연관 규칙을 사용한 상품평 오피니언 마이닝

김원영, 류준석, 김응모
성균관대학교 전자전기컴퓨터공학과
e-mail: wykim@ece.skku.ac.kr

Opinion Mining of Product Reviews using Association Rules

Won-Young Kim, Joon-Suk Ryu, Ung-Mo Kim
Dept of Electrical and Computer Engineering, Sungkyunkwan University

요 약

사용자가 웹 상에 작성한 상품평은 다양한 정보를 포함하고 있는 데이터이다. 대부분의 사람들이 상품을 구입하기 전에 상품평을 통해서 상품에 대한 많은 정보를 얻는다. 이에 따라 대량의 상품평 데이터로부터 유용한 정보를 추출하여 요약하는 오피니언 마이닝에 대한 연구가 활발히 진행되고 있다. 본 논문에서는 사용자가 많은 상품평들을 모두 읽어보지 않고 상품에 대한 오피니언과 장점과 단점을 쉽게 알 수 있도록 연관 규칙 마이닝을 적용하는 오피니언 마이닝 방법을 제안한다.

1. 서론

월드 와이드 웹(World Wide Web)이 널리 보급되기 이전부터 대부분의 사람들은 상품을 구매할 때 다른 사람의 의견을 많이 물어보고 어떤 상품이 좋은지 추천을 받아서 구매를 결정하였다. 이처럼 사람들은 그들이 상품을 구입하기 전에 상품에 대해서 많은 정보를 얻으려고 한다. 이때 다른 사람의 상품에 대한 의견은 아주 중요하게 작용한다. 오늘날 웹2.0의 발전과 함께 웹 상에 많은 사용자 데이터가 축적이 되었다. 이러한 사용자 데이터들 중에는 사람들이 상품이나 서비스에 대한 의견을 작성해놓은 블로그나 리뷰 같은 데이터가 있다. 그 결과로 블로그나 리뷰 데이터로부터 유용한 정보를 추출해내는 오피니언 마이닝에 대한 연구가 활발히 진행되고 있다.

본 논문에서는 연관 규칙을 이용하는 오피니언 마이닝 방법을 제안한다. 상품평 데이터를 구문 분석을 통해 본 논문에서 정의하는 트랜잭션 T의 형태로 저장하여 트랜잭션 데이터로부터 연관 규칙 마이닝 기법을 적용하여 유용한 정보를 추출한다.

본 논문은 다음과 같이 구성된다. 2장에서 오피니언 마이닝에 대해 알아보고 기존 연구에 대해 설명한다. 3장에서는 제안하는 오피니언 마이닝 방법에 대해서 설명하고, 4장에서 결론 및 향후 연구 과제를 제시한 후 본 논문을 마친다.

2. 관련 연구

오피니언 마이닝은 사람들이 작성한 데이터로부터 의미있는 오피니언을 추출해서 유용한 정보를 제공하는 기법이다. 오피니언 마이닝을 수행하기 위해서는 크게 세가지 하위 단계를 필요하다. 첫째, 문장의 구문을 분석하여 대상의 특징과 오피니언을 추출하는 것이다. 기존의 연구 결과들이 간단한 문장에 대해서는 높은 정확도를 보이지만 복잡한 문장의 문맥을 고려한 정확성 측면에서는 만족스럽지 못하다. 이에 따라 자연언어처리 기법을 이용하여 정확도를 향상시키는 연구[1][2]가 시도되고 있다.

둘째, 대상의 특징에 대한 오피니언의 극성을 분석하는 것이다. 간단한 방법으로 PMI(Pointwise Mutual Information) 기법을 적용할 수 있다. [3]에서 PMI 지수를 계산하여 오피니언의 극성을 판별한다. [4]에서는 기계학습 기법을 이용한다.

셋째, 대상에 대한 오피니언을 나타내는 언어학적 자원을 정의하고 구축하는 것이다. 양질의 언어 자원 구축은 오피니언 마이닝의 정확성을 높일 수 있다. 영어권에서는 WordNet을 확장하여 언어 자원을 구축하려는 시도가 존재하였다[5]. 아직 한국어를 기반으로 하는 오피니언 마이닝 연구는 활발히 이루어지지 않고 있어 앞으로 이 부분에 대한 지속적인 연구가 필요하다.

3. 제안하는 오피니언 마이닝 방법

본 논문에서 제안하는 오피니언 마이닝 방법은 크게 3가지 단계로 구성되어 있다.

3.1. 전처리과정

리뷰를 모은 후에, 전처리 과정이 필요하다. 전처리 과정에서는 파싱으로 상품평의 문장에 대해 구문 분석을 하여 트랜잭션 T의 형태로 저장하는 일을 수행한다. 전처리 과정은 크게 두 가지 단계로 나누어진다. 우선 파서를 이

용해서 상품평의 각 문장의 구문을 분석한다. 그리고 분석된 구문 트리로부터 상품에 대한 특징(feature)과 오피니언 어휘를 뽑아내어 트랜잭션 T의 형태로 저장한다. 트랜잭션 T의 형태는 다음과 같이 정의한다.

$T = (product, [feature_1, opinion_1], \cdots, [feature_n, opinion_n], opinion_1, \cdots, opinion_i)$

product: 상품의 이름
feature: 상품평에서 언급한 상품의 특징
opinion: 상품에 대한 오피니언
[feature, opinion]: 특징-오피니언 집합

3.2. 연관 규칙 마이닝

상품평 속에서 상품에 대한 신뢰성 있는 정보를 얻기 위해서는 많은 사람들이 언급하는 특징과 오피니언이 필요하다. 본 논문에서는 신뢰성 있는 정보를 얻기 위해서 전처리 과정을 통해 생성된 트랜잭션 T로 이루어진 데이터에 연관 규칙 마이닝 기법을 적용한다. 트랜잭션 T 데이터로부터 추출하는 연관 규칙을 통해서 빈발하게 언급되는 상품의 특징과 오피니언을 찾아낼 수 있다.

본 논문에서 제안하는 방법은 상품평 속에서 다음과 같은 형태의 연관 규칙을 찾아내어 유용한 정보를 추출한다.

Type 1: *product* ⇒ *opinion*
Type 2: *product* ⇒ *feature*
Type 3: *feature* ⇒ *opinion*
Type 4: *product* ⇒ [*feature* ⇒ *opinion*]

Type 1 규칙은 사용자들의 상품에 대한 전반적인 오피니언을 나타낸다. 이 규칙은 상품의 만족도에 대해서 알고 싶어하는 사용자들에게 유용한 정보를 제공한다. Type 2 규칙은 상품평에서 빈발하게 언급되는 특징을 나타내는 것이고, 빈발하게 언급되는 특징은 중요한 정보가 된다. Type 3 규칙은 상품의 각 특징에 대한 오피니언을 나타내고, 상품의 각 특징에 대한 오피니언을 알고 싶어하는 사용자들에게 유용한 정보를 제공한다. Type 4 규칙은 Type 2 규칙과 Type 3 규칙을 조합한 규칙으로 각 특징에 대한 오피니언의 극성을 계산하는데 사용되어 그 특징이 장점인지 단점인지 결정할 수 있다.

3.3. 오피니언 요약

이 섹션에서는 본 논문에서 제안하는 오피니언을 요약하는 방법에 대해 설명한다. 연관 규칙 마이닝을 적용하여 발견한 네 가지 형태의 연관 규칙을 이용하여 상품에 대한 전반적인 오피니언과 상품의 장점과 단점을 결정하여 요약한다. 신뢰도가 높은 Type 4의 특징의 오피니언에 대해 PMI-IR 알고리즘을 사용하여 긍정적인지 부정적인지 계산하여 상품의 장점과 단점을 결정한다.

4. 결론

본 논문에서 상품평을 분석하기 위하여 제안하는 방법은 파싱을 통해 상품의 특징과 오피니언을 트랜잭션 T 형태로 저장하는 방법을 제안하였다. 또한 저장된 트랜잭션 데이터의 분석은 연관 규칙 마이닝 기법을 활용하여 내가지 형태의 연관 규칙을 찾고, 그 연관 규칙에 대해서 PMI 기법을 사용하여 오피니언을 요약한다. 본 논문에서 제안하는 오피니언 마이닝 방법을 통해서 사용자들이 많은 상품평들을 모두 읽지 않고도 상품에 대한 전반적인 오피니언과 장점과 단점을 쉽게 알 수 있다.

오피니언 마이닝을 하기 위해서는 자연언어처리 기법을 통해 문장을 분석하는 것이 중요하다. 하지만 기존의 자연언어처리 기법으로는 간단한 문장에 대해서는 비교적 높은 정확도로 구문을 분석할 수 있지만 복잡한 형태의 문장은 분석하기 어렵다. 따라서 좀 더 정확한 구문 분석을 할 수 있는 자연언어처리 기법에 대한 연구가 필요하다. 그리고 오피니언의 극성에 따른 세기를 계산하여 각 특징에 대해 좀 더 자세하게 표현할 수 있는 방안에 대한 연구도 필요하다.

감사의 글

이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. 2009-0075771).

참고문헌

[1] Jeonghee Yi, Wayne Niblack, "Sentiment Mining in WebFountain," Proceedings of the 21st International Conference on Data Engineering, pp. 1073-1083, 2005.
[2] Ana-Maria Popescu, Oren Etzioni, "Extracting Product Features and Opinoins from Reviews," HLT/EMNLP, pp. 339-346, 2005.
[3] Turney, P. D., "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," Proceedings of the 40th Anual Meeting of the Association for Computational Linguistic (ACL'02), Philadelphia, Pennsylvania, pp. 417-424, 2002.
[4] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," Proceedings of the ACL 02 conference on Empirical methods in natural language processing, Volume 10, 2002.
[5] Esuli A., Sebastiani F., "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining," In Proceedings of LREC-06, 5th Conference of Lnaguage Resources and Evaluation, pp. 417-422, 2006.