규칙과 통계 정보에 기반을 둔 상품평 분석 시스템1)

김민호*, 최현수*, 권혁철*
*부산대학교 컴퓨터공학과
e-mail:{karma, gl2een33, hckwon}@pusan.ac.kr

A Product Review Analysis System using Rules and Statistical Information

Minho Kim*, Hyunsoo Choi*, Hyuk-Chul Kwon*
*Dept of Computer Science & Engineering, Pusan National University

요 약

상품평은 구매 예정자의 의사 결정에 큰 도움을 준다. 그러나 하나의 상품에 관한 상품평의 수가 많고 의견도 다양하여 모든 상품평을 읽고 상품에 대한 직관적인 판단을 내리기가 어렵다. 본 논문에서는 하나의 상품에 대한 모든 상품평을 분석하고 각각의 속성별로 극성(긍정, 부정) 정보를 추출하여 구매 예정자에게 제공함으로써 해당 상품이 어떠한 평가를 받고 있는지 빠른 판단이 가능하게 한다. 한국어의 언어적 특징을 반영하여 속성별 어휘 자질을 추출하고 이를 바탕으로 상품의 속성별 극성을 판단한다. 또한, 기구축한 속성별 어휘 사전의 자료부족 문제로 말미암아 상품평을 분석할 수 없을 때는 전체 어휘의 극성정보를 이용하여 상품의 전체 극성을 판단한다.

1. 서론

특정 상품을 구매하고자 하는 구매 예정자의 의사 결정에 가장 많은 영향을 끼치는 요소 중 하나는 구매자의 상품 평가이다. 인터넷과 모바일 환경이 발달하기 전에는 가까운 지인의 의견만을 들을 수 있었기 때문에 접할 수 있는 정보의 양이 제한적이었으나, 최근에는 인터넷 커뮤니티, 블로그, 소셜 네트워크 서비스 등 다양한 매체를 통해 접할 수 있는 정보의 양이 폭발적으로 늘어났다. 이 때문에 하나의 상품에 관한 상품평의 수가 많고 의견도 다양하여 모든 상품평을 읽고 상품에 대한 직관적인 판단을 내리기가 어렵다. 이러한 다양한 의견을 분석하여 해당 주제에 대한 평판을 도출해내고자 시작된 연구가 '오피니언 마이닝(opinion mining)'이다.

일반적으로 오피니언 마이닝에서 상품평을 분석하고자 가장 많이 사용되는 방법은 속성별 어휘 사전에 기반을 둔 극성(긍정 또는 부정) 분석이다. 예를 들어, '배송'에 관한 긍정적인 어휘인 '빠르다'와 부정적인 어휘인 '느리다'를 사전에 미리 구축하고, '배송이 빠르다.'라는 문장에서 '배송'에 관한 평가가 긍정적이라는 정보를 추출한다.

속성별 어휘 사전에 기반을 둔 극성 분석에서 분석의 정확도를 떨어트리는 요인은 크게 세 가지이다. 첫 번째는 속성별 어휘 사전 구축의 어려움이다. 속성별 어휘 사전은 하나의 어휘가 상품과 속성에 따라 다른 의미로 사용되기 때문에 개별적인 구축이 필요하다. 예를 들어, '빠르다'는 '배송'에서는 긍정적인 어휘로 사용되지만, '배터리 소모가 빠르다.'에서와 같이 '배터리'에서는 부정적인 어휘로 사용될 수 있다. 두 번째는 속성별 어휘의 문맥 정보를 반영해야 한다는 점이다. 예를 들어, '배터리 소모가 빠르지는 않다.'와 같이 문형에 따라 '빠르다'라는 부정적인 어휘가 다른 요소와 결합하여 긍정적인 의미로 사용되기도 한다. 마지막으로 자료부족 문제이다. 속성별 어휘 사전에 포함된 단어가 하나라도 나타나지 않았을 때 극성 분석을 할 수 없다는 문제가 있다.

본 논문에서는 속성별 어휘 사전을 구축하고 한국어의 언어적 특징을 반영하여 속성별 어휘 자질을 추출한다. 또한, 기구축한 속성별 어휘 사전의 자료부족 문제로 말미암아 속성별로 상품평을 분석할 수 없을 때는 전체 어휘의 극성정보를 이용하여 상품의 전체 극성을 판단한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 상품평 분석에 관한 기존 연구를 소개하고, 3장에서는 속성별 어휘 사전에 기반을 둔 속성별 극성 분류와 통계적 정보에 기반을 둔 상품평 극성 분류에 관해 기술한다. 4장에서는 제안한 방법을 이용한 실험 결과를 분석하고, 5장에서 결론 및 향후 연구에 대해서 논의한다.

2. 관련연구

상품평 분석에 관한 연구는 속성별 어휘 사전에 기반을 둔 극성 분석을 중심으로 이루어졌다. 초기 연구에서는 속

1) 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2012R1A2A2A06046730).

성별 어휘 사전을 구축하고 어휘 간 비교를 통해 극성 정보를 추출하였다[1, 2, 3]. 극성 정보를 추출할 때 수식어에 따라 극성 강도를 달리하거나[1], 추출된 정보를 조합하여 극성을 판단하는 방법[2, 3]의 차이는 있다. 그러나 이들 방법 모두 어휘 간 일치를 전제로 극성 정보를 추출하므로 어휘에 철자 오류가 있거나 속성 어휘의 문맥이 달라지면 극성 정보를 추출할 수 없으므로 속성 어휘가 사용된 패턴을 분석하여 패턴 간 일치를 통해 극성 정보를 추출하려는 연구가 있다[4, 5]. 최근에는 속성별 어휘 사전의 자료부족 문제를 완화하고자 속성에 따른 의미 차이가 없는 어휘를 구분하여 공통 어휘 사전과 속성별 어휘 사전을 구분하고, 관형사와 부정어 처리를 통해 극성 추출의 정확도를 높인 연구도 있다[6]. 그러나 이들 연구 모두 한국어의 언어적 특성을 깊게 반영하지 못하여 속성별 어휘 정보 추출의 정확도가 낮고, 속성별 어휘 사전에 포함된 단어가 하나라도 나타나지 않았을 때 극성 분석을 할 수 없다는 문제가 있다.

본 논문에서는 이러한 문제를 해결하고자 부정어 처리, 수식어 처리, 다어절 처리 등 부분적인 구문 분석을 통해 속성별 어휘 추출 정확도를 높인다. 또한, 속성에 상관없이 항상 같은 극성을 가지는 어휘를 학습 자질로 이용하여 기계학습을 함으로써 전체 상품평에 대한 극성 분류가 가능하도록 하였다.

3. 규칙과 통계 정보에 기반을 둔 상품평 분석

3.1 속성별 어휘 사전 구축

앞에서 언급하였듯이 속성별 어휘 사전은 개별 상품에 따라 각각 구축되어야 한다. 본 논문에서는 '스마트폰'을 주제로 '배터리, 디자인, 발열, 화질, 크기, 터치감'의 여섯 가지 속성을 정하고 속성별 어휘 사전을 구축하였다. 상품평 데이터를 수집하여 속성별 어휘를 분석한 결과, 극성 어휘만으로 속성별 긍정/부정 표현으로 판단이 가능한 경우와 속성 명사와 극성 어휘가 같이 쓰일 때 속성별 긍정/부정 표현으로 판단할 수 있는 경우가 존재했다. 예를 들어, '손난로'라는 단어는 배터리에 대한 부정 표현으로 바로 판단할 수 있고 '세련되다'라는 단어는 디자인에 대한 긍정 표현으로 바로 판단할 수 있다. 또한, '화면이 시원시원하다'는 화면이라는 속성 명사와 '시원시원하다'는 어휘가 같이 쓰일 때 화면 크기에 대한 긍정 표현으로 판단할 수 있다.

이에 따라, 극성 어휘만으로 속성별 긍정 혹은 부정 표현으로 판단하는 경우와 속성 명사와 극성 어휘가 같이 쓰일 때 판단하는 경우를 나누어 세분된 속성별 어휘 사전을 구축하였다.

3.2 속성별 어휘 사전을 이용한 극성 정보 추출

기존 연구에서는 속성별 어휘 사전을 이용하여 단순 어휘 간 일치 비교를 통해 극성 정보를 추출하였다. 이로 인해 '디자인이 좋다', '디자인이 좋지 않다', '디자인이 좋은 것만은 아니다', '디자인이 조금 좋다' 등에서 모두 같은 극성 정보를 추출하는 문제가 있다. 본 연구에서는 한국어의 언어학적 특징을 이용한 규칙을 생성하여 적용함으로써 극성 추출의 정확도를 향상한다. 극성 추출에 이용하는 통사 규칙은 다음과 같다.

부정어 결합: 속성별 긍정, 부정 어휘를 검색하였을 때 어휘 근처에 '안', '못', '않다', '없다', '아니다'와 같은 부정어가 존재하는 경우나 '-는데', '-지만'과 같이 표현을 반전시켜주는 역접 표현이 올 경우, 찾은 어휘가 긍정 표현이면 부정 표현으로, 어휘가 부정 표현이면 긍정 표현으로 처리한다. 또한, '없지는 않다'처럼 부정 표현이 두 번 나올 때 이중 부정으로 처리한다.

수식어 처리: 예를 들어, '배터리 진짜 오래가요.'처럼 어휘 근처에 '정말', '매우', '진짜' 같은 강조 어휘가 나오는 경우와 '디자인 별로예요. 그런데 배터리 용량은 진짜 대박!'처럼 '그런데', '하지만' 같은 역접 관계 접속사가 나오는 경우 찾은 긍정 혹은 부정 어휘 수치에 가중치를 준다. 그리고 '다른 스마트폰보다 화질이 선명해 보이더군요.' 처럼 '~보다', '~에 비해'와 같이 비교 표현으로 상대적으로 강조되는 어휘가 나오는 경우도 찾은 긍정 혹은 부정 어휘 수치에 가중치를 준다.

다어절 처리: '디자인은 마음에 드네요.'에서 '마음에 든다.'처럼 두 개 이상의 어휘가 모여 긍정 혹은 부정 표현이 되는 경우는 별도로 처리하도록 한다.

3.3 기계학습을 이용한 극성 분류

기존 연구에서는 속성별 어휘 사전을 이용하여 극성 정보를 추출하고, 속성별 극성을 분류한다. 그러나 속성별 어휘 사전에 포함된 단어가 나타나지 않을 때는 해당 속성의 극성은 중립이 된다. 본 연구에서는 속성별 어휘 사전의 자료부족 문제로 생기는 문제를 최소화하고자 고빈도 어휘에 대한 극성을 모두 분류한다. 특정 상품평에서 속성별 어휘 사전에 포함된 단어가 나타나지 않으면 기계학습을 통해 상품평 전체의 극성을 판단하도록 한다. 기계학습에는 SVM(Support Vector Machine)을 이용한다.

4. 실험

4.1 실험환경

웹 크롤러와 블로그 open API를 이용하여 1,209개의 상

품평을 수집하였다. 각 상품평을 분석하여 여섯 가지 속성에 따라 극성 분류를 하였다. 이 중 1,000개는 학습 데이터로 나머지 209개는 평가 데이터로 활용하였다. 각 데이터의 속성별 극성 정보는 표 1과 같다.

<표 1> 실험 데이터의 속성별 극성 분류

• 학습데이터 수 : 1,000개

feature	배터리	디자인	발열	화질	크기	터치감
긍정 개수	171	368	110	364	305	201
부정 개수	325	240	173	183	150	110
합계	496	608	283	547	455	311

• 테스트데이터 수 : 209개

feature	배터리	디자인	발열	화질	크기	터치감
긍정 개수	34	120	23	98	52	41
부정 개수	39	32	33	24	31	44
합계	73	152	56	122	83	85

4.2 실험결과

그림 1은 속성별 어휘 사전 중 어휘만으로 속성의 극성을 판단할 수 있는 어휘들로 실험한 결과이다. 부정어와 수식어 처리 등을 하지 않았기 때문에 전반적인 성능이 매우 낮은 것을 알 수 있다.

Feature	긍정		부정	
	Precision	Recall	Precision	Recall
배터리	0.00 %	0.00 %	66.66 %	3.84 %
디자인	90.47 %	58.16 %	84.61 %	21.15 %
발열	100.00 %	1.96 %	90.00 %	17.64 %
화질	72.72 %	50.63 %	83.33 %	9.43 %
크기	100.00 %	16.07 %	50.00 %	2.00 %
터치감	0.00 %	0.00 %	83.33 %	9.80 %

구분	Accuracy
배터리	1.90%
디자인	45.33%
발열	9.80%
화질	34.09%
크기	9.43%
터치감	4.95%

(그림 1) 속성별 어휘만 이용한 결과

그림 2는 부정어와 수식어 그리고 다어절 처리 등을 통해 속성별 어휘 추출의 정확도를 높였을 때의 실험 결과이다. 단순히 속성 어휘 간 일치만을 이용하여 극성 정보를 추출하였을 때 보다 성능이 높아진 것을 알 수 있다.

Feature	긍정		부정	
	Precision	Recall	Precision	Recall
배터리	90.90 %	56.60 %	72.72 %	76.92 %
디자인	87.87 %	59.18 %	60.00 %	63.46 %
발열	64.28 %	17.64 %	51.85 %	82.35 %
화질	80.30 %	67.08 %	78.04 %	60.37 %
크기	71.87 %	41.07 %	68.42 %	52.00 %
터치감	83.78 %	62.00 %	78.26 %	70.58 %

구분	Accuracy
배터리	66.66%
디자인	60.66%
발열	50.00%
화질	64.39%
크기	46.22%
터치감	66.33%

(그림 2) 부정어/수식어/다어절 처리

그림 3은 극성 분류를 한 고빈도 어휘를 학습 자질로 이용하여 평가 데이터를 극성 분류 실험을 한 결과이다. 학습 모델로 SVM을 사용하였으며, SVM Tool은 SVMlight[7]를 사용하였다. 다수의 실험을 통한 경험적

극성	Precision	Recall	Accuracy
긍정	65.97% (95/144)	85.58% (95/111)	68.90%
부정	75.38% (49/65)	50.00% (49/98)	

(그림 3) 전체 상품평의 극성 분류

정보를 이용하여 가장 좋은 결과 값을 보인 커널함수는 polynomial 커널 함수로서 degree(d)는 4였다.

5. 결론 및 향후 연구

본 논문에서는 속성별 어휘 사전을 구축하고 한국어의 언어적 특징을 반영하여 속성별 어휘 자질을 추출하였다. 또한, 기구축한 속성별 어휘 사전의 자료부족 문제로 말미암아 속성별로 상품평을 분석할 수 없을 때는 전체 어휘의 극성정보를 이용하여 상품의 전체 극성을 판단한다. 한국어의 언어적 특징을 반영하여 속성별 어휘 자질을 추출하였을 때가 그렇지 않았을 때보다 분류 성능이 월등히 높았으며, 속성별 어휘 사전에 포함된 단어가 나타나지 않았을 때도 극성 분류가 가능하다는 것을 실험적으로 보여주었다.

향후 연구에서는 속성별 어휘 사전을 구축하는데 드는 비용을 최소화하도록 반자동으로 속성별 어휘 사전을 구축하는 연구를 할 예정이다.

참고문헌

[1] 명재석, 이동주, 이상구, “반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템”, 정보과학회논문지: 소프트웨어 및 응용, 제35권 제6호, pp.392-403, 2008.06

[2] 장재영, “온라인 쇼핑몰의 상품평 자동분류를 위한 감성분석 알고리즘”, 한국전자거래학회지, 제14권 제4호, pp.19-33, 2009.11

[3] 윤홍준, 김한준, 장재영, “오피니언 마이닝 기술을 이용한 효율적 상품평 검색 기법”, 정보과학회논문지: 컴퓨팅의 실제 및 레터, 제16권 제2호, pp.222-226, 2010.02

[4] 신준수, 김학수, “강건한 한국어 상품평의 감정 분류를 위한 패턴 기반 자질 추출 방법”, 정보과학회논문지 : 소프트웨어 및 응용, 제37권 제12호, pp.946-950, 2010.12

[5] 김정호, 차명훈, 김명규, 채수환, “어미변화를 고려한 감성 구문 패턴을 이용한 상품평 의견 분류”, 2010 한국컴퓨터종합학술대회 논문집, 제37권 제1호(C), pp.285-290, 2010.06

[6] 송종석, 이수원, “상품평 극성 분류를 위한 특징별 서술어 긍정/부정 사전 자동 구축”, 정보과학회논문지: 소프트웨어 및 응용, 제38권 제3호, pp.157-168, 2011.03

[7] SVMLight, http://svmlight.joachims.org/