제점이 존재한다.

2.2 SNS 데이터 기반의 대학 평판도 평가

[15]는 특수목적 대학교들의 이미지 분석을 위해 트위터, 페이스북, 블로그 등 SNS에서 언급된 데이터를 수집하여 연구를 수행하였다. 빅데이터 분석 프로그램을 사용하여 연관어를 기반으로 시간의 흐름에 따른 이미지 결과를 분석하였다. 그 결과, SNS 데이터를 사용하는 연령층과 단기간에 형성된 대학에 관한 여론을 파악할 수 있었다. 적은 데이터 크기로 간접적인 평판 추측이 가능했지만 통계 분석을 통해 연관어의 정확한 감성을 도출하기 어렵다는 한계가 있다.

본 논문에서는 뉴스 데이터 기반의 대학 평판도 평가 연구의 문제점 해결을 위해 SNS 데이터를 수집한다. 감성 분석을 통하여 [6]에서 구축한 AR-KNU+ 확장한다. 확장한 감성 사전을 통해 감성이 포함된 게시물만을 이용하여 방향 그래프를 구축한다. 구축한 방향 그래프를 이용하여 페이지랭크를 통해 허브 노드를 추출한다. 마지막으로, 추출한 허브 노드를 기점으로 감성 분석을 수행해가며 대학과 관련된 데이터를 수집한다.

III. 제안 방안

기존 연구의 문제점을 해결하기 위해 주관적인 게시물을 공유하는 SNS의 데이터 수집을 제안한다. 기존 연구에서는 대학 평판도 분석을 위해 온라인 뉴스 데이터를 수집하였다. 온라인 뉴스 데이터는 객관적인 내용 전달을 위해 감정을 최대한 배제하기 때문에 감성 어휘의 비율이 낮고 개인의 의견이 반영되지 않는 문제점을 보였다. 그러므로 20~30대가 가장 많이 사용하는 인스타그램을 수집매체로 선정한다.

감성 어휘의 포함 비율이 높은 허브 노드를 추출하기 위해 그래프 구축 시 감성 게시물을 이용하는 방법을 제안한다. 기존 페이지랭크 기법은 하이퍼링크만을 고려하여 그래프를 구축한다. 감성 어휘가 적은 노드가 허브 노드로 선정되는 문제점이 발생한다. 그러므로 감성사전을 이용하여 감성 게시물을 분류하고 그래프 구축데이터로 사용한다.

그림 1은 제안 방법의 전체 과정을 나타내며 인스타그램 데이터 수집, 데이터 전처리, 그래프 구축, 허브 노드 추출, N-Hop 데이터 수집으로 구성한다.

첫 번째로, 인스타그램의 해시태그를 이용하여 대학과 관련 있는 게시물을 수집한다. 두 번째로, SNS 맞춤형 감성사전 구축을 위해 수집한 데이터를 감성 분석하여 감성 어휘를 도출한다. 도출한 감성 어휘를 추가하여 기존의 감성사전인 AR-KNU+를 확장한다. 세 번째로, 감성 어휘를 고려한 페이지랭크 방식을 사용하기 위해 감성 게시물의 작성 정보를 이용하여 방향 그래프를 구축한다. 네 번째로, 구축한 방향 그래프를 기반으로 페이지랭크 방식을 이용하여 네트워크에서 상위의 가중치를 가지는 노드를 추출한다. 마지막으로. 추출한 허브 노드를 기점으로 대학과 관련된 감성 게시물을 수집한다.

3.1 인스타그램 데이터 수집방안

이 절에서는 대학 평판도 분석을 위한 인스타그램의 데이터 수집방법을 제안한다. 이를 위해 인스타그램의 해시태그를 이용한 크롤러를 구축한다.

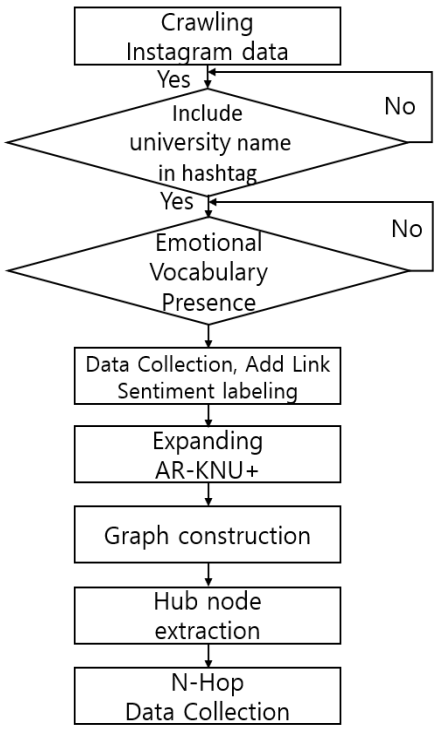

그림 1. 제안 방안의 흐름도
Fig. 1. Flow chart of proposed method

크롤러는 사용자가 수집을 원하는 대학명을 입력할 시 입력한 대학명을 해시태그로 하는 게시물을 검색하여 수집한다. 이때 입력한 대학의 이름은 약어로 사용하는 경우가 존재한다. 그러나 약어로 표현되는 대학명의 경우 다른 단어의 뜻을 가지는 경우가 존재한다. 예를 들어 연세대학교를 연대라고 표현하는 경우이다. 연대는 행동을 같이하는 뜻이 있다.

그러므로 해시태그를 이용한 대학명 검색 시 약어를 사용하지 않고 전체 대학명을 사용한다. 수집을 진행하는 방식은 인스타그램이 제공해준 알고리즘을 통해 다음 게시물로 이동해가며 게시물을 수집한다.

3.2 데이터 전처리

이 절에서는 수집한 인스타그램 데이터를 형태소 분석하여 품사를 구분하고 감성 평가를 통해 감성 어휘를 추출하여 감성사전을 확장한다. 수집한 인스타그램의 게시물의 해시태그를 대상으로 기존 감성사전인 AR-KNU를 이용하여 감성 게시물을 분류한다. 분류한 감성 게시물의 해시태그 및 댓글을 대상으로 네이버 맞춤법 검사를 수행하여 오타를 교정한다.

맞춤법 검사를 수행한 해시태그와 댓글을 형태소 분석기인 MeCab을 사용하여 명사, 동사, 형용사, 부사, 감탄사, 사전에 정의되어 있지 않은 단어로 분류한다. 형태소 분석을 완료한 단어를 감성 분석하여 기존 감성사전에 포함되지 않는 단어를 추가한다.

표 1은 감성 어휘의 감성 정도를 매기는 기준이다. 세 명의 평가자는 분류한 어휘를 표1을 기준으로 감성 평가하여 과반수에 해당하는 점수를 부여한다. 만일 평가자의 점수가 각각 다르다면 재평가를 통해 점수를 부여한다.

표 1. 감성에 따른 감성 정도의 값
Table 1. Sentiment degree by sentiment

Very neg	Neg	Obj	Pos	Very pos
-2	-1	0	1	2

3.3 그래프 구축

이 절에서는 감성 어휘 비율이 높은 허브 노드를 추출하기 위해 감성 어휘를 포함한 게시물을 이용하여 방향 그래프를 구축한다. 이를 위해 확장한 감성사전을 이용하여 게시물의 해시태그에 대학명과 감성 어휘가 포함된 게시물을 분류한다. 분류한 감성 게시물을 NetworkX 패키지를 이용하여 방향 그래프를 구축한다.

그래프 구축 방식은 그림 2와 같이 진행된다. 수집한 감성 게시물을 대상으로 게시물을 작성한 계정의 URL에서 댓글을 작성한 계정의 URL로 이동할 수 있다는 가정하에 진행한다. 또한, 댓글을 작성한 계정이 게시물을 작성한 계정의 URL로 이동할 수 있는 것을 배제하였다.

그림 3은 그림 2를 기반으로 그래프를 구축한 예시이다. 구축한 그래프는 n개의 깊이를 가지며 인접 노드를 통하여 특정 노드에 접근할 수 있다. 또한, 허브 노드 및 인접 노드에 링크가 없어 네트워크가 분리된 경우도 존재한다.

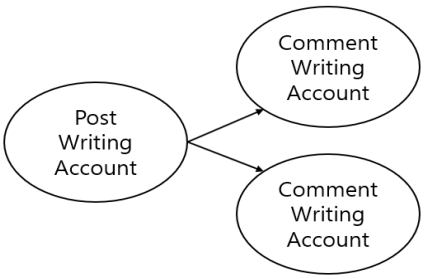

그림 2. 제안 방안을 위한 방향 그래프
Fig. 2. Direction graph for proposed method

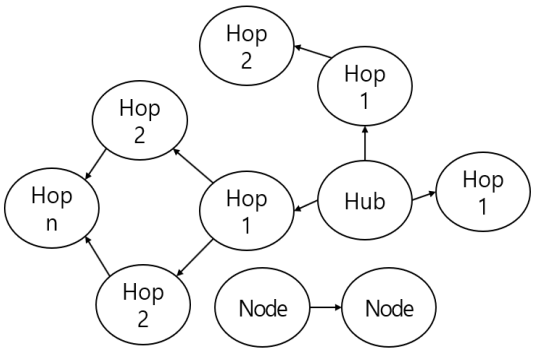

그림 3. 제안 방안의 방향 그래프 예시
Fig. 3. Example of a direction graph of the proposed method

3.4 허브 노드 추출

이 절에서는 허브 노드를 추출하기 위해 구축한 방향 그래프를 이용한다. 구축한 방향 그래프를 NetworkX 패키지를 이용하여 페이지랭크를 구현한다. 페이지랭크의 결과로 상위의 가중치를 가진 계정을 허브 노드로 추출한다.

허브 노드 추출 시 단일 노드가 아닌 다수의 노드를 대상으로 한다. 이를 통해 허브 노드를 기점으로 하는 데이터 수집 시 광범위한 네트워크를 구축하도록 한다.

식 (1)은 페이지 A의 페이지랭크를 계산하는 알고리즘이다. 주변 페이지의 가중치를 이용하여 계산하며 d는 다른 페이지로 가는 링크를 클릭할 확률의 값으로 0.85로 지정하여 85%의 확률로 다른 페이지를 클릭하도록 했다. 식의 결과는 페이지 A로 접속하는 확률을 얻을 수 있다.

$$PR(A) = (1-d)/N + d(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) \quad (1)$$

3.5 N-Hop 데이터 수집

이 절에서는 네트워크에서 영향력 있는 노드인 허브 노드를 기점으로 인접한 노드로 확산해가며 데이터를 수집한다. 이를 위해 그래프 탐색 방식으로 BFS(Breadth First Search) 기법을 사용하여 기준 노드를 시작으로 인접한 모든 노드를 탐색한다.

그림 4는 허브 노드를 기점으로 데이터를 수집하는 크롤러의 프로세스다. 사용자는 데이터 수집을 위해 대학명, 연도, 허브 노드의 링크를 입력한다. 입력한 허브 노드의 링크로 이동하여 게시물을 수집한다. 게시물 수집 시 입력한 연도와 비교 후 해시태그에 대학명의 포함 여부를 비교한다. 연도와 대학명의 조건에 충족하는 게시물을 대상으로 확장한 감성사전을 이용하여 감성 어휘를 분석한다. 해시태그 및 댓글에 감성 어휘를 포함할 시 해당 게시물의 해시태그, 댓글, 댓글을 작성한 계정 링크를 수집한다. 해당 허브 노드와 관련된 계정 링크가 존재하지 않을 때까지 반복한다.

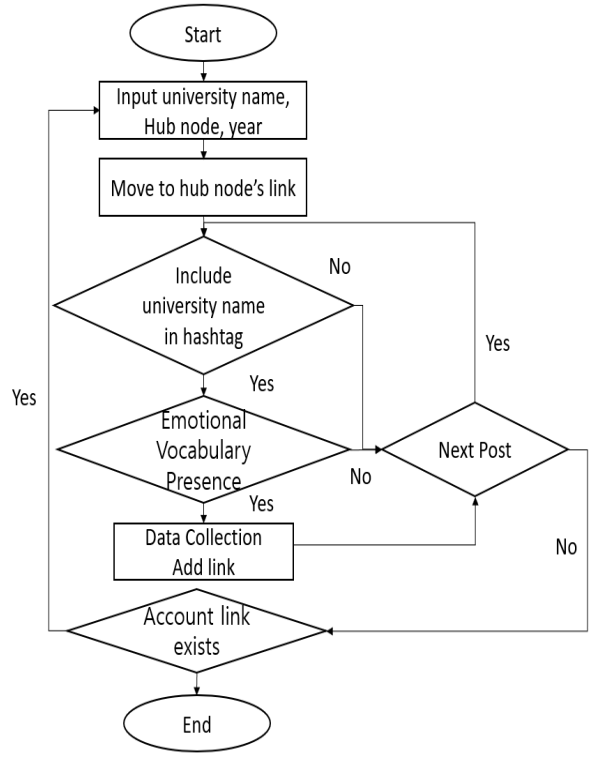

그림 4. 크롤러 프로세스
Fig. 4. Crawler process

IV. 실 험

4.1 실험 환경

실험 환경은 표 2와 같다. 크롤러 구현을 위한 개발 언어는 Python을 사용하였으며, 라이브러리는 Selenium을 사용한다. 크롬 드라이버를 사용하여 크롬을 통한 데이터 수집 환경을 구축하였다.

표 2. 실험 환경
Table 2. Experimental environment

Feature	Specification
CPU	Intel(R) Xeon(R) CPU E3-1270 v5 3.60GHz
RAM	16GB
SSD	128GB
HDD	931GB
GPU	NVIDIA Quadro K620
OS	Windows 10
Crawler	Selenium
Tools	Python
Browser	Chrome

4.2 수집 데이터

표 2는 실험을 수행할 대학의 인스타그램 게시물 수집개수이다. U1, U2, U3, U4, U5는 대학명을 익명으로 나타내었다. 데이터 수집 기간은 2020년 12월부터 2021년 4월이다. 그러나 U5는 특정 기간 매우 적은 양의 데이터 개수를 보여 수집 기간을 늘렸다. 수집 기간은 2019년 8월부터 2020년 8월이다. 감성 어휘를 포함한 게시물의 개수가 아닌 대학명을 해시태그로 포함하는 게시물의 개수이다.

4.3 감성 사전 확장

표 3은 표 2의 대학의 데이터를 맞춤법 검사와 형태소 분석기를 이용하여 분류한 감성 게시물을 세 명의 평가자가 매우 부정, 부정, 감정 없음, 긍정, 매우 긍정으로 단어 감성 평가를 완료한 결과를 보여준다. 감정 없음을 제외한 단어를 AR-KNU+에 추가하여 확장하였다. 그러나 73,937개의 단어 중 765개를 추가하여 낮은 빈도의 감정 어휘를 보였다.

4.4 허브 노드 개수 별 감성 어휘 비율

효율적인 감성 게시물 수집을 위해 허브 노드 개수에 따른 게시물의 감성 어휘 비율을 분석한다. 실험 데이터는 기존 방안인 페이지랭크를 이용하여 수집한 방식과 제안 방안인 감성 어휘를 고려한 페이지랭크로 수집한 데이터로 구분한다.

허브 노드가 접근할 수 있는 횟수를 변경하여 진행하였으며 감성 어휘 검출을 위해 AR-KNU+와 SNS 감성 어휘를 추가한 사전을 사용하여 비교한다. 샘플링 데이터를 100개로 하여 10회 실시한다.

표 2. 인스타그램 데이터 수집 결과
Table 2. Collection results from hub nodes

U1	U2	U3	U4	U5
9,308	11,185	9,519	7,406	10,000

표 3. 감성 평가 결과
Table 3. Emotional evaluation result

Very neg	Neg	Obj	Pos	Very pos
43	164	73,172	497	61

실험 결과는 정밀도(Precision)와 재현율(Recall)의 값을 이용하여 F1 score의 값을 계산한다. 정밀도는 대학과 관련된 게시물의 전체 해시태그에서 감정 어휘를 포함한 해시태그의 비율을 계산한 값이다. 재현율은 대학과 관련된 게시물의 해시태그에서 중복을 제거하고 감정 어휘를 포함한 해시태그의 비율을 계산한 값이다. F1 score는 정밀도와 재현율을 이용하여 계산한 값으로 대학과 관련된 게시물 당 감성 어휘를 포함한 비율을 나타낸다.

$$(Precision) = \frac{TP}{TP + FP} \tag{2}$$

$$(Recall) = \frac{TP}{TP + FN} \tag{3}$$

$$(F1 - score) = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

그림 5는 기존 방안과 제안 방안의 실험 결과를 보여준다. 기존 방안인 기존 페이지랭크를 사용한 방식은 감성 어휘를 추가한 사전이 더 높은 F1 score를 보였으며 허브 노드를 3개 사용한 방식이 5개를 사용한 방식보다 높은 F1 score를 보였다. 그러나 가장 높은 F1 score는 0.019를 보였다.

제안 방안인 감성 어휘를 고려한 페이지랭크를 사용한 방식은 감성 어휘를 추가한 사전이 더 높은 F1 score를 보였으며 허브 노드를 3개 사용한 방식이 5개를 사용한 방식보다 높은 F1 score를 보였다. 또한, 가장 높은 F1 score는 0.09로 기존 방안과 비교하였을 때 약 4.5배 높았다. 따라서 허브 노드를 기점으로 하는 데이터 수집은 허브 노드의 개수를 3개를 사용하여 수집한다.

표 4는 그림 5의 제안 방안의 대학별 F1 score의 값을 보여준다. Method A와 B는 각각 허브 노드를 3개, 5개로 하여 확장한 감성 사전을 사용하였으며, Method C와 D는 각각 허브 노드를 3개. 5개로 하여 기존 감성 사전을 사용하였다. 표 5는 그림 5의 기존 방안의 대학별 F1 score의 값을 보여준다.

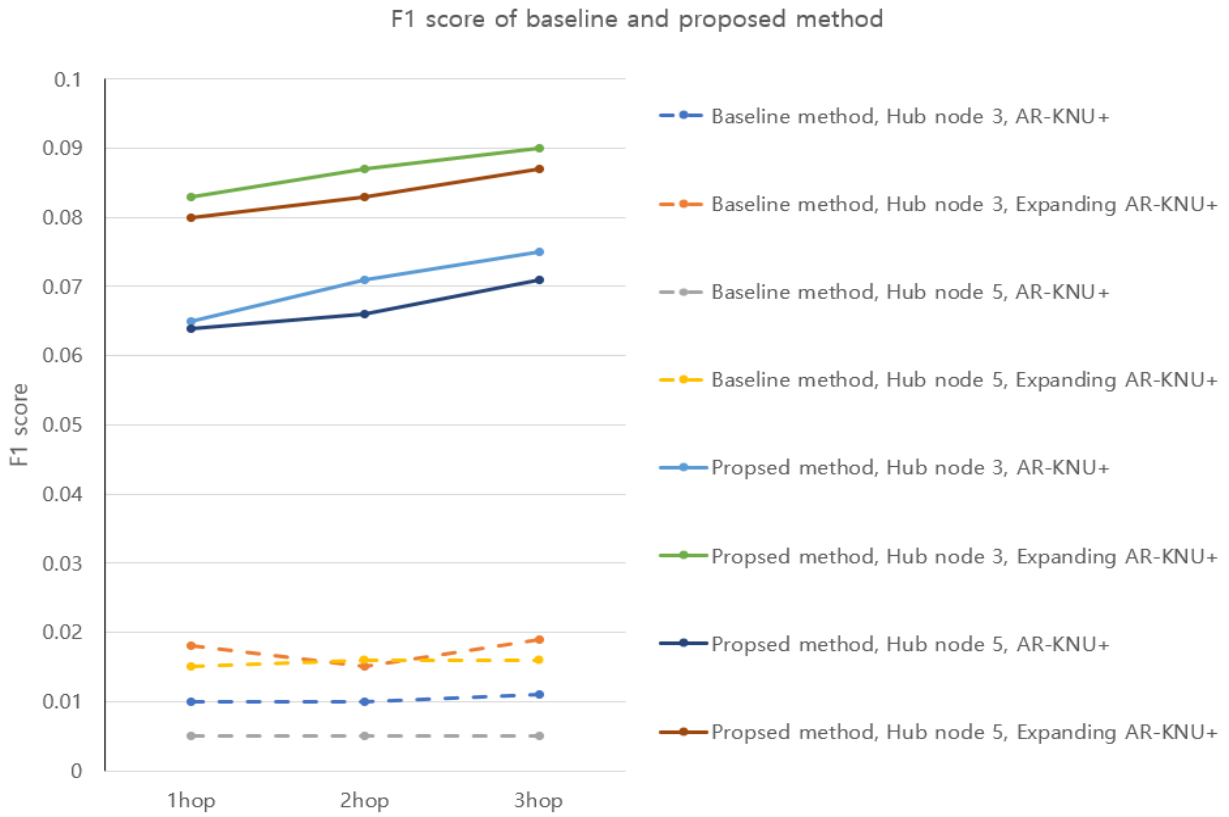

그림 5. 기존 방안 및 제안 방안의 F1 score
Fig. 5. F1 score of baseline and proposed method

표 4. 제안 방안의 대학별 F1 score
Table 4. F1 score of proposed method by university

Hop	Univ	Method A	Method B	Method C	Method D
1	U1	0.094	0.092	0.076	0.073
	U2	0.091	0.089	0.073	0.079
	U3	0.073	0.071	0.055	0.048
	U4	0.066	0.063	0.049	0.055
	U5	0.093	0.089	0.074	0.069
2	U1	0.099	0.096	0.079	0.078
	U2	0.095	0.091	0.078	0.075
	U3	0.076	0.072	0.062	0.05
	U4	0.072	0.068	0.059	0.054
	U5	0.096	0.092	0.078	0.075
3	U1	0.102	0.099	0.088	0.08
	U2	0.097	0.095	0.08	0.078
	U3	0.079	0.076	0.063	0.054
	U4	0.074	0.071	0.068	0.066
	U5	0.098	0.095	0.08	0.079

표 5. 기존 방안의 대학별 F1 score
Table 5. F1 score of baseline method by university

Hop	Univ	Method A	Method B	Method C	Method D
1	U1	0.094	0.092	0.076	0.073
	U2	0.091	0.089	0.073	0.079
	U3	0.073	0.071	0.055	0.048
	U4	0.066	0.063	0.049	0.055
	U5	0.093	0.089	0.074	0.069
2	U1	0.099	0.096	0.079	0.078
	U2	0.095	0.091	0.078	0.075
	U3	0.076	0.072	0.062	0.05
	U4	0.072	0.068	0.059	0.054
	U5	0.096	0.092	0.078	0.075
3	U1	0.102	0.099	0.088	0.08
	U2	0.097	0.095	0.08	0.078
	U3	0.079	0.076	0.063	0.054
	U4	0.074	0.071	0.068	0.066
	U5	0.098	0.095	0.08	0.079

Method A와 B는 각각 허브 노드를 3개, 5개로 하여 확장한 감성 사전을 사용하였으며, Method C와 D는 각각 허브 노드를 3개, 5개로 하여 기존 감성 사전을 사용하였다.

4.5 기존 방안과 감성 게시물 수집 비교

표 6은 기존 연구인 [2]와 수집 데이터의 개수를 비교한 결과를 보여준다.

표 6. 기존 방안과 제안 방안의 수집 게시물 비교
Table 6. Comparison of collection posts of existing and proposed methods

Univ	Baseline method	Proposed method
U1	1,367	7
U2	553	54
U3	323	12
U4	98	5
U5	0	1

제안 방안은 인스타그램을 활용하여 수집한 허브 노드를 기점으로 인접 노드를 탐색하여 감성 게시물을 수집하였다. 그러나 기존 방안과 비교하여 매우 적은 데이터양을 보였다. 이를 통해 인스타그램은 대학과 관련된 감성 게시물의 수가 매우 적은 매체임을 알 수 있다.

4.6 실험 결과

수집한 인스타그램 데이터 중 대학과 관련된 감성 게시물을 수집하는 실험을 수행하였다.

표 7은 허브 노드의 감성 게시물 수집결과를 보여준다. Univ는 대학을 익명으로 나타낸 값이며 Adjacent Node는 허브 노드에서 접근할 수 있는 노드로 감성 게시물에 댓글을 작성한 계정이다. Total Content는 허브 노드에서 수집된 게시물의 개수이다. 대학별로 최소 10개 이상의 인접 노드가 존재하였으며, U4, U5를 제외한 나머지 대학에서는 감성 게시물의 인접 노드로 20개 이상이 존재하였다. 이를 통해 허브 노드의 감성 게시물에 댓글을 작성하는 인접 노드가 많이 존재하는 것을 알 수 있다.

표 8은 허브 노드의 감성 게시물에 댓글을 작성한 노드인 인접 노드의 감성 게시물 수집결과를 보여준다.

표 7. 허브 노드의 게시물 수집결과
Table 7. Collection results from hub nodes

Univ	Adjacent node	Sentimental content	Total content
U1	21	7	112
U2	54	54	119
U3	23	12	86
U4	12	5	86
U5	11	1	27

표 8. 인접 노드의 게시물 수집결과
Table 6. Collection results from adjacent nodes

Univ	Adjacent node	Sentimental content	Total content
U1	21	0	483
U2	54	0	640
U3	23	0	517
U4	12	0	394
U5	11	0	81

허브 노드와 인접한 노드는 평균 24개 정도였지만, 대학과 관련된 감성 게시물을 검출할 수 없었었다. 이를 통해 허브 노드의 인접 노드는 게시물의 댓글만을 작성하고 대학과 관련된 게시물을 작성하지 않는 것을 알 수 있다.

V. 결 론

본 논문에서는 대학 평판도 분석 시 온라인 뉴스 데이터 사용으로 적은 감성 어휘와 다양한 대학이 언급되는 문제점을 해결하기 위해 인스타그램 데이터를 이용하여 대학 평판도 분석을 위한 데이터 타당성을 분석하였다. 대학명을 해시태그로 하는 인스타그램 데이터를 수집하고 맞춤법 검사와 형태소 분석을 수행하였다. 감성 사전을 이용하여 해시태그 감성 분석을 수행하였고 감성이 존재하는 어휘를 수집하였다. 기존 감성사전인 AR-KNU+에 수집한 어휘를 추가하여 SNS 맞춤형 감성 사전을 구축하였다. 구축한 감성사전을 기반으로 감성 게시물을 분류하여 방향 그래프를 구축하였으며 감성 어휘를 고려한 페이지랭크를 이용하여 허브 노드를 추출하였다. 추출한 허브 노드를 기점으로 데이터를 수집하였다.

기존 페이지랭크와 감성 어휘를 고려한 페이지랭크를 비교하였을 때 약 11배 높은 F1 score를 보였다. 그러나 추출한 허브 노드를 기점으로 데이터 수집 시 허브 노드는 대학과 관련된 감성 게시물을 일부 수집할 수 있었지만, 인접 노드에서는 수집할 수 없었다. 따라서 대학 평판도 분석 시 수집 매체로서 적합하지 않은 결과를 보였다.

향후 연구로는 수집 데이터 매체로 대학과 관련된 커뮤니티를 선정하여 데이터의 크기를 늘려 실험을 수행할 것이다.

References

[1] University Population Shortage, http://www.jeollailbo.com/. [accessed: Jun. 27, 2021]

[2] QS World University Rankings, https://en.wikipedia.org/wiki/QS_World_University_Rankings. [accessed: Jun. 27, 2021]

[3] Joongangilbo university reputation, http://univ.joongang.co.kr. [accessed: Jun. 27, 2021]

[4] S. M. Park, C. M. Eom, B. W. On, and D. W. Jeong, "An AR-KNU Sentiment Lexicon-based University Reputation Assessment Using Online News Data", The Journal of Korean Institute of Information Technology, Vol. 17, No. 3, pp. 11-21, Mar. 2019. http://dx.doi.org/10.14801/jkiit.2019.17.3.11.

[5] S. H. Chae, D. W. Jeong, B. W. On, and J. W. Gim, "University Reputation Assessment through AR-KNU Expansion", The Journal of Korean Institute of Information Technology, Vol. 18, No. 11, pp. 35-45, Nov. 2020. http://dx.doi.org/10.14801/jkiit.2020.18.11.35

[6] M. S. Jung, J. H. Jung, S. H. Chae, J. W. Gim, and D. W. Jeong, "Extending AR-KNU for Analysis of University Reputation", Proceedings of KIIT Conference, Cheongju, Korea, pp. 380-383, Oct. 2020.

[7] I Love Korean, http://www.ilovekorean.net/koreanlanguage3/chapter1/1-2-2.html. [accessed: Jun. 20, 2021]

[8] SNS utilization, https://biz.chosun.com/. [accessed: Jun. 20, 2021]

[9] M. J. Nam, E. J. Lee, and J. H. Shin, "A Method for User Sentiment Classification using Instagram Hashtags", Journal of KMMS, Vol. 18, No. 11, pp. 1391-1399, Aug. 2015. http://dx.doi.org/10.9717/kmms.2015.18.11.1391.

[10] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification

Approach", Proceedings of the 20th ACM internationalconference on Information and Knowledge management(CIKM), Glasgow, United Kingdom, pp. 1031-1040, Oct. 2011. https://doi.org/10.1145/2063576.2063726.

[11] H. S. Kim, H. W. Kim, H. J. Seo, and Y. K. Lee, "An Efficient Distributed PageRank Technique on Summarized Graph", Korean Institute of Information Scientists And Engineers, pp. 2026-2028, Jun. 2018.

[12] J. X. Parreira, C. Castillo, D. Donato, S. Michel, and G. Weikum, "The Juxtaposed approximate PageRank method for robust PageRank approximation in a peer-to-peer web search network", The Journal of VLDB, Vol. 17, No. 2, pp. 291-313, Feb. 2007. http://dx.doi.org/10.1007/s00778-007-0057-y.

[13] Pagerank, https://en.wikipedia.org/wiki/PageRank. [accessed: Jun. 20, 2021]

[14] E. A. Kim and Y. S. Lee, "The College Reputation System using Public Data and Sentiment Analysis", The Journal of convergence security, vol. 18, No. 1, pp. 103-110, Mar. 2018

[15] Y. K. Kim and W. S. Kang, "A Study of Special-purpose Academy Image Positioning Strategy Utilizing Big Data Analysis", Journal of Social Science, Vol. 35, No. 1, pp. 33-70, Jun. 2018.

저자소개

정 민 수 (Minsu Jung)

2021년 ~ 현재 : 군산대학교 소프트웨어융합공학과(학부생)
관심분야 : 빅데이터 분석, 소셜 네트워크 분석

김 장 원 (Jangwon Gim)

2005년 : 상명대학교 컴퓨터소프트웨어공학과(학사)
2008년 : 고려대학교 컴퓨터학과(석사)
2012년 : 고려대학교 컴퓨터·전파·통신공학과(박사)
2013년 : 한국과학기술정보연구원 선임연구원
2017년 ~ 현재 : 군산대학교 소프트웨어융합공학과 교수
관심분야 : 빅데이터 분석, 자연어처리, 지식 그래프

온 병 원 (Byung-Won On)

2007년 : Pennsylvania State U. 컴퓨터공학과(박사)
2008년 : U. of British Columbia 포스닥연구원
2010년 : U. of Illinois at Urbana-Champaign, Advanced Digital Sciences Center 선임연구원
2011년 : 차세대융합기술연구원 선임연구원
2014년 ~ 현재 : 군산대학교 소프트웨어융합공학과 교수
관심분야 : 데이터 마이닝, 빅데이터, 인공지능, 강화학습

정 동 원 (Dongwon Jeong)

1997년 : 군산대학교 컴퓨터과학과(학사)
1999년 : 충북대학교 전산학과(석사)
2004년 : 고려대학교 컴퓨터학과(박사)
2005년 ~ 현재 : 군산대학교 소프트웨어융합공학과 교수
관심분야 : 데이터베이스, 시맨틱 서비스, 빅데이터, 사물인터넷, 엣지컴퓨팅, 지능형 융합 서비스