JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 18, No. 2, pp. 339-346, Apr. 2017

교수-학습지원시스템에서 학습자 질의응답 자동분류를 위한 토픽 모델링

김 경 록[1] · 송 혜 진[2] · 문 남 미[2*]
[1]호서대학교 전자디스플레이공학부
[2]호서대학교 컴퓨터정보공학부

Topic modeling for automatic classification of learner question and answer in teaching-learning support system

Kyungrog Kim[1] · Hye jin Song[2] · Nammee Moon[2*]
[1]Department of Electronic Display Engineering, Hoseo University, Chungcheongnam-do 31499, Korea
[2*]Department of Computer Engineering, Hoseo University, Chungcheongnam-do 31499, Korea

[요 약]

기사와 댓글, 질의응답과 같은 비정형 데이터에 기반한 텍스트 분석에 대한 관심이 증가하고 있다. 이는 사람들의 견해인 비정형 텍스트 데이터로부터 특징을 파악하고, 평가, 예측 및 추천에 활용할 수 있기 때문이다. TEL 분야에서도 MOOC 서비스의 확대로 교수학습지원시스템 기반 토론, 질의응답 서비스를 자동화하기 위한 관심이 증가하고 있다. 시스템에 축적된 질의응답 데이터를 기반으로 질의 토픽을 생성하고, 새로운 질의에 대해 토픽을 자동분류하기 위해서이다. 따라서 본 연구에서는 새로운 질의 토픽을 자동분류 할 수 있도록 LDA기법을 활용한 토픽 모델링을 제안하고자 한다. 이를 바탕으로 질의 토픽 사전을 생성하고 새로운 질의에 대해 토픽을 자동분류 할 수 있다. 일부 질의에서는 0.7 이상의 높은 자동 분류를 보였으며, 새로운 질의가 여러 토픽에 포함될수록 좀 더 좋은 자동분류 결과를 보였다.

[Abstract]

There is increasing interest in text analysis based on unstructured data such as articles and comments, questions and answers. This is because they can be used to identify, evaluate, predict, and recommend features from unstructured text data, which is the opinion of people. The same holds true for TEL, where the MOOC service has evolved to automate debating, questioning and answering services based on the teaching-learning support system in order to generate question topics and to automatically classify the topics relevant to new questions based on question and answer data accumulated in the system. Therefore, in this study, we propose topic modeling using LDA to automatically classify new query topics. The proposed method enables the generation of a dictionary of question topics and the automatic classification of topics relevant to new questions. Experimentation showed high automatic classification of over 0.7 in some queries. The more new queries were included in the various topics, the better the automatic classification results.

색인어 : 데이터마이닝, 질문과대답, 유사성, 텍스트마이닝, 토픽모델링
Key word : Data Mining, Question and Answer, Similarity, Text Mining, Topic Modeling

http://dx.doi.org/10.9728/dcs.2017.18.2.339

Received 30 March 2017; **Revised** 06 April 2017
Accepted 25 April 2017

***Corresponding Author; Nammee Moon**

Tel: +82-041-540-9665
E-mail: it4all@hoseo.edu

Ⅰ. 서 론

자신 또는 주변의 문제들을 해결하기 위해 주변 지인들에게 묻고 답하는 일련의 행위가 우리에게는 아주 자연스러운 일부분이다. ICT환경이 발달하면서 유선과 무선의 통합, 통신과 방송의 융합, 온라인과 오프라인의 결합이 이루어지고, 인터넷과 소셜 미디어 사용이 일상 생활화 되고 있다. 이에 따라, 질의응답 행위도 온라인상에서 커뮤니티를 형성하여 자연스럽게 이루어지고 있다.

온라인상에서의 질의응답의 경우는 주로 텍스트를 기반으로 하며, 이는 비정형의 데이터이다. 신문기사처럼 정제된 글이 아닌 자유롭게 기술되는 비형식(informal)의 텍스트 문서이다. 이러한 비형식 텍스트 문서에서 일관된 규칙이나 패턴을 찾는 일은 형식(formal) 문서 경우에 비해 용이하지 않기 때문에, 이를 해결하기 위해 질의 데이터로부터 정보 추출 및 연관성 분석 연구와 기계학습을 이용하여 자동분류하고 응답하기 위한 연구들이 증가하고 있다[1]. 질의 주제와 특성을 이해하고 이를 분류할 수 있다면, 자동응답 체계를 구축하여 보다 신속하게 질의 사항에 대해 응답을 할 수 있기 때문이다[2].

TEL(Technology Enhanced Learning)분야에서도 교수학습지원시스템을 이용하여 교육 내용에 관한 것 뿐만 아니라 교육 환경, 시스템 사용, 수업 등에 대한 다양한 질의응답이 이루어지고 있다 [3][4]. 이러한 활동은 지속적이며 반복적으로 이루어지고 있다. 이렇게 생성된 질의는 집단지성의 결과물이다. 그러나 이러한 질의응답은 텍스트 형태의 데이터로 비정형의 자연언어의 형태로 작성되고 있으며, 체계적으로 관리하기가 어렵다. 교수학습지원시스템 질의응답 코너를 보면, 같은 질문이 여러 번 반복하여 올라온 경우를 볼 수 있다. 이미 답변이 완료된 질문들을 쉽게 찾을 수 없기 때문이기도 하다. 또한, 유사한 질문에 대한 답변이 담당자에 따라 조금씩 다른 경우도 있다. 이 같은 문제를 해결하기 위해서 메타데이터를 이용하여 질의들을 분류하고 군집화하는 작업을 통해 좀 더 정확한 답변을 할 수 있을 것이다. 그러나 모든 질의사항에 대해 반복되는 작업을 좀 더 효율적으로 하기 위해서는 새로운 방법이 필요하다[5].
이를 위해 본 연구에서는 교수학습지원시스템에서 발생하는 질의 데이터를 바탕으로 자동 분류를 위한 Topic-driven 모델링을 제안하고자 한다. 이를 통해 질의를 정확히 파악하고 분류하여 답변을 함으로써 응답 시간을 줄이고 답변의 정확도도 높일 수 있을 것이다.

본 논문의 구성은 다음과 같다. 2장에서 이론적 배경인 토픽 마이닝, 토픽 모델링, LDA 에 대해 알아보고, 3장에서는 LAD를 기반으로 토픽 모델링을 설계한다. 4장에서는 토픽모델을 적용하여 교수학습활동 질의 데이터에 대해 자동분류 실험을 하고 이를 해석한다. 그리고 5장에서 결론을 도출하고자 한다.

Ⅱ. 관 련 연 구

2-1 토픽 마이닝

토픽 마이닝은 문서에 사용된 단어 분포를 이용하여 문서들을 군집화하기 위해, 문서에 사용된 모든 단어를 말뭉치로 취급한다. 하나의 문서는 여러 토픽이 구성되어 만들어진 것이라 할 수 있다. 즉, 토픽1은 20%, 토픽2는 30%, 토픽3은 50%의 분포를 가질 수 있다. 문서에서 등장하는 토픽 분포 값에 따라 군집화 할 수 있다. 따라서 문서의 특성을 고려한 모델을 설계하여 토픽을 분석할 수 있다. 즉, 토픽 마이닝을 위한 토픽 모델은 분석을 목표로 하는 자료와 파악하기 위한 내용에 따라 각기 다른 방식으로 설계할 수 있다[5].
Nubbi (2009) 는 두 가지 종류의 토픽을 분석해 내기 위해, 하나는 각 문서에서 등장하는 각 개체의 단어 분포도를 이용하고, 다른 하나는 각 개체 쌍의 단어 분포도를 이용하는 방법을 제안했다[6]. Rosen-Zvi (2004) 은 논문에서 각 저자의 관심분야에 대한 토픽 분포를 연구하기 위해 Author Topic Model을 제안했다[7]. Chang (2009)은 문서 내 개체 사이의 관계를 파악하기 위해 개체 토픽 모델을 제안했다. 이는 문서 내에서 등장하는 개체 간에 어떠한 관계를 가지는지를 분석하는 것을 목적으로 설계된 것이다[5][6].

2-2 토픽 모델링

토픽 모델링은 구조화되지 않은 방대한 문헌에서 주제를 찾아내기 위한 알고리즘으로, 맥락과 관련된 단서들을 이용하여 유사한 의미를 가진 단어들을 군집화하는 방식으로 주제를 추론하는 방법이다[8]. 즉, 토픽 모델링은 비구조적 데이터를 분석하기 위한 알고리즘이다. 토픽 모델링은 크게 벡터기반(Vector-based techniques)과 확률기반(Probabilistic techniques)로 나눌 수 있다. 벡터 기반으로는 Latent Semantic Analysis (LSA)이 있다. 확률 기반으로는 Probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Allocation (LDA) 등이 있다. 특히, LDA는 무감독 생성 주제 모델(Unsupervised generative topic model)로, 주제 분류나 문서 사이 유사도 계산에 많이 이용되는 알고리즘이다[9][10][11].

문서(Document)는 토픽(Topic)의 집합이다. 토픽(Topic)은 단어(word)의 집합이다. 즉, 단어가 모이면 토픽이 되고, 토픽이 모이면 문서가 된다. 각 단어는 토픽을 대표할 수 있고, 토픽의 집합이 문서가 된다. 여기서, 토픽(Topic)은 숨어있기 때문에 이를 찾기 위해 LDA 알고리즘을 활용한다[10][12]. 즉, 문서에서 각 단어가 몇 번이나 토픽에 속하는지를 파악해내는 것이다. 질의 및 응답 관점에서 질의-토픽-단어 사이의 관계를 그림으로 표현하면, 아래 그림1과 같다. 나아가 질의와 질의 사이의 관계는 그림2와 같이 표현할 수 있다. 본 연구에서는 이를 이용하여 질문에 대한 토픽을 모델링 하고자 한다.

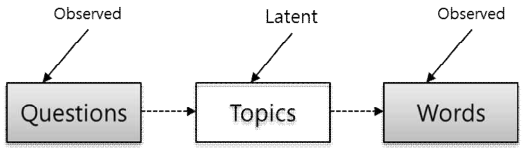

그림 1. Questions-Topics-Words 관계도
Fig. 1. Questions-Topics-Words relationship diagram

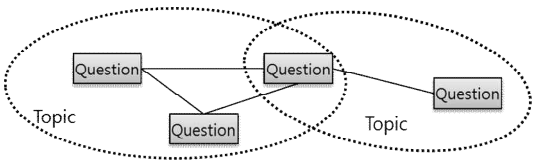

그림 2. Topic-Topic 연관 관계도
Fig. 2. Topic-Topic association diagram

2-3 잠재 디리클레 할당(Latent Dirichlet Allocation)

잠재 디리클레 할당(Latent Dirichlet Allocation)은 문서 내 혹은 문서 사이 단어들의 분포를 통해 토픽을 찾아내기 위한 생성적 확률 모델(generative probabilistic model)이다. 즉, 주어진 문서가 잠재적으로 갖는 주제들을 추론해 내는 확률 모델로, 단어 집합과 같은 이산형 데이터를 모델링하기 위한 것이다. LDA에서 한 문서는 여러 토픽들의 혼합으로 구성되어 있으며 이에 대한 확률분포를 갖는다고 가정한다. 또한 하나의 토픽은 여러 단어들의 혼합으로 구성되며 이에 대한 확률분포를 갖는다고 가정한다. 이를 바탕으로 문서 내 혹은 문서 사이의 단어 빈도수를 분석하여 문서에 대한 토픽 분포, 토픽에 대한 단어 분포를 추정한다[13]. 이러한 특성 때문에, 주제별, 특성별, 분석을 위해 많이 이용된다.

LDA는 단어의 분포와 토픽의 분포를 추정하여 각 문서를 생성한다. 그림3에서 (ω)는 하나의 문서를 의미하고, (z)는 토픽을 의미한다. 단어는 토픽으로부터 생성되며 해당 문서가 어떠한 토픽비율(Θ)을 가질 것인지는 디리클레 분포로부터 생성된다. 이때 해당 분포의 파라미터는(α)이다. 또한 (β)는 해당 토픽이 어떤 단어들로 구성될지를 결정하는 디리클레 분포의 파라미터이다. 이 과정의 반복을 통해 주어진 문서에 대하여 적절한 토픽의 분포와 생성된 토픽에서 적절한 단어의 분포를 추정할 수 있다[14][15].

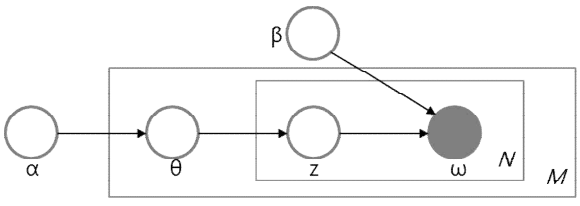

그림 3. 잠재 디리클레 할당 모델
Fig. 3. Latent Dirichlet Allocation Model

2-4 유사도

유사도 측정은 각 문서들 속에 포함되어 있는 단어들의 빈도수 분석을 통해 각 문서가 얼마나 유사한지를 계산하는 방법이다. 이는 정보 검색이나, 텍스트 마이닝 등에서 널리 사용되고 있다[16].

유사도 측정 방법은 크게 상관도 기반 방법과 벡터 기반 방법으로 구분한다. 상관도 기반 기법으로는 Pearson Correlation, Spearman Rank Correlation가 있다. 벡터 기반 기법으로는 Cosine Similarity, Euclidean distance, Inner product 등이 있다. 특히, 코사인 유사도(Cosine Similarity)는 데이터를 좌표평면상의 벡터로 취급하여 데이터 사이의 코사인 각을 구하여 연관성을 수치로 표현하는 방법으로, 단순하고, 계산이 빨라 벡터 기반에서 가장 널리 사용되는 방법이다. 코사인 유사도의 계산법은 아래 수식(1)과 같다[17].

$$COS(A,B) = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)} \sqrt{\sum_{i=1}^{n} (B_i)}} \quad (1)$$

본 연구에서는 토픽 내에서 구성 단어 사이의 유사도를 계산하기 위해 사용된다. 즉, 문서 집합에서 중요한 키워드를 추출하여 순서에 상관없이 하나의 말뭉치(bag-of-words)로 표현하고, 각 문서마다 키워드들의 빈도수(Term Frequency)를 측정하여 키워드에 해당하는 말뭉치의 원소에 저장한다. 이를 바탕으로 두 문서 내용 사이의 유사도를 계산한다.

Ⅲ. LDA를 이용한 질의 데이터 토픽 모델링

3-1 질의 토픽 모델링 개요

토픽 모델링은 문서 집합에서 맥락과 관련된 단어들을 이용하여 의미론적으로 단어들을 군집화 하여 주제를 추론하는 방법이다. 이는 같은 주제 의식을 가지고 있다면 유사한 단어가 출현할 것이라는 가정에 기반 한다. 교수학습지원시스템에는 많은 질의가 반복적으로 축적되고 있다. 대분의 교수학습지원시스템에서의 질의응답 서비스는 담당자가 눈으로 질의을 확인하고 주제를 분류하여 응답하는 방식으로 진행되었다. 이러한 절차를 자동화하기 위해서는 먼저 질의를 자동분류 할 수 있어야 한다. 이를 위해 본 연구에서는 Blei et al.(2003)의 디리클레(Dirichlet)분포 기반의 확률적 토픽 모델링인 잠재디리클레 할당(Latent Dirichlet Allocation) 알고리즘을 기반으로 한다[10][18]. LDA는 단순하고, 데이터의 차원을 축소하는 데 유용하며, 의미적으로 일관성이 있는 주제들을 추출하는 장점을 지니고 있다. 이 때문에 텍스트 분석에서 많이 사용되고 있다[10]. 본 연구에서는 교수학습 질의 데이터에서 토픽 분류를 위한 모

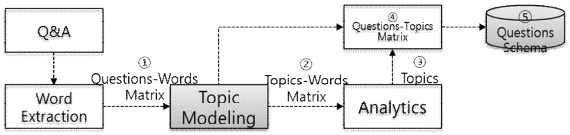

그림 4. 토픽 모델링 기반 사전 구축 절차
Fig. 4. Topic modeling based dictionary build procedure

델링에 활용하고자 한다. 또한, 토픽별 구성 단어 유사도 측정을 위해 코사인 유사도를 사용하고자 한다.

3-2 질의 토픽 모델링

교수학습지원시스템의 질의응답에서 반복적으로 생성되는 질의 데이터를 기반으로 한다. 각 질의는 단어들의 집합으로 이루어져 있다. 이러한 단어들의 집합은 특정 주제를 내포하고 있다. 그리고 각 주제는 질문마다 공유될 수 있다. 즉, 질의 항목을 토픽에 따라 유사질문으로 분류할 수 있다.

시스템에 축적된 질의 데이터를 기반으로 질의 토픽 사전을 생성하는 절차는 그림4와 같다.

① 기존 질의를 바탕으로 단어를 추출하여 질의-단어 매트릭을 구성한다.
② 초기 토픽 N개를 설정 후 반복 출력하여 구성 단어 M개를 선정한다. 이를 바탕으로 토픽-단어 매트릭을 구성한다. 이때 단어의 동질성과 토픽별 이질성을 고려한다.
③ 각 토픽별 구성 단어의 유사도를 바탕으로 최종 토픽을 선정한다. 이때 유사도 계산은 기계학습을 이용한다.
④ 토픽과 질의 매트릭을 구성한다.
⑤ 이를 바탕으로 질의 사전에 대한 스키마를 생성할 수 있다.

이를 바탕으로 새로 유입되는 질의에 대해 해당 질의의 토픽을 추정 분류할 수 있다.

세부적으로, 먼저, 텍스트 전처리 과정을 수행한다. 데이터가 비정형의 텍스트 형태이므로 텍스트에 대한 전처리 과정이 필요하다. 의미 있는 단위의 주제어를 추출하여 키워드 사전을 구축한다. 이 과정을 반복하면서 사전을 지속적으로 추가, 삭제한다. 추출된 주제어를 기반으로 문서 벡터를 생성한다. 그림5는 문서 벡터 생성 과정이다.

다음으로 토픽 모델링을 한다. 문서 벡터를 이용하여, LDA 기반의 토픽 모델링을 활용하여 교수학습활동 질의응답 데이터로부터 주제어를 추출한다. LDA 결과로 생성된 행렬을 이용

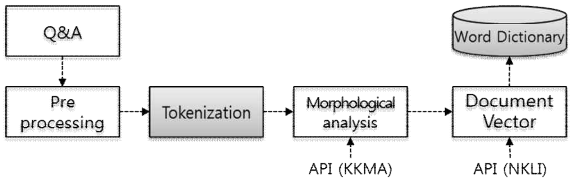

그림 5. 문서 벡터 생성
Fig. 5. Generate document vector

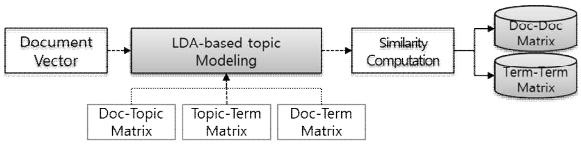

그림 6. 토픽 모델링
Fig. 6. Topic modeling

하여 문서 사이, 주제어 사이 유사도 행렬을 생성한다. 그림6은 유사도 생성 과정이다.

다음으로, 문서(d)에 단어(w)가 나타날 조건부 확률 P(w|d)을 독립 다항 분포로 가정하고 은닉 변수(hidden variable)인 토픽(z)를 추가하여 단어-토픽 간 조건부 확률 P(w/z)과 토픽-문서 간 조건부 확률 P(z/d)로 분해하여 수식(2)를 이용하여P(w/d)를 추정한다.

$$p(w|d) = \sum p(w|z)p(z|d) \quad (2)$$

마지막으로, LDA 수행 결과로 발생한 문서-토픽, 토픽-단어 행렬을 기반으로 문서-문서, 단어-단어의 유사도 행렬을 토픽 기반의 코사인 유사도를 이용하여 계산한다. 먼저, 문서-문서 사이의 유사도는 문서-토픽 행렬을 이용하여 토픽 기반의 문서 벡터를 생성한 후 수식(3)의 코사인 유사도를 이용한다. 다음으로, 주제어를 나타내는 단어-단어 사이의 유사도 행렬도 토픽-단어 행렬을 이용하여 토픽 기반의 단어 벡터를 생성한 후 수식(4)의 코사인 유사도를 이용 한다.

$$Doc-DocSim = COS(\theta) = \frac{D_i D_j}{\| D_i \| \| D_j \|} \quad (i \neq j) \quad (3)$$

$$Term-TermSim = COS(\theta) = \frac{T_i T_j}{\| T_i \| \| T_j \|} \quad (i \neq j) \quad (4)$$

만약 새로운 문서가 입력되면, 새로운 문서의 단어 분포를 분석하여 가장 비슷한 군집을 찾아낸다. 이를 통해, 교수학습지원시스템에서 질의응답 데이터를 지식 데이터로 재사용 가능하다. 교수학습 활동에서 발생하는 질의는 활동에 따라 유사하다. 비슷한 단어 분포를 보이는 질의끼리 군집화 한다면, 질의자가 먼저 유사 그룹의 질을 쉽게 찾아 볼 수 있을 것이다. 이를 통해 기존에 축적된 데이터의 재사용성을 높일 수 있다[5].

Ⅳ. 데이터 실험

본 연구에서는 H대학교 온라인 교수학습지원시스템의 질의응답 코너를 통해 진행되는 질의에 대한 주제 분류를 위해 토픽 모델링을 적용하고 검증한다.

4-1 사전 전처리

온라인 교수학습지원시스템의 최근 3년간의 질의 8,577개를 바탕으로, 한국어 형태소 분석기를 이용해 접두사, 접미사 등 불용어를 처리하고 명사, 동사, 형용사를 기본형 단어로 기본 사전을 구성한다. 문법, 맞춤법, 띄어쓰기, 비표준어 등의 오류가 포함된 자질은 다양하고 불규칙적이어서 실험자가 부분적으로 제거하는 작업을 반복했다.

4-2 LDA 기반 토픽 모델링

LDA에서는 하나의 질의를 그 질의 토픽 분포로 표현한다. 본 논문에서는 LDA를 학습시키기 위해 mallet Toolkit에서 제공하는 오픈소스를 사용하였다. 초기 토픽 모델의 파라미터 (α)= 0.02, 디리클레 분포의 파라미터(β)= 0.02, K=10으로 하였다.

토픽 K값을 결정하기 위해 실험을 반복하였다. K 값이 작은 경우는 한 토픽 안에 다양한 단어로 구성되어 있어 토픽의 주제에 대한 추측이 어려웠고, K값이 큰 경우는 토픽별로 겹치는 단어가 증가하였다. 반복 실험을 통해, K=10인 경우는 토픽이 무엇인지 추측하기 용이하였다. 그림7은 K=10인 경우 LDA 토픽 모델링 결과를 시각화 한 것이다. 그림7의 각 원은 하나의 토픽을 표현하며, 원의 넓이는 단어집합(corpus) 내에서 n개의 전체 토픽들에 대한 비율을 의미하며, 유사도가 높은 토픽끼리 근접해 있다. 막대 목록은 토픽을 선택하지 않았을 때에는 단어집합(corpus) 내에서 우선순위 단어를 나타낸다.

4-3 질의 데이터 사전 구축

LDA 토픽 모델링 결과를 바탕으로, 질의(Q)-단어(W)-토픽(T) 사이의 관계를 개선한다. 먼저, 질의(Q)의 단어 중 토픽(T)에 해당하는 비율을 TQ = P(topic T | question Q)로 계산한다. 다음으로, 단어(W)를 포함한 모든 질의 중 토픽(T)이 할당된 비율을 WT = P(word W | topic T)로 계산한다. 이를 바탕으로, 새로운 토픽(T)을 Tnew = TQ * WT 로 찾는다. 측정하고 있는 단어(W) 외에 다른 단어들이 알맞게 할당되면, 확률을 계산하여 현재 단어(W)를 업데이트 한다. 이러한 과정을 충분히 반복하여 초기 질의 데이터 사전을 안정화 시킨다. 그럼에도 불구하고, 전혀 예상하지 못한 자질도 출현하였다. 이는 질의가 비정형 데이터로 문법, 맞춤법, 띄어쓰기 등의 오류가 포함되어 있음을 의미한다.

4-4 새로운 질의에 대한 분류

질의 데이터 사전을 바탕으로 새로운 질의에 대해 토픽을 자동분류 하는 절차는 아래 그림8과 같다.

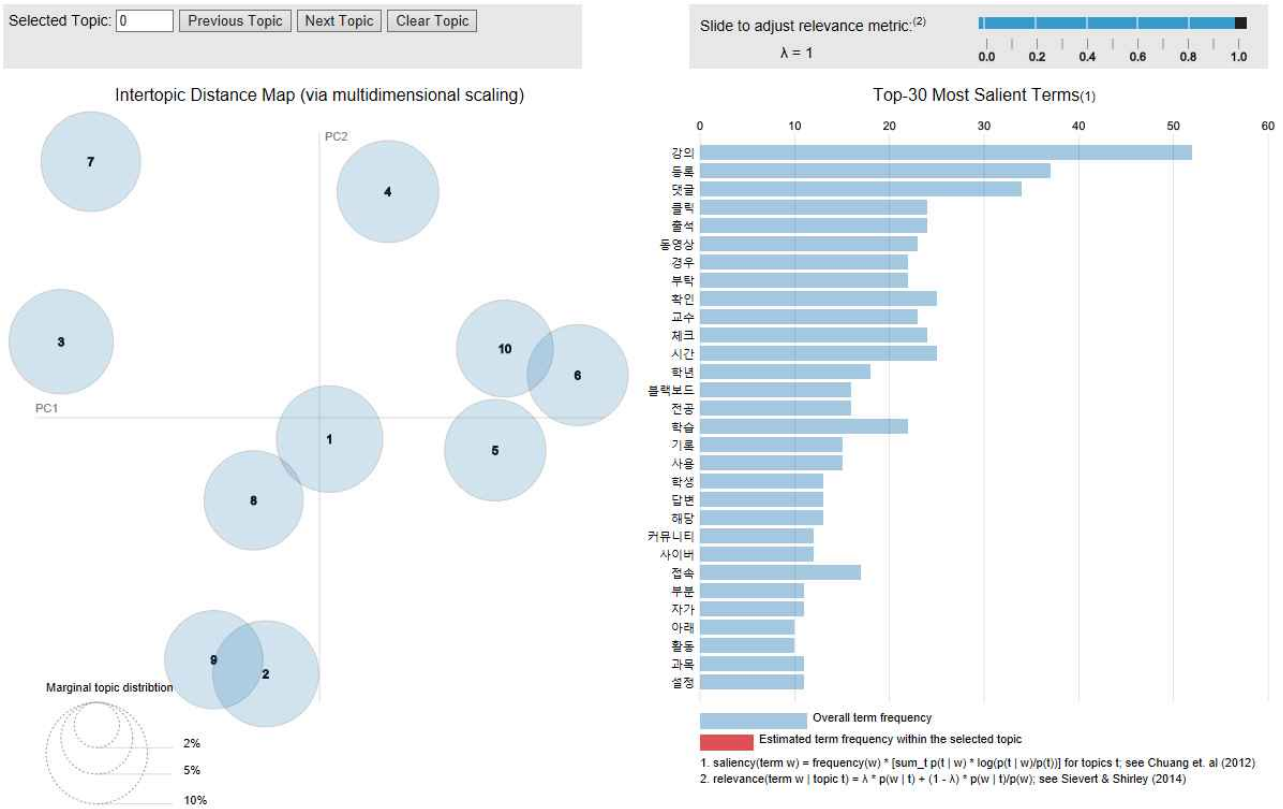

그림 7. 토픽 시각화
Fig. 7. Topic visualization

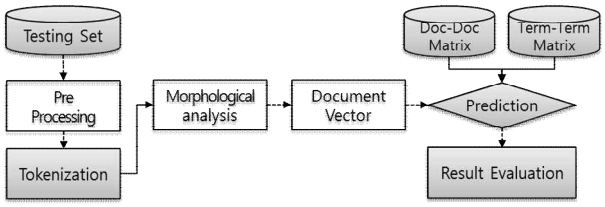

그림 8. 토픽 자동 분류
Fig. 8. Automatically classify topics

토픽 모델링 결과는 테이블 형태로 저장되어 있다. 만약 새로운 질의 글이 입력되면 토픽과 비교하기 위해 새로운 질의 단어 벡터를 리스트 형식으로 생성하여 기존 토픽에 추가한다. 그리고 추가 생성된 리스트를 기준으로 빈도수와 가중치를 구하고, 이를 정규화 한다. 이를 바탕으로, 질의 데이터 사전에 있는 10(K=10)개의 토픽과 코사인 유사도 기법을 이용하여 유사도를 계산하여 토픽을 자동분류 한다.

예를 들어, 표1에서처럼, 새로운 질의 "강의 제목을 바꾸기도 하고 별짓을 다했는데 계속 첨부파일과 같은 창만 뜨고 강의 재생이 안 되네요."가 입력되면, 토픽 모델링에서 생성한 질의 데이터 사전을 바탕으로 유사도를 찾아낸다. 이를 바탕으로 새로운 질의가 토픽7로 자동분류 된다. 토픽7은 학습, 학생, 부분, 활용, 재생, 좌측, 마이, 탐색, 담당, 인터넷 단어를 포함하고 있다.

4-5 F-척도 기반 정확도 검증

자동 분류된 토픽에 대한 정확도와 재현율을 측정하기 위해 데이터마이닝 분야에서 일반적으로 널리 이용되는 F-척도 측정 방법은 수식(5)와 같다[19]. 이를 이용한다.

$$F1-Measure = 2 \times \frac{\mathrm{Precision} \cdot \ Recall}{\mathrm{Precision} + Recall} \quad (5)$$

본 실험에서 실험용 데이터 5개를 추출하여 반복 실험한 결과 표2와 같은 F1값을 얻었다.

본 실험에서는 다양한 질의 데이터의 특성을 고려하여, 전문가에 의한 정답을 추가로 구성하지 않았다. 토픽 모델링 결과만을 가지고, 새로운 질의에 대해 토픽을 자동분류하고 F1값은 측정한 것이다. 일반적인 환경에서 키워드 추출의 정확률이 0.2에서 0.4 정도 나오는 것을 고려하면, 2번을 제외하고는 0.4 이상의 정확률로 추출해 내는 것은 낮지 않음을 의미한다.

표 1. 새로운 질의에 대한 토픽 분포
Table. 1. Topic distribution for new queries

구분	토픽1	토픽2	토픽3	토픽4	토픽5
Q	0	0	0	0	0
구분	**토픽6**	**토픽7**	**토픽8**	**토픽9**	**토픽10**
Q	0	0.738141	0	0	0

Ⅴ. 결 론

온라인 교수학습 활동에서는 다양한 질의 및 응답 데이터가 생성되고 이러한 데이터는 시스템에 자동으로 축적된다. 이렇게 축적된 데이터를 재목적화하여 활용하기 위해서는 자동 분류 모델이 필요하다. 우리는 본 연구에서 새로운 질의를 자동분류하기 위해 Topic-driven 모델을 제안하였다. 이는 초기 분류를 위한 기준 데이터가 없더라도 유용하게 활용될 수 있는 모델이라는 특징을 지닌다. 또한, 시간이 지나 질의가 증가할수록 토픽이 점점 더 정교화 된다.

본 연구에서 질의 및 응답 데이터를 자동분류하기 위해 토픽을 찾는 과정을 모델링하였다. 즉, 시스템에 축적된 질의 데이터를 기반으로 질의 단어를 분석하여 초기 질의 사전을 생성하고, 새로운 질의에 대해 토픽을 자동 추정 분류하도록 설계했다. 그리고 토픽 모델링 결과로 생성된 토픽 중심의 리스트를 이용하여 단어 중심이 아닌 토픽 중심으로 유사도를 도출하여 의미 상 연관관계 추출의 정확도를 향상시켰다.

새로운 질의에 대해 정확도 0.4 이상을 보였으며, 이는 전문가에 의한 정답을 추가로 구성하지 않는 상태에서 나온 결과로 질의에 대한 자동분류 결과가 양호하다고 할 수 있다. 일부 질

표 2. 새로운 질의에 대한 F1-값
Table. 2. F1-value for new query

No.	새로운 질의	토픽	정확도	재현율	F1-measure
1	교수 면담 동영상 강의 관련해서 질문 드립니다	3,4,8,9	0.4	0.010	0.0195122
2	동영상 학습 시 로딩만 계속되고 재생이 안됩니다 자주 이래서 블랙보드 사용하는게 너무 불편하네요	4,7,8	0.3	0.008	0.0146346
3	출석 확인 버튼 클릭이 안 되는데 확인하고 답변 바랍니다	1,5,6,9	0.4	0.010	0.0195122
4	학습 완료 창이 없는 경우 동영상 강의는 출석이 어떻게 되나요	1,3,4,5,7,8,9	0.7	0.018	0.0341463
5	블랙보드 연결이 자주 안 되서 강의 듣기에 너무 불편해요 이 경우 어떤 문제인건가요 실행이 잘 되도록 부탁 드립니다	1,3,8,9,10	0.5	0.013	0.0243902

의에서는 0.7 이상의 높은 자동 분류를 보였으며, 새로운 질의가 여러 토픽에 포함될수록 좀 더 좋은 자동분류 결과를 보였다.

본 연구에서 LDA를 훈련시키기 위해 토픽 개수를 K=10으로 설정했다. 이는 경험적으로 실험을 반복하면서 K=10인 경우 토픽이 무엇인지 추측하기 용이했기 때문에 적정한 토픽 개수로 정한 것이지 최적의 개수는 아니다. 향후 토픽의 최적 개수를 구하는 HDP-LAD[20]를 추가 연구하여 새로운 질의에 대한 자동 분류 성능을 개선하고자 한다. 또한, 질의 자동 분류를 바탕으로 사용자가 응답을 바로 받아 볼 수 있도록 맞춤서비스와 연계된 연구도 지속하고자 한다.

감사의 글

본 연구는 2016년 대한민국 교육부와 한국연구재단의 개인연구지원사업(신진연구)의 지원을 받아 수행된 연구(한국연구재단-2016년-2016015499)로서, 관계부처에 감사드립니다.

참고문헌

[1] Hokyung Lee, Seon Yang, Youngjoong Ko. "Feature Expansion based on LDA Word Distribution for Performance Improvement of Informal Document Classification", Journal of Korea Institute of Information Scientists and Engineers, 2016

[2] Wang, Gang, et al. "Wisdom in the social crowd: an analysis of quora", Proceedings of the 22nd international conference on World Wide Web, ACM, 2013.

[3] Cerulo, Luigi, and Damiano Distante, "Topic-driven semi-automatic reorganization of online discussion forums: a case study in an e-learning context.", Global Engineering Education Conference (EDUCON), IEEE, 2013.

[4] Ezen-Can, Aysu, et al. "Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach", Proceedings of the fifth international conference on learning analytics and knowledge, ACM, 2015.

[5] Lee, Won-Jo, Oh, KyoJoong, and Choi, Ho-Jin. "Comparison Method of Topic Flows for Reusing Experience." 15th Korea Conference on Software Engineering. KIISE, 2013.

[6] Chang, J., Boyd-Graber, J. L., and Blei, D. M. “Connections Between the Lines: Augmenting Social Networks with Text.” ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Paris, France, 2009.

[7] Rosen-Zvi, M., Griffiths, T. L., Steyvers, M., and Smyth, P., “The Author-Topic Model for Authors and Documents.” Uncertainty in Artificial Intelligence (UAI), Baff, Canada, 2004.

[8] Hu, Yuening, et al. "Interactive topic modeling." Machine learning 95.3, 2014.

[9] Lee Sang Yeon, and Keon Myung Lee. "A Reply Graph-based Social Mining Method with Topic Modeling." Journal of Korean Institute of Intelligent Systems 24.6, 2014.

[10] Blei, David M., Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation", Journal of machine learning research 3.Jan, 2003.

[11] Park, Jong Do. "A Study on Mapping Users' Topic Interest for Question Routing for Community-based Q&A Service", Journal of the Korean Society for information Management, 2015.

[12] Anoop, V. S., S. Asharaf, and P. Deepak, "Learning Concept Hierarchies through Probabilistic Topic Modeling", arXiv preprint arXiv:1611.09573, 2016.

[13] H Misra, O Cappe, and F Yvon, “Using LDA to detect semantically incoherent documents”. In Proc. of CoNLL, pages 41–48, Manchester, England, 2008.

[14] Jeong Byeongki, Kim Jungwook, Yoon Janghyeok. "A Semantic Patent Analysis Approach to Identifying Trends of Convergence Technology : Application of Topic Modeling and Cross-impact Analysis." The Journal of Intellectual Property, 2016.

[15] Taemin Cho, Jee-Hyong Lee, “Latent Keyphrase Extraction Using LDA Model”. Journal of Korean Institute of Intelligent Systems, 25(2), 2014.

[16] Lin, Yung-Shen, Jung-Yi Jiang, and Shie-Jue Lee, "A similarity measure for text classification and clustering", IEEE transactions on knowledge and data engineering , 2014.

[17] Sidorov, Grigori, et al. "Soft similarity and soft cosine measure: Similarity of features in vector space model." Computación y Sistemas, 2014.

[18] Hokyung Lee, Seon Yang, Youngjoong Ko. “Feature Expansion based on LDA Word Distribution for Performance Improvement of Informal Document Classification”. Journal of KIISE, 2016.

[19] Young-Sung Cho, Song-Chul Moon, Yeon S. Ahn. A Study of Recommending Service Using Mining Sequential Pattern based on Weight. Journal of Digital Contents Society, 15(6), p.711-719. 2014.

[20] TEH, Yee Whye, et al. Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. In: NIPS. p. 1385-1392. 2004.

김경록(Kyungrog Kim)

1999년 : 아주대학교 공학사
2006년 : 서울벤처정보대학원대학교 (공학석사 - 디지털미디어)
2012년 : 호서대학교 벤처전문대학원 (공학박사 - IT응용기술)

1999년～2000년: 동원시스템스(주)
2002년～2008년: 사단법인 차세대학습산업기반센터
2009년～2009년: ㈜조선에듀케이션
2004년～현 재: 호서대학교 전자디스플레이공학부 조교수
※관심분야 : 데이터 마이닝, 학습분석, 이러닝, Information Learing, 등

송혜진(Hyejin Song)

2016년 : 호서대학교 컴퓨터소프트웨어전공 (공학사)
2016년~현재 : 호서대학교 대학원 컴퓨터공학과 석사과정

※관심분야 : 데이터마이닝, 추천시스템, 패턴인식 등

문남미(Nammee Moon)

1985년: 이화여자대학교 컴퓨터학과 공학사
1987년: 이화여자대학교 공학석사
1998년: 이화여자대학교 공학박사

1999년~2003년: 이화여자대학교 조교수
2003년～2008년: 서울벤처정보대학원대학교 디지털미디어학과 교수
2008년～현 재: 호서대학교 컴퓨터소프트웨어전공 교수
※관심분야 : Social Learning, 필터링, HCI, 메타데이터, User Centric