오피니언 마이닝을 통한 학습자 상태 분류 및 활동 모니터링 시스템

김동현[1] 장두수[2] 최용석†
[1 2]한양대학교 컴퓨터.소프트웨어학과
† 한양대학교 공과대학 컴퓨터공학부
[1]unicorndrill@naver.com, [2]tim0225@hanmail.net, † cys@hanyang.ac.kr

Classifying learner's states and Monitoring it by using opinion Mining

Dong hyun Kim[1] Doo Soo Chang[2] Yong SuK Choi†
[1 2]Department of Computer and Software. Hanyang University
† Division of Computer Science and Engineering. Hanyang University

요 약

오피니언 마이닝은 객관적인 정보를 필요로 하는 많은 분야에서 쓰이는 기법이다. 그러나 표현의 자유도가 높은 한글 Text를 분석하는 것은 상당히 어려운 일이다. 또한 한글 파괴 현상도 하나의 원인으로 대두되고 있다. 본 논문에서는 Text를 음소단위로 분할하는 Trigram-Signature 기법과 구문태그 패턴 기법을 통합한 새로운 상태 분류 기법을 제안했고, 만족, 불만, 낙담, 의문, 흥분 5가지 감정 분류를 시도했다. 이를 토대로 사용자의 정보를 그래프로 보여주는 시각화 시스템을 제안한다.

1. 서 론

오피니언 마이닝은 상품의 후기를 분석하여 소비자의 의견과 취향을 상품 개발에 반영하거나, SNS 글들을 분석하여 객관적인 데이터로 만들 때 사용되곤 한다. 온라인 학습 사이트의 글도 상품 후기이며 이를 분석한다면 본 논문에서 소개할 극성 사전을 구축하는데 사용할 수 있다. 또한 학습자의 상태판별에 쓰이는 알고리즘이 극성 사전을 참조한다면 더 정확한 상태 분류를 할 수 있다. 기존 연구[1,2,3]에서는 시그니처 기법과 구문태그패턴 기법을 통합한 새로운 오피니언 분류 기법을 통해서 주어진 Text를 만족, 불만, 흥분, 의문 4가지 감정 상태로 분류하는 알고리즘을 제안했다.

본 논문에서는 '낙담'이라는 새로운 상태 추가 및 기존 분류기법을 개선한 알고리즘을 소개한다. 또한 알고리즘을 통해 상태를 분류한 데이터들을 토대로 사용자의 정보를 시각화 할 수 있는 모니터링 방법을 개선하였다.

이를 위해 본 논문 2장에서는 오피니언 마이닝과 관련된 연구들을 소개한다. 3장에서는 본 논문에서 사용한 알고리즘 및 사전 구축에 대해서 소개한다. 4장에서는 5가지 상태 분류 성능을 평가하고, 5장에서는 결론 및 추후 연구에 대해서 논의한다.

2. 관련 연구

오피니언 마이닝에 관한 연구는 여행지 정보나 고객의견 및 상품에 대한 각종 후기 등의 데이터에서 더 객관적인 정보를 얻어내는 방식으로 이뤄진다. 이전에 진행된 연구에서는 OSAR(Opinion Association Rules Network) 알고리즘을 적용하여 감성 사전을 구성[6]하거나, PMI(Point-wise Mutual Information)를 이용하여 어휘의 의미에 따른 극성을 판단하는 기법[7] 등 다양한 연구가 진행되었다.

기존연구[1,2,3]에서는 시그니처 기법과 구문태그패턴 기법을 통합한 기법이 다른 기법들에 비해 성능이 우수하다는 것을 입증했다. 특히, 한글 Text를 음소단위로 분할하여 Trigram -Signature를 얻어내고 이를 분석하는 시그니처 기법의 장점과 구문태그패턴으

* 본 연구는 산업통상자원부의 재원으로 기술혁신사업의 지원을 받아 수행한 연구 과제임 (No. 10060086, 개인 서비스용 로봇을 위한 지능-지식 집약 ·개방 ·진화형 로봇지능 소프트웨어 프레임워크 기술 개발)
* 본 연구는 2016년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초 연구사업임 (No.NRF-2015R1D1A1A01060950).
† 교신저자(Corresponding author): 한양대학교 공과대학 컴퓨터공학부 교수 최용석(cys@hanyang.ac.kr)

로 어휘쌍을 추출해 극성을 학습시키는 구문태그패턴 기법의 장점을 가지는 새로운 기법은 한글 Text 분석에 매우 적합했다.

본 연구에서는 다수의 학습자가 '학습 의욕 저하'를 나타내는 것을 발견하게 됐고, 관련 자료[8]를 참조하여 '낙담' 이라는 새로운 상태를 추가하게 되었다. 실제로 '낙담'을 하나의 상태로 지정하여 분류를 진행한 논문은 거의 없으며, 본 논문에서 기존 방식[1,2,3]개선을 목표로 하였다. 또한 만족, 불만, 낙담, 흥분, 의문 5가지 감정 상태 로 분류한 결과를 토대로 학습자의 상태를 그래프로 시각화해주는 시스템을 제안한다.

3. 상태 분류

3.1 사전 구축 및 설계

사전은 구문태그패턴과 어절 쌍의 극성값을 가지는 구문태그패턴 극성 사전과 어절의 극성값을 가지는 어절 극성사전으로 구성된다. 극성사전 구성에 사용될 Text는 특수기호 제거 및 띄어쓰기 작업이 선행된다. 본 논문의 실험 대상은 한글 Text이므로 영문도 특수기호로 취급하여 제거한다. 그리고 KLT 형태소 분석기[4]를 사용, 자동 띄어쓰기 및 문장 종결 어미의 유무를 판별하여 문장을 분리한다. 그리고 NLPHUB 의존구문분석기[5]로 학습자의 글에서 구문 태그를 뽑아내고 구문 태그 패턴과 일치하는 어절쌍을 찾아낸다.

상태 A에 대한 극성 사전 구축시 text(doc)는 띄어쓰기 단위로 나뉘어져 들어오게 된다. 예를 들어, text(doc)의 i번째 어절인 '안녕하세요'가 상태 A인 text(doc)에서 나왔다면 +1, 상태 A가 아닌 text(doc)에서 등장했다면 -1을 극성값에 더해준다. 식(1)은 이를 나타낸다.

$$Word(i)=\begin{cases}1: doc\text{상태가 }A\text{에 해당}\\-1: doc\text{상태가 }\lfloor A\text{에 해당}\end{cases}\quad(1)$$
(상태 A에 대한 극성사전 구축시)

식(2)에서 Word(i).Polarity는 text(doc)의 i번째 어절 또는 어절쌍의 최종적인 극성값을 말하며, 해당 어절이 포함된 text(doc)의 비율로 정의하였다. M은 해당 어절/어절쌍을 포함하는 text(doc)의 개수이다.

$$Word(i).Polarity=\frac{\sum_{i=1}^{M}Word(i)}{M}\quad(2)$$

기존 연구에서는 만족-불만 판별에만 극성 사전을 사용했다. 그러나 본 논문에서는 이를 개선하여 모든 상태 분류기에서 사용할 어절 극성사전과 구문태그패턴 극성사전을 구축했다. 사용되는 어절 및 어휘가 한정적인 의문과 흥분 상태의 경우, 이전에 수집해 놓았던 단어사전을 기반으로 극성사전을 구축하였다.

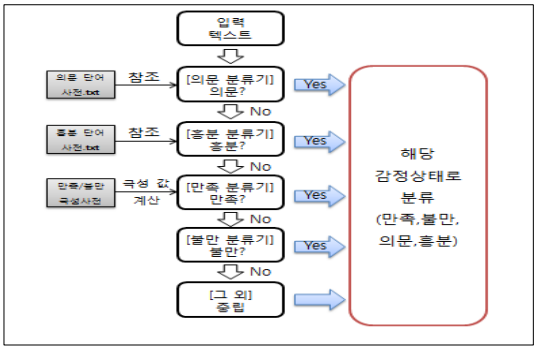
의문 극성 사전을 구축하는 경우, 의문을 나타내기 위해서 사용되는 어절(Ex. 까요, 나요, 니까) 에 초점을 맞춰서 극성 값을 부여했다. 또한 흥분은 일반적인 불만보다 상대적으로 과격한 어휘 표현이나 심한 욕설이 들어있는 어절에 초점을 맞춰서 극성값을 부여했다.

3.2 상태 분류기 설계

분류 작업 전, 대상 text(doc)에 대하여 자동 띄어쓰기 및 특수기호처리와 문장분리 작업과 구문 태그 패턴 추출을 시행한다. 이후 text(doc)에 포함된 어절 및 구문태그 패턴의 음소단위인 Trigram-Signature를 만들고 어절 극성사전과 구문태그패턴 극성사전에 저장된 시그니처와의 유사도를 계산한다[2]. 계산된 유사도가 Threshold 값 이상인 어절과 구문태그패턴의 극성값(Word(i).Polarity)을 추출한다. 그리고 식(2)를 포함한 식(3)을 사용하여 유사도 가중합을 계산해 최종적인 text(doc)의 극성값을 추출한다. 식(3)의 유사도는 사전에 있는 어절/어절쌍의 시그니처와 분석 대상이 되는 text(doc)의 어절 시그니처간의 유사도 값이다. n은 text(doc) 의 유사도 값이 Threshold 이상인 어절들의 개수이다.

$$TextPolarity=\sum_{j}^{n}(Word(j).Polarity*\text{유사도})\quad(3)$$

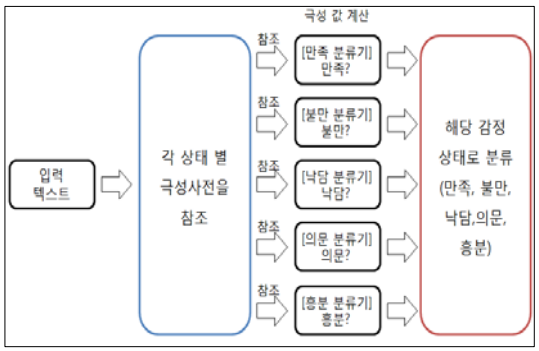
기존 연구에서는 그림 1처럼 만족-불만 판별시에만 (3)식을 사용했다. 의문 분류기와 흥분 분류기에서 걸러지지 않은 text(doc)의 최종적인 극성값이 양수일 때 만족, 음수일 때 불만, 0일 때 중립으로 분류했다.



[그림 1: 기존 분류 방식 순서도]

본 논문에서는 흥분과 의문 상태를 단어의 출현 유무로 판별하는 기존의 방식을 그림 2처럼 극성사전을

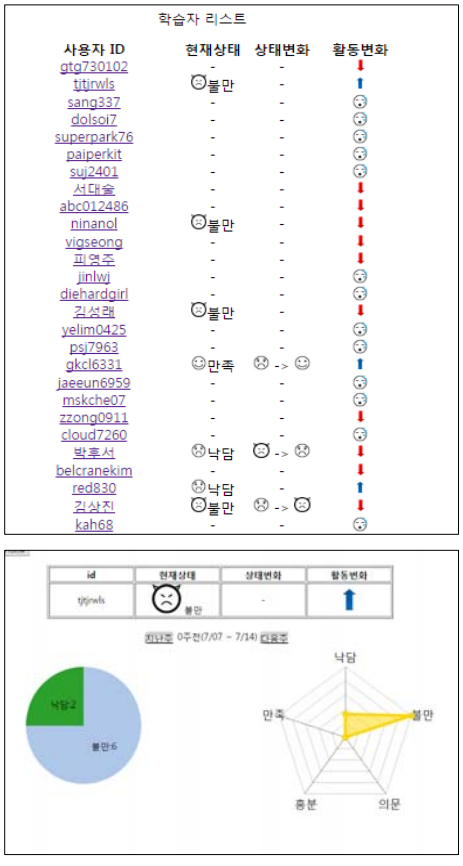
이용하는 방식으로 개선하였다. 결과적으로 모든 상태분류기에서 각 상태별 사전을 참조하여 극성값을 계산하고 이를 상태별 Threshold값과 비교해 상태를 판별했다. 어떤 상태도 나타나지 않는다면 상태가 없으며, 2개 이상의 상태가 나오는 경우엔 극성값이 가장 큰 상태로 대표하도록 했다.

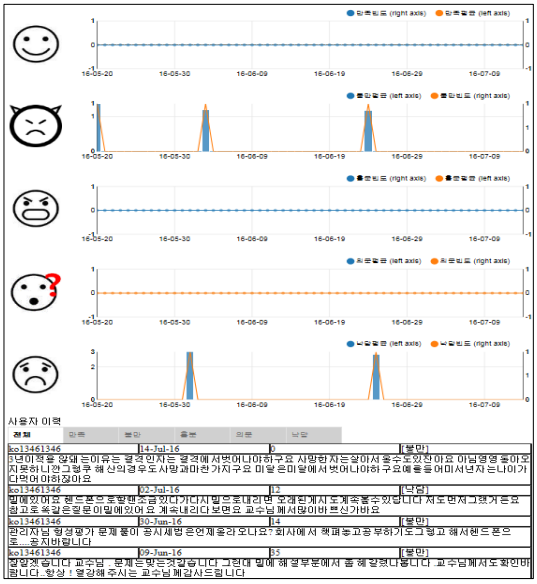


[그림 2: 현재 분류 방식 순서도]

3.3 시각화 시스템 설계

온라인 학습자가 작성한 글과 날짜를 토대로 학습자의 상태를 간단한 이모티콘과 그래프로 나타낸다. 또한 학습자의 현재 상태를 비롯, 상태변화 및 활동변화를 원형 그래프 및 오각형 형태의 방사형 그래프와 꺾은선 그래프 형태로 나타낸다. 그림 3처럼 초기 화면에서는 사용자의 ID 및 상태를 볼 수 있으며, ID를 클릭하게 되면 학습자의 상태를 자세히 볼 수 있다.





[그림 3: 그래프 및 게시글 리스트]

4. 실험

4.1 실험 데이터

사전 구축을 위한 데이터들은 다수의 학습 사이트에 게시되어 있는 강의 후기 및 평가 게시판 글들을 위주로 수집했다. 수집한 Text들을 수동분류기를 사용하여 만족, 불만, 낙담, 흥분, 의문 중 해당하는 상태로 분류했다. 그 결과 총 10,900건의 Text를 수집했고 이를 분석하여 약 18만개의 극성값 정보를 구성했고 극성 사전 구성에 사용될 트레이닝 데이터와 성능평가에 쓰일 테스트 데이터를 구성했다.

트레이닝 데이터는 상태 A에 해당하는 글과 상태가 Not A인 글의 비율을 1:1로 구성했고, 테스트 데이터는 트레이닝 데이터에 포함되지 않는 나머지 데이터들을 사용하여 마찬가지의 방법으로 구성했다.

4.2 척도

알고리즘의 성능을 판단하기 위해 본 실험에서는 Precision(정확도), Recall(재현율), f-measure(조화평균)을 사용했다. 정확도나 재현율 중 하나의 지표에 집중하여 평가하지 않고, 두 가지 지표 모두 중점을 두고 계산된 조화평균을 기준으로 성능을 평가했다. 정확도, 재현율, 조화평균을 구하는 식은 다음과 같다.

$$Precision = \frac{\text{실제 해당 상태로 분류된 글의 개수}}{\text{해당 상태로 분류된 글의 개수}}$$

$$Recall = \frac{\text{실제 해당 상태로 분류된 글의 개수}}{\text{실제로 해당 상태인 글의 개수}}$$

$$f1-measure = \frac{2*Precision*Recall}{Precision+Recall}$$

4.3 실험 결과 및 해석

본 논문에서 제안한 기법을 사용해서 만족, 불만, 낙담, 의문, 흥분 분류를 시행했고 Precision, Recall, F-measure 값은 다음과 같다. 본 연구에서 새로 추가한 낙담은 다른 연구에서는 분류를 시도하지 않은 상태라는 점을 감안하더라도 높은 성능을 보여준다.
기존 연구에서는 특정 연산 없이 단어의 유무만을 판단하여 흥분/의문 상태를 분류하였으나 방법자체가 극성값을 계산하는 것보다 신뢰도가 떨어진다. 따라서 본 논문에서 보인 흥분/의문 분류기 성능은 상대적으로 신뢰도가 높으며 성능면에서도 크게 차이가 나지 않기 때문에 타당하다고 볼 수 있다. 그리고 기존 연구에서는 만족/불만 극성사전 구축 시 만족데이터와 불만 데이터를 1:1의 비율로 맞춰서 사용했으나, 본 논문에서는 한 상태에 대한 극성사전 구축 시 해당 상태 데이터와 나머지 4가지 상태의 데이터를 1:1 비율로 사용해서 사전을 구축했다. 따라서 상태마다 일부 중복되는 속성이 있기 때문에 일부 상태에 대해서 성능 저하가 발생했다. 예를 들어, 불만 극성사전 구축 시 불만이 아닌 데이터에는 만족,흥분,의문,낙담 데이터가 포함되며, 일부 흥분 및 낙담 데이터에서 불만에 해당하는 어절 및 어절 쌍들이 다수 포함되어 있음이 확인되었다. 이는 불만, 흥분, 낙담 상태가 공통적으로 부정적인 속성을 가지기 때문이다.

기법	상태	Precision	Recall	F-measure
만족/불만 극성사전 만 사용	만족	0.83	0.9	0.8
	불만	0.89	0.82	0.85
	흥분	0.90	0.78	0.83
	의문	0.95	0.79	0.86
각 상태별 극성사전 사용	만족	0.89	0.66	0.76
	불만	0.58	0.885	0.70
	흥분	0.83	0.60	0.70
	의문	0.81	0.81	0.81
	낙담	0.70	0.91	0.79

[표 : 상태 분류 성능 평가표]

5. 결론 및 향후 연구

한글 Text를 대상으로 오피니언 마이닝을 할 때 가장 문제가 되는 것은 한글이 가지는 표현의 자유성이다. 본 논문에서는 이러한 문제를 해결하기 위해서 한글을 음소 단위로 분리하여 분석하고 상태를 분류하는 방법을 제안했다. 그러나 학습 사이트에 게시된 글이라는 카테고리 내에서 오피니언 마이닝을 실시한 것이기 때문에 본 실험에서 구성한 극성사전이 일반적인 글을 대상으로 높은 성능을 보일 지는 미지수이다. 그러나 시그니처 기법과 구문태그패턴 기법을 통합한 새로운 기법은 온라인에서 많이 발생하는 한글 파괴 현상 속의 Text 분석에 강점을 보일 것이다. 또한 학습 사이트라는 카테고리를 벗어나서 일반적인 글들을 충분히 수집하여 표본을 더 크게 구성하고 본 논문에서 제안한 바를 적용한다면 원하는 목적에 맞게 오피니언 마이닝을 수월하게 진행할 수 있을 것이다.

향후에는 극성사전 구축 시 활용되는 데이터 정제를 통해 중복 속성의 글이 트레이닝 데이터로 활용되지 않도록 하여 성능 저하 문제를 해결할 예정이다. 또한 학습 도메인에서 자주 검출되는 격려, 요청 성향의 글에 대한 상태 분류를 위해 상태 분류기를 확장 개발할 계획이다.

6. 참고 문헌

[1] 김도연,장두수,최용석, “오피니언 구문 태그 패턴을 이용한 감정상태 분류 성능 향상”, 한국정보과학회 2015 한국컴퓨터종합학술대회 논문집, p.978-980, 2015

[2] 장두수,김도연,최용석, “한글 음소단위 Trigram-Signature 기반의 오피니언 마이닝 기법”, 한국정보과학회 2015 한국컴퓨터 종합 학술대회 논문집, p.811-813, 2015

[3] 김도연,장두수,최용석, “오피니언 마이닝을 이용한 온라인 학습자 모니터링” 한국정보과학회 2015년 동계학술발표회 논문집 p.795-797, 2015

[4] 강승식, “한글 문장의 자동 띄어쓰기를 위한 어절 블록 양방향 알고리즘”, 정보과학회 논문지(B), 27RNJS 4호, P.441-447

[5] D. Choi, J. Park, K Choi, ″Korean Treebank Transformation for Parser Training″, Proceedings of the 50th Annual Meeting of the Association for Comput-
ational Linguistics, ACL 2012, pp. 78-88, 2012

[6] 이윤주, 서지훈, 최진탁, “SNS 텍스트 콘텐츠를 활용한 오피니언 마이닝 기반의 패션 트랜드 마케팅 예측 분석” , 한국정보기술학회논문지 제12권 제12호 (JKIIT, Vol.12, No.12), 2014.12, 163-170

[7] 문찬영,인관호,김응모, “오피니언 마이닝을 이용한 여행지 정보 비교/분석”, 한국정보과학회 2012 가을 학술발표논문집 제39권 제2호(C), 2012.11, 92-94

[8] Paul Ekman, Wallace V.Friesen, “Basic Emotions”, 1972