토픽 모델링을 활용한 교양 ICT 활용과정 서술형 강의평가 분석

[1]김효숙

Analysis of Descriptive Lecture Evaluation on Liberal Arts ICT utilization using Topic Modeling

[1]HyoSook Kim

요 약

본 연구의 목적은 교양 ICT활용 과정의 서술형 강의 평가에 대해 텍스트 마이닝의 토픽 모델링 분석을 실시하여 수강생의 강의 선택 요인과 강의에 대한 긍정적 · 부정적 요소 파악을 하고자 하는데 있다. 이를 위해 M 대학교의 2019년 2학기에 개설된 ICT활용 과정 강의에 대해 '강의를 신청한 이유', '강의에서 개선되어야 할 점'과 '강의에서 좋았던 점'에 대한 데이터 전처리부터 키워드 빈도 분석, 워드 클라우드 시각화 및 토픽 모델링 분석을 실시하였다. 연구결과 M 대학의 2019년 2학기 ICT활용 과정은 자격증 취득을 위해 강의를 신청하며, 동시에 자격증을 취득할 수 있어 강의가 좋았다는 긍정적 분석을 알 수 있다. 부정적 요소로 강의실 사용 환경 불편에 대한 것을 알 수 있다.

Abstract

The purpose of this study is to identify factors in selecting the elective ICT utilization lecture and to find positive and negative elements of the lecture through conducting topic modeling analysis of text mining of the narrative lecture evaluation. In order to do so, from pre-processing of data, keyword frequency analysis to wordcloud visualization and topic modeling analysis have been conducted from 'reasons of selecting the lecture,' 'improvements to be made on the lecture,' and 'what I liked about the lecture' categories regarding the ICT utilization lecture which was opened in the second semester of 2019 at M University. The analysis results show that students mostly registered for the ICT utilization lecture at M University to obtain a certificate and the fact being certified and taking the lecture can be done simultaneously is a positive element of taking the lecture. On the other hand, negative element included inconvenience of the classroom setting environment.

Keywords: *Topic Modeling, ICT utilization, Text Mining, Text analysis, Lecture Evaluation*

[1] *목원대학교 지능정보융합학과 박사과정(kkamti@mokwon..ac.kr)*

Received: Feb. 17, 2020, Revised: Mar. 09, 2020, Accepted: Mar. 29, 2020

I. 서론

대학에서 강의평가는 선택형 문항과 서술형 문항으로 구성되어 평가되고 있다. 양적 평가인 선택형 문항 평가는 과목의 특성을 반영하지 못한 획일적인 문항과 이로 인한 무성의한 답변으로 낮은 신뢰도가 발생되는 문제를 가지고 있다[1]. 또한 평가 항목 이외에 나타날 수 있는 학생들의 다양한 의견이 제한 될 수 있다[2]. 이런 선택형 문항의 제한점을 극복하기 위해 질적인 서술형 문항을 통해 학생들의 의견을 파악 할 필요가 있다.

서술형 문항은 학생들이 강의에서 중요하게 생각하는 부분에 대해 핵심적으로 서술하기에 학생들이 특히 어떤 부분에 관심을 두는지를 파악할 수 있는 장점을 가지고 있다 [3]. 즉, 강의 전반에 대한 심층적인 학생들의 의견을 다양하게 알아 볼 수 있을 것으로 기대된다. 반면 특정 수업이 아닌 대학 수업 전반의 개선점을 제공하는 데는 한계가 있다. 이는 학생들의 공통된 사항을 범주화 하기 어렵고 의미 있는 관계성을 파악하는 것이 쉽지 않기 때문이다[2]. 하지만 이러한 점에도 불구하고 학생들이 특정 수업에서 어떠한 요소들을 중요하게 고려하고 있고, 요소들간의 어떠한 관계를 가지고 있는지 파악할 수 있는 장점을 지닌다[2].

이런 텍스트 자료 분석을 위한 방법인 텍스트 마이닝(text mining)은 텍스트 요약, 정보 추출, 텍스트 분석, 문서 군집화, 언어 인식, 핵심문구 식별 등이 포함된다[4]. 특히, 텍스트 마이닝 방법 중 하나인 주제 분석 시 유용한 토픽 모델링(topic modeling)은 텍스트 데이터 내 단어들의 빈도수를 통계적으로 분석하여 전체 데이터에서 잠재적 주제인 토픽(topic)들을 자동으로 추출하여 분류한다. 즉, 방대한 텍스트 자료로부터 특정 주제를 추출하는 알고리즘으로 문서에 잠재된 토픽의 확률을 추정하는 통계적인 텍스트 처리기법이다[5].

따라서 본 연구에서는 M 대학교에 개설된 ICT 활용 교과의 서술형 강의평가 문항을 분석하여 학습자의 수강과목에 대한 만족도, 교수자의교수-학습 활동의 개선, 강좌 선택의 중요 정보 획득, 수업 환경의 요인들을 도출하여 분석의 결과를 ICT 활용 교과 운영에 반영하고자 한다.

이를 위해 학생들이 갖고 있는 ICT 활용 교과 강의에 대한 긍정적・부정적 요소 파악에 중점을 두어 다음과 같이 연구문제를 설정하였다.

첫째, ‘강의를 신청한 이유’, ‘강의에서 개선되어야 할 점’과 ‘강의에서 좋았던 점’에 나타난 주요 키워드와 빈도는 어떠한가?

둘째, ‘강의를 신청한 이유’, ‘강의에서 개선되어야 할 점’과 ‘강의에서 좋았던 점’의 응답에 나타난 주요 토픽특성은 어떠한가?

II. 관련 연구

2.1 서술형 강의평가

서술형 강의평가 응답을 분석한 연구에는 강의평가의 총점을 활용하여 상위 30%의 과목을 분류하여, 10 년간 매 학기별 강의평가 서술형 문항을 활용하여 전공계열별로 좋은 강의의 특성과 패턴을 분석하였다[6]. [7]에서는 강의평가 내용에서 문헌 연구를 통해 얻은 학습자 상호작용과 관련한 키워드를 추출하고 상호작용 점수를 도출하고 강의평가 점수와 비교하였다. 강의평가 내용에 대해 학생 변인, 교과목 변인에 따라 어떤 특성이 있는지를 분석하여 구체적인 수업 개선을 위한 시사점을 도출하였다[8]. 이 세 연구 모두 텍스트 마이닝 방법을 활용하였다는 점에서는 이 연구와 유사하나 ‘좋은 강의’ 나 ‘학습자 상호작용’, ‘학생 변인, 교과목 변인에 따른 특성 분석’에 대한 한정적인 주제를 다루었다는 측면에 차이점을 보인다. [3]는 ‘강의에서 개선되어야 할 점’과 ‘강의에서 좋았던 점’에 대한 문항 분석을 함에는 공통적이다. 그러나 대학 수업 전반에 대한 문항 평가와 일정한 표본 크기가 확보된 단과대학 대상 강의평가 분석이라 특정 수업 즉, ICT 활용 과정의 강의 평가 분석에 대한 긍정적•부정적 요소를 분석하는 본 연구와는 차이점을 갖는다. 각 연구자별 서술형 강의평가를 표 1 과 같이 나타내었다.

Table 1. Comparisons by Descriptive Lecture Evaluation Researcher

Researcher	Methodology	Data	Differences from Research results
HaeDeum Lee MinWoo Nam (2018)	Frequency analysis of Text Mining	Frequencies of those words out of upper 30% evaluation scores from 10 years longitudinal data	· Extract words with High Frequency · Characteristics and patterns of 'Better Class'
JungWoong choi DongKyu An (2016)	Cosine Similarity Text Mining	Extract Keywords Related to Student Interaction	· Derive Student-Interaction Score · Comparison of 5-point Lecture Evaluations
JongHo Shin Jaewon Choi (2019)	Text Mining	Evaluation of Descriptive Lectures from the first semester of 2014 to the first semester of 2018	·Analysis of Student variable and Subject variable
Minho Kwak Hyeree Min Meereem Kin (2019)	Topic Modeling	1,500 courses in the first semester of 2015	· Subject Analysis by College

2.2 토픽 모델링

토픽 모델링은 키워드 수에서 단일 단어의 의미 파악이 어렵다는 한계를 보완하기 위해 대량의 문서들에서 잠재되어 있는 전반적인 주제를 찾아내기 위한 데이터 마이닝 기법으로, 구조화되어 있지 않은 문서에서 중심 주제를 추출하는 알고리즘을 구성하여 주제를 찾아내고, 유사한 단어들끼리 군집화 하여 문서의 주제를 찾는데 사용된다[9][10][11]. 즉, 기존의 키워드 네트워크 분석만으로 알 수 없던 의미를 탐색할 수 있다.

토픽 모델링은 비구조화된 문서집합에서 잠재된 토픽들을 추출해주는 확률적 모델 알고리즘이다[10]. 또한 방대한 텍스트 자료로부터 특정 주제를 추출하는 알고리즘으로써 문서와 단어로 구성된 행렬(Document Term Matrix)을 사용하여 문서에 잠재된 토픽의 등장 확률을 추정하는 통계적인 텍스트 처리기법이다[5].

그림 1은 LDA 과정을 형식적으로 표현 한 것이다. D는 문서 집합, K는 토픽의 개수, α는 θ값을 결정하는 파라미터이며 η는 β값을 결정하는 파라미터이다. θ는 문서별 토픽의 비율, β는 토픽별 단어 w의 생성비율이며, $Z_{d,n}$은 문서 d의 n번째 단어의 토픽, $W_{d,n}$은 문서 d의 n번째 단어로 문서에서 관측되는 변수를 의미한다.

관측 값 $W_{d,n}$을 통하여 잠재변수인 θ_d, $Z_{d,n}$과 β_k를 찾는 것이 목적이다. θ_d의 파라미터 값은 α이고 θ_d의 분포로 디리클레 (Dirichlet) 분포를 따른다. 문서 d의 n번째 단어에 대한 토픽 $Z_{d,n}$은 θ_d를 통해 얻어지고, 단어 $W_{d,n}$은 $Z_{d,n}$과 전체 토픽 $\beta_{1:k}$로부터 얻을 수 있다[11].

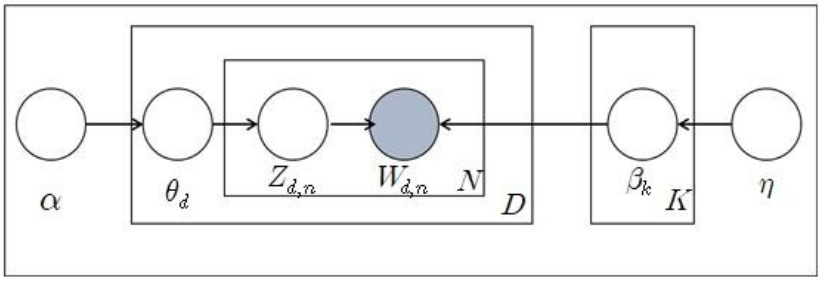

Figure 1. LDA Topic Modeling Process

이는 LDA 기법이 기계학습(Machine Learning)의 비지도 학습(Unsupervised Learning)방식을 사용하여 문서의 토픽을 얻는 것이 특징으로 다른 기법에 비해 결과 해석이 쉽고[10], 과적합(overfitting) 문제들을 해결하기 때문에 방대한 비정형 데이터로부터 다양한 토픽들을 도출하는데 유리하기 때문이다[13].

III. 토픽 모델링을 활용한 서술형 강의평가 분석

3.1 연구 대상

본 연구의 대상은 M 대학교 2019-2 학기 교양 ICT 활용 과정의 교과인 컴퓨터활용능력특강, ITQ 특강(한글/파워포인트), ITQ 특강(엑셀/인터넷), GTQ 특강, MOS 특강(Excel/PowerPoint), MOS 특강(Word/Access) 6 개 강의에 대한 서술형 강의평가 응답을 대상으로 하였다. '강의를 신청한 이유'와 '강의에서 개선되어야 할 점', '강의에서 좋았던 점'에 대한 257 명의 응답을 분석하였다. R 의 'topicmodels' 라이브러리를 활용하여 LDA 알고리즘 기반 토픽모델링을 수행하였다. 분석의 단위는 ICT 활용 과정 강의에 남긴 모든 학생들의 응답을 하나의 분석 단위로 간주하였다.

3.2 데이터 전처리

텍스트 데이터는 텍스트 마이닝을 활용한 분석이 쉬운 형태로 변환하기 위해서 그림 2 와 같이 데이터 전처리를 수행하였다.

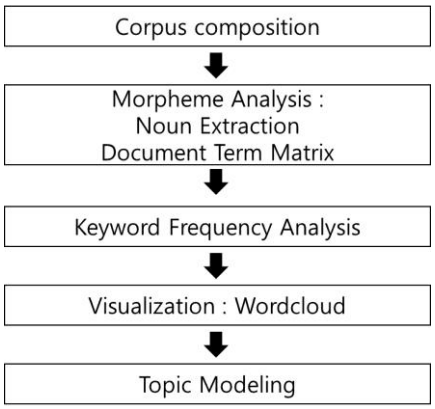

Figure 2. Data preprocessing Process

첫째, 응답 데이터를 말뭉치(Corpus)로 변환한 후, 불용어(stopword)를 제거하는 단계를 거친다. 불용어는 '그리고 또는 및'과 같은 분석에 불필요한 단어나 어구를 삭제하고 각종 문장부호와 특수문자 등을 제거하였다.

둘째, 형태소 분석은 전처리 과정을 거친 데이터를 품사 분석을 실행하여 명사 추출(extract noun) 작업과 길이가 두 글자 이상인 단어들만 걸러내어 단어-문서 행렬(DTM: Document Term Matrix)을 구성하였다. 위와 같은 작업을 위해 통계프로그램인 R 의 'tm', 'KoNLP', 'stringr'패키지를 사용하였다.

셋째, 전체 문서 내에 사용된 단어의 빈도에 따라 의미를 해석하는 방법으로, 서술형 강의평가 문서에서 자주 등장하는 키워드를 찾기 위해 문서 내 나타나는 단어의 총 빈도수를 사용하는 단어빈도(Term Frequency, TF)를 통해 키워드 빈도 분석과 워드 클라우드 분석으로 시각화하여 표 2 와 그림 1, 그림 2, 그림 3 과 같이 나타내었다.

넷째, 전체 문서의 토픽을 파악하기 위해 R 의 깁스 샘플링(Gibbs sampling) 알고리즘의 lda.collapsed.gibbs.sampler 함수로 단어들의 확률분포를 계산하여 LDA 분석을 실시하였다. 사후확률의 업데이트 횟수는 100 으로, 토픽 수는 2 개부터 5 개까지 변화시키면서 분석을 하였다.

IV. 연구 결과

4.1 키워드 빈도 분석

전체 문서 내에 사용된 단어의 빈도에 따라 의미를 해석하는 방법으로, 서술형 강의평가 문서에서 자주 등장하는 키워드를 찾기 위해 문서 내 나타나는 단어의 총 빈도수를 사용하는 단어 빈도수(Term Frequency, TF)를 통해 빈도수가 높은 주제어 10 개를 표 2 과 같이 키워드 빈도 분석을 하였다. 이때 '자격증', '자격'과 '취업', '취직'과 같은 하나의 의미로 생각되는 단어는 하나로 통일하였다. 또한 빈도수를 기준으로 워드 클라우드 분석을 진행한 결과 강좌를 신청한 이유의 워드 클라우드 그림 1 과 강의에서 개선할 점 워드 클라우드 그림 2, 강의에서 좋았던 점의 워드 클라우드 그림 3 와 같이 나타내었다.

Table 2. Keyword Frequency Analysis

Rank	reason of taking the course		Lesson Improvements		What was good about lecture	
	word	Frequency	word	Frequency	word	Frequency
1	Certificate	184	Computer	20	Certificate	53
2	acquisition	50	Lecture	12	Professor	42
3	Employment	21	Professor	10	Real time	12
4	Computer	9	Lesson	10	Lesson	12
5	help	6	Time	10	Explanation	10
6	apply	6	Good	10	acquisition	9
7	Photoshop	5	None	9	Lecture	8
8	Grades	5	Exam	8	Understand	8
9	PowerPoint	3	Lab	7	Exam	6
10	ability	3	Certificate	7	Questions	6

Figure 3. reason of taking the course

Figure 4. Lesson Improvements

Figure 5. What was good about lecture

표 2 와 그림 1, 그림 2, 그림 3 에서 알 수 있듯이 빈도수가 높은 주제어일수록 워드 클라우드에 나타나는 단어의 크기도 비례하는 것을 알 수 있다.

4.2 토픽 모델링 분석

토픽 모델링(topic modeling)은 언어 텍스트 집합을 가장 잘 표현하는 토픽(topic) 또는 주제 범주를 구분하는 기법이다. 전체 문서의 주된 주제를 파악하기 위해 R 의 'topicmodels' 패키지에서 제공하는 LDA 함수를 사용하여 토픽 모델링을 실시하였다. 토픽의 수는 주제어 간 중복 및 간섭이 발생하지 않고 주제 범주화가 가장 잘 되었다고 판단되어지는 2 개로 정하였다[14]. 토픽 모델링의 파라미터 값은 α : 0.1, β : 0.01, iterations: 100 으로 설정하였다[15][16].

Table 3. Topic modeling based analysis

Topic Number	Issue	Main Keyword and Content	
1	reason of taking the lecture	Keyword	Certificate, acquisition, Employment, Computer, ability
		content	Get a certificate for work
2	reason of taking the lecture	Keyword	Help, Photoshop, Certificate, PowerPoint, Grades
		content	Certificates and Grades Enrollment Preparing Photoshop and PowerPoint Skills
3	Lesson Improvements	Keyword	Lecture, Time, Classroom, Mic, Keyboard
		content	Improved classroom environment
4	Lesson Improvements	Keyword	Computer, Lesson, Professor, Exam, Lab, Certificate, schedule
		content	Improvements to computer, lecture, exam, and progress
5	What was good about lecture	Keyword	Professor, Real time, Explanation, Understand, Exam, Question, Kind
	What was good about lecture	content	Real time question, Professor's explanation is easy and kind. Positive elements of instructors and teaching-learning methods
6	What was good about lecture	Keyword	Certificate, acquisition, Lesson, Lecture, Time Exam, Response
	What was good about lecture	Content	Positive factors in certification, teaching, practice, and response

표 3 에 '강의를 신청한 이유', '강의에서 개선되어야 할 점'과 '강의에서 좋았던 점'에 대한 주제 분석 결과를 제시하였다.

토픽 별 단어 분포에 따른 주제 내용을 부여해 보면, 토픽 1 은 '취업을 위한 자격증 취득하기'로, 토픽 2 는 '자격증 취득과 학점 신청'이라는 주제 내용을 부여하였다.

토픽 3 은 '마이크', '키보드'와 단어 분포로 보아 '강의실 사용 환경 개선'. 토픽 4 는 '강의·시험·진도에 대한 개선사항'으로 명명하였다.

토픽 5 에서는 교수님, 실시간, 이해, 시험. 질문, 친절의 의미를 가진 단어들이 나타났다. 교수학습 방법에서 실시간 질문과 교수자에 대한 설명이 친절하고 이해가 좋다와 같은 '교수자와 교수학습 방법의 긍정적 요소'로 주제 내용을 부여 할 수 있고, 마지막으로 토픽 6 에서는 '자격증 취득과 강의·실습·응답·에 대한 긍정적 요소'로 명명하였다.

V. 결론

서술형 강의평가는 강의에서 학생들이 특히 어떤 부분에 관심을 두는지 핵심적으로 서술하기에 강의 전반에 대한 심층적인 학생들의 의견을 다양하게 알아 볼 수 있다. 본 연구에서는 M 대학교 2019-2 학기에 개설된 ICT 활용 교과의 서술형 강의 평가 분석 결과를 정리하면 다음과 같다.

첫째, ‘자격증’에 대한 응답이 눈에 뜨게 가장 많다. 이는 학생들이 강의를 신청한 이유가 ‘자격증 취득’을 위한 것임을 알 수 있다. 또한 강의에서 좋았던 점도 ‘자격증 취득’이라는 분석을 볼 수 있다. 이는 강의를 신청한 이유와 강의에서 좋았던 점 모두 ‘자격증 취득’이라는 공통 사항이 있음을 알 수 있다.

둘째, 강의실 사용 환경 불편, 실습실 컴퓨터의 개선과, 강의 · 시험 · 진도에 대한 개선사항이 필요함을 알 수 있다.

이와 같은 연구결과를 통해 교양 교육과정으로 운영되고 있는 ICT 활용 과정에 대해 학생들은 취업을 위해서는 컴퓨터 관련 자격증이 필요하고 자격증 취득을 위해 강의를 신청한다는 것을 알 수 있었다.

그러나 이 연구는 한 개 대학의 ICT 활용 과정의 강의평가를 활용했기 때문에 연구 결과를 일반화하기 어렵다는 한계가 있다. 하지만 서술형 강의평가를 통해 학생들의 강의 선택요인이 무엇인지 파악 할 수 있고 교양 교육과정 편성 및 강좌 구성에 중요 변인으로 활용될 것으로 기대된다.

VI. 참고문헌

[1] O-Young Kwon, Young-Tae Park, Il-Kyu Hwang, Tae-Won Ahn, Kyeong-Sook Kim, “A Study on the Reliability Improvement for the Course Evaluation System,” Journal of Engineering Education Research, Vol. 17, No. 2, pp. 35~41, March, 2014.

[2] Lee, Hoo-Hee, Lee, Sang-Soo, Lee, Soo-Sang, Kim, Eun-Jung, “Semantic Network Analysis of College Students,” Journal of Educational Innovation Research, Vol. 28, No. 2, pp.237-262, 2018.

[3] Minho Kwak, Hyeree Min, Meereem Kim, “Analysis of Students Open-Ended Course Evaluation Using Topic Modeling,” Asian Journal of Education, Vol. 20, No. 2, pp.491-522, June, 2019.

[4] Khan, R.A. and Kanth, S. “Text Mining: Knowledge Discovery from Unstructured data, Artifical Intelligent Systems and Machine Learning”, 8(2), 71-77, 2016

[5] Baek Young, Y. Text Mining using R, Paju: HanulAcademy, 2017.

[6] HaeDeum Lee, MinWoo Nam, “Better Class’s Characteristics by Major Field based on the Analysis of Text Mining from College Course Evaluation Subjective Results”, Research Institute for Early Childhood Education, Vol.20, no.2, pp. 21-41, 2018

[7] JungWoong Choi, Dongkyu An “A Study on the Data Analysis of the Written Comments in Lecture Evaluation”, Journal of Digital Convergence Vol, 14, no.11, pp. 101-106, 2016

[8] JongHo Shin, Jaewon Choi, “Text mining Analysis of College Students’ Descriptive Course Evaluation”, Journal of Learner-Centered Curriculum and Instruction, Vol.19, no.16, pp. 77-99, 2019

[9] SungHoon Seo, HakYeon Lee, “Fintech trend analysis using topic modeling of BM patents”, The Korean Institute of Industrial Engineers fall conference, pp. 471-480, 2015

[10] David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol 3, pp. 993-1022, JAN, 2003

[11] JuSeop Park, SoonGoo Hong, JongWeon Kim. “A Study on Science Technology Trend and Prediction Using Topic Modeling”, Journal of the Korea Industrial Information Systems Research, Vol 22, No 4, pp. 19-28, 2017

[12] Blei, D., Ng, A. and Jordan, M., ‘Latent Dirichlet Allocation,’ Journal of Machine Learning Research, Vol. 3, pp. 993- 1022, 2003.

[13] Blei, D. M. (2012). Probabilistic topic models, Communications of the ACM, 55(4), 77-84.

[14] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics, P roceedings of the National academy of Sciences , 101(suppl1), 5228-5235.

[15] SooSang Lee, “A Study on the Application of Topic Modeling for the Book Report Text”, Journal of Korean Library and Information Science Society, Vol 47(4), pp. 1-18, 2016

[16] Marwa Naili, Anja Chaibi, Henda Ghézala. “Arabic topic identification based on empirical studies of topic models”, Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées, 27. 45-49. 2017

저자 소개

김효숙(*HyoSook Kim*)

2018 년 3 월~ 현재 목원대학교 일반대학원 지능정보융합학과 박사과정

관심분야 : 인공지능, 자연어처리, 텍스트마이닝, 소프트웨어교육,융합교육