Opinion Mining 을 이용한 신문 기사 사용자 의견 추출

황치훈, 류준석, 김응모
성균관대학교 전자전기컴퓨터공학과
e-mail : chihoon@skku.edu

Opinion Mining from Internet Article

Chi hoon Hwang, Joon suk Ryu, Ung mo Kim
* Dept of Electrical and Computer Engineering, Sungkyunkwan University

요 약

오늘날 인터넷의 발달 때문에 인터넷으로 쉽게 신문을 볼 수 있게 되었다. 또한, 해당 기사에 대한 의견을 사용자끼리 쉽게 교환할 수 있다. 본 논문에서는 이러한 인터넷 기사의 사용자 의견들에 Opinion Mining 기술을 활용하여 해당 기사 대상의 특징을 올바르게 파악하는 방법을 제안한다.

1. 서론

오늘날 인터넷이 발달함에 따라 인터넷상으로 쉽게 볼 수 있는 인터넷 신문도 함께 발달하였다. 이러한 인터넷 신문의 발달 때문에 사용자끼리 해당 기사에 관한 의견들을 쉽게 교환할 수 있게 되었다. 또한, 기사에 대한 의견들은 해당 기사의 대상에 대해 잘 모르는 사람에게 대상의 특징을 빠르게 파악할 수 있는 중요한 정보로 이용되고 있다. 하지만, 너무 많은 사용자의 의견이 혼합되어 있기 때문에 대상에 대한 올바른 정보만을 선택하여 대상의 특징을 올바르게 파악하기가 어려운 상황이다. 따라서 본 논문에서는 해당 기사에 관한 사용자 의견들을 수집하는 방법으로 Opinion Mining 기술을 활용하여 해당 대상의 특징을 올바르게 파악할 수 있는 시스템을 제안한다.

본 논문의 구성은 2 장에서는 Opinion Mining 의 배경지식을 소개하고, 3 장에서는 본 논문에서 제안하는 방법에 대해서 설명한다. 4 장에서는 제안한 결과와 향후 연구에 대해서 언급한다.

2. 배경 지식

본 연구는 POS tagging 과 Association Rule Mining 을 사용하였다.

2.1 POS Tagging

POS tagging 은 자연어 처리 기법을 사용하여 한 문장을 입력받아 그 문장을 분석하여 단어에 명사, 동사, 형용사 등 품사를 저장하는 기법이다. 이 기법을 이용하여 문장의 구성 요소를 파악할 수 있다. 예를 들면 “When I watched this film, I hoped it ended as soon as possible” 라는 문장이 parser 로 입력되면, 다음과 같이 출력된다. “When/WRB I/PRP watched/VBD this/DT film,/NN I/PRP hoped/VBD it/PRP ended/VBD as/IN sooon/NN as/IN possible/JJ.” POS tagging 를 통한 Opinion mining 적용사례는 [3]에서 찾아볼 수 있다.

2.2 Association Rule Mining

Association Rule Mining[1]은 $X \Rightarrow Y$ 로 나타내며 X 는 $Y \subset I$ 와 $X \cap Y = \phi$ 이다. 여기서 rule 을 찾아내기 위해선 두 가지 threshold 를 사용한다. 첫 번째는 Minimum support 로 $(X \cup Y)/N$ 으로 표현한다. 즉, 전체 트랜잭션 중 X 와 Y 를 모두 가지는 트랜잭션의 개수를 나타낸다. 두 번째는 minimum confidence 로 $sup(X \cap Y)/sup(X)$로 표현한다. 즉, X 를 포함하는 트랜잭션 중 X 와 Y 를 모두 포함하는 트랜잭션의 개수를 나타낸다. Association Rule Mining 은 기본적으로 두 가지 단계로 이루어져 있다. 첫 번째 단계에서는 트랜잭션내에 존재하는 itemset 중 frequent 한 것들을 추출해내는 단계로 frequent itemset 이란 사용자가 정의해 놓은 minimum support 를 만족하게 하는 itemset 이다. 다음 단계에서는 이전 단계에서 찾아낸 frequent itemset 을 가지고 rule 을 만들어 낸다. 표 1 은 Association Rule Mining 의 예제이며 minimum support 는 0.4 이다. Minimum support 를 만족하는 룰은 {milk, bread},{bread, butter}가 된다. 본 논문에서는 frequent itemset 을 찾는 단계까지만 이용된다.

표 1 트랜잭션 데이터

Transaction ID	milk	bread	butter	Beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

3. 의견 추출

이번 장에서는 신문 기사에서 사용자의 의견을 추출하는 방법을 설명한다.

3.1 신문기사에서 Feature 추출

신문기사에 대한 reply 는 해당 신문과 관련 없는 정보를 가지고 있을 가능성이 있다. 따라서 우리는 reply 에서 feature 를 찾는 방법이 아닌 신문기사에 존재하는 고유 명사를 feature 로 저장하는 방법을 사용한다. 그림 1 과 같이 신문 기사를 전송 받은 후 기사의 문단을 각 문장을 따로 POS tagging 을 적용한다. POS tagging 을 통해 나온 결과 중 고유 명사를 feature 로 지정하여 저장한다. 그림 1 의 결과로는 "피겨 요정, 김연아, 네티즌, 유튜브, 미국, NBC" 등이 feature 로 저장된다.

'피겨 요정' 김연아의 영어실력이 네티즌들의 이목을 사로잡고 있다. 동영상 사이트 유튜브 메인 화면에 김연아가 미국 NBC 방송과의 영어 인터뷰 방송이 공개되면서다. 2분 7초 길이의 영상은 김연아가 앵커의 질문에 대해 유창한 영어로 대답하는 장면을 담았다.

S1 '피겨 요정' 김연아의 영어실력이 네티즌들의 이목을 사로잡고 있다.
S2 동영상 사이트 유튜브 메인 화면에 김연아가 미국 NBC 방송과의 영어 인터뷰 방송이 공개되면서다.
S3 2분 7초 길이의 영상은 김연아가 앵커의 질문에 대해 유창한 영어로 대답하는 장면을 담았다

그림 1

3.2 Reply 에서 후보 문장 추출

신문기사에 대한 Reply 를 모두 모은다. 그 후 3.1 과 같은 방식으로 문장을 따로 저장하여 POS tagging 을 적용시킨다. 사람들은 어떠한 사물에 대한 의견을 표현할 때는 대부분을 형용사로 표현하기 때문에 우리는 형용사가 존재하는 모든 문장을 후보 Opinion 을 갖은 문장으로 저장한다. 예를 들어 "김연아 너무 예쁘다.", "어떤 팬들의 매너가 너무 안 좋았다."라는 reply 는 "예쁘다"와 "안 좋았다"라는 형용사가 존재하기 때문에 후보문장으로 들어가게 된다.

3.3 후보 문장에서 Opinion 추출

3.2 절에서 생성된 후보 문장들과 3.1 절에서 생성된 feature list 를 이용하여 Opinion 을 추출하는 단계이다. 3.2 절에서 생성된 후보 문장에서 후보 feature 를 먼저 찾게 된다. Feature 는 후보 문장의 형용사를 찾아내고 나서 가장 가까이 있는 명사를 후보 feature 로 저장한다. 다음 이렇게 저장된 후보 feature 들과 3.1 절에서 생성된 feature list 를 비교하여 feature list 에 같은 명사가 존재하면 feature 와 opinion 을 저장하게 된다. 예를 들면 "김연아 너무 예쁘다"는 "예쁘다"라는 형용사와 가장 가까운 명사인 김연아는 feature list 내에 존재 여부를 확인하게 되고 김연아라는 단어는 feature list 내에 존재함으로 그 feature 에 대한 Opinion 으로 저장하게 된다. 반면 "어떤 팬들의 매너가 너무 안 좋았다"라는 후보 문장은 어떤 팬들의 매너가 list 에 존재하지 않으므로 feature 로 저장되지 않는다.

3.4 Association Rule Mining 을 통한 Opinion 추출

3.3 절을 통해 Feature 와 Opinion 이 추출되었다. 하지만, 3.3 절에서 추출되지 않은 문장 중 의미가 있는 문장이 존재할 수 있다. 그래서 우리는 Association Rule Mining 을 통해 3.3 절에서 버려진 문장들을 중 의미가 있는 Opinion 과 Feature 를 찾아낸다. 문장들을 <명사, 형용사>와 같은 방법으로 분류하고 나서 이것을 Association Rule 을 통해서 Rule 을 찾아내고 나서 이전 단계에서 만들어진 Feature list 에 새로운 feature 와 Opinion 을 추가한다.

3.5 의견 요약

본 장에서는 앞에서 찾아낸 feature 와 그와 관련된 opinion 을 장점/긍정과 단점/부정으로 나타내게 된다. Feature 는 같은 단어가 많이 쓰이는 것이 일반적이나 opinion 은 같은 표현이라 하여도 많은 다른 단어를 사용할 수 있다. 예를 들면 "멋진 스케이팅", "환상적인 스케이팅"는 두 문장 모두 스케이팅을 긍정적으로 표현한 것이다. 이처럼 같은 긍정도 다른 단어로 표현할 수 있다. 그래서 우리는 우선 기본이 되는 단어를 뽑고 그와 동의어를 저장하는 방식으로 단어 사전을 구축하고 나서 opinion 단어가 속한 그룹을 찾아 의견의 orientation 을 찾아낸다.

4. 결과 및 향후 연구

본 논문은 신문기사에 대한 Reply 를 대상으로 사용자들의 의견을 추출하고 요약하는 방법을 제안하였다. 이런 방법을 제안함으로써 신문기사의 내용을 정독하지 않고도 해당 기사 대상의 특징을 빠르고 올바르게 파악할 수가 있게 된다.

현재 본 논문에서 제시한 방법은 한글에 적용하기는 조금 부적합하다. 그 이유는 아직 WordNET[4]같이 단어의 semantic 을 중심으로 구축된 단어 사전이 존재하지 않을뿐더러 아직 한글의 POS tagging 을 사용한 문장 분석에 어려움을 많이 겪고 있기 때문이다. 이처럼 앞으로는 자연어 기법에 대한 연구가 더욱 많이 필요하고 WordNET 같은 단어 사전을 구축해 나갈 필요가 있다.

감사의 글

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

참고문헌

[1] Agrawal, R. and Srikant, R. 1994. "Fast algorithm for mining association rules." VLDB'94, 1994
[2] Minqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews", KDD'04, Washington, 2004
[3] Lipika Dey and SK Mirajul Haque. "Opinion mining from noisy data," Proceedings of 2nd workshop on Analytics for noisy unstructured text data, Singapole. pp 83-90, 2008
[4] http://wordnet.princeton.edu/