비정형 데이터를 활용한 감기 판단 사전 구축

김광민, 남기훈
*서경대학교 컴퓨터공학과

e-mail : ip9894@naver.com

Constructing the Dictionary of Flue using unstructured data

KimKangMin,NamKiHun
*Dept. of Computer Engineering, Seo-Kyeong University

요 약

최근에 비정형 데이터의 잠재적 가치를 유용한 데이터로써 사용하려는 경우가 많아지고 있다. 특히 트위터는 사용자의 상태나 이벤트가 잘 나타나 있어서 하나의 사용자의 이벤트로서 간주될 수 있다. 본 논문은 트위터에서 발생하는 이벤트에 주목하여, 감기라는 이벤트를 트위터 내에서 추적하고자 한다. 추적을 위해서는 트위터를 판단할 필요가 있는데, 이를 위해 기존의 감성 사전 방식 중 하나인 통계적 사전 구축을 기반으로 키워드를 활용하여 감기 판단 사전을 구축하는 방식을 제안한다.

1. 서론

스마트 폰 사용자들은 집이나 혹은 사무실 같은 제한된 장소만이 아닌 언제 어디서나 컴퓨터를 사용 할 수 있게 되었다. 이로 인해 비정형 데이터가 급속도로 증가 하게 되었다. 비정형 데이터란 관계형 데이터베이스에서 주로 사용되는 정형 데이터와는 다르게 정해진 틀이 없는 텍스트 데이터를 의미한다. 스마트 폰의 등장 후 SNS 라는 소셜 네트워크 서비스가 빠르게 발전하였다. 소셜 네트워크란, 일종의 실시간 인터넷 게시판이라고 생각하면 되는데, 사용자가 실시간으로 자신이 본 것, 먹은 것, 느낀 것 등을 기록할 수 있다는 점이 소셜 네트워크의 가장 큰 특징이라 할 수 있다. 이러한 특징 때문에 사용자들은 실시간으로 많은 비정형 데이터를 생산하게 됐다. 비정형 데이터는 별도의 가공을 거치면 새로운 데이터를 만들어 낼 수 있는 잠재적 가지를 가진 데이터이기 때문에 정부나 많은 기업들은 이 점을 파악하여 활발히 연구를 진행 하고 있다. 예를 들면 A 기업은 SNS 에서 나오는 비정형 데이터를 분석하여 개봉한 영화의 흥행을 예측했으며, B 기업은 인터넷 뉴스 비정형 데이터 자료와 SNS 비정형 데이터를 활용하여 주식의 증감률을 예측하기도 하였다.

본 논문은 이러한 트위터의 잠재적 가치를 기준으로 의학분야, 그 중 감기라는 질환에 대하여 대중들에게 유용한 정보를 추출하고자 한다.

2. 관련 연구

트위터 데이터를 활용하여 특정 이벤트와의 상호작용을 분석하는 논문으로 도쿄 대학교에서 진행 된 지진에 관련된 연구[1]를 꼽을 수 있다. 이 연구에서 지진이 일어나는 시점에서 발생하는 트위터를 분석하여 지진 발생 지점을 알아내는 방식을 고안해 냈다.

<그림 1: 구글 독감 지도>

질병 관련 연구로는 대표적으로 구글의 독감 지도를 꼽을 수 있다. 구글은 검색사이트를 사용하는 사용자들이 검색하는 키워드 중, 독감에 관련된 키워드를 분석하여 <그림 1> 을 사용자들에게 제공해 주고 있다.

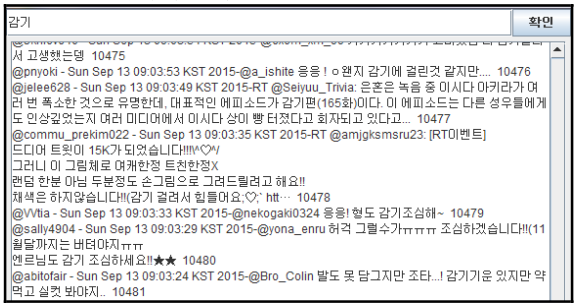
트위터를 대상으로 연구한 논문으로는 '트위터에서 추출한 감기 증상의 사회적 신호와 영향요인과의 상관분석' [2]을 참고 하였다. 참고한 논문은 감기를 키워드로 잡아 데이터를 수집한 후, 데이터 량과 기상청에서 제공하는 감기 지수와의 상관계수를 분석하여 트위터 데이터 수와 감기지수는 상관관계가 높다는 것을 알게 되었다.

본 논문은 이러한 연구들을 바탕으로 트위터에서 감기 키워드를 가지는 데이터를 수집하고 더 나아가 감기에 걸린 트위터 사용자를 판단하는 사전을 구축, 활용하여 감기에 걸린 트위터 사용자의 수를 알아내

려 한다.

3. 비정형 데이터를 활용한 감기 판단 사전 구현

3.1 비정형 데이터 수집

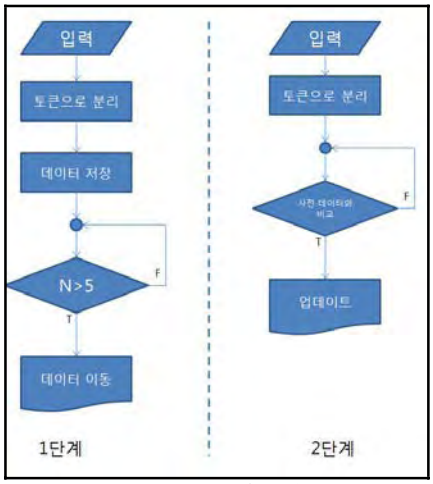


<그림 2: 트위터 데이터 수집>

감기데이터 수집을 위해 본 논문은 twitter4j 라는 검색 API 를 사용하였다. twitter4j 는 트위터에서 원하는 데이터를 검색하게 해주는 API 로 접근이 쉬우며 구현하기가 간단하다는 점에서 채택하게 되었다. <그림 2>는 검색 조건으로는 '감기' 키워드가 들어간 트위터로 설정한 모습을 보여 준다. 설정한 검색 조건을 기준으로 '감기' 키워드를 가지는 트위터를 검색한 후, 검색한 데이터는 파싱 속도가 빠른 JSON 을 활용하여 저장하였다.

3.2 사전 구축 과정

감기 데이터를 수집 한 후, '감기' 키워드를 가진 트위터를 감기에 걸린 사용자로 판단하기 위해서 감성 사전과 비슷한 감기 판단 사전을 구축할 필요성이 있다. 감기 판단 사전은 '맵리듀스를 이용한 통계적 접근의 감성 분류'[3]에서 제시한 방식을 기반으로 구축하였다. 논문에 따르면, 사람들이 사용하는 단어들의 변수가 너무 많기 때문에 형태소 처리를 하지 않고 문장의 단어들을 토큰화 하여 사전에 저장한다. 토큰화한 단어는 정해진 기준을 바탕으로 긍정과 부정점수가 판단되어 사전에 저장된다. 이 방식은 영화 댓글에서 나오는 평점을 기준으로 사전 저장을 진행하였다. 평점이 높은 댓글에서 나오는 단어 토큰들은 긍정점수가 부여하고 평점이 낮은 경우 부정점수를 부여하였다. 이 방식은 단어 저장 기준이 명확히 존재하기 때문에 사전에 단어를 일괄적으로 저장하는 것이 가능하지만, 기존방식을 감기 판단 사전에 적용하는데 몇 가지 문제가 발생한다. 감기 판단 사전은 '감기' 키워드가 들어간 트위터를 가지고 사전 저장을 진행한다. 하지만 영화 댓글의 평점과 다르게 트위터 글은 판단할 기준이 존재하지 않으므로 사전에 단어를 기존 방식처럼 일괄적으로 저장할 수가 없다. 그래서 감기에 걸린 트위터인지, 그렇지 않은 트위터인지를 직접 보면서 사전에 등록하는 과정을 가질 필요가 있다. 구축할 감기 판단 사전은 기존 감성 사전처럼 긍정적, 부정적인 글을 구별해 내는 것이 아닌 단순히 감기에 걸린 트위터 사용자를 판단하기 때문에 판단 과정에 있어 불필요한 단어 역시 저장될 수 있다. 기존 방식을 그대로 사용하면 위 같은 문제가 발생하기 때문에 트위터에서 나오는 단어들을 토큰화하여 임시 저장소에 저장한 후, 임의로 지정한 빈도수를 기준으로 임시 저장소에서 감기 판단 사전으로 데이터를 옮겨오는 작업을 수행하는 키워드 사전 방식을 제시한다. 알고리즘은 <그림 3> 과 같다.



<그림 3: 사전 구축 알고리즘>

1) 입력 데이터(트위터)를 받는다.
2) 데이터를 토큰화 한다.
3) 토큰화 한 데이터를 임시 저장소에 등록한다.
4) 임의로 지정한 빈도수를 기준으로 임시 저장소에서 감기 사전으로 데이터를 옮긴다.
5) 새롭게 입력 데이터를 받는다
6) 데이터를 토큰화 한다.
7) 토큰화 한 데이터와 저장된 감기 사전과 비교하여 사전 내용을 업데이트한다.

다만 이러한 작업을 수행하면, 사전 볼륨이 작아지며 트위터를 판단 할 때 사전에 적중되지 않는 단어가 상당 수 등장하는 문제가 발생하게 된다. '강건한 한국어 상품평의 감정 분류를 위한 패턴 기반 자질 추출 방법' [4] 에서는 초/중성 데이터를 활용한 방법은 제시하는데, 비슷한 의미의 단어는 초성과 중성이 일치하는 성향을 보이게 된다.

	초성	중성
좋아해요	ㅈㅇㅎㅇ	ㅗㅏㅐㅛ
조아해요	ㅈㅇㅎㅇ	ㅗㅏㅐㅛ
조아해용	ㅈㅇㅎㅇ	ㅗㅏㅐㅛ

<표 1: 초성/중성 비교 표>

<표 1>은 같은 의미지만 맞춤법이 틀린 단어들을 비교해본 예이다. 좋아해요/조아해요/조아해용 전부 '좋아해요'와 같은 의미로 맞춤법은 틀리지만 초성과 중성은 전부 같은 것을 확인할 수 있다. 본 논문은 사전에서 적중하지 않는 단어들과 비슷한 단어를 찾아내기 위해 강건한 한국어 상품평의 감정 분류를 위한 패턴 기반 자질 추출 방법' [4]에서 제시한 방식을 채택하여 사전에 적용하여 사용하였다.

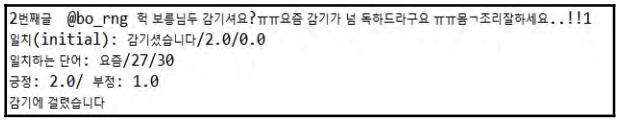
Field	Type	Null	Key	Default	Extra
text	varchar(50)	NO	PRI	NULL	
fhash	int(11)	YES		NULL	
initial	varchar(50)	YES		NULL	
count_true	int(11)	YES		NULL	
count_false	int(11)	YES		NULL	
count	int(11)	YES		NULL	

<그림 4: 감기 판단 사전 테이블>

<그림 4>은 감기 판단 사전의 테이블 구조를 나타낸다. ‘text’의 경우, 저장된 단어를 뜻한다. ‘fhash’는 본 사전에서 인덱스나, 해쉬 탐색을 사용하지 않기 때문에 임의로 해쉬 항목을 만들어 쿼리 수행 시 빠른 접근을 하기 위해 사용 하였으며 text 의 첫 글자에 해당하는 hash 코드가 저장되어 있다. ‘initial’ 경우, text 의 초/중성이 저장 되어있다. count_true/count_false는 ‘감기에 걸린/걸리지 않은’ 트위터에 해당하는 단어들의 빈도수를 저장하였다.

3.3 감기 판단 방법

트위터 내용을 가지고 감기에 걸렸는지를 판단하는 과정은 count_true/count_false 의 합 연산을 가지고 판단한다.



<그림 5: 트위터 판단 결과의 예>

<그림 5>는 트위터에서 나온 단어를 감기 판단 사전과 비교하여 감기에 걸렸는지를 판단한 모습이다. ‘긍정>부정’ 인 경우 감기에 걸렸다고 판단하며 ‘긍정<부정’인 경우 감기에 걸리지 않았다고 판단한다. 마지막으로 ‘긍정=부정’인 경우 판단 불가로 카운팅에서 제외하였다.

4. 실험 및 결과

본 논문은 실험을 위해, 사전 구축시 참고했던 ‘맵리듀스를 이용한 통계적 접근의 감성 분류’의 방식과 비교하였다. 참고한 논문은 통계적으로 사전을 구축하는 방식은 본 논문과 같지만 키워드 측면으로 접근하는 점과 초/중성을 활용한 점은 다르다.

참고한 ‘맵리듀스를 이용한 통계적 접근의 감성 분류’[3] 기반으로 구성한 사전을 적용한 적중률은 다음과 같다.

	비적중	판단불가	적중률
2015.4.21	20	4	0.208333
2015.4.22	20	7	0.215053
2015.4.24	20	5	0.210526
2015.4.25	22	4	0.219166
20.15.4.26	21	4	0.21875

<표 1: 기존 방식으로 구현한 사전의 적중률>

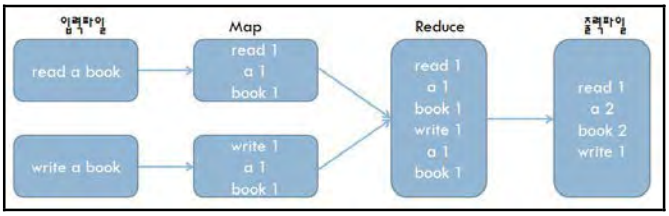
새로 제안한 방식의 적중률은 다음과 같다.

	비적중	판단불가	적중률
2015.4.21	15	8	0.163043
2015.4.22	13	11	0.146067
2015.4.24	10	7	0.107526
2015.4.25	10	11	0.112359
2015.4.26	11	13	0.126436

<표 2: 제안한 사전의 적중률>

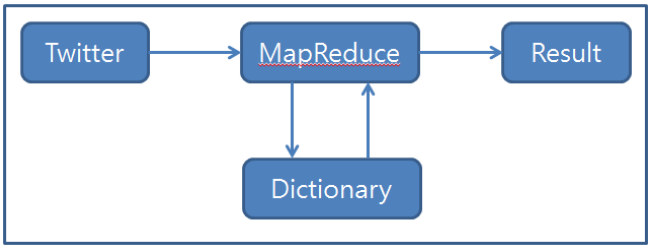
같은 사전 크기에서 총 100 개씩 5 일을 기준으로(‘판단불가’는 카운팅에서 제외) 적중률을 알아본 결과, 기존의 방식은 평균 78.2 퍼센트, 제안한 방식은 평균 약 86.5 퍼센트의 적중률을 보이며 약 8.3 퍼센트 높은 적중률을 보인 것을 알 수 있었다.

제안한 방식의 사전을 구현한 후, 최종적으로 감기에 걸린 트위터 환자수를 알아내기 위해 완성된 사전으로 수집된 트위터를 처리하였다. 트위터 데이터의 수가 매우 많기 때문에 일반적인 시스템에서는 효율적인 처리를 기대하기 힘들다. 그래서 본 논문은 효율적인 처리를 위해 MapReduce 하둡 시스템을 채용했다.



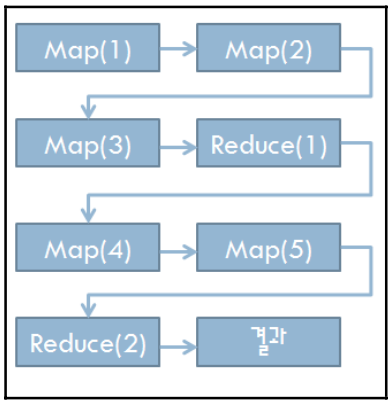
<그림 6:MapReduce 의 데이터 처리 과정>

<그림 6>는 MapReduce 내의 Map 과 Reduce 함수의 처리 과정을 보여주고 있다. Map 은 입력 파일을 한 줄씩 읽어 데이터를 변형하고 리듀스는 맵의 결과 데이터를 집계했다.



<그림 7: 트위터 판단 과정>

<그림 7>는 본 논문의 트위터 판단 과정 모델을 보여주고있다. 트위터 데이터를 MapReduce 로 넘겨준 후, MapReduce 는 전달받은 트위터를 미리 구현한 감기 사전과 데이터를 비교하여 결과 값을 만들어 낸다. 본 논문은 총 다섯 개의 Map 과 두개의 Reduce 를 활용하여, 하나의 Map 과 Reduce 로 많은 비정형 데이터의 전처리 과정을 처리해야하는 비효율성을 해결하였다.



<그림 8: MapReduce 시스템 모델>

<그림 8>은 수행할 MapReduce 시스템 모델이다. 여러 개의 Map 을 사용하기 위해서는 ChinMapper 와 ChinReducer 기법을 사용해야 된다. Chain 을 사용할 경우, MapReduce 를 여러 번 사용하는 것보다 입출력이 감소하기 때문에 여러 번 MapReduce 를 반복적으로 수행할 경우보다 더욱 효율적으로 수행될 수 있다.

Map(1)~Map(3)까지는 단어를 전처리 과정을 나타낸

다. 1 번~2 번은 트위터 문장에서 단어를 분리한 후, 3 번 Map 에서 사전 DB 에 접속하여 단어의 점수를 판단한다. Reduce(1)에서 Map(1)과 Map(3)까지 수행한 결과 값을 바탕으로, 문장 별로 점수를 계산한다. Map(4)는 Map(5)로 데이터를 넘겨주는 과정에서의 전처리를 수행해준다. Map(5)는 Chain 시스템의 특성상 Reduce 가 한번 밖에 수행되지 않기 때문에, 앞에서 나온 결과 값을 Reduce(2)처리를 하기위해, 새로운 MapReduce 과정을 거쳐야만 된다. 그렇기 때문에, Map(5)는 의미 없이 Reduce(2)를 수행하기 위한 다리 역할만 수행한다.

Reduce(2)에서 최종적으로 감기에 걸린 트위터 사람의 수를 계산해 낸다.

331	786$1102$218
401	743$897$174
402	275$339$59

<그림 9: MapReduce 결과>

<그림 9>에서 2015.3.31 에 감기에 걸린 트위터 사용자는 786 명인 것을 알 수가 있다. 1102 는 감기에 걸리지 않았지만, 트위터에 감기에 관련된 글을 올린 사람이고 218 은 판단 불가 트위터를 나타낸다.

5. 결론

SNS 는 사용자의 감정 상태, 행동 등 여러 이벤트가 발생한다. 본 논문은 이러한 이벤트에서 나타나는 정보를 활용하여 트위터 사용자 내에서 감기에 걸린 환자를 추적해보았다. 트위터 판단 과정에 기존 감정 사전의 방식 중 통계적 방식을 적용하였는데 통계적 방식은 기존 자연어 처리를 활용해 규칙 기반으로 사전을 구축하는 방식과는 다르게 자연어 처리를 수행하지 않고 단어를 토큰화하여 사전을 구축하는 기법이다.

하지만 기존 방식은 단어 저장의 기준이 명확하여 사전에 단어를 저장하는 과정이 매우 간단하지만 감기 판단 사전의 경우, 사전 저장의 단어 선별의 기준이 모호하며 불필요한 단어가 판단 과정에 영향을 줄 수 있기 때문에 기존 방식을 사용하기에는 어려움이 있다. 그래서 본 논문은 키워드 방식으로 일단 단어를 토큰화 하여 임시 저장소에 저장한 후, 빈도수가 높은 단어를 선별하여 사전을 구축하는 방식을 제안했고 기존 방식보다 약 8.3 퍼센트 높은 적중률을 보여주었다.

이번에 제안한 감기 판단 사전을 활용하면 트위터 등의 SNS 데이터에서 감기 환자를 파악 할 수 있으며, 나아가 사전을 활용하여 감기 환자 분포도, 전파 경로 등 응용될 것으로 보이며 또한 감기뿐만 아니라 다른 질병 혹은 새로운 이벤트를 기준으로 사전을 새롭게 구현할 수 있다는 점에서 다양하게 사용 가능할 것이라고 기대된다.

참고 문헌

[1] **T. Sakaki, M. Okzaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," Proc. of the 19th Int'l Conf. on World Wide Web, pp. 851-860, 2010.**

[2] 윤진영, 김석중, 이범석, 황병연 " 트위터에서 추출한 감기 증상의 사회적 신호와 영향요인과의 상관분석" , 멀티미디어학회 논문지 제16권 제6호, pp.667-677, 2013.

[3] 백승희,"맵리듀스를 이용한 통계적 접근의 감성 분류", 한국감성과학회 제15권 제4호, pp.425-450, 2012.

[4] 신준수,김학수, "강건한 한국어 상품평의 감정 분류를 위한 패턴 기반 자질 추출 방법", 정보과학회논문지 : 소프트웨어 및 응용 제 37 권 제 12 호,pp.946-950,2010.

[5] **Vasileios Lampos, Tijl De Bie, and Nello Cristianini, "Flu detector - Tracking epidemics on Twitter", Proc. of the European Conf. on Machine Learning and Principles and Practice on Knowledge Discovery in Database, pp.599~602, 2010**

[6] 손영우, "Social Networks과 Twitter 서비스에 관한 고찰," 멀티미디어 학회 논문지, 제14권, 제4호, pp. 546-553, 2011.

[7] 이범석, 김석중, 윤진영, 황병연, "온라인 소셜트렌드의 감지 및 시각화" 제30회 한국멀티미디어 학회 추계학술대회 논문집, 제15권, 제2호, pp. 95, 2012.

2010~ 현재 서경대학교 컴퓨터 공학과 학사과정

관심분야:빅 데이터, Java,DB , 소셜 네트워크 분석

2013~ 현재 서경대학교 컴퓨터 공학과 초빙교수

관심분야:빅 데이터, IoT, SoC