KIPS Tr. Software and Data Eng.
Vol.3, No.1 pp.31~36 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2014.3.1.31

Related Term Extraction with Proximity Matrix for Query Related Issue Detection using Twitter

Je-Sang Kim† · Hyo-Geun Jo† · Dong-Sung Kim† · Byeong Man Kim†† · Hyun Ah Lee†††

ABSTRACT

Social network services(SNS) including Twitter and Facebook are good resources to extract various issues like public interest, trend and topic. This paper proposes a method to extract query-related issues by calculating relatedness between terms in Twitter. As a term that frequently appears near query terms should be semantically related to a query, we calculate term relatedness in retrieved documents by summing proximity that is proportional to term frequency and inversely proportional to distance between words. Then terms, relatedness of which is bigger than threshold, are extracted as query-related issues, and our system shows those issues with a connected network. By analyzing single transitions in a connected network, compound words are easily obtained.

Keywords : Related Term, Proximity Matrix, Query Related Issue, Issue Detection, SNS

트위터를 이용한 질의어 관련 이슈 탐지를 위한 인접도 행렬 기반 연관 어휘 추출

김 제 상† · 조 효 근† · 김 동 성† · 김 병 만†† · 이 현 아†††

요 약

트위터와 페이스북 등의 SNS(Social Network Service)는 일반 대중의 관심사나 트렌드 등의 이슈를 탐지하기 좋은 지식원이다. 본 논문에서는 검색 질의어에 관련된 이슈나 화제를 질의어에 대한 연관 어휘로 보고, 이를 트위터에서 추출하기 위한 방법을 제안한다. 제안하는 방법에서는 질의어와 연관성이 높은 단어는 질의어와 가까운 위치에서 자주 발생한다고 가정하고, 단어 간 거리에 반비례하고 공기 빈도에 비례하는 단어 간 인접도의 합으로 단어간 연관도를 구한다. 구해진 연관도 값이 임계치를 넘는 어휘를 연관 어휘로 보고 네트워크의 형태로 관련 이슈를 제시한다. 제안한 방법에서는 네트워크의 특성을 분석하여 복합어를 손쉽게 탐지할 수 있다.

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

1. 서 론

SNS와 스마트 디바이스의 보급으로 실시간으로 생성되는 비정형 텍스트 데이터와 빅데이터에 대한 관심이 크게 증대되고 있다. 빅데이터에 대한 다양한 접근에서 가장 큰 난제는 실시간으로 생성되는 데이터의 양에 있다. 실제 7일간 트위터의 트윗을 수집한 결과 25기가바이트의 데이터가 얻어졌으며, 이러한 대량의 데이터에서 원하는 정보를 수집하기 위해서는 빠른 처리 속도가 요구된다.

SNS를 활용하는 다양한 연구는 사용자간 친밀도 분석에서 오피니언 마이닝까지 폭넓게 이루어지고 있으며, 이 중 실시간 이슈 탐지가 근래에 활발하게 시도되고 있다[1]. SNS에서의 이슈나 토픽을 추출하기 위한 기존 연구에서는 용어의 시간적 추이를 기준으로 권위성 등의 자질을 활용하여 토픽을 파악하는 방법[2][3]이나, 특징적인 트렌드 추출을 위해 변동성, 지속성, 안정성, 누적량의 속성을 활용하여 트렌드 순위 결정 방법[4] 등이 제안되었다. 키워드에 대한 연관성을 통해 SNS의 비정형 문서들을 분석하여 정보를 제공해주는 다음소프트의 소셜 인사이트[5]에서는, 키워드와 함께 등장한 빈도에 기반하여 연관 명사의 순위가 제시되며, 색깔을 사용하여 해당 단어의 속성 정보를 제공한다. 하지만 이 접근에서는 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 연관규칙 마이닝

※ 본 연구는 금오공과대학교 학술연구비에 의하여 연구된 논문임.
† 준 회 원 : 금오공과대학교 컴퓨터공학부 학부생
†† 종신회원 : 금오공과대학교 컴퓨터소프트웨어공학과 교수
††† 종신회원 : 금오공과대학교 컴퓨터소프트웨어공학과 부교수
논문접수 : 2013년 8월 12일
수 정 일 : 1차 2013년 10월 22일
심사완료 : 2013년 11월 14일
* Corresponding Author : Hyun Ah Lee(halee@kumoh.ac.kr)

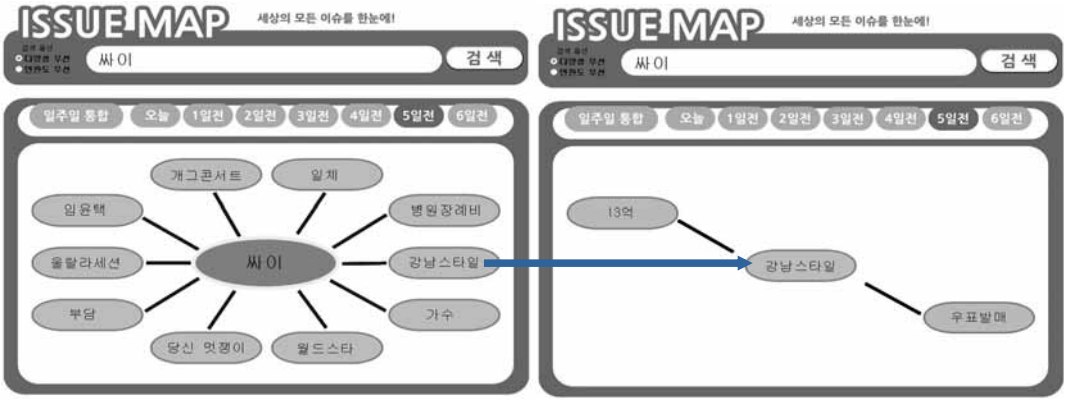

Fig. 1. Example of 1st related issue extraction for '싸이'

을 이용하였다. [7]에서는 공기 어휘의 인접도와 빈도, IDF를 결합하여 질의어 확장(query expansion)을 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. [8]은 학술 논문 간의 지식 관계를 지도 형태로 표현하기 위해 키워드 연관 네트워크를 사용하였다. 이는 전문 분야에 대해서는 좋은 성능을 보장할 수 있으나, 대용량의 데이터에 대한 빠른 처리를 보장하기 어렵다는 단점이 있다. [9]와 [10]에서는 질의어에 대한 대량의 클릭 로그를 이용하여 "070 인터넷전화"와 같은 연관 키워드를 추출했다. 이러한 연구는 언어 독립적으로 연관 키워드를 수집할 수 있으며 높은 성능을 보장한다는 장점이 있지만, 과거 데이터와는 독립적인 새롭게 발생하는 이슈에 대해서는 취약한 단점이 있다.

본 논문에서는 SNS 중 트위터를 기반으로 사용자가 입력한 질의어에 관련된 이슈를 자동으로 탐지하는 방법을 제안한다. 시스템에서는 인접도 행렬을 이용하여 빠른 시간에 단어간 연관도를 계산한다. Fig. 1은 시스템의 구동 예를 보인다. 시스템에서는 트위터에서 제공하는 검색 기능을 이용하여 사용자 질의어를 포함하는 1주일 이내의 트윗을 자동으로 분석하여, 질의어에 관련된 최근 이슈를 자동으로 분석하여 제시한다. 최초에 제시되는 1차적인 연관 이슈를 클릭하면 2차 연관 이슈들이 네트워크 형태로 제시된다.

본 논문은 다음과 같이 구성된다. 2장에서는 인접도와 연관도를 이용한 연관 어휘 탐지에 의한 질의어 관련 이슈 탐색 방법을 설명한다. 3장에서는 실험 및 평가를 보이고 4장에서는 결론을 맺는다.

2. 연관 어휘를 이용한 이슈 탐지 시스템

본 논문에서는 질의어에 대한 관련 이슈 어휘를 실시간으로 추출하기 위한 방법을 제시한다. 앞에서 살펴보았듯이 기존의 방법은 대량의 데이터를 필요로 하여 실시간성이 강한 이슈 추출에 적합하지 않은 단점이 있다. TF-IDF를 사용하는 방식은 비교적 빠르게 결과를 낼 수 있지만 실시간으로 생성되는 대량의 데이터에서 적합한 IDF를 구하기 쉽지 않고, [11]의 연구 결과에서 본 바와 같이 검색 결과 문서에서의 IDF는 정보 판별력에 악영향을 미칠 수 있다. 본 논문의 연관 어휘 자동 추출에서는 어휘 간 거리와 공기 빈도, 빈도의 비율 정보만을 사용하여 빠른 속도를 보장한다.

Fig. 2는 본 논문의 연관 어휘 추출의 단계를 보인다. 첫 번째 단계에서는 사용자가 입력한 질의어를 대상으로 질의어를 포함한 트윗을 추출한다. 추출된 트윗을 질의어 즉 키워드를 포함한 문서로 보고 이를 키워드 문서라 한다. 두 번째 단계에서는 키워드 문서를 형태소 분석하여 빈도와 인접도를 기준으로 질의어에 대한 공기 어휘의 연관도를 계산한다. 비교적 짧게 작성되는 트윗에서는 중요 이슈가 대부분 명사로 표현되므로, 본 연구에서는 형태소 분석 결과 중 명사만을 연관 어휘 후보로 사용한다. 제안하는 방식은 인접도 행렬을 통해 연관도를 계산하여 빠른 시간을 보장하는 장점을 가진다. 아래에서는 인접도 계산과 이후 단계에 대해서 상세하게 설명한다.

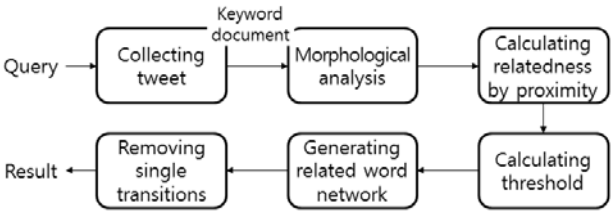

Fig. 2. Flowchart of extracting related terms

2.1 인접도에 기반한 연관도 측정

키워드 문서에서 키워드를 제외하고 가장 많이 등장하는 명사는 키워드 즉 질의어와 관련성이 높을 것으로 예상할 수 있다. Fig. 3은 트윗에서 '싸이'와 '악동뮤지션'의 공기 어휘를 빈도의 내림차순으로 나열한다. 그림에서 볼 수 있듯이 공기 빈도가 높은 명사들이 키워드와 관련성 있는 이슈가 될 확률이 높다. 하지만 빈도만을 사용하여 연관성을 측정하는 경우 부정확한 결과를 얻을 수 있다. [11]에서는 검색 결과의 요약(snippet)을 이용한 클러스터링에서 공기 어휘와 질의어 간 거리의 역수인 인접도의 합을 사용하여, 공기 어휘의 IDF보다 좋은 성능을 얻었다. 이는 짧게 구성되는 요약의 특성상 중요 어휘가 다수 추출되어(예를 들어 '이클립스'의 경우 '자바'나 '설치'), IDF가 부정적인 영향을 미치기 때문이었다.

본 논문에서는 문장 길이가 짧은 SNS의 문장 특성에 맞게 공기 어휘간 거리 역수인 인접도를 사용하여 어휘 연관도를 측정하고자 한다. 연관도 측정에서는 공기 어휘 간 행

Word	Freqency
싸이	361
싸이블로그	227
싸이코	206
강남스타일	197
우표	168
싸이월드	164
트위터	135
신천지	103
싸이코패스	99
카톡	96
1위	96
방해하지마	93
진짜	86
유튜브	83

(a) 질의어 '싸이'

Word	Freqency
악동뮤지션	1703
라면인건가	435
투표	164
네티즌	138
조윤선	137
김종훈	136
생방송	108
K팝스타2	103
앤드류최	102
노래	88
근신7	77
부족	77
징계처분	77
최예근	74

(b) 질의어 '악동뮤지션'

Fig. 3. Word frequency in keyword document for a query

럴 모델을 제안하여, 문맥의 방향성과 어휘 간 연접 여부를 효율적으로 표현한다.

키워드 문서의 문장 s에서 지정된 문맥 크기 *windowSize* 안에서 발생하는 i번째 명사 w_i와 j번째 명사 w_j간의 인접도 $proximity_s(w_i, w_j)$는 $1/(j-i)$을 그 값으로 한다. 인접도로 두 단어 사이의 거리 $j-i$의 역수를 사용하여 인접한 단어는 1에 가까운 값을, 인접도가 떨어지는 값은 0에 가까운 값을 얻는다. $j-i$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 중복 발생하는 경우에 의한 노이즈를 제거한다. 검색 문서 전체에서의 인접도 $proximity(w_i, w_j)$는 전체 키워드 문서 내 모든 문장의 인접도 합으로 구한다.

Table 1은 문장 "개성공단 출입 정상 북한군 귀순 여파 주시"에 대하여 *windowSize*가 1인 경우와 4인 경우의 인접도 계산의 예를 보인다. *windowSize*가 1인 경우는 바로 뒤 단어에 대한 정보만을, 4인 경우에는 뒤의 네 단어까지의 정보를 인접도 행렬로 저장한다. Table 1의 (b)에서 (북한군, 여파)는 '북한군' 뒤에 나타나는 '여파'가 0.5의 연관도로 나타나며, (출입, 북한군)은 '북한군' 앞에 나타나는 '출입'의 연관도를 0.5로 나타낸다. 이처럼 하나의 행렬을 행 중심으로 보면 문맥의 우측 방향의 인접도를 파악할 수 있게 하고 열 중심으로 보면 문맥의 좌측 방향의 인접도를 파악할 수 있다.

얻어진 인접도를 기반으로 단어 간 연관도를 구한다. 단어 w_i와 w_j간 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.

Table 1. Relatedness extraction for "개성공단 출입 정상 북한국 귀순 여파 주시"

	개성공단	출입	정상	북한군	귀순	여파	주시
개성공단		1.0					
출입			1.0				
정상				1.0			
북한군					1.0		
귀순						1.0	
여파							1.0
주시							

(a) windowSize = 1 의 경우

	개성공단	출입	정상	북한군	귀순	여파	주시
개성공단		1.0	0.5	0.33	0.25		
출입			1.0	0.5	0.33	0.25	
정상				1.0	0.5	0.33	0.25
북한군					1.0	0.5	0.33
귀순						1.0	0.5
여파							1.0
주시							

(b) windowSize = 4 의 경우

Table 2는 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. 예를 들어 '싸이'의 행과 열에서 '임윤택'→'부담'→'전액'→'장례비용'→'취임식'의 순서로 연관 어휘를 얻을 수 있다. 또한 '싸이'에 대한 키워드 문서에서의 '임윤택'에 대해서 '장례비용'→'전액'→'부담'→'단독'을 얻을 수 있다. Fig. 4는 이를 네트워크 형태로 표현한다. 추출된 네트워크에서는 '싸이'→'임윤택'→'장례비용'→'전액'→'부담' 그리고 '당신'→'멋쟁이', '대통령'→'취임식'의 연관관계를 쉽게 파악할 수 있다. 본 논문에서는 빈도의 상대 비율을 기준으로 인접도의 임계치(threshold)를 계산하여 연관 어휘를 최종적으로 결정한다. 아래에서 임계치 계산 방법을 설명한다.

Table 2. Relatedness Matrix Extraction for '싸이'

	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

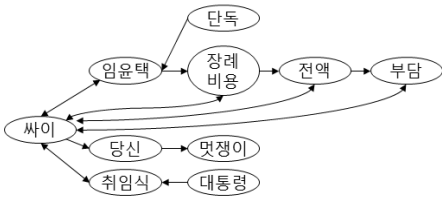

Fig. 4. Related word network for '싸이'

2.2 연관 어휘 판별 및 네트워크 생성

시스템에서는 연관 어휘를 판별하는 기준으로 질의어가 키워드 문서에서 나타난 빈도에 대비한 연관도의 비율을 임계치로 사용한다. 1차 연관 어휘와 2차 연관 어휘 추출을 위한 임계치는 질의어 빈도에 각각을 위한 가중치를 곱하여 구한다. 각 가중치는 마다 기준이 되는 상수값으로 시스템에서 설정한다.

시스템에서는 질의어 query에 대한 연관도 $relatedness(query, w_i)$가 1차 연관 어휘 임계치 이상인 w_i를 구하고, 구해진 w_i에 대해서 연관도 $relatedness(w_i, w_j)$가 2차 연관 어휘 임계치 이상인 w_j, 연관도 $relatedness(w_j, w_k)$가 3차 연관 어휘 임계치 이상인 w_k를 찾는다. 얻어진 1차 연관 어휘 w_i, 2차 연관 어휘 w_j, 3차 연관 어휘 w_k에 대해서 Fig. 5의 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.

Fig. 5에서 level1a는 level2a 이외의 연결 간선이 없는 단일 간선을 구성한다. Fig. 4에서는 '당신'과 '멋쟁이', '대통령'과 '취임식'이 단일 간선을 구성하며, 이러한 단일 간선의 두

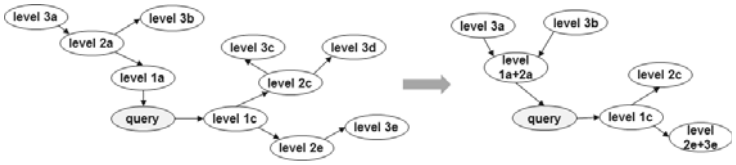

Fig. 5. First and Result Network without Single Transition and the 3rd Relation

단어는 연관되는 다른 어휘가 없다고 볼 수 있다. 이러한 단일 간선에 놓인 단어를 방향 그래프의 간선 방향을 고려하여 조합하면 '당신 멋쟁이', '대통령 취임식'과 같은 복합어를 구성할 수 있다. Fig. 5에서 level2c와 level3c와 3d는 단일간선이 아니기 때문에 다른 노드와 조합되지 않고, level3c와 level3d는 3차 관계로 그대로 남겨진다. 시스템에서는 단일 간선을 통합한 뒤에 2차 관계까지 만을 사용하여 Fig. 5의 오른쪽과 같은 결과 네트워크를 구성한다.

3. 실험 및 평가

Fig. 1과 같이, 제안하는 시스템에서는 사용자 가독성을 고려하여 연관도 상위 10개까지의 1차와 2차 연관 어휘를 제공한다. 시스템은 최신성이 중요한 이슈의 특성에 맞추어 7일간의 트위터의 정보만을 사용하며, 날짜별 캐시를 사용하여 기분석된 결과를 저장하여 사용한다. 결과에서 '일주일 통합'은 캐시에 저장되어 있는 7일간의 데이터의 연관도를 단순 합산하여 상위 10개의 결과를 보인다. 시스템에서는 형태소 분석기로 [12]를 사용하고 트위터 검색에서는 [13]을 사용하였다. 시스템에서는 형태소 분석기에서 대명사를 제외한 모든 종류의 명사를 연관어휘 후보로 추출한다. [12]는 세종 코퍼스 임의의 100 문장에 대하여 명사 추출의 정확도로 79.9%를 보였다.

추출된 연관 어휘의 정확성 평가를 수동으로 진행하였다. Fig. 6은 추출된 어휘에 대한 정확성 평가의 예로, 2013년 10월 21일에 '아이폰'을 검색어로 사용하여 얻어진 키워드 문서에 대하여, *windowSize*의 값을 1, 연관 어휘 추출을 위한 가중치를 5%로 설정하여 얻은 1차 연관 어휘 추출 결과이다. 그림에서는 이슈 여부를 기준으로 이슈가 아닌 것으로 평가된 단어는 어둡게 표시하였다. 그림에서 볼 수 있듯이 "애플 공개, 아이폰5S 첫 TV 광고 영상","빈티지 카메라를 아이폰으로 조작한다?"등과 같은 기사화 된 사실은 이슈성이 있다고 보았고, "[위치추적앱 진돗개] 트윗 로고"같은 이슈화된 사실이 포함되어 있지 않는 트위에 대해서는 이슈성이 없다고 보았다. 평가에서는 이와 같이 추출된 연관 어휘가 이슈를 반영하는지를 엄격하게 판단하였다.

제안한 방식의 문맥 크기와 연관도의 임계치에 따른 성능을 평가하기 위해 2013년 2월 14일부터 2월 20일까지의 기간 중 10개 키워드(악동뮤지션, 싸이, 동아쇼핑, 무한도전, 아이리스2, 택시, 윤하, 유재석, 장혁, 패션)를 중심으로 평가를 시행하였다.

2013.10.21	1day before	2days before	3days before	4days before
날씨 메인	아이패드	사진 경우	사람	판매자
광고	출시	출시	예약	중고나라
조작	화면	광대역LTE	충전기	접속
갤럭시	갤럭시	갤럭시	갤럭시	갤럭시
애플	애플	애플	애플	애플
사람	조작	스마트폰	안드로이드	G2 모바
개발자	사진	판매	친일인 명사전	사용자
빈티지 카메라	빈티지 카메라	예약	스마트폰 어플로	생산
위치추적앱 진돗개	역시	예약판매	위치추적앱 진돗개	위치추적앱 진돗개
아이폰5s 예약 대한민국	Unknown 이게 지금	역시	아이폰5s 오늘	고급형 아이폰5S

Fig. 6. 1st related word extracted for query '아이폰'

문맥의 크기에 따른 성능을 비교하기 위한 실험을 수행하였다. Fig. 7은 1차와 2차 연관 어휘의 가중치를 1% 즉 0.01으로 고정하고, *windowSize*를 1, 2, 4, 6, 9로 변화시켜, 인접 체언의 범위가 연관 어휘 추출에 어떤 영향을 미치는지 보인다. 날짜별로 10개의 연관어휘를 추출할 수 있어 7일간에는 최대 70개의 결과를 얻을 수 있지만, 여러 날짜 동안 지속되는 연관어가 많은 경우 70보다 작은 결과를 낸다. 3개의 키워드(싸이, 택시, 패션)를 제외한 나머지 키워드에 대해서는 *windowSize*가 커짐에도 불구하고 변화가 뚜렷하지 않아, 키워드별 차이를 보였다. '동아쇼핑'은 추출되는 연관 어휘 수가 작게 나오는데, 이는 일회성 이슈인 '동아쇼핑 화재'에 대하여 트윗에서 발생되는 연관 어휘 자체가 작은 이유로 분석되었다. 이 결과는 짧은 SNS 문장의 특성으로 문맥 크기를 키우더라도 연관성이 떨어지는 어휘는 추출하지 않는다는 장점을 입증한다고 볼 수 있다.

Fig. 8은 문맥 크기에 따른 시스템의 정확도를 보인다. 평가에서는 위에서 설명한 바와 같이 이슈 여부를 엄격하게 판단하였으며, 수동 평가의 어려움으로 5개의 단어에 대하여 평가를 수행하였다. 문맥의 크기를 키우더라도 Fig. 7에서 본 바와 같이 추가적인 연관 어휘가 추출되지 않아 성능 차이가 크지 않았다. 또한 평균적인 정확률이 60%이상으로 실용성 있는 이슈 추출이 가능함을 볼 수 있다.

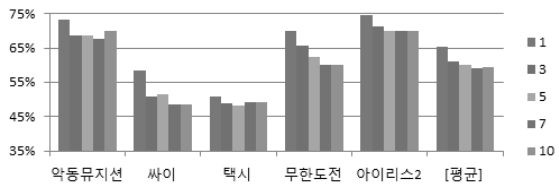

Fig. 8. Correctness of related word extraction according to windowSize

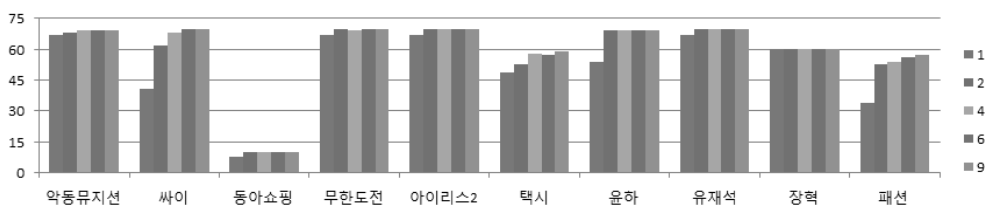

Fig. 7. Number of related words according to windowSize with top 10 words

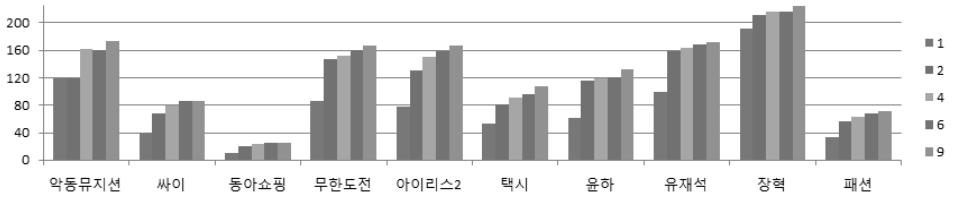

Fig. 9. Number of related words according to windowSize with top 50 words

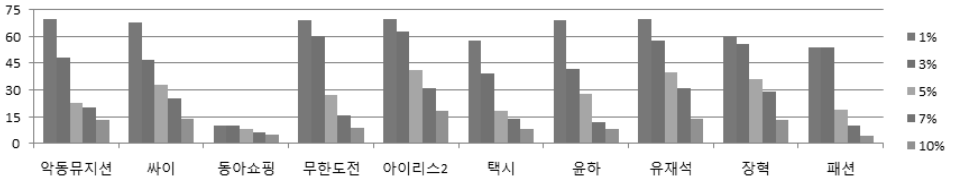

Fig. 10. Number of Related Words according to weight with top 10 words

보다 정확한 분석을 위하여 연관도 상위 50개를 추출한 경우를 분석하였다. 결과는 Fig. 9와 같다. 추출되는 연관 어휘 개수는 최대 7일 * 50 = 350개가 나올 수 있으나, 지속성이 큰 이슈들로 최대 230여개가 추출되는 것을 볼 수 있다. 결과에서는 *windowSize*가 증가함에 따라 추출되는 연관 어휘 개수가 증가하는 모습이 보다 뚜렷하다.

상위 50개에 대한 정확도 분석에서는 '아이리스'에 대한 결과를 수작업으로 분석하였다. 결과에서는 문맥 크기가 1인 경우 80%의 정확도를, 문맥 크기가 2~9의 경우 65%~61%의 정확도를 보여 문맥 크기를 키워 연관 어휘 개수가 많아지는 만큼, 부정확한 연관 어휘도 많이 추출되는 것으로 나타났다. 큰 문맥에 의해서는 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.

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. 가중치 변동에 따른 '악동뮤지션'에 대한 추출 결과에서는 임계치를 1~3%로 한 경우 연관 어휘로 "자작곡, 생방송 2등, 라면인건가, 폰매장, 올레송, 생방송, 고백하려해, T월드, 말도안돼 알티공약, 노래"가 추출되었으나, 10%로 한 경우 "자작곡, 생방송 2등, 라면인건가"가 연관 어휘로 추출되었다. 임계치를 높일수록 연관도가 높은 어휘만 추출되어 결과 개수는 줄어들고 정확도는 높아짐을 볼 수 있다. Fig. 10은 가중치 변화에 따른 연관 어휘 추출 개수를 보여준다. 가중치의 임계치 값을 높게 할수록 연관 어휘 개수가 점차적으로 줄어든다는 것을 알 수 있다.

임계치에 따른 정확도에 대한 평가는 Fig. 11에서 보인다. 결과에서 임계치가 높을수록 높은 정확도를 보였으며, 평균 정확도는 57%~82%를 보여 인접도 행렬을 이용한 빠른 처리에도 불구하고 실용성 있는 결과를 얻을 수 있었다. 결과 중 '싸이'는 관련 이슈가 매우 다양하여, '택시'는 이슈성이 떨어져서, 낮은 성능을 보이는 것으로 분석되었다. Fig. 10에서 가중치가 1%에서 3%로 증가되는 경우 연관 어휘 개수는 약 20% 감소하는 것에 비해, Fig. 11에서는 정확도의 차이는 크지 않은 것으로 나타났다. 또한, 가중치가 1%에서 10%로 증가되는 경우 추출되는 연관 어휘가 1/10의 수준으로 감소하는 것에 비해, 정확도는 약 20% 증가하는 것으로 파악되었다. 문맥 크기와 마찬가지로 응용 분야에 맞는 가중치 설정이 필요할 것으로 보인다.

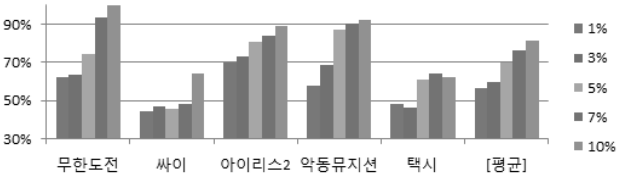

Fig. 11. Correctness of related word according to weight value

Fig. 12는 2013년 10월 21일부터 일주일간의 "아이유"에 대한 본 시스템의 결과와 [5]의 결과를 비교한다. 본 시스템의 *windowSize*의 값을 1, 연관 어휘 추출을 위한 가중치를 5%로 설정하여 결과를 추출하였다. 결과에서 [5]는 연관어에 복합어가 등장하지 않으며, 1차 연관어만을 결과로 제시하고 있다. 본 시스템은 "트로피 노긍정형님", "화성 영스트리트"와 같이 질의어와 관련한 다중 차수 연관어를 제공함으로써 추출된 연관어에 정보를 더 구체화 할 수 있었다.

Related word by [5]	Related word extracted by our system	
	1st related word	2nd related word
작곡가	안무서워	
직찍	직찍	쭈아삼촌, 한장
뮤비	트로피 노긍정형님	찬열 백현, 카이
앨범	공개방송	화성 영스트리트
시계	반디앤루니스	
나인	사랑	
블락비	1위	트로피, 제보, 인기가요
티아라	노래	
샤이니	분홍신	3집
노래	버스커	

Fig. 12. Comparison of results of our system and [5]

4. 결론 및 향후과제

본 논문에서는 질의어에 대한 트윗 문서에서의 연관 어휘 추출 방식을 제안하였다. 연관 어휘는 질의어와 가까운 거리에서 자주 발생한다는 점에 착안하여, 인접도의 합으로 단어 간 연관도를 얻었다. 인접도 행렬을 이용하는 방법으로 빠른 시간 안에 결과를 얻을 수 있으며, 연관 관계 네트워크의 간선을 이용하는 손쉬운 방법으로 질의어에 따라 달라지는 복합어를 비교적 정확하게 추출할 수 있었다. 실험에서는 임계치와 문맥의 크기에 따른 영향성을 분석하였으며, 비교적 간단한 방법을 사용하였음에도 실용성 있는 결과를 얻을 수 있었다.

이번 연구에서는 예산 및 시간의 한계로 인해 실시간으로 많은 정보를 수집, 분석하는 빅데이터 처리 시스템을 갖추지 못하였으며, 비용적인 한계로 재현율 분석 등을 수행하지 못했다. 향후 연구로는 연구 결과와 응용 분야나 질의어의 특성에 따른 성능 분석을 통해 연관 어휘 추출의 정확성을 높이고, 빅 데이터에 대한 수집 및 처리 시스템을 갖추어 더욱 효율적이고 방대한 정보를 추출을 예정하고 있다.

참 고 문 헌

[1] Pum-Mo Ryu, Hyeon Jin Kim, HyunKi Kim, Sang Kyu Park, “Social Media Issue Detection & Monitoring based on Deep Language Analysis Techniques”, *Korea Information Science Society Review*, Vol. 30, No. 6, 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] Heung-Seon Oh, Yoonjung Choi, Wookhyun Shin, Yoonjae Jeong, Sung-Hyon Myaeng, “Trend Properties and a Ranking Method for Automatic Trend Analysis”, *Journal of KIISE: Software and Applications*, Vol. 36, No. 3, pp. 236-243, 2009.

[5] Daumsoft Ltd., “http://insight.some.co.kr/searchKeyword Map.html”

[6] Han-joon Kim, Jaeyoung Chang, “Discovering News Keyword Associations Using Association Rule Mining”, *Journal of Institute of Webcasting, Internet and Telecommunication*, Vol. 11, No. 6, 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] Bo Seok Jung, Yung-Keun Kwon, Seung Jin Kwak, “A Knowledge Map Based on a Keyword-Relation Network by Using a Research Paper Database in the Computer Engineering Field”, *KIPS Transaction:PartD*, Vol. 18, No. 6, pp. 501-508, 2011.

[9] Kwang-Mo Ahn, Young-Hoon Seo, Jeong Heo, Chung-Hee Lee, Myung-Gil Jang, “Relevant Keyword Collection using Click-log”, *KIPS Transactions:PartB*, Vol. 19, No. 2, pp 149-154, 2012.

[10] Kil-Hong Joo, Joo-Il Lee, and Won-Suk Lee, “An Associated Keywords Extraction and a Spread Clustering Methods for an Efficient Document Searching”, *Journal of Korean Institute of Information Technology*, Vol. 9, No. 6, pp. 155-166, 2011.

[11] Seok-pal Jung, Seong-Hyeon Lim, Jin-Hyeong Jeon, Byeong Man Kim and Hyun Ah Lee, “Web Search Result Clustering using Snippets”, *Journal of KIISE: Database*, Vol. 39, No. 5, pp. 321-331, 2012.

[12] Lucene Core 4.0 and SolrTM 4.0 Available, “http://lucene.apache.org”, 12 October 2012.

[13] Twitter search system, “https://dev.twitter.com/”

김 제 상

e-mail : oiu124@naver.com

2007년~현 재 금오공과대학교 컴퓨터공학부 재학중

관심분야 : 자연언어처리, 빅데이터, 정보검색

조 효 근

e-mail : whitesky0109@naver.com

2007년~현 재 금오공과대학교 컴퓨터공학부 재학중

관심분야 : 임베디드시스템, 빅데이터, 정보검색

김 동 성

e-mail : azuregale@hotmail.com

2007년~현 재 금오공과대학교 컴퓨터공학부 재학중

관심분야 : 정보검색, 유비쿼터스

김 병 만

e-mail : bmkim@kumoh.ac.kr

1987년 서울대학교 컴퓨터공학과(학사)

1989년 KAIST 전산학과(공학석사)

1992년 KAIST 전산학과(공학박사)

1992년~현 재 금오공과대학교 컴퓨터소프트웨어공학과 교수

관심분야 : 인공지능, 정보검색, 정보보안

이 현 아

e-mail : halee@kumoh.ac.kr

1996년 연세대학교 컴퓨터과학과(학사)

1998년 KAIST 전산학과(공학석사)

2004년 KAIST 전산학과(공학박사)

2004년~현 재 금오공과대학교 컴퓨터소프트웨어공학과 부교수

관심분야 : 자연언어처리, 정보검색, 지식공학