공공데이터와 감성분석을 이용한 대학평판시스템

김 은 아*, 이 연 식**

요 약

현대 사회는 인터넷과 SNS를 통해 발생하는 복합적이며 대량의 데이터를 수집·집계·분석하는 빅데이터 처리 기술이 여러 분야에서 요구되고 있다. 그 중 대표적인 활용분야가 기업이나 대학에 대한 평판을 평가하는 평판시스템이다. 대학평판을 측정하고 수치화하기 위해서는 공정하고 객관적인 자료와 효율적인 데이터 처리가 무엇보다도 중요하다. 이를 위하여 공공데이터 지표를 활용하여 정량지수를 구하였고, 뉴스 기사를 활용한 감성분석을 통해서 정성지수를 구한 후 혼합 대학평판 지수를 산출하였다. 본 논문에서는 대학평판을 측정하기 위하여 정량지수로 객관성을 확보하면서 감성적 평판을 반영한 혼합 대학평판 지수를 산출하였고 이를 바탕으로 혼합 대학평판 시스템을 제안하였다.

The College Reputation System using Public Data and Sentiment Analysis

Eun-Ah Kim*, Yon-Sik Lee**

ABSTRACT

Modern society is increasingly demanding in many areas of big data processing technology to collect, aggregate, and analyze large amounts of data over the Internet and SNS. A typical application is to evaluate the reputation of a company or college. To measure and quantify a reputation, fair and precise data and efficient data processing are very important. For this purpose, a quantitative quotient was obtained using public data, a qualitative quotient was obtained through sentiment analysis using news articles, and a complex college reputation quotient was calculated. In this paper, a complex college reputation quotient was calculated based on the quantitative index, reflecting the sentimental reputation, and based on the proposed mixed university system. In this paper, the Complex College Reputation System(CCRS) was proposed, which produced the Complex College Reputation Quotient with an objective quantitative quotient and qualitative quotient reflecting the sentimental reputation to measure the college reputation.

Key words : Reputation, College, Public data, Sentiment Analysis, Reputation Quotient

접수일(2018년 3월 12일), 수정일(1차: 2018년 3월 23일), 게재확정일(2018년 3월 30일)

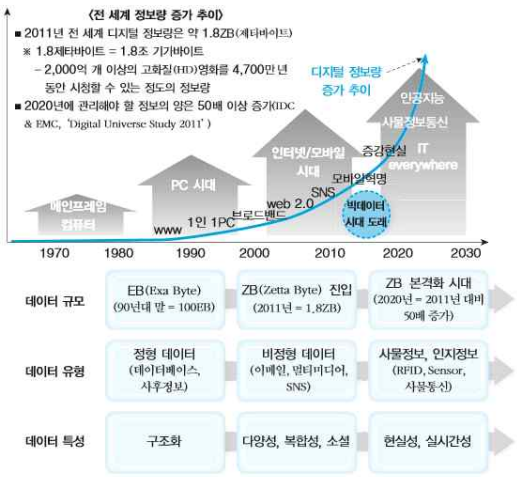
* 한국폴리텍대학 인천캠퍼스 / 컴퓨터정보과

** 국립군산대학교 / 컴퓨터정보통신공학부

1. 서 론

최근 모바일 중심[1]의 온라인과 가상환경으로 인하여 개인은 물론 정부, 기업과 대학들은 홈페이지를 통하여 정보와 서비스를 제공하고 있으며, 자료의 온라인화로 정보검색이 중요해지면서 포털을 중심으로 하는 검색사이트들의 비중이 높아졌다. 현재 많은 사람들이 정보 검색과 인터넷의 시작을 포털 중심으로 하고 있고 여기에 SNS 서비스의 등장으로 사소한 잡담에서 의미 있는 정보 제공까지 혼합되어 사용되면서 데이터의 양은 기하급수적으로 증가하였다.

IDC는 2020년 전 세계 디지털 정보의 양이 2011년에 비해 50배 정도 증가할 것으로 전망하고 있으며, 2011년 전 세계에서 생성되는 디지털 정보량은 1.8제타바이트로, 매 2년마다 2배씩 증가한다고 발표하였다. 전 세계 정보량 증가 추이를 그림 1에 나타내었다.[2] 빅 데이터의 활용에서 데이터가 많다고 모두 양질의 정보를 제공하는 것은 아니므로, 어떤 데이터를 어떻게 수집·집계·분석하여 원하는 분야에 적용해야 하는지가 중요하게 되었다.



(그림 1) 전 세계 정보량 증가 추이

온라인과 오프라인에 흩어져있는 대량의 데이터를 통한 조직이나 개인에 대여 평가하는 것은 일반적인 현상이 되었으며, 개인, 기업, 대학과 다양한 조직에 있어 평판은 선택을 위한 최우선적 고려사항이 되어가고 있다. 평판에 활용되는 데이터는 신문기사 같이 정형화 되어있는 정형데이터와 트위터나 페이스북과 같이 형식 없이 자유롭게 작성된 비정형 데이터가 있는데 모두 평판에 중요한 요소들이다. 정형데이터는 수집과 분석에 유리한 반면, 비정형데이터는 텍스트 정보뿐만 아니라 이모티콘이나 약어, 특수기호 등의 사용으로 해석에 어려움을 겪고 있다. 현재 평판에 활용되고 있는 자료는 신문기사, 트위터, 페이스북 등인데 사회 환경이나 유행에 따라 크게 변하여 정확한 평판이 쉽지 않다. 따라서 정성적인 요소와 조직이나 개인에 대한 판단에 활용할 수 있는 정량요소를 활용하면 객관성을 높일 수 있다고 판단된다.

본 논문은 대학평판 시스템에 대한 연구이며 기존의 대학평판 시스템들과는 다르게 공공데이터를 활용한 정량지수와 빅 데이터의 정형데이터 중 뉴스기사를 중심으로 감성분석을 한 정성지수를 활용하고자 한다. 대학평판에 정량지수와 정성지수를 활용한 혼합 대학평판 지수(CCRQ : Complex College Reputation Quotient)를 산출하고 이를 기반으로 한 혼합 대학평판 시스템(CCRS : Complex College Reputation System)을 제안한다.

본 논문의 구성은 2장에서 평판시스템과 대학평판을 소개하고, 3장에서는 평판지수로 활용하고자 하는 정량지수와 정성지수에 대하여 정의한다. 4장에서는 정의한 정량지수와 정성지수를 활용하여 3단계로 구성된 새로운 혼합 대학 평판시스템을 설계 제안하고, 시스템의 특징을 제시한다. 5장에서는 시스템 구현 환경과 실험 결과를 보이고, 마지막으로 6장에서 결론을 제시한다.

2. 평판시스템과 대학평판

평판은 1950년대 이래 경제학, 조직이론, 마케팅 등 여러 분야에서 연구 되어온 학문으로 현대사회에서는 평판이 매우 중요하며 신용점수와 마찬가지로 평판지수로 평가되어 다양한 분야로 활용될 전망이다.[3] 평판은 정체성, 이미지, 브랜드 등의 개념과 혼용되어 사용하고 있다. 여기서 정체성은 개인에 있어 신체적 특징, 활동적 특징, 사회적 특징, 심리적인 특징을 나타내고 있으며, 기업에 있어서는 기업 브랜드, 제품 브랜드를 나타내고 있다.[4]

평판의 좋고 나쁨을 지수화하려는 노력이 계속되었는데 일반적으로는 설문조사와 같은 정성적인 방법을 통하여 이것을 계량화하려는 시도가 주를 이루었다. 이것을 평판지수라 하며 기업의 경우 기업이 과거에서부터 현재까지 오랜 기간 수행해온 기업행위에 대한 평가를 의미한다. 대표적인 평판지수로 미국 평판 연구소와 여론조사 업체 해리스 인터랙티브의 평판지수로 측정항목은 감성적인 매력, 제품과 서비스, 근무환경, 재무성과 비전과 리더십, 사회책임으로 평가하고 있다.[5]

평판을 쌓는 것도 중요하지만 평판을 관리하는 중요성이 높아지게 되었는데, 평판관리의 핵심은 부정적 평가를 줄이고 긍정적 평가를 늘려 사회적 신뢰자본을 쌓아 나가는 활동이라 할 수 있다. [6]

평판이 중요한 조직 중 하나가 대학이며 좋은 대학과 아닌 대학의 구분은 평가자 마다 기준이 다르지만 기업, 일반인이나 수험생들에게 중요한 요소가 되었다. 그러나 대학평판이 중요하다고 생각하는 것에 비해 학문적 연구는 활발하지 못한 실정이다. 여러 가지 이유가 있겠지만 대표적인 것은 교육부에서 추진하는 대학구조개혁과 재정지원사업 등의 대학 평가 때문이며 일반인들은 그 결과를 일방적으로 수용하고 있는 실정이다. 다른 이유 중 하나는 평가의 공정성이나 객관성에 문제를 제기하는 부분도 있지만 언론에서 발표하는 대학평가(사례 중앙일보 대학평가)가 있다. 이러한 평가결과는 파급력이 커서 대학구성원이나 일반인들도 중요한 자료로 활용하고 있다. 이러한 이유들로 대학평판에 대한 연구가 어려움이 커지고 연구자들의 관심을 받지 못한 것으로 보인다.

대학 이미지는 대학의 명성이나 평판과 깊은 관련이 있거나 때때로 동일한 개념으로 파악되고 있다.[7] 대학의 교육 서비스, 경영 스타일, 대외활동 노력, 그리고 글로벌 활동 정도 등 대학의 대내외적인 활동들이 반영되어 이미지가 형성된다고 하였으며 대학의 이미지 측정도구인 Colovim(College Overall Image) 척도 개발로 7개 요인에 21개 측정항목으로 구성하였는데, 보완할 부분은 측정 방법이 정성적이란 것과 조사 대상이 재학생이었다는 부분이다.

SNS 데이터를 활용한 국내대학 인식 및 선호도 분석에서는 SNS 빅 데이터를 수집하고 분석하여 이전까지 알 수 없었던 새로운 가치를 발견하려 하였고 국내대학들에 대한 일반인들의 인식을 파악 및 분석하여 각 대학의 발전에 도움이 될 시의성 있는 정보와 의견을 제시하였다. 이 연구의 결과는 전체대학에 대한 키워드 빈도분석과 해당 대학의 연관 키워드 빈도분석과 감성분석을 실시하였다.[8] 연구결과는 우수하나 트위터 같은 비정형데이터만으로 데이터 분석과 감성분석을 실시한 부분은 아쉬운 부분으로 생각되며 여기에 정형데이터가 포함한다면 더욱 좋을 것이다.

대학 평가지표들에 대한 상관분석 연구에서는 지표들 간의 연관성과 통계적 모형을 추정하였는데 사용된 평가지표는 12개로 대표적으로는 신입생 최종등록률, 재학생 충원율, 건강보험 DB연계 취업률 등이었다.[9] 평가지표는 공공데이터 중 대학알리미 자료를 활용하였고 이들 지표들 간의 연관성을 분석하였다. 대학 평가를 목표로 하면 평가지표는 더욱 정교해져야 할 것인데 대학 평판을 목표로 한다면 단순하고 일반인들도 알 수 있는 주요지표인 몇 개로 선정하는 것이 필요하다고 판단된다.

평판지수로 정성지수 뿐만 아니라 정량지수를 혼합하여 사용하고, 대학 평판 측정을 위해서 뉴스기사 위주의 정형데이와 대학 평가지표 중 신입생 취업률, 재학생 충원율, 취업률(건강보험 DB연계) 등 3가지 대표 지표를 활용함으로서, 평판시스템의 신뢰성과 객관성을 향상시키고자 한다.

3. 평판지수

3.1 공공데이터와 정량지수

공공데이터(Public data)를 민간에 개방하는 오픈데이터 정책이 세계적으로 확산되고 있다. 미국은 오픈데이터정책, 영국은 정보경제전략, 일본은 전자정부 오픈데이터 전략 등의 정책을 시행하며 공공데이터 플랫폼을 구축하는 등 적극적인 개방 정책을 추진하고 있다. 공공 부문의 데이터 공개를 통해 행정의 효율성을 높이고, 국민의 참여를 활성화시키며 경제 활성화 등의 파급 효과를 기대하고 있다. 정부의 데이터 공개 정책은 정보화 시대에 소통과 공유, 협업 전략이 무엇보다 중요하다는 것을 의미한다.[10]

공공데이터 공개는 과거에도 추진되었지만 특히

최근에 주목받는 이유는 빅 데이터 분석기술의 발달과 스마트 생활환경의 확산에 기인한다고 볼 수 있다. 과거에는 연구나 정책 개발의 목적으로 연구자나 업무 담당자가 공공데이터를 제한적이고 폐쇄적으로 활용했으나 최근에는 대중교통정보 서비스 등과 같이 많은 사람들이 수시로 자신의 정보단말기로 공공데이터를 활용하고 있다. 공공데이터 중에서도 대학정보공시센터의 '대학알리미' (http://www.academyinfo.go.kr)는 교육관련 연구자들과 대학 관계자들을 중심으로 많이 활용되고 있다. 대학알리미 사이트 정보를 그림 2에 나타내었다. 현재 대학알리미에는 대학과 관련된 정보와 대학 평가 지표들이 있는데 공공기관이나 정부에서 제공하는 자료로 객관성과 신뢰성이 있는 자료들이다.[9]

(그림 2) 대학알리미 정보

본 논문에서 사용할 정량지수 산출을 위한 지표로는 신입생 충원율, 재학생 충원율, 취업률이다. 많은 평가지표가 있지만 학생대상의 지표를 중심으로 일반인들이 알고 있는 입학, 대학생활, 취업이라는 3가지 중요한 항목을 선정하여 대학평판 지표로 활용하였다. 대학알리미에서는 지표별로 3년간의 데이터를 제공하고 있으며 최근데이터를 기준으로 차등적으로 정량지수로 활용한다. 신입생 충원율은 매년 입시가 끝난 3월 최종 등록률을 기준으로 산출하고 8월에 공시하고 있다. 신입생 충원율에 대한 정량지수는 2018년 1월 기준으로 2017년도를 50%, 2016년도를 30%, 2015년도를 20% 비율로 계산한다. 재학생 충원율은 매년 4월 1일을 기준일로 하여 전체학생과 편제정원을 기준으로 산출하고 있으며 매년 8월에 공시한다. 대학평가에서도 주로 편제정원 대비 정원 내 재학생 충원율을 활용하고 있어 그 기준에 따라 정량지수로 활용하며 편제정원을 기준으로 정원 내 재학생의 비율로 계산하는데 100%가 넘는 경우 100%로 계산한다. 캠퍼스 통합 등의 사유로 재학생 충원율이 일시적으로 과도한 비율(100%를 훨씬 넘기는)을 나타내는 문제가 있어 그 부분을 보완하기 위해서이다. 재학생 충원율도 대학평가 지표에서 활용하는 정원 내 재학생 충원율을 2017년 공시 50%, 2016년 공시 30%, 2015년 공시 20% 비율로 계산한다. 취업률은 2017년 공시자료인 경우는 전년도 졸업자(2016년 2월)를 대상으로 건강보험DB연계 자료와 교내취업에 대한 자료를 6월에 공시하고 건강보험DB연계자료, 교내취업, 개인창작활동, 1인 창업과 프리랜서등의 자료를 추가하여 12월에 공시한다. 취업률도 2017년 공시 50%, 2016년 공시 30%, 2015년 공시 20%의 비율로 계산한다.

3.2 감성분석과 정성지수

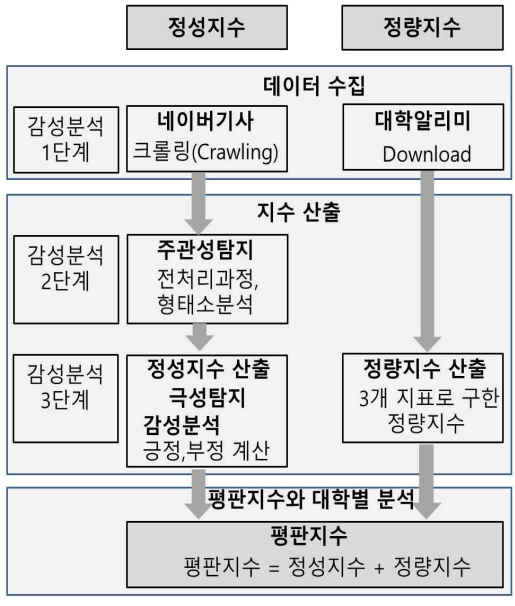
뉴스기사나 소셜 네트워크(Social Network)상의 정보를 기반으로 평판 대상에 대한 사람들의 생각이나 좋고 나쁨에 대한 감성 상태를 알고 싶어 한다. 특정 대상에 대해 좋고 싫음에 관한 감성적인 내용을 분석하는 것이 감성분석(Sentiment Analysis)이다. 감성분석은 글이나 텍스트 정보상의 사람들의 감정, 태도 같이 주관적인 데이터를 처리하고 분석하는 자연어 처리 기술을 의미한다. 감성분석과 유사한 형태가 페이스북의 '좋아요', 영화 평의 '별점'같은 경우인데 이런 정보들은 수치화하기 쉬운 장점이 있는 반면에 나쁜 의도로 조작할 수 있다는 단점을 가지고 있으며 이는 결국 '좋아요'와 '별 점'자체의 신뢰도를 떨어트릴 수 있다. 이러한 신뢰도 문제를 해결하기 위해서는 평가자가 작성한 텍스트 정보의 의도를 정확히 파악하여 긍정, 부정 여부를 파악해야 한다. 인터넷 기사나 사용자들이 작성하는 SNS 자료를 수집한 뒤, 체계적으로 수치화 하는 작업이 필요하며 이를 통해 긍정, 부정이나 중립적인 의견을 해석할 수 있어야 한다. 인터넷과 소셜미디어상의 자료는 뉴스, SNS나 블로그 등의 형태로 제공되며 그림 1에 나타난 바와

같이 방대한 양의 자료에 대해 사람을 투입해 일일이 추적하기란 사실상 불가능에 가깝다. 이런 방대한 작업을 자동화하는 기술을 감성 분석 기술이라 한다.[11] 감성분석은 3단계로 이루어지며 이는 데이터 수집, 주관성 탐지(Subjectivity Detection), 극성 탐지(Polarity Detection)이다. 1단계 데이터 수집은 인터넷상의 각종 블로그, 게시판, 트위터, 페이스북이나 신문기사들의 방대한 양의 데이터를 가지고 오는 것을 말하며 검색엔진을 통해 자료를 다운로드하거나 크롤링을 통해서 데이터를 수집한다. 수집된 데이터는 파일로 보관하거나 데이터베이스에 저장 할 수 있다. 2단계 주관성 탐지는 데이터를 수집하고 난 다음에는 감성분석에 사용될 요소들을 분리, 분류하는 작업을 말하며 감성분석과 무관한 데이터를 처리하면서 생길 수 있는 여러 가지 손실을 막기 위한 필요한 부분이다. 감성분석과 연관이 있는 요소에서도 처리과정에 불필요한 부분을 제거하는 것도 포함된다. 특수문자나 이모티콘 등의 감성분석에서 처리 못하는 요소들을 제거한다. 3단계 극성탐지는 감성분석에 필요한 부분에서 주어진 데이터가 '긍정'인지, '부정'인지를 판단하는 부분이다. 긍정적, 부정적인 단어를 탐지, 이를 정량화 한 뒤 통계적 기법을 적용한다. 극성탐지에는 기본적인 방법으로 감성어 사전(lexicon)을 이용하는 방법이 활용되고 있으며, 최근에는 기계학습을 사용하여 긍정, 부정을 판단하는데 사용하고 있다. 가장 많이 사용하는 방법으로는 긍정, 부정 어휘를 분류한 후 문서 내에서 이런 극성 어휘들의 빈도를 연산하는 방법이 주를 이뤄왔다. 이를 위해서는 감성어 사전이 구축되어 있어야 한다. 영어의 경우 약 만개의 문장 안에 나타나는 감성 표현들로 구축된 사전 MPQA(Muiti-Perspective Question Answering)가 있다. 한글의 경우 서울대학교에서 구축한 한국어 감성 코퍼스 KOSAC(Korean Sentiment Analysis Corpus)[11]이라는 감성사전이 있어 한국어 감성분석연구에 활용되고 있다. 감성어 분석의 질은 사전이 얼마나 많은 감성어를 포함하고 있는가에 따라 달라진다. 한글은 영어와 달리 언어자체의 복잡성 때문에 감성어 분석이 어렵다. 보통 자연어 처리는 '형태소 분석', '구문 분석' 그리고 '의미 분석'순으로 이루어지는데 한국어는 아직도 형태소 분석과정에서 많은 어려움을 겪고 있다.

본 논문에서는 네이버 교육 뉴스기사를 대상으로 삼았으며 이는 많은 뉴스 기사 중 네이버의 기준으로 선별된 기사로 신뢰성이 있다고 판단된다. 이 기사들을 대상으로 KOSAC의 감성사전을 이용해서 감성분석의 3단계인 극성탐지를 실시하며 신문기사의 극성어휘의 빈도수를 측정(30개 추출)하고 이를 통해 신문기사의 긍정, 부정을 판단하고자 한다. 긍정의 기사에는 +1, 부정의 기사에는 -1, 중립인 경우는 0점을 부여하여 정성지수(Qualitative Quotient)를 산출하는데 각 대학별 월별 기사 중 모든 기사의 극성을 계산하고 합산하여 산출한다.

4. 혼합 대학평판 시스템(CCRS)

혼합 대학평판 시스템은 다른 연구들에 비해 정량데이터를 활용하여 대학평판에 이용하는 점이 특징이며, 정형데이터 위주의 신문기사를 중심으로 감성사전을 활용하여 기사의 긍정, 부정을 계산한 결과를 정성지수로 활용한다. 혼합 대학평판 지수는 정량지수와 정성지수의 합으로 산출하며, 산출과정은 다음 그림 3과 같다.



(그림 3) 혼합 대학평판지수 산출과정

혼합 대학평판 시스템은 3단계로 구성된다. 1단계는 데이터 수집 단계로 정량지수 산출을 위한 데이터 수집단계이며 공공데이터인 대학알리미를 통해서 대학의 3년간 신입생 충원율, 재학생 충원율, 취업률을 산출한다. 정성지수 산출을 위한 데이터 수집은 네이버 신문기사를 수집한다. 2단계는 지수 산출 단계로 정량지수에서는 3년간 데이터 중 최신(2017년 기준) 데이터의 비중을 50%, 전년도는 30%, 전전년도는 20%의 비율로 계산한 후 신입생 충원율(15%), 재학생 충원율(20%), 취업률(15%)의 비율로 정량지수를 산출한다.

정성지수는 전처리(Preprocessing) 과정으로 크롤링된 전체 기사 중에서 먼저 전문대학을 제외한 대학 관련 기사를 추출하고, 추가적으로 제외할 필요가 있는 단어들을 포함하여 필터링함으로서 '일반대학'과 관련된 기사만을 처리대상으로 한다. 한글 형태소 분석은 KoLNPy[13]의 API중 mecab을 활용하여 빈도가 높은 형태소 30개를 추출하고 이 형태소를 대상으로 감성사전의 극성을 참조하여 긍정, 부정과 중립을 계산한다. 여기서 극성이란 각 형태소별로 긍정, 부정의 사용 확률을 계산한 결과를 의미한다.[12]

본 논문에서는 KOSAC의 감성사전을 기반으로 형태소가 없는 단어는 추가했고, 극성 결과가 대학과 현저히 맞지 않은 부분에 대해서는 극성 값을 조정하여 사용하였다. 감성분석 중 형태소 분석 시 약어에 의한 동의어 처리로 인한 혼란이 우려되어 가장 일반적으로 사용되는 몇 가지의 (대)명사들로 통일하여 적용한다.

예를 들면. "고려대학교'의 약어인 '고대'는 다른 의미도 있어 사용하지 않고 '고려대학', '고려대학교' 등 2가지 형태를 하나의 통일된 의미로 적용한다.
혼합 대학평판 지수는 정량지수와 정성지수를 각각 50%씩 반영하여 계산한다.

교육 관련 대학기사들은 각 신문사별 기사 내용 중에서 선별되어 네이버에 실리기 때문에 정성지수 계산 시 한 건의 기사도 실리지 못하는 대학들이 많아 정량지수가 어느 정도 보완하는 기능도 가지고 있다. 혼합 대학평판 지수는 월별로 계산하고 매월 평균치로 계산되어 급격한 상승과 하락을 보완하는데 월별로는 대학별로 큰 차이가 나타날 수 있다. 평판지수 산출 후 각 대학별로 해당 월에 많이 언급된 단어들을 대상으로 빈도수를 나타내어 대학별 월별 변화를 관찰할 수 있다. 이 결과는 빈도수 높은 단어 위주로 보일 수도 있고 시각화를 통해 단어 구름(Word Cloud) 형태로도 나타낼 수 있다. 따라서 혼합 대학평판 지수와 대학별 맞춤형 자료를 통하여 각 대학의 평판을 정량적인 결과와 정성적인 결과로 확인 할 수 있는 장점을 가진다.

5. 구현 및 결과

혼합 대학평판 시스템은 대학의 정량지수와 정성지수로 계산된 혼합 대학평판 지수를 산출하고 각 대학별 이슈를 여러 가지 형태로 나타낼 수 있다. 구현은 파이썬 언어를 사용하여 구현하였고 크롤링은 파이썬 라이브러리(BeautifulSoup)를 사용하여 구현하였으며, 한글 형태소 분석은 KoNLPy의 API중 mecab으로 수행하였다. 감성분석은 KOSAC의 감성사전에 기반을 두어 긍정 부정을 판단하였고, 데이터 처리와 혼합 대학평판 지수 산출은 파이썬 라이브러리인 팬더스(Pandas)에서 수행하였다.

정량지수는 대학알리미의 2017년 12월 31일 기준으로 산출하였고 결과를 보면 전국적으로 교육대학들의 정량지표(신입생 충원율, 재학생 충원율, 취업률)가 높게 나왔는데 3개 지표에서 골고루 높게 나왔고 특히 취업률에서는 10개 대학 중 8개 대학이 75% 이상을 나타냈다. 수도권을 보면 일반대학 평가와 비슷한 결과를 나타냈다.

정성지수는 2017년 12월, 2018년 1월 네이버 교육란의 기사를 기준으로 산출하였는데 2017년 12월 네이버 교육관련 기사 7,369개 중 전처리를 거친 일반대학 관련 기사는 1,044개가 대상이었고, 2018년 1월은 네이버 교육관련 기사 6,306개 중 전처리를 거친 일반대학 관련 기사는 753개가 대상이었다. 결과를 보면 긍정, 부정의 점수 계산도 중요하지만 일반적으로 기사 노출 빈도수에 비례하는 결과를 보이고 있다.

정량지수와 정성지수를 합한 혼합 대학평판 지수의 일부 결과는 그림 4에 나타내었다. 왼쪽 그림은 2017년 12월 결과이며, 오른쪽은 2018년 1월 결과이다.

지역	대학명	형태	혼합대학평판지수
서울	서울대학교	국립대법인	95.55
서울	세종대학교	사립	63.33
서울	건국대학교	사립	59.24
서울	고려대학교	사립	56.41
전북	전북대학교	국립	56.16
경기	아주대학교	사립	54.24
서울	숭실대학교	사립	53.54
서울	한양대학교	사립	53.36
인천	인천대학교	국립대법인	53.03
서울	상명대학교	사립	52.44
서울	연세대학교	사립	51.35
강원	강원대학교	국립	50.65
충남	한국기술교육대학교	사립	50.44
서울	동국대학교	사립	50.34
서울	이화여자대학교	사립	50.27
서울	성균관대학교	사립	50.20
인천	인하대학교	사립	49.78

지역	대학명	형태	혼합대학평판지수
서울	서울대학교	국립대법인	95.55
부산	한국해양대학교	국립	81.08
서울	성균관대학교	사립	67.97
전북	전북대학교	국립	66.51
광주	전남대학교	국립	65.53
서울	연세대학교	사립	63.84
서울	한양대학교	사립	62.23
서울	건국대학교	사립	60.97
울산	울산과학기술원	특별법법인	60.21
경기	수원대학교	사립	59.78
울산	울산대학교	사립	57.45
인천	인하대학교	사립	56.90
인천	인천대학교	국립대법인	56.90
전북	원광대학교	사립	56.84
서울	숭실대학교	사립	56.76
서울	고려대학교	사립	54.35

(그림 4) 혼합 대학평판 지수 결과

각 대학별 월별 정보제공은 대학별로 이슈가 되는 단어 중 빈도수가 높은 단어를 표나 시각화 정보로 표시하며 이를 통해서 관심사항의 변화를 파악할 수 있다. 서울대학교에 대한 예를 그림 5에 나타내었다.

2017년 12월	2018년 1월
서울대	교수
뉴스	서울대
캠퍼스	논문
학생	대학
시흥	조사
교수	연구
총장	학생
서울	자녀
제보	교육
금지	경제

(그림 5) 각 대학별 월별 이슈 결과(서울대 사례)

본 논문의 결과는 혼합평판지수뿐만 아니라 정량지수, 정성지수를 따로 추출해 사용할 수 있는 장점을 가지고 있으며 정성지수에서는 대학별 특정 기간 기사 빈도수도 추출할 수 있다. 기존 연구 중 비정형 데이터를 중심으로 국내대학 인식 및 선호도 분석[8] 연구결과와 비교해 보았다. 동일한 기간(2013년 10월 15일~11월 23일)의 정형데이터의 빈도수와 비교한 결과를 표 1에 나타내었다. 이 기간 중 네이버 기사는 6,177건이었고 이중 대학과 관련된 기사는 1,156건이었다. 빈도수 결과를 보면 10개 대학 중 5개 대학이 비슷한 결과를 보였으며, 또한 차이도 보였는데 원인은 정형데이터와 비정형데이터의 차이로 나타난 결과이다. 혼합평판시스템에서는 혼합평판지수뿐만 아니라 정성지수 중 기사빈도수까지 포함하고 있어 결과를 다양하게 나타낼 수 있는 장점이 있다.

<표 1> 대학별 빈도수 결과 비교

전체대학빈도분석		평판시스템(CCRS)	
대학교명	빈도수	대학교명	빈도수
서울대학교	67,095	**서울대학교**	99
고려대학교	47,202	**고려대학교**	36
국민대학교	25,650	**건국대학교**	33
경북대학교	24,478	성균관대학교	27
강원대학교	22,641	이화여자대학교	25
서강대학교	21,332	동국대학교	23
연세대학교	20,224	**서강대학교**	23
한양대학교	15,113	**연세대학교**	16
건국대학교	12,570	조선대학교	15
경찰대학교	11,790	목원대학교	13

6. 결 론

국내에서 대학평판에 관한 연구는 대학 이미지 인식 측정이나 선호도 분석으로 수행되었으며 대학평가와 관련하여 평가지표 연구로 진행되었다. 이러한 연구들의 결과를 대학평판 시스템으로 통합하면서 보완하였는데 보완사항은 비정형 데이터로 인한 신뢰성 문제와 정성적 데이터로 인한 객관성 문제들이었다. 본 논문에서는 대학평판을 위해서 정형데이터 위주로 신뢰성을 높였고, 정량지수와 정성지수를 혼합하여 객관성을 높인 혼합 대학평판 시스템과 평판 지수를 제안하였다. 공공데이터 중 대학알리미의 정량지표를 활용하여 정량지수를 구하였고, 정형데이터인 네이버 뉴스 기사를 기반으로 감성분석을 통해서 정성지수를 구한 후 혼합 대학평판 지수를 산출하였다. 이를 통해 신뢰성과 객관성을 가진 대학평판 시스템을 구축하고자 하였다.

빅 데이터 시대의 특징을 살린 대학평판 시스템이 되기 위해서는 많은 데이터의 활용과 많은 데이터를 처리하면서도 정확한 결과를 얻을 수 있어야 한다. 본 연구에서는 정형데이터 위주로 다루었는데 비정형 데이터까지 취급하는 확장성과 감성사전을 활용한 방법을 개선하여 머신러닝 등을 활용하여 정확성을 높인다면 향상된 대학평판 시스템이 될 것이다.

참고문헌

[1] 문송철, 안연식, “스마트폰 가치의 사용자 인식에 관한 연구”, 한국융합보안학회, 제11권 3호, pp55~66, 2011.6.

[2] 이진형, “데이터 빅뱅, 빅 데이터(BIG DATA)의 동향”, 방송통신전파저널, 2012

[3] 박홍식, “평판관리”, 커뮤니케이션스, 2016.

[4] Formburn. C. & Shanley, “What’s in a name? Reputation building and corporate strategy”, Academy of Managing Journal. 33(2), p233~258, 1990.

[5] Formburn. C. & C. Van Rie. “The Reputational landscape”, Corporate Reputation Review,1(1), p5~13, 1997.

[6] 박슬라 역, “디지털 평판이 부를 결정한다”, 중앙북스, 2015.

[7] 이원준,“대학의 종합적 이미지 인식 측정을 위한 실증적 연구”, 예술인문사회융합멀티미디어논문지, 2015.

[8] 양민혁 외 3인,“SNS 데이터를 활용한 국내대학 인식 및 선호도 분석”, 한국빅데이터서비스학회 논문지(1/1), 2014.

[9] 송필준, 김종태, “대학 평가지표들에 대한 상관분석과 변수선택에 의한 선형모형추정”, 한국데이터정보과학회지, 2012.

[10] 홍연웅, “공공데이터 이용 활성화를 위한 정책에 관한 연구”, 한국데이터정보과학회지, 2014.

[11] IDG, “감성분석의 이해”, IDG Tech Report, 2014.

[12] http://word.snu.ac.kr/kosac/index.php

[13] http://konlpy.org/ko/latest/#api

〔저자소개〕

김 은 아 (Eun-Ah Kim)
1990년 2월 광운대학교 전자계산학과
1998년 2월 광운대학교 전자계산학과 이학석사
2008년 8월 군산대학교 컴퓨터공학과 박사과정 수료
1998년 3월 ~ 현재 한국폴리텍대학 인천캠퍼스 컴퓨터정보과 교수
관심분야 : 에이전트시스템, 평판시스템, 데이터분석
email : eakim@kopo.ac.kr

이 연 식 (Yon-Sik Lee)
1982년 2월 전남대학교 전자계산학과
1984년 2월 전남대학교 대학원 전자계산학 전공(이학 석사)
1994년 2월 전북대학교 대학원 전산응용공학 전공(공학박사)
1986년 3월 ~ 현재 국립군산대학교 컴퓨터정보통신공학부 교수
관심분야:Sensor network middleware, active rule system, agent system and edge computing.
email : yslee@kunsan.ac.kr