상품평 데이터와 웹 검색엔진을 이용한 상품별 평가항목 자동 추출

이우철 이현아
금오공과대학교 컴퓨터공학부
lee256@naver.com halee@kumoh.ac.kr

Automatic Product Attribute Extraction from Reviews Using Web Search Engine

Woo Chul Lee, Hyun Ah Lee
School of Computer & Software Engineering,
Kumoh National Institute of Technology

요 약

상품평은 인터넷 쇼핑 이용자들의 최종 구매결정에 큰 영향을 미치는 것으로 알려져 있다. 많은 쇼핑몰에서 상품평 활성화를 위해 노력하고 있지만, 상품평을 모으는 것에만 주력할 뿐 기존에 수집된 상품평을 제공하는 방법에 있어서는 원시적인 수준에 그치고 있다. 상품평을 좀 더 효율적으로 제공하려면 사용자들이 상품평에서 찾게 될 평가항목들을 미리 예측하여 그 항목에 따라 상품평을 분류/요약해서 제공하는 방법을 생각할 수 있다. 본 논문에서는 상품평과 웹 검색엔진을 이용하여 각 상품별 평가항목들을 자동으로 추출하는 방법을 제안한다. 상품평 데이터의 특성상 노이즈가 많기 때문에 먼저 데이터를 정제하고, 정제된 상품평 데이터를 형태소 분석하여 후보명사들을 선택한다. 선택된 후보명사를 웹 검색엔진에 질의하여 반환된 결과 값으로 상품 카테고리와 후보명사 간 연관도를 계산하여 평가항목을 추출한다. 실험은 5개 상품 카테고리의 170,294개 실제 상품평을 대상으로 각 카테고리별 평가항목을 추출하였다.

1. 서론

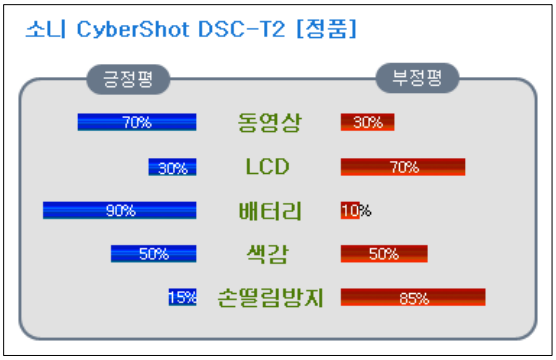
인터넷쇼핑을 이용하는 사람들이 매년 늘어나고 있다. 이들 중 절반 정도인 45.4%가 최종 구매결정 시 다른 사람이 남긴 상품평에 영향을 받는 것으로 나타났다[1]. 상품평이 쇼핑몰의 흥망성쇠를 결정할 정도로 중요해지면서 판매자들이 상품평을 마케팅 수단으로 이용하여 상품평 조작 등의 부작용으로 나타나기도 하지만 이것은 초기의 일부 소형 쇼핑몰에 국한된 문제였다. 현재의 쇼핑몰은 자체규정을 통해 입주한 판매자가 상품평을 임의로 조작하는 것을 방지하고 있다. 이 외의 몇몇 문제점에도 불구하고 사용자 입장에서는 상품평이 실구매자를 통해 얻을 수 있는 가장 믿음직한 정보라는 것에는 변함이 없다.

상품평을 '구입한 상품에 대한 피드백 데이터'로 정의했을 때 크게 세 가지로 구분할 수 있다[2]. 해당 정보에 직접 접근한 횟수를 나타내는 단순 클릭수, 제품의 품질이나 디자인, 배송서비스 등을 평점이나 별점 등으로 수치화한 폐쇄형 상품평, 직접 사용해본 소감이나 평가를 텍스트 형태로 남긴 개방형 상품평의 세 가지 형태이다. 이중 단순 클릭수나 폐쇄형 상품평은 통계적 수치로서 표현하기 좋은 장점은 있지만 자세한 정보를 제공하는 것이 불가능하므로 개방형 상품평에 비해 영향력이 떨어진다고 볼 수 있다. 따라서 좀 더 구체적인 정보를 가진 개방형 상품평에 주목할 필요가 있다. 본 논문에서 언급하는 상품평도 개방형 상품평을 의미한다.

많은 쇼핑몰에서 상품평의 활성화를 위해 상품평 작성자에게 포인트나 쿠폰을 제공하는 등의 노력을 기울이고 있지만 이런 노력들은 모두 상품평을 더 많이 모으기 위한 노력일 뿐 존재하는 상품평들을 보다 효과적으로 제공할 수 있는 방법에 대한 노력은 많이 미흡한 것으로 보인다. 기존 대다수의 쇼핑몰들은 <그림 1>과 같이 등록된 시간 순으로 나열하는 단순한 방식으로 상품평을 제공하고 있다. 사용자가 원하는 정보를 찾기 힘들고 기존 구매자들의 전반적인 분위기를 파악하려면 상품평 전체를 다 읽어야 하기 때문에 상품평을 읽는데 많은 시간을 소모할 수밖에 없다. 본 논문에서는 이런 비효율적인 접근성을 개선하고자 <그림 2>와 같은 새로운 형태의 상품평 분류/요약 시스템을 제안한다. <그림 2>와 같이 상품평을 요약하여 제공하려면 먼저 적절한 분류항목을 마련해야 한다. 분류 항목은 각 상품별 평가항목을 사용한다. 각 상품별로 평가해야하는 항목이 다르기 때문에 상품평 데이터 내에서 평가항목을 추출하고, 추출된 평가항목을 기준으로 상품평을 분류한다. 분류된 상품평들을 극성 판별하여 백분율 막대로 표시하고, 각 극성 막대를 눌렀을 때 실제 긍정 또는 부정으로 판별된 상품평만 따로 보여준다면, 일일이 상품평을 뒤적일 필요 없이 원하는 정보에 쉽게 접근할 수 있다. 뿐만 아니라 상품평을 하나하나 다 읽어보지 않

<그림 1> 기존의 상품평 제공 방식

아도 주요 관심 요소별 평가 극성을 한눈에 파악할 수 있고, 상품에 대한 구매자들의 전반적인 성향도 빠르게 알 수 있으므로 온라인 쇼핑 시 상품평을 읽는데 소모하는 시간을 상당부분 절약할 수 있을 것이다.



<그림 2> 제안하는 상품평 제공 방식

본 논문에서는 <그림 2>와 같이 상품평을 요약하여 제공하기 위한 '상품별 특수 평가항목 추출' 부분에 대해서 다룬다. 특수 평가 항목이란 모든 상품에 적용되는 품질, 디자인, 배송, 가격 등의 일반적인 평가항목을 제외한 특정 상품에만 해당되는 평가항목을 말한다. 디지털 카메라를 예로 들면 손떨림, 액정(LCD), 색감, 배터리, 삼각대, 접사 등의 항목을 꼽을 수 있다.

특수 평가 항목을 추출하기 위해 우선 상품평 내에서 명사들을 추출하고, 추출된 명사를 웹 검색엔진에 질의하여 반환된 결과 값을 바탕으로 현재 상품과의 연관도를 계산한다. 계산된 연관도와 상품평 내 출현빈도를 기반으로 점수화하고 사용자가 설정한 임계치에 따라 평가 항목들을 제공한다.

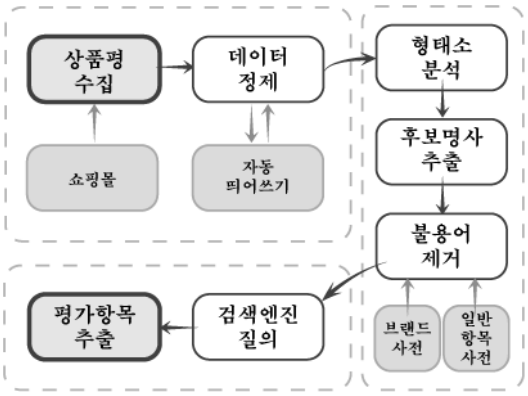
2. 기존연구

상품평 분류에 대한 기존 연구는 대부분 상품평에 해당되는 문장을 대량의 문서 내에서 찾아내거나 각 상품에 대한 극성을 판별하는 것에 치중하고 있다[3,4,5]. 국내의 연구 중 극성 판별을 위해 상품의 특징을 추출하는 방법이 있었으나, 상품평에서 추출된 어휘들을 고빈도순으로 정렬하여 관리자에게 추천하는 반자동방식을 채택하고 있다[6]. 이러한 방법이 정확도 면에서 유리한 장점이 있지만 특징 추출의 객관성이 결여되고 사용자가 일일이 확인해야하는 번거로움이 있다.

상품평의 집합을 하나의 문서로 본다면, 상품평에서 평가항목을 추출하는 것은 기존의 전문용어 자동 추출 연구와 유사점이 있다. 전문용어 추출에서 통계에 기반한 방법에서는 문서 내 용어들의 출현빈도를 이용하는데, 이 때 문서 내 빈도만을 이용하는 경우 빈도가 낮은 용어를 추출하기 어려운 점이 있고 연관도가 낮은 단어가 추출되는 경우도 많다[7]. 언어학적 정보를 이용하는 방법으로는 빈도와 내포관계에 기반한 전문용어 추출기법[8], 사전계층관계에 기반하여 분야 간 유사도와 통계기법을 이용한 전문용어 자동 추출기법[9] 등이 있다. 하지만, 이들은 단일 어절의 전문용어를 추출하지 못하거나 기 구축된 방대한 양의 전문용어 사전이 필요한 제약이 있다.

3. 검색엔진을 이용한 상품 평가항목 자동 추출

본 논문에서 제안하는 평가항목 자동 추출의 전체 과정은 <그림 3>과 같이 크게 세 단계로 나눌 수 있다. 첫 번째 단계는 쇼핑몰에서 상품평 데이터를 수집/저장하고 데이터를 정제하는 과정으로, 반복패턴을 제거하고 띄어쓰기 보정을 한다. 두 번째 과정에서는 형태소 분석기를 이용하여 후보명사를 추출한다. 평가항목이 될 수 없는 브랜드명과 모든 상품에 적용되는 일반평가항목들을 후보명사에서 제외한다. 세 번째 과정에서는 후보명사들을 웹 검색엔진에 질의하여 카테고리와의 연관도를 계산하고 상품평 내 출현빈도와 조합하여 최종 점수를 산출한다. 본 논문에서는 후보 명사의 단일 검색 결과와 후보명사와 카테고리명의 복합 검색의 두 결과 값을 조합하여 연관도가 높은 평가 항목을 추출하고자 한다.



<그림 3> 평가항목 자동추출 구조 및 흐름도

3.1 데이터 정제

상품평의 특성상 신문기사나 책과 다르게 문장이 단정하지 못하다. 맞춤법, 문법, 띄어쓰기의 오류에서부터 불필

요한 이모티콘이나 특수문자, 오타, 동일문장(문자)의 여러 번 붙여넣기 등의 노이즈들이 무수히 존재한다. 특히 동일문자 붙여넣기는 상품평 활성화를 위해 상품평에 등록할 수 있는 최소 글자 수를 제한함으로써 생성되는 노이즈이다. 동일문자가 반복적으로 나타날 경우 올바른 빈도 측정을 방해하기 때문에 반드시 제거해 주어야 한다.

상품평에서 반복문장을 판단하기 위해 비교적 간단한 방법을 사용하였다. 문장 중간부분에서 특정 크기의 텍스트를 추출하여 패턴으로 설정하고 해당 패턴이 문장 내에서 얼마나 반복되는지를 카운트하여 반복문장을 판별한다. 반복문장으로 판단된 문장은 상품평 데이터에서 제외된다. 아래 예문에서는 '감사합니다.'가 8회 반복되고 있으며 추출된 패턴과 7회 매칭된다.

예문) 직원분들도친절하고 좋습니다. 감사합니다.감사합니다.감사합니다.감사합니다.감사합니다.감사합니다.감사합니다.감사합니다.
pattern : [다.감사합니] 7회 반복됨. -> 제거

띄어쓰기가 제대로 되어 있지 않으면 형태소 분석과정에서 올바른 결과를 얻을 수 없다. 예를 들어 '정말 마음에들어요'라는 문장과 같이 띄어쓰기가 바르지 않은 상태일 경우 형태소 분석기는 '마음에들어' 까지를 하나의 명사로 태깅하여 평가항목 후보로 추출한다. 이런 오류를 막기 위해 띄어쓰기 보정이 필요하다.

상품평 데이터에서 문장 당 공백비의 평균을 활용하여 띄어쓰기가 제대로 되지 않은 문장을 판별하였다. 상품평의 공백비 측정 결과 평균 20.89%, 뉴스 데이터는 평균 25.99%의 공백비를 보이는 것을 알 수 있었다. 엄격한 띄어쓰기를 적용하는 뉴스 데이터에는 못 미치지만 상품평 데이터도 어느 정도 띄어쓰기가 이루어지고 있다고 보고, 상품평의 문장 당 평균글자수인 22글자를 기준으로 공백비 20%선에서 공백불량문장을 판별하였다. 글자 수가 평균글자수와 같거나 많을 경우 20% 고정 적용하고, 못 미치는 문장의 경우엔 기준 글자 수에서 1글자씩 줄어들 때마다 공백비를 0.5%씩 감소시켜 적용했다. 6글자 이하로 이루어진 문장은 공백을 포함하지 않아도 비문처리 하지 않는다. 이렇게 비문으로 판단된 문장은 네이버랩의 자동띄어쓰기[10]를 이용하여 띄어쓰기 처리하였다.

3.2 형태소 분석과 후보 명사 추출

정제된 상품평을 형태소 분석기[11]를 통하여 형태소 분석하였다. 형태소 분석 결과 파일에서 명사만을 추출하여 후보 명사 집합으로 분류하였으며, 상품평가항목이 될 수 없는 의존명사와 이모티콘, 특수기호 등을 제거하고 해당 상품 카테고리의 브랜드명 역시 브랜드 사전을 이용하여 제거했다. 브랜드 사전은 쇼핑몰[12]의 조건검색 항목에서 '제조사' 부분에 등록된 어휘를 자동 추출하여 생성하였다. 그리고 카테고리에 관계없이 광범위하게 적용되는 일반적인 평가항목들은 실험에서 카테고리와의 연관도 계산 결과가 낮아 평가항목으로 추출되지 못하는 문제점이 있어 후보 추출과정에서 사전을 이용하여 모두 제거한 뒤 추출과정이 모두 끝난 후에 일괄 추가한다.

3.3 웹 검색결과에 의한 연관도 계산

위 과정에서 추출된 후보명사 집합에서 상품평가항목에 적합한 명사만 추출하기 위해 상품카테고리명과 후보명사 간 연관도를 계산한다. 예를 들어 '노트북' 카테고리의 상품평에서 '생각', '맘', '감사', '사양', '발열', '불량화소'가 후보명사로 추출되었다면 이들 중 노트북의 평가항목으로 적합한 것은 '사양', '발열', '불량화소'이다.

위 예시항목을 상품평 내 빈도수로 내림차순 정렬하면 <표 1>과 같다. '생각'이나 '감사'와 같은 카테고리 연관성이 떨어지는 어휘가 상위에 분포하고 있어, 기존 연구에서 사용하고 있는 단어 빈도를 이용한 방식은 평가 어휘 추출에 적합하지 않음을 알 수 있다. 본 논문에서는 후보 명사와 대상 카테고리명이 동시에 나타나는 문서의 빈도가 높을수록 두 단어의 연관성이 높다는 점에 착안하여 연관도가 높은 평가 항목을 추출하고자 한다.

추출된 후보명사를 t라 하고, 카테고리명을 c라고 하자. 카테고리 c의 리뷰에서 추출된 전체 후보명사들이 가지는 빈도값 중 가장 높은 빈도값을 $rf(T)$라고 하고, 각 단어 t가 상품평에 나타난 빈도(review frequeny)를 $rf(t)$라 하자. 웹 검색엔진에서 얻어지는 단어 t의 단일검색결과(single word frequency)를 $sf(t)$라 하고 카테고리명과 후보 명사를 AND연산으로 함께 검색한 복합검색결과(composite word frequency)를 $cf(c,t)$라 하자.

추출된 후보명사 t와 카테고리명 c를 검색엔진에 질의하여 $sf(t)$와 $cf(c,t)$를 구한다. $cf(c,t)$와 $sf(t)$의 값이 근접할수록 두 어휘 간 연관도가 큰 것이므로 후보명사 t는 함께 검색한 카테고리에서만 사용되는 특수 평가항목일 확률이 높다.

<표 1> 후보 명사 연관도 측정(빈도 내림차순)

* $rf(T)$ = 7,309

후보명사	빈도	단일검색	복합검색	연관도	적합성점수
생각	2,863	81,326,859	1,157,656	0.014	④ 0.558
맘	1,922	11,885,152	195,255	0.016	⑤ 0.432
감사	1,760	73,068,521	919,693	0.013	⑥ 0.303
사양	1,153	7,330,223	717,030	0.098	③ 1.543
발열	985	709,804	100,948	0.142	② 1.917
불량화소	380	1,115,299	952,762	0.854	① 4.441

질의어 예시) t = 발열, c = 노트북
- 단일 검색 $sf(t)$ 시 : ["발열"]
- 복합 검색 $cf(c,t)$ 시 : ["노트북" "발열"] (AND 연산)

단어 t의 평가항목 적합성 점수는 $rf(t)$를 최고 빈도값

인 *rf(T)*로 나누어 정규화한 수치와, *sf(t)*에서 *cf(c,t)*가 차지하는 비율 값인 연관도를 곱하여 계산한다.

$$score(t) = \frac{rf(t)}{rf(T)} \times \frac{cf(c,t)}{sf(t)}$$

위 수식을 적용하여 얻어진 적합성 점수를 기준으로 내림차순 정렬하면 순위는 '불량화소', '발열', '사양', '생각', '맘', '감사'가 된다(표1참조). 빈도순 정렬일 때 상위에 있던 연관성이 적은 단어들이 순위가 떨어져 연관성이 높은 순서대로 정렬된 것을 볼 수 있다.

실험 결과에서 점수별로 정렬했을 때 상품의 브랜드명이 상위권에 다수 분포되어 있는 것을 발견할 수 있었다. 브랜드명은 해당 카테고리에 가장 연관도가 높은 어휘라고 볼 수 있으며, 이런 결과는 위의 방법이 특정분야의 키워드를 추출하는데 효과적이라는 것을 증명해준다. 그러나 상품 평가항목으로는 적절치 않기 때문에 후보명사 추출과정에서 브랜드 사전을 구축하여 필터링했다. 카테고리에 관계없이 광범위하게 적용되는 일반적인 평가항목들 역시 후보명사 추출과정에서 제외하고 처리한 뒤 모든 평가항목 추출이 완료되면 일괄적으로 추가해 준다.

4. 실험

상품평 데이터는 현재 온라인상에서 운영 중인 가격비교 사이트 베스트바이어[12]에서 수집하였다. 노트북, 디지털카메라, 스커트, 인라인, 남성용화장품의 다섯 개 카테고리에 대한 상품평을 수집하여 특수 평가항목 추출 실험을 수행하였다. 웹 검색엔진으로는 네이버[13]의 웹문서 검색을 사용하였다. 본 추출법에 대한 실험결과는 <표 2>에 요약하였다.

<표 2> 실험결과

항 목	화장품	디카	노트북	인라인	스커트
상품평 수	85,989개	63,681개	19,023개	11,842개	805개
빈도 기반기법	15.62%	13.95%	14.58%	11.76%	12.12%
연관도 기반기법	75.00%	62.79%	52.08%	58.82%	61.61%

정확률은 시스템 처리 결과 중 상위 50위 범위 내에 정답 평가 항목이 포함된 비율로 계산하였다. 정답 평가 항목은 다섯 개의 카테고리에 대하여 적절한 평가 항목을 수작업으로 선별하여 사용하였다. 범위를 상위 50위로 제한한 것은 상품에 대한 평가항목을 요약하여 사용자에게 보여줄 때 한눈에 파악할 수 있는 적정 개수이기 때문이다. <표 2>에서 기존의 빈도 기반 기법과 본 논문에서 제안한 연관도 기반 기법의 정확률을 비교하였다. 실험 결과, 본 논문에서 제안한 기법이 평균 62.06%의 정확률을 보여 단순 출현빈도 기반 기법의 평균 정확률 13.6%보다 우월한 성능을 보였다. 결과에서 볼 수 있듯이, "후보 어휘와 카테고리 어휘의 복합 검색"에 기반한 본 논문의 방식이 웹에서 손쉽게 얻은 정보를 이용한 간단한 방법임에도 불구하고 높은 성능을 보임을 알 수 있다.

5. 결론 및 향후 연구

본 논문에서는 실제 상품평에서 상품별 평가항목을 추출하는데 웹 검색엔진을 이용하는 새로운 방법을 제안하였다. 상품평에서 후보어휘를 추출하고 이를 웹 검색엔진에 검색한 후, 반환된 페이지 수를 이용하여 상품 카테고리와 평가항목간의 연관도를 계산한다. 제안한 방법의 성능 분석을 위해 수동으로 작성된 정답과 비교하여 평균 62.06%의 정확률을 보였으며, 기존의 단순 빈도수 기반의 용어 추출 기법과 비교한 결과 월등히 우수한 결과를 보여주었다.

본 논문을 바탕으로 향후 각 평가항목에 따른 상품평 분류와 극성 판별에 관한 연구를 통하여 서론에서 제안한 상품평 분석/요약시스템을 구현할 예정이다. 현재 이에 대한 연구를 수행 하고 있다.

참고문헌

[1] "2007년 상반기 정보화실태조사 요약보고서", 한국인터넷진흥원, 2007. 8

[2] 김승훈, 강희택, "온라인 피드백 메커니즘으로서 상품평 게시판의 지각된 효과성과 신뢰간의 관계 구조 분석" 한국경영과학회지, 2007

[3] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques", EMNLP. 2002.

[4] Kushal Dave, Steve Lawrence, David M. Pennock, "Mining the peanut gallery : opinion extraction and semantic classification of product reviews", WWW 2003: 2003.

[5] E. Breck, Y, Choi and C. Cardie, "Identifying expressions of opinion in context", IJCAI, 2007.

[6] 명재석, 이동주, 이상구, "반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템" 한글 및 한국어 정보처리 학술대회 발표논문집 제 19권, 2007

[7] Dagan, I. and K. Church, "Termight: Identifying and translating technical terminology", EACL95, 1995.

[8] Frantzi, K.T. and S.Ananiadou, "The C-value/NC-value domain independent method for multi-word term extraction", Journal of Natural Language Processing, Vol. 6, No.3, 1999.9

[9] 오종훈, 이경순, 최기선, "분야간 유사도와 통계기법을 이용한 전문용어의 자동 추출" 정보과학회논문지:소프트웨어 및 응용 제 29권 제 4호, 2004. 4

[10] http://lab.naver.com/autospacing

[11] 강승식, "한국어 형태소 분석기와 한국어 분석 모듈", 국민대학교 자연언어 정보검색연구실, http://nlp.kookmin.ac.kr.

[12] http://www.bb.co.kr

[13] http://www.naver.com