상품특징별 점수화를 이용한 상품리뷰요약 시스템의 설계 및 구현 *

양정연, 명재석, 이상구

서울대학교 컴퓨터공학부

A product review summarization system using a scoring of features

Jung-Yeon Yang, Jaeseok Myung, Sang-goo Lee

요 약

온라인 마켓에 수많은 상품정보가 공개됨에 따라, 소비자들은 장소나 시간에 구애 받지 않고 자신이 원하는 상품을 구매할 수 있게 되었다. 하지만, 온라인 마켓의 경우 소비자들이 직접 상품을 살펴볼 수 없기 때문에, 다른 사람의 상품리뷰가 구매 의사결정에 많은 영향을 미친다. 한편, 많은 수의 리뷰를 모두 살펴보는 것은 구매자에게 부담으로 느껴진다. 이에 따라 많은 양의 상품리뷰를 분석하여 소비자에게 정제된 정보를 제공할 필요성이 제기되고 있다.

본 논문에서는 자연어처리 및 통계적 분석을 활용하여 상품의 특징을 추출하고, 각 특징별 평가점수를 소비자에게 제공하여 상품의 장단점을 보다 쉽고 정확하게 알 수 있도록 하는 상품평가 시스템의 설계 및 구현에 대하여 다루었다. 상품특징별 평가를 소비자에게 제

* 본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 육성·지원사업 (IITA-2008-C1090-0801-0031)의 연구결과로 수행되었음.

공함으로써, 소비자는 자신의 취향에 맞는 상품을 선택할 수 있는 기회를 얻을 수 있으며, 기업은 소비자의 상품에 대한 선호정보를 보다 구체적으로 파악할 수 있을 것으로 기대된다.

Abstract

As a number of product information is increasing in online markets, customers can purchase products with no spatial and time problems. However, in case of an online market, since customers can't see products directly, others' reviews make a big influence to customers. Meanwhile, it is a burden to read all reviews about some products. Therefore, we need to provide refined information to customers as summarizing whole product reviews.

In this paper, we explain about the product review summarization system which can provide to customers as show evaluation scores of product features. Natural Language Processing skills and computational statistics are utilized for summarization. Customers can get chances to buy a feasible product that he wants to get through this system. Moreover, Enterprises can find out the needs of customers deeply.

1. 서론

인터넷 쇼핑몰을 통한 구매활동이 활발해지면서, 점점 많은 수의 상품들이 구매자들에게 보여지고 있다. 하지만, 인터넷의 특성상 구매자들은 상품을 직접 살펴보고 구매할 수 없기 때문에, 다른 구매자들의 경험을 간접적으로 참고하게 된다. 이러한 간접경험을 나타낸 것이 상품리뷰인 것이다.

상품리뷰의 중요성이 커짐에 따라, 인터넷상에는 수 많은 상품리뷰가 나타나게 되었고, 구매자는 상품을 결정하는 것뿐만이 아닌, 상품리뷰를 살펴보는 것에도 많은 어려움을 겪고 있다. 이러한 문제를 해결하기 위하여 많은 수의 상품리뷰를 효율적으로 요약하여 구매자에게 중요한 정보만을 전달할 필요성이 대두되었다.

상품리뷰와 같은 문서로 이루어진 정보에서 의미 있는 정보를 뽑아내는 분야로써 오피니언마이닝 기술이 많이 연구되어 왔다. 주로 자연어처리 기법을 활용하여 상품리뷰에서 말하고 있는 평가 내용을 요약하거나, 통계적 기법을 통한 상품특징별 요약을 수행하는 시스템에 대한 연구가 있었다. 하지만, 각 시스템들은 자연어처리가 갖는 현실

성의 문제점 및 요약 방법 등에서의 단점들을 가지고 있다.

본 논문에서는 상품리뷰에 나타난 사용자의 의견정보를 추출 및 분석하고, 상품특징별 점수화를 통해 상품에 대한 평가를 내리는 방법에 대하여 연구하고, 이러한 과정을 수행하는 상품리뷰요약시스템을 설계 및 구현한 내용을 다루었다.

본 논문의 구성은 다음과 같다. 2장에서는 기존에 연구된 상품리뷰요약시스템에 대하여 설명한다. 3장에서는 본 연구의 상품리뷰 요약 방법 및 시스템의 설계를 다룬다. 4장에서는 실제 시스템의 구현 예를 보이고, 5장에서 결론 및 향후 연구방향에 대하여 논하였다.

2. 관련연구

2.1 상품리뷰요약 방법 [1, 2, 6, 7, 8]

상품리뷰를 요약하는 방법으로는 크게 자연어처리기법과 통계학적 접근법이 있다. 오피니언마이닝 기술로 가장 많이 연구되어진 방법이 자연어처리기법이다. 리뷰문서의 내용을 파악하기 위해서 POS 태그를 활용하여 문서의 문장들을 문장성분 단위로 파싱한 뒤, 문장성분들 사이의 관계 및 문장성분의 패턴정보 등을 활용하여 리뷰에서 사용자가 말하고자 하는 내용을 파악한다. 이 때, 가장 큰 역할을 수행하며 의미판단의 기초가 되는 것으로써 워드넷이 활용된다. 최근의 연구에서는 어휘의 긍정적 또는 부정적 의미의 정도를 정량화하여 표현한 센티워드넷도 활용되고 있다.

통계적 방법은 자연어처리기법이 전문가의 손이 많이 필요하며, 많은 시간을 소비한다는 단점을 해결하기 위한 대체방안으로 주목되었다. 문서 내의 어휘의 출현빈도 및 TF/IDF, PMI 등의 방법을 이용하여 리뷰문서 내의 중요한 상품특징 및 의견을 나타내는 어휘를 추출하고, 그 의미를 판단하였다. 특히, 최근에는 사용자가 리뷰와 함께 상품에 대해 평가정도를 나타낸 점수(별점)정보를 활용하여, 사용자의 의견을 파악하는데 활용한 연구가 이루어졌다. 점수정보는 사용자가 자신의 평가정도를 명확하게 나타난 지표의 역할을 하기 때문에 중요하게 활용될 수 있다.

2.2 상품리뷰요약 시스템[1, 2, 3, 6]

앞 절에서 설명한 두 가지 대표적인 접근방식과 관련하여 크게 3 가지 형태의 대표적인 상품리뷰요약 시스템이 존재한다. 첫번째 시스템은 자연어처리기법에 기반한 것으로, 그림 1 의 (가)와 같이, 각 상품특징별로 리뷰 내에서 긍정적 또는 부정적 의미를 가진 어휘의 수를 계산하여 누적 정보를 보여준다. 두번째 시스템은 사용자의 평가점수를 활용한 것으로써, 각 상품특징 어휘의 출현빈도와 상품에 대한 평가점수를 활용하여 상품특징별로 점수화하여 (나)와 같이 사용자에게 제공한다. 마지막 형태의 시

스템은 상품특징을 나타내는 어휘들의 출현 빈도 정보를 활용하여, 특징별 중요성을 트리맵의 형태로 사용자에게 제공하는 것으로, (다)와 같은 형태를 보인다.

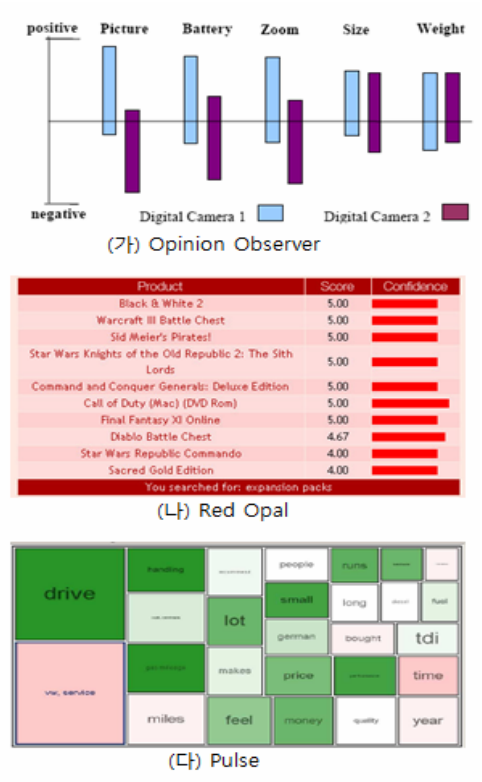

그림 1. 상품리뷰요약 시스템

3. 리뷰요약 방법 및 시스템 설계

3.1 상품특징별 리뷰요약 방법[4, 5]

본 연구에서는 상품리뷰요약을 수행할 때, 상품의 대표적인 특징별로 점수화하여 사용자에게 제공한다. 이것은 사용자들이 상품을 구매할 때 자신에게 필요한 적합한 특징을 고려하여 구매할 수 있도록 하기 위함이다. 본 연구에서는 상품정보리뷰를 그림 2와 같이 정의하였다.

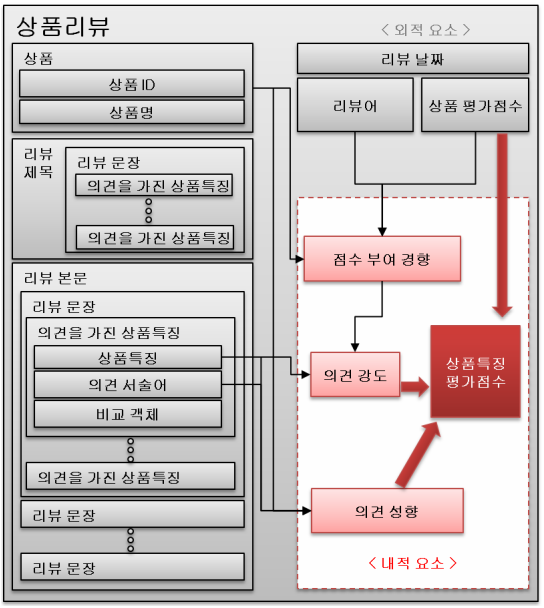

그림 2 상품리뷰 모델

상품리뷰요약은 크게 3단계, 상품특징추출, 의미분석, 특징별 점수화 단계로 이루어진다. 각 단계별 상호 역할 및 프로세스 단계는 그림 3과 같다.

상품특징추출 단계에서는 RDB 형태의 리뷰데이터를 파싱을 통해 POS태그가 부여된 어휘들로 변환한다. 변환된 어휘들로부터 패턴, 출현빈도, 어휘간 거리 등의 정보를 활용하여 요약에 사용될 상품특징세트를 완성하게 된다. 또한, 추출된 상품특징 어휘들을 수식하고 있는 의미를 가진 어휘들도 함께 추출하게 된다.

의미분석 단계에서는 상품특징추출단계

에서 뽑아진, <상품특징 어휘, 의미표현 어휘>의 쌍으로 이루어진 어휘들이 의미하는 것이 긍정적인지, 또는 부정적인지를 분석한다. 이를 위해서 기 구축된 의미사전을 활용하여 PMI값을 계산한 후, 의미적 성향을 판별한다.

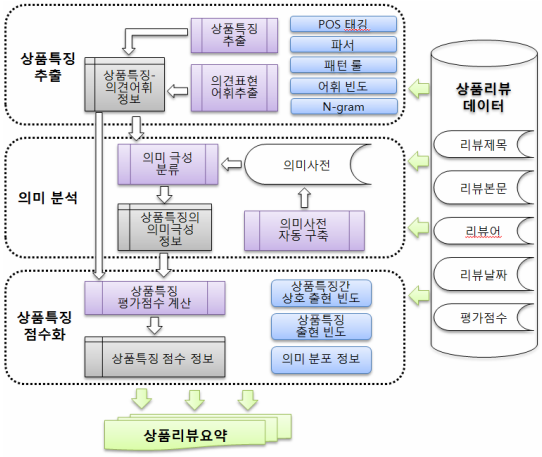

그림 3 상품리뷰요약 프로세스

특징별 점수화 단계에서는 의미분석 단계에서 얻어진 의미적 성향정보 및 사용자가 리뷰에서 부여한 상품에 대한 평가점수를 활용하여 상품특징별로 평가점수를 계산한다. 한 리뷰 안에는 여러 가지의 상품특징들이 언급되고 개별적인 평가가 이루어지게 된다. 그러나, 리뷰에 부여된 평가점수는 종합적인 의미를 나타낸다. 따라서, 특정 상품특징에 대한 평가점수를 계산하기 위해서 본 연구에서는 상품특징의 의미정보와 리뷰의 평가점수를 활용하여 새로운 점수를 계산한다.

3.2 상품리뷰요약시스템 설계

앞 절에서 설명한 각 단계별 처리를 위해서 그림 4와 같이 시스템 구성을 설계하였다. 구성도의 하단에서부터 각각 상품특징 추출, 의미분석, 상품특징별 점수화 기능을 담당한다.

전처리 모듈에서는 RDB 형태로 존재하는 상품리뷰문서를 파서 및 태깅모듈이 사용할 수 있는 형태로 처리한다. 파서 및 태깅모듈을 통해서 리뷰데이터는 POS 태그가 부여된 문장성분단위로 저장된다. 기본, 복합 추출모듈을 통해 간단한 추출 알고리즘을 통한 특징 선별 및 두 개의 알고리즘을 조합한 특징 선별 과정을 거친다. 마지막으로, 선별된 상품특징에 대하여 각각의 서술어를 추출하게 된다.

상품특징 및 서술어가 추출된 후, 어휘의미 판단모듈은 의미사전을 참조하여 각 상품특징들이 어떤 의미로 표현되었는지에 대한 평가의미의 극성을 판별한다. 의미사전 구축모듈은 상품리뷰데이터를 활용하여 긍정적 경우 및 부정적 경우의 의미를 대표하는 어휘사전을 생성함으로써 어휘 의미 판단모듈이 활용할 수 있도록 한다.

점수계산 모듈은 이전 단계에서 마련된 상품 특징별 평가 극성 정보 및 리뷰 내 평가 극성의 분포정보, 상품특징의 리뷰 내 출현 빈도 등을 활용하여 각 상품특징별로 사용자가 내린 평가 정도를 유추하여 점수로 표현한다. 이 정보를 요약조회모듈을 통

해서 사용자에게 최종적으로 제공하게 된다.

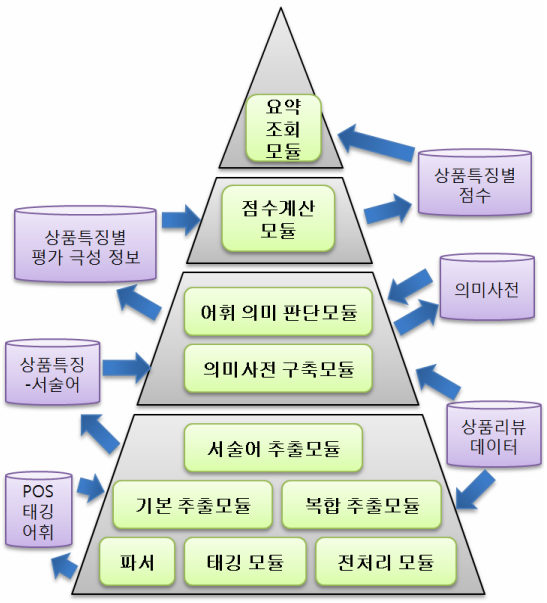

그림 4 상품리뷰요약시스템 구성

4. 시스템 구현 예시

본 장에서는 앞에서 설계한 상품리뷰요약 시스템의 실제 구현 예시를 보이고, 각 기능에 대한 설명을 한다.

그림 5는 상품리뷰요약의 첫 단계로써, 상품을 대표하는 특징들을 추출하는 기능을 보이고 있다. RDB에 존재하는 데이터소스를 선택하여 다양한 추출 알고리즘을 적용한 후, 추출된 어휘들을 활성화 또는 비활성화 할 수 있다.

그림 6은 추출된 상품특징 어휘를 수식하고 있는 서술어를 추출하는 기능을 보이고 있다. 이 때 수식하는 범위를 문장단위 또는 특징을 나타내는 어휘와의 거리단위 등으로 선택하여 추출할 수 있으며, 추출 대상이 되는 서술어의 문장성분도 선택할 수 있다.

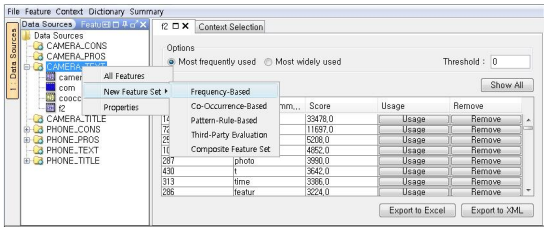

그림 5 상품특징 추출기능 화면

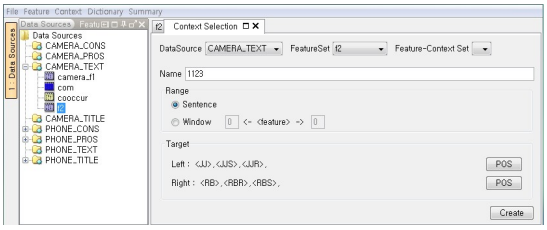

그림 6 상품특징 서술어 추출기능 화면

그림 7은 의미사전을 구축하는 기능을 보이고 있다. RDB에 저장된 데이터소스를 선택한 후 의미사전들을 생성 및 삭제하는 기능을 수행한다.

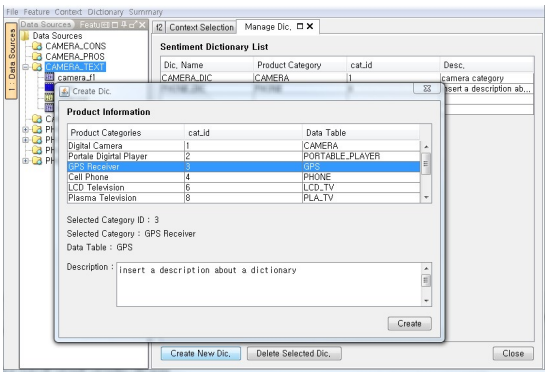

그림 7 의미사전 구축기능 화면

그림 8에서 보는 것과 같이, 최종 리뷰요약 과정은 대상이 되는 <상품특징, 서술어> 쌍 정보와 의미사전을 선택한 후 점수화 과정을 거치게 됨으로써 완성된다.

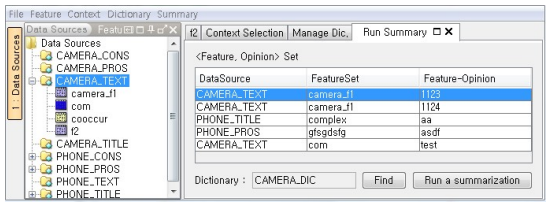

그림 8 리뷰요약 실행 화면

그림 9는 완성된 상품리뷰요약 결과를 조회하는 화면이다. 사용자는 상품분류를 선택 한 후, 조회하고자 하는 특정 상품을 선택할 수 있다. 시스템은 선택된 상품에 관련된 리뷰를 요약한 내용을 상품특징별로 평가된 점수로 표현해 준다.

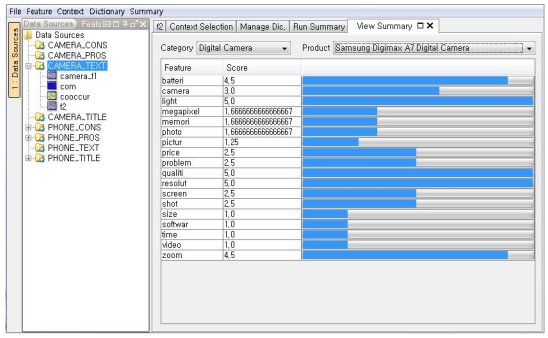

그림 9 상품리뷰요약 조회기능 화면

5. 결론

본 논문에서는 상품특징별 점수화를 활용한 상품리뷰요약 시스템의 설계 및 구현에 대하여 다루었다. 이를 위해 각 단계별 프로세스를 설명하고, 그 역할을 수행하는 모듈의 구성 및 기능에 대하여 설명하였다. 기존의 시스템 및 방법들과 비교했을 때, 전문가의 큰 노력이 필요 없이 다양한 알고리즘을 쉽게 적용해 볼 수 있으며, 사전 구축의 자동화를 통해 유연성을 높였다. 또한 리뷰의 평가점수를 그대로 보여주는 것이 아닌, 상품특징별 평가점수를 계산하여 사용자에게 제공함으로써 좀 더 정확한 리뷰 요약을 수행하였다. 따라서, 제시된 시스템을 통하여 대용량의 상품리뷰를 사용자에게 효율적으로 요약하여 보여줄 수 있게 되었다.

향후, 웹을 통한 상품리뷰요약 조회가 가능하도록 하고, 국내 온라인 상품리뷰 데이터를 대상으로 리뷰요약시스템을 개발할 예정이며, 국내 데이터의 특성에 맞는 의미사전 구축 및 상품특징별 점수화 방법에 대한 고도화 연구를 지속할 예정이다.

참고문헌

[1] B. Liu, M. Hu, J. Cheng, “Opinion Observer: Analyzing and Comparing Opinions on the Web”, the 14th Intl. World Wide Web Conf. 2005

[2] C. Scaffidi, K Boerhoff, et al, “Red Opal: Product-Feature Scoring from Reviews”, ACM Conf. on Electronic Commerce (EC’07), June 11-15, 2007

[3] M. Gamon, A Aue, et al, “Pulse: Mining Customer Opinions from Free Text”, the 6th Intl. Symposium on Intelligent Data Analysis, Sep. 8-10, 2005

[4] J. Yang, J. Myung, S. Lee, "The method for a summarization of product reviews using the user’s opinion”, Intl. Conf. on Information, Process, and Knowledge Management (eKNOW 2009), Feb 1-7, 2009

[5] J. Yang, J. Myung, S. Lee, “PicAChoo: A Tool for Customizable Feature Extraction Utilizing Characteristics of Textual Data”, Intl. Conf. on Ubiquitous Information Management and Communication (ICUIMC 2009), Jan. 15-16, 2009

[6] 명재석, 이동주, 이상구, “반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템”, 정보과학회논문지: 소프트웨어 및 응용 제 35권, 제6호, 2008.

[7] R. Ghani, K. Probst, et al, “Text Mining for Product Attribute Extraction”, ACM SIGKDD Explorations Newsletter, 2006

[8] A. Popescu, B. Nguyen, O. Etzioni, “OPINE: Extracting product Features and Opinions from Reviews”, HLT/EMNLP 2005, Oct. 2005.

저자소개

양정연(e-mail: jyyang@snu.ac.kr)은 2002년 충남대학교 컴퓨터공학부 학사를 취득하고, 2003년부터 현재까지 서울대학교 대학원 컴퓨터공학부 석박사통합과정에 재학 중이다. 관심분야는 오피니언 마이닝, 상황인지 서비스, e-비즈니스 기술, 시맨틱 웹이다.

명재석(e-mail: jsmyung@snu.ac.kr)은 2007년 성균관대학교 정보통신공학부 학사를 취득하고, 2007년부터 현재까지 서울대학교 대학원 컴퓨터공학부 석박사통합과정에 재학 중이다. 관심분야는 상황인지 서비스, 시맨틱 웹, e-비즈니스 기술이다.

이상구(e-mail: sglee@snu.ac.kr)은 1985년 서울대학교 계산통계학과 학사를 취득하고, 1987년 Northwestern Univ. 컴퓨터과학과 석사를 취득하였으며, 1990년 Northwestern Univ. 컴퓨터과학과 박사를 취득하였다. 1992년부터 현재까지 서울대학교 교수로 재직중이며, e-비즈니스 기술연구 센터장을 역임하고 있다. 관심분야는 e-비즈니스 기술, 데이터베이스, 상황인지 서비스이다.