실시간 트위터 분석을 통한 트렌드 및 연관키워드 추출[1]

김대용, 김대훈, 황인준
고려대학교 전자전기전파공학부
e-mail : {ritgd05, kdh812, ehwang04}@korea.ac.kr

Trend and related keyword extraction based on real-time Twitter analysis

Daeyong Kim, Daehoon Kim, Eenjun Hwang
School of Electrical Engineering, Korea University

요 약

최근 Twitter를 비롯한 소셜 네트워크 서비스의 급속한 확산으로 인해, 많은 수의 SNS 메시지가 실시간으로 생성되고 있다. 이러한 SNS상에서의 단문 글들을 실시간으로 분석하여 최신의 트렌드를 추출해 낼 수 있다면, 사용자에게 유용한 정보를 제공하는 것이 가능하다. 본 논문에서는 다량의 Tweet글들에 대한 실시간 분석을 바탕으로 트렌드를 추출하고 연관된 키워드를 제공하는 기법을 제안한다. 제안하는 기법은 실시간으로 생성되는 Tweet내에서 영어의 언어적 특성을 활용하여 최근 이슈화된 트렌드 키워드를 추출해낸다. 또한, Tweet 내에서 각 트렌드 키워드간 관계를 분석하여 연관 키워드를 제공하며, 동시에 Wikipedia와 Google에서의 검색을 통하여 다른 형태의 연관 키워드도 추출한다. 이 모든 과정은 제안된 트렌드 추출 알고리즘을 통해 실시간으로 제공된다. 제안된 기법을 바탕으로 시스템을 구현하고 다양한 실험을 통하여 키워드의 유효성 및 처리 속도 면에서 시스템의 성능을 평가한다.

1. 서론

최근 Twitter를 비롯한 다양한 소셜 네트워크 서비스(SNS)가 급속도로 확산되고 있다. Twitter에서는 짧은 글 위주로 작성되며 축약어가 많이 등장하며, 대부분 모바일 환경에서 작성되기 때문에 오타도 흔히 발생한다. 따라서, 효과적인 트렌드 추출을 위해서는 위와 같은 제약을 극복할 수 있어야 한다.

이를 위한 다양한 연구가 진행되어 왔다. 대량의 Tweet을 수집하여 장문으로 병합 후 LDA를 활용하거나[1], 단시간에 출현 빈도가 급등한 Single Tag 정보를 이용하여[2] 트렌드 키워드를 추출하는 방식이 제안되었다. 하지만 LDA를 활용한 방법은 긴 처리 시간으로 인해 실시간 활용에 적합하지 않다. 또한, Single Tag만을 활용하는 경우 추출된 특정 키워드 및 연관 키워드 별로 검토 할 수 있는 방안이나 특정 지역 혹은 시간대별 추세를 제공하긴 어렵다.

본 논문에서는 클라이언트-서버 환경에서Tweet을 분석하여 트렌드를 추출하고, 연관된 키워드를 제공하는 시스템을 구현한다. 먼저 클라이언트는 Tweet에 대해 후보 키워드를 찾아내어 서버에 전달한다. 서버에서는 키워드 병합 및 지역 정보를 기반으로 인덱스한다. 이후 연관된 키워드 추출 및 Stopword를 제거하며, Wikipedia및 Google에서 검색으로 기존의 연관 키워드도 함께 제공한다.

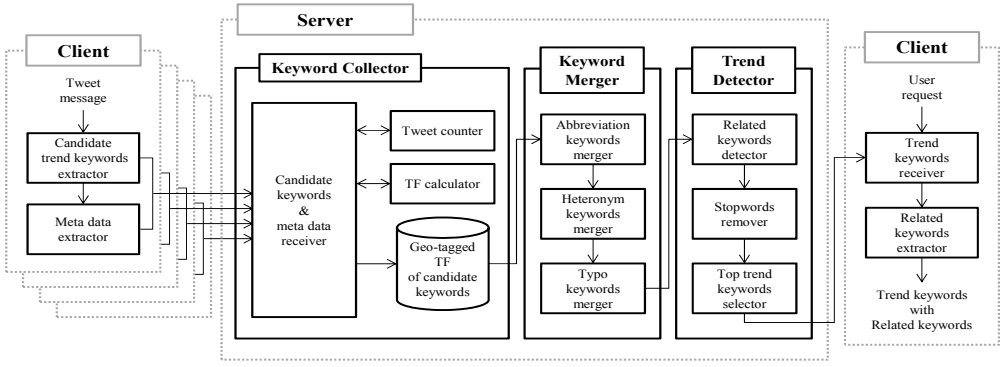
2. 시스템 구성

전체적인 시스템 구성은 그림 1과 같다. 모바일 클라이언트는 Tweet을 GPS센서 등을 통한 메타 데이터와 함께 제공하며, 이는 사용자의 위치나 상황을 판단하는 데 활용된다[3]. 한편, 서버는 해당 정보들을 이용해 트렌드 키워드를 찾아내는 역할을 한다. 이렇게 찾아진 키워드들은 사용자가 필요로 할 때 사용자의 클라이언트로 전달 된다.

키워드 대부분이 고유명사라는 점을 활용하여 후보 키워드 추출한다. 따라서, 대문자로 시작되는 연속적인 단어, 따옴표로 강조된 단어 등이 후보 키워드에 해당된다. 한편, 추출된 후보 키워드의 메타데이터로는 GPS센서를 사용하여 얻을 수 있는 사용자의 현재 위치가 포함된다.

다음으로, 전송받은 데이터를 취합하여 각 후보 키워드 단위로 Term Frequency를 계산하는 키워드 수집 과정을 수행한다. 이때, 키워드의 TF는 지역 구분을

[1] 본 연구는 지식경제부 및 정보통신사업진흥원의 대학 IT 연구센터 지원사업(NIPA-2012-H0301-12-3006)과 2012 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업(2012-0007202) 지원을 받아 수행된 것임



(그림 1) 시스템 구성도

위해 구축한 2차원 벡터의 형태로 저장된다. 또한, 각각 Tweet에 대해 고유한 ID를 부여하여 키워드 별로 ID 리스트를 생성한다.

다음으로 줄임말, 이철 동의어, 오탈자 병합 과정이 수행된다. 우선, 속도 및 정확도 향상을 위하여 낮은 TF의 단어들을 제거한다. 줄임말 병합 과정은 단어의 수가 3개 이상인 모든 후보 키워드에 대해 각 단어의 앞자리를 이용하여 병합한다. 이철 동의어의 경우 동일한 의미를 가진 의도로 사용된 단어들이라면 형태가 틀려도 비슷한 빈도로 쓰일 것이라고 가정 하에 키워드 간 빈도수가 차가 적고 한 키워드가 다른 키워드를 온전히 포함하면 하나로 병합한다. 오탈자 병합은 키워드의 글자 수 차이와 문자 히스토그램의 차이를 기반으로 판단하여 병합한다.

최종적으로, 연관 키워드 검출 및 Stopword 제거가 이루어지며 이때 방향성을 가진 그래프를 이용한다. 그래프 상 노드는 하나의 키워드를 의미한다. 에지는 키워드간의 연관성을 나타내며 키워드 A의 ID리스트 중 20% 이상의 항목이 키워드 B와 중복이 되면 A에서 B로 에지를 생성한다. 이는 파라미터를 수정해 가면서 실행한 실험 결과 가장 좋은 결과를 보인 파라미터 값을 반영한 결과이다. 특정 노드에 대한 에지의 방향을 기반으로, 다른 노드에서 해당 노드로의 에지가 반대 방향의 에지보다 월등히 많을 때 해당 키워드를 Stopword로 판단하여 제거한다. 두 노드간 서로 다른 방향의 에지 개수 차이가 크지 않고 에지 개수 역시 일정수 이상 존재할 때 이 두 노드를 연관 키워드로 간주한다. 또한, Wikipedia의 검색 API와 Google검색 결과의 HTML파싱을 이용하여 연관 키워드 확장을 실시한다.

3. 실험결과

본 실험은 실험은 Intel Core 2 Duo 2.67Ghz, 8GB 메모리 사양의 Windows 7 환경에서 수행되었으며 모바일 클라이언트에서 수행하는 작업들은 물리적인 한계 때문에 서버에서 한번에 진행하였다. 또한, 모든 실험에 사용된 Tweet들은 Twitter Streaming API로부터 얻어서 사용하였다.

표1은 2012년 8월 16일 4,795,186개의 Tweet에 대해서 Top 15 트렌드 키워드와 연관 키워드로 검색된 다양한 결과를 보여준다. 이를 통하여 당일 이슈화된 트렌드와 연관 키워드를 통하여 해당 트렌드에 대해서 직관적인 이해가 가능하다. 2012년 9월 11일 키워드의 시간대별 트렌드 점수를 분석한 결과 ANDY MURRAY가 우승한 시점에 해당 키워드가 가파르게 상승했으며, 9/11 Victims의 경우 당일 점수가 꾸준히 유지됨을 볼 수 있었다. 또한, Tweet 데이터를 분석하여 인덱스 구성에 걸리는 시간을 측정해 본 결과 단일 코어 기준 10만개의 Tweet당 평균적으로 2분이 소요되었다. 4 코어에서 대략 3.4배의 Speed up을 보임으로서 병렬화 성능이 우수하며 대용량 Tweet의 경우에도 실시간 처리가 가능할 정도로 속도 향상이 가능하다고 판단된다.

<표 1> 2012 년 8 월 16 일 Tweet 분석 결과

Trend keywords	Related keywords sets
ROBIN VAN PERSIE MANCHESTER UNITED ARSENAL	ROBIN VAN PERSIE ARSENAL MANCHESTER UNITED
JUSTIN BIEBER ONE DIRECTION JULIAN ASSANGE MEXICO LONDON	FELIX HERNANDEZ SEATTLE MARINERS RAYS PERFECT GAME MLB
ALEVEL HARRY POTTER	MELKY CABRERA TESTOSTERONE
NIALL PAUL RYAN MITT ROMNEY USA TAYLOR SWIFT	JULIAN ASSANGE WIKILEAKS ECUADOR

4. 결과

본 논문에서는 실시간으로 사용자의 Tweet을 활용하여 트렌드를 추출하고, 연관된 키워드를 제공하는 클라이언트-서버 모델을 구축하였다. 제안된 시스템으로 Tweet을 분석하여 트렌드 키워드를 추출하며 Tweet 내부 키워드간의 연관성을 통하여 Stopword 제거 및 연관 키워드를 검출해낸다. 특히 기존의 기법과 달리 빠른 처리가 가능하여 실시간으로 활용할 수 있으며, 또한 트렌드 키워드의 변동 추이 및 지역별 트렌드 검색이 가능토록 수행할 수 있음을 보였다.

참고문헌

[1] D. Ramage, S. Dumais, D. Liebling. “Characterizing microblogs with topic models”. In Proc. of the International AAAI Conf. on Weblogs and Social Media, 2010. pp. 130–137.

[2] F. Alvanaki, M. Sebastian, K. Ramamritham, G. Weikum. “EnBlogue: emergent topic detection in web 2.0 streams” In Proc. of the ACM SIGMOD International Conference on Management of data, 2011, pp. 1271-1274.

[3] D. Kim, S. Rho, E. Hwang. “Location-based large-scale landmark image recognition scheme for mobile devices,” In Proc. of the International Conf. on Mobile, Ubiquitous, and Intelligent Computing, 2012, pp. 47-52.