빅데이터 기반 의사결정 지원을 위한 이슈이벤트 검색시스템: 소셜위즈덤

(An Issue Event Search System based on Big Data for Decision Supporting: SocialWisdom)

허 정 [†] 류 법 모 [††] 최 윤 재 [†††] 김 현 기 [††††] 옥 철 영 [†††††]
(Jeong Heo) (Pum-Mo Ryu) (Yoon Jae Choi) (Hyun Ki Kim) (Cheol Young Ock)

요 약 본 논문은 빅데이터 기반 의사결정 지원 시스템인 '소셜위즈덤'에 포함된 이벤트 추출 및 이슈 이벤트 검색에 대해서 소개한다. 의사결정 지원 시스템은 경제적, 사회적 중요사항을 결정할 수 있도록 관련 정보와 인사이트(insight)를 제공하는 정보시스템을 이른다. 기존 시스템은 단지 특정 키워드 빈도나 공기하는 키워드들의 관계만을 제공하였다. 그러나 소셜위즈덤은 이벤트로 정의되는 주체(subject), 이벤트 속성(event property), 객체(object)의 트리플 집합인 템플릿을 추출하여 이를 기반으로 이벤트 정보를 제공한다. 본 논문에서는 이벤트 추출 성능 개선과 이슈 이벤트 순위화 모델에 대해서 소개한다. 이벤트 추출 성능 개선을 위해 온톨로지와 이벤트 통합 기술에 기반한 필터링 방법에 대해서 소개하고, 변동계수를 이용한 이슈 이벤트 순위화 모델을 소개한다. 또한, 변동계수가 이슈 이벤트 순위화 성능에 미치는 영향을 분석하였다.

키워드: 이슈 이벤트, 의사결정지원 시스템, 빅데이터, 소셜위즈덤

Abstract In this paper, we introduce the issue event search system in 'SocialWisdom', a decision support system based on big data. A decision support system is a computer-based information system that gives information and insight to help make decisions on financial and social issues. Existing systems have given simply keyword frequency or keyword co-occurrence information. But 'SocialWisdom' gives information by extracting event templates from big data. An event template is a triple comprised of subject, property and object. We propose methods to improve issue event extraction and ranking based on knowledge engineering and statistical approaches. For event extraction, we introduce a filtering method based on ontologies and an event-merging technique. For event ranking, we introduce a ranking model based on the coefficient of variation. We have also conducted evaluation of the impact of the coefficient of variation on the performance of event ranking.

Keywords: issue-event, decision support system, big data, socialwisdom

· 이 논문은 2012년도 제24회 한글 및 한국어 정보처리 학술대회에서 '소셜미디어 기반 의사결정 지원을 위한 이벤트 템플릿 추출'의 제목으로 발표된 논문을 확장한 것임

† 비 회 원 : 한국전자통신연구원 지식마이닝연구팀 선임연구원
jeonghur@etri.re.kr
(Corresponding author임)
†† 정 회 원 : 한국전자통신연구원 지식마이닝연구팀 선임연구원
pmryu@etri.re.kr
††† 비 회 원 : 한국전자통신연구원 지식마이닝연구팀 연구원
mp2893@etri.re.kr
†††† 비 회 원 : 한국전자통신연구원 지식마이닝연구팀 책임연구원
hkk@etri.re.kr
††††† 종신회원 : 울산대학교 컴퓨터정보통신공학 교수
okcy@ulsan.ac.kr

논문접수 : 2013년 3월 4일
심사완료 : 2013년 5월 7일

정보과학회논문지: 소프트웨어 및 응용 제40권 제7호(2013.7)

1. 서 론

소셜 미디어와 모바일 기기의 활성화로 다양한 형태의 정보가 폭발적으로 축적되고 있다. 로그, 텍스트 콘텐츠, 멀티미디어 콘텐츠 등 다양한 형태의 대용량 데이터를 처리하고 관리하는 것의 중요성이 부각되면서 '빅데이터(big data)' 분석에 대한 요구가 급증하고 있으며 많은 연구가 진행되고 있다. 빅데이터 분석은 도구, 플랫폼, 분석방법 등 포괄적인 기술적 개념으로 사용되고 있다. 빅데이터 분석을 통해 전문가의 의사결정 지원을 위한 인사이트(insight)를 제공하는 기술이 차세대 구글(next google)을 실현할 수 있는 기술로 언급되고 있다[1].

빅데이터 분석의 중요한 소스인 소셜 미디어에는 정치, 경제, 사회문화적인 이슈에 대한 다양한 의견 및 사회적 행동 패턴이 잠재되어 있다. 이와 같이 소셜 미디어에 잠재되어 있는 중요 정보를 파악하여 분석할 수 있다면, 기업 및 공공단체에서 선행적인 의사결정을 통해 많은 기회를 창출할 수 있다. 이와 같은 특성으로 인해 많은 의사결정 지원 시스템은 소셜 미디어를 핵심 분석 대상으로 인식하고 있다. 소셜 미디어 상의 여론 추이 및 특정 브랜드, 인물, 정책 등의 호불호를 파악하거나, 리스크를 감지하기 위해서 다양한 유형의 소셜 미디어 분석 기술이 연구되고 있다. 소셜 미디어 사용자의 집단지성 및 여론을 분석, 모니터링, 예측할 수 있는 의사결정 지원 시스템은 소셜 비즈니스, 온라인 여론 분석 시스템, 온라인 광고/홍보/마케팅, KMS 등의 다양한 지능형 소프트웨어 분야에 적용될 수 있는 기술로 각광을 받고 있다.

의사결정 지원을 위한 소셜 미디어 분석 기술은 크게 두 가지 유형으로 구분된다. 첫째, 소셜 미디어의 네트워크 분석이다[2]. 트윗의 전파 양상 및 경로를 분석하여 영향력자가 누구이며, 트윗이 얼마나 이슈성이 있는지를 분석하는 기술로서 여론의 추이와 확산에 기반한 의사결정에서 중요한 역할을 하고 있다. 그러나, 주로 트위터와 같이 개인 의견이 네크워크를 통해 원활히 전파될 수 있는 플랫폼만을 대상으로 할 수 있다. 이는 트위터를 사용하는 사람들 만의 닫힌 공간에 대한 여론을 반영하는 단점이 있다. 둘째, 소셜 미디어 컨텐츠에 대한 내용분석이다[3]. 즉, 소셜 미디어 상에서 주요한 키워드에 대한 시간대 별 언급 추이와 감성 변화를 분석하는 기술이다. 마케팅과 관련하여 통제어휘(controlled vocabulary)를 선정하고, 해당 어휘들의 변화 추이를 비교 분석하여 의사결정을 위한 인사이트를 제공한다. 그러나, 키워드들 간의 연관정보를 분석하지 않고, 독립적인 키워드의 빈도 변화 추이와 감성 변화만을 제시하여 키워드들 간의 연관성을 파악하기 어려운 단점이 있다.

본 논문은 사회전반적인 여론 추이와 변화에 대한 인사이트를 제공하기 위해서 소셜 미디어 뿐만 아니라, 뉴스와 블로그 컨텐츠도 분석을 수행한다. 그리고, 기존 소셜 미디어 분석의 한계인 키워드들 간의 연관성 정보를 분석하기 위해 관계추출 기술을 이용하여 개체들간의 연관성을 분석한다. 개체들간의 주요한 연관정보는 이벤트 템플릿으로 정의되고, 관계추출을 통해 개체들간의 연관정보를 추출하여 이벤트 템플릿을 구성한다. 그리고, 생성된 이벤트 템플릿을 기반으로 개별 이벤트들의 변화 추이와 이슈성 정보를 제공함으로써 보다 연관정보가 많은 인사이트를 제공할 수 있다.

본 논문의 구성은 2장에서 관련 연구를 소개하고, 3장에서 이슈이벤트 검색시스템의 세부적인 기술을 설명한다. 4장에서는 소개된 기술들에 대한 평가 방법과 결과에 대해서 분석하고, 5장에서는 이슈이벤트 검색 시스템인 소셜위즈덤에 대해서 소개한다. 마지막 6장은 본 논문에 대한 결론과 향후 연구방향에 대해서 기술한다.

2. 관련 연구

빅데이터 분석의 대표적인 사례는 IBM에서 개발한 DeepQA인, Watson이 있다[4]. Watson은 2011년 2월 TV 퀴즈 프로그램인, 'Jeopardy! 퀴즈쇼'에서 인간 챔피언들과의 경쟁에서 이기면서 크게 주목을 받았다. Watson은 고성능의 하드웨어(90대의 IBM Power 750 서버)를 기반으로 대용량의 콘텐츠(약 200억 페이지 분량)를 대상으로 고정밀 자연어 분석을 통해 구조화된 지식을 구축하여, 질문에 대한 정답을 제시하는 질의응답 시스템(question answering system)이다. Watson은 엄청난 컴퓨팅 파워가 필요한 기술로서, IBM의 슈퍼컴퓨터를 이용하고 있다. 또한, 'Jeopardy! 퀴즈쇼'에 특화되어 있어서, 의사결정 지원을 위한 질의응답 및 검색 기술과는 차이가 있다. 그러나, 대용량 콘텐츠로부터 지식을 구축하기 위해 수행하는 다양한 언어처리기술은 의사결정 지원시스템과 다르지 않다. 단지, 언어처리된 결과에 대한 활용의 차이일 뿐이다.

[5]는 지진과 태풍 같은 이벤트가 트위터 상에서 시/공간적으로 전파되는 양상을 파악하는 확률 모델에 대해서 연구하였다. 트위터 사용자를 센서로 간주하고 이를 기반으로 트위터 상에 포스팅되는 지진관련 자질들을 분석하여 지진 발생여부를 경고하는 시스템을 구축하였다. 평가를 통해 일본정부보다도 지진에 대한 경보 전파 속도가 빠르다는 것을 결론으로 도출하였다. [5]는 지진 관련 트윗의 전파경로를 네트워크 분석을 통해 파악하고 이를 기반으로 경고를 하는 성공적인 사례이다. 그러나, 지진이나 태풍과 같은 기정의된 자연재해 이벤트로 특화되어 있어, 사회 전반적인 여론 동향이나 다양한 도

메인에 대한 의사결정 지원을 위한 인사이트 제공에는 한계가 있다.

[6]은 선거철에 소셜 미디어 상의 유권자 의견을 분석하여 선거의 결과를 예측하는 기술에 대해서 소개하고 있다. [6]에서는 소셜 미디어 상의 특정 개체의 언급 빈도와 감성정보를 기반으로 한 선거결과 예측과 선거 유권자의 투표 결과의 차이에 대해서 분석하였고 소셜 미디어를 이용한 보다 현실성 있는 선거예측 시스템을 만들기 위한 방법론을 소개하고 있다. 선거예측을 위한 소셜 미디어 분석 기술은 많은 국가에서 활용되고 있다. 그러나, 대부분의 시스템은 단지 특정 키워드(대부분 후보자 이름)에 대한 볼륨과 감성정보에 기반하고 있다. 즉, 해당 후보자에 대한 긍정적인 이벤트가 무엇인지, 부정적인 이벤트가 무엇인지 파악할 수 없는 단점이 있다. 따라서, 단순 키워드에 대한 볼륨과 감성정보가 아닌, 키워드 쌍의 연관성을 구조화한 이벤트의 볼륨과 감성정보를 분석하여야 보다 정확한 선거예측이 가능할 것이다.

다른 사례로는 Recorded Future가 있다[7]. Recorded Future는 웹 데이터를 기반으로 웹 인텔리전스 및 예측분석을 전문으로 하는 회사이다. 비구조화 텍스트 콘텐츠를 대상으로 기 정의한 이벤트에 대해서 정보를 추출, 분석하고, 시각화하여 제시하고 있다. 이벤트 별 긍/부정 정보와 사용자의 관심도, 흥미도, 중요도의 관점에서 모멘텀(momentum) 정보를 시간 축을 기준으로 제시하고 있다. 그러나, 분석대상 컨텐츠가 뉴스와 블로그로 국한되어 있어 즉각적인 여론 추이를 파악할 수 없는 단점이 있다. 또한, 정의된 이벤트도 단순히 What, Who, Where, When 에 기반한 개체만을 대상으로 하고 있다. 추출되는 이벤트가 시/공간 정보에 기반한 것으로써 다양한 개체들에 대한 연관성 분석에는 한계가 있다.

응용사례로는 다음소프트의 소셜 매트릭스를 비롯한 다양한 시스템들[1]이 있다. 소셜 매트릭스는 자연어처리 기술과 텍스트마이닝 기술을 바탕으로 블로그, 트위터 문서를 분석하여 사용자의 요구에 따라 해당 키워드의 빈도 추이 및 감성정보들을 모니터링할 수 있다. 소셜 매트릭스는 단지 개체 단위의 정보만을 제공하고 있으며, 개체들 간의 연관성에 기반한 이벤트 단위의 정보 분석은 지원하지 않는다. 이로 인해 개체들 간의 의미적 연관성 파악이 어려워 의사결정 지원에 한계가 있다.

앞서 언급된 논문들은 대부분 단일 개체에 대한 볼륨 정보와 해당 개체가 포함된 문장의 감성정보를 기반으로 하여 소셜 미디어를 분석하고 있다. 이로 인해 개체들 간의 의미적 연관성 파악이 어렵다. 또한, 특정 개체의 호불호에 대한 원인 파악도 어렵다.

이벤트의 정의 및 표현은 1987년부터 1997년까지 진행된 MUC(Message Understanding Conference)을 기반으로 한다[8,9,10]. MUC에서는 정보추출(information extraction)의 주요 기술로 개체명인식(named entity recognition), 상호참조해결(co-reference resolution), 관계추출(relationship extraction), 템플릿 생성(template generation), 시나리오 생성(scenario generation) 등을 언급하고 있다. 개체명을 인식하고 개체들간의 연관관계를 분석하기 위해 관계추출을 수행한다. 개체명들 간의 관계성은 기정의된 템플릿 생성에 활용된다. 생성된 템플릿들간의 연관성을 기반으로 시나리오를 생성하는 과정을 거치게 된다. 본 논문에서는 MUC에서 정의하는 정보추출의 기술을 이용하여 이벤트 템플릿을 정의한다. 그러나, 템플릿 추출의 범위를 문서가 아닌 문장으로 제한한다. 또한, 템플릿들간의 관계성을 이용한 시나리오는 정의하지 않는다.

본 논문에서는 MUC에 기반하여 개체들 간의 의미적 관계를 정의한 이벤트들을 템플릿 구조로 정의하며 소셜 미디어에서 이벤트 템플릿을 추출하고 검색하는 기술에 대해서 소개한다. 개체들간 의미적 관계를 기정의한 템플릿을 대상으로 하여 사용자의 요구에 부합하는 인사이트 정보를 제공함으로서 기존 시스템이나 논문의 키워드 기반 정보 제공의 한계를 극복하고 한다. 또한, 개체들을 시간정보에 기반하여 검색할 수 있도록 함으로서 명시적으로 정의하지 않은 템플릿들간의 연관성 정보인 시나리오 정보를 사용자가 파악할 수도 있다.

3. 이슈이벤트 검색시스템

그림 1은 이슈이벤트 검색시스템의 구성도로써, 크게 빅데이터 수집, 이벤트 추출과 이벤트 검색으로 나뉜다. 본 논문에서는 이벤트 추출과 검색에 대해서만 소개한다.

본 논문에서는 이벤트를 '주체(subject)', '이벤트속성(event-property)', '객체(object)'로 구성된 트리플(triple)의 집합인 템플릿(template)으로 정의한다. 이벤트속성은 이벤트 이름과 속성 이름으로 구성된다.

이벤트 템플릿 선정은 기업과 공공분야를 대상으로 의미적 가치[2]와 출현 빈도를 고려하여 선정하였으며, 31개의 이벤트 템플릿은 온톨로지로 구성되었다. 표 1은 이벤트 템플릿의 목록으로서, 속성에서 **볼드체**로 표시된 속성명은 주체에 해당되는 속성이다.

1) 솔트룩스 - 트루스토리(http://politician.truestory.co.kr/)
코난테크놀로지 - 펄스 K(http://www.pulsek.com/)
그루터 - 씨날(http://www.seenal.com/)

2) 관계추출의 기술적 한계(엔티티들간의 관계만 대상으로 함)에 기반하여 일반 사용자들이 알고 싶어하는 정보를 마케팅 전략 분석 전문가의 도움으로 선정

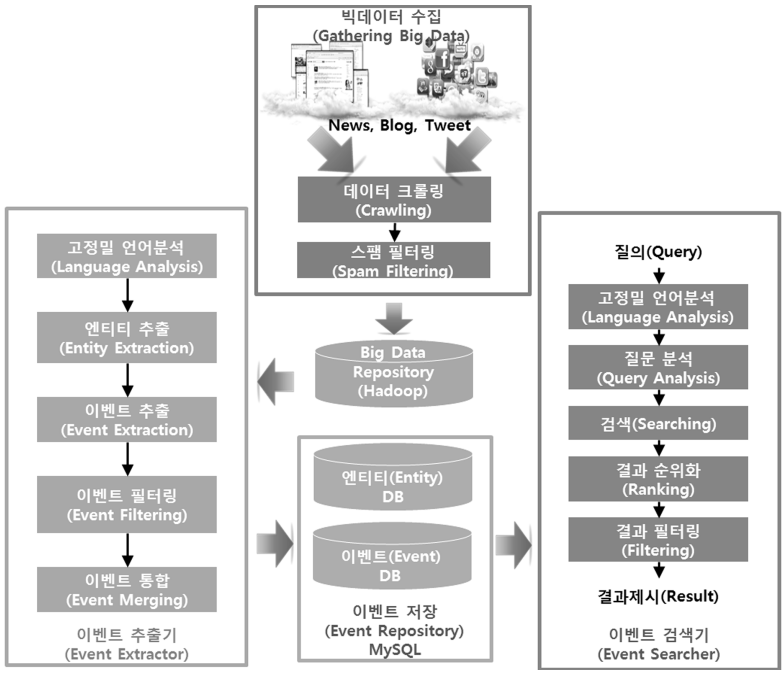


그림 1 이슈이벤트 검색시스템 구성도

Fig. 1 Architecture of the issue-event search system

3.1 이벤트 추출 및 저장

템플릿 기반 이벤트 추출은 고정밀 언어분석을 기반으로 수행된다. 언어분석은 형태소분석, 개체명인식, 관계추출이 수행된다. 개체명인식은 약 180개의 다양한 개체명을 분류하고 태깅하는 ETRI 개체명 인식기를 이용하였다[11].

관계추출은 인식된 개체명들 간의 관계를 분석하는 기술로서, 복수개의 관계추출(n-ary relation) 결과를 제공한다. 추출된 관계에서 주체(subject) 또는 객체(object)로 사용된 개체들을 개체연결(entity linking) 과정을 통하여 도메인 온톨로지의 클래스와 연결한다. 도메인 온톨로지는 표 1에서 나열한 이벤트의 속성으로 올 수 있는 '사람', '제품', '회사', '조직', '장소', '정책', '사건' 등의 클래스를 정의하고 있다. 따라서 ETRI 개체명 인식기로 인식된 개체들을 온톨로지의 클래스로 맵핑하는 과정이 필요하다. 예를 들어, 이벤트 ProductRelease의 주체 속성인 product는 ETRI 개체명 인식기의 개체분류 중에서 TMI_HARDWARE(하드웨어), TMI_MODEL(제품 모델명), ARTIFACT(인공물) 등으로 맵핑될 수 있다.

또한, 관계 분석된 개체명들 간의 관계는 이벤트 속성으로 매핑이 수행된다. 매핑관계는 N:1로 이벤트 매핑 모호성(ambiguity)이 발생한다. 표 1의 Award 이벤트의 경우, 주체(subject)로 인식될 수 있는 속성이 person과 organization이 있다. 관계 추출에서 Award.person과 Award.organization이 모두 인식되었을 경우, 구체적인 속성인 person을 주체로 인식하는 것이다. 예를 들면, "한국대 홍길동 교수가 대통령상을 수상했다."라는 문장에서 'Award.person - 홍길동', 'Award.organization - 한국대'와 'Award.prize - 대통령상'이 추출되었을 경우, Award이벤트의 주체로 'Award.person - 홍길동'이 선택된다.

추출된 이벤트의 속성값은 의미적으로 동일한 개체(entity)로 인식될 수 있으나, 표현형태가 다른 많은 이형태로 구성된다. 예를 들면, 'IPHONE 4S'와 '아이폰 4S'는 동일한 개체이지만 그 표현 형태는 다르다. 이를 처리하기 위해서 정규화를 수행한다. 정규화 모듈은 개체명 정규화, 날짜 정규화, 지역명 정규화, 가격 정규화로 구성된다. 개체명 정규화는 위키피디아 리다이렉션(redirection) 정보를 활용하여 반자동으로 구축한 사전을 이용한다. 위키피디아는 리다이렉션 페이지를 이용하여 의미적으로 동일하거나 관련이 있는 다른 페이지로 이동할 수 있는 방법을 제공하고 있다. 위의 예에서 'IPHONE 4S' 페이지는 동일한 의미를 가지는 '아이폰 4S' 페이지로 리다이렉션된다. 그러나 '메소포타미아문명' 페이지는 '메소포타미아' 페이지로 리다이렉션되지만 동일한 의미가 아니고 의미적 연관성을 표현하고 있다.

표 1 이벤트 템플릿 목록
Table 1 List of Event templates

Event Name	Description	Properties
Acquisition	인수합병	**acquirer**, beingacquired, date, price
CompanyCompetitor	경쟁기업	**company1**, company2
CompanyEarning Announcement	기업실적	**company**, date, earning
CompanyExpansion	기업확장	**company**, date, expansiontype
CompanyInvestment	기업투자	**company**, date, target, price
CompanyLegalIssue	법적이슈(기업)	**company_plaintff, person_plaintiff**, date, company_sued
CompanyMeeting	사업미팅	**company**, date, meetingtype
CompanyTechnology	보유기술	**company**, technology
CreditRating	신용평가	**rated_org**, date, rating_org
EmploymentChange	직책변동	**person**, date, positiontitle, organization
ProductRecall	제품리콜	**product**, date, company
ProductRelease	제품출시	**product**, company, date, price
AdvertisingStart	광고	**product**, date, company
Announce	발표	**organization, person**, date, location
Award	수상	**person, organization**, prize, date, product
EventOpen	행사	**event**, date, organization, person, location
Investigate	조사	**person**, event, date, organization
LegalAct	법적규제	**organization**, date, legalaction
MarketShare	시장점유율	**company**, date, ratequantity, product
PersonDie	부고	**person**, date, reason
PersonTravel	여행출장	**Person**, date, destination
PolicyEnforce	정책시행	**policy**, date, organization
PolicyOpposition	정책반대	**policy**, date, organization, person
PolicySupport	정책지지	
PriceDecline	가격하락	**product, fee**, date, organization, pricechange
PriceRise	가격상승	
Recommend	추천	**product**, person, organization
StockDecline	주가하락	**company**, date, stockrisecontent
StockRise	주가상승	
StockList	주식상장	**company**, date
Vote	투표/선거	**vote**, date

또 '양양' 페이지가 중국의 '샹판시'로 리다이렉션되는데, 이 경우도 한국의 '양양시'와 구분하지 못하기 때문에 정규화 사전으로 사용하기 어렵다. 따라서 전체 한국어 위키피디아 리다이렉션 리스트를 대상으로 의미 모호성이 없고 동일한 개념을 표현하는 리스트를 수작업으로 추출하여 개체명 이형태 사전을 구축한다. 날짜와 가격 정규화는 정규표현식(regular expression)을 기반한 규칙을 이용하고, 지역명 정규화는 주소사전(gazetteer) 기반의 규칙을 활용한다. 날짜 정규화는 TAC(Text Analysis Conference) – KBP(Knowledge Base Population)의 TSF(Temporal Slot Filling)에서 많은 연구가 진행되고 있다[12].

추출된 이벤트 템플릿은 재현율보다는 정밀도가 중요하다. 따라서, 오류에 해당하는 템플릿에 대해서 필터링을 수행하는 것은 상당히 중요하다. 이를 위해서, 이벤트 템플릿의 속성들과 온톨로지의 관계 정보, 이벤트들 간의 관계 정보를 이용하여 오류로 추정되는 이벤트 템플릿을 필터링한다. 또한 이벤트 속성값을 일(하루) 단위로 빈도정보에 기반하여 통합한다. 상대적으로 저빈도에 해당하는 속성값은 필터링된다. 이벤트 필터링과 일 단위 통합의 상세 설명은 3.1.1과 3.1.2에서 기술한다.

3.1.1 이벤트 제약 및 필터링

이벤트 속성 매핑이 완료되면, 템플릿은 주체, 이벤트 속성, 객체의 트리플 집합으로 구성된다. 구성된 이벤트 템플릿은 일관성(consistency) 유지를 위하여 제약규칙(constraint rule)에 따른 필터링(filtering)을 수행한다. 제약규칙은 크게 네 가지 유형으로 나뉜다.

가) 온톨로지에 기반한 제약

온톨로지에 기반한 제약은 개체와 온톨로지 연결정보에 따른 제약으로써, 개체 카테고리(entity category[3])

와 개체 유형(entity type[4])에 따른 제약으로 구분된다.

템플릿의 속성별로 유효한 개체 클래스가 있다. 예를 들면, ProductRecall.company의 속성값으로 '회사' 클래스만이 유효하다. '회사' 클래스가 아닌 다른 클래스의 경우 필터링된다. ProductRecall.company의 속성값으로 '서울시'가 추출되었을 경우, 인스턴스인 '서울시'는 온톨로지의 '지역' 클래스로 연결될 것이다. 그러나, ProductRecall.company의 속성값으로 '회사' 클래스만 유효하기 때문에 추출된 템플릿은 필터링된다.

또한, 템플릿의 속성별로 유효한 개체 유형이 있다. 예를 들면, ProductRelease.product의 속성값은 인스턴스 유형만이 유효하다. 즉, 개체 유형이 클래스인 경우는 필터링 된다. ProductRelease.product의 속성값으로 '아이폰 4s'와 '스마트폰'이 추출되었다고 할 경우, '아이폰 4s'는 온톨로지 내 '스마트폰' 클래스의 인스턴스로 연결되기 때문에 개체유형이 Instance로 인식된다. 반면, '스마트폰'은 '스마트폰' 클래스 그 자체를 언급하는 것이기 때문에 개체유형이 Class로 인식된다. 따라서 ProductRelease.product의 속성값으로 유효한 개체유형이 Instance이므로, 속성값으로 '스마트폰'이 추출된 템플릿은 필터링된다.

나) 배타적 관계(exclusive relation)에 기반한 제약

배타적 관계에 해당하는 이벤트 템플릿은 다음과 같다.

StockRise.company – StockDecline.company
PriceRise.product – PriceDecline.product
PriceRise.fee – PriceDecline.fee
PolicySupport.organization – PolicyOpposition.organization
PolicySupport.person – PolicyOpposition.person

이벤트 템플릿은 일 단위로 통합되어 빈도정보를 저장한다. 따라서 배타적 관계에 해당하는 템플릿은 빈도정보에 의해서 하나의 이벤트 템플릿을 선택하고, 선택되지 않은 템플릿은 필터링 된다. 예를 들어, 'PriceRise.product – 아이폰 4s'의 빈도가 2이고, 'PriceDecline.product – 아이폰 4s'가 5번이라면, 'PriceRise.product – 아이폰 4s'가 오류일 가능성이 많으므로 필터링하는 것이다.

다) 필수 속성(required property)에 기반한 제약

이벤트 템플릿 별로 주체를 제외한 객체 속성들 중, 반드시 값이 있어야 하는 속성을 정의하고 있다. 해당 속성에 대한 값이 추출되지 않으면 필터링이 된다. 예를 들면, 문장에서 'EmploymentChange.person – 박근혜'와 'EmploymentChange.date – 2013년 2월 25일'가 추출되었다고 가정하자. 이 경우에 EmploymentChange 이벤트의 필수 속성인 positiontitle은 반드시 값이 채워져야 하지만, 해당 속성값이 채워지지 않았다. 따라서, 해당 템플릿은 필터링 된다.

라) 사전(dictionary)에 기반한 제약

사전(dictionary)에 기반한 제약은 템플릿 추출에서 고빈도로 발생하는 오류를 사전에 규칙으로 기술하고, 이를 기반으로 템플릿을 제약한다. 예를 들면, 트위터에서 '문재인은 대통령으로 뽑읍시다."라는 트윗이 많이 있을 경우, 현재 관계추출 기술에서는 '문재인 – EmploymentChange.positiontitle – 대통령'으로 추출된다. 이처럼 기술적 한계로 사실(fact)이 아닌 정보가 빈번하게 추출될 경우에 사전에 해당 관계를 자동 필터링할 수 있도록 등록하는 것이다. 트윗의 경우, 위의 예와 같이 사실과 관계없이 청유형이나 명령형 등의 문장이 많다. 그러나, 이처럼 사실이 아닌 다양한 문장유형을 자동으로 분류하기에는 기술적 한계가 있다.

3.1.2 일 단위 이벤트 통합

이벤트의 속성들은 일(하루) 단위로 통합된다. 하루 동안 수집되어 추출된 이벤트들은 동일한 사건을 언급한 것으로 가정하고 통합한다. 이벤트 통합 결과를 기반으로 추출된 이벤트의 오류를 필터링할 수 있다. 데이터 잉여성(data redundancy)에 기반하여 "고빈도로 추출되는 이벤트는 올바르게 추출된 이벤트일 가능성이 높다"는 전제를 바탕으로 한다. 그림 2는 이벤트 통합와 필터링의 관계를 예제로 표현한 그림이다. ProductRelease 이벤트에서 '갤럭시 S3'의 ProductRelease.company속성값으로 '삼성전자', 'LG전자', '애플'이 추출되었을 때, 이벤트 통합에서 고빈도로 추출된 '삼성전자'를 ProductRelease.company 속성값으로 통합한다. 즉, 필터링에서 '갤럭시 S3'에 대한 ProductRelease 이벤트의 company 속성으로 'LG전자'와 '애플'은 필터링되는 것이다.

이벤트의 속성은 통합이 가능한 속성(constant property)과 통합이 불필요한 속성(variable property)으로 구분된다. constant 속성은 절대 변하지 않는 하나의 사실(fact)만을 속성값으로 가지는 속성이다. 예를 들면, 인물의 사망일을 의미하는 PersonDie.date 속성 값은 항상 하나일 수 밖에 없다. 반면, variable 속성은 여러 값을 가질 수 있고, 시간에 따라 값이 변화한다. 예를 들면, 제품의 시장점유율을 의미하는 MarketShare.rate-quantity 속성값은 시간에 따라 변화된다. 통합에 활용되는 자질은 이벤트 속성값의 빈도, 이벤트 속성값이 추출된 문장들의 중복정도, 이벤트 속성값이 추출된 문장의 소스정보가 활용된다.

문장들의 중복정도는 Q-gram 유사도 알고리즘를 응

3) Entity Category : entity가 연결(linking)된 온톨로지 클래스(class) 이름

4) Entity Type : entity가 온톨로지의 class인지, instance인지 여부

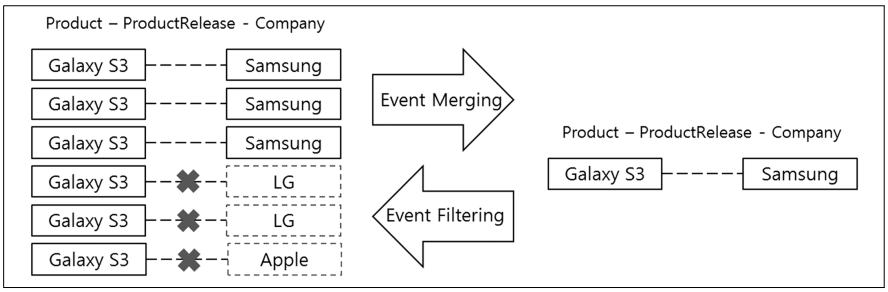


그림 2 이벤트 통합에 의한 이벤트 필터링 예제
Fig. 2 An example of merging-based event filtering

용하였다[13,14].

$$\text{Sentence Set}(SC_i) = \{S_1, S_2, S_3, \dots, S_l\} \quad (1)$$

$$\text{Sentence}(S_j) = \{T_1, T_2, T_3, \dots, T_m\},\ (1 \le j \le l) \quad (2)$$

$$\text{Token Set}(TC_i) = \{S_1 \cup S_2 \cup S_3 \cup \dots \cup S_l\} = \bigcup_{j=1}^{l} S_j = Unique\{T_1, T_2, T_3, \dots, T_n\} \quad (3)$$

$$\text{Sentences Similarity In Set}_i(SS_i) = \frac{\sum_{k=1}^{n} \frac{ISF_i(T_k)}{l}}{n} \quad (4)$$

$$\text{Duplication Degree}_i(DD_i) = \frac{((l \times SS_i) - 1)}{(l-1)} \times 0.9 \quad (5)$$

i번째 문장집합 SC_i는 문장 $S_j, (1 \le j \le l)$의 집합으로 구성된다(식 (1)). 각 문장 S_j는 문장을 구성하는 문자열을 bi-gram으로 분할한 토큰 $T_k, (1 \le k \le m)$들의 집합으로 구성된다(식 (2)). 이를 기반으로 문장집합 SC_i는 토큰의 집합 TC_i로 변환할 수 있다. 즉, 문장집합 SC_i를 구성하는 문장 S_j들을 문장 별 토큰집합(식 (2))으로 변환하고 이를 교집합($\bigcup_{j=1}^{l} S_j$)하여 문장집합에 대한 토큰집합 TC_i ($Unique\{T_1, T_2, T_3, ..., T_n\}$)로 변환할 수 있다(식 (3)). 문장집합 SC_i의 유사도 SS_i는 토큰집합 TC_i를 구성하는 개별 토큰 T_k에 대한 ISF(Inverted Sentence Frequency)인 $ISF_i(T_k)$를 전체문장 수 l로 나눈 비율에 대한 전체 토큰의 평균으로 계산된다.[5](식 (4)). 집합의 중복도(DD_i)는 집합의 유사도(SS_i)를 0과 0.9사이로 정규화(식 (5))시킨 것으로써, 집합 내 문장이 모두 동일하면 0.9, 완전히 다르면 0의 값을 가진다. 집합을 구성하는 문장이 하나일 경우, 1의 값을 부여한다.

일 단위 이벤트 속성 통합은 이벤트명, 주체와 이벤트 속성을 키(key)값으로 하여 객체 속성값을 통합한다. 속성값 통합은 constant 속성만을 대상으로 한다.

$$arg\max_i(\log_{10}(F_i + 1) \times (2 - DD_i) \times ST_i) \quad (6)$$

F_i는 객체 속성값 i의 빈도이고, DD_i는 객체 속성값 i가 추출된 문장 집합의 중복도이다. ST_i는 개체 속성값 i가 추출된 소스 유형(뉴스, 블로그, 트위터)의 수로써 1~3의 값을 가진다. 즉, 다양한 소스에서 추출된 객체 값에 가중치를 부여하는 것이다. 예를 들면, '김대중-PersonDie.date'의 속성값으로 아래와 같이 두 개의 값이 추출되었다고 가정하자.

① '2009년 08월 18일'
- 추출빈도(F_1) : 20
- 속성값이 추출된 문장들의 중복정도(DD_1) : 0.5
- 속성값이 추출된 소스유형의 수(ST_1) : 3 (뉴스, 블로그, 트위터 모두에서 추출)
- $\log_{10}(F_1 + 1) \times (2 - DD_1) \times ST_1 = 5.95$

② '2009년 08월 23일'
- 추출빈도(F_2) : 15
- 속성값이 추출된 문장들의 중복정도(DD_2) : 0.6
- 속성값이 추출된 소스유형의 수(ST_2) : 2 (뉴스, 블로그, 트위터 모두에서 추출)
- $\log_{10}(F_2 + 1) \times (2 - DD_2) \times ST_2 = 3.37$

위의 두 속성값에 대한 $\log_{10}(F_i + 1) \times (2 - DD_i) \times ST_i$ 값은 5.95와 3.37로 '2009년 08월 18일'의 값이 더 크다. 따라서, '김대중-PersonDie.date'의 속성값은 '2009년 08월 18일'로 통합된다.

3.2 이슈 이벤트 검색 및 순위화 알고리즘

주요 개체(entity)[6]에 대한 이벤트 변화 추이를 알기 위해 사용자가 질의를 입력하면, 질의분석을 통해 개체와 이벤트를 인식하게 된다. 개체의 인식은 개체명 인식(named entity)을 기반으로 하며, 이벤트 인식은 이벤트 단서사전에 기반한다. 예를 들면, "삼성의 주가"라는 질의에서 개체명으로 태깅된 '삼성(OGG_BUSINESS(회사))'이 개체로 인식되고, '주가'라는 단서 어휘를 통해 StockDecline, StockRise, StockList 이벤트를 인식하게 된다. 사용자의 질의가 없을 경우는 모든 개체와 이벤트

5) $ISF_i(T_k)$와 l은 중복정도를 계산하고자 하는 문장집합 SC_i을 대상으로 함

6) 주체의 속성값과 객체의 속성값

를 대상으로 순위화된 결과를 제시한다.

질문분석에서 인식된 개체와 이벤트에 대한 검색을 수행한 후, 해당 이벤트의 순위화를 수행한다. 이벤트의 순위화는 이벤트의 단위기간(time span)내 이슈성(hotness)에 대한 순위(ranking)이다. 이벤트 이슈성 계산은 이벤트 중요성(importance), 이벤트 변동성(event variation), 개체의 이슈성(entity hotness), 문장의 중복성(duplication degree), 객체의 필수성(object necessariness) 자질을 이용한다.

이벤트의 중요성은 TDT(Topic Detection and Tracking)에서 중요어휘를 선정하기 위해 사용하는 TF*PDF 알고리즘을 이용한다[15-18]. TF*IDF는 정보검색을 위한 키워드의 중요성에 대한 가중치에 해당한다. 즉, 빈도가 높고, 적은 문서에서 언급되는 키워드는 정보검색에서 중요한 키워드로 평가된다. 해당 키워드로 적합한 문서를 효과적으로 찾아서 순위화할 수 있다는 것을 의미한다. 반면, TF*PDF는 빈도가 높으면서 많은 문서에 언급되는 키워드가 더 중요하다. 즉, 키워드의 이슈성은 키워드의 확산이 중요한 기준이기 때문에 많은 문서에 언급되어야 한다. 이로 인해, 본 논문에서 개체와 이벤트의 중요성 평가를 위해서 TF*PDF가 적합하다.

$$TFPDF_j = \sum_{s=1}^{s=|S|} |F_s(j)| exp\left(\frac{n_s(j)}{N_s}\right) \quad (7)$$

$$|F_s(j)| = \frac{F_s(j)}{\sqrt{\sum_{k=1}^{k=K} F_s^2(k)}} \quad (8)$$

$TFPDF_j$ 는 이벤트 j 의 중요성 가중치이다. $F_s(j)$는 채널(뉴스, 블로그, 트위터) s 내의 이벤트 j 의 빈도이다. $n_s(j)$ 는 이벤트 j 가 포함된 채널 s 내의 문서 수이다. N_s 는 채널 s 내의 전체 문서의 수이다. $|S|$ 는 채널의 수로써 본 논문에서는 3이다. K 는 채널 s 에 포함된 전체 이벤트의 수이다. 개체의 중요성은 이벤트의 중요성 가중치 계산과 동일하다.

이벤트 변동성은 단위기간 동안의 이벤트의 빈도 변화에 대한 변동계수(coefficient of variation)로 정의한다. "단위기간 동안 변화의 폭이 많은 이벤트가 변화의 폭인 적은 이벤트보다 이슈로서의 가치가 높다"는 직관에 기반한다. 그림 3은 이벤트의 변동성과 이슈성의 상관관계 평가를 위한 예제로써, 단위기간 동안 이벤트 A와 B의 빈도 평균은 동일하지만, 변동의 폭이 큰 이벤트 B가 이벤트 A보다 이슈로서의 가치가 높다는 것이다.

$$CV_t(j) = \frac{SD_t(j)}{\overline{x_t(j)}} \quad (9)$$

$CV_t(j)$ 는 단위기간 t 동안의 변수 j 의 변동계수이고, $SD_t(j)$ 는 j 의 표준편차(standard deviation), $\overline{x_t(j)}$ 는 j 의 산술평균(arithmetic mean)이다.

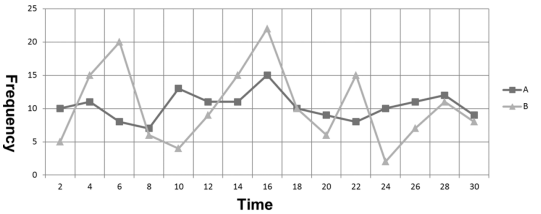


그림 3 이벤트 변동성(coefficient of variation)과 이슈성(hotness)의 연관관계

Fig. 3 Relation between event hotness and coefficient of variation

개체 j 의 이슈성 가중치 $EnH_t(j)$ 는 개체 j 의 중요성 $TFPDF_j$ 와 변동성 $CV_t(j)$ 의 곱으로 계산된다.

$$EnH_t(j) = TFPDF_j \times CV_t(j) \quad (10)$$

이벤트 j 의 이슈성 가중치 $EvH_t(j)$ 는 이벤트 j 의 중요성 $TFPDF_j$, 단위기간 t 동안의 변동성 $CV_t(j)$, 이벤트 j 가 추출된 문장들의 중복성 DD_j, 이벤트 j 의 주체인 j_{sub} 의 이슈성 가중치 $EnH_t(j_{sub})$, 이벤트 j 의 객체인 j_{obj} 의 이슈성 가중치 $EnH_t(j_{obj})$ 과 객체의 필수성 ON_j 의 곱으로 계산된다. 객체의 필수성(ON_j)은 온톨로지에 정의된 이벤트 객체의 속성유형(property type)[7] 에 따라 가중치를 부여한다.

$$EvH_t(j) = TFPDF_j \times CV_t(j) \times \sqrt{(2 - DD_j)} \times EnH_t(j_{sub}) \times EnH_t(j_{obj}) \times ON_j \quad (11)$$

4. 실험 및 평가

실험은 이벤트 추출 및 필터링의 성능 평가, 변동계수를 이용한 이슈성 순위화 성능 평가와 일 단위 이벤트 검색 성능에 대한 평가를 수행하였다.

4.1 이벤트 추출 및 필터링 성능 평가

평가를 위해서 뉴스, 블로그, 트위터로 구성된 평가데이터를 구축하였다. 표 2는 평가데이터를 구성하는 소스별 분포와 성능에 대해서 기술하고 있다. 평가데이터는 문서 생성 시간(document created time)이 없는 관계로 이벤트 제약 규칙 중, 나)는 적용하지 않고 평가하였다. 그림 4는 평가방법으로써, 필터링 전/후의 정밀도(rrecision)과 재현율(recall) 계산방법, 필터링의 정확도(accuracy) 계산방법을 벤다이어그램으로 설명하고 있다.

7) 객체의 속성 유형
필수속성(required object) : 반드시 필요한 속성, 가중치 1.2
중요속성(important object) : 중요한 속성, 가중치 1.1
일반속성(general object) : 일반 속성, 가중치 1.0

표 2 이벤트 추출 및 필터링 성능 평가
Table 2 Evaluation results of event extraction and event filtering

(P: Precision, R: Recall, F1: F1-Score)

Source	Sentence Count	Event Count	Average Event Count in a Sentence	Before Filtering (Micro-Average)			After Filtering (Micro-Average)			Filtering Accuracy
				P	R	F1	P	R	F1	
News	1,118	2,060	1.84	0.723	0.452	0.556	0.743	0.430	0.544	0.726
Blog	806	1,058	1.31	0.599	0.250	0.352	0.616	0.234	0.339	0.612
Tweet	806	757	0.94	0.628	0.240	0.348	0.640	0.218	0.325	0.621
All	2,730	3,875	1.42	0.682	0.358	0.470	0.707	0.346	0.465	0.722

제약에 따른 필터링은 정밀도 향상에는 도움이 되나, 재현율의 성능을 저하시켰다. F-Score도 재현율의 성능 저하로 인해 필터링 후, 성능이 하락하는 경향을 보였다. 그러나 재현율보다는 정밀도의 성능이 중요한 정보추출 분야에서는 제약규칙에 따른 필터링이 중요하다.

블로그와 트위터보다 뉴스가 문장당 템플릿 수가 많으며 성능도 우수하였다. 뉴스는 문법적으로 올바른 문장으로 구성되기 때문에 비문이 많은 블로그와 트위터에 비해서 좋은 성능을 보였다. 또한, 필터링의 성능도 우수하였다.

오류가 사용자에게 제시되지 않아야 하는 정보추출 분야에서는 빅데이터에 기반한 정밀도 중심의 접근법이 효율적일 수 있다. 이를 위해서는 재현율의 성능 저하는 있지만, 정밀도의 성능 개선에 도움이 되는 필터링이 중요하며, 비교적 비문이 적은 뉴스나 백과사전(위키피디아) 등을 활용하는 것이 바람직할 것이다.

4.2 변동계수에 기반한 이슈 순위화 성능 평가

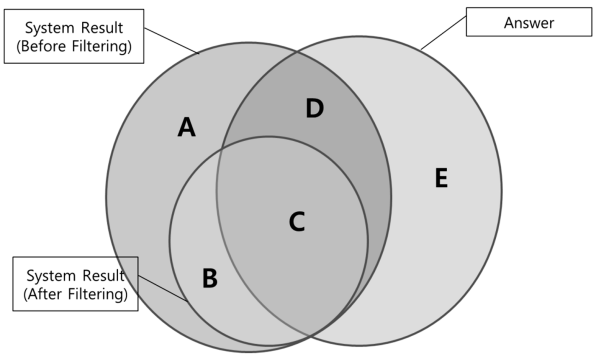
이슈 순위화 모델의 성능을 객관적으로 평가하기 위해서 네이버 오픈 API[8]를 활용하여 추출한 실시간 인물 급상승 검색어를 이용하였다. 그림 5는 네이버의 실시간 급상승 검색어의 순위 선정기준이다.

"지정된 기간동안 많이 언급된 검색어는 해당 기간동안의 주요한 이슈를 반영하고 있고, 주요한 이슈는 해당 기간동안 생성되는 컨텐츠에 많이 언급된다"는 전제를 기반으로 실시간 급상승 인물 검색어 순위와 본 시스템에서 추출된 인물 개체에 대한 순위화 결과를 비교하였다.

실시간 급상승 인물 검색어의 순위는 시간에 따라 변경된다.[9] 평가를 위해 시간단위로 인물 검색어를 수집하고, 일 단위로 통합하여 인물 검색어를 순위화하였다.

$$RankWeight_i = \sum_t^p 1/KeywordRank_i \quad (12)$$

검색어 i의 일 단위 순위화 가중치 $RankWeight_t^i$는 1을 특정시간 t에 수집된 검색어 i의 순위, $Keyword$-



	Event Before Filtering	Event After Filtering	Filtering
Precision(Accuracy)	(C+D)/(A+B+C+D)	C/(B+C)	(A+C)/(A+B+C+D)
Recall	(C+D)/(C+D+E)	C/(C+D+E)	N/A

그림 4 이벤트 추출 및 필터링 성능 평가 방법
Fig. 4 Evaluation metrics of event extraction and event filtering

8) http://dev.naver.com/openapi/apis/search/rank
9) 실시간 순위는 10위까지만 제공

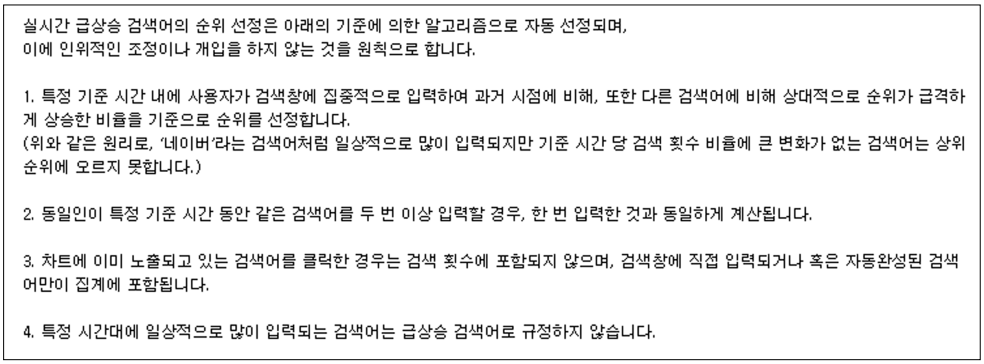


그림 5 네이버 실시간 급상승 검색어의 순위 선정 기준
Fig. 5 Ranking criterion of Naver's real-time hot keywords

$Rank_i$ 로 나눈 값의 합으로 계산된다. t 는 1이고, D 는 일 단위 시간으로 24이다. 일 단위 키워드 순위는 $RankWeight_t^i$ 값의 내림차순으로 결정된다.

$$Inclusion\ Ratio_d = \frac{시스템\ 제시\ 개체 \cap 실시간\ 검색어}{시스템\ 제시\ 수} \quad (13)$$

$$Rank\ Distance_d = \frac{\sum_i^{Inclusion\ Keywords_s}(1-\frac{|시스템제시순위_i - 실시간검색어순위_i|}{실시간\ 검색어\ 수})}{시스템\ 제시\ 수} \quad (14)$$

평가는 TF*PDF에 기반한 이슈 순위화 모델과 변동계수에 기반한 이슈 순위화 모델을 비교하였다. 평가 방법은 특정 날짜 d에 시스템이 제시한 상위 10위까지의 인물 개체 중, 실시간 급상승 검색어에 포함된 비율($Inclusion\ Ratio_d$)과 시스템에 제시한 개체 순위와 실시간 급상승 검색어의 순위 간의 평균 거리($Rank\ Distance_d$)를 이용한다. (식 (14))의 $Inclusion\ Keywords_s$ 는 시스템 제시 개체와 실시간 검색어에 모두 포함된 키워드의 집합(***시스템 제시 개체 ∩ 실시간 검색어***)이다. 본 실험에서 시스템에서 제시한 상위 10위까지의 개체만을 대상으로 하였으므로, '실시간 검색어 수'는 10이다. 변동계수를 계산하기 위한 식 (9)의 단위기간 t 는 실험에 의해서 5로 결정하였다. 평가기간은 2013년 1월 11일부터 24일까지 2주간의 결과를 이용하였다.

표 3은 TF*PDF 모델(TF*PDF)과 변동계수에 기반한 TF*PDF 모델(CV based TF*PDF)의 성능비교 평가 결과이다. 평가에서는 변동계수(CV)에 기반한 TF*PDF 모델의 성능이 TF*PDF 모델과 비교하여 우수한 성능을 보이고 있다. 소스 별 성능비교에서도 변동계수에 기반한 것이 성능향상에 도움이 됨을 알 수 있었다. 즉, "단위기간 동안 변화의 폭이 많은 이벤트가 변화의 폭이 적은 이벤트보다 이슈로서의 가치가 높다"는 전제가 실험을 통해 확인되었다.

네이버 실시간 급상승 검색어 선정은 특정 기준 시간 내에 사용자가 검색창에 집중적으로 입력하여 과거 시점에 비해 순위가 급상승한 비율이 기준이다. 이는 본 논문에서 제시한 변동계수의 특성을 일부는 반영한다. 그러나, 네이버 실시간 급상승 검색어의 선정 기준은 상승에 해당하는 비율만을 고려하지만, 변동계수는 값의 상승뿐만 아니라 하락의 정도를 모두 반영한 변동의 정도를 의미하는 계수이다. 네이버의 기준은 단위기간 내에서 급상승과 급하락을 반복하는 패턴의 검색어를 반영하기에는 한계가 있다. 또한 네이버는 검색창에 입력되는 검색 키워드를 대상으로 한 것이고, 본 논문의 개체는 수집된 문서로부터 추출한 어휘라는 차이가 있다.

전반적인 성능이 낮은 이유는 실시간 검색어 순위의 경우, 사용자들이 입력하는 검색 키워드의 변동을 기준으로 하며, 본 시스템에서 제시하는 개체의 이슈성 순위는 수집된 컨텐츠에서 추출한 개체의 빈도와 변동계수를 기반으로 제시된다는 근원적 차이에 따른 것으로 분석된다. 또한 본 시스템에서는 빅데이터 수집기에 의해서 수집된 데이터만을 대상으로 수집된 날짜를 기준으

표 3 TF*PDF와 CV-based TF*PDF의 성능비교
Table 3 Performance comparison between TF*PDF and CV-based TF*PDF

Evaluation Measure	Model	All	News	Blog	Tweet
$Inclusion\ Ratio_d$	TF*PDF	0.1714	0.3071	0.1	0.1143
	CV-based TF*PDF	0.2929	0.3143	0.1357	0.1857
$Rank\ Distance_d$	TF*PDF	0.1333	0.2566	0.0844	0.0758
	CV-based TF*PDF	0.2325	0.264	0.1162	0.1626

로 빈도를 계산한다. 이로 인해, 문서의 생성 시간(document created time)과 수집 시간(crawling time)의 시차로 인한 왜곡과 수집 시 누락된 정보에 의한 왜곡이 있을 수 있다.

뉴스를 대상으로 한 실험 결과가 블로그, 트위터와 비교하여 우수한 이유는 다음과 같이 분석된다. 사회적 이슈는 뉴스기사로 반영되던지, 뉴스로 기사화되어 이슈로 부각되는 것으로 분석된다. 그리고, 블로그는 트위터와 비교하여 상대적으로 실시간 이슈정보를 반영하지 못하는 것으로 분석되었다.

4.3 일 단위 이벤트 검색 성능 평가

일 단위 이벤트 검색 성능은 이슈 이벤트 순위화를 통해 제시되는 이벤트의 정밀도(precision)를 평가하였다. 2013년 1월 11일부터 24일까지 2주간, 일 단위로 순위화된 이슈 이벤트 상위 10위까지의 이벤트 추출 정밀도를 평가하였다. 또한 일 단위 이벤트 통합 결과에 기반한 이슈 이벤트 필터링을 적용한 순위화 결과에 대한 정밀도도 평가하였다.

표 4의 결과와 같이 뉴스를 대상으로 한 결과에서는 이벤트 통합에 의한 필터링이 적용되었을 때 다소 우수한 성능을 보였다. 그러나, 나머지의 경우 오히려 성능이 떨어졌다. 이는 뉴스는 대부분 문법적으로 올바른 정문인 반면, 블로그와 트위터는 비문이 많고 생략된 표현이 많아 상대적으로 관계추출의 오류가 많다는 점과 블로그의 펌질 현상과 트위터의 리트윗 현상에 따른 문서의 중복도가 높다는 것이 원인으로 분석된다. 즉, 관계추출의 오류를 야기하는 문장이나 문서가 펌질과 리트윗을 통해 많은 채널로 전파되고 이 문서의 중복 수집으로 인해 빈도에 기반한 이벤트 통합에서 오류를 양산하는 것으로 분석된다. 표 5는 소스 별 수집 문서의 중복정도를 평가한 것이다.

$$\text{소스별 중복정도} = \frac{(\text{전체문장수} - \text{중복제거후문장수})}{\text{전체문장수}} \quad (13)$$

표 4의 실험결과는 일 단위 전체 이벤트에 대한 이슈 순위 중, 상위 10위만을 대상으로 한 결과이다. 즉, 해당 이벤트들은 고빈도로 추출되는 이벤트들로서, 빈도에 기반한 이벤트 통합 결과가 이벤트 필터링에 크게 영향을 미치지 못하는 것으로 판단된다. 그래서, 비교적 이벤트 빈도가 적을 수 있는 특정 개체와 관련된 이벤트를 검색하여 평가하였다. 평가 대상 기간 동안의 중요한 개체 10개(질의 키워드)를 선정하고 이와 관련된 이슈 이벤트를 상위 10위까지 추출하여 평가하였다. 평가에 사용된 질의 키워드는 다음과 같다.

- 질의 키워드 목록 : LG전자, SK텔레콤, 김용준, 문재인, 박근혜, 삼성전자, 이동흡, 이명박, 포스코, 현대자동차

표 6의 결과에서 이벤트 통합에 기반하여 검색 결과 필터링을 반영한 것이 개별 소스(뉴스, 블로그, 트위터)만을 대상으로 한 검색에서 우수한 성능을 보였다. 그러나, 전체 소스를 대상으로 한 성능 비교에서는 성능향상이 없었다. 현재 시스템에서는 소스 별 이벤트 검색을 수행하고, 각 소스 별 이벤트 가중치를 단순히 합산하여 이

표 4 이벤트 통합 결과에 기반한 이슈 이벤트 필터링이 이슈 이벤트 검색에 미치는 성능 비교 평가
Table 4 Evaluation results of event-filtering effectiveness on issue-event search

	All	News	Blog	Tweet
Issue-event search precision before filtering (%)	84.3	84.29	80.71	76.43
Issue-event search precision after filtering (%)	83.6	85.71	73.57	74.29

표 5 소스 별 수집 문서의 중복정도
Table 5 Ratio of document duplication
(2012년 12월 14일 ~ 31일 수집 데이터[10])

Source	Total Sentence Count	Sentence count after filtering duplicate documents	Duplication ratio
News	5,323	3,487	0.345
Blog	13,297	6,357	0.522
Tweet	32,965	9,389	0.715

표 6 질의 키워드 10개에 대한 이슈 이벤트 검색 성능 비교 평가
Table 6 Evaluation results of issue-event search for 10 keywords

	ALL	News	Blog	Tweet
Issue-event search precision before filtering (%)	71	73.9	79.2	64.4
Issue-event search precision after filtering (%)	70.2	75.2	85.5	68.2

10) 이벤트가 추출된 문장만을 대상으로 한 결과

그림 6 소셜위즈덤의 이슈 이벤트 검색 화면
Fig. 6 Screenshot of issue-event search in SocialWisdom

벤트를 재순위화한다. 이로 인해, 소스 별 관계추출의 난이도 및 성능, 이벤트의 양(volume), 소스의 신뢰도(confidence) 등의 다양한 특성을 반영하지 못하고 있다. 이는 전체 소스를 대상으로 한 이벤트 검색에서 이벤트 통합에 기반한 검색 결과 필터링이 성능향상에 영향을 미치지 못하는 원인으로 분석된다. 향후 소스 별 주요한 특성을 이벤트 재순위화 모델에 반영할 예정이다.

5. 소셜위즈덤

그림 6은 소셜위즈덤의 이슈 이벤트 검색 화면을 갈무리한 것이다. 지정된 기간에 대해서 일별 이슈 이벤트 순위를 사용자에게 제시한다. 상단의 이슈검색 창에 특정 키워드를 입력하면 해당 키워드와 연관된 모든 이벤트를 순위화하여 제시하게 된다. 이슈 이벤트는 일 단위 이슈 이벤트(상단)와 시간 단위 이슈 이벤트(하단)로 구분하여 제공한다. 시간 단위 이슈 이벤트는 현재 시간을 기준으로 최근 몇 시간 동안의 이슈 이벤트 추이를 파악할 수 있다. 시간 단위 이슈 이벤트의 단위 시간은 3시간이며, 이벤트 통합은 수행되지 않는다. 추출된 이벤트의 목록은 좌측에 제시되며, 사용자가 원하는 이벤트만을 선택하여 볼 수 있도록 UI를 제공하고 있다. 이벤트 순위화에 반영될 소스도 선택할 수 있다. 그리고, 해당 이벤트를 클릭하면, 해당 이벤트를 추출한 원문 정보와 함께 URL을 제공한다.

표 7은 소셜 위즈덤 시스템에서 2011년 10월 01일부터 2012년 12월 31일까지 수집한 문서의 양이다. 해당 문서를 대상으로 하둡(hadoop) 플랫폼을 이용하여 이슈 이벤트를 분석하였고, 분석된 이벤트는 MySQL에 저장하였다. 소셜 위즈덤은 오늘 날짜를 기준으로 3개월 전까지의 정보만 제공한다.

그림 7은 소셜위즈덤의 이슈이벤트 검색을 이용한 의사결정 시나리오이다. 제품 A의 출시부터 브랜드 가치 상승에 이르는 과정 동안, 광고(AvdertisingStart) 및 홍보행사(EventOpen)와 연관된 의사결정 과정에서 이슈이벤트 검색이 이용되는 과정을 시나리오로 표현한 것이다. 제품 A가 출시된 후, 출시이벤트에서 이슈화 추이를 분석하고 광고를 결정한다. 그리고, 광고 후 제품 A의 광고와 관련된 이슈 추이를 분석하고 홍보행사를 결정하여 제품 A의 브랜드 가치 상승을 이끌 수 있다.

표 7 수집 문서의 수(2011년 10월 01일 ~ 2012년 12월 31일)
Table 7 The number of crawling documents (01-Oct-2011 ~ 31-Dec-2012)

	뉴스(news)	블로그(blog)	트위터(tweet)	총합(total)
문서 수 (the number of documents)	1,372,956	25,021,945	695,165,728	721,256,629

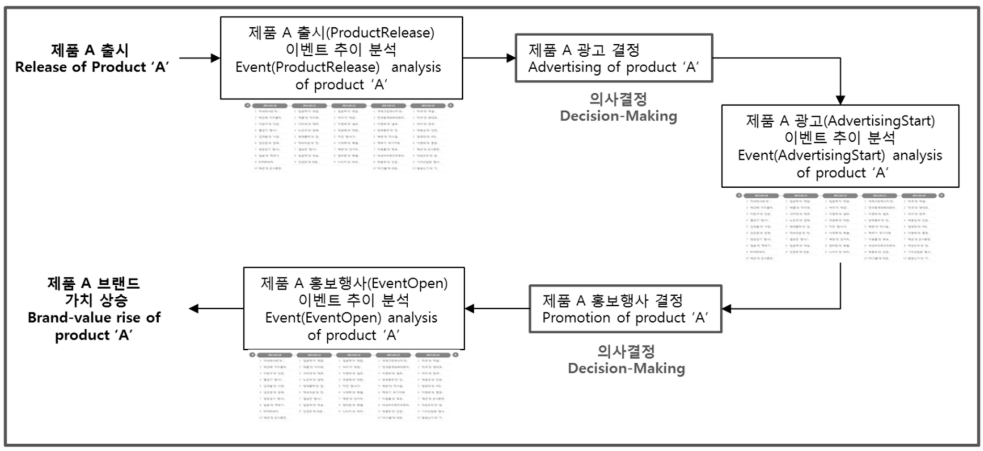


그림 7 이슈 이벤트 검색기를 이용한 의사결정의 예
Fig. 7 An example of decision-making by using the issue-event search system

6. 결론 및 향후 연구방향

본 논문에서는 빅데이터에 기반한 의사결정 지원 시스템인 소셜위즈덤의 이슈 이벤트 추출 및 검색에 대하여 소개하였다. 이슈 이벤트 추출 및 검색의 세가지 기술에 대해서 다음과 같이 소개하였고, 그 성능을 평가하였다.

첫째, 온톨로지에 기반한 이벤트 제약 및 필터링 방법 4가지를 소개하였다. 이벤트 온톨로지의 의미적 제약 및 관계에 기반한 이벤트 필터링은 이벤트 추출 정밀도 성능을 향상시키나, 재현율의 성능은 다소 하락시켰다. 정보의 신뢰성이 중요한 정보추출에서는 재현율보다는 정밀도 향상에 도움이 되는 필터링 기술이 중요하다. 그리고, 빅데이터의 데이터 잉여성(data redundancy)에 기반한다면, 재현율보다는 정밀도에 중점을 두는 전략이 필요할 것이다. 소스 별 성능을 비교한 결과, 비교적 정문으로 구성된 뉴스에서 이벤트 추출 성능이 좋으며, 다양한 제약에 기반한 필터링이 효과적임을 알 수 있었다.

둘째, 이벤트 통합 결과에 기반한 이벤트 검색 필터링의 효과에 대해서 소개하였다. 빈도와 다양한 자질에 기반한 이벤트 통합 결과를 기반으로 이벤트를 필터링하는 방법은 고빈도 이벤트에서는 효과가 없으나, 비교적 빈도가 낮은 이벤트들에서는 효과적인 것을 알 수 있었다.

셋째, 변동계수를 이용한 이슈 이벤트 순위화 방법에 대해서 소개하였다. 일정기간 동안 빈도 추이 변화가 많은 이벤트가 적은 이벤트보다 이슈성(hotness) 가치가 높을 것이라는 전제를 기반으로 변동계수를 이용한 이슈 이벤트 순위화 알고리즘을 개발하였다. 단순 TF*PDF만을 이용한 순위화 알고리즘과 변동계수를 함께 이용한 알고리즘의 성능비교 결과, 변동계수를 함께 이용한 경우 많은 성능향상이 있었다. 정보의 이슈성을 평가할 때, 특정 기간 동안의 빈도 추이에 대한 변동계수가 이슈성 평가의 중요한 자질임을 알 수 있었다.

본 연구에서 소개한 이슈 이벤트 추출 및 검색 기술은 소셜 미디어 상의 주요한 이슈, 이벤트, 감성정보 등을 제공하는 의사결정 지원 시스템인 소셜위즈덤 서비스에 포함되어 있다. 기존 시스템들의 경우, 주요 개체에 대한 빈도 및 감성 정보만을 제공하여 개체들 간의 의미적 연관성에 기반한 정보를 파악하기 쉽지 않은 단점이 있었다. 그러나 본 논문에서는 이벤트 템플릿에 기반하여 키워드와 연관된 이벤트를 순위화하여 제시함으로써, 시간대별 개체들 간의 연관성을 파악할 수 있어서 의사결정에 많은 도움을 제공할 것으로 판단된다.

향후 연구계획은 이벤트의 추가, 확장이 용이할 수 있도록, 이벤트 추출 기술에 대한 적응성(adaptation) 향상을 위한 오픈 정보 추출(open IE) 기술에 대한 연구를 진행할 예정이다. 또한, 정보추출에서 가장 중요한 두 가지 정보인 시간정보(temporal information)와 지역(spatial information)정보를 이용한 정보 시각화 기술에 대해서 연구할 예정이다. 특히, 시간정보를 기반으로 주

요 개체에 대한 이벤트들을 타임라인 상에 연대기로 요약하여 제공하는 기술은 빅데이터에 기반한 정보추출 및 요약에서 중요한 기술이 될 것이다.

참 고 문 헌

[1] "Special Report : The next Google," *Nature*, vol.455, Sep. 2008.

[2] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, Duncan J. Watts, "Everyons's an Influencer : Quantifying Influence on Twitter," *In Proceedings of the fourth ACM international conference on Web Search and Data Mining (WSDM)*, pp.65-74, 2011.

[3] Sitaram Asur, Bernardo A. Huberman, "Predicting the Future With Social Media," *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference*, vol.1, pp.492-499, 2010.

[4] IBM Watson [Online]. Available: http://www-03.ibm.com/innovation/us/waston/

[5] Takeshi Sakaki, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," WWW 2010, 2010.

[6] Panagioris T. Metaxas, "How (Not) to Predict Elections," *2011 ieee third international conference on social computing*, 2011.

[7] Recorded Future [Online]. Available: http://www.recordedfuture.com

[8] Grishman, Ralph, Beth Sundheim, "Message understanding conference-6: A brief history," *Proceedings of COLING*, vol.96, 1996.

[9] Douglas E. Appelt, "Introduction to information extraction," *Ai Communications*, pp.161-172, 1999.

[10] Grishman, Ralph, "Information extraction: Techniques and challenges," *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology, Springer Berlin Heidelberg,* pp.10-27, 1997.

[11] C. Lee, "Named Entity Recognition with Structural SVMs and Pegasos algorithm," *Korean Journal of Cognitive Science*, vol.21, no.4, 2010.

[12] Heng Ji, Ralph Grishman and Hoa Trang Dang, "An Overview of the TAC2011 Knowledge Base Population Track," *In Proceedings to the Third Text Analytics Conference (TAC2011)*, 2011.

[13] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios, "Duplication Record Detection: A Survey," *IEEE Transactions on Knowledge And Data Engineering*, vol.19, no.1, Jan. 2007.

[14] E. Ukkonen, "Approximate String Matching with q-Grams and Maximal Matches," *Theoretical Computer Science*, vol.92, no.1, 1992.

[15] Yan Gao, Jin Liu, PeiXun Ma, "The Hot Keyphrase Extraction based on TF*PDF," *2011 International Joint Conference of IEEE TrustCom-11/IEEE ICESS-11/FCST-11*, 2011.

[16] Zhiqi Fang, Yue Ning, Tingshao Zhu, "Hot Keyword Identification for Extracting Web Public Opinion," *Pervasive Computing and Applications (ICPCA)*, 2010.

[17] Kuan-Yu, Luesak Luesukprasert, Seng-cho T. Chou, "Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling," *IEEE Transactions on Knowledge and Data Engineering*, vol.19, no.8, Aug. 2007.

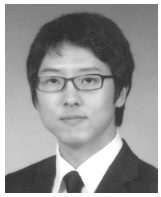
[18] Khoo Khyou Bun, Mitsuru Ishizuka, "Topic Extraction from News Archive Using TF*PDF Algorithm," *Web Information Systems Engineering*, 2002.

허 정

1999년 울산대학교 전자계산학과(공학사) 2001년 울산대학교 전자계산학과(공학석사). 2001년~현재 한국전자통신연구원 지식마이닝연구팀 선임. 2013년~현재 울산대학교 전자계산학과(박사과정). 관심분야는 자연어처리, 정보검색, 텍스트마이닝

류 법 모

1995년 경북대학교 컴퓨터공학과(공학사) 1997년 POSTECH 대학원 컴퓨터공학과(공학석사). 2009년 KAIST 대학원 전자전산학과(공학박사). 1997년~1999년 한국전자통신연구원 언어이해연구팀 연구원. 1999년~2004년 (주)케이포엠 기술연구소 연구원. 2009년~현재 한국전자통신연구원 지식마이닝연구팀 선임. 관심분야는 정보검색, 자연어처리, 온톨로지



최 윤 재

2007년 서울대학교 컴퓨터공학과(공학사) 2009년 KAIST 대학원 전산학과(이학석사). 2010년~현재 한국전자통신연구원 지식마이닝연구팀 연구원. 관심분야는 정보검색, 자연어처리

김 현 기

1994년 전북대학교 컴퓨터공학과(공학사) 1996년 전북대학교 컴퓨터공학과(공학석사). 2005년 Univ. of Florida 컴퓨터공학과(공학박사). 1995년~현재 한국전자통신연구원 지식마이닝연구팀 팀장. 관심분야는 자연어처리, 기계학습, 정보검색, 지식마이닝

옥 철 영

정보과학회논문지 : 소프트웨어 및 응용 제 39 권 제 5 호 참조