KIPS Tr. Software and Data Eng.
Vol.3, No.12 pp.553~564 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2014.3.12.553

Automatic Generation of Issue Analysis Report Based on Social Big Data Mining

Heo Jeong[†] · Lee Chung Hee[†] · Oh Hyo Jung[††] · Yoon Yeo Chan[†††] · Kim Hyun Ki[††]
Jo Yo Han[††††] · Ock Cheol Young[†††††]

ABSTRACT

In this paper, we propose the system for automatic generation of issue analysis report based on social big data mining, with the purpose of resolving three problems of the previous technologies in a social media analysis and analytic report generation. Three problems are the isolation of analysis, the subjectivity of experts and the closure of information attributable to a high price. The system is comprised of the natural language query analysis, the issue analysis, the social big data analysis, the social big data correlation analysis and the automatic report generation. For the evaluation of report usefulness, we used a Likert scale and made two experts of big data analysis evaluate. The result shows that the quality of report is comparatively useful and reliable. Because of a low price of the report generation, the correlation analysis of social big data and the objectivity of social big data analysis, the proposed system will lead us to the popularization of social big data analysis.

Keywords : Social Big Data Mining, Automatic Report Generation, Issue Analysis Report, Correlation Analysis of Social Big Data

소셜 빅데이터 마이닝 기반 이슈 분석보고서 자동 생성

허 정[†] · 이 충 희[†] · 오 효 정[††] · 윤 여 찬[†††] · 김 현 기[††] · 조 요 한[††††] · 옥 철 영[†††††]

요 약

본 논문은 지금까지의 소셜미디어 분석과 분석보고서 생성의 세 가지 문제점을 해결하기 위해서 소셜 빅데이터 마이닝에 기반한 이슈 분석보고서 자동 생성 시스템을 제안한다. 세 가지 문제점은 분석의 고립성, 전문가의 주관성과 고비용에 기인한 정보의 폐쇄성이다. 시스템은 자연언어 질의분석, 이슈분석, 소셜 빅데이터 분석, 소셜 빅데이터 상관성분석과 자동 보고서 생성으로 구성된다. 생성된 보고서의 유용성을 평가하기 위해, 본 논문에서는 리커트척도를 사용하였고, 빅데이터 분석 전문가 2명이 평가하였다. 평가결과는 리커트 척도 평가에서 보고서의 품질이 비교적 유용하고 신뢰할 수 있는 것으로 평가되었다. 보고서 생성의 저비용, 소셜 빅데이터의 상관성 분석과 소셜 빅데이터 분석의 객관성 때문에, 제안된 시스템이 소셜 빅데이터 분석의 대중화를 선도할 것으로 기대된다.

키워드 : 소셜 빅데이터 마이닝, 보고서 자동 생성, 이슈 분석보고서, 소셜 빅데이터 상관성 분석

1. 서 론

웹 환경의 급속한 변화는 디지털 콘텐츠 시장(digital contents market)의 생태계를 크게 변화시키고 있다. 웹 1.0의 시기에는 대형 미디어 매체들이 일방적으로 콘텐츠를 제공하고, 사용자들이 콘텐츠를 단순히 소비하는 형태였다. 웹 2.0 환경에서는 집단지성에 기반한 다양한 형태의 콘텐츠 생산 플랫폼(platform)이 제공되었고, 이로 인해 웹 1.0에서 단순 소비자였던 사용자들이 콘텐츠를 생산, 유통 및 소비를 하게 되었다. 또한, 스마트폰(smart-phone)과 태블릿컴퓨터(tablet PC)을 중심으로 한 모바일(mobile) 환경으로의 급속한 변화로 다양한 소셜미디어(social media)가 성장하게 되었다. 웹 환경 변화로 인한 디지털 콘텐츠 생산자와 소비자의 통합은 디지털 콘텐츠 내에 사회문화적 다양한 의견 및 여론 추이를 파악할 수 있는 많은 정보가 내포될 수 있다는 것을 의미한다. 이런 이유로 많은 기업과 기관을 중심으로 기업과

† 정 회 원 : 한국전자통신연구원 선임연구원
†† 정 회 원 : 한국전자통신연구원 책임연구원
††† 비 회 원 : 한국전자통신연구원 선임연구원
†††† 비 회 원 : Carnegie Mellon University 석사과정
††††† 종신회원 : 울산대학교 전기공학부 IT융합전공 교수
Manuscript Received : July 11, 2014
First Revision : November 19, 2014
Accepted : November 24, 2014
* Corresponding Author : Ock Cheol Young(okcy@ulsan.ac.kr)

상품 브랜드 및 기관에 대한 여론 동향을 파악하고 의사결정을 지원하기 위한 소셜미디어 분석이 활발하게 이루어지고 있다.

소셜미디어 분석은 정보추출(information extraction)에 기반한 콘텐츠 내용분석과 소셜미디어의 구조적 연관성을 분석하는 네트워크 분석으로 구분할 수 있다. 콘텐츠 내용분석은 주로 콘텐츠에 기술된 주요한 개체들(entities)의 노출(buzz)추이 및 감성분석(sentiment analysis)이 중심이고, 네트워크 분석은 트위터(twitter)나 페이스북(facebook)과 같은 소셜미디어 플랫폼에서 사용자들 간의 콘텐츠 유통 및 확산 추이 분석이 핵심기술이다[1].

소셜미디어 콘텐츠 내용분석은 특정 개체의 시간별 빈도 및 중요도 변화를 분석하는 노출추이 분석, 감성정보 분석, 사건(event) 분석과 연관어 및 경쟁어 분석 등이 있다. 네크워크 분석은 콘텐츠의 확산 추이 분석 및 예측, 영향력자 분석 등이 있다.

소셜미디어 데이터는 모바일 환경의 활성화로 인해 기하급수적으로 늘어나고 있으며, 다양한 사건과 주제가 혼재되어 있어서 쉽게 트렌트를 파악하거나 통찰(insight)을 얻기가 힘들어지고 있다. 이와 같은 문제를 해결하기 위해 소셜미디어 데이터를 요약(summarization)하는 기술들이 연구되고 있다[2,3,4,5]. 문서로부터 주요한 문장을 인식하고 이 문장들을 통합하여 요약 제시하는 것이 일반적인 방법이다[2,3]. 그러나 최근에서는 시간에 따른 주제변화 및 사건을 파악하기 위해 트윗(tweet)의 시간대별 노출 변화추이를 순차적으로 요약하는 트윗 분석기술도 연구되고 있다[4].

다양한 소셜미디어 분석 기술들은 서로 독립적인 기술로서 개별 분석결과들은 소셜미디어 상의 단편적인 부분에 대한 통찰만을 제공한다. 앞서 언급된 요약기술들은 텍스트에 기반한 문서요약 및 시계열 상의 특정 사건 변화추이만을 요약하여 제시하고 있기 때문에 그래프에 대한 요약정보를 생성하는 것에는 한계가 있다. 따라서 개별 분석결과들은 데이터분석 전문가들에 의해서 수집 및 재분석되어 보고서로 요약/제공되어 의사결정을 지원하게 된다. 그러나 이러한 개별 분석결과들은 어쩌면 다른 시각에서 상호 보완적인 입장을 제시할 수 있지만, 이러한 상호보완적인 입장은 전문가의 통찰력에 의해서만 분석될 수 있다. 이로 인해 전문가에 의해서 제공된 보고서는 전문가의 지식수준 및 분석 성향에 따라 주관적인 경향이 있다. 또한, 전문가의 노동력에 의한 비용부담으로 인해 분석결과에 대한 공유 및 접근에 한계가 있다.

본 논문에서는 기존 소셜미디어 분석 기술의 고립성(isolation), 데이터 분석 전문가의 주관성(subjectivity), 비용부담에 의한 정보의 폐쇄성(closure of information)을 극복하기 위해서 개별 기술들의 분석 결과에 대한 상관성 분석을 통해 객관적이고 누구나 접근할 수 있는 소셜 빅데이터[1] 마이닝에 기반한 이슈 분석보고서 자동 생성 시스템에 대해서 소개한다. 제2장은 개별 소셜 빅데이터 분석기술과 관련 연구에 대해서 소개하고, 제3장은 이슈 분석보고서 자동 생성 방법과 알고리듬에 대해서 설명한다. 제4장에서는 이슈 분석보고서에 대한 사용자 인터페이스(UI)에 대해서 소개한다. 제5장에서는 시스템에 대한 평가방법과 결과에 대한 분석을 기술하고 제6장에서 결론 및 향후 연구 방향에 대해서 제시한다.

2. 관련 연구

소셜미디어 분석의 중요성과 더불어 많은 도구 개발 및 연구가 진행되었다. 앞서 언급한 바와 같이 크게 정보추출에 기반한 내용분석과 메타데이터의 구조적 정보에 기반한 네크워크 분석으로 나뉘어서 진행되었다.

정보추출에 기반한 내용분석과 관련된 대표적인 기술은 감성분석과 노출추이 분석이다.

감성분석은 시간대별 특정 개체와 연관된 감성의 변화를 분석하는 기술이다. 감성분석은 극성(polarity)에 기반하여 긍정(positive), 부정(negative) 및 중립(neutral)으로 범주(category)를 구분하고, 사용자의 텍스트 콘텐츠를 해당 범주로 분류하는 것이 전형적인 감성분석의 방법론이다. 최근에는 감성을 보다 세분화된 범주(fine-grained category)로 구분하여 분류하는 연구와 이를 위한 학습데이터 구축 방법론에 대한 연구도 활발히 진행되고 있다[6,7,8,9]

노출추이 분석은 시간대별 특정 개체의 빈도나 중요도의 변화를 분석하여 해당 개체(entity)나 사건 등이 이슈화되고 있는지 여부를 분석하는 기술이다. 노출추이 분석의 대상은 일반적으로 개체명인식기(NE recognizer)나 기정의된 사전(predefined dictionary)에 기반한 키워드(keyword)를 중심으로 분석하여 이슈개체 또는 이슈키워드로 결과를 제시하기도 하고, 관계추출(relation extraction)에 기반한 SPO[2] 트리플로 구성되는 이슈사건을 인식하기도 한다[10,11,12,13].

구조정보에 기반한 네트워크 분석(network analysis)은 대부분 트윗의 확산분석 및 영향력자 분석(influencer analysis) 기술에 집중되어 있다.

트윗의 확산 분석은 특정 트윗의 확산유형을 네트워크 그래프로 분석(network graph analysis)하는 기술이다. 트윗들의 확산형태를 분석하고 그래프의 유형을 분류함으로써, 해당 트윗이 로봇(robot)에 의해 생성된 스팸인지 여부를 알 수 있고, 특정 트윗의 초기 확산형태의 유형으로 향후 트윗의 확산양상을 예측을 할 수도 있다[14,15,16].

영향력자 분석은 트위터리안(twitterian)[3]들 중 특정 주제

1) 소셜미디어는 일반적으로 트위터, 페이스북과 같이 사용자 간의 네트워크가 구성되는 미디어를 이른다. 그러나, 본 시스템은 기존의 소셜미디어뿐만 아니라 블로그와 뉴스를 포함하여 분석을 수행하므로 '소셜 빅데이터'라는 용어를 사용한다.

2) SPO : Subject - Predicate - Object (주어 - 술어 - 목적어)

3) 트위터리안은 트위터를 사용하는 사람을 지칭하는 단어로, 한국에서만 사

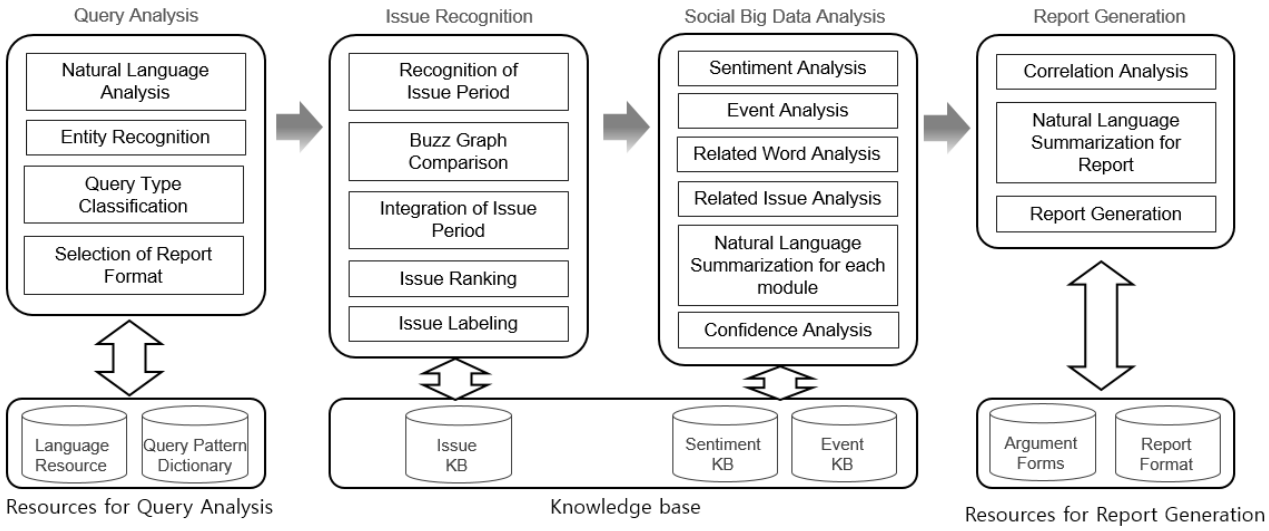


Fig. 1. System architecture for automatic generation of issue analysis report based on social big data mining

에 대해서 가장 영향력이 큰 사람을 순위화(ranking)하는 기술이다. 영향력자 분석은 트윗의 확산분석과 밀접하게 연관되며, 개별 트윗들의 확산분포를 분석하여 해당 트윗을 작성한 트위터리안이 영향력자인지 여부를 결정한다. 그리고, 확산된 트윗의 내용이 어떤 주제에 해당하는지 기정의된 주제로 분류한다[17].

문서요약기술은 다양한 텍스트 문서를 대상으로 어휘의 연관성 및 문장의 유사도를 기반으로 주요한 문장을 추출하고 통합하여 요약한다. [2]는 요약대상 문서 집합에서 어휘의 연관성(word association) 정도에 의존하여 문서요약을 수행하는 방법을 제시하고 있으며, [3]은 문장 유사도(sentence similarity)에 기반하여 블로그의 논평을 요약하는 2단계 문장 유사도 측정 방법을 소개하고 있다.

소셜미디어에 대한 요약기술로 시계열 상의 트렌드 토픽(trend topic) 변화를 요약하는 기술이 연구되고 있다[4]. [4]에서는 스트림(stream)과 의미(semantic) 기반의 접근법을 이용하여 트윗을 대상으로 시계열상의 토픽별로 순차적인 요약(sequential summarization)을 제공하는 기술을 제시하고 있다.

앞서 언급된 관련연구들은 소셜미디어의 다양한 측면(aspect)들 중 하나를 분석하거나 텍스트에 국한된 요약기술들로써, 개별 분석 결과만으로는 소셜미디어 전체의 양상이나 흐름을 파악하기에는 한계가 있고 텍스트가 아닌 그래프를 위한 텍스트 요약기술로는 활용될 수 없다. 이런 문제점을 해결하기 위해서 독립적인 다양한 소셜미디어 분석도구를 이용하여 데이터를 분석하고, 개별 분석 결과를 취합하고 상관성을 파악하여 소셜미디어 전체 양상과 흐름을 파악하는 빅데이터 분석 전문가들의 중요성이 대두되었다.

전문가에 의한 소셜미디어 분석는 개별 분석 도구의 분석오류를 필터링하고, 분석 결과에 대한 보충자료 수집 및 제시가 가능한 장점이 있는 반면, 다음과 같은 단점도 있다. 첫째, 전문가에 의한 분석은 많은 비용이 부담되어야 한다. 둘째, 전문가의 분석은 전문가의 지식수준 및 접근 가능한 정보의 양에 따라 결과가 상이할 수 있으며 주관적인 판단이 개입될 수 있다. 셋째, 분석에 많은 시간이 소요되므로, 시간적으로 시급성을 요구하는 경우 대응할 수 없다. 넷째, 많은 수의 개체나 키워드들에 대한 분석을 수행하기가 힘들다.

본 논문에서는 앞서 언급된 전문가에 의한 소셜미디어 분석의 문제점을 해소할 목적으로 개별 소셜 빅데이터 분석 결과에 대한 상관성을 자동으로 분석하고 자연어 요약과 다양한 그래프가 포함되는 이슈 분석보고서를 자동으로 생성하는 기술에 대해서 소개한다.

3. 이슈 분석보고서 자동 생성 시스템

그림 1은 본 논문에서 제안하는 소셜 빅데이터 마이닝 기반 이슈 분석보고서 자동 생성 시스템의 구성도를 보여주고 있다. 시스템은 질의분석, 이슈인식, 소셜 빅데이터 분석과 보고서 생성으로 구성이 되고, 다양한 언어자원, 지식베이스와 규칙사전 등이 필요하다.

3.1 질의분석

질의분석은 사용자가 분석을 요구하는 개체와 분석대상 기술[4]을 인식하는 모듈이다. 본 시스템에서는 표 1과 같이 두 종류의 구분기준에 기반하여 네 가지의 질의유형으로

용되는 단어이고, 외국에서는 트위터러(twitterer) 또는 트위플(tweeple)로 사용됨.

4) 3.3절에서 언급되는 소셜웹 분석 기술들 중, 어떤 기술로 분석한 결과를 제시해야 하는지 여부

Table 1. Query types and examples

Query Type	Example Query	Entity	Result
One entity - Designated module	삼성전자의 버즈추이는? (Trend of Samsung's buzz)	삼성전자(Samsung)	Buzz Analysis Graph
One entity - Report	삼성전자 (Samsung)	삼성전자(Samsung)	Analysis Report
Two entity - Designated module	애플과 삼성전자에 대한 감성변화 추이를 알려줘 (Let me know the trend change of sentiment between Apple and Samsung)	애플과 삼성전자 (Apple and Samsung)	Comparison of Sentiment Regression Graph
Two entity - Report	애플과 삼성전자 비교 (Comparison between Apple and Samsung)	애플과 삼성전자 (Apple and Samsung)	Comparative Analysis Report

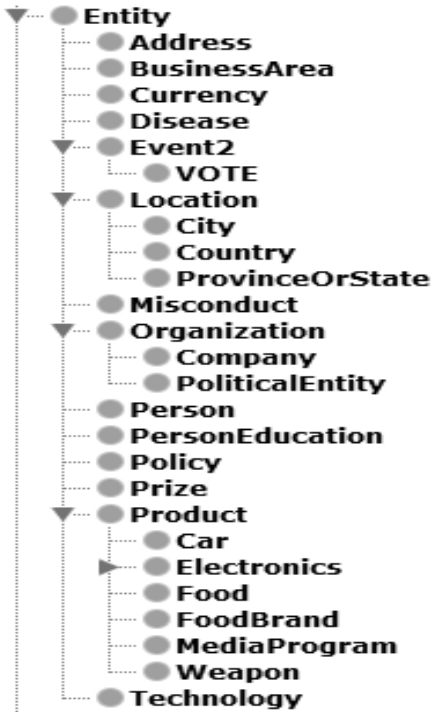


Fig. 2. Entities of Zodiac ontology

분류한다. 구분기준은 개체의 수와 분석대상 기술로 구분된다. 개체의 수는 하나 또는 둘로 구분되며, 개체가 둘인 경우 개체들 간의 비교분석을 수행하게 된다. 분석대상 기술은 그림 1의 소셜 빅데이터 분석에서 언급된 다양한 기술 중, 특정 기술을 언급하는 단서가 질의에 포함된 경우와 그렇지 않은 경우로 분류된다. 예를 들어, "애플과 삼성전자에 대한 감성변화 추이를 알려줘"라는 질의에서는 두 개체에 대한 감성변화 분석 결과만을 제시하면 된다. 반면, "삼성전자"라는 질의만 입력될 경우, 분석대상 기술을 언급하는 단서 없이 하나의 개체만 언급되었기 때문에 '삼성전자'에 대한 이슈 분석보고서를 제공한다.

질의분석에서는 형태소분석과 개체명인식 결과에 기반하여 앞서 언급된 질의유형으로 질의를 분류한다. 분석대상 기술을 인식하기 위해서, 다양한 질의를 수집하고 수작업으로 분석하여 구축한 단서사전(clue dictionary)과 어휘의미규칙(lexico-semantic rules)을 이용하였다.

보고서 서식(report format)은 질의에서 인식된 개체의 유형에 의존적이다. 다양한 소셜 빅데이터 분석 기술들은 분석대상이 되는 개체의 유형에 따라 그 성능의 차가 두드러진다. 예를 들어, 감성원인 분석 기술과 속성감성 분석 기술은 주로 상품과 관련된 개체에 적합한 기술이므로, 질의에서 인식된 개체가 상품일 경우 더 좋은 결과를 제시한다. 이처럼 질의에서 인식된 개체의 유형별로 적합한 분석 기술들을 보고서 서식에 할당하고 정의하여 자동 생성된 보고서의 품질을 보장하고자 하였다. 본 시스템에서 분류하는 개체의 유형은 인물, 기관, 상품, 정책이다. 개체의 유형 분류를 위해서 네 가지 유형을 포함하는 그림 2의 조디악온톨로지(Zodiac ontology)[5]를 구축하고 인스턴스화(instantiation)기술을 수행하였다.

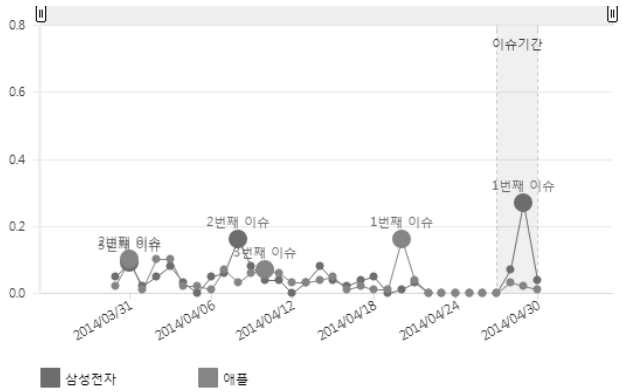
3.2 이슈인식

이슈는 시계열상의 특정 시점에 급속히 언급되는 개체를 의미한다. 가장 기본적인 자질은 개체의 빈도정보이다. 그러나 개체의 인지도에 따라 평균빈도의 편차가 심하기 때문에 단순 빈도만으로 이슈를 인식할 경우, 인지도가 높아 자주 언급되는 개체들만 이슈로 추출되는 문제가 발생한다. 본 논문에서는 인지도에 따른 왜곡 문제를 완화하기 위해서 다음과 같이 5개의 주요한 이슈속성을 정의하고, 이를 기반으로 날짜별 개체의 이슈정도를 계산한다.

- 신규성 : 시계열상의 빈도차이 계산을 통한 신규성 평가
- 중요성 : 개체의 중요도 평가
- 파급력 : 유입량/안정성/변동성 등 파급력 평가
- 신뢰성 : 이슈의 출처에 대한 신뢰도 평가
- 관심도 : 감성도, 댓글, 리트윗 수에 대한 평가

이슈기간은 이슈정도가 높은 날짜를 기준으로 선정된다. 두 개체를 비교할 경우, 개별 개체의 이슈정도를 날짜별로 통합하여 이슈기간을 선정한다. 인식된 이슈기간은 순위화되고, 이슈원인을 추측할 수 있도록 인식된 이슈기간에 개

5) Zodiac 온톨로지는 ETRI에서 소셜웹 분석용 이벤트 추출을 위해서 정의한 온톨로지임.



순위	이슈명	이슈기간
1	[뉴스] 매출 실적 영업이익 삼성전자 스마트 / 애플 삼성전자 소송 재판 [트위터] 삼성전자 공개 스마트폰	2014.04.27 ~ 2014.05.03 분석
2	[뉴스] 갤럭시 출시 삼성전자 영업 [블로그] 삼성전자 갤럭시 시장 기업 발표 / 애플 제품 폰 사용 삼성그룹 [트위터] 삼성전자 영업 발표 이익	2014.04.06 ~ 2014.04.12 분석
3	[뉴스] 삼성전자 등기이사 회장 사장 / 미국 연봉 애플 사장 삼성 [블로그] 삼성전자 연봉 회장 기업 / 애플 제품 삼성 시작 [트위터] 삼성전자 연봉 공개 회장 삼성그룹 / 애플 삼성그룹 시작	2014.03.29 ~ 2014.04.04 분석

상세분석 보기 >

Fig. 3. Comparison of issue graph between Samsung and Apple during april 2014

체와 가장 많이 공기한 어휘를 이슈명으로 제시한다. 그림 3은 2014년 4월 동안 '삼성전자'와 '애플'의 이슈그래프를 보여주고 있다.

3.3 소셜 빅데이터 분석

본 시스템에서 분석하는 소셜 빅데이터 분석 기술들은 다음과 같다.

- 감성시계열분석 : 시계열 상에 감성의 긍/부정 변화를 분석하는 기술
- 세부감성분석 : 기정의한 20개의 세부 감성분류[6]에 기반하여 감성을 분류하는 기술
- 속성감성분석 : 상품의 다양한 속성별로 긍/부정의 변화를 분석하는 기술
- 감성원인분석 : 긍/부정 문장을 대상으로 클러스터링(clustering)을 수행하여 긍/부정의 주요원인별로 그룹핑(grouping)하고 레이블링(labelling)하는 기술
- 이슈이벤트분석 : 관계추출 기술에 기반하여 SPO 트리플을 추출하고 기정의된 이벤트 템플릿을 생성하는 기술

```
  BODY+1 = <bullet1>"<bold><EVENT></bold>"
이벤트가 <bold>상승</bold> 중입니다.</bullet1>
  BODY-1 = <bullet1>"<bold><EVENT></bold>"
이벤트가 <bold>하락</bold> 중입니다.</bullet1>
  BODY-0 = <bullet1>"<bold><EVENT></bold>"
이벤트의 상태변화가 적습니다.</bullet1>
...
```

Fig. 4. An example of template for natural language summarization [Issue event analysis]

- 영향력자분석 : 특정 개체에 대해서 소셜미디어 상에서 영향력이 큰 사용자를 분석하는 기술
- 연관/경쟁키워드분석 : 입력된 키워드와 연관관계나 경쟁관계가 있는 키워드를 분석하는 기술
- 연관/경쟁이슈분석 : 입력된 키워드와 연관관계나 경쟁관계가 있는 이슈를 분석하는 기술

개별 분석 기술에서는 분석결과에 대한 신뢰도를 정규화된 형태로 제공한다. 이는 향후 보고서 생성 시에 결과를 포함할지 여부를 판단하는 기준이 된다. 또한 개별 분석 결과에 대한 자연어요약 정보를 제공한다. 자연어요약은 사용자들에게 제공되는 분석 그래프를 요약하는 것으로서, 일반 사용자의 그래프 이해력을 증진시키는 역할을 한다. 그리고 개별 기술에서 제공되는 자연어 요약정보는 전체 보고서에 대한 자연어 요약에서 활용된다. 자연어 요약은 기정의된 템플릿(template)의 각 슬롯(slot)에 값을 채우는 방식인 템플릿 기반 자연어 생성기술을 이용하였다. 그림 4는 이슈이벤트분석의 자연어 요약을 생성하기 위한 템플릿의 예제이다. 'BODY±숫자(0/1)'은 자연어문장의 분류코드이고, 대문자 영문('EVENT')이 슬롯에 해당하는 부분이다.

3.4 상관성 분석 및 보고서 생성

상관성 분석은 앞 절에서 언급된 다양한 소셜 빅데이터 분석 기술들 중, 감성분석과 관련된 기술들만을 대상으로 하였다. 본 시스템에서는 다음과 같이 세 가지 유형의 상관성분석을 수행하였다.

- 감성시계열과 감성원인 상관성 분석
- 감성시계열과 속성감성 상관성 분석
- 세부감성과 감성원인 상관성 분석

특정 개체에 대해서 시계열상에 감성변화를 보는 것만으로도 가치가 있지만, 보다 정확한 통찰력을 얻기 위해서는 감성변화의 원인이 무엇인지 파악하는 것이 중요하다. 이를 파악하기 위한 것이 감성시계열과 감성원인 상관성 분석 기술이다. 이슈인식기술로 인식된 이슈기간을 중심으로 이슈

6) 긍정세부감성 : 자신감, 감동, 감사, 기대감, 좋아함, 기쁨, 안심, 신뢰, 선의
부정세부감성 : 두려움, 화남, 싫어함, 슬픔, 실망, 수치심, 곤란, 미안함, 부러움, 반대, 의심

기간 이전의 감성변화와 이슈기간의 감성변화를 비교분석하고 변화폭이 큰 감성에 대한 감성원인을 제시하는 것이다. 이슈기간 이전에 분석되지 않았던 감성원인이 이슈기간에 인식된 경우에 사용자에게 제시함으로써 감성변화의 원인을 쉽게 파악할 수 있다.

상품개체는 그림 2의 'Product' 하위 개념들 별로 주요한 속성을 정의할 수 있다. 예를 들어, 스마트기기는 A/S, 속도, 가격, 디자인, 배터리, 스피커, 버그, 카메라, 디스플레이과 터치감으로 속성을 정의한다. 상품개체에 대한 감성시계열 분석에서 긍/부정 변화에 영향을 미치는 상품의 속성들을 파악하기 위해서, 감성시계열과 속성감성 상관성을 분석하였다. 이 결과는 상품의 품질개선 및 홍보를 위한 의사결정에 중요한 자료가 된다. 본 시스템에서는 감성속성을 예에서 언급한 스마크기기 외에 영화와 화장품에 대해서도 정의하였다.[7] 속성분류는 단서어휘 사전에 기반한 규칙으로 처리하였다.

감성은 긍/부정별로 세부감성을 계층적으로 분류할 수 있다. 앞 절에서 언급한 바와 같이 긍정에 9개의 세부감성이 부정에 11개의 세부감성이 있다. 감성시계열과 감성원인 상관성 분석보다 더 구체적인 통찰력을 얻기 위해서는 긍/부정의 세부감성별로 감성의 원인을 파악하는 것이 필요하다. 이를 위해서 세부감성과 감성원인 상관성 분석을 수행하였다. 즉, 감성분류의 계층을 한 단계 더 깊이 들어가서 감성원인을 분석함으로써 의사결정에 도움이 될 수 있는 미묘한 감성차이를 파악할 수 있다.

보고서생성은 질문분석에서 선택된 보고서 서식과 소셜 빅데이터 분석에서 제공하는 개별 분석결과의 신뢰도 값에 기반하여 동적으로 구성된다. 분석결과의 신뢰도 값이 지정된 임계값(threshold)을 만족하지 못하는 경우, 선택된 보고서의 서식에서 제외된다. 지정된 개수보다 많은 분석결과가 신뢰도 값에 의해서 제외되면, 개별 분석결과의 임계값을 조정한다. 이렇게 함으로써, 일정 개수 이상의 결과가 보고서로 생성될 수 있도록 하였다. 보고서에는 전체 분석결과에 대한 자연어 요약이 포함된다. 자연어 요약은 개별 소셜 빅데이터 분석결과에서 제시한 요약문과 자연어 생성 템플릿의 슬롯 값을 채울 때 사용된 변수 값을 이용한다.

4. 시스템 UI

시스템 UI는 그림 5와 같다. 자연어 질의를 입력받을 수 있는 입력창(①)이 최상단에 위치한다. 자연어 입력창 우측에는 기준 날짜를 지정할 수 있도록 달력을 제공한다. 주요 이슈 키워드(②)는 지정된 날짜에 가장 이슈가 된 개체 리스트를 순위화하여 제시한다. 질의 입력창(①)에서 사용자가

Fig. 5. User interface of system [Issue QA in Social Wisdom]

자연어를 입력하고, 분석을 수행하면 분석된 결과가 분석설정 창(③)에 제시된다. 자연어 질의를 분석하여 질의에서 언급된 주요한 개체를 인식하여 인식개체와 비교개체로 제시한다. 또한 분석 대상 미디어(뉴스, 블로그, 트위터) 선정 및 분석기간(1달, 2달, 3달)을 선택할 수 있다. 분석설정 창(③)에서 분석하기 버튼을 클릭하면, 지정한 기간과 개체들에 대한 이슈분석을 수행한다. 분석된 이슈 그래프와 이슈기간은 이슈분석 창(④)에서 제시된다. 이슈분석 창(④)에서는 인식된 이슈기간의 3위까지 제시되고, 사용자는 원하는 이슈기간을 클릭함으로써 정밀한 소셜 빅데이터 분석을 요청할 수 있다. 분석이 완료되면, 소셜 빅데이터 분석 결과창(⑤)에 전체분석에 대한 자연어 요약정보, 3.3절의 소셜 빅데이터 분석 결과와 3.4절의 상관성 분석 결과가 제시된다.

4.1 상관성 분석 결과 UI

3.4절에서 세 가지 유형의 상관성 분석에 대해서 소개하였다. 세 유형의 분석결과를 사용자들이 쉽게 이해할 수 있도록 시각화하여 결과를 제시하는 것은 중요하다.

감성시계열과 감성원인 상관성 분석은 그림 6과 같이 결과가 제시된다. 감성시계열상에서 이슈기간 이전과 이슈기간의 감성변화를 분석하여 상승한 감성(긍정 or 부정)의 원인을 제시한다. 또한 분석된 감성원인에 대한 자연어요약정보와 대상 문장을 함께 제시한다.

감성시계열과 속성감성 상관성 분석은 그림 7과 같다. 이

7) 영화 속성 - 영상, 사운드, 배우, 소재, 스토리, 연출, 결말, 분위기, 독창성
화장품 속성 - 가격, 색, 향, 사용감, 발림성, 세정력, 수분감, 유분감, 양, 디자인, 지속력, 적합성

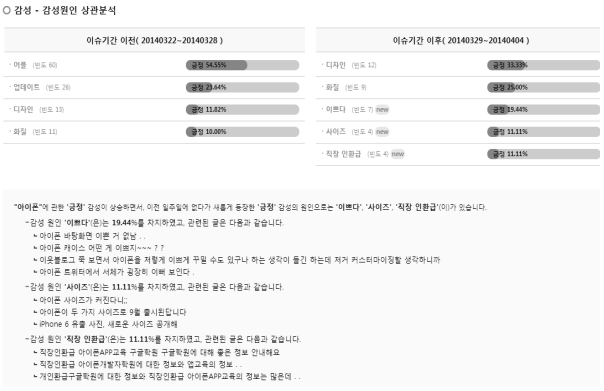


Fig. 6. Correlation analysis between sentiment and cause on time series [query : iPhone]

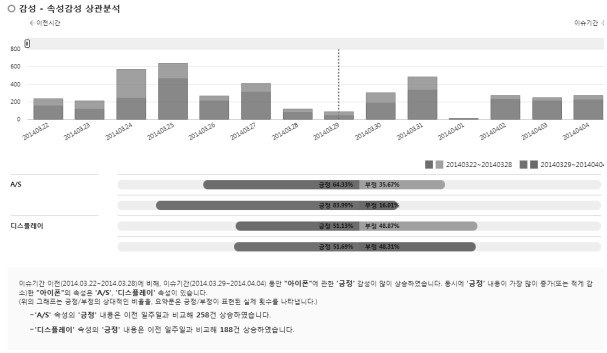


Fig. 7. Correlation analysis between sentiment and aspect on time series [query : iPhone]

Fig. 8. Correlation analysis between fine-grained sentiment and cause [query : iPhone]

슈기간 이전과 이슈기간 동안의 감성분석의 변화를 시계열상에 제시하고, 감성변화에 영향을 미친 주요한 속성들의 감성변화를 함께 제시한다. 분석 대상 개체가 상품이면, 상품의 호불호에 영향을 미치는 주요한 속성을 쉽게 파악할 수 있어서 마케팅 및 상품의 품질개선에 대한 통찰력을 제공할 수 있다.

그림 8은 세부감성과 감성원인의 상관성 분석 결과이다. 20개의 세부감성 중 긍정과 부정에 가장 큰 영향을 미친 세부감성을 분석하고, 해당 세부감성의 원인을 분석하여 제시한다.

Table 2. The distribution of evaluation sentences for fine-grained sentimental analysis

	News	Blog	Tweet	Total
Person	646	1,053	2,239	3,941
Organization	555	2,005	1,560	4,120
Product	99	847	499	1,445
Policy	209	297	229	735
Total	1,512	42,02	4,527	

Table 3. The distribution of evaluation sentences for sentimental aspect analysis

	News	Blog	Twitter	Total
Product	121	720	200	1,041

Table 4. The distribution of evaluation data for sentimental cause analysis

	# of sentences	# of clusters	# of entities
Person	224,153	26	7
Organization	19,807	41	13
Product	3,833	44	6
Policy	4,878	392	15
Total	252,671	503	41

세 유형의 상관성분석 결과는 분석 대상 개체의 감성변화를 다양한 측면(facet)으로 분석하여 직관적으로 제시함으로써 상품과 브랜드의 마케팅 전략 등의 의사결정을 지원할 수 있는 유용한 정보를 제공한다.

5. 평 가

평가는 상관성 분석의 핵심기술들에 대한 개별 평가를 수행하고, 이슈 분석 보고서에 대한 평가를 수행하였다.

세부감성을 위한 평가데이터는 표 2와 같이 세부분류로 인물, 기관, 상품, 정책으로 분류하고 각 분류별로 뉴스, 블로그, 트위터에서 문장을 수집하여 평가데이터를 구축하였다. 표 3은 속성감성을 위한 평가데이터의 분포이다. 속성감성은 상품만을 대상으로 하는 분석기술이므로 상품분류에 속하는 문장만을 대상으로 선정하였다. 표 4는 감성원인분석 평가를 위한 평가데이터의 분포정보이다.

세부감성과 속성감성의 평가는 크게 네 가지 측면을 평가한다. 첫째, 해당 문장이 긍/부정에 대해서 올바르게 분류하였는지 여부이다. 둘째, 긍/부정의 대상(target)이 되는 개체

를 올바르게 인식했는지 여부이다. 셋째, 세부감성분석에서 기정의된 20개의 세부감성으로 잘 분류했는지 여부이다. 마지막으로 속성감성분석에서 속성별로 잘 분류했는지 여부이다. 평가척도는 정확률(precision)이다.

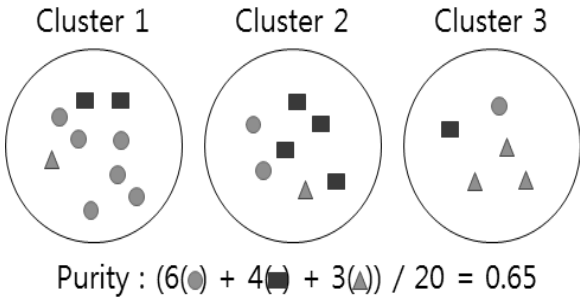


Fig. 9. An example of how to compute Purity

감성원인분석의 평가척도는 클러스터링에서 정확률을 이용하는 외부기준의 평가척도인 purity를 사용한다[18].

$$Precision(C_i, L_j) = \frac{|C_i \cap L_j|}{|C_i|} \quad (1)$$

C는 평가할 클러스터들의 집합이고, L은 클래스들의 집합이다. 정확률은 클러스터의 문장들 중에 동일한 클래스에서 온 문장의 비율로서, 모든 문장이 동일한 클래스에 포함되는 경우 1 값을 갖는다.

$$Purity = \sum_i \frac{|C_i|}{n} max\ Precision(C_i, L_j) \quad (2)$$

n은 총 문장의 수이다. 그림 9는 Purity를 계산하는 방법에 대한 예제이다. 클러스터 1에는 ●가 6개, 클러스터 2에서는 ■가 4개, 클러스터 3에서는 ▲가 3개로 개별 클러스터에서 최대값(max)을 가지는 객체들이다. 따라서 Purity는 13/20으로 0.65의 값을 가진다.

개별 소셜 빅데이터 분석 기술에 대한 평가는 앞서 언급한 바와 같이 평가데이터를 구축하여 다양한 방법으로 진행될 수 있고, 객관적인 평가데이터를 구축하는 것도 상대적으로 쉽다. 그러나 특정 개체에 대한 이슈 분석보고서의 유용성을 평가하기 위한 평가데이터를 구축하는 것은 어려움이 많고 객관성 확보도 쉽지 않다. 이와 같은 문제점으로 인해서 본 시스템에 대한 평가는 리커트척도(Likert scale)를 이용하였으며, 평가를 위한 질의는 2013년 10월 1일부터 10월 31일 한 달간, NHN의 실시간 검색 키워드 순위와 시스템에서 추출한 개체들의 빈도를 고려하여 389개를 선정하였다. 선정된 질의는 먼저 기업과 국가(정부) 관련 카테고리(category)로 분류하였고, 세부분류로 인물, 기관, 상품, 정책으로 분류하여 평가를 진행하였다. 표 5는 평가에 사용된 질의의 분포이다. 표 5에서 괄호 내의 숫자는 질의들 중, 두 개체의 비교분석을 요청하는 질의의 수이다. 예를 들어, '아이폰5와 갤럭시S4'와 같은 질의로서, 질의 내에 두 개체가 포함된 경우를 의미한다. 평가는 외부 전문가 2명[8]이 진행하였으며, 리커트척도는 다음과 같다.

Table 5. The distribution of evaluation queries for issue analysis report

	Person	Organization	Product	Policy
Company	116 (4)	127 (14)	60 (6)	4 (0)
Government	24 (10)	13 (0)	0 (0)	56 (6)

(): the number of queries for comparison between two entities.

- 질의를 분석하여 생성된 보고서의 품질을 평가해주세요.

① 결과에 대해서 신뢰할 수 있으며 만족스럽다.
② 결과에 신뢰할 수 없는 정보가 일부 있으나, 유용한 정보를 제공한다.
③ 결과에 대해서 참고 자료로 활용할 수 있는 정도의 정보를 제공한다.
④ 결과에 대한 신뢰성에 의문이 있으며, 보충 정보가 요구된다.
⑤ 질의와 연관성이 없는 결과를 제시하여, 전혀 신뢰할 수 없다. (결과를 제시하지 못한 경우도 포함)

평가는 시스템에서 제시하는 소셜 빅데이터 분석의 개별 결과들에 대해서 리커트척도로 평가를 하고, 이를 매크로 평균(macro-average)으로 계산하였다.

표 6과 표 7은 세부감성분석에 대한 평가결과이다. 표 6은 세부분류별 평가결과를 정리한 것이고, 표 7은 평가문장의 출처별로 평가결과를 정리한 것이다. 감성대상(target) 인식성능이 0.65로 긍/부정과 세부감성 인식 성능보다 낮았다. 일반적으로 문장에서는 복수 개의 개체와 함께 감성정보가 제시된다. 이때, 해당 감성에 대한 대상이 어떤 개체인지를 인식하는 것은 구문관계에 기반한 인식문제로서 구문분석의 성능에 의존적이다. 이로 인해 상대적으로 성능이 낮은 것으로 분석된다. 평가문장의 출처별 평가결과는 긍/부정과 세부감성 인식에서는 뉴스 문장에서 가장 성능이 낮았다. 반면, 감성대상 인식은 제일 성능이 높았다. 뉴스에서는 다양한 개체에 대한 객관적인 사실을 중심으로 기술된 문장이 대부분이다. 즉, 특정 개체에 대한 감성적인 표현이 적다. 따라서 긍/부정과 세부감성 인식 성능이 저조한 것으로 분석된다. 반면, 블로그와 트윗은 비문(非文)이 많지만, 뉴스는 올바른 문장인 경우가 비교적 많다. 이는 구문분석의 성능 향상에 도움이 된다. 이로 인해 뉴스에서 감성대상의 성능이 높은 것으로 분석된다.

8) 전산언어학 전공자들로 빅데이터 분석 경력이 있는 자.
평가자 A : 전산언어학 석사 수료 후 경력 13년
평가자 B : 전산언어학 석사 수료 후 경력 9년

Table 6. The evaluation results of fine-grained sentimental analysis classified by category

	Positive / negative	Fine-grained category	Sentimental target
Person	0.85	0.79	0.76
Organization	0.85	0.82	0.55
Product	0.87	0.85	0.68
Policy	0.77	0.72	0.63
Total	0.84	0.81	0.65

Table 7. The evaluation results of fine-grained sentimental analysis classified by source

	Positive / negative	Fine-grained category	Sentimental target
News	0.59	0.52	0.8
Blog	0.86	0.82	0.56
Tweet	0.92	0.88	0.69

Table 8. The evaluation results of sentimental aspect analysis classified by source

	Positive / negative	Sentimental aspect category	Sentimental target
News	0.76	0.81	0.83
Blog	0.76	0.87	0.68
Tweet	0.82	0.83	0.8
Total	0.85	0.85	0.72

Table 9. The evaluation results of issue analysis report by each distribution
(개별분포 별 이슈 분석보고서의 평가결과)

	Person	Organization	Product	Policy
Company	1.43	1.45	2.04	3.25
Government	1.72	1.34	n/a	1.98

표 8은 속성감성분석에 대한 평가결과이다. 감성속성분류의 성능은 0.85이고, 블로그에서 가장 좋은 성능을 보였다. 속성감성은 상품의 속성별 감성분석을 목적으로 한 것이다. 따라서 파워 블로거(power blogger)의 상품비교 포스트(post)가 많은 블로그에서의 성능이 우수한 것으로 분석된다.

표 9는 개별 분포집합별로 이슈 분석보고서의 리커트척도 평균을 정리한 표이다. 사람과 기관으로 분류되는 개체에 대한 결과가 상품과 정책에 대한 결과보다 우수하다. 이는 크게 두 가지 원인으로 분석된다. 첫째, 개체인식 성능과 연관된 것으로 일반적으로 개체명 인식에서 사람, 기관명과 상품명이 정책명보다 상대적으로 인식 정확률이 높기 때문이다. 평가 질의에 포함된 개체에 대한 인식 정확률을 평가한 결과, 사람, 기관, 상품, 정책 각각의 정확률이 95%, 97.14%, 95%, 86.67%이었다. 둘째, 개체유형별 콘텐츠 특성이 상이하다는 것이다. 일반적으로 상품에 대한 콘텐츠는 복수개의 상품을 비교하는 경우가 많다. 이런 경우, 감성분석에서 감성에 대한 대상(target)을 정확히 인식하는 것이 중요하다. 예를 들어, "갤럭시S가 아이폰보다 화면이 크고 좋다"라는 문장에서 '화면'이라는 속성(aspect)에 대해서 '아이폰'이 아닌 '갤럭시S'가 좋다는 것을 인식해야 한다. 즉, 한 문장에서 복수 개의 감성분석 대상이 출현하는 콘텐츠가 상품관련 콘텐츠가 상대적으로 많아서, 평가점수가 저조한 것으로 분석된다.

Table 10. The evaluation results of correlation analysis

	CSA	CSC	CFC
# of queries including report	25	60	58
average of Likert score	1.96	1.42	1.21

CSA : Correlation analysis between sentiment and aspect on time series (감성시계열과 속성감성의 상관성 분석)
CSC : Correlation analysis between sentiment and cause on time series (감성시계열과 감성원인의 상관성 분석)
CFC : Correlation analysis between fine-grained sentiment and cause (세부감성과 감성원인의 상관성 분석)

표 10은 본 시스템에서 분석한 세 가지 유형의 상관성 분석에 대한 평가 결과이다. 상관성 분석결과는 개체유형과 결과의 신뢰도 값에 의해서 필터링 될 수 있다. 이로 인해, 평가질의 중 25개 질의만이 감성시계열과 속성감성의 상관성 분석결과를 제시하였고, 감성시계열과 감성원인의 상관성 분석이 가장 많은 60개의 질의에서 상관성 분석결과를 제시하였다.

상관성 분석에 대한 평가자의 평가점수는 감성시계열과 속성감성의 상관성 분석이 다른 두 유형보다 상대적으로 낮았다. 이는 속성감성분석의 대상이 기업도메인의 상품유형으로 제한되어 있기 때문에 보고서에 결과를 제시한 질문의 수도 25개로 적고, 평가자의 평가점수도 상대적으로 저조한 것으로 분석된다.

6. 결론 및 향후연구

본 논문에서 개별 소셜 빅데이터 분석 결과를 통합하고 자동으로 상관성을 분석하여 이슈 분석보고서를 생성하는 시스템에 대해서 소개하였다. 본 시스템은 다음과 같은 특징을 가지고 있다.

- 입력 질의 분석을 통한 대상 개체의 시계열 분석과 이슈기간 자동 분석
- 입력 질의의 유형에 따른 보고서 서식 변경 및 분석 신뢰도에 기반한 동적 보고서 생성
- 개별 소셜 빅데이터 분석 결과들의 상관성 분석을 통한 통찰력 제공

- 템플릿에 기반한 자연어 요약 생성
- 개체에 대한 소셜 빅데이터 분석에 기반한 이슈 분석보고서 자동 생성

위와 같은 특징은 개별 소셜 빅데이터 분석 기술의 고립성과 기존 이슈 분석보고서의 다음과 같은 단점을 보완하고 있다.

- 개별 소셜미디어 분석의 상관성은 전문가의 주관에 의해서 진행되므로 객관성 확보가 어려움.
- 특정 개체에 대한 이슈 분석보고서는 전문가에 의해서 개별 소셜미디어 분석 결과를 취합하고 분석하여 생성되기 때문에 많은 시간과 노력이 요구됨.
- 기존 이슈 분석보고서는 데이터 분석 전문가의 시간과 노력에 따른 고비용 문제로 대중화에 한계가 있음.

기존 소셜미디어 분석보고서가 가지고 있는 단점을 보완하기 위해서 설계 및 구현된 본 시스템의 성능은 리커트 척도에 기반한 평가에서 1과 2 사이의 평균 점수를 받았다. 이는 결과가 비교적 신뢰할 수 있고, 유용함을 의미한다. 즉, 기존 소셜미디어 분석보고서 작성의 문제점을 충분히 보완하여 향후 소셜 빅데이터 분석의 대중화에 기여할 수 있는 기술임을 확인할 수 있었다.

향후연구 방향은 개별 소셜 빅데이터의 상관성 분석을 확대하는 것이다. 본 논문에서는 감성과 관련된 세 유형의 상관성만을 분석하고 있지만, 다양한 모듈 간의 상관성 분석으로의 확대는 소셜 빅데이터 분석을 통한 더욱 많은 통찰력을 얻을 수 있는 기회를 제공할 것이다. 또한, 템플릿에 기반한 자연어 요약은 템플릿의 구조에 너무 의존적인 한계가 있다. 다양한 분석 결과를 다양하게 요약하여 사용자에게 직관적인 정보를 제공할 수 있도록 자연어 요약에 대한 연구가 진행되어야 할 것이다.

References

[1] Jeong Heo, Pum-Mo Ryu, Yoon-Jae Choi, Hyun-Ki Kim and Cheol-Young Ock, “An Issue Event Search System based on Big Data for Decision Supporting: Social Wisdom”, Journal of KIISE: Software and Application, Vol.40, No.7, 2013.07.

[2] Oskar Gross, Antoine Doucet and Hannu Toivonen, “Document Summarization Based on Word Associations”, Proceedings of the 37th international ACM SIGIR conference on Research and Development in Information Retrieval. ACM, 2014.

[3] Hongjie Li, Lifu Huang, Qifeng Fan and Lian’en Huang, “Comments-Oriented Summarization in Blogsphere Using a Two-Stage Sentence Similarity Measure”, In Web-Age Information Management. Springer International Publishing, pp.480-483, 2014.

[4] Dehong Gao, Wenjie Li, Xiaoyan Cai, Renxian Zhang, and You Ouyang, “Sequential Summarization: A Full View of Twitter Trending Topics”, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), Vol.22, No.2, pp.293-302, 2014.

[5] Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhong Su, and Juanzi Li, “Social Context Summarization”, In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, pp.255-264, 2011.

[6] Yo-Han Jo, Hyo-Jung Oh, Chung-Hee Lee, and Hyun-Ki Kim, “Fine-grained Sentiment Lexicon Construction via Semi-supervised Learning”, 25th Annual Conference on HCLT, 2013.

[7] Moon-Soo Chang, “Empirical Sentiment Classification Using Psychological Emotions and Social Web Data”, Journal of Korean Institute of Intelligent Systems, Vol.22, No.5, pp.563-569, 2012.

[8] Yong-Min Park, Su-Jeong Kwak, Daniel Lee, Bo-Gyum Kim, Yeo-Chan Yoon, and Jae-Sung Lee, “Construction of Korean Test Collection for Social Media Text Sentiment Analysis”, Proceeding of the KIISE Fall Conference, Vol.39, No.2, pp.118-120, 2012.

[9] Kong-Joo Lee, Jee-Eun Kim, and Bo-Hyun Yun, “Extracting Multiword Sentiment Expressions by Using a Domain-Specific Corpus and a Seed Lexicon,” ETRI Journal, Vol.35, No.5, pp.838-848. 2013.

[10] Pum-Mo Ryu, Hyun-Jin Kim, Hyun-Ki Kim, and Sang-Kyu Park, “Social Media Issue Detection & Monitoring based on Deep Language Analysis Techniques,” Journal of Computing Science and Engineering, Vol.30, No.6, pp.47-58, 2012.

[11] Chung-Hee Lee, Hyun-Jin Kim, Hyo-Jung Oh, Jeong Hur, Pum-Mo Ryu, and Hyun-Ki Kim, “Social WISDOM: An Issue Detection/Monitoring System”, Proceedings of the Korea Information Processing Society Conference, Vol.19, No.2, 2012.

[12] Jeong Heo, Pum-Mo Ryu, Yoon-Jae Choi, and Hyun-Ki Kim, “Event Template Extraction for the Decision Support based on Social Media”, 24th Annual Conference on HCLT, 2012.

[13] Yoonjae Choi, Pum-Mo Ryu, Hyunki Kim, and Changki Lee, “Extracting Events from Web Documents for Social Media Monitoring using Structured SVM”, IEICE, Vol.E96-D, No. 6, 2013.

[14] Min-Chul Yang, Jung-Tae Lee, and Hae-Chang Rim, “Using Link Analysis to Discover Interesting Message Spread Across Twitter”, Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing. Association for Computational Linguistics, pp.15-19, 2012.

[15] 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 the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012.

[16] Yong-Jin Bae, Pum-Mo Ryu, and Hyun-Ki Kim, "Predicting Popular Tweets based on Similarity Analysis from Collaborative Features", Journal of KIISE: Software and Application, Vol.40, No.7, pp.405-416, 2013.

[17] Eytan Barkshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts, "Everyone's an Influencer : Quantifying Influence on Twitter", Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

[18] Kyeongtaek, Kim, "F_n-Measure: An External Cluster Evaluation Measure", Journal of Society of Korea Industrial and Systems Engineering, Vol.35, No.4, pp.244-248, 2012.

허 정

e-mail : jeonghur@etri.re.kr
1999년 울산대학교 전자계산학과(학사)
2001년 울산대학교 전자계산학과(석사)
2001년~현 재 한국전자통신연구원 선임연구원
2013년~현 재 울산대학교 정보통신공학전공 박사과정
관심분야 : 자연어처리, 정보검색, 텍스트마이닝, 빅데이터처리

이 충 희

e-mail : forever@etri.re.kr
1996년 한양대학교 전자계산학과(학사)
2001년 연세대학교 컴퓨터과학과(석사)
2001년 충북대학교 컴퓨터공학과(박사)
2012년~현 재 한국전자통신연구원 선임연구원
관심분야 : 자연어처리, 정보추출, 정보검색, 질의응답

오 효 정

e-mail : ohj@etri.re.kr
2000년 충남대학교 컴퓨터과학과(석사)
2008년 한국과학기술원 컴퓨터공학과(박사)
2000년~현 재 한국전자통신연구원 책임연구원
관심분야 : 정보검색, 질의응답, 빅데이터 정보처리, 소셜웹마이닝

윤 여 찬

e-mail : ycyoon@etri.re.kr
2004년 고려대학교 컴퓨터학과(학사)
2007년 고려대학교 컴퓨터학과(석사)
2007년~현 재 한국전자통신연구원 선임연구원
관심분야 : Text Analysis

김 현 기

e-mail : hkk@etri.re.kr
1995년 전북대학교 컴퓨터공학부(석사)
2005년 University of Florida 전산학(박사)
1995년~현 재 한국전자통신연구원 책임연구원
관심분야 : 자연어 처리, 정보검색, 자연어 질의응답

조 요 한

e-mail : yohani@cs.cmu.edu
2009년 한국과학기술원 전산학과(학사)
2011년 한국과학기술원 전산학과(석사)
2011년~2014년 한국전자통신연구원 연구원
2014년~현 재 Carnegie Mellon University 석사과정
관심분야 : Language Technologies, Medical Data Mining, Learning Sciences

옥 철 영

e-mail : okcy@ulsan.ac.kr

1982년 서울대학교 컴퓨터공학과(학사)

1984년 서울대학교 컴퓨터공학과(석사)

1993년 서울대학교 컴퓨터공학과(박사)

1994년 러시아 TOMSK 공과대학 교환교수

1996년 영국 GLASGOW 대학교 객원교수

2007년~2008년 한국정보과학회 언어공학연구회 위원장

2008년 국립국어원 객원교수

1984년~현 재 울산대학교 전기공학부 IT융합전공 교수

2010년~현 재 울산대학교 국어국문학부 겸직교수

관심분야 : Korean Language Processing, Korean Homograph Tagging, Ontology, Knowledge Base, Document Clustering