소셜위즈덤: 소셜미디어 이슈 탐지/모니터링 시스템

이충희*, 김현진*, 오효정*, 허정*, 류법모*, 김현기*
*한국전자통신연구원
{forever, jini, ohj, jeonghur, pmryu, hkk}@etri.re.kr

Social WISDOM: A Issue Detection/Monitoring System

Chung-Hee Lee*, Hyun-Jin Kim*, Hyo-Jung Oh*, Jeong Hur*, Pum-Mo Ryu*, Hyun-Ki Kim*,
*Electronics and Telecommunications Research Institute

요 약

본 논문에서는 소셜 빅데이터에 대한 심층적 언어분석을 통해 이슈를 탐지하고 모니터링하는 소셜위즈덤 시스템을 소개한다. 소셜위즈덤은 키워드의 단순 빈도 정보 외에도 이슈의 신규성, 중요성, 파급력, 관심도, 신뢰도 등을 수치화한 이슈성지수에 기반한 이슈성 측정이 가능하여 정확한 이슈탐지가 가능하다. 또한, 추가적인 정보로 단순 긍부정 분석이 아닌 17 개의 세부감성을 분석해서 제공하고 긍부정에 대한 호불호의 원인분석 정보도 제공하므로, 소셜미디어 분석에 기반한 깊은 인사이트를 제공하여 사용자의 의사결정에 많은 도움을 줄 수 있다.

1. 서론

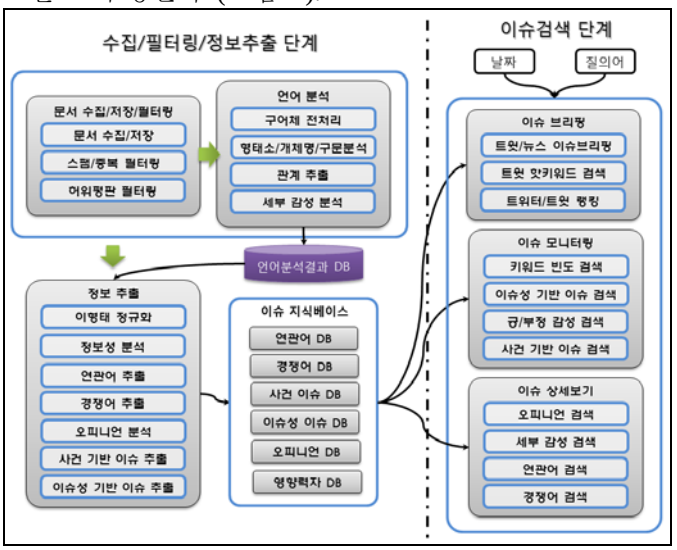
최근에는 스마트폰의 보편화된 사용과 클라우드 기술의 발전으로 개인과 조직의 활동기록이 축적되면서 활용할 수 있는 정보의 양이 폭발적으로 증가하고 있고, 특히 소셜미디어 기반의 인터랙션 데이터를 분석하여 활용하자는 요구가 급증하고 있다[1]. 소셜미디어 상의 인터랙션 데이터는 현재 시점의 사회 구조 특징 및 사회 구성원의 행동 패턴을 내재하고 있으므로, 소셜미디어를 이해하고 유용한 정보를 추출하여 현재의 주요 이슈를 탐지하고 모니터링하여 미래를 예측하기 위한 연구가 활발히 진행되고 있다[2].

기존의 소셜미디어 분석서비스는 키워드 단순 빈도 정보와 단순 긍부정 정보 등을 분석해서 결과로 제시하고 있다[3]. 하지만, 기존 서비스는 항상 많이 출현하는 단어나 계절적, 이벤트성으로 정기적으로 출현하는 고빈도 이슈와 실제 이슈성이 높아진 이슈를 자동으로 구분하기 힘든 문제점이 있다. 본 논문에서는 심층 언어분석 기술과 정보추출 기술을 활용하여, 현재 이슈를 탐지/모니터링하고 세부 오피니언 정보를 제공함으로써 사용자의 의사결정에 도움을 줄 수 있는 소셜위즈덤 시스템을 소개한다. 2 장에서는 소셜위즈덤 시스템에 대해서 상세히 설명하고, 3 장에서 소셜위즈덤 시스템을 외부에 오픈하여 시범서비스한 내용을 소개한다. 4 장에서 결론 및 향후 방향에 대해서 설명한다.

2. 소셜위즈덤 시스템

소셜위즈덤 시스템은 다양한 소셜웹 콘텐츠(뉴스/블로그/트위터)로부터 이슈를 탐지 및 모니터링하여, 이슈의 향후 전개과정에 대한 예측결과를 제시하는 것을 목적으로 한다.

소셜위즈덤은 문서 수집/저장/필터링 모듈, 언어분석 모듈, 정보추출 모듈, 이슈탐지 모듈, 이슈모니터링 모듈로 구성된다 (그림 1).



(그림 1) 소셜위즈덤 시스템 구성도

2.1. 수집/필터링/정보추출 단계

이슈 탐지는 뉴스, 블로그, 트위터를 대상으로 하고 있으며, 뉴스, 블로그 문서는 크롤러를 통해 자동으로 수집하고, 트위터 문서는 공개된 스트리밍 API 를 사용해서 수집한다. 수집된 문서들은 스팸/중복/허위평판 필터링을 통해 무의미한 문서들이 제거되고[4], 양질의 문서들에 대해서 언어분석을 수행해서 언어분석 결과를 지식베이스로 저장한다.

언어분석과정은 한 개의 입력 문서를 대상으로 문장분할, 띄어쓰기 오류수정[5]을 포함하는 구어체 전처리 과정과 형태소분석, 개체명인식[6], 구문분석, 의미관계분석[7], 감성분석[8] 단계를 순차적으로 진행한다.

이슈는 소셜미디어 상에서 빠르게 전파되고 생성/소멸이 빈번하게 발생하므로 이슈 탐지/모니터링에서 실시간성은 매우 중요하다. 그러므로, 소셜위즈덤은 바로 전날 데이터까지 이슈를 분석해서 서비스하고 있으며, 매일 수집되는 약 370 만 문서를 당일에 언어분석하고 정보추출까지 완료하기 위해서 대표적인 빅데이터 분산 처리 플랫폼인 Hadoop[9]과 비관계형 분산 데이터베이스인 HBase[10]를 사용해서 정보추출 시스템을 구성하고 있다. Hadoop/HBase 시스템은 1 개의 Master 와 30 개의 Slave 로 구성된 30 대의 서버로 운용된다. 하루에 수집되는 문서 개수와 언어분석 시간은 <표 1>과 같다.

<표 1> 수집문서 및 언어분석시간(하루 단위)

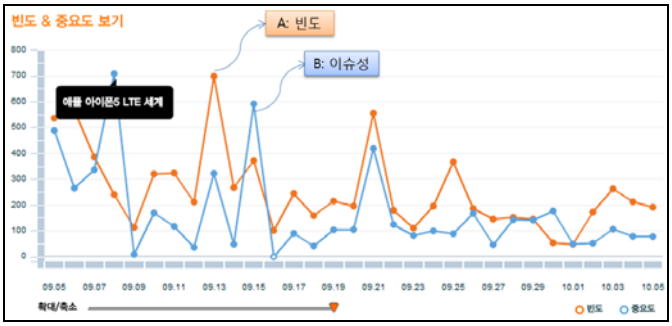
매체	수집량	언어분석 시간[1]	언어분석 결과
뉴스	3,800 문서 (8.5MB)	3 분	45MB
블로그	210,000 문서 (240MB)	73 분	2.25GB
트위터	3,500,000 문서 (1.35GB)	113 분	3.74GB
합계	3,713,800 문서 (1.59GB)	189 분	6.04GB

정보추출과정은 언어분석결과를 기반으로 이슈 모니터링에 필요한 다양한 정보를 추출하는 과정이며, 이형태 정규화 모듈, 정보성 분석 모듈, 연관어 추출 모듈, 경쟁어 추출 모듈, 오피니언 분석 모듈, 사건 기반 이슈추출 모듈, 이슈성 기반 이슈추출 모듈로 구성된다. 이형태 정규화 모듈은 “주식회사 애플”, “Apple”, “Apple Inc.”, “(주)애플”과 같은 이형태들을 원형인 “애플”로 복원해 준다. 원형은 가장 많이 사용되는 형태가 선정하였고, 약 15 만개의 원형/이형태 어휘쌍이 구축되어 사용된다. 정보성분석 모듈은 트위터의 파급력을 분석하기 위한 모듈로써, 트위터 사용자와 트윗 문서 자체의 영향력을 분석한다[11,12]. 연관어추출 모듈은 문장 단위를 기반으로 연관성이 높은 (복합)명사, 개체명들을 추출해서 지식베이스로 저장한다. 단어 연관도를 측정하기 위해서 Jaccard’s coefficient 와 Chi-square 변형수식을 적용하였고, 기간별/미디어별 가중치 랭킹 모델을 추가적으로 적용하였다. 경쟁어추출 모듈은 “애플↔삼성”과 같이 경쟁관계에 있는 2 개의 개체명 쌍을 추출해서 지식베이스로 저장한다. 경쟁어 추출 방법은 한 문장에 출현한 동일한 유형의 개체명이 경쟁관계에 있는지 없는지를 분류하는 바이너리 분류문제로 접근하였다[13]. 분류모델은 Support Vector Machines(SVMs)을 사용하였고, 경쟁표현이 있는 문장과 없는 문장을 각각 positive 와 negative 학습데이터로 구축해서 SVMs 모델 학습에 이용하였다. 오피니언분석 모듈은 정책/기업/상품/인물에 대한 찬성/반대, 장점/단점, 호/불호의 이유를 분석한다. 오피니언분석 방법은 개체명에 기반해서 대상을 선정하고, 서술형인식기를 통해 선정된 대상의 원인에 대한 문서를 추출하고[14], 추출된 서술형 문서에 대해서 클러스터링 방법을 통해 유사한 문서들의 클러스터를 생성하고, 마지막으로 각각의 클러스터에 레이블을 생성해서 지식베이스로 저장한다. 사건 기반 이슈 추출 모듈은 현재 기업, 공공분야에서 많이 이슈화 되고 있는 사건유형 30 여 개를 미리 정의하고, 모든 문서를 대상으로 매칭되는 사건을 이슈로 추출해서 템플릿으로 저장한다[15]. 이슈성 기반 이슈추출 모듈은 단순 빈도수에 기반해서 이슈어를 추출하는 것이 아니라 이슈성이 반영된 이슈어를 추출한다. 이슈성 측정방법은 이슈 대상어들의 신규성, 중요성, 파급력, 관심도 및 미디어 별 신뢰도 등을 수치화하여 이슈성 지수로 사용한다.

2.2. 이슈검색 단계

이슈검색 단계는 사용자가 날짜와 질의어를 입력하면 질의어와 관련된 이슈를 탐지해서 모니터링 결과를 보여주는 과정으로 구성된다. 세부적으로는 이슈 모니터링, 이슈 상세보기 과정으로 구분된다.

이슈 모니터링 과정은 입력된 질의어와 관련된 이슈들을 상세히 보여준다.



(그림 2) 빈도 및 이슈성 기반 이슈 검색 화면

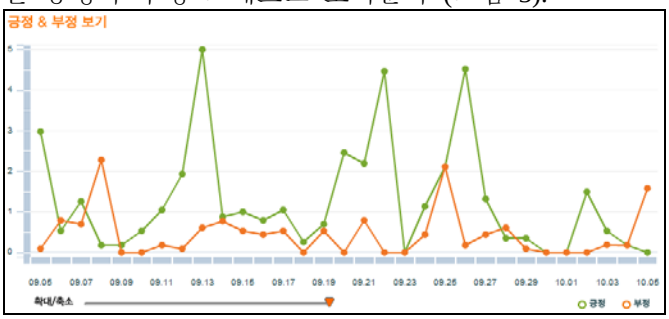
(그림 2)는 “애플” 질의어에 대한 빈도 및 이슈성 결과를 보여준다. A 그래프는 질의어의 단순빈도 정보를 나타내고, B 그래프는 이슈성 정보를 나타낸다. B 그래프의 특정날짜를 클릭하면 해당 날짜의 이슈성이 가장 높았던 이슈를 레이블로 보여준다. (그림 2)에서는 9 월 8 일의 이슈로 “애플 아이폰 5 LTE 세계”가 추출된 것을 보여주는데, 아이폰 5 가 LTE 로 나온다는 것이 빈도 정보는 낮지만 실제로 이슈화가 되었음을 알 수 있다.

키워드 빈도 검색 모듈은 질의어의 매체 별 출현 빈도정보를 날짜 별로 보여주며(그림 2-A), 기존 소셜미디어 분석서비스들에서 보여주는 정보와 유사하다. 이슈성 기반 이슈 검색 모듈은 질의어의 날짜 별 이슈를 단순 빈도가 아닌 이슈성 지수에 기반해서 보여준다(그림 2-B). 기존의 빈도기반 이슈측정 방법에서는 항상 많이 출현하는 단어나 계절적, 이벤트성으로 정기적으로 출현하는 고빈도 단어와 실제 이슈화된 단어를 구분하기 힘든 반면, 본 모듈에서는 빈도 정보 외에도 신규성, 중요성, 파급력, 관심도, 신뢰도 등을 수치화한 이슈성 지수에 기반하므로 이슈성 여부를 정확히 분석할 수 있다.

긍/부정 감성 검색 모듈은 질의어의 날짜 별 감성도

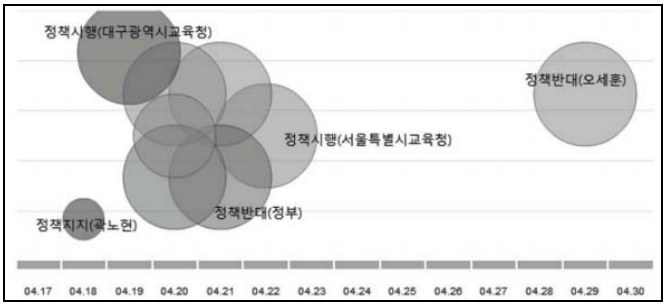
[1] 30 대 Hadoop 시스템을 기준으로 계산된 것임

를 긍정과 부정 그래프로 보여준다 (그림 3).



(그림 3) "애플"의 긍정/부정 검색 화면

사건 기반 이슈 검색 모듈은 미리 정의된 30 여 개의 사건유형을 기반으로 질의어와 관련된 사건 중 이슈성이 높은 상위 5 개를 <사건유형, 사건주체, 사건대상> 트리플 형태로 날짜 별로 보여준다(그림 4). 각각의 원은 "무상급식"과 관련된 사건을 나타내고, 동일한 수평선에 위치한 원은 동일 사건으로써 날짜 별 정보를 보여준다. 원의 크기는 해당 사건의 이슈성 정도를 나타낸다. 사건 단위 모니터링은 특정 개체에 대한 구체적인 현상들을 세분화하여 관찰할 수 있기 때문에 단순 빈도 기반 모니터링보다 깊은 인사이트를 제공한다.



(그림 4) "무상급식"의 사건 기반 이슈 검색 화면

이슈 상세보기 과정은 이슈정보 외에 인사이트를 얻을 수 있는 세부 정보를 제공한다.

(그림 5) 오피니언 검색 화면

오피니언 검색 모듈은 질의어에 대한 호/불호의 원인 정보를 제공한다. 예를 들어, "아이폰" 질의어에 대해서 좋아하는 이유로 "최고성능 카메라" 등이 나오고, 싫어하는 이유로 "수리비용이 많이 청구됨" 등이 제공되며(그림 5), 분석결과를 통해서 전반적인 호감도 뿐 아니라 여론의 세부적 분석을 통해 호/불호의 원인파악이 가능하다.

세부 감성 검색 모듈은 질의어에 대한 감성분석 결과로 17 개의 세부 감성 분석 결과를 제공한다 (그림 6).

(그림 6) "아이폰"의 세부 감성 검색 화면

(그림 7)은 긍정/부정의 감성분류에서 보다 세분화된 17 개 세부 감성 분류표를 보여준다.

(그림 7) 소셜위즈덤 세부감성 분류표

연관어 검색 및 경쟁어 검색 모듈은 질의어에 대한 연관어 및 경쟁어 정보를 제공한다. (그림 8)은 "애플" 질의어에 대한 연관어와 경쟁어 검색 화면을 보여준다.

(그림 8) 연관어, 경쟁어 검색 화면

3. 시범서비스 및 소셜미디어 추이 분석

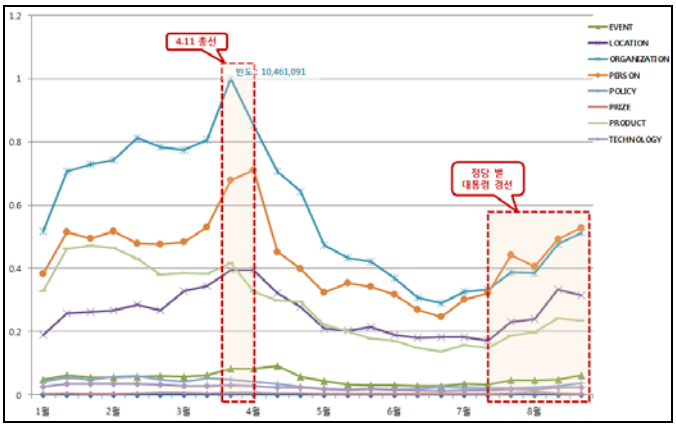
소셜위즈덤 시스템은 2011 년부터 개발되었고, 2012 년 6 월 20 일부터 9 월 19 일까지 3 개월 동안 외부에 오픈해서 시범서비스를 진행하였다. <표 2>는 시범서비스 기간 동안 소셜위즈덤에 접속한 사용자 접속빈도 정보이다. 3 개월 동안, 36,369 개의 IP 가 접속되었고, 77,545 개의 질의어가 입력되었다.

(그림 9)는 전체 기간 중 소셜미디어 상에 많이 나타난 상위 8 개 엔티티 유형들의 추이 변화 그래프이다. 전체적으로 가장 많이 나타난 유형은 기관, 인물, 상

품, 지역 4 개이며, 4.11 총선 기간 중에는 기관과 인물이 두드러지게 많이 출현하였다.

<표 2> 시범서비스 접속현황

항목		접속빈도
전체 접속 IP	전체	36,369
	주 평균	466
엔진유형별 접속 IP	이슈 브리핑	1,510
	이슈 모니터링	3,158
	이슈 상세보기	1,160



(그림 9) 소셜미디어 추이 분석

3 개월 동안 입력된 전체 질의어는 4,318 개이며, 중복을 제외한 단일한 질의어는 975 개이다. 고빈도 질의어는 (그림 10)과 같다.

애플 -> 1240개	갤럭시S3 -> 107개	알칼리 -> 59개
박근혜 -> 405개	마계촌온라인 -> 103개	단어장 -> 59개
한일군사협정 -> 323개	옵티머스G -> 94개	KB카드 -> 59개
안철수 -> 284개	삼성물산 -> 90개	카카오톡 -> 57개
삼성 -> 277개	VIPS -> 86개	기품원 -> 57개
경동나비엔 -> 267개	녹조 -> 81개	코카콜라 -> 56개
삼성전자 -> 211개	갤럭시노트2 -> 80개	삼성카드4 -> 55개
탐론 -> 197개	보일러 -> 78개	김연아 -> 55개
국방기술품질원 -> 180개	지해 -> 76개	티아라 -> 54개
방위사업청 -> 179개	아이패드 -> 73개	엔씨소프트 -> 54개
화장품 -> 176개	옥션 -> 72개	박지원체포 -> 54개
삼성특허소송 -> 155개	싼타페 -> 70개	민주당 -> 53개
문재인 -> 154개	폴무원 -> 68개	모의고사 -> 53개
손학규 -> 147개	전투복 -> 67개	조완섭 -> 52개
닉쿤 -> 140개	이투스 -> 66개	공천헌금 -> 52개
기아자동차 -> 140개	귀뚜라미보일러 -> 66개	K9 -> 52개
대성셀틱 -> 138개	하나은행 -> 65개	최창곤 -> 51개
올림픽 -> 137개	캐시비 -> 64개	저녁이있는삶 -> 51개
기아차 -> 128개	아이폰 -> 64개	봉투모의고사 -> 51개
아이폰5 -> 126개	경제끝장토론 -> 64개	갤럭시s3 -> 51개
이명박 -> 122개	성범죄 -> 62개	락앤락 -> 50개
처음처럼 -> 117개	박태환 -> 62개	롯데 -> 48개
FTA -> 117개	tgif -> 62개	대성에너시스 -> 48개
알파스캔 -> 112개	미션플러스 -> 61개	이무인 -> 46개
4대강 -> 111개	차세대전투기 -> 60개	박지원 -> 46개
스마트폰 -> 109개	피부고민 -> 59개	교보생명 -> 46개
삼성카드 -> 109개	청계재단 -> 59개	갤럭시3 -> 46개
불법사찰 -> 109개	이효리 -> 59개	g마켓 -> 46개

(그림 10) 시범서비스 현황: 고빈도 질의어 정보

4. 결론

기존에 서비스되고 있는 대부분의 소셜미디어 분석 시스템은 단순 빈도 정보와 단순 긍부정 정보만을 기반으로 이슈를 탐지하므로, 항상 많이 출현하는 단어나 계절적, 이벤트성으로 정기적으로 출현하는 고빈도 이슈와 실제 이슈성이 높아진 이슈를 자동으로 구분하기 힘든 문제점이 있다. 그에 비해서, 소셜위즈덤에서는 빈도 정보 외에도 신규성, 중요성, 파급력, 관심도, 신뢰도 등을 수치화한 이슈성 지수에 기반한 이슈성 측정이 가능하여 정확한 이슈탐지가 가능하고, 17 개 세부감성분석 및 호/불호에 대한 원인분석 정보 제공을 통해 사용자의 의사결정에 많은 도움을 줄 수 있다. 향후 연구방향으로는 과거 및 현재의 소셜미디어 분석 결과를 이용해서, 이머징 이슈를 사전에 탐지하기 위한 기술을 개발할 계획이다.

참고문헌

[1] Informatica,“Harnessing Big Data”, http://www.informatica.com/us/vision/harnessing-big-data/

[2] Sheng Yu, Subhash Kak, “A Survey of Prediction Using Social Media”, The Computing Research Repository (CoRR), 2012

[3] 류법모, 김현진, 김현기, 박상규, “심층 언어분석 기반 소셜미디어 이슈 탐지 및 모니터링 기술”, 정보과학회지, 제 30 권, 제 6 호, 2012, pp. 47-58.

[4] Yeo-Chan YOON, Myung-Gil JANG, Hyun-Ki KIM, and So-Young PARK, "Detecting Partial and Near Duplication in the Blogosphere," IEICE TRANSACTIONS on Information and Systems, Vol.E95-D, No.2, 2012, pp.681-685.

[5] 이창기, 김현기, “Structural SVM 을 이용한 한국어 자동 띄어쓰기”, 2012 한국컴퓨터종합학술대회, pp. 270 -272.

[6] C.K. Lee and M.G. Jang, "A Prior Model of Structural SVMs for Domain Adaptation," ETRI J., vol. 33, no. 5, 2011, pp. 712-719.

[7] C. Lee, Y. Hwang, and M. Jang, “Fine-Grained Named Entity Recognition and Relation Extraction for Question Answering,” Proceedings of SIGIR 2007, pp. 799-800.

[8] Yoonjung Choi, Hyo-Jung Oh, and Sung-Hyon Myaeng, "A Generate and Test Method of Detecting Negative-Sentiment Sentences," Proceedings of CICLing 2012, pp. 500-512.

[9] Apache Hadoop, http://hadoop.apache.org/

[10] Apache HBase, http://hbase.apache.org

[11] Min-Chul Yang, Jung-Tae Lee, and Hae-Chang Rim, “Using Link Analysis to Discover Interesting Messages Spread Across Twitter,” Proceedings of ACL 2012, pp. 15-19.

[12] Min-Chul Yang, Jung-Tae Lee, Seung-Wook Lee, and Hae-Chang Rim, “Finding Interesting Posts in Twitter Based on Retweet Graph Analysis,” Proceedings of SIGIR 2012, pp. 1073-1074.

[13] 이충희, 김현진, 류법모, 김현기, 서영훈, “기계학습 기반 경쟁자 자동추출 방법”, 제 24 회 한글 및 한국어정보처리 학술대회, 2012 (2012.10.13 게재 예정).

[14] Yeo-Chan YOON, Chang-Ki Lee, Hyun-Ki KIM, Myung-Gil JANG, Pum Mo RYU, and So-Young PARK, “Descriptive Question Answering with Answer Type Independent Features,” IEICE TRANSACTIONS on Information and Systems, Vol.E95-D, No.7, 2012, pp.2009-2012.

[15] 허정, 류법모, 최윤재, 김현기, “소셜미디어 기반 의사결정 지원을 위한 이벤트 템플릿 추출”, 제 24 회 한글 및 한국어정보처리 학술대회, 2012