특집 06

SNS에서 오피니언마이닝 연구

박경미 · 박호건 · 김형곤 · 고희동 (한국과학기술연구원)

목 차 »

1. 서 론
2. 오피니언마이닝
3. SNS에서 오피니언마이닝
4. SNS 기반 핫토픽 추출
5. 결 론

1. 서 론

새로운 환경인 최근 스마트 디바이스의 발달은 다양한 변화를 가져오고 있다. 특히, 언제 어디서나 소셜 네트워크 서비스(SNS: Social Network Service)에 접속할 수 있게 되자, SNS는 급속히 성장하게 되었고 사용자가 급증하였다. 실시간 영화나 공연 리뷰, 특정 제품 사용 후기 등 개인적인 오피니언이 트위터 등에서 다양하게 생산되고 있으며 스마트 시대와 맞물려 그 양과 속도가 점차 가속화되고 있다. 이처럼 SNS에서영화나 맛집 리뷰 및 제품 후기 등이 끊임없이 쏟아지고 있기 때문에 사용자들이 이처럼 방대한 데이터에서 원하는 정보를 찾는 작업은 점차 어려워지고 있으며 그 효율성 측면에서도 많은 문제를 발생시키고 있다. 따라서 SNS의 대량의 리뷰로부터 사용자가 원하는 정보를 빠르게 분석해 주고, 유의미한 정보를 지능적으로 유추해내는 오피니언 마이닝(opinion mining) 기술의 중요성은 그 어느 때보다도 커지고 있는 실정이다.

스마트폰을 활용하여 맛집이나 영화, 제품에 대한 다양한 실시간 리뷰들을 트위터, 인스턴트 메시지 보드 등과 같은 SNS에 작성할 수 있다. 예를 들어, “근처에 데이트 하기 좋은 음식점 없나요?”, “광화문 메드 포 갈릭 괜찮아요”처럼 질의 응답 형식이나 “정말 우리동네 수치”, “가격대비 완전 강추!!”처럼 댓글 형식 등으로 기술한다. 리뷰 대상에 대한 오피니언은 계속해서 변경될 수 있기 때문에, SNS의 텍스트를 활용하여 최신의 오피니언을 반영하는 것이 중요하다. 또한 사용자 의도를 파악한 리뷰 요약을 통해 제품 및 서비스 등에 대한 추천 시스템을 구축하는 것이 유용하다.

사용자가 특정 제품이나 서비스에 대해 전체적으로 부정적인 인상을 받았지만, 특정 평가요소에 대해서는 긍정적인 리뷰를 작성할 수 있다. 예를 들어, 사용자가 특정 맛집에 대해 전체적으로 부정적인 인상을 받았지만, 이 맛집은 양이 많고

주차하기 좋다고 리뷰를 작성할 수 있다. 따라서, 무엇에 대한 오피니언인지, 평가 대상과 오피니언을 정확하게 연결하는 것이 중요하다. 그러나, 아주 다양한 평가요소와 각각의 오피니언의 연결관계를 정확하게 인식하는 것은 쉽지 않다. 예를 들어, 맛집에 대한 리뷰에는 맛, 서비스, 분위기, 가격, 위생, 주차 등 여러 가지 세부 평가요소가 존재할 수 있고,디지털 카메라에 대해서는 렌즈, LCD, 메모리, 해상도, 동영상 촬영, 디자인 등의 다양한 평가요소가 있을 수 있다. 따라서, 사용자가 리뷰를 작성할 때 대상 전체에 대한 긍정/부정 평가뿐만 아니라 세부 평가요소에 대한 평가도 작성할 수 있다는 것을 감안하여 오피니언마이닝을 수행하는 것이 필요하다.

앞으로, 2장에서는 오피니언마이닝의 개념 및 기존연구에 대하여 알아보고, 3장에서는 SNS에서 오피니언마이닝 기술을 적용하여 관련 정보를 추출하는 연구 사례를 살펴본다. SNS에서 오피니언은 시간이 지남에 따라 오피니언 흐름의 변화가 일어날 수 있다. 특히 실시간으로 SNS의 이슈에 따라 영향을 받는다. 따라서 오피니언 분석과 밀접한 관련이 있는 핫토픽 문제를 4장에서 다루고 5장에서 결론을 맺는다.

2. 오피니언마이닝

오피니언마이닝은 사람들이 특정 제품 및 서비스를 좋아하거나 싫어하는 이유를 분석한다. 또한, 어떤 사안에 대해 여론이나 대중의 관심이 실시간으로 어떻게 변하는지 확인한다. 현재의 검색 방법은 질의어와 메타 데이터 간의 일치하는 단어를 찾아서 결과를 보여준다. 하지만, 사용자의 의도가 반영이 안돼 전혀 다른 결과를 내놓을 가능성이 높다. 사용자의 의도가 무엇인지 파악하여 질의를 처리하는 것이 필요하다. 그래서, 주어진 조건과 상황에 따라 그에 맞는 '추천'을 오피니언의 흐름을 반영하여 제시하는 것이 중요하다.

오피니언마이닝은 일반적으로 다음의 단계를 거치게 된다: (1) 긍정 및 부정을 표현하는 단어 정보를 추출하고, (2) 세부 평가요소와 그것이 가리키는 오피니언의 연결관계를 포함한 문장을 인식하고, (3) 긍정/부정 표현의 수 및 유용한 문장들을 추출하여 리뷰 요약을 생성한다. 각 단계에서 사용된 방법에 대한 세부적인 설명은 다음과 같다.

첫째로, 오피니언마이닝에서 긍정/부정 표현에 해당하는 어휘 정보를 추출하는 것은 중요하다. 기존에 구축된 사전 등의 리소스를 이용하거나 수작업을 통해서 해당 도메인의 고빈도 긍정/부정을 표현하는 단어들을 확인할 수 있다. 예를 들어, 기존의 WordNet에 오피니언 정보를 추가로 부착한 SentiWordNet[1]과 WordNet Affect[2]를 활용할 수 있다. 각 부착한 레이블이 갖는 값의 범위는 0.0~1.0이며 synset별로 점수의 총합은 1.0이다. 이러한 리소스들은 영어에 국한된 정보들로, 한국어에 대해서는 아직 활용할 만한 리소스가 존재하지 않는다.

또한, 학습 데이터에 대한 유용한 통계 정보를 활용하여 자동으로 어휘 정보를 얻을 수도 있다. 통계적인 방법을 적용하여 어휘 정보를 추출하는 경우, mutual information과 같은 평가척도를 사용할 수 있다. 예를 들어, 긍정/부정 오피니언을 주어진 단어와 'excellent' 사이의 mutual information에서, 주어진 단어와 'poor' 사이의 mutual information을 뺀 값으로 계산할 수 있다[4,5]. 그러나, 주어진 특정 도메인에 대해 긍정/부정 표현에 해당하는 단어들을 자동으로 생성하는 것은 쉽지가 않다.

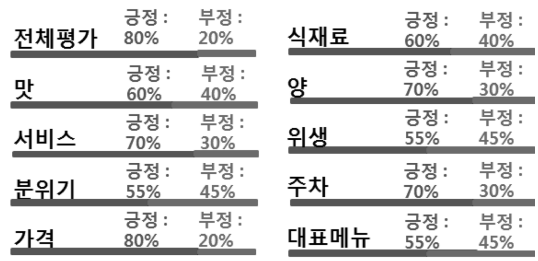
한국어에 대해 사용자 별점이 아주 높은 리뷰

에서 고빈도 단어를 긍정 표현으로, 사용자 별점이 아주 낮은 리뷰에서 고빈도 단어를 부정 표현으로 추출할 수 있다[3]. 즉, 사용자 별점 1-2점에서 자주 발생하는 서술어는 부정을 나타내고, 사용자 별점 4-5점에서 자주 발생하는 서술어는 긍정을 나타낸다고 간주할 수 있다. 한국어 서술어 긍정/부정 감정(positive/negative sentiment) 사전을 자동으로 구축하였을 때, 정확도는 약 80% 로 아주 높지는 않았다. 수작업을 최소화 하면서 유용한 어휘별 긍정/부정 감정 정보를 자동으로 생성하는 방법에 대한 연구가 계속해서 수행되어야 한다.

둘째로, 세부 평가요소와 오피니언으로 구성된 문장을 인식하는 것이 중요하다. 이 때 첫 번째 단계에서 구축된 어휘 정보를 사용하여 세부 평가요소와 긍정/부정 표현을 찾게 된다. 또한, 긍정적인 오피니언인지 부정적인 오피니언인지 문장 단위로 분류하기 위해서 여러 가지 방법을 적용할 수 있다. 규칙기반 방법과 통계기반 방법을 동시에 사용할 수 있다. 형용사를 오피니언 단어로 간주하고 긍정/부정을 결정하기 위하여 다양한 규칙 및 통계량을 활용한다. 또한, 대량의 레이블이 부착된 학습 데이터를 생성하여, Naive Bayes, Maximum Entropy(ME) model, Support Vector Machine (SVM)과 같은 알고리즘을 적용하여 기계학습을 수행한다.

셋째로, 긍정/부정 표현의 수 및 중요 문장을 추출하여 리뷰 요약을 생성하는 것이 필요하다. 각 세부 평가요소에 대한 긍정 표현과 부정 표현의 차를 통하여 사용자들의 선호도를 (그림 1)처럼 제시할 수 있다. 또한 세부 평가요소와 관련된 오피니언을 포함하는 문장들 중 유의미한 문장들을 긍정/부정 평가별로 추출하여 중요 문장으로 구성된 리뷰 요약을 생성할 수 있다.

오피니언마이닝의 결과는 긍정/부정 평가의 정도를 나타내거나 요약 형태로 제시될 수 있다. (그림 1)은 특정 맛집의 여러 평가요소에 대한 긍정/부정 표현의 비율을 나타낸다. 이러한 오피니언마이닝 결과를 통해 사람들이 그 맛집의 세부 평가요소에 대하여 좋아하거나 싫어하는 정도를 얻을 수 있다. <표 1>은 특정 맛집의 평가요소 '가격'에 대한 리뷰 요약의 예제를 나타낸다. 리뷰 요약은 사용자들의 리뷰를 대표하면서 유익한 정보를 제공해 줄 수 있어야 하고 읽기 편하고 길지 않아야 한다. 여기서 강조된 단어들은 평가요소 또는 오피니언에 속하는 단서 단어들이다.



(그림 1) 맛집의 평가요소들에 대한 리뷰 요약의 예

〈표 1〉 맛집의 평가요소 '가격'에 대한 리뷰 요약의 예

맛집의 평가요소-가격	
긍정평가	1. 순수함 그 자체육수도 서비스로 한사발 더 주신 주인 아저씨도 친절하시고 가격대비대만족^^ 2. 아줌마들이모임하기에 안성맞춤이고 가격도 적당하며 무엇보다 자극적이지 않은 음식…… 3. 분위기, 서비스, 가격 등 레스토랑으로서 손색이 없어 글을 올립니다……
부정평가	1. 굉장히 유명한 편인데—가격도 싼 편은 아닌데— 그에 비해 맛은 쏘쏘— 서비스는 중하 정도? 2. 가격만 비싸지고 맛은 조미료 범벅 어렸을 때 먹었던 그 맛이 그립네요 3. 가격대비 별로 먹을 것 정말 없구요.

2.1 연구 사례

오피니언마이닝과 관련한 기존 연구들은 다양한 도메인에 대해 리뷰 요약을 생성하였다[4,5]. 먼

저, 영화 도메인에서 리뷰 요약을 추출하였다[4]. 영화에 대한 세부 평가요소로 감독, 작가, 배우, 음악 등을 정의하였고 그 수는 총 12개이다. 문장 단위로 평가요소가 가리키는 오피니언을 찾아 영화의 각 평가요소 별로 긍정 표현과 부정 표현의 수를 계산하고 평가요소와 그것의 오피니언의 연결관계를 포함한 문장들을 제시하였다. 또한 영화 도메인 이외에, 음식점 및 호텔 등 2개의 다른 도메인에서 리뷰 요약을 생성할 수 있다[5]. 이것은 세부 평가요소와 그것이 가리키는 오피니언의 연결관계로 구성된 문장들을 찾았다. 최종 리뷰 요약을 생성하기 위하여 문장 점수에 따른 순위를 이용하였다. 사전으로부터 각 단어가 긍정 표현일 가능성과 부정 표현일 가능성을 구하였다. 각 문장의 점수는 이러한 단어들의 긍정/부정 극성 점수의 합이다. 문장 점수가 큰 순으로 몇 개의 문장을 리뷰 요약으로 제시하였다.

신문기사 등을 사용하는 전통적인 텍스트 요약(text summarization)에서는 결과를 평가하기 위해서 수작업으로 만들어놓은 요약과 시스템이 생성한 요약을 비교하였다. Document Understanding Conference(DUC)[6], Text Summarization Challenge (TSC)[7], Text Analytics Conference(TAC)[8] 등의 shard task에서 정답 요약을 구축하여 ROUGE[9]와 같은 평가 프로그램을 적용하였다. 한편, 길이가 짧은 댓글들의 집합체인 제품 및 서비스 리뷰 요약(Review summarization)에서는 ROUGE를 사용하거나 수작업으로 성능을 평가하였다.

오피니언마이닝 시스템의 성능을 향상시키기 위하여 리뷰 요약의 결과를 적절하게 평가하는 것은 중요하다. 그러나, 명확한 정답이 존재하지 않기 때문에, 시스템이 리턴한 리뷰 요약의 결과가 좋은 것인지 결정하기는 쉽지가 않다. 그래서 리뷰 요약 자체가 아니라 두 가지 모델의 결과를 평가자에게 제시하고 더 선호하는 요약을 선택하게 하여 선호하는 정도를 시스템의 성능으로 제시하였다[10]. 한편, 평가자에게 더 어려울 수 있는 방법으로, 직접적으로 여러 명의 평가자를 통하여 1~5의 별점처럼 평가자가 요약 결과 자체에 대한 점수를 주어 평균 결과를 보이는 방법도 가능하다[11].

사용자 선호도가 높거나 낮은 리뷰 요약의 특징은 다음과 같다[10]. 사용자는 장점과 단점이 리스트 형태로 나타난 요약과 문법적으로 올바르게 작성된 요약을 선호하였다. 반면에 사용자는 긍정/부정 표현이 없는 요약, 좋다/나쁘다는 표현만 있는 요약, 사용자가 준 별점의 평균과는 대조적인 평가를 기술한 요약을 선호하지 않았다.

휴대폰과 디지털 카메라에 대한 영어 리뷰에 대하여, 사용자 별점과 제안한 시스템 평가 간의 상관 관계에 대하여 분석할 수 있다[10]. 실험 결과 Pearson 상관 계수(correlation coefficient) 0.76만큼의 관련성을 보였다. 이것은 제안했던 방법이 사용자 평점과 유사한 점수를 추정한다는 것을 나타낸다.

오피니언마이닝은 특정 제품 및 서비스에 대한 사실(fact)을 얻고자 하는 것이 아니기 때문에 오피니언이 없는 객관적인 문장은 분석 대상에서 제외하는 것이 필요하다. 전처리 단계로써 마이닝을 수행하기 전에 먼저 각 문장이 객관적인 문장인지 주관적인 문장인지 파악할 필요가 있다[11]. 주관적인 문장들을 찾아내고 그 안에서 긍정/부정 표현의 정도를 분석한다. 이러한 방법을 통하여 오피니언을 포함할 가능성이 적은 문장을 제거함으로써 학습 비용을 낮출 수 있는 반면, 추가적인 단계의 수행으로 오류가 전파되는 문제가 있다.

한국어에 대하여 오피니언마이닝을 수행한 다양한 연구들이 있다[3,12]. 예를 들어, 맛집 리뷰로부터 평가요소-오피니언 연결관계를 포함한 문장

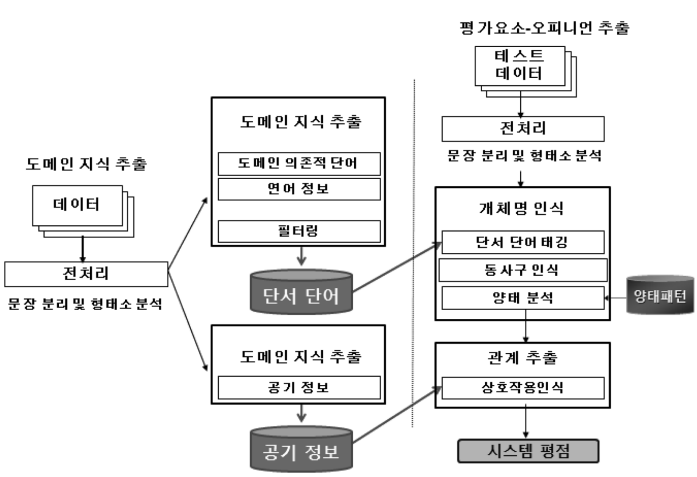
을 추출할 수 있다[12]. 맛집에 대한 세부 평가요소는 맛, 서비스, 분위기, 가격, 식재료, 양, 위생, 주차, 대표메뉴 등이다. 시스템에 대한 자세한 설명은 (그림 2)에서 주어진다.

(그림 2)는 도메인 지식 추출과 평가요소-오피니언 추출로 구분할 수 있다. 두 단계 모두 입력 데이터에 대하여 문장 분리 및 형태소 분석의 전처리를 수행한다. 도메인 지식 추출은 각 도메인의 유의미한 정보를 인식하는 단계로 여기서 추출된 정보는 평가요소-오피니언 추출에서 활용된다. 도메인 의존적 단어는 특정 단어가 주어진 도메인에서 발생한 빈도와 일반 도메인의 신문 기사에서 발생한 빈도를 비교하여 추출한다. 연어 정보는 likelihood ratio를 평가 척도로 활용하여 추출한다. 도메인 의존적 단어와 연어에 대한 수작업 필터링을 통하여 최종적으로 평가요소와 오피니언에 해당하는 단서 단어를 확인한다. 이 정보는 개체명 인식에서 단서 단어를 찾는데 활용된다. 또한, 평가요소와 오피니언간의 공기 정보를 추출하여 관계 추출에서 가장 적합한 평가요소-오피니언 연결관계를 찾는데 사용한다.

평가요소-오피니언 추출은 개체명 인식과 관계 추출로 구성된다. 여기서 개체명 인식은 평가요소 또는 오피니언에 해당하는 단어열을 인식하는 단계이고 관계 추출은 평가요소-오피니언 관계 중 관련성이 존재하는 연결관계만을 인식하는 단계이다. 개체명 인식에서는 정의된 지식을 활용하여 단서 단어 태깅을 수행하여 평가요소를 나타내는 단어열과 오피니언을 나타내는 단어열을 확인한다. 한국어 동사구에서 복잡한 어미 활용을 보완하기 위하여 동사구를 인식하고 본동사와 보조동사의 관계를 분석하여 양태분석을 수행한다. 양태분석을 위하여 정의된 패턴을 활용하여 긍정 또는 부정의 오피니언을 확인한다. 관계 추출에서는 가능한 평가요소-오피니언의 연결관계 후보들 중에서 공기 빈도에 기반하여 적합한 연결관계를 추출한다. 시스템 평점의 결과는 평가대상에 대하여 긍정 표현의 수와 부정 표현의 수로 요약될 수 있다.

3. SNS에서 오피니언마이닝

SNS를 사용하는 수많은 사람들은 일상생활의 다양한 측면에 대한 의견을 매일 공유한다[13].



(그림 2) 시스템 구성도의 예

SNS 웹사이트를 통하여 사용자는 무엇을 좋아하고 싫어하는지, 삶의 많은 양상에 대한 그들의 의견을 게재한다. 따라서 SNS 사이트는 오피니언 마이닝을 위한 풍부한 데이터 소스이다. 비교적 최근에 등장하였고, 가장 인기있는 마이크로 블로깅 플랫폼인 트위터를 사용하여 오피니언 분석이 수행되고 있다. 트위터로부터 자동으로 학습 데이터를 수집하고, 그 데이터에 대하여 자연어 분석을 수행하고, 그 결과를 통하여 오피니언을 발견한다. SNS는 사람들의 의견과 감정에 대한 소중한 자원이고 이러한 데이터들은 마케팅이나 사회 여론 분석에 효율적으로 사용될 수 있다.

오피니언마이닝을 위한 학습 데이터로 트위터를 사용하는 구체적인 이유는 다음과 같다[13]. 첫째로, 트위터는 여러 가지 다른 주제에 대한 자신의 의견을 표현하는 다양한 사람들이 존재한다는 점에서 상당히 가치 있는 정보 소스이다. 둘째로, 트위터는 엄청난 수의 텍스트 게시물을 포함하고, 매일 급속하게 성장한다는 것이다. 셋째로, 트위터 이용자는 매우 다양하여 서로 다른 사회적 관심 그룹에서 텍스트 게시물을 수집하는 것이 가능하다. 넷째로, 트위터 이용자는 많은 나라의 사용자들이기 때문에, 여러 언어의 데이터를 수집하는 것도 가능하다.

트위터에서 수작업 없이 자동으로 행복, 오락, 즐거움 같은 긍정적인 감정을 포함하는 텍스트 및 슬픔, 분노, 실망 같은 부정적인 감정을 포함하는 텍스트, 사실만을 나타내는 객관적인 텍스트를 수집하기 위하여 이모티콘을 많이 활용하였다[13,14]. 행복함(":-)", ":)", "=)", ":D") 또는 슬픔(":-(", ":(", "=(", ";(")을 나타내는 이모티콘을 활용하여 텍스트의 긍정적 또는 부정적 감정을 인식하였다. 이모티콘 기반의 감정 분류 성능은 70~80% 사이의 정확도를 보였다[13,14].

트위터는 매순간 엄청난 수의 사용자가 이용할 수 있기 때문에 긍정/부정 오피니언의 변화가 지속적으로 일어날 수 있다. 시간에 따라 변화하는 SNS상의 이슈도 오피니언을 결정하는데 중요한 역할을 한다. 현재의 이슈가 무엇이냐에 따라 오피니언의 흐름은 영향을 받기 때문이다. 오피니언마이닝과 밀접한 관련이 있고, 사용자들이 가장 많이 검색하는 용어인 핫토픽에 대한 내용은 4장에서 설명한다.

실시간 트위터 메시지를 통하여 오피니언마이닝을 수행할 수 있다. 이 경우 긍정/부정 오피니언 분류 모델은 제한된 학습 데이터 샘플로부터 생성된 모델을 적용하여 분류를 수행한다. 그러나, 리뷰 및 이슈에 대한 긍정/부정 오피니언의 정도는 시간이 지남에 따라 예기치 못한 방식으로 변경될 수 있다. 제한된 샘플 데이터를 이용하는 경우 긍정/부정 오피니언의 흐름을 정확하게 예측하는 능력이 감소할 수 있다. 따라서, 분류 모델의 학습 데이터가 항상 업데이트 되어서 최신의 정보로부터 오피니언마이닝을 수행하는 것이 필요하다[15].

3.1 연구 사례

트위터로부터 수작업 없이 긍정/부정의 오피니언을 표현하는 텍스트와 사실만을 나타내는 객관적인 텍스트를 수집하는 방법이 있다[13,14]. 텍스트 분류 과정은 수작업 없이 이모티콘 기반으로 이루어진다. 또한, 자연어 분석을 활용하여 학습 데이터로부터 유용한 통계량을 추출한다. 이러한 트위터 데이터는 자동으로 오피니언 분류를 위한 시스템의 학습 데이터로 사용된다.

트위터 데이터는 중국어에서 형용사의 긍정/부정 감정의 정도를 구별하는데 사용되었다[14]. 제안하는 방법은 트위터로부터 수집한 대량의 문맥 정보에 따라 형용사의 의미를 긍정적 또는 부정

적인 감정의 극성으로 자동으로 분류한다. 각 형용사가 긍정/부정을 표현하는 정도를 수치화함으로써 각 문장에 대한 오피니언은 각 형용사의 긍정/부정 극성의 가중치 합으로 추정할 수 있다.

대부분의 기존 연구들은 고정된 학습 데이터를 가지고, 테스트 데이터의 오피니언을 효과적으로 분석하는데 초점을 맞추었다. 제한된 트위터 샘플을 활용하여 적절한 분류 모델을 학습하는 방법에 집중하였다. 그러나, SNS의 실시간성을 적용하여 오피니언의 변화 흐름을 파악하는 것이 필요하다. 즉, 분류 모델은 새로운 데이터를 확인하고 업데이트 되어야 한다. 이와 관련해서 최신의 데이터에 기반하여 효율적으로 분류 모델을 업데이트하는 방법이 제안되었다[15]. 이것은 사용자 요구 중심의 학습 프레임을 만들고, 각 프레임을 구성하는 메시지를 분석한다. 각 프레임에 나타나는 질적 정보에 따라 해당 메시지는 자동으로 선택된다. 또한 학습 메시지의 기한을 정하여 학습 메시지가 오피니언 분류를 위하여 효용 가치가 없어지면 제거한다. 대량의 데이터에 기반한 신뢰할 만한 판단이 가능해질 때까지 오피니언 예측은 수행하지 않는다.

분류 모델은 새로운 학습 데이터의 정보로 통합되면서 자체 보강이 되는 방법을 사용한다[15]. 결과적으로 메시지 스트림을 추가하면 자동적으로 학습 데이터가 확대된다. 분류 모델은 점진적으로 최신의 규칙을 유지함으로써 즉시 업데이트된다. 그래서 메시지 스트림에서 다음 메시지는 최근에 포함된 정보를 잠재적으로 활용할 수 있게 된다. 견고한 메시지 스트림에 의존하면, 분류 모델은 일시적으로 작은 블록에 대하여 메시지를 만들어내면서 신뢰할 만한 예측을 수행할 수 있다. 현재 이용 가능한 학습 데이터만을 가지고 신뢰할 수 있는 예측을 수행할 수는 없는 반면에, 이러한 최신의 트위터 메시지들은 가장 최근에 얻어진 학습 정보의 효과를 볼 수 있다.

또한, 분류 모델을 실시간으로 업데이트하기 위하여, 추가된 각각의 메시지가 학습을 통하여 사용자 요구 중심의 예측을 수행할 수 있다. 학습 기반의 예측을 수행하기 위하여 분석 중인 주제에 따라 적절한 학습 메시지를 자동으로 선택한다. 이렇게 업데이트가 가능한 분류 모델은 학습 메시지로써 의미가 없는 메시지들을 자동으로 제거하기 때문에, 오피니언의 변화 흐름을 인식하는데 효과적이다. 대부분의 분류 모델은 엄청난 양의 트위터 데이터에 기반하기 때문에, 모델을 제안할 때는 대량의 의미 없는 정보가 학습 데이터에 존재하더라도 최신의 정보에 초점을 맞추어 효과적으로 오피니언 흐름의 변화를 인식할 수 있도록 설계되어야 한다.

오피니언마이닝을 수행할 때 리뷰 데이터 자체의 긍정/부정 표현의 정도뿐만 아니라, 신뢰할 만한 사용자가 작성한 것인지 결정하는 것도 중요하다[16]. 모든 리뷰를 동일한 선상에서 고려하는 것이 아니라 신뢰할 만한 사용자가 작성한 리뷰에 가중치를 부여하여 전체 오피니언의 결정에 더 영향을 발휘하도록 하는 것이다. 대표적인 온라인 소셜 네트워크 사이트인 Epinions (epinions.com), Slashdot(slashdot.org), Wikipedia (wikipedia.org) 등에서는 각각의 사용자 링크가 긍정인지 부정인지 고려하고 있다[16]. Epinions는 제품 리뷰 웹사이트로써 사용자 커뮤니티가 매우 활성화되어 있다. 사용자들은 신뢰와 불신의 네트워크로 연결되고 어떤 리뷰가 가장 권위가 있는지 결정한다. Slashdot는 기술 관련 뉴스 웹사이트이다. Slashdot는 2002년에 사용자가 서로를 “친구” 또는 ”원수”로 태그할 수 있는 Slashdot Zoo를 개발하였다. 친구 관계는 사용자가 다른 사용자의 코멘트를 좋아한다는 것을 의미하고, 원수 관계는 사용자가 다른 사용자의 코멘트를 흥미없어 하는

것을 의미한다. Wikipedia는 백과사전으로써 개인에게 관리자의 역할을 부여하기 위해, 선거에서 사용자에 의한 투표를 진행한다. +는 찬성표이고 −는 반대표로써, 사용자에 의해 긍정 또는 부정 투표를 나타낸다. 이처럼 개인 간의 긍정/부정의 링크는 각 개인의 신뢰도를 측정하는데 이용될 수 있고, 그 개인이 작성한 리뷰의 가치에도 영향을 미칠 수 있다.

4. SNS 기반 핫토픽 추출

정당, 기업 등 여러 단체는 현재 이슈에 관한 사람들의 의견을 확인하는 것이 중요하다. 따라서, 오피니언마이닝뿐만 아니라 SNS기반의 핫토픽 추출과 같은 다양한 정보 서비스가 개발되고 있다. 그 중 대표적인 것이 실시간으로 현재 트위터 상에서 이슈를 찾는 것이다. 사용자들은 다양한 메시지를 통하여, 다양한 주제에 대한 의견을 공유하고, 현재의 이슈를 논의한다. 예를 들어, 야구 경기가 진행되는 동안 트위터 상의 키워드 그래프에서 기하급수적으로 증가하는 피크가 있다면, 각 피크의 시작은 각 개별 핫토픽 이벤트의 시작에 해당한다[17]. 따라서 현재 트위터에서 가장 대표적인 용어를 찾을 수 있다. 그리고 SNS의 이슈는 관련이 있는 제품 및 서비스 리뷰의 오피니언에 영향을 미칠 수 있기 때문에 핫토픽 추출과 오피니언마이닝은 서로 밀접한 관련이 있고 핫토픽에 따라 오피니언의 흐름이 달라질 수 있다.

SNS는 정보 공유를 위한 필수적인 커뮤니케이션 플랫폼이 되고 있다. 특히, 실시간 이벤트를 다루는 트위터에서 실시간 이벤트를 이해하고 그 순간의 핫토픽을 발견하는 것이 필요하다. 또한, SNS에서 핫토픽 추출은 쉽게 하이라이트 영상을 만들 수 있게 한다. 예를 들어, 사용자가 다른 채널에서 야구 경기를 보고 있거나 그 순간에 자리에 없었을 때, 사용자는 홈런이나 득점같은 몇 가지 중요한 장면을 놓칠 수 있다. 그러나, SNS의 핫토픽을 통해서 어떤 이슈가 있는지 혹은 왜 사람들이 그런 방식으로 행동하는지를 이해할 수 있게 된다. 즉, 야구 경기 비디오를 가지고 있다면, 핫토픽 이벤트의 피크 기간을 통하여 이벤트가 발생한 기간의 야구 하이라이트 영상을 자동으로 얻을 수 있다. 이것을 통하여 사람들은 그 이슈가 확산될 가치가 있을 때는 적극적으로 참여하면서 다른 사람들과 현재의 이슈를 공유한다.

트위터뿐만 아니라 인스턴트 메시지는 실시간 화제가 되는 이슈를 다룰 가능성이 높기 때문에 실시간 이벤트의 하이라이트를 발췌하는 것이 가능하다. 트위터와 같은 소셜 네트워크, 인스턴트 메시지 보드 등의 소셜미디어는 누구나 정보를 게시할 수 있도록 접근가능한 도구이기 때문에, 수많은 실시간 소셜미디어콘텐츠를 생성하도록 한다. 그래서, 같은 이벤트를 다루는 소셜미디어의 집합은 현재 가장 대표적인 용어인 핫토픽을 찾는데 좋은 소스를 제공할 가능성이 높다. 이런 특징들은 해당 실시간 이벤트에 대하여 즉각적이지만 상당한 이해를 할 수 있도록 해준다. 또한, 이벤트에 대한 설명을 확장하는데 사용될 수 있다. 각 개인은 다른 앵글에서 혹은 다른 관점으로부터 사건을 볼 수 있다. 소셜미디어플랫폼은 더욱 쉬운 방식으로 다양한 스토리를 공유하고 간접적으로 경험하는 것을 용이하게 해주기 때문에, 이러한 전체 데이터는 그 순간의 풍부한 관점을 공급해줄 수 있다. 게다가, 모든 소셜미디어데이터는 단지 생성되기만 하는 것이 아니라 종종 다른 사람들에게 영향을 끼치기도 한다. 따라서 사람들이 많은 관심을 보이는 핫토픽은 SNS 상에서 오피니언 흐름의 변화를 이끌 수 있다.

5. 결 론

SNS는 오늘날 인터넷 사용자들 사이에서 아주 인기 있는 커뮤니케이션 도구가 되고 있다. 트위터, 페이스북처럼 인기 있는 웹사이트에는 매일 수백만 개의 메시지가 게재되고 있다. 인터넷 사용자들이 메시지들의 자유로운 형식과 SNS의 쉬운 접근성 때문에 전통적인 커뮤니케이션 도구인 블로그 등에서 SNS 서비스로 이동하고 있다. 점점 더 많은 사용자들이 자신들이 사용하는 제품과 서비스에 대해 댓글을 게재하거나 자신의 정치적, 종교적 견해를 표현할 것이다. 따라서, 더욱 방대해지는 데이터를 효율적으로 요약, 정리 및 추론을 하여 가시화 하여주는 데이터마이닝 시스템의 개발은 필수적이다. 그러나, 아직까지 데이터 마이닝의 전반적인 과정에서 볼 때, SNS에서 오피니언마이닝 관련 기술은 초기 단계에 머물러 있다. 이것은 SNS 관련 데이터에서의 마이닝 전반을 자동화하거나 기계학습과 같은 학습 알고리즘을 이용한 고급 마이닝기법에 관한 연구가 아직 미비하기 때문이다. 그러므로, 이 분야에 대한 지속적인 연구와 투자가 필요하고 그 결과로 SNS 분야의 유용한 시스템이 개발된다면, 이러한 오피니언마이닝 관련 기술은 마케팅을 비롯한 여러 분야에서 매우 유용하게 사용될 것이다. 또한, 이러한 오피니언마이닝 기술을 정보 과다로 정보습득의 효율성과 정보에 대한 접근에 많은 문제를 겪고 있는 SNS상의 리뷰 및 제품 후기에 활용하여 최신 정보를 습득하는 능률을 향상시키고 최신 정보의 효율적 제공을 가능하게 하는 것이 필요하다.

본 논문은 한국과학기술연구원 미래원천연구사업(Tangible 소셜 미디어 플랫폼 기술개발)의 일환으로 수행되었음.

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2011년도 콘텐츠산업기술지원사업의 연구결과로 수행되었음.(모바일 혼합현실 기반 체험투어 기술 개발)

참 고 문 헌

[1] A. Esuli and F. Sebastiani, “SentiWordNet: A publicly available lexical resource for opinion mining,” in Proceedings of Language Resources and Evaluation(LREC), 2006.

[2] C. Strapparava and A. Valitutti, “Wordnet-affect: an affective extension of wordnet,” In Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, 2004.

[3] 송종석, 이수원, “상품평 극성 분류를 위한 특징별 서술어 긍정/부정 사전 자동 구축,” 한국정보과학회논문지 B-소프트웨어 및 응용, Vol. 38, No. 3, pp. 157- 168, 2011.

[4] L. Zhuang, F. Jing, and X.Y. Zhu, “Movie review mining and summarization,” In Proceedings of the International Conference on Information and Knowledge Management (CIKM), 2006.

[5] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G.A. Reis, and J. Reynar, “Building a sentiment summarizer for local service reviews,” In WWW Workshop on NLP in the Information Explosion Era, 2008.

[6] http://duc.nist.gov/

[7] http://lr-www.pi.titech.ac.jp/tsc/index-en.html

[8] http://www.nist.gov/tac/

[9] C-Y. Lin and E.H. Hovy, “Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics.”In Proceedings of Language Technology Conference(HLT-NAACL), 2003.

[10] K. Lerman, S. Blair-Goldensohn, and R. McDonald, “Sentiment summarization: Evaluating and learning user preferences,” In Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), 2009.
[11] H. Nishikawa, T. Hasegawa, Y. Matsuo, and G. Kikui, “Opinion summarization with integer linear programming formulation for sentence extraction and ordering,” In COLING, 2010.
[12] Kyung-Mi Park, Hogun Park, Hyoung-Gon Kim, and Heedong Ko, “Review Mining Using Lexical Knowledge and Modality Analysis,” In Proceedings of the 5th International Universal Communication Symposium(IUCS), 2011.
[13] Alexander Pak and Patrick Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining,” In Proceedings of theEuropean Language Resources Association (ELRA), 2010a.
[14] Alexander Pak and Patrick Paroubek, “Twitter based system : Using Twitter for Disambiguating Sentiment Ambiguous Adjectives,” In Proceedings of International Workshop of Semantic Evaluations, 2010b.
[15] Ismael S. Silva, JanainaGomide, Adriano Veloso, Wagner Meira Jr., Renato Ferreira, “Effective Sentiment Stream Analysis with Self-Augmenting Training and Demand-Driven Projection,” In SIGIR, 2011.
[16] J. Leskovec, D. P. Huttenlocher, and J. M. Kleinberg, “Predicting positive and negative links in online social networks,”In WWW, pp.641-650. ACM, 2010.
[17] Hogun Park, Sun-Bum Youn, Eugene Hong, Changhyeon Lee, Yong-moo Kwon, HeedongKo, Myon- Woong Park, Young Tae Sohn, and Jae Kwan Kim, “Sharing of Baseball Event through Social Media,” In Proceedings of the 11th ACM SIGMM International Conference on Multimedia Information Retrieval, pp.389-392, 2010.

저 자 약 력

박 경 미

이메일 : kmpark@imrc.kist.re.kr

- 2008년 고려대학교 컴퓨터학과(박사)
- 2009년~2010년 숭실대 컴퓨터학부 박사후연구원
- 2010년~현재 한국과학기술연구원 박사후연구원
- 관심분야 : 자연어처리, 오피니언마이닝

박 호 건

이메일 : hogun@imrc.kist.re.kr

- 2008년 한국정보통신대학교 전산학과(석사)
- 2008년~현재 한국과학기술연구원 연구원
- 관심분야 : 정보검색, 자연어처리, 오피니언마이닝

김 형 곤

이메일 : hgk@imrc.kist.re.kr

- 1974년 한국항공대학교 전자공학(학사)
- 1982년 Univ. of Kent 전자공학(석사)
- 1985년 Univ. of Kent 전자공학(박사)
- 1977년~현재 한국과학기술연구원 책임연구원
- 관심분야 : 인터랙티브 미디어, 몰입형 혼합 환경, 컴퓨터 비전 기반 콘텐츠 생성, 멀티모달인터랙션, 지능형 공간, 디지털 라이프

고 희 동

이메일 : ko@imrc.kist.re.kr

- 1989년 일리노이대학교 전산학과(박사)
- 1988년~1990년 죠지메이슨대학교 전산학과 객원조교수
- 1990년~현재 한국과학기술연구원 책임연구원
- 관심분야 : 가상현실, 인공지능, HCI