# 리뷰분석을 통한 온라인교육자 신뢰도 파악 자동화 시스템 설계

이기훈*, 문남미*
*호서대학교 컴퓨터정보공학부
e-mail:happy51738@gmail.com

# Designing an automated system to grasp the reliability of online educators through review analysis

Ki-Hoon Lee*, Nammee Moon*
*Dept of Computer Science, Hoseo University

요 약

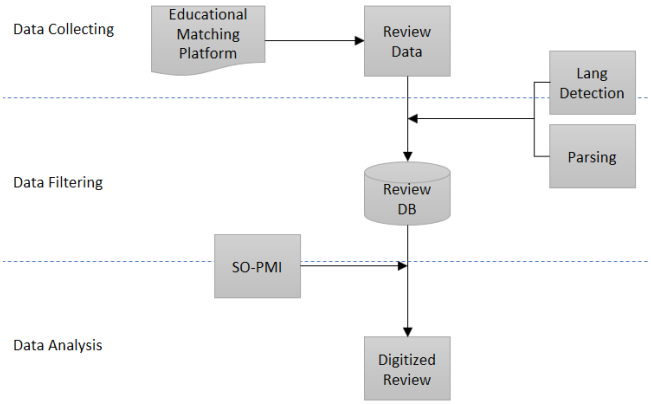
본 논문은 온라인 교육매칭 플랫폼의 교육자에 대한 신뢰도 파악을 위한 리뷰분석 자동화 시스템을 설계한 논문이다. 웹 크롤링을 통해 비정형 데이터인 교육자에 대한 리뷰를 수집 및 파싱을 통해 데이터 베이스화 한다. 수집한 리뷰 데이터와 SO-PMI를 이용해 온라인 교육자 신뢰도 파악을 위한 맞춤형 감성사전을 구축하고자 한다. 구축한 감성사전을 이용해 리뷰를 수치화해 교육자와 피교육자 매칭 시 신뢰성 향상에 도움을 주고자 한다.

## 1. 서론

21세기를 지식정보화사회라 부르며 최근 많은 전략가나 학자들이 지식정보화사회에서의 역할을 강조하고 있고, 인적 자원 발굴 및 관리가 21세기의 중요한 과제임을 주장하고 있다[1]. 인적 자원 개발 분야의 교육을 살펴보면, 과거의 집합 교육이나 기술 위주의 교육은 점차 축소되고 있다. 반면에, 개인별 요구를 일대일 관계에서의 교육이 증가되는 추세이다. 즉 인성과 지식을 모두 고려하는 교육이 주목받고 있다[2].

과거에는 이러한 일대일 관계의 교육을 오프라인광고나 커뮤니티 형태의 플랫폼에서 매칭을 해줬다면, 현재는 수많은 O2O 교육 매칭 플랫폼이 주목받고 있다. 예술, 체육 등 특정 분야의 교육의 경우 온라인에 비해 오프라인 교육이 효과적이라는 연구가 있다. 이러한 특성상 온라인 매칭이 오프라인으로 이루어지는 경우가 빈번하다. 이러한 오프라인 교육이 진행될 경우 튜터, 튜티 모두 서로에 대한 신뢰성은 매우 중요한 매칭의 요소가 될 수 있다[3].

본 논문에서는 온라인 교육매칭에서 튜터에 대한 신뢰성을 비정형 데이터인 튜터에 대한 리뷰데이터를 SO-PMI를 이용한 감성사전을 구축, 감성사전을 이용해 수치화해주는 시스템을 제안할 것이다[4-5]. 이를 통해 튜터에 대한 신뢰성 파악에 도움이 되고, 자동화와 수치화를 통해 원활한 매칭을 돕고자 한다.




(그림 1) 전체 시스템 구조도

## 2. 본론

### 2-1. 전체 시스템 구조

본 논문에서 제안하는 온라인 교육자 신뢰도 파악 시스템은 그림 1과 같다. 시스템은 크게 데이터 수집, 데이터 필터링, 데이터 분석(수치화) 세 단계로 나뉜다. 그림 1의 데이터수집(Data Collecting)은 온라인 교육매칭 플랫폼에서 튜터에 대한 후기데이터를 수집하는 단계를 나타낸다.

두 번째 단계에서는 수집한 Review Data가 문장들의 집합으로 되어있으므로, 이를 문장단위, 더 나아가 형태소 분석기를 이용해 형용사 단위로 파싱해 데이터베이스화 한다.

마지막 단계에서는 형용사단위로 처리된 ReviewDB를 SO-PMI기법을 이용해 감성사전을 구축한다. 구축된 감성사전을 이용해 리뷰데이터를 수치화 하는 시스템이다.

## 2-2. SO-PMI

SO-PMI는 PMI의 개별 단어에 대한 편향을 줄인 기법이다. PMI는 Point-wise Mutual Information의 약자로 두 점 x, y 사이의 상관도를 확률론을 적용해 수치화하는 기법이다. PMI는 Mutual Information을 기반으로 한다. 각각의 발생 확률 P(x), P(y)를 갖는 두 점 x, y에 대해 두 점 사이 $PMI(x,y)$를 다음 식 1과 같이 정의할 수 있다.

$$PMI(x,y) = \log \frac{P(x,y)}{P(x)P(y)} \quad (1)$$

x와 y가 독립적으로 발생했다 가정했을 때 동시에 발생할 확률과, 측정된 x, y의 동시발생 확률을 비교함으로써 두 변수가 얼마만큼의 상관관계를 가지고 있는지 판단한다.
SO-PMI를 계산하시는 식은 다음 식 2와 같다.

$$SO-PMI(x) = \sum_{px \in PX} PMI(x,px) - \sum_{nx \in NX} PMI(x,nx) \quad (2)$$

여기서 $PX$는 긍정기준단어의 집합, $NX$는 부정기준단어의 집합을 의미한다. $SO-PMI(x)$는 단어 x와 긍정단어 집합의 PMI합에서 부정단어집합과의 PMI합을 뺀 결과값을 의미한다.
본 논문에서는 온라인교육 플랫폼에 등록되어있는 튜터 개개인에 대한 리뷰를 수치화해서 제공하고자 한다.

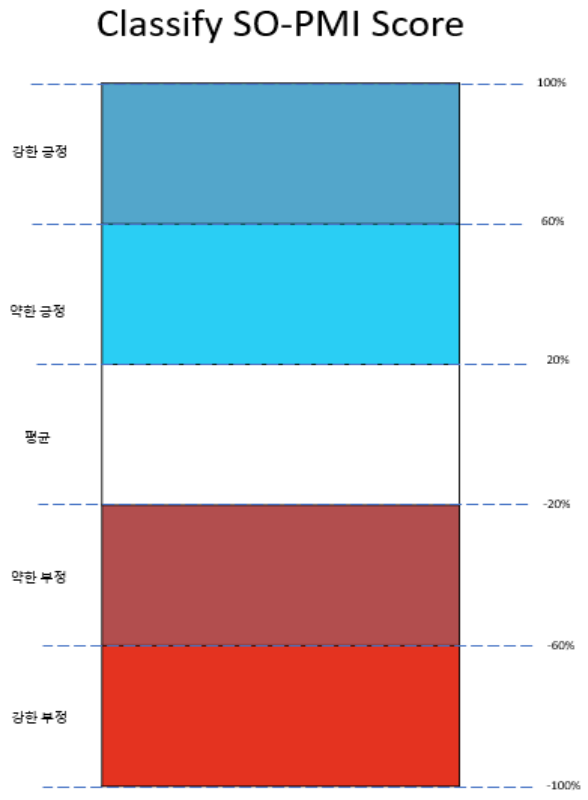
## 2-3. 단어의 극성 판단

감성사전을 구축하기 위해 본 연구에서는 수집된 리뷰에서 단어 간의 극성이 명확하게 구분되는 '형용사' 만을 활용하였다. 형용사 추출에 앞서 웹 크롤링 과정에서 발생한 불필요한 단어와 의미를 알 수 없는 단어를 제거하였으며 숫자, 특수기호 등의 불용어를 제거한다. 최종적으로 추출된 형용사들의 SO-PMI 점수를 계산해 단어의 극성을 판단한다. PMI는 두 단어 간의 유사성을 나타내는 점수로 PMI값이 높으면 두 단어 사이의 유사성이 높다는 것을 의미한다. 계산한 PMI값은 SO-PMI계산에 활용한다. SO-PMI 적용에 앞서 분석 데이터의 긍정기준단어 집합과 부정기준단어 집합을 설정한 후 분석대상 단어와 긍정기준단어와 PMI 합에서 부정기준단어와의 PMI합을 뺀 점수인 SO-PMI 점수를 계산한다. SO-PMI점수가 높으면 긍정기준단어 집합과 유사성이 높으며 반대로 부정기준단어 집합과 유사성이 낮으므로 긍정단어로 해석할 수 있다. 마찬가지로 SO-PMI점수가 낮으면 부정단어로 해석할 수 있다.

SO-PMI 값은 기준단어집합에 영향을 받으므로 기준단어집합을 적절하게 설정하는 것이 중요하다. 본 연구에서 기준단어집합을 만들기 위해 훈련데이터의 별점평균과 리뷰수를 이용해 상위 30%를 긍정단어집합, 하위 30%를 부정단어집합으로 설정하고자 한다. 두 집합에서 생성된 단어의 빈도수를 비교하여 해당 단어의 빈도수가 더 높은 집합에 귀속시킨다. 이렇게 생성된 긍정과 부정 상위 15개의 단어를 기준단어집단으로 만들었다. 기준단어집단의 개수선정은 기존연구에 근거했다[6].

## 2-3. 감성사전 구축

SO-PMI를 통해 단어별 극성점수를 판단한 후 이를 토대로 감성사전을 구축한다. 감성사전에서 각 단어는 긍부정의 단계별로 '강한 긍정', '약한 긍정', '강한 부정', '약한 부정,' '평균' 5개로 구분한다. SO-PMI가 0 이상이면서 그 값이 0 이상 값 중 상위 40%에 해당하면 '강한 긍정',상위 80%에 해당하면 '약한 긍정', 나머지는 '평균'으로 정의했으며 0 이하에서도 마찬가지의 규칙을 적용했다. SO-PMI를 기준으로 단어의 극성을 분류한 규칙은 다음 그림2와 같다.




(그림 2) SO-PMI 점수를 이용한 분류

## 3. 결론

본 논문에서는 SO-PMI를 사용한 단어의 극성 판별을 통해 감성사전 구축하는 시스템을 설계하고자 했다. 또한 감성사전을 이용해 온라인 교육 매칭플랫폼 이용 시 튜터에 대한 신뢰도 파악을 위한 자동화시스템을 설계했다.

본 연구는 SO-PMI를 활용하여 튜터 리뷰글에 대한 극성을 새로 정의했으며, 이를 감성사전 구축에 사용해 감성분석의 성능 향상을 도모하고자 했다. 그러나 각 분야별 감성분석에 필요한 맞춤형감성사전 구축을 위해 본 연구가 갖는 연구 한계점은 다음과 같으며 이후 추가적인 연구가 필요하다.

본 논문에서 훈련데이터를 활용한 맞춤형 감성사전 구축 시 수집하는 리뷰 중 단어의 극성이 뚜렷하다고 예상되는 형용사만을 한정하여 사용하려고 했다. 그러나 동사, 부사등 다른 여러 품사의 단어도 감성분석에 활용 가능할 것으로 예상되며 여러 품사를 포함한 감성사전 구축에 대한 연구가 진행되어야할 것으로 생각한다.

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2018년도 문화기술 연구개발 지원사업(R2018020083)으로 수행되었음.

## 참고문헌

[1] 이성주, 정진욱 “Substitute and Complementary Relationships between Online and Offline Courses in Foreign Language Education” 한국경제학보 25권 1호 pp.45-60 2018

[2] 김성빈, 임규연 “온라인 고등·평생교육 학습자의 학습참여 동기와 학습 만족도 관계에서 지각된 유용성, 자기조절학습 능력의 조절효과 검증” 평생학습사회 제13권 제3호, 2017.8, 85-107 (23 pages)

[3] 이수민, 이종혁, 김우제 “유아체육 O2O 매칭 플랫폼 설계“ 한국IT서비스학회 2018 춘계할술대회, 354~356p

[4] Alistair Kennedy, Diana Inkpen ″Sentiment Classification of Movie Reviews Using Contextual Valence Shifters″ Computational Intelligence, Volume 22, Number 2, 2006

[5] 이상훈, 최정, 김종우 ″영역별 맞춤형 감성사전 구축을 통한 영화리뷰 감성분석″ J Intell Inform Syst 2016 June: 22(2): 97~113

[6] Song J. S., and S. W. Lee, “Automatic Construction of Positive/Negative FeaturePredicate Dictionary for Polarity Classification of Product Reviews,” Journal of KIISE: Software and Applications, Vol.38, No.3(2013), 157~168