자연어 처리 기법을 이용한 상품평 분석에 관한 연구

Analyzing Product Reviews by Consumers using Natural Language Processing Techniques

전소은, So Eun Jeon*, 이영구, Young-gu Lee**, 박경철, Kyeong Cheol Park**, 백우진, Woojin Paik**

요약 많은 소비자는 특히 온라인상에서 물건을 구입하고 그 물건에 대하여 자신이 좋아하는 점이나 싫어하는 것을 포함하는 평가를 웹에 올린다. 이 평가를 분석하여 소비자가 물건을 구매할 때 무엇에 관심을 가지고 중요하게 여기는가에 대하여 알 수있다. 에를 들어 노트북을 구매할 때 작성된 평가를 분석하면 어떤 기능이 중요한 구매 결정 요소이며 어떤 것들을 고쳐야 하는지에 대하여 알 수 있다. 하지만 많은 양의 자료를 수동으로 분석하기에는 시간이 많이 걸린다. 따라서 대량의 자료를 쉽게 분석할 수 있는 두 개의 자연어처리 기법을 이용한 자동 분석 방법을 구현하였다. 두 가지 방법은 자동 문서 분류와 자동 정보 추출이다. 네이버 정보 포털에 있는 상품평을 개발한 시스템으로 분석하였고 평가 결과를 도출했다. 자동 분석시스템의 정확율과 재현율 측면에서 유사한 시스템이 다른 자료유형 분석에 적용했을 때와 비교하여 비슷하였다.

Abstract Consumers express how they evaluate what they purchased by writing reviews especially when they purchased products online. By analyzing the reviews about a product, it will be possible to find out what the consumers liked and disliked about the product. It will be also possible to identify the general consensus on what matters in purchaing certain product type such as a laptop if many reviews about many instances of a particular product type is analyzed. However, it takes a lot of time to manually analyzing the reviews. Thus, we propose to use two natural language processing oriented computational techniques to analyze a large number of reviews. The techniques are text classification and information extraction. We developed 무 review analysis system and conducted experiments against the reviews about the laptop computers posted on the Naver information portal.

핵심어: Natiral Language Processing, Text Classification, Information Extraction, Discourse Analysis: 자연어 처리, 문서 분류, 정보 추출, 담론 분석

*주저자 : 건국대학교 충주캠퍼스 컴퓨터학과 대학원생 e-mail: wjsthdms@kku.ac.kr

**공동저자 : 건국대학교 충주캠퍼스 컴퓨터학과 학부생 e-mail: sniper209@nate.com, depose@nate.com

***교신저자 : 건국대학교 충주캠퍼스 컴퓨터학과 교수; e-mail: wjpaik@kku.ac.kr

1. 서론

인터넷은 일반 이용자의 경험과 의견을 조사하고 이해할 수 있는 전례 없는 기회를 만들었다. 심각한 위기 상황이나

새로운 제품을 소개하는 것과 같은 일들이 발생했을 때, 이런 사건들에 관한 여론을 수집하고 요약하는 것은 중요하다. 이런 여론 수집기능은 새로운 제품을 개발하거나 서비스를 제공하고자 하는 기업이나 개인들이 제품/서비스 사용자의 반응을 이해하여 제품이나 서비스를 개선하는데 쓰일 수 있다. 이 연구에서는 상품평을 자동으로 분석하는 정보추출, 활용적인(performative) 정보추출, 대화형 문체분석 (text discourse analysis)과 같은 자연언어처리 기법들이 적용된 시스템의 개발과 평가에 대한 결과를 보고한다.

양적 조사 방법은 일반 여론을 측정하는 데 널리 이용되어 왔다. 그러나 설문을 개발하고, 데이터를 모으고, 결과를 분석하는 작업은 많은 경제적, 시간적 비용을 요구하기 마련이다. 또한, 일반여론을 적시에 평가하는 것도 어려운 일이다. 이러한 문제는 설문이 이루어진 시점과 분석결과를 이용할 수 있는 시점의 차이에서 비롯된다. 더욱 중요한 것은 어떤 유형의 문제들은 설문방법으로써 직접 일반인들에게 물어볼 수 없다는 것이다. 이러한 문제는 응답자들이 길고 상세한 설문을 기피하거나 이를 사생활 침해라고 여기기 때문이다. 또한, 설문작업을 이용하기에는 부적절한 유형의 문제들도 있다. 구매자들이 물건을 구매한 사이트에 올리는 상품평은 설문조사의 과정을 거치지 않으나 정형화되지 않은 정보가 포함되어 분석에 어려움이 특히 시간적 과점에서 크다.

본 연구에서는 상품평 분석 시스템을 개발하고 개발된 시스템의 효과성을 하나의 주제 영역을 대상으로 평가했다. 선정된 주제영역은 노트북이다. 노트북은 웹에서 많이 팔리는 제품이고 고가이기 때문에 많은 상품평이 올라온다. 따라서 쉽게 자료를 구할 수 있어 주제영역으로 선정되었다.

2. 자동 상품평 평가 시스템

본 연구의 첫 번째 단계는 상품평에 대한 활용적 (performative) 정보 모델을 개발하는 것이다. 두 번째 단계는 대화형 문체분석 (text discourse analysis) 모델의 개발이다. 세 번째 단계는 연습 데이터와 실험 데이터를 생성하는 것이다. 기본적인 정보추출기법을 대화형 문체에 맞게 만드는 단계는 자체적으로 개발된 분석시스템과 언어학적 분석시스템을 이용하여 수행하였다 [1].

문헌 분류를 위한 자연 언어 분석 증거 조합 방법 개발 단계는 이 연구에서 가장 핵심적인 부분이다. 지금까지 어형론적으로나 구문론적 의미론적인 주석달기로부터 나온 다양한 자연언어처리기법을 어떻게 조합하여 최적의 문헌 분류를 할 것인지에 대한 연구는 많이 없었다. 전통적으로 단어가 문헌 분류의 기본으로 사용되었다. 본 연구에서는 문헌 내에서 모든 단일어의 발생빈도를 가지고 그 문헌의 자질 벡터를 생성하고 이러한 자질 벡터를 나이브 베이지안 확률 분류기 (Na ve Bayuesian Probabilistic Classifier) 이용하여 구현하였다 [2].

나이브 베이지안(Na ve Bayesian)학습 기법은 신경망(Neural Network) 또는, 결정 트리(Decision Tree)와 같은 알고리즘 비교연구에서 전자 뉴스기사, 전자편지와 같은 텍스트 문서 분류를 위한 방법으로 나이브 베이지안 학습기법이 가장 효과적인 알고리즘들 중에 하나라고 알려져 있다. 나이브 베이지안 분류 학습은 기법은 베이지안 네트워크를 분류기에 적용한 것으로 베이즈 정리(Bayes Theorem)에 기초한 확률 모델을 이용하는 기법이다. 그림 1 은 베이지언 네트워크를 보여주는데 노드 C 는 클래스를 의미 하고 노드 C 에 대한 각각의 특성(feature)을 wi 라 표시했다. 그림 1 은 각각의 특성들 간의 서로 조건부 독립(conditionally independent)이라는 나이브 베이지안 학습기법을 보여준다.

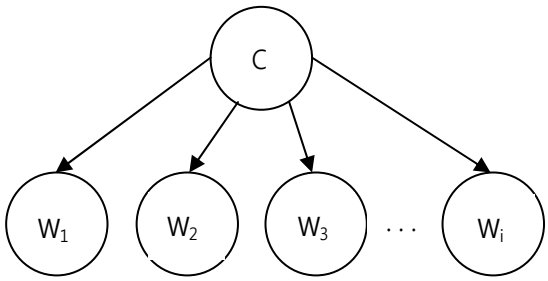

그림 1. 베이지안 네트워크

하나의 문서 d 가 $w_1,\ldots,w_n$의 특성들로 이루어 졌을 때 베이지안 학습기법은 식 (1)과 같이 문서 d 에 대한 조건부 확률이 가장 큰 클래스로 분류한다 [5].

$$\arg\max P(c|d) = \arg\max P(c|w_1,w_2,\cdots,w_n) \quad (1)$$

다음 단계에서는 활용적 (performative) 정보 추출 알고리즘을 개발하는 것이다. 이 단계에서는 전 단계에서 결정된 자연 언어 분석 기반 증거 조합 방법을 활용적 (performative) 정보 추출에 어떻게 사용할 것인지 결정한다. 다음 단계에서 대화형 문체분석 (text discourse analysis) 모델링 알고리즘을 개발했다 이 단계에서는 대화 범주와 의미론적 관계의 형식 안에서 대화형 문체 (text discourse) 구조를 발견하기 위한 방법을 이용하여 시스템을 개발하였다 [3].

정보 추출은 사용자가 원하는 정보를 검색하거나 추출하는 연구 분야로서 대용량의 정보자료들로부터 미리 정의된 주제나 관심분야의 정보만을 인식하여 요약된 형태로 가공을 한다. 정보 추출 시스템은 자연언어 처리기술을 바탕으로 문서의 내용을 분석하여 주어진 주제영역에서 유용한 정보들을 데이터베이스와 같이 구조화된 형식으로 저장한다. 정보 추출 시스템은 텍스트에서 적합한 부분을 효과적으로 찾아서 추가적인 처리를 한다.

정보 추출 단계는 도메인과 관련이 있는 엔티티를 인식하는 것으로 전체적인 정보 추출 시스템의 구조 중 첫 번째로 도메인에 종속적인 부분이다. 본 연구에서는 정보 추출을 위한 도메인을 알고 있기 때문에 각 문장별로

어떠한 정보를 추출해야 되는지 알고 있다. 예를 컴퓨터 기자재에 관한 상품평에서는 기자재의 이름이 가장 중요한 추출할 정보이고 구매자가 불망이 있을 때 무엇에 관하여 불만을 가지고 있는가 하는 것이 정보 추출의 목적이 것이다. 본 연구에서는 대화형 문체분석 (text discourse analysis) 모델링에 기반을 둔 정보 추출을 하는데 이는 많은 상품평이 사람간의 대화와 같은 형태를 취하기 때문이다.

문장분석 결과를 이용하여 정보를 추출하는 단계에서는 패턴일치 기법(pattern matching technique)이 사용되었다. 특정 주제에 관한 정보를 추출하기 위한 추출 패턴(extraction pattern)들은 상품평으로부터 자동으로 학습되는데 추출 패턴들은 정확한 정보들을 추출할 수 있도록 하기 위해 범용성(generality)과 부적합한 정보의 추출을 방지할수 있는 구체성(specialty)을 겸비하도록 하였다. 이러한 목적으로 말뭉치 기반 통계적 기법을 사용하였다 [6].

마지막 단계인 상품평 요약/보고서 생성 알고리즘 개발에서 특정한 상품에 대한 모든 추출된 정보를 종합하여 보고서를 생성하는 시스템을 개발하였다.

평가 단계에서는 기본 정보 추출, 활용적 (performative) 정보 추출, 대화형 문체 (text discourse) 구조 모델링의 효과성 평가를 하였다. 미리 수집한 실험 데이터를 이용하여 전반적인 시스템의 효과성 평가를 하였다.

3. 상품평과 감정 분류 모델

상품평의 예는 다음과 같다. “처음엔 옥션을 통해 노트북을 사기가 조금은 망설여 졌는데, 생각보다 빠른 배송과 아주 좋은 상품질에 만족입니다!^^ 포장도 나름 깔끔했고, 본 제품도 사진에서 보는 것처럼 귀엽고, 성능도 괜찮은 것 같습니다! 그냥 간단한 일처리 하기에는 적당한 노트북이 아닌가 싶습니다.”

상품평에 있는 각 문장의 의 수동 분석에 의하여 개발된 감정 분류 모델은 매우 긍적적, 긍적적, 보통, 부정적, 매우 부정적의 다섯개의 부류로 이루어졌다. 각 부류에 대한 설명은 다음과 같다.

매우 긍정적은 ‘〈vp〉’ 즉 ‘very positive’로 표시가 되는데 상품평을 구성하는 문장에 ‘훌륭하다’, ‘우수하다’, ‘잘’, ‘좋다’, ‘발휘’, ‘추천’, ‘친절’, ‘기분’, ‘역시’, ‘참’, ‘대만족’, ‘강추’, ‘적극’, ‘부럽다’, ‘감사’, ‘놀랍다’, ‘깔끔’, ‘최고’, ‘시원시원’, ‘고급’, ‘최상’, ‘완벽’, ‘매력’과 같은 단어가 있는 경우에 그 문장은 매우 긍정적으로 부류되는 경우가 많다. 다음 문장들은 매우 긍정의 예이다. ‘음...디자인 부터 시작해서 맘에 드네요ㅋㅋ 메모리도 괜챠낳고 ㅋㅋ^^’. ‘상품 잘받았고요 포장 뜯고나서 너무 좋았습니다. ㅋㅋ’. ‘사업부가 많아 다르게 구매했는데 상당히 마음에 듭니다.’ ‘주로 동영상, 워드작업용으로 구매했는데 정말 크기가 딱 적당합니다.’

긍정적은 ‘〈p〉’ 즉 ‘positive’로 표시가 되는데 상품평을 구성하는 문장에 ‘그리’, ‘괜찮다’, ‘맘(마음)에 든다’, ‘부드럽다’, ‘빠르다’, ‘만족’, ‘반가운’, ‘넓다’, ‘쓸만하다’, ‘무난하다’, ‘이만한’, ‘튼튼’, ‘견고’, ‘저렴’, ‘싸다’, ‘안정’, ‘선명하다’, ‘즐겁다’와 같은 단어가 있는 경우에 그 문장은 긍정적으로 부류되는 경우가 많다. 다음 문장들은 긍정의 예이다. ‘아무튼 일단은 만족이고 추가적으로 사용하면서 봐야겠네요’. ‘그냥 간단한 일처리 하기에는 적당한 노트북이 아닌가 싶습니다’. ‘들고 다니다 보니 생각보다 어깨에 무리가 가는 듯 하지만...그래도 와이브로 깔고 나니 인터넷도 잘 되고 좋네요~’. ‘저의 첫 넷북이었는데 ASUS 와 DELL 제품을 놓고 무척이나 고민했지만 LCD 화면의 크기 / 하드용량 / 자판 / 무게 등을 고려해서 이 제품을 골랐습니다.’

보통은 ‘〈u〉’ 즉 ‘usual’로 표시가 되는데 상품평을 구성하는 문장에 ‘별로’, ‘고민’, ‘모르겠어요’, ‘상관없다’와 같은 단어가 있는 경우에 그 문장은 보통으로 부류되는 경우가 많다. 다음 문장들은 보통의 예이다. ‘하루좀 늦기는 했지만, 실수 때문에 그런거니 괜찮습니다.’ ‘MSI 가 많이 알려지지 않아서 잘 모르긴 하지만...’ ‘전반적으로 가격경쟁력과 실용성을 갖춘 본격 넷북의 초기형으로 그런대로 무난한 편입니다.’ ‘핑크 있었음 더 좋았을껄..ㅠㅠ’.

부정은 ‘〈n〉’ 즉 ‘negative’로 표시가 되는데 상품평을 구성하는 문장에 ‘투박한’, ‘불가’, ‘해오름’, ‘아닌지’, ‘맘에 안 드네요’, ‘아쉬운 점’, ‘없다’, ‘글쎄다’, ‘감수’, ‘부담’, ‘걱정’, ‘맛이 갔다’, ‘비싸다’, ‘흐리멍텅’, ‘망설이다’, ‘약간’, ‘수긍’, ‘조심’, ‘어둡다’, ‘각오’, ‘실망’, ‘힘들다’, ‘거슬리다’, ‘참다’, ‘뿌옇다’, ‘좁다’, ‘우려’, ‘걱정’, ‘어렵다’, ‘부담’, ‘부족’, ‘부실하다’와 같은 단어가 있는 경우에 그 문장은 부정으로 부류되는 경우가 많다. 다음 문장들은 부정의 예이다. '배송이 쪼금 느렸구요’. ‘하드가 정확시 160 기가가 아니라서 좀 난감해 했습니다만 쓰는데는 별 문제’. ‘메뉴얼이 웹사이트에 있다면, 다운받을 웹사이트 주소도 같이 알려주시면 더 좋을껴 같아요.’ ‘근데 그래픽이 생각보다 별로인 듯.’

매우 부정은 ‘〈vn〉’ 즉 ‘very negative’로 표시가 되는데 상품평을 구성하는 문장에 ‘아직도’, ‘취소’, ‘오류’, ‘자주’, ‘짜증’, ‘영’, ‘짱나는’, ‘한참’, ‘지루’, ‘찜찜하다’, ‘후회’, ‘비추’, ‘황당’, ‘화나다’, ‘퍽’, ‘피해’, ‘심하다’, ‘거짓’, ‘무성의’, ‘치밀다’, ‘강제’, ‘불만족’, ‘불편’, ‘무리다’, ‘환불’, ‘열 받다’, ‘심하다’와 같은 단어가 있는 경우에 그 문장은 매우 부정으로 부류되는 경우가 많다. 다음 문장들은 매우

부정의 예이다. '급해서 오전내내 일도 못하고 웨이코스트에서 하자확인서까지 직접 받고 다시 판매처인 엔씨디지텍까지 제품반납까지 했는데, 아무런 이야기도 없이 약속한 시간에 왜 입금을 안시킨 것인지 이해가 안갑니다.' '소형이라는 점을 감안하더라도 음량이 너무 딸립니다.' '휴대의 편리성때문에 구입한제품인데 초기부터 오류가있어서 한번 return 시켜서 정상화가되었지만 그동안 불편했던것 생각하면 짜증납니다'.

4. 실험 결과 및 결론

본 연구에서는 200 개의 상품평을 수동 분석하여 분류 모델을 만들었다. 수동 분석 결과에 의하면 상품평은 감정측면에서 매우 긍적, 긍정, 부정, 매우 부정, 기타의 5 개 부류를 가진 모델로 만들어 졌다. 내용 측면에서는 가격, 무게, 성능(속도), 화면 사이즈, 배터리 시간, 디자인, 배송의 7 개의 부류의 모델이 도출되었다.

추가로 400 개의 상품평을 수동으로 분석하여 각 문장에 하나 이상의 감정 부류와 내용 부류를 부여하였다. 처음 20 개의 상품평을 대상으로 단일어 추출을 하고 나이브 베이지안 확률 분류기를 이용하여 자동 문서 분류 시스템을 구현하였다. 즉 자동 문서 분류 시스템은 각 문장을 분석하여 하나 이상의 감정 부류와 내용 부류를 부여한다. 이 시스템을 추가로 수동 분석한 400 개의 상품평에 적용하여 정확도를 산정하였다.

정확율은 맞게 하나의 부류로 분류된 상품평의 수를 시스템이 그 부류로 분류된 총 상품수로 나눈 것이다. 재현율은 맞게 하나의 부류로 분류된 상품평의 수를 그 부류로 분류되어야 하는 총 상품평 수로 나눈 것이다 [4]. 전체적인 평가결과는 정확율이 82%였고 재현율이 78%였다.

본 연구에서 개발된 시스템은 웹을 지속적으로 모니터하여 구매자의 의견 요약을 가능하게 할 것이다. 이 시스템의 중요기능인 자연언어처리 모듈은 개인들이 의사소통 시에 주고받는 메시지에 함축되어 있는 의미와 견해들을 보다 정확하게 자동적으로 해석하도록 할 것이다.

현재 여론조사 데이터가 많은 기관들에 의해 이용되는 방식과 마찬가지로, 자동적으로 생성되는 상품평 요약 자료는 제품이나 서비스 제공자가 자신의 제품이나 서비스에 대한 많은 양의 여론 데이터를 효율적으로 접근하고 이해하도록 할 것이다. 이를 토대로 제품이나 서비스 제공자가 일반인들의 제품이나 서비스에 관한 잘못된 인식을 바로잡는 새로운 정보를 전파하는 것을 가능하게 할 것이며 궁극적으로 제품이나 서비스의 이용을 높일 수 있을 것이다.

참고문헌

[1] W. Paik and J. Lee, "From whom, about what, and reason why: Capturing public's perception using Natural Language Processing (NLP)" Proceedings of AAAI Fall Symposium Series: Intent Inference for Users, Teams, and Adversaries. North Falmouth, MA. AAAI Press, Nov. 15, 2002.

[2] S.Kotsiantis, P. Pintelas, "Increasing the Classification Accuracy of Simple Bayesian Classifier" Lecture Notes in Artificial Intelligence, AIMSA 2004, Springer-Verlag Vol 3192, pp. 198-207, 2004.

[3] W. Paik, S. Harwell, S. Yilmazel, E. Brown, M. Poulin, S. Dubon, and C. Amice, "Applying Natural Language Processing Based Metadata Extraction to Automatically Acquire User Preferences" Proceedings of the First International Conference on Knowledge Capture. Victoria, British Columbia, Canada. Oct 21, 2001

[4] M. Buckland and F. Gey, "The relationship between Recall and Precision" Journal of the American Society for Information Science, Vol 45.1, pp12-19, 1994.

[5] A. McCallum and K. Nigam, "A Comparison of Event Models for Na ve Bayes Text Classification", In AAAI-98 Workshop on Learning for Text Categorization, 1998.

[6] Weischedel, R., D. Ayuso, S. Boisen, H. Fox, T. Matsukawa, C. Papageorgiou, D. MacLaughlin, T. Sakai, H. J. H. Abe, Y. Miyamoto, and S. Miller, "BBN's PLUM Probabilistic Language-Understanding System", Proceedings, TIPSTER Text Program(Phase I), Morgan Kaufmann, pp.195-208, 1993.