JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 21, No. 2 : 423~428 Feb. 2017

빅데이터 분석을 이용한 이러닝 수강 후기 분석

김장영[1*] · 박은혜[2]

e-Learning Course Reviews Analysis based on Big Data Analytics

Jang-Young Kim[1*] · Eun-Hye Park[2]

[1*]Department of Computer Science, The University of Suwon, Hwaseong 18323, Korea
[2]Department of Computer Science, The University of Suwon, Hwaseong 18323, Korea

요 약

인터넷과 스마트 기기의 사용량 증가로 인해 다양한 교육정보와 많은 양의 데이터가 생성되어 빠르게 확산되고 있다. 최근 이러닝 이용률이 증가하면서 발생하는 빅데이터를 활용하여 학습자들의 교육 성과와 교육 시스템의 효과성을 극대화 하는 것을 목표로 하는 교육 데이터 관련 연구 분야에 대한 관심이 높아지고 있으며 온라인에서 학습자들이 학습한 수많은 기록과 데이터들이 정보로 쌓이게 된다. 이에 본 논문에서는 이러닝 학습자들이 시스템에 남긴 수강 기록을 기반으로 학습자 현황에 대해 객관적으로 파악할 수 있도록 신경망 알고리즘인 Word2Vec을 적용하여 단어 간 유사도를 구하고 클러스터링 알고리즘을 이용하여 군집화 하였다. Word2vec을 이용하여 학습을 시키면 연관된 의미의 단어가 나타나게 되고 학습을 반복해 나가는 과정에서 점차 가까운 벡터를 지니게 된다. 또한 클러스터 알고리즘을 이용하여 명사, 동사, 형용사, 부사가 중심점에서 최소의 거리를 두고 같은 거리에 위치해 있음을 실험 검증하였다.

ABSTRACT

These days, various and tons of education information are rapidly increasing and spreading due to Internet and smart devices usage. Recently, as e-Learning usage increasing, many instructors and students (learners) need to set a goal to maximize learners' result of education and education system efficiency based on big data analytics via online recorded education historical data. In this paper, the author applied Word2Vec algorithm (neural network algorithm) to find similarity among education words and classification by clustering algorithm in order to objectively recognize and analyze online recorded education historical data. When the author applied the Word2Vec algorithm to education words, related-meaning words can be found, classified and get a similar vector values via learning repetition. In addition, through experimental results, the author proved the part of speech (noun, verb, adjective and adverb) have same shortest distance from the centroid by using clustering algorithm.

키워드 : 이러닝, 빅데이터, Word2Vec, 신경망, 클러스터링

Key word : e-learning, Big Data, Word2Vec, Neural network, Clustering

Received 07 November 2016, **Revised** 08 November 2016, **Accepted** 21 November 2016

* **Corresponding Author** Jang-Young Kim (E-mail: jykim77@suwon.ac.kr, Tel: +82-31-229-8345)
Department of Computer Science, The University of Suwon, Hwaseong 18323, Korea

Open Access http://doi.org/10.6109/jkiice.2017.21.2.423 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

IT 기술의 발전과 사회적 변화로 교육의 영역이 확대되고 있으며 각종 기관이나 학교, 회사 등에서 '이러닝'이라는 온라인 학습이 활발히 사용되고 있다. 이러닝은 시간과 장소에 관계없이 언제, 어디서, 누구나 자유롭게 학습할 수 있는 환경을 말하며, 그 이용도가 높아지고 있어 지속적인 성장이 기대된다[1, 2].

더불어 인터넷과 스마트 기기의 사용량 증가로 인해 다양한 정보와 많은 양의 데이터가 생성되어 빠르게 확산되고 있다. 특히 정보통신기술의 발달로 사람뿐만 아니라 사물 간에도 네트워크로 연결되어 대량의 데이터를 발생시키고 있으며, 이러한 데이터를 바탕으로 가치 있는 정보를 찾아내서 분석하고 효율적으로 사용하는 것을 빅데이터(Big Data)라고 한다. 최근 이러닝 이용률이 증가하면서 발생하는 빅데이터를 활용하여 학습자들의 교육 성과와 교육 시스템의 효과성을 극대화 하는 것을 목표로 하는 교육 데이터 관련 연구 분야에 대한 관심이 높아지고 있다[3, 4].

또한 전통적인 방식의 교육에서는 학습활동과 관련된 대부분의 데이터가 교육이 끝난 뒤 모두 없어지는데, 온라인에서는 학습자들이 학습한 수많은 기록과 데이터들이 디지털 정보로 쌓이게 된다. 이에 본 논문에서는 이러닝 학습자들이 시스템에 남긴 수강 기록을 기반으로 학습자 현황에 대해 객관적으로 파악할 수 있도록 신경망 알고리즘인 Word2Vec을 적용하여 단어 간 유사도를 구하고 클러스터링 알고리즘을 이용하여 군집화 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 알아보고, 3장에서는 연구 방법을 설명한다. 그리고 4장에서는 연구 방법을 통한 실험 결과와 평가를 기술하고 5장에서는 결론 및 향후 연구 과제를 제시한다.

Ⅱ. 관련연구

2.1. 빅데이터

빅데이터의 3가지 특징은 데이터의 양(Volume), 데이터의 속도(Velocity), 데이터의 다양성(Variety)에 가치(Value)를 추가하여 4V라고 한다.

빅데이터가 주목받는 이유는 IT를 활용한 다양한 분야의 대용량의 데이터가 급증하면서 정형화된 데이터에서 비정형 데이터까지 범위가 넓어지고 있으며 가공되지 않은 데이터의 가치가 높아지고 있기 때문이다 [5].

그리고 교육 환경을 변화시키는 방안으로 데이터 안에 숨겨진 패턴을 찾고 다양한 분석 기술을 사용하여 의미 있는 데이터를 찾아내는 것이 중요하며, 이러한 데이터와 정보를 효율적으로 관리해야 한다.

2.2. 이러닝

학습이란 어느 시대에나 존재해왔고 학습에 있어 시공간의 제약을 벗어나기 위한 노력으로 우편 교육에서 방송 교육으로 환경이 변화하였으며, 정보통신기술의 발달로 다양한 전자 기기를 사용해 학습자들이 학습하고 관련 지식과 정보에 접근할 수 있는 이러닝 환경이 조성되었다.

기존의 전통적인 교육에서는 교수자가 일방적으로 지식을 전달하는 강의식 수업이었고 학습자들과 함께 상호작용하는 과정은 사라지고 시험의 결과인 성적만 기록되었다[6]. 하지만 이러닝에서는 교수자와 학습자, 학습자와 학습자들 간의 학습 활동과 관련된 모든 데이터들이 디지털화 되어 쌓이고 있으며, 여기서 발생하는 데이터를 분석하여 학습자들을 파악하고자 한다.

Ⅲ. 연구 방법

3.1. Crawler

검색 엔진의 근간으로 웹 크롤러, 스파이더 로봇 등 다양한 이름으로 불리며 웹 페이지에서 각종 정보를 자동적으로 수집하는 프로그램이다. 사용자가 웹페이지 링크를 일일이 따라가 정보를 얻는 작업을 대신하여 자동적으로 순회하며 내용을 분석하고 수집한다. Java의 Jsoup 라이브러리를 사용해 크롤러를 만들고 이러닝 학습이 가능한 지안에듀 (http://www.algisa.com) 사이트를 대상으로 선정하였다. 데이터를 수집할 때 제일 먼저 고려한 사항은 학생들의 의견이 정확히 나와 있고 유의미한 연구 결과를 보여 줄 수 있는 데이터 규모를 고려하여 1,200여 건의 수강 후기를 수집하였다.

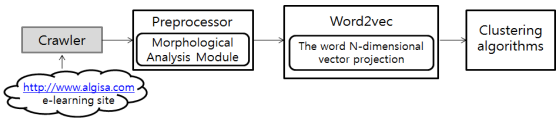

Fig. 1 Course analysis reviews overview using Machine Learning Techniques

그림 1은 전체적인 연구 과정을 도식화한 것이다.

3.2. 형태소 분석

형태소 분석은 어절을 형태소 단위로 분리하고 각 형태소에 적절한 품사를 부여하는 과정을 말한다. 구문 분석의 전 단계로 문장을 구성하는 어절에 들어 있는 형태소를 분리하고, 분리된 형태소 간의 결합 관계를 분석하여 불필요한 부분을 제거한다.

Crawler에서 수집된 데이터를 분석하여 전 처리하고 표 1과 표 2에 긍정적 극성을 보이는 어휘와 부정적 극성을 보이는 어휘로 분류하여 정리하였다.

Table. 1 Examples of positive polarity

Part of speech	Positive characteristic vocabulary
Noun	jackpot, impression, thank
Verb	enjoy, concentrate, appreciate
Adjective	kind, good, beautiful
Adverb	perfectly, truly

Table. 2 Examples of negative polarity

Part of speech	The negative characteristics Vocabulary
Noun	disappointment, burden, threat
Verb	disgust, dislike
Adjective	bad, exhausting
Adverb	untruly

3.3. Word2vec

인공 신경망(Neural network) 연구에서 시작되었으며 단어 간 유사도를 구하기 위해 Word2vec을 이용하였다 [7,8]. 형태소 분석을 통해 나온 단어들을 인공 신경망에 학습을 시키면 연관된 의미의 단어가 나타나게 되고 학습을 반복해 나가는 과정에서 점차 가까운 벡터를 지니게 된다 [9].

따라서 아주 추상적인 동사나 형용사는 학습이 명사에 비해 어려울 수 있다. 다만 수 없이 많은 데이터를 보면 동사들이 어떤 목적어를 가지는지 규칙성을 파악함으로 어느 정도 동사들 간의 의미 관계도 학습이 가능하다고 볼 수 있다. Word2vec 모델은 C, Python, Java, Go, Scala 등 다양한 언어로 구현이 가능하며 본 논문에서는 Java program으로 구현하였다.

긍정적 극성을 보이는 단어 3개와 부정적 극성을 보이는 단어 3개를 품사별로 2차원의 벡터로 표현을 하니 그림 2와 같이 학습 전에는 랜덤으로 분포하였으나, 학습 후에는 그림 3과 같이 단어들은 극성끼리 모여 있는 것을 볼 수 있다. 또한 Word2vec 알고리즘은 비지도 학습이므로 그래프 상의 위치는 임의로 결정된다. 그림 2는 학습 전 랜덤 분포를 나타내고 그림3은 학습 후 긍정과 부정, 즉 같은 극성끼리 모인 것을 보여준다. 따라서 이러닝에서 학습자들의 긍정적인 반응과 부정적인 반응을 파악하여 이러닝 강의를 수정하거나 보완가능하다는 결론이다.

Fig. 2 Before learning (random distribution)

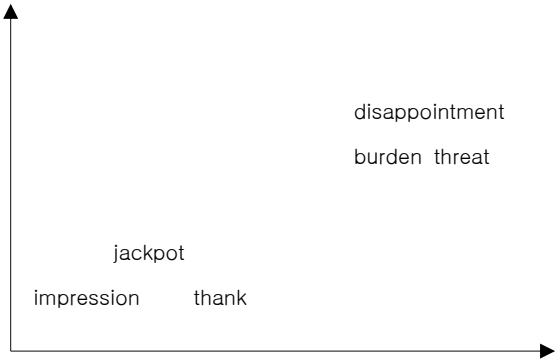

Fig. 3 After learning

Ⅳ. 실험 결과

표 1, 표 2에서 형태소 분석기로 분류한 명사, 동사, 형용사, 부사로 구분된 어휘들이 품사와 관계없이 같은 수준으로 클러스터링 되는지 아니면 품사에 따라 다른 수준으로 클러스터링 되는지 알아보기 위해 실험을 진행하였다.

V= $\{v_{1,} v_2, \ldots v_n\}$ (예를 들면 v1=impression, v2=jackpot, v3=thank, v4=disappointment, v5=burden, v6=threat 이다.) (1)

위 (1) 식에서 V를 2개의 클러스터로 나누는데 오류를 최소화하여야 한다. 즉 군집 중심점과 각 객체간의 거리의 합이 최소가 되어야 한다. 이를 수학식 (2) 로 표현하면 다음과 같다.

$$\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\ldots+(p_n-q_n)^2}$$

$$=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \quad (2)$$

(여기서 p는 v의 좌표이고 q는 중심점의 좌표이다.)

클러스터링 수준을 알아보기 위해 배타적(Exclusive) 군집분석 방법인 K-Means Clustering을 사용하게 되면 모든 객체에 대하여 각 군집 중심점과의 거리를 측정하여 가장 가까운 군집에 할당하고 할당된 군집을 대상으로 새로운 군집 중심점을 할당하는 작업을 군집 간 이동이 없을 때까지 계속 반복하게 된다.

$$\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} = \text{Error} \quad (3)$$

Error의 최소 값을 구하는 과정 (3)은 무게 중심(center of gravity)을 구하는 알고리즘과 같다. 무게 중심을 구하는 알고리즘을 사용하여 Word2vec이 클러스터링 마친 객체들의 무게 중심점과 각 객체간의 거리의 합을 구해서 표 3,표 4에 정리하였고 무게 중심(center of gravity)을 구하는 알고리즘은 다음과 같다(4), (5).

Algorithm
Input : k (the number of cluster), Set V of n objects
Output : A set of k clusters which minimizes the sum of distance error criterion (4)

Method :
Choose k objects as the initial cluster centers; set i =0
Loop for each object, p, in V
For each object, p, in V
    the NearestCenter(p), and assign p to it
Compute mean of cluster as the new centers (5)

Table. 3 Positive polarity results

Part of speech	Positive characteristic vocabulary
Noun	0.187642121015318035
Verb	0.187642121015318035
Adjective	0.187642121015318035
Adverb	0.187642121015318035

Table. 4 Negative polarity results

Part of speech	The negative characteristics Vocabulary
Noun	0.184312126970434154
Verb	0.184312126970434154
Adjective	0.184312126970434154
Adverb	0.184312126970434154

실험 결과 표 3, 표 4와 같이 명사, 동사, 형용사, 부사가 모두 동일한 결과를 보였다. 중심점에서 최소의 거리를 두고 세 개의 단어가 같은 거리에 위치해 있다. Google에서 제공하는 패키지를 사용해서 Word2vec을 만들게 되면 훈련을 위해 입력된 단어의 품사와는 관계없이 동일한 기계학습 환경이 구성된다는 것을 알 수 있다. 표 3, 표 4의 실험 결과는 중심점과 객체간의 거리의 합을 나타낸다.

이때 거리의 합이 0이면 모든 개체가 중심점에서 일치함을 뜻한다. 표3, 4 값은 동일 클러스터링 적용 후 모든 단어들이 같은 거리에 위치해 있음을 보여준다. 따라서 기계학습을 적용할 경우 다양한 알고리즘을 사용해야 할 것이다.

Ⅴ. 결 론

본 논문에서는 이러닝 학습자들이 시스템에 남긴 수강 기록을 기반으로 Word2Vec을 적용하여 단어 간 유사도를 구하고 클러스터링 알고리즘을 이용하여 군집화 하였다. 크롤러를 사용하여 수강 후기 데이터를 모은 다음 형태소 분석을 통해 어휘를 구분하였다.

Word2vec을 이용하여 학습을 시키면 연관된 의미의 단어가 나타나게 되고 학습을 반복해 나가는 과정에서 점차 가까운 벡터를 지니게 된다. 또한 클러스터 알고리즘을 이용하여 명사, 동사, 형용사, 부사가 중심점에서 최소의 거리를 두고 같은 거리에 위치해 있음을 실험 검증하였다. 향후에는 형태소 분석기로 문장을 단어로 나눌 필요 없이 직접 문장을 학습하게 함으로써 정확도의 향상을 기대해 본다. 또한 빅데이터 분석을 이용한 알고리즘 구현에도 중요한 연구가 될 것이다.

Acknowledgements

The paper was supported by The research grant of the University of Suwon in 2016.

REFERENCES

[1] J. Lee, “What Drives a Successful e-Learning: Focusing on the Critical Factors Influencing e-Learning Satisfaction,” *Korea Jounal of Business Administration*, vol. 24, no. 4, pp. 2245-2257, Aug. 2011.

[2] J. Park, and D. Lee, “Proposal of Smart era Online Learning Model with BigData,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 19, no. 4, pp. 991-1000 Apr. 2015.

[3] J. Shin, J. Choi, and W. Koh, “A study on the Use of Learning Analytics in Higher Education: Focusing on the perspective of professors,” *Journal of Educational Technology*, vol. 31, no. 2, pp. 223-252, Feb. 2015.

[4] Y. Yun, H. Ji, “A development of Open Social Learning Platform for learning analytics and educational data mining,” *Journal of Korean Institute of Information Scientists and Engineers*, vol. 23, no. 12, pp. 1349-1351, Dec. 2015.

[5] H. Yoon, “Research on the Application Methods of Big Data within the Cultural Industry,” *Academic Association of Global Cultural Contents*, vol. 10, no. 1, pp. 157-179, Feb. 2013.

[6] J. Choi, “Applications of Educational Big Data Generated in Smart Education,” *Journal of Korea Intelligent Information System Society*, vol. 20, no. 5, pp. 144-148, May 2015.

[7] J. Lee, “Three-Step Probabilistic Model for Korean Morphological Analysis,” *Journal of KIISE: Software and Applications*, vol.38, no.5, pp.257-268, May 2011.

[8] L. Wolf, Y. Hanani, K. Bar, and N. Dershowitz, “Joint word2vec networks for bilingual semantic representations,” *International Journal of Computational Linguistics and Applications*, vol. 5, no.1, pp. 27-44, Feb. 2014.

[9] S. Kim, and S. Lee, “Automatic Extraction of Alternative Word Candidates using the Word2vec model,” *Journal of Korean Institute of Information Scientists and Engineers*, vol. 23, no. 12, pp. 769-771, Dec. 2015.

김장영(Jang-Young Kim)

2005년 2월: 연세대학교 컴퓨터과학 공학사
2010년 5월: Pennsylvania State Univ. 공학석사
2013년 7월: State University of New York 공학박사
2013년 8월: University of South Carolina 조교수
2014년 3월: 수원대학교 컴퓨터학과 조교수
※관심분야 : Big data, Cloud computing, Networks

박은혜(Eun-Hye Park)

2011년 2월: 수원대학교 컴퓨터학과 공학사
2016년 3월: 수원대학교 컴퓨터학과 교육대학원
※관심분야 : Big data, 컴퓨터교육