J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

문서 중요도를 고려한 토픽 기반의 논문 교정자 매칭 방법론☆

A Proofreader Matching Method Based on Topic Modeling Using the Importance of Documents

손 연 빈[1] 안 현 태[1] 최 예 림[1*]
Yeonbin Son Hyeontae An Yerim Choi

요 약

최근 국내외 연구기관에서는 논문을 저널에 제출하는 과정에서 연구결과를 효과적으로 전달하기 위해 외부 기관을 통해 논문의 문맥, 전문 용어의 쓰임, 스타일 등에 대한 논문 교정을 진행하는 경우가 증가하고 있다. 하지만 대다수의 논문 교정 회사에서는 매니저의 주관적 판단에 따라 수동으로 논문 교정자를 할당하는 시스템이며, 이에 따라 논문의 주제에 대한 전문성이 부족한 교정자를 할당하여 논문 교정 의뢰인의 만족도가 떨어지는 사례가 발생하고 있다. 따라서 본 논문에서는 효과적인 논문 교정자 할당을 위해 논문의 토픽을 고려한 논문 교정자 매칭 방법론을 제안한다. Latent Dirichlet Allocation을 이용하여 문서의 토픽 모델링을 진행하고, 그 결과를 이용하여 코사인 유사도 기반으로 사용자간 유사도를 계산하였다. 특히, 논문 교정자의 토픽 모델링 과정에서, 대표 문서로 간주되는 문서의 중요도에 따라 가중치를 부여하여 빈도수에 차별을 둬 정밀한 토픽 추정을 가능하게 한다. 실제 서비스의 데이터를 이용한 실험에서 제안 방법론의 성능이 비교 방법론보다 우수함을 확인하였으며, 정성적 평가를 통해 논문 교정자 매칭 결과의 유효성을 검증하였다.

☞ 주제어 : 논문 교정, 논문 교정자 추천, 토픽 모델링, 문서 중요도

ABSTRACT

In the process of submitting a manuscript to a journal in order to present the results of the research at the research institution, researchers often proofread the manuscript because it can manuscripts to communicate the results more effectively. Currently, most of the manuscript proofreading companies use the manual proofreader assignment method according to the subjective judgment of the matching manager. Therefore, in this paper, we propose a topic-based proofreader matching method for effective proofreading results. The proposed method is categorized into two steps. First, a topic modeling is performed by using Latent Dirichlet Allocation. In this process, the frequency of each document constituting the representative document of a user is determined according to the importance of the document. Second, the user similarity is calculated based on the cosine similarity method. In addition, we confirmed through experiments by using real-world dataset. The performance of the proposed method is superior to the comparative method, and the validity of the matching results was verified using qualitative evaluation.

☞ keyword : Manuscript proofreading, proofreader matching, topic modeling, importance of document

1. 서 론

국내외 연구기관에서는 연구결과 발표를 위해 국제 저널에 논문을 투고하는 과정에서 연구결과를 보다 효과적으로 전달하기 위해 논문 교정 과정을 거치는 경우가 빈번하다. 특히 국제 저널에 논문을 투고하는 경우, 만국공통어인 영어를 이용하는데 모든 연구자의 모국어가 영어인 것은 아니므로 논문 교정 과정이 필요한 것이다. 논문 교정 과정에서는 단순히 철자, 문법적 오류 교정만이 이루어지는 것이 아니라 의미적 전달이 모호한 용어, 문장 등의 스타일 교정 등이 이루어진다. 따라서 논문 교정자는 교정을 위해 논문에 대해 정확한 이해를 하고 있어야 한다.

일반적으로 논문 교정 과정은 논문 교정 회사[1]를 매개체로 의뢰자와 논문 교정자를 매칭하는 과정이 핵심적으로 이루어진다. 의뢰자가 회사로 논문 교정을 의뢰하면 매니저가 논문을 읽고 내용을 파악하여 회사에 속한 적

1 Industrial and Management Engineering, Kyonggi University, Suwon, Republic of Korea
* Corresponding author: yrchoi@kgu.ac.kr
[Received 4 April 2018, Reviewed 25 April 2018(R2 14 June 2018), Accepted 6 July 2018]
☆ 본 연구는 경기도의 경기도 지역협력연구센터 사업의 일환으로 수행하였음.[GRRC 경기 2017-B01, 지능형 제조 빅데이터 분석 연구]
☆ 본 논문은 2017년도 한국인터넷정보학회 추계학술발표대회 우수 논문 추천에 따라 확장 및 수정된 논문임.

절한 논문 교정자를 할당한다. 할당하는 과정에서 매니저가 기준으로 삼는 것은 논문 교정자의 전공, 현재 논문 교정이 가능한 상태인지, 가격이 논문 교정자가 제안한 금액과 적절한지 등이다. 우리는 이와 같은 논문 교정자 매칭 과정에서 문제점을 파악하였다.

매니저는 모든 분야에 전문 지식을 가지고 있지 않기 때문에, 논문을 읽고 논문 교정자를 할당하는 과정은 전적으로 매니저의 판단 하에 이루어진다. 하지만 대부분 논문에 대한 정확한 이해를 기반으로 한 논문 교정자 할당이 아닌 단순 카테고리화에 의한 할당이 이루어진다. 이러한 매칭 시스템은 주관적이며, 할당의 기준에 일관성이 존재하지 않을 가능성이 높다. 따라서 고객만족도 또한 일관적이지 못하며, 회사는 서비스 만족도 향상을 위해 일관성이 있는 매칭 시스템을 이용하는 것이 바람직하다.

일관성 있는 매칭 시스템을 제안하기 위해 일반적인 토픽 기반 추천 시스템[2]을 조사한 결과, 텍스트 분석이 널리 사용되고 있으며 이에 대한 성능은 연구를 통해 검증된 바 있다[3]. 이러한 시스템은 의미론적 분석 방법으로 문서의 주제를 파악하고, 이를 통해 추천하는 방법이 일반적이다[4]. 더 나아가, LDA(Latent Dirichlet Allocation) 방법론이 높은 성능을 보이며, 이는 사용자 대표 문서를 이용한 토픽 모델링을 통해 사용자의 토픽을 추출하는 방법론이다[5]. 따라서, 본 논문에서는 LDA 기반의 토픽 모델링을 통한 토픽 추출 후, 코사인 유사도를 이용해 사용자간 연관도를 계산하여 적절한 논문 교정자를 매칭하는 방법론을 제안한다[6].

토픽 추출 기반 텍스트 분석을 수행하기 위해 사용자별 대표 문서를 지정한 후 이를 이용해 분석 과정을 거치는데, 대표 문서가 2개 이상인 경우 일반적으로 모든 문서의 빈도수를 동일하게 간주하여 분석이 진행된다. 따라서 특정 문서가 사용자를 설명하는데 비교적 영향력이 높은 문서이더라도 이를 고려하지 않고 텍스트 분석이 진행된다.

논문 교정 분야에서는 의뢰자가 서비스를 제공받은 이후 평점을 주는 과정이 포함되며, 평점에 따라 의뢰자의 서비스에 대한 만족도 판단이 가능하다. 서비스 사용자의 리뷰를 시스템 개선에 반영하면 만족도가 확연히 높아진다는 연구 결과에 따라[7,8], 평점을 추천 시스템에 반영한다면 높은 기대효과를 얻을 수 있을 것으로 예상하였다. 평점이 낮다면 논문 교정자는 그 논문에 대해 전문성을 가지고 있지 않다고 판단할 수 있으며 이를 이용하여 각 문서의 중요도를 알 수 있다. 이를 추천에 이용한다면 높은 성능의 추천을 할 수 있다[9]. 이에 따라 논문 교정자를 대표하는 문서를 설정할 때, 평점을 높게 받은 논문은 가중치를 높게 책정, 낮게 받은 논문은 낮게 책정하여 가중치가 높은 경우 상대적으로 빈도수를 높게 지정하여 각 사용자의 대표 문서를 재가공하였다.

본 논문에서는 LDA를 이용한 토픽 기반 전문가 추천 시스템을 제안한다. 특히, 토픽 추출 과정에서 사용자를 대표하는 각 문서의 중요도에 따라 가중치를 다르게 할당하여 빈도수에 차별을 두는 방법을 적용한다. 일반적으로 문서의 중요도를 똑같이 간주하고 토픽 모델링을 진행하는 LDA 방법론으로부터 발전된 제안 방법론은 크게 두 가지 과정으로 이루어져 있다. 첫째, 사용자를 대표하는 문서의 토픽 모델링을 진행하는 부분, 이 과정에서 특정 사용자의 여러 문서가 모여 하나의 대표 문서로 형성되는 경우 각 문서의 중요도를 고려하여 가중치를 부여하여 빈도수에 차이를 둔다. 둘째, 토픽 모델링 결과에 따라 코사인 연관도 계산법을 이용하여 사용자 간의 연관도를 계산하는 부분으로 이루어져 있다.

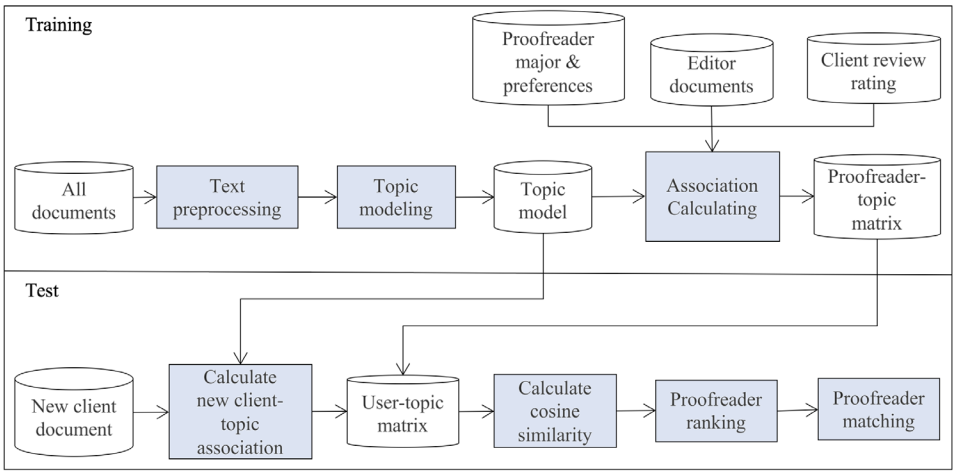
2장에서는 토픽 기반의 논문 교정자 매칭 방법론의 전체 구조와 더불어 각 단계 별 상세설명을 하며, 3장에서는 이를 기반으로 실험을 수행하여 비교 방법론의 실험결과를 비교하여 제안 방법론의 성능을 평가한다. 마지막으로, 4장에서는 본 연구에 대한 결론에 대해 논의한다.

2. 제안 방법론

본 논문에서는 논문 교정 분야에 적용 가능한 논문 교정자 추천 시스템을 제안하며, LDA를 이용한 토픽 모델링 방법에 논문 교정자의 전문성을 반영하는 평점 데이터를 이용해 문서의 중요도를 고려하여 가중치를 부여하는 방법을 이용한다. 그림 1로 도식화한 제안 방법론은 (a)데이터 전처리, (b)LDA 기반 토픽 모델링, (c)코사인 유사도 기반 사용자간 연관도 계산 및 논문 교정자 추천 과정으로 구성되어 있다. 첫번째로 기존 데이터를 추천 시스템에 적절하게 전처리를 진행하며, 두번째는 이를 이용하여 LDA 기반의 토픽 모델링을 진행한다. 마지막으로 토픽 모델링 결과를 통해 의뢰자와 논문 교정자 간의 연관도를 계산하여 적절한 논문 교정자 추천을 하는 과정으로 이루어진다.

2.1 데이터 전처리

토픽 모델링을 진행하기 위하여 사용할 문서의 전처리를 진행한다. 토픽에 반영된다고 볼 수 없는 단어 및 의미



(그림 1) 제안 방법론 도식
(Figure 1) Framework of the proposed method

가 없는데 단어 빈도수가 높은 단어 등을 문서에서 제거하는 과정을 거친다. 예를 들어, 'a', 'the', 'she', 'I' 등의 단어는 문장 구성을 위해 필수적으로 사용되지만 문서의 토픽을 반영한다고 볼 수 없다. 연구 논문의 특성상 자체적으로 만든 약어가 다수 존재하나, 이는 특정 전문 분야의 특징을 반영한 일반적인 약어가 아니므로 이 또한 제거하며 특수 문자와 영어 이외의 언어도 제거한다. 그림, 표의 경우 토픽 모델링 과정에서 고려될 수 없으므로 제거한다.

다음으로 평점 데이터의 전처리를 진행한다. 의뢰자가 동일한 논문 교정자에게 2회 이상 서비스를 제공받은 경우, 평점을 여러 번 부여하는 경우가 존재할 가능성이 있다. 이 평점을 날짜순으로 정렬하여 가장 최근의 평점을 이용한다.

2.2 토픽 모델링

본 단계에서 사용자란, 의뢰자와 논문 교정자를 통칭한다. 텍스트 분석을 위해 모든 사용자의 문서를 하나의 문서로 만든 뒤 벡터화 하며, 이를 이용해 LDA 기반 토픽 모델링을 진행한다. LDA 기반 토픽 모델링은 단어 분포를 기반으로 특정 문서의 토픽을 추출하는 방법론으로, 텍스트를 이용한 데이터 분석 분야에 널리 사용된다. 이 과정에서 두개의 변수가 이용된다. 첫째, 텍스트 데이터를 분석하기에 앞서 벡터화 과정에서 단어를 빈도수 순으로 정렬하여 하나의 단어 집합에 몇 개의 단어가 속할 것인가에 대해 결정하는 변수가 필요하다[10]. 둘째, 벡터화된 텍스트 데이터를 이용하여 토픽 모델링 하는 과정에서는 토픽 별 단어 집단을 몇개 생성할 것인지에 대한 변수가 필요하다. 변수에 따라 특정 개수의 토픽 별 단어 집단이 형성된다.

토픽 별 단어 집단을 기반으로 각 사용자의 문서에 대한 토픽을 추출하는 과정이 진행된다. 먼저 사용자를 대표하는 문서를 지정한 뒤 토픽 추출 과정이 이루어지는데, 사용자를 대표하는 문서란 의뢰자의 경우 의뢰를 요청한 논문이며 과거 의뢰한 경험이 2회 이상인 경우는 여러 논문을 하나의 문서로 합친다. 또한 논문 교정자를 대표하는 문서는 과거 교정한 논문들과 전문 분야의 명칭을 하나의 문서로 합친 것이다. 논문 교정자의 대표 문서 설정 시, 과거 교정한 논문들에 대한 평점이 존재하는데 1,2,3점의 평점을 받았을 경우 이 점수는 의뢰자에게 만족을 주었다고 간주하기 어려우므로 데이터로 이용하지 않았으며, 4,5점의 평점을 받은 경우의 문서만 이용하였고, 4점보다 5점을 받은 문서의 가중치를 높게 부여하였다. 이 과정을 통해 사용자와 토픽 간 연관도를 얻을 수 있다.

2.3 사용자간 연관도 계산 및 논문 교정자 추천

토픽 모델링 결과로 얻은 사용자 별 토픽과의 연관도를 이용하여 새로운 논문이 들어온 경우의 의뢰자와 논문 교정자 간의 연관도를 계산한다. 연관도 계산으로는 코사인 기반 계산과 유클리디안 기반 계산이 일반적으로

이용되는데[6], 본 연구에서의 사용자간 연관도 계산은 코사인 유사도를 이용하였다. 코사인 유사도란 내적 공간의 두 벡터 간 각도의 코사인 값을 이용하여 측정된 벡터 간의 유사한 정도를 의미한다[11].

사용자간 연관도를 이용하여 특정 의뢰자에게 가장 적절한 논문 교정자를 추천한다. 특정 의뢰자와 가장 적절한 논문 교정자를 추천하기 위하여 특정 의뢰자와 연관도가 가장 높은 논문 교정자를 찾아야 한다. 사용자간 연관도를 계산하였기 때문에, 가장 높은 연관도를 찾는 과정에서 의뢰자와의 연관도는 제외하고 고려한다.

3. 실 험

3.1 실험 데이터

제안 방법론의 추천 성능을 평가하기 위해 논문 교정 서비스 회사 '워드바이스[1]'의 데이터를 이용하여 실험을 수행하였다. 표 1은 수집 데이터에 대해 요약한 표이다. 의뢰자가 서비스를 제공받은 후 매길 수 있는 평점은 1~5점이며, 평점을 매기지 않아도 된다. 워드바이스에 속한 문서 중 5,000개의 문서를 고려하였으며, 이 중 평점을 받은 문서 수는 1,861개이다. 평점을 받은 문서 중 1~3점을 받은 문서는 논문 교정자가 교정 서비스를 만족스럽게 수행한 것이 아니므로, 논문 교정자의 전문 분야의 특징을 대표할 수 없다고 판단하여 이용하지 않았다.

데이터 전처리 과정에서는 의미를 가진다고 볼 수 없는 단어의 제거 과정이 이루어졌다. 표 2는 제거된 단어의 예시를 나타낸 것으로, 연구 논문에서 전반적으로 사용되는 단어, 문장 구성을 하기 위해 사용된 필수 단어 등

(표 1) 수집된 데이터 요약
(Table 1) Summary of the collected dataset

The number of documents		5,000
The average number of words		1,644
The greatest number of words in document		18,083
The number of documents along to ratings	5	1,189
	4	415
	3	162
	2	45
	1	50
	Total	1,861

(표 2) 제거된 단어의 예시
(Table 2) Example of the removed words

by	Figure	printed	best	Kim
so	do	amount	Seoul	…

(표 3) 제거한 약어, 특수문자의 예시
(Table 3) Example of the removed abbreviations and special characters

xxx	vey	pp	vb	th
gen	NPNL	PWA	TMAH	…

이다. 표 3는 제거된 약어, 특수 문자의 예시로, 이때의 약어는 통용되는 약어가 아니라 연구 논문 자체적으로 만든 약어를 의미한다. 이러한 약어 자체는 의미가 없기 때문에 제거하였다. 또한 영어 이외의 단어, 그림, 표 등 텍스트가 아닌 데이터를 제거하였다.

3.2 실험 환경

논문 교정자의 전문 분야를 판단하는데 있어 평점이 효과적으로 반영될 수 있으므로 평점을 이용하여 문서에 가중치를 부여하여 논문 교정자의 대표 문서를 설정하였다. 1~3점의 평점을 받은 경우, 의뢰자가 논문 교정자의 서비스에 대해 만족하지 못했다고 판단하여 이 경우의 논문은 대표 문서 구성에 이용하지 않았다. 서비스를 받았지만 의뢰자가 평점을 주지 않은 경우도 존재하는데, 이 경우는 가중치 1을 부여하였으며, 4점의 평점을 받은 경우 가중치를 2로, 5점의 경우 가중치를 3으로 부여하였다. 또한 논문 교정자가 회사에 고용된 때 지정한 '자신 있는 분야', '가장 자신 있는 분야' 의 경우 각각 30, 100으로 가중치를 부여하였다.

실험을 수행하는 과정에 있어서, 두 가지 변수를 고려했다. 첫째, 빈도수 기반 고려해야하는 단어의 수를 나타내는 변수, 둘째로 토픽 모델링 과정에서의 토픽의 수를 나타내는 변수가 있다. 여러 경우의 수에 따라 실험을 진행하였으며, 가장 적절한 변수를 크기를 찾고자 하였다.

벡터화 과정에서 빈도수 기반으로 고려해야 할 단어의 수는 변수로써 사용되었는데, 실험에서 이용되는 문서의 단어 수를 고려하여 변수를 설정하였다. 문서의 단어 수가 가장 큰 경우는 18,083개로, 이를 고려하여 단어의 벡터화 과정에서 고려하는 단어의 수는 10,000개로 지정하였다. 토픽 모델링 과정에서 지정하는 생성되는 토픽의 개수는 '워드바이스'에서 실제로 분류한 연구 분야 리스

트의 개수를 근간으로 변수를 설정하였다. 연구 분야의 리스트가 22개의 항목을 가지고 있으므로 분야가 추가될 수 있다고 가정하여 변수를 25개로 설정하였다.

제안 방법론의 성능을 평가하기위한 비교 방법론으로는 선호도 기반 추천 시스템에 널리 사용되는 MF(matrix factorization)[12,13]를 이용하였다. 의뢰자가 3명 이상의 논문 교정자에게 평점을 매긴 경우만 성능 평가의 데이터로 이용하였다. 이 경우만 순위를 3등 이상으로 매길 수 있기 때문이다. 방법론의 성능을 평가하기 위한 수식은 (1)과 같은데, 테스트 데이터를 지정하여 0으로 변경한 후 제안 및 비교 방법론을 수행하였다. 예측된 평점을 이용하여 다시 순위를 매긴 후 이를 rank(B)로 칭한다. 원래 순위인 rank(A)와 rank(B)의 비교를 통해 성능을 평가하였다.

$$\text{Error rate} = \frac{\sum_{n=1}^{N}\left|\frac{rank(A)-rank(B)}{n(row(i))}\right|}{N} \quad (1)$$

3.3 실험 결과

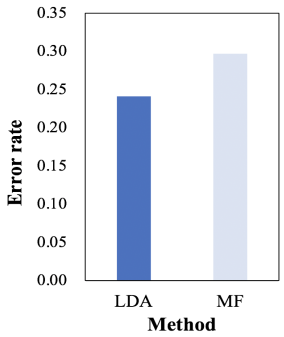
표 4는 토픽의 갯수를 100개로 지정하여 LDA 기반 토픽 모델링 실험 결과 토픽에 대한 단어 집합을 나타낸다. 토픽 모델링 실험 결과, 첫번째 토픽의 단어를 정성적으로 분석하여 보았을 때, 의학 분야에서 주로 사용되는 단어이므로 첫번째 토픽은 의학 논문과 유사도가 높을 것이란 것을 예측할 수 있다. 이로써 정성적으로 토픽 모델링의 결과가 유의성이 있음을 판단하였다.

성능 평가 비교 척도에 따라 제안 및 비교 방법론의 성능을 측정한 결과는 그림 2와 같다. 제안 방법론의 경우 에러율이 0.24이며, 비교 방법론의 경우 0.30이었다. 따라서 추천 시스템에서 일반적으로 많이 사용되는 비교 방법론보다 제안 방법론의 성능이 우수함을 알 수 있었다.

(표 4) 100개 토픽에 대해 토픽 모델링을 했을 때 각 토픽을 대표하는 단어 리스트

(Table 4) List of words representing each topic when execute topic modeling for 100 topics

Topic number	Word
0	hydrogel, ear, construct, tissue, laden, …
1	alloy, electrode, hydrogen, discharge, …
2	wage, coping, tenure, maltreatment, …
…	…
99	senators, fans, ottawa, mensa, downtown, …



(그림 2) 제안 방법론 LDA와 비교 방법론 MF의 에러율 평가 결과

(Figure 2) Results of error rate between the prposed method and thecompared method

추가적으로, 비교 방법론 MF의 경우 사용자의 선호도를 반영하는 별점을 이용해 추천을 진행하는 방법론이다. 이는 컨텐츠 기반의 추천이 아닌 수치적 선호도을 이용해 진행한다. 반면에 제안 방법론의 경우 LDA 방법론을 이용하여 문서의 토픽 추출을 이용해 추천을 진행한다. 따라서 제안 방법론은 컨텐츠를 이용해 추천을 진행한다는 점에서 의의가 있다.

의뢰자와 논문 교정자 간의 연관도 계산은 코사인 유사도를 이용하여 이루어졌다. 의뢰자와 연관도가 가장 높은 논문 교정자 추천이 실제로 유의성이 있는지 의뢰자와 논문 교정자의 전문 분야에 대한 정성적 판단을 진행하였다. 표 5는 임의로 선택한 의뢰자 1에 대해 논문 교정자와의 코사인 유사도 결과이며 이에 따른 순위를 나타내는 리스트이다. 의뢰자1에게는 논문 교정자9488이 추천되는 것이 가장 적절함을 알 수 있다. 정성적 분석 결과, 의뢰자1과 논문 교정자9488은 둘 다 화학 공학 분야가 전문 분야로, 적절한 매칭이 이루어졌음을 확인하였다.

(표 5) 의뢰자1과 각 논문 교정자 간의 토픽에 대한 코사인 유사도 및 그에 따른 순위

(Table 5) Cosine similarities between User1 and proofreaders and their ranks

Proofreader number	Cosine similarity	Rank
9488	0.990	1
9663	0.982	2
6959	0.972	3
2048	0.969	4
7666	0.968	5
…	…	…

4. 결 론

본 연구는 문서도의 중요도를 고려하지 않은 기존의 주관적 논문 교정자 할당 시스템의 한계점에 주목하였다. 특히 전문가 추천이 유용하게 사용되는 논문 교정 분야에서 연구를 진행하였으며, 문서도 별 중요도에 따라 상이한 가중치 부여를 통한 논문 교정자 자동 추천 방법론을 제안하였다. 그 중에서도 제안 방법론의 성능을 평가하기 위해 평점을 수집하는 추천 시스템에서 일반적으로 널리 사용되는 MF를 비교 방법론으로 이용하였으며, 에러율 측정 결과 제안 방법론의 성능이 더 우수함을 알 수 있었다. 정성적 평가 결과 의뢰자와 매칭된 논문 교정자의 전문 분야가 동일하여 방법론이 유의함을 판단하였다. 본 연구에서 제안된 논문 교정자 자동 추천 방법론을 이용하여, 논문 교정 분야에서의 객관적이고 일관적인 논문 교정자 추천을 통한 고객의 서비스 만족도 향상을 기대한다.

참고문헌(Reference)

[1] Wordvice Inc., https://essayreview.co.kr/intro/.

[2] Hyun, Y., Shun, W. W. X., & Kim, N., “Methodology for Issue-related R&D Keywords Packaging Using Text Mining.”, Journal of Internet Computing and Services, Vol.16, No.02, pp.57-66, 2015.
https://doi.org/10.7472/jksii.2015.16.2.57

[3] Yu, Y., Mo, L., & Wang, J., “Identifying Topic-Specific Experts on Microblog.”, Transactions on Internet & Information Systems, Vol.10, No.06, 2016.
https://doi.org/10.3837/tiis.2016.06.010

[4] Hofmann, T., “Probabilistic Latent Semantic Analysis.”, In Proceedings of the conference on Uncertainty in Artificial Intelligence, pp.289-296, 1999.
https://dl.acm.org/citation.cfm?id=-2073829

[5] Blei, D. M., Ng, A. Y., & M. I., “Latent Dirichlet Allocation.”, Journal of Machine Learning Research, pp.993-1022, 2003.
http://www.jmlr.org/papers/v3/blei03a.html

[6] Pang, N. T., Michael, S., and Vipin, K., “Introduction to Data Mining.”, Addison-Wesley, 2007.

[7] Kim, M., Song, E., & Kim, Y., “A Design of Satisfaction Analysis System for Content Using Opinion Mining of Online Review Data.”, Journal of Internet Computing and Services, Vol.17, No.03, pp.107-113, 2016.
https://doi.org/10.7472/jksii.2016.17.3.107

[8] Jung, S., Lee, H., & Suh, Y., “The Influence of Negative Emotions on Customer Contribution to Organizational Innovation in an Online Brand Community.”, Journal of Internet Computing and Services, Vol.14, No.04, pp.91-100, 2013.
https://doi.org/10.7472/jksii.2013.14.4.91

[9] Yoo, S. Y., & Jeong, O. R., “Social Category Based Recommendation Method.”, Journal of Internet Computing and Services, Vol.15, No.05, pp.73-82, 2014.
https://doi.org/10.7472/jksii.2014.15.5.73

[10] Wallach, H. M., “Topic Modeling: Beyond Bag-of-words.”, In Proceedings of the 23rd International Conference on Machine Learning, pp.977-984, 2006.
https://doi.org/10.1145/1143844.1143967

[11] Karnik, A., Goswami, S., & Guha, R., “Detecting Obfuscated Viruses Using Cosine Similarity Analysis.”, Modelling & Simulation, pp.165-170, 2007.
https://doi.org/10.1109/ams.2007.31

[12] Koren, Y., Bell, R., & Volinsky, C., “Matrix Factorization Techniques for Recommender Systems.”, Computer, Vol.42, No.08, 2009.
https://doi.org/10.1109/mc.2009.263

[13] Baltrunas, L., Ludwig, B., & Ricci, F., “Matrix Factorization Techniques for Context Aware Recommendation.”, In Proceedings of the fifth ACM conference on Recommender systems., pp.301-304, 2011.
https://doi.org/10.1145/2043932.2043988

◐ 저 자 소 개 ◑

손 연 빈(Yeonbin Son)
2018년 경기대학교 산업경영공학과(공학사)
2018년~현재 경기대학교 일반대학원 산업경영공학과(공학석사)
2017년~현재 주식회사 디저팅 대표
관심분야: 머신러닝, 추천시스템
E-mail: yeonbin517@gmail.com

안 현 태(Hyeontae An)
2013년~현재 경기대학교 산업경영공학과
관심분야: 텍스트마이닝
E-mail: hyeontaean94@gmail.com

최 예 림(Yerim Choi)
2010년 서울대학교 산업공학과(공학사)
2016년 서울대학교 산업공학과(공학박사)
2016년~2017년 네이버랩스 Data Scientist
2017년~현재 경기대학교 산업경영공학과 조교수
관심분야: 인공지능/머신러닝, 빅데이터 기반의 인간 모델링
E-mail: yrchoi@kgu.ac.kr