정확한 이슈를 찾기 위한 트위터 기반 정제기법 제안

최봉준*, 우호진*, 이원석*
*연세대학교 컴퓨터과학과
e-mail : {bongdalc, judas, leewo}@database.yonsei.ac.kr

Tweet-Based Filtering and Refinement for Finding Accurate Issues

BongJun Choi*, Ho Jin Woo*, Won Suk Lee*
*Dept. of Computer Science, Yonsei University

요 약

스마트 디바이스 산업의 발전으로 소셜미디어 데이터의 양은 기하급수적으로 증가하고 있다. 이렇게 증가한 데이터와 함께 분석을 통해 발견할 수 있는 정보의 양도 다양해지면서 여러 산업분야에서 소셜미디어 데이터 분석을 위한 연구가 진행되고 있다. 소셜미디어는 종류가 다양하고 하루 평균 발생량이 너무 많기 때문에 분석시간이 오래 걸릴 뿐 아니라, 불필요한 불용어 및 방해요소 때문에 적절한 정제작업이 필요하다. 본 논문에서는 소셜미디어의 한 종류인 트위터 분석을 위해 여러 가지 기법으로 데이터를 정제한다. 정제과정은 분석에 용이한 형태로 데이터를 변형시킨 후 의미없는 데이터와 분석에 방해가 되는 불용어를 제거한다. 이 정제를 통해 데이터 정보의 질을 높이고 분석 시간을 단축시켜 빠르고 신뢰성 높은 분석결과를 도출할 수 있다.

1. 서론

최근 스마트 모바일 디바이스의 발전으로 소셜미디어에 대한 사용이 급증하였다. 이를 기반으로 온라인 소셜 미디어 사이트의 급속한 발전과 함께 엄청난 양의 콘텐츠가 발생하였다[1]. 이렇게 많은 양의 웹 콘텐츠는 다양한 정보를 내포하고 있고 이 정보를 기반으로 하는 사용자 패턴분석이나 웹로그 데이터 마이닝 등의 연구가 활발하게 진행되고 있다[2][3]. 하지만 많은 양의 데이터를 분석하기에는 많은 시간이 필요하며 불필요한 데이터가 많이 포함되어 있다. 특히 많은 사용자가 이용하는 트위터 데이터는 빅데이터로써 데이터의 크기가 쉽게 분석할 수 없는 수준이다. 데이터의 양과 비례로 데이터 분석에 불필요한 정보도 많기 때문에 좋은 정보를 얻기가 쉽지가 않다. 따라서 트위터 데이터를 효과적으로 분석하기 위하여 본 논문에서는 트위터 분석을 위한 정제기법을 제안한다.

본 논문에서 사용하는 정제기법은 데이터 분석을 위해 필요한 형태로 수정하거나 불필요한 데이터와 방해요소를 제거하는 것이다. 데이터에서 정보를 추출하는 알고리즘들은 데이터의 형식이 알고리즘마다 다르기 때문에 분석 가능한 형태로 변형시켜야 한다. 트위터 데이터는 기본적으로 JSON 형식으로 제공되는데 기존의 분석 도구들은 JSON 형식을 바로 사용할 수 없는 경우가 많다. 그래서 분석에 용이하게 필요한 요소들만 추출하고 형태소 분석기를 이용하여 의미는 같지만 다른 형태의 단어들을 기본 형태로 만든다. 이후 분석에 불필요한 항목은 필터링 과정을 통해 제거한다. 이렇게 정제된 트위터 데이터는 불필요한 단어와 불용어가 제거되고 전체 데이터의 크기가 줄어들어 분석시간은 감소하고 분석된 정보의 의미는 보다 정확해 진다.

본 논문의 구성은 다음과 같다. 먼저 2 장에서 트위터 데이터를 분석하여 가공하는 방법에 대하여 살펴보고, 3 장에서 불필요한 데이터를 필터링으로 분석에 필요한 데이터를 얻는 방법을 설명한다. 4 장에서는 결론을 제시한다.

이 논문은 2014 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2011-0016648).

2. 트위터 데이터 가공

2-1. 트위터 데이터 파싱

트위터에서 데이터를 제공하는 기본 형식은 JSON(JavaScript Object Notation)으로 컬럼과 값 형식이다. 트위터 데이터는 많은 정보를 가지고 있는데 분석에 주로 사용되는 컬럼은 트위터 사용자가 작성한 텍스트 문구이다. JSON 형태로 구성된 트위터를 트랜잭션(Transaction)단위로 단위로 변형한다. 트랜잭션 형태는 JSON 형태의 트위터 데이터로부터 필요한 항목인 트윗 ID, 날짜, 텍스트를 구분자를 이용하여 표현한다. 트랜잭션 형태로 변형시키는 이유는 데이터를 분석할 때 필요한 정보만을 표시하는 트랜잭션형태가 데이터 분석에 용이한 경우가 많기 때문이다. 아래 표 1 은 트위터의 JSON 형태와 트랜잭션 형태로 변형한 예시이다.

<표 1> 트위터 JSON 및 트랜잭션 형태

형태	예시
JSON	{"id":9999092, "text":"주말에 소풍을 간다.", "oriTweetId":null, "date":"Sun Mar 17 07:36:52 +0000 2013", "user" : 1950350039}
트랜잭션	9999092|20130317|주말에 소풍을 간다.

2-2. 형태소 분석

위의 과정으로 추출된 트랜잭션의 텍스트는 바로 분석하기에 문제점이 많다. 한글의 특성상 대한민국이라는 단어에 여러 가지 어미가 붙을 수 있기 때문에 하나의 단어가 많은 형태로 표현된다. 예로 '대한민국에, 대한민국의, 대한민국은, 대한민국을, 대한민국과, 대한민국이' 등과 같이 대한민국을 표현하는 방법이 많기 때문에 이러한 단어들 모두 '대한민국'이라는 단어로 통일시켜야 한다. 이러한 문제점을 해결할 수 있는 방법은 형태소 분석기를 이용하는 것이다

본 논문에서는 꼬꼬마 형태소 분석기[4]를 이용하여 분석할 문장을 형태소 분석기에 입력하면 불필요한 어미는 제거되고 본래의 의미만 가지는 단어가 추출된다. 이렇게 추출된 단어들을 공백 구분자를 이용하여 아래 표 2 와 같은 최종 형태로 출력한다.

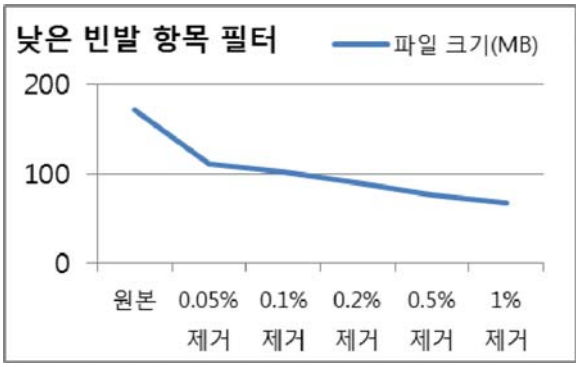
<표 2> 트랜잭션 최종 형태

최종 형태
9999092|20130317|주말 소풍 간다

3. 트위터 데이터 필터링

3-1. 낮은 빈발 항목 필터

본 논문에서 사용한 트위터 데이터는 약 300 만개의 트랜잭션으로 데이터 마이닝을 수행하기엔 과도한 양이다. 분석 시간을 단축시키기 위해서 낮은 빈발 항목을 걸러내는 작업을 수행한다. 낮은 빈발 항목은 주제어가 될 확률은 희박하면서도 전체 데이터의 상당량을 차지하고 있기 때문에 적절한 필터링을 거치면 전체 수행시간을 큰 폭으로 줄일 수 있게 된다.



(그림 1) 낮은 빈발 항목 제거 그래프

그림 1 은 매개 변수에 따른 파일 크기 변화로 비교적 선형으로 감소하는 것을 볼 수 있다. 분석 결과에 영향을 미치지 않으면서 분석속도를 향상시키기 위해 파일 크기가 급격하게 떨어지는 구간인 0.05~0.1 을 매개 변수로 잡는다. 이 매개변수는 도메인의 성향과 분석방향에 따라 변할 수 있다. 예를 들어 분석할 목표를 빈발도가 높은 항목뿐 아니라 낮은 항목에서도 갑자기 증가한 항목을 찾기를 원한다면 낮은 빈발 항목을 제거하는 방법은 뒤에 설명할 일(日) 수를 이용하여 제거해야 한다. 낮은 빈발 항목의 제거로 인해 분석해야 할 데이터의 크기가 반으로 줄어들었다. 이 과정을 통해 전체 파일의 크기가 크게 줄어들고 분석 결과의 정보도 질이 높아진다.

3-2. 높은 빈발 항목 필터

트위터에서 이슈를 찾을 때 매일 빈발하는 항목은 주제어로 보기 힘들다. 예를 들어 '오늘', '친구'와 같은 단어는 매일 빈발하게 나오지만 이를 주제어로 뽑기에는 무리가 있다. 이를 해결하기 위해 매일 빈번하게 나오는 높은 빈발 항목을 제거한다. 하지만 특정 하루에 엄청난 이슈로 특정 항목이 전체이슈에서도 높은 빈발을 보일 수 있는데 이를 구분해 내기 위해서 일별로 집계를 한 뒤 높은 빈발의 단어를 집계값을 빼고 단어만 모아 다시 집계를 한다. 일별로 구해진 집계에서 높은 빈발을 정의하기 위한 매개변수를 설정해야 한다. 매개변수에 따라 중요한 단어가 제거될 수 있기 때문에 매개변수 설정은 매우 중요하다. 필터링에 사용되는 매개변수는 2 종류가 있다. 하나는 30 일중 필요이상의 빈발일(日) 수로 일정 일 이상 검출된 단어를 목록으로 작성하여 분석전에 걸러낸다. 두 번째는 각 일자별로 빈발한 단어를 선택할 때 사용되는 매개 변수로 높은 빈발의 구간을 퍼센트로 표시한다. 아래 표 3 은 일자별 상위 빈발도에 따른 필터링 목록의 크기를 나타낸 것이다.

<표 3> 높은 빈발 항목 표

필터링 매개 변수(%)	필터링 목록(개)
1	355
2	717
3	1064

빈발도가 높은 항목들은 이슈가 될 가능성이 높고 주요한 구역이지만 매일 빈발하는 키워드들은 이슈와 관련이 없는 경우가 많기 때문에 정밀한 정제기법이 필요하다. 높은 빈발 항목을 많이 제거하면 전체 파일의 크기는 줄어들겠지만 필요한 항목이 제거될 우려가 있다. 그렇기 때문에 적절한 매개 변수를 찾기 위해서 데이터의 특성을 잘 파악하는 것이 중요하다.

<표 4> 높은 빈발 항목 매개변수 표

빈발도	1%	2%	3%
결과 1	안철수 출마 부산 원병 민주당 정부 대통령 찬 희망 선거	출마 원병 찬 정의 노회 교수 대선 결정 진보	출마 원병 노회찬 진보 영도 담임 길이
결과 2	1000 일 인피니트 학년 컴백 천일 데뷔	1000 일 데뷔 가방 쌍 리전	1000 일 데뷔 천일 덤 대한문 자기소개 특별

높은 빈발 항목에서 더 높은 매개 변수를 설정할수록 필터링 목록이 늘어난다. 첫 번째 매개변수와는 달리 두 번째 매개변수는 정확도에 영향을 미칠 수 있다. 표 4 는 각 빈발도 매개변수 별로 분석한 뒤에 나온 결과이다. 검출 일수 매개변수는 15 로 고정하였다. 결과 1, 2 에서 각각 중요한 키워드가 되는 '안철수'와 '인피니트'가 빈발도 매개변수를 2%, 3%로 설정하면, 필터링시 제거되어 주제어에 나타나지 않게 된다. 그 이유는 2%, 3%로 했을 때 '안철수'와 '인피니트'가 15 일 이상 빈발하게 발생하지만 1%로 했을 때는 15 일 이상 빈발하지 않기 때문에 1%로 했을 때만 나타나는 현상을 볼 수 있다. 첫 번째 매개변수와 두 번째 매개변수는 분석하는 대상에 따라 조금씩 변할 수 있지만 본 방식은 다른 대상에도 적용할 수 있다.

4. 결론

본 논문에서는 소셜미디어를 분석하기 전 원본데이터를 정제함으로써 전체 파일크기를 줄이고 분석결과의 정확도를 높이는 것을 보였다. 이슈로 분류되지 않으면서 전체 파일을 많이 차지하고 있는 낮은 빈발 항목을 제거하여 문서크기를 절반가량 감소시킬 수 있었고 높은 빈발 항목을 제거할 때는 매개변수 조정을 통해 최종 결과값의 정확도를 높였다.

현재 정제기법으로 분석시간은 줄이고 정확도는 높였으나 아직 정제되지 않은 부분이 있다. 트위터에는 봇(BOT)이 존재하는데 봇은 반복적으로 광고를 위한 키워드를 작성한다. 이러한 문제점 때문에 불필요한 트윗이 증가하게 되고 분석된 결과의 의미를 불분명하게 한다. 이러한 문제점만 해결한다면 더 높은 수준의 결과를 도출해 낼 수 있을 것이다.

참고문헌

[1] Xuning Tang, Christopher C. Yang "TUT: A Statistical Model for Detecting Trends, Topics and User Interests in Social Media" Proceedings of the 21st ACM international conference on Information and knowledge management, pa.972-981 October 29 - November 02, 2012

[2] J. Pei, J. Han, B. Mortazavi-Asl and H. Zhu, "Mining Access Patterns Efficiently from Web Logs", PAKDD'00, pp. 396-407, 2000

[3] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg, "Top 10 algorithms in data mining.", Knowl. Inf. Syst. vol.14, no.1, pp.1-37, Dec. 2007.

[4] 이동주, 연종흠, 황인범, 이상구, 꼬꼬마: 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구, 2010, 정보과학회논문지: 컴퓨팅의 실제 및 레터 (Journal of KIISE: Computing Practices and Letters), Volume 16, No.11, Page 1046-1050