영화 리뷰에 대한 한국어 오피니언 마이닝 기법

방수라*, 김원영*, 류준석*, 김응모*
*성균관대학교 정보통신공학부
e-mail : tnfk06@skku.edu

Mining Movie Reviews in Korean

Soo-Ra Bang*, Won-Young Kim*, Joon-Suk Ryu*, Ung-Mo Kim*
*School of Information and Communication Engineering, SungKyunKwan University

요 약

인터넷 시장이 빠르게 성장함에 따라 사용자들의 참여도가 매우 높아졌다. 인터넷 사용자들은 시장의 상품과 더불어 영화와 같은 문화 생활에 관한 의견을 웹 상에 표현하기 시작했고, 현재에 들어 그 양이 방대해 졌다. 본 논문에서는 사용들이 작성한 영화에 관한 리뷰를 모아 방대한 양에서 유용한 정보를 효율적으로 도출하고 요약해서 사용자에게 제공하는 방법을 제안한다.

1. 서론

사용자의 참여를 강조하는 Web2.0 의 성장과 발전으로, 웹 상에 각종 리뷰가 크게 증가 하였다. 사용자의 리뷰는 다른 잠재적인 사용자의 결정에 많은 영향을 미치기 때문에 그것이 매우 중요하고 유용한 정보가 되었다. 하지만 사용자들은 범람하는 리뷰들 가운데 그들에게 유용한 정보만을 가려내서 수렴하는 것이 어려워졌다. 이러한 배경으로 리뷰를 마이닝하고 요약하는 기법이 필요하게 되었다. 오피니언 마이닝의 연구는 주로 영어에 대해서 이루어 졌기 때문에 한글로 작성된 리뷰에 대한 마이닝 기술을 제안하는 것을 목표로 연구했다. 특히 지금까지의 연구는 상품에 대한 리뷰의 마이닝과 요약이 주를 이루고 있기 때문에 본 논문에서는 다른 도메인인 영화에 대한 리뷰를 분석해 보았다. 사용자가 어떤 영화에 대한 정보를 얻고자 할 때, 모든 리뷰를 다 읽고 판단하기 보다는 특정 영화에 대한 요약된 정보를 접함으로써 판단을 쉽게 할 수 있다.

본 논문은 다음과 같이 구성된다. 2 장에서 오피니언 마이닝 기법과 기존의 대표적인 연구에 대해 설명한다. 3 장에서는 본 논문에서 제안하는 영화 리뷰에 대한 한국어 오피니언 마이닝 기법에 대해서 설명하고, 4 장에서 결론과 발전된 연구를 위한 향후 연구 과제를 제시하며 논문을 마친다.

2. 관련연구

오피니언 마이닝이란 사용자가 웹 상에 작성한 문서를 분석해서 작성자의 주관적인 의견을 추출해 내는 작업이다.[1] 본 논문에서 오피니언 마이닝 기법을 이용해서 영화 리뷰들을 분석하고 요약 하는데 관련 있는 연구가 Sentiment Classification 에 관한 것과 Text Summarization 에 대한 연구이다.
여기서 이 연구들과 우리가 제안할 기술과의 관련성과 더불어 차이점에 대해서 설명할 것이다.

2.1 Sentiment Classification 기법

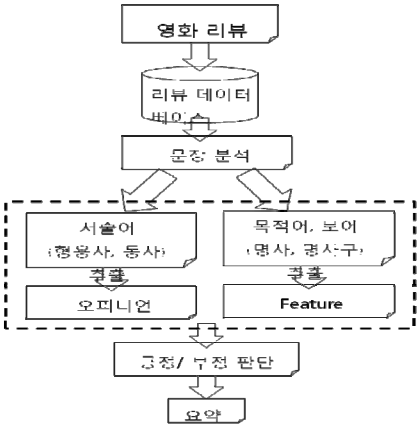
Sentiment Classification 기법이란, 단어, 문장, 문서의 의미적인 방향을 결정하는 것인데 여기서 의미적인 방향이란, 그 단어, 문장, 문서가 긍정적인지 부정적인지에 대한 태도를 말한다. Hearst[2] 와 Scak [3]의 연구에서는 전체 문서를 인지언어학에 의해서 도출된 모델을 바탕으로 분류하였다. Tong[4]은 영화에 대한 의견을 추적해서 시간에 따른 긍정적 의견의 수와 부정적 의견의 수를 보여주었다. 예를 들어서 "연기가 좋았다", "화면이 멋있었다" 등과 같은 영화에 대한 의견을 수동적으로 특별한 사전에 추가한 후에 그것의 긍정, 부정을 판단하는 과정을 거친다.
이 논문에서는 의견을 모아서 즉시 문장과 단어의 긍정과 부정을 판단하고 그 수를 세는 방법을 제시한다. 초기의 sentiment classification 은 문서 단계에서만 이루어졌고 그 방법과 분류의 범위가 점점 자세해 졌다. 그러나 이 기법으로는 구체적인 feature(특징)을 발견할 수 없다. 즉, 어떤 feature 에 대한 의견인지 알 수 없는 것이다. 이 논문에서는 문장 단계에서 표현된 각각의 특징에 대한 의견을 추출해 내도록 한다.

2.2 Text Summarization 기법

Text Summarization 기법은 문서에서 중요하다고 생각되는 문장을 추출해서 문서를 요약하는 기법이다. 대부분의 연구는 한 개의 문서에 대한 요약에 초점을 맞추었다. 비슷한 정보를 담고 있는 다수의 문서를 요약하는 방법에 대한 연구도 있었지만, 이러한 연구의 가장 큰 목적은 그 정보들 사이의 유사성과 차이점을 찾는 것이었다[5]. 본 논문에서 제안할 기술의 목표는 리뷰들 사이의 유사성과 차이점을 요약하는 것이 아니고 다수의 리뷰에서 언급한 주요한 특징을 찾는 것이기 때문에 text summarization 기법과 관련이 있지만 차이가 많이 존재한다.

3. 제안내용

본 논문에서 제안하는 한국어로 작성된 각종 리뷰에 대한 마이닝 기법은 다음과 같이 구성된다.



(그림 1) 제안기법의 구조

(1) 영화에 대한 리뷰를 각종 포털 사이트(네이버, 다음, 엠파스 등)와 공개된 블로그에서 가져와 리뷰 데이터베이스에 모은다[6].

(2) 리뷰 데이터베이스에 모은 문장에 대해 분석을 한다. 기본적으로 문장 구성의 필수 성분인 주어, 목적어, 보어, 서술어로 이루어진 문장에 대해서 feature(특징)와 오피니언을 추출하도록 한다. 서술어는 문장의 주체를 설명하는 성분으로, 문장의 기본 골격에서 '어찌한다, 어떠하다, 무엇이다' 에 해당한다. 이것은 형용사, 동사로 표현되기 때문에 문장에서 형용사나 동사가 그 문장의 오피니언이 될 수 있다. 이렇게 추출된 오피니언으로 문장의 긍정/부정 방향을 예측 할 수 있다. 그리고 문장에서 목적어와 보어에 존재하는 명사나 명사구는 feature 로 추출 될 수 있다.
이 기법에 대한 간단한 예를 몇 가지 들어 보겠다. 여기 영화에 대한 세 문장이 있다.

"나는 이 영화가 정말 재미있었다."
"이 영화의 배경음악은 영화와 잘 어울렸다."
"주연 배우의 연기가 너무 어색했다."

리뷰들 에서 서술어는 각각 '재미있었다', '어울렸다', '어색했다'가 된다. '재미있었다'와 '어울렸다'는 긍정적인 형용사와 동사이기 때문에 이 문장의 오피니언은 긍정적일 것이라는 예측이 가능하게 된다. 세 번째 문장에서의 서술어인 '어색했다'는 부정적인 의미의 형용사이기 때문에 이 문장의 방향은 부정적인 것으로 예측되는 것이다.
이 문장들에 존재 하는 명사나 명사구를 찾아 보면 '영화', '배경음악', '배우의 연기'가 있다. 이것들이 각각 '재미있었다', '어울렸다', '어색했다'의 feature 가 되는 것이다.

(3) 영화 리뷰에 대한 오피니언을 요약한다. 특정한 영화의 여러 리뷰를 분석한 결과를 간결하게 보여주는 작업을 한다. 여러 가지 feature 에 대한 긍정도/부정도를 계산하여 수, 우, 미, 양, 가로 표현한다. 해당 feature 에 대한 긍정도가 90% 이상이면 '수', 80%이상 90%이면 '우', 60%이상 80% 미만이면 '미', 40%이상 60%미만이면 '양', 40% 미만이면 '가'로 표현한다.

영화	오펀-천사의 비밀				
Feature	배우	연기	내용/스토리	배경음악	...
결과	우	수	수	미	...
총점	우				

(그림 2) 예시 결과

4. 결론

본 논문에서 제안하는 기술은 영화에 관한 각종 리뷰를 모아 거기에서 작성자의 주관적인 의견을 추출해 내는 것이다. 한글 문장의 필수 성분 요소인 주어, 목적어, 보어, 서술어를 분석하여 각 문장에서 feature 와 오피니언을 추출하는 방법을 제시 하였다. 또한 오피니언과 feature 를 추출하는 과정에서 긍정과 부정의 방향을 예측할 수 있다. 이렇게 특정 영화에 대한 리뷰를 분석한 결과를 간결하게 요약해서 보여줌으로써 사용자가 웹 상의 정보를 쉽고 빠르게 볼 수 있도록 한다.

향후 과제로는 한글 문장에서 관용어구로 이루어져 서술어 만으로는 오피니언의 판단이 불가능한 문장을 분석 가능하도록 하는 연구와 더불어 웹 상에서 많이 쓰이는 인터넷 용어 또는 줄임 말을 추출하여 그것의 사전적 의미뿐만 아니라 문맥적 의미를 분석할 수 있는 기술의 개발이 필요하다. 또한 그 가운데 오피니언으로써 가치 없고, 유용하지 않은 정보를 걸러내는 기술의 연구가 필요할 것이다.

감사의 글

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

참고문헌

[1] Bo Pang and Lillian Lee, "Opinion mining and Sentiment Analysis"
[2] Hearst, M, Direction-based Text Interpretation as an Information Access Refinement. In Paul Jacobs, editor, *Text-Based Intelligent Systems*. Lawrence Erlbaum Associates, 1992.
[3] Sack, W., On the Computation of Point of View. *AAAI'94*, Student abstract, 1994.
[4] Tong, R., An Operational System for Detecting and Tracking Opinions in on-line discussion. *SIGIR 2001. Workshop on Operational Text Classification*.
[5] Mani, I., and Bloedorn, E., Multi-document Summarization by Graph Search and Matching. *AAAI'97*.
[6] Minqing Hu and Bing Liu, Mining and summarizing customer reviews. *Association for Computing Machinery.2004.*