상품평 분석을 통한 상품 평가 요약 시스템

김제상O,정군영, 권인호, 이현아
금오공과대학교, 컴퓨터소프트웨어공학과
oiu124@naver.com, gunyoung20@naver.com, kih6412@naver.com, halee@kumoh.ac.kr

Product Review Summarization through Review Sentence Analysis

Je-Sang KimO, Gun-young jung, In-ho Gwan, Hyun-Ah Lee
Kumoh National Institute of Technology, Dept. of Computer Software Engineering

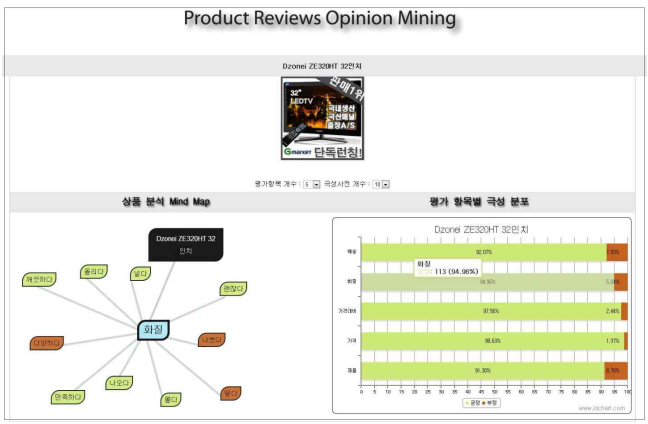
요 약

다수의 상품평 요약은 인터넷 쇼핑몰 고객에게 편의를 제공할 수 있다. 본 논문에서는 상품평 요약 시스템의 성능 향상을 위한 방안을 제안한다. 시스템은 크게 상품평의 평가 항목 추출과 극성 사전 생성, 극성 판별 단계로 구성된다. 평가 항목 추출에서는 외부 연관도의 영향력을 줄이고, 극성 사전 생성에서는 단어 거리 평균을 적용한다. 제안한 방식을 사용하였을 때 평가 항목에 대한 문장의 극성 판별 시 90.8%의 정확율을 보였다.

주제어: 평가 항목, 상품평 요약, 극성 판별, 감성 분류

1. 서론

본 논문에서는 상품평을 자동으로 요약하기 위한 시스템을 제안한다. 본 시스템에서는 평가 항목, 즉 '색상', '크기' 등의 상품의 다양한 특성별로 상품을 긍정과 부정 중 어느 극성(polarity)인지 판별하여 사용자가 원하는 정보를 알아보기 쉽게 제시한다. 본 시스템은 상품평을 수집하고 정형화한 뒤, 상품평 내 어휘 정보와 웹 검색을 활용하여 평가 항목을 추출하고, 언어 처리기법을 이용하여 평가 항목별 평가 어휘와 평가 어휘의 극성을 자동으로 추출하여, 평가 항목별 극성을 [그림 1]에서와 같이 그래프 형태로 요약하여 출력한다. 각 단계는 상품평과 웹 검색 결과를 활용하여 모두 자동으로 처리된다.



[그림 1] 상품평 요약 시스템
(평가 항목별 평가어휘 출력 / 평가 항목별 극성 분포)

2. 관련 연구

상품평의 의견 추출에서는 세부적인 상품 특징(예를 들어, 스커트의 경우 '사이즈', 디지털 카메라의 경우 '액정')에 대한 기존 구매자의 평가가 구매 예정자의 구매 결정에 중요한 판단 기준이 될 수 있어 특징 기반 요약이 필요하다. 국내에서 상품 특징 단어를 추출하는 방법은 k-Structure[1]를 이용한 방법과 PMI-IR 방식[2]을 이용한 방법 등이 있다. [3]과 [4]에서는 평가 항목을 추출하기 위해서 특징 추출에 사용되는 통계적 기법으로 PMI를 이용하며, [5]에서는 ME-NAS모델을 사용하여 상품평의 속성과 감정 데이터를 추출한다. 이러한 연구들은 평가 항목에 따라 달라지는 평가 어휘의 극성을 처리 하기 위한 연구들으로, 특징 기반 요약 방법을 사용하고 있다.

[6]는 도메인별로 평점이 포함된 상품평을 수집하여 상품평의 평점 데이터를 활용하여 극성 사전을 생성하고, [7]은 평가 항목에 의존적이지 않은 소량의 시드(seed) 극성 사전을 이용하여 자동 확장 방식으로 극성 사전을 생성한다. [5]는 어절 유니그램 모델을 제안하여 극성 사전을 생성한다. 도메인별 평점을 활용할 경우 평점 데이터의 불분명성이 문제가 되고, 어절 유니그램 및 바이그램 이상의 구를 포함한 확률 토픽 모델을 활용한 [5]는 amazone.com의 상품평을 활용한 실험 모델이다.

본 연구는 [3]과 [7]의 방법에 기반하여, 개선된 극성 사전의 자동 구축과 평가 요약을 목표로 한다.

3. 상품 평가 요약 시스템

상품 평가 요약 시스템은 크게 상품평 수집과 정형화, 상품 평가 항목 추출, 극성 사전 생성, 극성 판별의 네 부분으로 구성된다. 대상 상품평은 웹에서 수집되며, [7]의 구어체 보정과 정형화를 수행한다. 다음에서는 본

논문에서 제안하는 평가 항목 추출과 극성 사전 생성 방식을 설명한다. 최종적인 극성 판별은 [7]와 동일한 방식을 사용한다.

3.1 상품 평가 항목 추출

[3]의 경우, 상품 평가에 의미없는 단어의 영향력을 줄이기 위해 웹문서에서의 공기 빈도를 외부 연관도로 사용한다. 이에 대한 결과 분석에서 외부 연관도로 인하여 정확하지 않은 단어가 평가 항목으로 추출되는 경우가 발견되었다. 본 연구에서는 외부 연관도의 순위별 편차를 줄여 평가 항목 추출의 정확도를 향상시키고자 한다. 수식 (1)에서는 [3]의 PMI-RTF의 외부 연관도에 log를 취하여 외부 연관도 값의 편차를 감소시킨다.

$$PMI-RTF(t_i c_{name}) = \frac{f_{review \in c}(t_i)}{MAX(f_{review \in c})} \times \frac{\log(f_{web}(c_{name}, t_i))}{\log(f_{web}(t_i))} \quad (1)$$

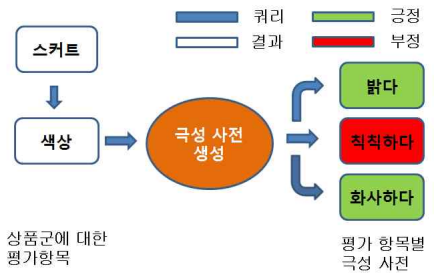
3.2 극성 사전 생성

극성 사전의 생성을 위해서 [7]을 참조하여 수동으로 극성을 분류한 소량의 시드(seed) 집합에 기반한 극성 사전 자동 확장을 수행한다. [7]에서는 긍정 어휘와 부정 어휘에 대해 극성 어휘와의 거리 합의 역수를 사용하여 극성 점수를 계산한다. 이 경우 극성 어휘가 등장할 때마다 $Score(e,p+)$와 $Score(e,p-)$가 점차 감소하여 0에 가까워지는 결과를 얻게 된다. 이로 인하여 사용빈도가 높은 단어가 확장 사전에서 제외되어, 충분한 크기의 확장 사전을 얻기 어렵다. 본 연구에서는 사전 확장의 효율성을 높이기 위해 평가 항목 e와의 거리 값의 평균 $\overline{dist(e,p+)}$와 $\overline{dist(e,p-)}$을 사용한 수식 (2)와 (3)을 제안한다. 수식 (2)와 (3)의 결과 중 임계치(threshold)가 0.2 이상인 결과 어휘에 한해서 긍정 어휘 또는 부정 어휘로 판단한다.

$$Score(p,+) = \frac{f(p+)}{f(p+)+f(p-)} \times \frac{1}{\overline{dist(e,p+)}} \quad (2)$$

$$Score(p,-) = \frac{f(p-)}{f(p+)+f(p-)} \times \frac{1}{\overline{dist(e,p-)}} \quad (3)$$

생성된 극성 사전은 [그림 2]과 같이 평가 항목에 대한 극성을 가진 평가 어휘들의 쌍으로 생성된다. 예를 들어, “색상” 항목에 대해서 “밝다”, “화사하다”는 긍정 어휘, “칙칙하다”는 부정 어휘로 분류된다.



[그림 2] 극성 사전 생성의 입출력

4. 실험 및 평가

시스템 구축에는 형태소 분석기[8], 네이버 웹 문서 검색 Open API를 활용했다. 실험을 위한 데이터는 온라인 쇼핑몰[9]의 데이터를 수집하였으며, 실험 데이터에 대한 상세한 정보는 [표 1]와 같다.

[표 1] 실험 데이터 상세 정보

카테고리 명	상품평 (개)	문장 수 (문장)
전자사전	5,364	13,631
스커트	3,681	6,849
LED TV	3,258	6,416
여성 가방	3,391	7,652
여성 티셔츠	1,751	3,571
자전거	3,148	7,334
Total	20,593	45,453

[표 2]는 평가 항목 추출을 위한 기존의 PMI-RTF[3]와 본 논문의 결과를 비교한다. 결과에서 본 논문의 개선된 PMI-RTF이 높은 정확률을 보이는 동시에, 정답 평가 항목들이 상위 등수에 더 밀집해 있는 것을 확인할 수 있다.

[표 2] ‘전자사전에 대한 평가 항목 상위 20위’

	개선한 PMI-RTF	기존 PMI-RTF	외부 연관도	내부 연관도
1	기능	기능	mp3	배송
2	배송	사전	추천	기능
3	사전	가격	전자	제품
4	가격	mp3	기능	디자인
5	제품	동영상	사전	사전
6	디자인	샤프	딕플	가격
7	생각	추천	배터리	생각
8	상품	전자	동영상	마음
9	사은품	제품	샤프	물건
10	동영상	배터리	아이리버	상품
11	아이	영어	가격	사은품
12	샤프	사용	영어	동영상
13	사용	아이리버	라디오	선물
14	마음	딕플	구매	아이
15	발음	아이	메모리	만족
16	중국어	누리안	사용	발음
17	영어	구매	누리안	중국어
18	선물	하나	하나	사용
19	누리안	메모리	터치	샤프
20	하나	상품	아이	성능

평가 항목 추출의 정확도를 평가하기 위하여, 스커트 상품에 대한 상품평에서 수동으로 평가 항목을 추출하여 정답 평가 항목을 구축하였다. 평가 항목 자동 추출 상위 20위에 대한 평가에서 [3]은 평균 28.7%의 재현율을 보였으나, 본 논문의 방식은 36.7%의 재현율을 보여, 외부 연관도의 순위별 외부 연관도 값의 편차를 줄인 것이 효과적임을 알 수 있다.

개선된 방법으로 추출된 평가 항목을 이용한 극성 판별 결과는 [표 3]와 같다. 평가에서는 상품평 500문장에 대하여 평가 항목별로 극성 정확도를 수동으로 평가하였다. 극성 판별의 정확도는 약 89.9%로, [7]의 결과인 81.8%의 정확도와 비교하여 봤을 때 7.1% 정도 높은 수

치를 확인 할 수 있다.

[표 3] 극성 판별 시스템 평가

상품	평가항목	정확률
위드이폰 (스커트)	배송	93.33
	재질	75.00
	사이즈	82.14
TIVA (LED TV)	배송	81.81
	화질	93.44
	가격	95.74
누리안 (전자사전)	기능	92.68
	가격	96.87
	디자인	98.14
평 균		89.90

5. 결론

본 논문에서는 특징 기반 요약에 특화된 단일 문서인 상품평에 대하여 분석하는 방법을 제안하였다. 제안한 특징 기반 요약 방법을 온라인 쇼핑몰의 상품평 뿐만 아니라 다양한 장르의 단일 문서의 댓글에 적용시킨다면 효과적인 데이터 마이닝 기법으로 활용될 수 있을 것으로 기대된다. 향후 연구로는 비정형 데이터들의 비문법적 표현의 추가적인 보정과 분석된 데이터에 대한 자동 학습이 필요하다.

참고문헌

[1] "k-Structure를 이용한 한국어 상품평 단어 자동 추출 방법",강한훈, 유성준, 한동일,정보과학회논문지: 소프트웨어 및 응용 제 37권 제 6호(2010.6)
[2] “PMI-IR을 이용한 국내 소셜커머스 상품 평가”, 임지연,김이준, 성균관대학교,2011 한국컴퓨터종합학술대회 논문집 Vol.38, No.1(C)
[3] “효율적인 상품평 분석을 위한 어휘 통계 정보 기반 평가 항목 추출 시스템”, 이우철, 이현아, 이공주, 정보처리학회논문지 16-B권 6호 , 2009.12, 497~502p
[4] “상품리뷰요약을 위한 대체어 자동추출”, 안미희, 백종범, 이수원, 한국컴퓨터종합학술대회 논문집 Vol.39, No.1(B). 2012
[5] “상품평에서 속성과 감정을 나타내는 어절 n-gram 추출을 위한 MaxEnt 확률 토픽 모델”, 이영록, 박성배, 이상조, 2012년 가을 학술발표논문집 Vol.39, No.2(B)
[6] “상품평 극성 분류를 위한 특징별 서술어 긍정/부정 사전 자동 구축”, 송종석, 이수원, 정보과학회논문지 소프트웨어 및 응용 제 38권 제 3호(2011.3)
[7] “상품평의 언어적 분석을 통한 상품 평가 요약 시스템”, 이우철, 이현아, 이공주, 정보처리학회논문지 17-B권 1호, 2010.2, 93~98p
[8] 한나눔 한국어 형태소 분석기, JHanNanum-0.8.4-kover, http://kldp.net/projects/hannanum/, 2010/7/31
[9] 비교 쇼핑몰 Basket - http://www.basket.co.kr/