DOI: 10.3745/KIPSTB.2009.16B.6.497

효율적인 상품평 분석을 위한 어휘 통계 정보 기반 평가 항목 추출 시스템

이 우 철[†] · 이 현 아[††] · 이 공 주[†††]

요 약

본 논문에서는 상품평의 효율적인 분석을 위한 평가 항목 추출 시스템을 제안한다. 시스템은 크게 상품평 수집-보정과 평가 항목 추출의 두 단계로 구성된다. 상품평 수집-보정에서는 인터넷 쇼핑몰에서 상품평을 수집하고 상품평 특유의 구어체 표현과 맞춤법 오류 등을 처리한다. 평가 항목 추출에서는 스커트 상품 카테고리의 경우 '사이즈', '스타일'과 같이 상품을 평가하는 기준이 되는 항목을 상품평과 인터넷 상의 웹 문서를 활용하여 자동으로 추출한다. 상품평에 나타나는 명사들을 평가 항목 후보로 설정하고, 각 후보 명사의 상품평에서의 어휘 통계인 내부연관도와, 후보 명사와 상품 카테고리명의 웹 문서에서의 공기 빈도에 기반하여 계산된 외부연관도를 결합하여 상품과 평가 항목 후보의 연관도를 계산한다. 본 논문의 평가 항목 추출 방식은 평균 재현율 90%를 보여 기존 연구보다 우수한 결과를 보였다.

키워드 : 상품평, 평가 항목 추출, 특징 기반 요약, 어휘 통계, 전자상거래

Automatic Product Feature Extraction for Efficient Analysis of Product Reviews Using Term Statistics

Woo Chul Lee[†] · Hyun Ah Lee[††] · Kong Joo Lee[†††]

ABSTRACT

In this paper, we introduce an automatic product feature extracting system that improves the efficiency of product review analysis. Our system consists of 2 parts: a review collection and correction part and a product feature extraction part. The former part collects reviews from internet shopping malls and revises spoken style or ungrammatical sentences. In the latter part, product features that mean items that can be used as evaluation criteria like 'size' and 'style' for a skirt are automatically extracted by utilizing term statistics in reviews and web documents on the Internet. We choose nouns in reviews as candidates for product features, and calculate degree of association between candidate nouns and products by combining inner association degree and outer association degree. Inner association degree is calculated from noun frequency in reviews and outer association degree is calculated from co-occurrence frequency of a candidate noun and a product name in web documents. In evaluation results, our extraction method showed an average recall of 90%, which is better than the results of previous approaches.

Keywords : Product Review, Product Feature Extraction, Feature Based Summarization, Term Statistics, Electronic Commerce

1. 서 론

우수한 정보 인프라에 힘입어 우리나라의 인터넷 쇼핑 시장은 비약적인 발전 추세를 보이고 있다. 상품을 직접 살펴볼 수 없는 인터넷 쇼핑에서 기존 구매자의 상품평은 구매 예정자의 상품 구매 결정에 영향을 미치는 중요 정보원이다. 웹 2.0으로 대표되는 사용자 참여 중심의 인터넷 환경, 집단 지성(Collective Intelligence)의 발현 등으로 소비자가 능동적으로 컨텐츠를 생산하고 공유하고 지속적으로 가치를 부여하면서 상품평의 영향력은 더욱 광범위해지고 강력해지고 있다. 국내의 경우 상품평 및 이용 후기 이용자의 94.3%는 다른 이용자의 구매 경험과 평가를 기반으로 최종 구매 여부를 결정하고 있으며, 59%는 쇼핑 후 상품평 및 이용 후기 등을 작성함으로써 온라인 구전의 형성과 확산에 참여하고 있다[1].

인터넷 쇼핑 시장의 발전과 이용자의 증가에 따라 상품평의 개수도 지속적으로 증가하고 있다. 근래에는 (그림 1)과

※ 본 연구는 금오공과대학교 학술연구비에 의하여 연구된 논문입니다.
† 정 회 원 : (주) 유승토탈솔류션 대리
†† 종신회원 : 금오공과대학교 컴퓨터공학부 조교수(교신저자)
††† 정 회 원 : 충남대학교 전기정보통신공학부 부교수
논문접수 : 2009년 8월 18일
수 정 일 : 1차 2009년 9월 29일
심사완료 : 2009년 10월 7일

(그림 1) 기존의 상품평 제공방식과 별점 평점 오류의 예

같이 하나의 상품에 만 개 이상의 상품평이 등록되는 경우도 많아져, 기존 구매자의 평가를 분석하기 위해 상품평을 하나씩 살펴보는 것이 거의 불가능해졌다. 상품평 활용의 어려움을 해소하기 위해 대부분의 쇼핑몰에서 상품에 대한 구매자의 평가를 요약하는 별점 평점을 제공하고 있으며, 일부의 경우 별점 평점의 오름차순/내림차순으로 상품평을 조회하는 기능을 제공하고 있다. 하지만, 별점 평점은 배송이나 쇼핑몰 서비스와 같이 상품 외적인 부분에 대한 평가를 포함하는 경우가 많으며, (그림 1)과 같이 상품평에 맞지 않은 별점 평점을 부여한 경우도 있어, 구매 예정자에게 상품에 대한 적절한 정보를 제공하지 못하는 경우가 대부분이다.

오피니언 마이닝(Opinion Mining) 또는 감성 분류(Sentiment Classification) 연구 분야에서는 다양한 텍스트를 분석하여 의견을 추출하고 추출된 의견에 긍정 또는 부정 극성을 부여하여 구조화하는 연구를 수행한다. 근래의 급격한 인터넷 쇼핑 시장의 성장과 앞서 지적한 별점 평점의 문제로, 감성 분류를 적용하여 부정과 긍정의 사용자 평가를 상품평에서 직접 추출하려는 접근이 늘어나고 있다. 이 중 특징 기반 요약(Feature Based Summarization) 연구에서는 상품 특징을 추출하고 그 특징을 포함하는 문장을 식별하고자 한다.

상품을 직접 만져볼 수 없는 인터넷 쇼핑에서 사용자는 상품에 대한 총괄적인 평가와 함께 각 개인이 중점을 두는 상품의 세부 특성에 대해 분류된 평가를 필요로 한다. 예를 들어 디지털카메라의 경우 카메라의 화질을 구매 결정의 요인으로 삼는 구매자가 있는 반면, 카메라의 크기나 디자인을 중시하는 구매자도 있다. 상품에 따라 사이즈 기준이 각기 다른 스커트 등의 의류를 구매하는 경우, 인터넷 쇼핑에서 실패한 경험이 있는 사용자는 사이즈에 대한 기존 구매자의 평가를 조회하고 싶어 한다. 이러한 상품 특성에 따라 분류된 상품평이 제공된다면 구매 예정자는 매우 용이하게 상품평을 분석할 수 있다. 이를 위하여 상품 특징을 자동으로 추출하고자 하는 연구가 영어권에서는 다양하게 진행되었으나, 국내의 경우 인터넷 쇼핑 시장의 크기에 비해 상품평과 특징 추출에 대한 연구는 미진한 실정이다.

본 논문에서는 디지털카메라의 경우 '화질'이나 '크기', 스커트의 경우 '사이즈'나 '스타일'과 같이 상품 구매의 판단 기준이 될 수 있는 상품의 특성을 평가 항목으로 정의하고 한국어 상품평에서의 평가 항목 자동 추출 방법을 제안한다. 제안하는 방식에서는 상품평 내의 명사를 평가 항목 후보로 보고 각 명사의 상품평 내에서의 빈도와 웹 문서에서의 상품 카테고리 이름과 평가 항목 후보 어휘의 공기 빈도의 두 가지 정보를 결합하여 상품과의 연관도가 높은 단어를 평가 어휘로 추출한다. 시스템에서는 인터넷 상에 공개된 대량의 웹 문서에서의 어휘간 공기 빈도를 얻기 위해 인터넷 웹 검색 엔진을 활용한다.

본 논문은 다음과 같이 구성된다. 2장에서 평가 항목 추출에 대한 기존 연구를 살펴보고, 3장에서는 본 연구의 평가 항목 추출 방법에 대해 다룬다. 4장에서는 실험 결과를 살펴보고 5장에서 결론 및 향후 연구에 대해 논의한다.

2. 기존 연구

문서나 문장을 단위로 극성을 판별하여 요약하는 감성 분류에 비해, 특징 기반 요약에서는 특징을 표현하는 어휘나 어구를 추출하고 추출된 특징 어휘를 기준으로 의견을 기술한 문장들을 식별하고 극성 분류를 수행한다. 본 논문의 평가 항목 추출은 특징 기반 요약에서의 특징 추출과 유사하다. 기존의 특징 추출 연구에서는 주로 TF-IDF, 연관마이닝, PMI 등의 통계적 기법을 사용한다.

TF-IDF(Term Frequency - Inverse Document Frequency)를 사용하여 평가 항목을 추출하는 대표적인 연구는 Kim 외[2]의 연구가 있으며, Red Opal[3]은 TF-IDF와 유사한 방법으로 평가 항목을 추출한다. TF-IDF 가중치 기법은 가장 간단하게 사용할 수 있는 방법이지만 정확성이 떨어지는 단점이 있다. Hu, M. 외의 연구[4, 5]에서는 평가 항목 추출에 연관 마이닝(Association Mining)을 사용한다. 연관 마이닝에 의한 평가 항목 추출은 하나 이상의 단어로 이루어진 복합 명사나 명사구 형태의 평가 항목을 추출할 수 있는 장점이 있지만, 결합 규칙에 의해 명사를 결합하는 과정에서 불필요한 평가 항목이 많이 생성되는 단점이 있다.

PMI(Point-wise Mutual Information)는 정보이론과 통계학에서 사용되는 연관성 측정법이다. 용어 x와 y의 확률 $p(x)$, $p(y)$와 x, y의 공기 확률 $p(x,y)$이 주어졌을 때 PMI는 수식 (1)으로 계산된다.

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

PMI를 이용한 상품 특징 추출 연구에서는 신뢰성 있는 실세계의 단어 확률을 얻기 위해 대량의 통계 정보를 제공하는 인터넷 상의 웹 문서를 활용한다. 웹 문서에서의 어휘 빈도는 인터넷 검색을 통해 얻을 수 있으며, 이 경우 PMI 값은 수식 (2)와 같이 계산된다. 수식에서는 단어 t와 t_i를 각각 포함하는 문서를 검색 엔진에 검색한 결과의 수 $\#(t)$, $\#(t_i)$와 두 단어가 근접하여 등장하는 문서를 검색한 결과의 수 $\#hits(t\ NEAR\ t_i)$을 기반으로 PMI값을 계산한다. PMI를 이용하여 특징을 추출한 시스템은 KnowItAll[6]과 Opine[7]이 있다.

$$PMI(t,t_i) = \log \frac{hits(t\ \ NEAR\ \ t_i)}{\#(t)\#(t_i)} \quad (2)$$

한국어 상품평에 대한 연구에서는 특징 기반 요약을 사용하지 않거나, 상품평에 포함된 명사의 빈도순으로 상품 특징을 추천하고 관리자에 의해 수동으로 평가 항목을 생성[8]하는 등, 평가 항목 자동 추출에 대한 연구는 미진한 실정이다.

본 논문에서는, 다양한 분야에서 다양한 의미로 사용되는 상품 특징이라는 용어 대신, 상품평에서 상품을 평가하는 기준으로 사용되는 단어들을 상품 평가 항목으로 표현하고, 상품평에서 상품 평가 항목을 추출하기 위한 방법을 제안한다.

3. 어휘 통계 정보에 기반한 평가 항목 추출

본 장에서는 상품 평가 항목 추출을 위한 방법을 제안한다. 평가 항목 추출은 1) 상품평 수집-보정과 2) 평가 항목 추출의 두 단계로 구성된다. 상품평 수집-보정에서는 온라인상의 상품평을 수집하고, 문장 종단 오류, 띄어쓰기 등의 오류를 보정한 후 형태소를 분석한다. 평가 항목 추출부에서는 형태소 분석과 품사 태깅 결과에서 명사로 태깅된 어휘를 평가 항목 후보로 설정하고, 해당 명사의 빈도 정보와 웹 검색을 통해 산출된 연관도 점수를 이용하여 평가 항목을 추출한다.

3.1 상품평 수집-보정

본 논문에서는 여러 쇼핑몰의 상품평을 자동으로 수집하여 통합 제공하는 가격비교사이트[9]를 상품평 수집 대상 사이트로 사용하여, 여러 쇼핑몰에 산재해 있는 상품별 상품평을 일괄적으로 수집한다. 상품평 수집에서는 상품 코드나 카테고리 코드를 인자로 주어 수집할 수 있다. 수집된 상품평은 상품별로 구조화되어 XML 형태로 시스템에 저장된다. 상품평을 XML형태로 저장하여 사용함으로써 국내의 쇼핑몰에서도 Amazon.com과 같은 open API를 통한 상품평 제공 방식을 채택할 경우에 별도의 수정 없이 편리하게 연결할 수 있다.

일반인들이 별다른 제약없이 자유롭게 기술하는 상품평은 띄어쓰기 오류, 구어체, 오타, 깨진 문자, 불필요한 이모티콘이나 잘못된 문장 부호 등의 다수의 오류를 포함한다. 근래 상품평의 중요성을 인식한 인터넷 쇼핑몰에서 상품평 작성시 다양한 혜택을 제공하여, 작성되는 상품평의 개수가 늘어나는 동시에 오류가 포함된 상품평도 크게 늘고 있다. 이런 오류들은 평가 항목 추출과 그의 전처리 과정인 형태소 분석 등에서 더 큰 오류로 발전할 가능성이 있어 모든 처리 과정에 앞서 최대한 보정한다. 상품평 수집-보정의 처리 과정은 아래와 같다.

1) 수집 중 보정

상품평 수집 과정에서 HTML태그를 제거하고 이스케이프된 특수 문자와 인코딩 변환 과정에서 깨진 문자들을 원래 문자로 치환한다. 치환 불가능한 문자는 공백 처리한다.

원문: <td colspan="4" style="overflow: hidden;">가격에 비해 너무 뷁스러운 것 같아요 gt;_ lt;</td>
보정: 가격에 비해 너무 뷁스러운 것 같아요 >_<

2) 중복 표현 제거

충실한 상품평을 얻기 위한 정책으로 상품평 작성에서 최소 글자수를 제한하는 쇼핑몰이 늘어나면서 상품명 등의 동일한 문구를 반복하여 포함하는 상품평이 늘고 있다. 의미없는 반복 문구를 제거하기 위해 상품평에 동일한 문자열이 3번 이상 중복되면 이를 제거한다.

3) 문장 종단 보정

한국어 상품평은 문장 종단 기호를 잘못 사용하거나 아예 사용하지 않은 문장들이 많아 문장 구분이 힘들다. 마침표 없이 이모티콘(예: '^^', 또는 'ㅋㅋ') 형태로 종료되는 문장이 특히 많고, 문장 부호를 두 번 이상 반복 사용하는 경우(예: ,,,, 또는 ..., !!!, ??)도 흔하다. 시스템에서는 문장 종결로 사용가능한 어구 뒤에 이모티콘 리스트에 포함된 기호열이나 중복된 문장 부호가 등장하는 경우, 기호열을 마침표로 대체하거나 문장 부호 중복을 제거하여 문장 종단을 보정한다.

원문: 색상이...좀 그렇네요ㅎㅎ디자인은,,,,이쁘구요^0^ㅋㅋ많이파세요!!!!
보정: 색상이, 좀 그렇네요. 디자인은, 이쁘구요. 많이파세요!

4) 구어체 표현 보정

일반인들이 작성한 상품평에는 '하네요'에서 어미 '네요'를 '네욤', '네용', '네염', '네여' 등으로 표기한 구어체 표현들이 자주 발생하며 이는 형태소 분석에서 잘못된 결과를 내기 쉽다. 문장 기호 바로 앞에 나타나면서 미등록 명사로 태깅된 단어가 '여', '염', '영', '용', '욤'으로 종료되는 경우 이를 '요'로 치환한다.

원문: 화면으로 봤을 땐 이뻤는데염...받아보니 색깔이 넘 어둡네용-0-;;
보정: 화면으로 봤을 땐 예뻤는데요. 받아보니 색깔이 넘 어둡네요.

5) 띄어쓰기 보정 및 형태소 분석과 품사 태깅

보정 과정이 끝난 상품평은 자동띄어쓰기 모듈을 이용하여 띄어쓰기를 보정한다. 띄어쓰기 보정이 완료되면 문장별로 분리하여 형태소 분석과 품사 태깅을 수행한다.

3.2 상품 평가 항목 추출

상품평의 수집과 보정 과정이 끝나면 해당 데이터에서 상품 평가 항목을 추출한다. 상품을 평가할 수 있는 평가 항목은 대부분 명사이므로, 평가 항목 추출에서는 각 카테고리 단위로 분류된 상품평에서 명사로 태깅된 형태소들을 평가 항목 후보로 이용한다. 평가 항목에는 '가격'이나 '배송', '크기'와 같이 어느 상품에나 적용 가능한 평가 항목과 디지털카메라의 '화질', 스커트의 '스타일'과 같이 각 상품의 특징을 표현하는 상품 의존적인 평가 항목이 있다. 사용자 입장에서는 상품 의존적인 평가 항목과 상품 비의존적인 평가 항목에 대한 구분 없이 구매하고자 하는 상품의 다양한 평가 항목에 대한 세부 평가를 기대한다. 각 상품에 맞는 평가 항목 집합을 제공하기 위해 본 연구에서는 상품 카테고리별로 평가 항목을 추출한다. 이를 그림으로 나타내면 (그림 2)와 같다. 평가 항목 추출의 대상을 '니콘P90'과 같은 세부 상품이 아닌 '컴팩트 디지털카메라'과 같은 상품 카테고리로 지정하여, 각 상품의 특성에 맞는 평가 항목을 추출하되 가능한 많은 상품평을 활용하여 적합한 평가 항목을 얻

게 한다. 평가 항목 추출에서는 각 상품 카테고리의 상품평을 독립적으로 사용하여 상품 의존적인 평가 항목과 상품 비의존적인 평가 항목에 대한 구분 없이 동일한 방법으로 추출한다.

평가 항목은 해당 상품의 구매 결정에 영향을 미치는 특징을 의미한다. 평가 항목은 상품평에서 자주 발생하는 단어일 가능성이 높지만 상품평에서의 빈도만을 고려하면 상품평에 자주 발생하는 표현 - 예를 들어 '마음에 들다', '선물로 보내다' 등의 '마음', '선물' - 때문에 적절하지 않은 단어가 추출될 가능성이 높다. 이러한 단어를 제외시키기 위해서 TF-IDF 방식을 적용할 수 있지만, 상품평 내에서 역문서빈도를 적용하면 대부분의 상품평에 포함되는 '가격'이나 '품질'과 같은 중요 평가 항목까지 제외되는 문제가 발생한다. 본 논문에서는 상품평 내에서의 명사의 빈도와 웹 문서에서의 PMI를 결합하여 평가 항목을 추출하고자 한다.

PMI 방식에서는 상품 특징 명사는 상품 카테고리 이름을 나타내는 단어 즉 상품 카테고리명과 함께 등장할 확률이 높다고 가정하고, 대량의 문서에서 상품 카테고리명과 특징 명사가 공기하여 발생하는 빈도를 특징 명사의 빈도로 나눈 값으로 특징 명사의 점수를 구한다[6, 7]. 기존 연구에서는 대량의 웹 문서에 대한 통계를 추출할 수 있는 인터넷 검색을 통하여 단어 빈도를 획득하며, 특징 명사의 역빈도가 역문서빈도의 역할을 하게 되어 상품과 연관도가 높은 명사를 결과로 추출할 수 있다. 본 논문에서는 상품평 내에서의 특징 명사, 즉 평가 항목의 빈도와 PMI를 곱한 변형된 방법을 이용한다. 기존의 PMI 방법은 인터넷 검색을 통한 웹 문서에서의 단어 간 연관도에 의존하며 평가 항목 추출의 중요 정보원인 상품평의 특성을 반영하지 못 한다. 본 논문에서는 웹 문서에서 얻은 상품 카테고리명과 평가 항목 후보의 연관도 값인 PMI를 외부연관도로 정의하고, 상품평 내에서의 평가 항목 후보의 빈도를 내부연관도로 정의하고, 이를 결합하여 평가 항목 추출의 정확성을 높이고자 한다.

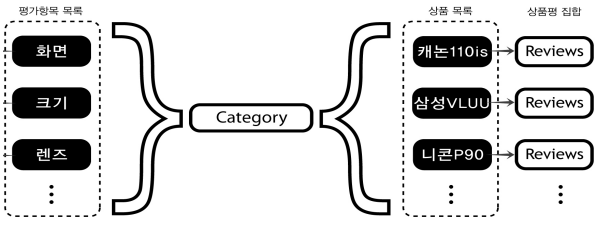
웹 문서에서의 통계 정보를 얻기 위해 인터넷 검색을 활용하여 외부연관도를 계산하는 과정에서, 각 카테고리별 상품평에서 명사로 태깅된 모든 형태소들을 평가 항목 후보로 사용하면 심각한 부하가 발생하다. 본 연구에서는 부하를 최대한 줄이기 위해 검색엔진 질의 과정 전에 아래와 같은 필터링 규칙을 적용한다.

1) 상품평을 문장 단위로 구분하여 형태소 분석과 품사 태깅한 후, '좋다' '나쁘다'와 같은 평가 표현에 이용되는 형용사나 동사가 포함된 문장에서만 후보 명사를 뽑는다. 의견을 나타내는 단어가 포함되지 않은 문장에 등장하는 명사는 평가 항목이 아닐 가능성이 높기 때문이다.
2) '추천하다'의 '추천'이나 '생각하다'의 '생각'처럼 '하다'와 결합하여 사용된 명사는 평가 항목 후보가 될 수 없으므로 추출하지 않는다. 단, "추천이 많아 구입했는데 별로네요."의 '추천'과 같이 동사로 사용되지 않는 경우는 추출한다.
3) 태깅된 품사를 이용하여 '것', '바', '소', '자루', '마리'와 같은 평가 항목으로 부적절한 각종 의존명사를 필터링한다.
4) 띄어쓰기 에러 등으로 인한 형태소 분석 오류로 잘못 생성된 명사들을 제거하기 위해, 후보 명사의 전체 출현 빈도 대비 최소 지지도(minimum support)가 0.1% 미만인 후보 명사들을 필터링한다.
5) 제품 제조사명, 브랜드명, 판매 쇼핑몰명 등은 상품과의 연관도는 높지만 상품 평가 항목은 될 수 없다. 이를 제거하기 위해 브랜드 사전을 사용하여 필터링한다. 브랜드 사전은 인터넷 쇼핑몰의 제품 분류 레이블을 추출하여 자동 생성한다.

위의 규칙에 의해 필터링된 후보 명사 집합에서 현재 카테고리에 적합한 평가 항목을 추출한다. 변형된 PMI기법을 적용하여 카테고리 c의 전체 상품평($review \in c$)에서 후보 명사 t_i의 평가 후보 점수를 계산한다. t_i의 상품평 내 출현 빈도 $f_{review \in c}(t_i)$를 상품평에서 등장한 후보 명사 빈도의 최고값 $MAX(f_{review \in c})$으로 나누어 0과 1사이의 값을 가지는 내부연관도를 산출한다. 인터넷 검색 엔진에 후보 명사 t_i를 단일 검색하여 얻어진 결과 개수 $f_{web}(t_i)$를 상품 카테고리명 c_{name}과 후보 명사 t_i를 결합하여 검색한 결과 $f_{web}(c_{name}, t_i)$로 나누어 PMI 값, 즉 외부연관도를 산출한다. 내부연관도와 외부연관도 두 값을 곱하여 얻어지는 값을 PMI-RTF(Point-wise Mutual Information - Review Term Frequency)로 정의하고 이를 적합성 점수로 보고 순위를 매긴다. 해당 수식은 아래의 수식 (3)과 같다.

$$PMI-RTF(t_i, c_{name}) = \frac{f_{review \in c}(t_i)}{MAX(f_{review \in c})} \times \frac{f_{web}(c_{name}, t_i)}{f_{web}(t_i)} \quad (3)$$

<표 1>은 PMI-RTF 수식을 통해 평가 항목 추출 과정을 처리하는 예이다. 표는 스커트 상품 카테고리에 대한 평가 항목 후보 명사 집합의 일부를 나타낸다. 표의 수치는 네이버의 웹문서 검색을 사용하여 구하였으며, 후보 명사의 출현 빈도는 필터링을 거친 후의 값이다. 표에서 사용된 c_{name}은 '스커트', $MAX(f_{review \in c})$는 '치마'의 상품평 내 빈도인 405이다. 내부연관도를 기준으로 정렬하면 스커트 상품 카테고리에 대한 평가 항목이 될 수 없는 '생각'이나 '느낌'이 높은 순위에 있음을 확인할 수 있다. 외부연관도를 기준으로 정렬할 경우도 적합한 평가 항목인 '사이즈'보다 적합하지 않은 '바지'의 연관도가 매우 높게 나오는 결과를 볼 수 있다. 두 연관도 값을 결합하여 산출된 PMI-RTF값을



(그림 2) 상품 카테고리별 평가 항목 추출

〈표 1〉 후보 명사의 PMI-RTF 계산

후보 명사 (t_i)	출현빈도 ($f_{review \in c}(t_i)$)	내부 연관도	복합검색 ($f_{web}(c_{name}, t_i)$)	단일검색 ($f_{web}(t_i)$)	외부 연관도	PMI-RTF (*100)
길이	320	① 0.79	842,748	20,440,325	⑤ 0.041	② 3.26
사이즈	313	② 0.77	3,688,191	44,176,408	④ 0.083	① 6.45
생각	211	③ 0.52	784,726	99,885,669	⑦ 0.008	⑦ 0.41
느낌	150	④ 0.37	1,032,719	28,032,805	⑥ 0.037	⑥ 1.36
스타일	103	⑤ 0.25	2,305,445	25,926,349	③ 0.089	④ 2.26
속치마	43	⑥ 0.11	46,866	198,701	② 0.236	③ 2.50
바지	28	⑦ 0.07	3,746,296	12,307,888	① 0.304	⑤ 2.10

기준으로 정렬하면 '사이즈', '길이', '속치마', '스타일'과 같은 평가 항목이 '느낌', '생각', '바지'의 적절하지 않은 항목보다 높은 순서를 보여 원하는 결과를 얻을 수 있다.

4. 실험 및 평가

본 논문에서 제안한 시스템의 성능을 평가하기 위해 실험을 진행하였다. 여러 쇼핑몰의 상품평을 자동으로 수집하여 통합 제공하는 가격비교사이트[9]에서 상품평을 수집하였으며, 문장 보정을 위해 자동띄어쓰기 모듈[10]을, 형태소 분석과 태깅을 위해 형태소분석기[11]를, 검색을 위해 네이버 웹 문서 검색기를 사용하였다. 각 상품별 정답 평가 항목은 상품평에 출현하는 모든 명사 리스트에서 평가 항목으로 적합한 것을 수동으로 분류하여 구축하였다. 실험 데이터에 대한 상세 정보는 <표 2>와 같다.

평가 항목 추출에 대한 실험은 시스템이 생성한 평가 항목 리스트의 상위 50위내에 수동 분류한 정답 항목이 포함된 개수를 측정하였다. 카테고리마다 평가 항목과 평가 항목의 적정 개수가 달라 카테고리별 평가 항목의 적정 개수를 미리 알 수 없으므로 한 화면에 보여주기 적당한 개수로 상위 50위까지를 추출 범위로 하였다. 특징 추출 실험에서는 시스템 추출 범위인 50위 이내에 정답 데이터가 얼마나 포함되어 있는지의 재현율을 측정하였다.

PMI-RTF의 성능 평가를 위해 상품평의 어휘 빈도를 이용하는 내부연관도 방법과 웹 문서의 어휘 공기 빈도를 이용한 PMI 즉 외부연관도 방법으로 각각 상품 평가 항목 추출 실험을 수행하고 그 결과를 PMI-RTF와 비교하였다. 실험 결과는 <표 3>과 같다. 표에서 내부연관도는 [8]의 결과로, 외부연관도는 [6]와 [7]의 결과로 볼 수 있다. 내부연관도 방법의 평균 재현율은 62.34%, 외부연관도 방법의 평균 재현율은 74.49%였다. PMI-RTF 방법의 평균 재현율은 90.16%로 세 가지 방법 중 가장 우수한 결과를 보였다.

〈표 2〉 실험 데이터 상세 정보 (단위 : 개)

카테고리	상품	상품평	문장	형태소	명사	상품평 당 평균 문장 수	정답 평가항목
스커트	567	2,990	8,748	103,980	16,627	2.92	23
쌍안경	82	486	1,406	15,534	2,384	2.89	17
립라이너	71	1,477	4,317	51,735	8,850	2.92	18
로만쉐이드	157	1,438	4,525	57,372	9,797	3.05	30

<표 4>는 스커트 항목에 대한 각 접근 방법의 상위 20위의 결과를 보인다. 표에서 색으로 표시된 부분이 정답을 추출한 경우에 해당한다. PMI에 기반한 외부연관도는 상품 카테고리명 명사 '스커트'와 일반적인 관련성이 높은 단어를 추출하고, 빈도에 기반한 내부연관도는 상품평에 많이 포함된 단어를 추출하는 경향을 보였다. 이에 반하여 제안하는 PMI-RTF 방식에서는 스커트 구매에 영향을 줄 수 있는 어휘를 상위에 포함함을 볼 수 있다.

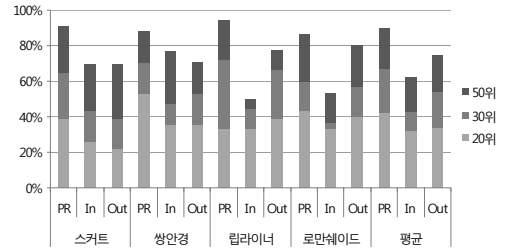
(그림 3)에서는 추출 범위를 상위 20위, 30위, 50위로 설정한 각 경우에 대한 본 논문의 방식과 내부연관도, 외부연관도의 재현율을 그래프로 나타낸다. 그림에서 PR는 PMI-RTF의 결과를, In은 내부연관도의 결과를, Out은 외부연관도의 결과를 보인다. 추출 범위를 다르게 했을 때도 본 논문의 방식이 립라이너에 대한 상위 20위에 대한 결과를 제외한 모든 경우에서 가장 좋은 결과를 보였다. 결과에서 대부분의 경우에 내부연관도에 비해 PMI를 이용하는 외부연관도가 좋은 성능을 보였다. 결과 분석에서는 상품평에 나타나는 표기 오류와 다양한 표현 방법이 내부연관도의 낮은 성능의 이유로 분석되었다. <표 4>의 내부연관도 결과에서 볼 수 있듯이 상품평은 띄어쓰기나 표기 오류로 '66사이즈', '입고', '제가'와 같이 형태소 오분석의 경우가 많았다. 또한 상품평에는 맞춤법 오류(예: 스탈-스타일, 초점-촛점)나 다양한 음차 표기(예: 배터리-밧데리, 베터리), 동의어로 표현되는 평가 항목(예: 색상-색깔, 크기-사이즈)이 발생하여, 상품평을 그대로 사용하는 경우에 동일한 평가 항목에 대한 빈도가 분산되어 평가 항목 추출의 성능이 떨어지는 현상을

〈표 3〉 평가 항목 추출 방법별 재현율 비교

	스커트	쌍안경	립라이너	로만쉐이드	평균
PMI-RTF	91.30%	88.24%	94.44%	86.67%	90.16%
외부연관도	69.57%	70.59%	77.78%	80.00%	74.49%
내부연관도	69.57%	76.47%	50.00%	53.33%	62.34%

〈표 4〉 스커트에 대한 평가 항목 상위 20위

	PMI-RTF	외부연관도	내부연관도
1	치마	미니스커트	치마
2	레깅스	블라우스	디자인
3	사이즈	레깅스	가격
4	블라우스	정장	길이
5	옷	쉬폰	사이즈
6	미니스커트	자켓	옷
7	정장	부츠	배송
8	색상	바지	색상
9	자켓	H라인	생각
10	길이	모직	제가
11	디자인	속치마	입고
12	쉬폰	밑단	느낌
13	배송	치마	저
14	속치마	66사이즈	키
15	주름	안감	사진
16	벨트	벨트	소재
17	스타일	네이비	화면
18	가격	정사이즈	상품
19	바지	허리부분	여름
20	부츠	골반	무릎



(그림 3) 추출 범위에 따른 재현율 비교

보였다. 이의 해결을 위해서는 잘 구축된 동의어나 상하위어 지식이 필요하며, 이는 <표 4>의 예제처럼 '스커트'의 동의어인 '치마'나 하위어인 '미니스커트'가 중요 연관 단어로 추출되는 문제도 해결할 수 있을 것으로 기대된다.

상품평의 어휘 빈도를 사용하는 내부연관도와 제안한 방식의 결과에서 상품평에 자주 등장하는 '배송'이나 '서비스'와 같이 상품이 아닌 쇼핑몰에 평가 항목이 상위에 추출되는 결과를 보였다. 추출된 평가 항목을 사용하는 응용 시스템에서는 평가 항목을 포함하는 상품평이나 상품평 내의 문장을 분류하여 조회하는 기능을 제공하게 되고, 이러한 평가 항목별 상품평 조회는 상품에 대한 평가와 '배송'이나 '서비스' 등의 쇼핑몰에 대한 평가가 혼재되어 제공되는 기존 상품평 제공 방식이 가지는 문제점을 해결할 것으로 기대된다. 또한 가격 비교 사이트나 쇼핑몰에서 '배송'이나 '서비스'에 대한 평가를 쇼핑몰이나 배송사 단위로 통합한다면 각 업체에 대한 평가를 용이하게 분석할 수 있는 추가 효과도 기대할 수 있다.

5. 결 론

본 논문에서는 근래 폭발적으로 증가하고 있는 상품평을 용이하게 활용하기 위한 평가 항목 자동 추출 시스템을 제안하였다. 실험 결과에서는 본 논문의 방식이 90.16%의 평균 재현율을 보여 기존의 방법에 비해 우월한 결과를 나타냈다. 제안된 시스템에서는 특징 추출 과정에서 한 단어 이상으로 구성된 명사 구(복합명사) 형태의 상품 특징을 추출하지 못하는 문제점이 남아 있다. 추후 연구를 통해 해결해야 할 과제이며, 추출된 평가 항목을 기준으로 상품평의 극성을 판별하여 상품평을 요약하는 연구도 진행 중이다.

참 고 문 헌

[1] 한국인터넷진흥원, "웹 2.0시대의 네티즌 인터넷 이용 현황 - 참여와 공유의 인터넷", http://www.nida.or.kr/doc/issue_sum.pdf, 2006.

[2] Kim, S.M., Pantel, P., Chklovski, T., Pennacchiotti, M, "Automatically Assessing Review Helpfulness," In Proc. of EMNLP, pp.423-430, 2006.

[3] Scaffidi, C., Bierhoff, K., Chang, E., Felker, M., Ng, H., and Jin, C. "Red Opal: product-feature scoring from reviews," In Proc. of the ACM Conference on Electronic Commerce, San Diego, California, USA, New York, pp.182-191, 2007.

[4] Hu, M. and Liu, B. "Mining opinion features in customer reviews," In Proc. of the 19th National Conference on Artificial Intelligence, San Jose, USA, pp.755-760, 2004.

[5] Hu, M. and Liu, B. "Mining and summarizing customer reviews," In Proc. of the 10th ACM SIGKDD Conf., pp.168-177, New York, NY, USA. ACM Press, 2004.

[6] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates, "Unsupervised named-entity extraction from the web: An experimental study," Artificial Intelligence, 165(1) pp.91-134, 2005.

[7] Popescu, A. and Etzioni, O. "Extracting product features and opinions from reviews," In Proc. of the Conference on HLT and EMNLP, pp.339-346, 2005.

[8] 명재석, 이동주, 이상구, "반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템," 정보과학회논문지 : 소프트웨어 및 응용, 제35권 제6호(2008. 6), pp.392-403, 2008.

[9] 온라인 가격비교 사이트 BB.co.kr, http://www.bb.co.kr.

[10] Naver Lab, 자동 띄어쓰기, http://s.lab.naver.com/autospacing/

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

이 우 철

e-mail : lee256@naver.com

2009년 금오공과대학교 소프트웨어공학과 (공학석사)

2009년~현 재 (주) 유승토탈솔류션 대리

관심분야 : 자연언어처리, 지식공학, 웹서비스

이 현 아

e-mail : halee@kumoh.ac.kr

1996년 연세대학교 컴퓨터과학과(학사)

1998년 한국과학기술원 전산학과(공학석사)

2004년 한국과학기술원 전산학과(공학박사)

2000년~2004년 (주)다음소프트 언어연구팀장

2004년~현 재 금오공과대학교 컴퓨터공학부 조교수

관심분야 : 자연언어처리, 정보검색, 지식공학, 기계번역

이 공 주

e-mail : kjoolee@cnu.ac.kr

1992년 서강대학교 전자계산학과(학사)

1994년 한국과학기술원 전산학과(공학석사)

1998년 한국과학기술원 전산학과(공학박사)

1998년~2003년 한국마이크로소프트(유) 연구원

2003년 이화여자대학교 컴퓨터학과 대우전임강사

2004년 경인여자대학 전산정보과 전임강사

2005년~현 재 충남대학교 전기정보통신공학부 부교수

관심분야 : 자연언어처리, 자연어인터페이스, 기계번역, 정보검색