오피니언 마이닝을 이용한 블로그/SNS 에서 익명 사용자의 직업 예측

유전국*, 윤재열**, 김응모
*성균관대학교 정보통신공학부
e-mail : begoniaj@daum.net

Guessing Job of Anonymous user in Blog/SNS by using the Opinion Mining

Yu Jeun Kuk*, Jae Yeol Yoon**, Ung-mo Kim
School of Information and Communication Engineering,
Sungkyunkwan University

요 약

최근 Blog/SNS가 새로운 커뮤니케이션 도구로 정착이 되면서 지위 고하를 막론하고 많은 사람들이 자신의 의견을 인터넷 상에 표현을 하고 있다. 이러한 많은 의견들은 다른 사람들과 의견을 교환하는 역할도 하고 있으나 여기에 신뢰할 수 없는 부정확한 정보 또한 다수 포함 되어 있다. 본 논문에서는 Blog/SNS의 익명의 사용자 직업을 추측하여 이용자로 하여금 이러한 부정확한 정보를 판단할 수 있는 하나의 수단과 이용의 판단에 근거를 제시한다. 이러한 방법을 통해 사용자는 무수한 정보들 속에서 신뢰성 있는 정보와 신뢰성이 없는 정보를 보다 정확하게 판단 할 수 있을 것이다.

1.서론

최근 SNS가 사회에 미치는 영향력을 보면 예전에는 상상도 하지 못할 만큼 그 파장 효과는 크다. 미국에서의 사례를 보면 오바마 현 대통령은 큰 지지율 차로 뒤지고 있는 상황에서 단 3개월 만에 SNS를 통한 선거 캠페인으로 대승을 거두었다. 이는 SNS가 이제 세상에서 보편적인 새로운 커뮤니케이션 도구로 정착하고 있다는 것을 공식화 시켜준 역사적인 사건이다. 이처럼 사회에 인지도가 있는 유명 연예인, CEO, 정치인 등 많은 사람들이 Blog 또는 SNS 상에서 자신의 의견을 펴며 많은 글들을 남기고 있다. 그러나 최근에 SNS의 신빙성을 의심하는 하나의 사례가 있었다. 한전의 정전사태당시 한 트위터 사용자가 정확하지 않은 사실을 확실한 정보인 마냥 올린 글이 바로 그 사례이다. 나중에 밝혀진 사실엔 그 사용자는 한 의원이었으며 그 글의 내용은 진실이 아니라는 사실이 밝혀져 큰 소동은 없었지만 이 사례를 보면 많은 정보들이 쏟아져 나오는 Blog/SNS 상에서 신빙성 있는 글을 판단하는 몫은 그 글을 읽는 사람의 몫이 되어 버린 것이다.

본 논문에서는 오피니언 마이닝 기법을 이용하여 이러한 Blog/SNS 상의 익명의 사용자의 직업을 추측하여 그 사용자가 작성한 의견에 얼마나 신뢰성이 있는지 이용자의 판단에 조금이나 도움이 되고자 하는 것이 목표이다.

본 논문은 다음과 같이 구성된다. 2장에서 오피니언 마이닝 기법과 기존의 대표적인 연구에 대해 설명한다. 3장에서는 본 논문에서 제안하는 익명의 사용자의 직업을 예측하는 방법에 대해 설명하고, 4장에서 결론과 발전된 연구를 위한 향후 연구 과제를 제시하며 논문을 마친다.

2.관련연구

2.1 오피니언 마이닝

오피니언 마이닝은 각종 대량의 정보 속에서 유용한 정보를 찾아내는 방법 중의 한 가지 이다. 이러한 오피니언 마이닝은 최근에 들어 활발히 연구되어 왔으며, 그 기반이 되는 기술은 자연어처리, 텍스트 마이닝, 통계 등의 분야로 기원한다. 이러한 오피니언 마이닝은 의견의 의미방향을 분류하는 분야와 언어적 자원을 구축하는 분야로 나눌 수 있다.

의견의 방향을 분류하는 연구로는 문장의 관계, 문장 패턴 등 언어학적 규칙들을 이용하여 의견을 분류하는 방법이 존재하며 PMI(Point-wise Mutual Information)값을 계산하여 의견을 분류하는 방법도 존재 한다 [1,2]

언어적 자원들을 구축하는 분야로는 형용사와 접속사들의 관계를 활용해 형용사의 의미방향을 분류하여

구축하는 연구가 있으며[3], 단어 간의 유의어 및 반의어 정보를 활용하여 언어적 자원을 구축하는 연구도 시도되었다[4].

초반 연구들은 자연어 처리기술을 기반으로 수행되어져 왔으며, 최근에는 통계적 확률론에 기초한 연구들이 수행되어져 왔으며 이 두 가지 모두를 융합한 연구방법도 있었다[5,7].

2.2 형태소 분석

형태소 분석이란 단어를 구성하는 각각의 형태소들을 인식하고 불규칙 활용이나 축약, 탈락 현상이 일어난 경우 원형으로 복원하는 전 과정을 말한다. 이러한 형태소 분석은 한글의 어절 단위로 단어를 나누어 그 어절의 문자열을 파악한다. 여기에 사용되는 방식은 크게 세가지가 있는데 하나는 모든 형태소 어휘와 일전 규칙을 사전에 담아 놓고 이를 탐색하여 결과를 반환하는 '사전 베이스', 오토마타 등 규칙의 연산을 통해 결과를 반환하는 '규칙 베이스' 그리고 마지막으로 이 두 가지를 적절히 융합한 '절충형' 이 있다. 그리고 색인어로서 가치가 있는 명사는 주로 고유명사와 복합명사가 있는데 이 복합 명사를 단일명사로 분리해 내야만 보다 정확하고 뛰어난 성능의 색인어 추출기가 될 수 있다[8].

3.직업 예측 기법

3.1 제안 배경

일반적으로 블로그나 트위터등 SNS 사용자들은 자신의 전문 분야의 지식을 다른 사람들과 공유하기를 즐기는 경향이 있다. 그러나 글을 읽는 독자들은 그 사람이 어떤 분야에서 종사 하는지 모르는 상황에서 글에 대한 신뢰성은 낮아 질 수밖에 없다. 본 논문은 전문 분야에 의견을 제시하는 여러 글들에 대해서 오피니언 마이닝과 형태소 분석기를 이용하여 사용자의 직업을 예측하여 그러한 글과 의견들에 대한 신뢰성을 높이는 방법을 제공 한다.

3.2 제안 내용

본 논문에서 개발한 시스템의 전반적인 구조는 그림1과 같다. 우선 한 사용자의 블로그 또는 SNS에서 작성된 글을 수집하여 형태소분석 작업을 시작한다. 전문 용어를 추출 하면 되므로 분석기에 입력된 글 중에 명사만을 추출하여 Database에 저장을 한다. 예를 들면 IT업계에 종사하는 사람들은 OS, C ,Java, 산출물, 로컬, SaaS, 클라우드 등과 같은 일반 사람들이 잘 사용하지 않는 전문 용어들을 보다 더 많이 사용하게 된다. 또한 여기에 비교 작업을 위한 직업에 따른 전문 용어 사전DB를 사전에 작성을 한다. 이는 직업 별로 사용되는 전문용어를 정리하여 저장한 DB로서 본 논문에서 직업을 예측하는데 중요한 역할을 한다. 형태소 분석기를 이용하여 추출한 명사를 직업에 따른 사전과 과 비교하여 해당 직업에 해당하는 전문 용어가 추출이 되었을 경우 해당 직업의 극성을 1씩 증가 시킨다. 예를 들어 SaaS라는 단어가 추출이 되어 사전과 비교 작업에 들어갔을 경우 IT분야의 직업 전문 용어에 SaaS가 포함 되어 있으므로 IT 분야의 극성을 1 증가 시킨다. 그렇게 분석이 완료가 되면 가장 높은 점수를 얻은 직업군을 해당 글쓴이의 직업으로 예측을 한다.

3.3실험

예를 들어 인터넷 블로그에 작성되어 있는 글[9]을 기준으로 작성된 글의 형태소 분석을 통한 명사를 추출해보도록 하겠다. 일반적으로 직업 분류에 따른 전문용어를 예를 들면 표 2와 같다. 이러한 미리 구축된 혹은 제공되는 전문용어 사전을 이용하여 위의 작성된 글을 분석하면 표3와 같은 결과를 얻을 수 있다. 이처럼 일방적으로 한 가지 직업군에 대한 전문용어가 추출 될 수도 있다. 이처럼 추출한 단어의 개수를 카운터 하면

직업	극성
IT	13

표 1 결과의 예

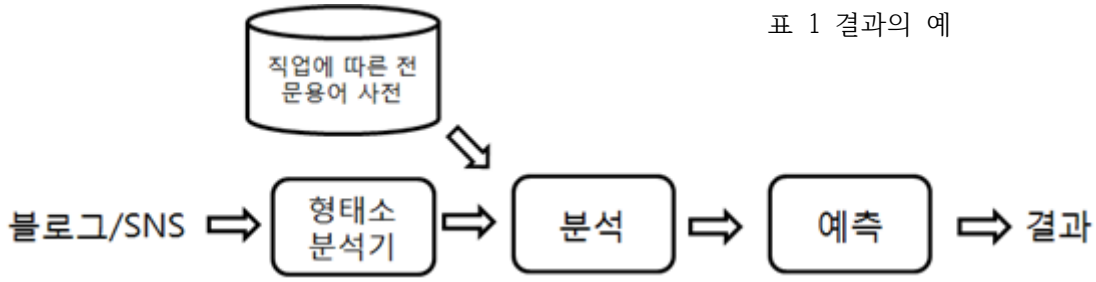

그림 1 시스템 구조

직업	용어
IT업계	Complier,API,Protocol,등
의료계	응급AR,부정맥
경제	자본주의,사회주의,등

표 2 직업군에 따른 전문용어 예

platform,무한루프,Compiler API,Network,Protocol,Browser,개발자,안드로이드,iOS,바다,플래시,SKAF

표 3 결과 예시

이라는 결과를 얻을 수 있다. 이런 결과를 기준으로 작성자의 많은 글들을 분석을 하여 통계를 얻으면 IT 업계의 종사자라는 결과를 추출할 수가 있다.

4.결론 및 향후 연구

본 논문에서는 익명의 사용자가 블로그나 SNS에 작성한 글을 형태소 분석과 오피니언 마이닝을 이용한 카운터 기법을 활용하여 그 직업을 예측 할 수 있는 방법을 제시 하였다. 이는 작성자의 전문성을 확인 할 수 있으며 나아가 글의 신뢰성을 높이는 하나의 척도가 될 수 있을 것이다.

향후 과제로는 직업 추출의 척도가 되는 중요한 직업에 따른 전문 용어 사전을 어떻게 자동/반자동 적으로 축적 할 것인가에 대한 연구가 필요 할 것이다.

참고문헌

[1] Minqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews," KDD'04, Seattle, Washington, USA., Aug. 2004.

[2] Peter D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, pp.417-424, Jul.2002.

[3] Hatzivassiloglou V., Mackeown K., "Predicting the Semantic Orientation of Adjectives," Proceedings of the 8th Conference on European chapter of the association for Computational Linguistics, pp.174-181, 1997.

[4] Xiaowen Ding, Bing Liu, "The Utility of Linguistic Rules in Opinion Mining," Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.811-812, 2007.

[5] Xiaowen Ding, Bing Liu. "The Utility of Linguistic Rules in Opinion Mining," pp.811-812, SIGIR2007,2007.

[6] M. Gamon, A. Aue, S. Corston-Oliver, E. Ringger. "Pulse: Mining Customer Opinions from Free Text," In Lecture Notes in Computer Science, Vol.3646. Springer Verlag. (IDA 2005), 2005.

[7] Wilson, T., Wiebe, J., Hoffmann, P. "Recognizing contextual polarity in phrase-level sentiment analysis," In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp.347-354, 2005.

[8] 강승식 지음, "한국어 형태소 분석과 정보검색", 홍릉과학출판사, 2002

[9] http://fstory97.blog.me