연관 어휘 추출을 통한 질의어 관련 이슈 탐지

김제상[0], 김동성, 조효근, 이현아
금오공과대학교
oiu124@naver.com, kaiserangel@live.co.kr, whitesky0109@naver.com, halee@kumoh.ac.kr

Query Related Issue Detection using Related Term Extraction

Je-Sang Kim[0], Dong-Sung Kim, Hyo-Geun Jo, Hyun-Ah Lee
Kumoh National Institute of Technology, Dept. of Computer Software Engineering

요 약

근래 트위터와 페이스북 등의 SNS(Social Network Service)에서 일반 대중의 관심사나 트렌드 등의 이슈를 탐지하는 많은 연구가 이루어지고 있다. 본 논문에서는 검색어에 대한 연관 어휘 추출을 통해 검색어에 연관된 이슈나 화제를 트위터에서 추출하기 위한 방법을 제안한다. 본 논문에서는 연관성이 높은 단어는 서로 가깝게 발생할 것으로 기대하고, 단어 간 거리가 가까울수록, 공기빈도가 높을수록 커지는 단어연관도 계산법을 제안한다. 연관도 값이 임계치를 넘는 어휘를 연관 어휘로 보고 네트워크의 형태로 관련 이슈를 제시한다.

주제어: 연관 어휘, 인접도 행렬, 질의어 관련 이슈, 이슈 자동 탐지, SNS.

1. 서론

SNS와 스마트 디바이스의 보급으로 인하여 실시간으로 생성되는 비정형 데이터가 급격히 늘어나고 있다. 이러한 빅데이터에 대해서는 실시간으로 생성되는 대량의 데이터를 빠르게 처리하기 위한 방법이 필요하다. SNS에 대해 근래 활발하게 시도되고 있는 분야로 실시간 이슈 탐지가 있다. [1]에서는 빅 인터렉션 데이터에서 이슈를 중심으로 현재를 모니터링하고 미래를 예측하기 위한 실제 개발 사례를 소개하고 있다.

본 논문에서는 SNS 중 트위터를 기반으로 사용자가 입력한 질의어에 관련된 이슈를 자동으로 탐지하는 방법을 제안한다. 키워드에 대한 관련 이슈를 탐지하기 위해 연관 어휘를 추출한다. 연관 어휘 추출에서는 단어 간 거리에 반비례하고 공기 빈도에 비례하는 단어 간 인접도의 합으로 연관도를 구한다.

2. 기존 연구

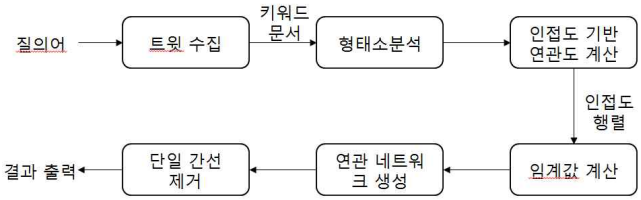
SNS에서의 이슈 또는 토픽 추출은 주로 어휘 빈도의 시간적 추이를 활용한다. 이 중 [2]와 [3]은 용어의 시간적 추이를 기준으로 권위성 등의 자질을 활용하여 떠오르는 토픽을 파악하기 위한 방법을 제안하였다. [4]에서는 특징적인 트렌드 추출을 위해 변동성, 지속성, 안정성, 누적량의 속성을 활용하여 트렌드 순위 결정 방법을 제안하여, 트렌드 탐지를 위한 방법을 제안하였다.

키워드에 대한 연관성을 통해 SNS의 비정형 문서들을 분석하여 정보를 제공해주는 서비스로는 다음소프트의 소셜 인사이트가 있다[5]. 이 서비스는 복합명사 검출과 단어 속성 등의 추출을 위해서 수작업을 필요로 하여 높은 비용이 필요하며, 최신 관련어나 복합어가 추출되지 않는다는 단점이 발견된다.

연관 단어 추출은 자동 시소러스(thesaurus) 구축이나 정보검색에서의 질의어 확장 등의 다양한 목적으로 연구되어 왔다. [6]은 뉴스에서 연관 인물명을 제시하기 위하여 문장 내 공기 어휘에 기반한 변형된 TF-IDF와 연관 규칙 마이닝을 이용한 방법을 제시하였다. [7]에서는 공기 어휘의 인접도와 빈도, IDF를 결합하여, 질의어 확장을 위한 연관어휘 추출 방식을 제안하였다. 이러한 방식은 단일 문서내 정보가 아닌 문서간 정보인 역문서빈도(IDF)를 사용하기 때문에, 대량의 실시간 문서가 발생하는 SNS환경에는 적절하지 않을 수 있다.

3. 연관 어휘를 이용한 이슈 탐지 시스템

본 논문에서는 질의어에 대한 관련 이슈 어휘를 실시간으로 추출하기 위한 방법을 제시한다. [그림 1]은 본 연구에서 제안하는 질의어에 대한 연관 어휘 추출의 단계를 보인다. 본 연구에서는 한국어 형태소 분석기를 사용하여 문서를 분석하고 형태소 분석 결과 중 명사만을 연관 어휘 후보로 사용한다.



[그림 1] 질의어에 대한 관계 명사 추출의 플로우차트

3.1 인접도에 기반한 연관도 측정

본 논문에서는 문장 길이가 짧은 SNS의 문장 특성에 맞게 [8]에서 사용한 공기 어휘간 거리 역수 즉, 인접도를 사용하여 어휘 연관도를 측정하고자 한다. 연관도 측정에서는 공기 어휘 간 행렬 모델을 제안하여, 문맥의

방향성과 어휘 간 연접 여부를 효율적으로 표현한다.

키워드 문서 내 문장 s에서 지정된 문맥의 크기 *windowSize* 안에서 발생하는 i번째 명사 w_i와 j번째 명사 w_j간의 인접도 $proximity_s(w_i, w_j)$를 1/(j-i)로 구한다. 인접도로 두 단어 사이의 거리 $j-i$의 역수를 사용하여 인접한 단어는 1에 가까운 값을, 인접도가 떨어지는 값은 0에 가까운 값을 얻는다. 이때 $j-i$의 값이 1보다 크고 *windowSize*보다 작도록 지정하여 문맥의 우측 방향에서 문맥 크기 이내의 공기 어휘만을 고려한다. 또한 w_i와 w_j가 같은 경우를 배제하여 같은 단어가 중복적으로 발생하는 경우에 의한 노이즈를 제거한다. 검색된 문서 전체에서의 인접도 $proximity(w_i, w_j)$는 전체 키워드 문서에 포함된 문장에 대한 인접도의 합으로 구한다.

얻어진 인접도를 기반으로 단어 간 연관도를 구한다. 단어 w_i와 w_j 사이의 연관도 $relatedness(w_i, w_j)$는 $proximity(w_i, w_j) + proximity(w_j, w_i)$로 구하여, 문맥의 우측 방향과 좌측 방향에 대한 인접도를 합산한다.

[표 1] '싸이'에 대한 연관 행렬 추출의 예

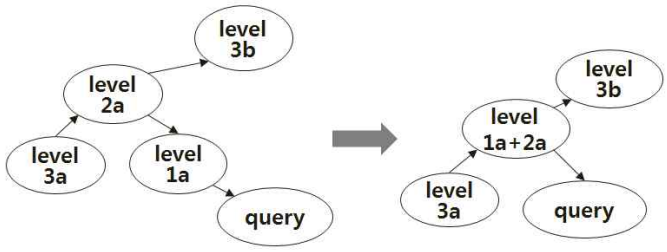
	임윤택	부담	싸이	전액	장례비용	대통령	취임식	당신	멋쟁이
임윤택	0	266.8	28	332.6	809.8	0	0	71.6	71.6
부담	0.2	0	28.4	0	0	0	0	358	143.6
싸이	441.4	90.2	0	31.2	13	0.8	8.8	71.6	0
전액	0.2	933	15.6	0	0	0	0	0	0
장례비용	4.4	307.2	28.2	638.6	0	0	0	0	0
대통령	0	0	7.4	0	0	0	384.4	0	0
취임식	0	0	18.2	0	0	0	0	0	0
당신	0	0	1.4	0	0	0	0	0	358
멋쟁이	0	0	1	0	0	0	0	0	0
단독	174.8	86.6	0	87.4	87.4	0	0	0	0

〔표 1〕은 2013년 2월 15일 기준으로 추출한 '싸이'에 대한 키워드 문서에서의 *windowSize*=4에서의 인접도 행렬을 보인다. *relatedness*(싸이, 임윤택)은 *proximity*(싸이, 임윤택) + *proximity*(임윤택, 싸이) = 441.4 + 28로 구할 수 있다. 인접도 행렬에서는 명사 간의 연관성의 경로를 쉽게 파악할 수 있다.

3.2 연관 어휘 판별 및 네트워크 생성

제안하는 시스템에서는 키워드 문서에서 인접도를 이용하여 얻어진 어휘간 연관도를 계산하고, 질의어와 연관도가 높은 단어를 '연관 어휘'로 결정하여 사용자에게 제시한다.

질의어 *query*에 대해서 연관도 $relatedness(query, w_i)$가 1차 연관 어휘 임계값 이상인 단어 w_i를 구하고, 구해진 w_i에 대해서 연관도 $relatedness(w_i, w_j)$가 2차 연관 어휘 임계값 이상인 w_k를 찾는다. 얻어진 1차 연관 어휘 w_i, 2차 연관 어휘 w_j, 3차 연관 어휘 w_k에 대해서 〔그림 2〕의 왼쪽과 같은 그래프를 구한다. 그림에서 level1은 1차, level2는 2차, level3는 3차 연관 어휘를 표시한다. 네트워크는 인접도 행렬의 좌우 문맥의 방향성에 따라 방향 네트워크의 형태로 생성한다. 〔그림 2〕에서 level1a는 level2a 이외의 연결 간선이 없는 단일 간선을 구성한다. 〔표 1〕의 '당신'과 '멋쟁이', '대통령'과 '취임식' 같은 단일 간선에 대해서, 방향 그래프의 간선 방향을 고려하여 명사를 조합하면 '당신 멋쟁이', '대통령 취임식'과 같은 복합어를 구성할 수 있다. 시스템에서는 단일 간선을 통합하고 2차 관계까지 만을 남겨 〔그림 2〕의 오른쪽과 같은 결과 네트워크를 구성한다.



[그림 2] 단일 간선과 3차 관계를 제거한 결과 네트워크

4. 실험 및 평가

4.1 질의어 관련 이슈 추출 시스템

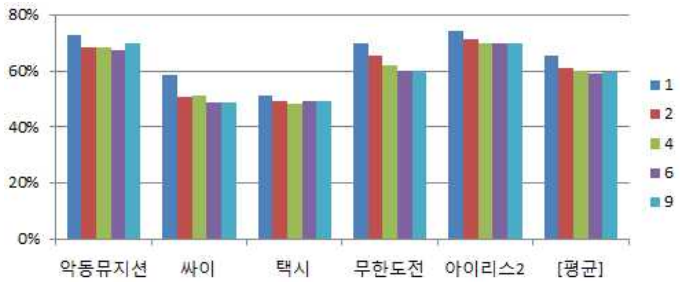
시스템에서는 사용자 가독성을 고려하여 1차와 2차 연관 어휘로 최대 10개의 결과를 제공한다. 최신성이 중요한 이슈의 특성에 맞추어 트위터와 동일하게 시스템도 7일간 자료만을 저장한다. 형태소 분석기로 [9]를 사용하고 트위터 검색에서는 [10]을 사용하였다.

제안한 방식의 성능을 평가하기 위해 2013년 2월 14일부터 2월 20일까지의 기간 중 5개 키워드(악동뮤지션, 싸이, 택시, 무한도전, 아이리스2)를 중심으로 평가를 시행하였다.

4.2 문맥 크기에 따른 성능 평가

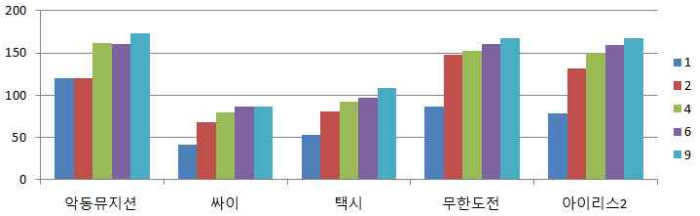
문맥의 크기에 따른 성능을 비교하기 위한 실험을 수행하였다. 10개의 키워드를 대상으로 실험군을 형성하기 위해 연관 어휘 분석 모듈의 1차와 2차 연관 어휘의 가중치를 1% 즉 0.01으로 고정하고, *windowSize*를 1, 2, 4, 6, 9로 변화시켜, 인접 체언의 범위에 따라 연관 어휘 추출에 어떤 영향을 미치는지 살펴보았다.

[그림 3]은 문맥 크기에 따른 시스템의 정확도를 보인다. 이슈성에 근거하여 엄격하게 수동 평가한 결과에서, 상위 10위까지에서 대부분 정확한 연관 어휘가 추출된 것으로 볼 수 있다. 문맥의 크기를 키우더라도 추가적인 연관 어휘가 추출되지 않아 성능 차이가 크지 않았다. 또한 평균적인 정확률이 60%이상으로 실용성 있는 이슈 추출이 가능함을 볼 수 있다.



[그림 3] windowSize에 따른 연관 어휘 정확도

보다 정확한 분석을 위하여 연관도 상위 50개를 추출한 경우를 분석하였다. 결과는 [그림 4]와 같다. 추출되는 연관 어휘 개수는 최대 7일 * 50 = 350개가 나올 수 있으나, 지속성이 큰 이슈들로 최대 230여개가 추출되는 것을 볼 수 있다. 결과에서는 windowSize가 증가함에 따라 추출되는 연관 어휘 개수가 증가하는 모습이 보다 뚜렷하다.



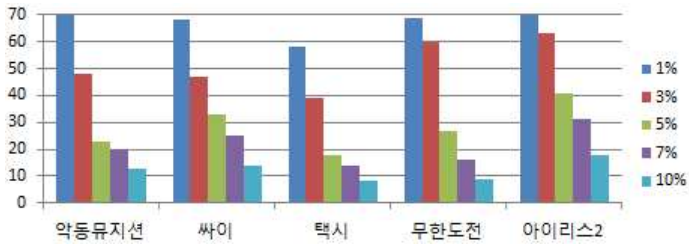
[그림 4] 상위 50개 대상 windowSize에 따른 연관 어휘 개수

상위 50개에 대한 정확도 분석에서는 수동 평가로 인해 키워드 '아이리스2'에 대한 결과만을 분석하였다. 결과에서는 문맥 크기가 1인 경우 80%의 정확도를, 문맥크기가 2~9의 경우 61%~65%의 정확도를 보여 문맥크기를 키워 연관 어휘 개수가 많아지는 만큼, 부정확한 연관 어휘도 많이 추출되는 것으로 나타났다. 큰 문맥에 의해서는 어휘에 노이즈가 포함될 여지가 커지는 동시에 인접도 행렬에 저장되는 값이 많아져서 시스템 속도가 느려질 수 있으며, 작은 문맥에 의해서는 명확한 연관 어휘만 추출되지만 폭넓은 연관 어휘를 추출할 수 없는 문제가 있어, 응용 분야에 맞는 적절한 문맥 크기의 선택이 필요한 것으로 분석되었다.

4.3 연관도 임계치에 따른 성능 평가

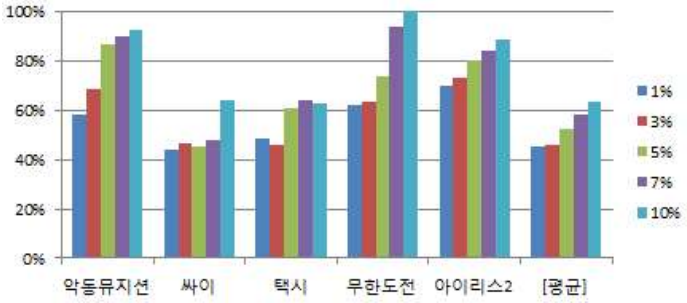
연관 어휘 추출을 위한 임계치에 따른 실험을 수행하였다. 실험 환경에서 제시한 10개의 키워드를 대상으로 *windowSize*의 값을 4로 고정시키고 1차 연관 어휘 추출을 위한 가중치를 1, 3, 5, 7, 10%로 변화함에 따른 연관 어휘 발생 빈도를 분석하였다.

[그림 5]는 가중치 변화에 따른 연관 어휘 추출 개수를 보여준다. 가중치의 임계치 값을 높게 할수록 연관 어휘 개수가 점차적으로 줄어든다는 것을 알 수 있다.



[그림 5] 가중치에 따른 상위 10개 연관 어휘 개수

임계값에 따른 정확도에 대한 평가는 [그림 6]과 같다. 결과는 키워드 5개에 대하여 가중치 변화에 따른 이슈성이 있는 사건의 비율 즉, 정확도를 나타낸다. 결과는 평균 57%~82%의 연관 어휘 추출의 정확도를 보여, 인접도 행렬을 이용한 빠른 처리에도 불구하고 실용성 있는 결과를 얻을 수 있었다. 또한 가중치와 임계치가 증가함에 따라 연관 어휘 개수가 줄어들지만 연관성이 높은 사건을 추출하는 정확도는 증가하는 것으로 나타났다. 이는 임계치가 높아 추출되는 연관 어휘 개수가 줄어들어 발생하는 효과로 볼 수 있다. [그림 5]에서 가중치가 1%에서 3%로 증가되는 경우 연관 어휘 개수는 20% 정도 감소하는 것에 비해, [그림 6]에서는 정확도의 차이는 크지 않은 것으로 나타났다. 또한, 가중치가 1%에서 10%로 증가되는 경우 추출되는 연관 어휘가 1/10의 수준으로 감소하는 것에 비해, 정확도는 20% 정도 증가하는 것으로 파악되었다. 문맥 크기와 마찬가지로 응용 분야에 맞는 가중치의 설정이 필요할 것으로 보인다.



[그림 6] 가중치에 따른 연관 어휘 정확도

5. 결론

본 논문에서는 질의어에 대한 트윗 문서에서의 연관 어휘 추출 방식을 제안하였다. 연관 어휘 추출에서는 질의어와 가까운 거리에서 자주 발생하는지의 여부가 중요하다는 점에 착안하여, 인접도의 합으로 단어 간 연관도를 얻었다. 형태소 분석된 문서를 순차적으로 한번만 탐색하면서 인접도 행렬을 구할 수 있어 빠른 시간 안에 결과를 얻을 수 있었다. 또한, 연관 관계에 따른 네트워크 형성 후 간선 조정을 통해 복합어를 추출하여 비교적 정확하면서 손쉬운 방법으로 동적으로 생성되는 복합어 처리 결과를 얻을 수 있었다. 실험에서는 임계치와 문맥의 크기에 따른 영향성을 분석하였다. 전반적으로 높은 성을 보여 비교적 간단한 방법을 사용하는데도 불구하고 실용성 있는 결과를 얻을 수 있었다.

참고문헌

[1] 류법모, 김현진, 김현기, 박상규, “심층 언어분석 기반 소셜미디어 이슈 탐지 및 모니터링 기술”, 한국정보과학회지, 제30권, 제6호, pp. 47-58, 2012.

[2] Mario Cataldi, Luigi Di Caro, Claudio Schifanella, "Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation", Proceedings of the 10th International Workshop on Multimedia Data Mining at KDD, 2010.

[3] Michael Mathioudakis, Nick Koudas, "TwitterMonitor: trend detection over the twitter stream", Proceedings of the ACM SIGMOD International Conference on Management of data, pp. 1155-1158, 2010.

[4] 오흥선, 최윤정, 신욱현, 정윤재, 맹석현, "자동 트

렌드 탐지를 위한 속성의 정의 및 트렌드 순위 결정 방법", 정보과학회논문지 : 소프트웨어 및 응용, 제36권, 제3호, pp. 236-243, 2009.
[5] 다음소프트, "http://insight.some.co.kr/searchKeywordMap.html"
[6] 김한준, 장재영, "연관규칙 마이닝을 활용한 뉴스기사 키워드의 연관성 탐사", 한국인터넷방송통신학회 논문지, 제11권, 제6호, pp. 63-71, 2011.
[7] Tetsuya Oishi, Shunsuke Kuramoto, Tsunenori Mine, Ryuzo Hasegawa, Hiroshi Fujita, Miyuki Moshimura, "A Method for Query Expansio Using the Related Word Extraction Algorithm", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008.
[8] 정석팔, 임성현, 전진형, 김병만, 이현아, "요약문을 이용한 웹 검색 결과 군집화", 정보과학회논문지:데이타베이스, 제39권, 제5호, pp.321-331, 2012.
[9] Lucene Core 4.0 and SolrTM 4.0 Available, "http://lucene.apache.org", 12 October 2012
[10] 트윗 검색 시스템, "https://dev.twitter.com/"