韓國IT서비스學會誌
第14卷 第3號
2015年 9月, pp.33-48

Journal of Information Technology Services
http://dx.doi.org/10.9716/KITS.2015.14.3.033

텍스트 분석을 활용한 과학기술이슈 여론 분석 방법론

김다솜* · William Xiu Shun Wong* · 임명수* · 류 신*
김남규** · 박준형*** · 길우영*** · 윤한술***

A Methodology for Analyzing Public Opinion about Science and Technology Issues Using Text Analysis

Dasom Kim* · William Xiu Shun Wong* · Myungsu Lim* · Chen Liu*
Namgyu Kim** · Junhyung Park*** · Wooyeong Kil*** · Hansool Yoon***

■ Abstract ■

Recently, many users frequently share their opinions on diverse issues using various social media. Therefore, many governments have attempted to establish or improve national policies according to the public opinions captured from the various social media. In this paper, we indicate several limitations of traditional approaches for analyzing public opinions about science and technology and provide an alternative methodology to overcome the limitations. First of all, we distinguish science and technology analysis phase and social issue analysis phase to reflect the fact that public opinion can be formed only when a certain science and technology is applied to a specific social issue. Next, we apply a start list and a stop list successively to acquire clarified and interesting results. Finally, to identify most appropriate documents fitting to a given subject, we develop a new concept of logical filter that consists of not only mere keywords but also a logical relationship among keywords. This study then analyzes the possibilities for the practical use of the proposed methodology thorough its application to discovering core issues and public opinions from 1,700,886 documents comprising SNS, blog, news, and discussion.

Keyword : Big Data, Social Network Analysis, Text Mining, Topic Modeling

Submitted : July 26, 2015 1st Revision : September 1, 2015 Accepted : September 4, 2015

* 국민대학교 비즈니스IT전문대학원
** 국민대학교 경영정보학부 부교수, 교신저자
*** 실전전략연구소

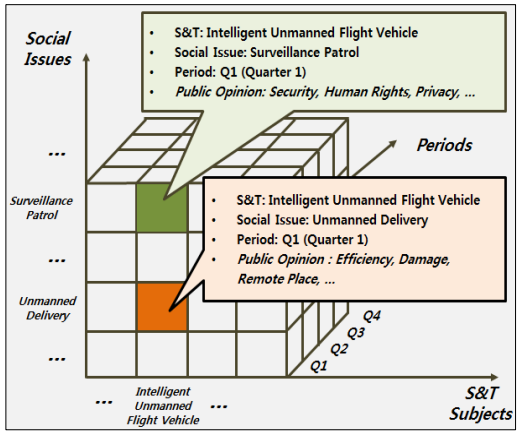
1. 서 론

최근 참여형 매체의 보급 및 활용이 급증함에 따라 많은 사용자들이 개인의 의견을 게시하고 타인과 토론할 수 있는 소통 공간이 증가하고 있다. 특히 기존에는 단방향 커뮤니케이션으로 여겨지던 정부의 정책에 대한 언론의 보도에 대해서도, 인터넷 뉴스의 댓글, 소셜 네트워크 서비스(Social Network Service), 토론 게시판 등 다양한 매체를 통해 국민들이 개인의 의사표현을 적극적으로 하는 현상이 두드러지게 나타나고 있다. 즉 다양한 매체들을 통해 사회 이슈에 대해 유사한 의견을 가진 개개인이 모여 여론을 쉽게 형성할 수 있게 되었으며, 이렇게 형성된 여론은 다시 정부로 전달되어 정책의 보완 및 새로운 정책 수립에 적지 않은 영향을 미치고 있다. 역설적으로 이러한 현상은 여러 매체에 이미 표출되어 있는 다양한 여론을 충분히 수렴하지 않은 정부의 정책에 대해 국민들이 더욱 큰 불만을 느끼게 하는 원인이 되기도 한다.

이렇게 소셜 네트워크 서비스를 통해 표출되는 데이터의 양은 실로 방대하여, 2020년에는 디지털 정보의 양이 35제타바이트(ZB)로 폭증할 것으로 전망되고 있다(KISA, 2014). 구체적으로 2014년 기준 국내 인구의 57%가 소셜 네트워크 서비스를 이용하고 있으며, 트위터(Twitter)는 전 세계적으로 하루 평균 4억 건, 페이스북(Facebook)은 시간 당 1천만 건 가량의 데이터가 발생하고 있다. 이처럼 사용자의 의견을 담은 텍스트 데이터의 양은 급증하고 있으며, 이로 인해 방대한 데이터로부터 필요한 정보만을 추출하고 가공 및 분석하여 새로운 지식을 창출하는 과정은 보다 더 많은 시간과 노력을 요구하게 되었다.

이러한 문제점을 해결하기 위해 최근 빅데이터(Big Data) 분석 기술에 대한 관심이 높아지고 있다. 빅데이터는 데이터의 규모가 방대하고(Volume), 생성에서 유통까지의 속도가 빠르며(Velocity), 속성과 형태가 매우 다양(Variety)하다는 특징을 갖는 정형 또는 비정형 데이터를 의미한다(Gartner, 2012). 빅데이터 관련 기술은 향후 IT업계의 핵심 기술로 자리 잡을 것으로 주목받아 왔으며, 현재 이에 대한 학계와 산업계의 연구 및 관련 솔루션의 상용화가 활발히 진행되고 있다. 특히 텍스트 형태의 비정형 데이터에 대한 분석을 통해 문서 내용을 요약하고, 주요 토픽(Topic)을 추출하고, 유사 문서를 분류하는 등의 작업을 수행하는 텍스트 마이닝(Text Mining) 분야에 대한 연구와 투자가 급증하고 있다.

최근에는 국가적 차원에서 과학기술에 대한 국민의 여론을 반영하여 정책을 수립 또는 개선하기 위한 다양한 시도가 이루어지고 있다. 하지만 지금까지의 노력은 국민의 여론이 과학기술 자체에 대해 형성되기 보다는 각 과학기술이 특정 사회이슈에 적용될 때 형성된다는 특징을 간과한 측면이 있다. 예를 들어 <Figure 1>에서 “드론”으로 알려진 “지능형 무인 비행체”에 대한 국민의 여론은 해당 기술이 “감시 정찰” 이슈에 적용될 때 “안보”, “인권”, “사생활”의 키워드로, 그리고 “무인 택배” 이슈에 적용될 때 “효율”, “파손”, “도서 산간”의 키워드로 서로 다르게 나타남을 알 수 있다.



〈Figure 1〉 Different Social Issues and Public Opinions for the Same Science and Technology

또한 국가 정책 수립을 지원하기 위해 텍스트 분석을 활용한 기존의 연구들은 문서 식별을 위한

키워드 선정 과정을 충분히 신중하게 다루지 않은 측면이 있다. 예를 들어 “뇌 질환 관련 과학기술”에 대한 여론을 분석하고자 하는 경우 기존 대부분의 연구는 우선 뉴스, 토론, SNS 등의 시장 데이터 가운데 “뇌 질환” 등의 키워드가 포함된 문서를 식별하고, 이에 대한 토픽 모델링(Topic Modeling) 또는 이슈 트래킹(Issue Tracking) 등의 분석을 수행하여 그 결과를 제시한다. 이 때 토픽 모델링과 이슈 트래킹 분석을 통해 도출된 결과가 어떤 주제에 대한 여론인지는 최초 분석 대상 문서를 식별하는 과정에서 제시된 키워드에 의해 정의된다. 따라서 연구 주제에 부합하는 충분한 키워드를 제시하지 못할 경우, 분석 대상이 되어야 할 문서가 분석 대상에서 누락되거나 주제와 직접적인 관련이 없는 문서가 분석에 포함되는 등의 부작용이 발생할 수 있다. 이처럼 분석 대상 문서의 식별을 위해 키워드를 선정하는 과정은 전체 분석 결과의 품질에 매우 큰 영향을 끼침에도 불구하고, 토픽 모델링, 이슈 트래킹, 시각화 등의 분석 과정에 비해 상대적으로 그 중요성이 덜 부각되어 왔다.

마지막으로 텍스트 분석을 다루는 대부분의 연구는 공통적으로 결과로 제시되는 어휘의 정제 수준 설정의 딜레마를 겪어왔다. 예를 들어 어휘가 과도하게 정제되는 경우 분석 결과가 상투적으로 나타나게 되고, 반대로 정제 수준이 지나치게 낮은 경우는 다듬어지지 않은 어휘가 그대로 결과에 포함되어 분석의 신뢰성을 떨어뜨리게 된다. 일반적으로 실무에서 높은 정제 수준이 필요한 분석의 경우 용어 사전(Start List)을 사용하고 다양한 어휘를 결과에서 제시하고 싶은 경우는 제한된 수준의 불용어 사전(Stop List)을 사용하게 된다. 이 때 용어 사전을 사용한 분석에서는 용어 사전이 충분히 많은 수의 어휘를 포함하지 않는 경우 분석 결과가 상투적으로 나타나게 되며, 불용어 사전을 사용한 분석에서는 불용어 사전이 충분히 많은 수의 어휘를 포함하지 않는 경우 분석 결과에 덜 다듬어진 어휘가 나타나게 된다.

본 연구에서는 과학기술에 대한 여론 분석 과정에서 앞서 언급한 기존 연구의 세 가지 한계를 극복하기 위한 방안을 제시하고자 한다. 구체적으로 각 과학기술과 특정 사회이슈와의 조합에서 여론이 형성되는 현상을 반영하기 위해 2단계 분석 방법론을 제안하며, 어휘 정제 수준 설정의 딜레마를 해결하기 위해 2단계 분석 과정에서 용어 사전과 불용어 사전을 순차적으로 사용하고자 한다. 또한 주제에 부합하는 문서를 식별하기 위해 키워드의 단순 나열이 아닌 키워드의 논리 조합으로 구성된 논리 필터(Logical Filter)를 활용하는 방안을 새롭게 제시한다.

본 논문의 이후 부분은 다음과 같이 구성된다. 다음 장인 제 2장에서는 본 연구와 관련된 선행 연구들을 요약하고, 제 3장에서는 본 연구의 전체적인 개요와 방법론을 소개한다. 제 4장에서는 제 3장에서 제시한 방법론을 실제 데이터에 적용한 실험 결과를 살펴보고 마지막 제 5장에서는 본 연구의 기여 및 한계를 요약한다.

2. 관련 연구

최근 사회 전반에서 방대한 양의 정형 또는 비정형 데이터가 생성되고 있다. 이러한 데이터의 수집, 저장, 검색, 분석, 그리고 시각화에 대한 연구가 여러 분야에 이루어지고 있으며, 이처럼 데이터의 양 자체가 문제의 일부분이 되는 분석 및 기법을 빅데이터 분석(McKinsey, 2011)이라 한다. 특히 빅데이터 분석은 개인이나 집단 또는 개념을 하나의 노드(Node)로, 그리고 각 노드간 관계를 링크(Link)로 표현하여 개체간 연결 형태나 구조를 분석하는 소셜 네트워크 분석(Social Network Analysis)(Kwahk, 2014) 또는 비정형 텍스트 데이터를 분석하여 가치 있는 정보를 추출하는 텍스트 마이닝(Witten, 2005)과 연계하여 복잡다단한 문제를 해결하기 위한 다양한 응용(Son et al., 2009)에 사용되고 있다. 특히 텍스트 마이닝은 온라인(On-line)과 오프라인(Offline)상에서 정보를 표현하고 교환하는 대표적인 방식인 텍스트를 직접 분석한

다는 점에서 관심과 활용도가 꾸준히 증가하고 있다.

더욱이 최근 인터넷과 모바일 기기의 발달로 인해 대중들이 직접 참여하여 자신의 의견이나 정보를 온라인 상에 표현하는 일이 매우 용이하게 되었다. 이로 인해 SNS, 인터넷 뉴스, 각종 인터넷 게시판 등에서 발생하는 텍스트 데이터의 양이 기하급수적으로 급증하고 있으며, 이러한 비정형 텍스트 데이터를 분석하여 기존의 정형 데이터마이닝에서는 다루지 못했던 사회 현상을 설명하기 위한 시도가 활발하게 이루어지고 있다.

특히 텍스트 마이닝의 핵심인 다량의 문서로부터 주요 주제를 식별하는 토픽 모델링(Huang et al., 2013)을 활용하는 연구가 최근 다양한 분야에서 활발하게 수행되고 있다. 구체적으로는 국가 현안 주제를 효과적이고 체계적으로 선정하기 위해 키워드 유사도 분석, 연관 관계 분석, 소셜 네트워크 분석을 활용하여 현안 관련 키워드를 도출하고 이에 대응되는 R&D 정보를 패키징하는 연구(Hyun et al., 2014), 이슈의 기간별 추이를 비교하여 이슈가 생성되고 소멸되며 변화하는 생명주기를 분석한 연구(Lim and Kim, 2014), 온라인 쇼핑몰 고객의 인터넷 사용 기록을 기반으로 고객의 실제 관심분야를 파악하고 사용자 중심의 카테고리(Category)를 설계하여 새로운 고객군을 정의한 연구(Kim et al., 2014), 그리고 QR코드에 워드 클라우드(Word Cloud) 기술을 접목하여 사용자가 능동적으로 디지털 콘텐츠를 생산하는 전략을 제안한 연구 등을 통해 다양한 문제 해결에 대한 토픽 모델링의 실효성이 입증되어 왔다.

정책 연구 분야에서도 빅데이터를 활용하여 과학기술 정책을 발굴 또는 개선하기 위한 많은 시도가 있어왔으며, 정부, 학계 뿐 아니라 산업계에서도 텍스트 마이닝을 활용한 다양한 정책 연구 방안이 제시되어 왔다. 빅데이터 분석을 활용한 과학기술 정책 연구의 최근 예로는 텍스트 데이터에 대한 키워드 빈도 분석 및 트렌드 분석을 통해 철강 산업계의 경쟁사 전략, 관심 국가의 시장 변화와 해외 사업장 여론 등을 파악한 연구(Min et al., 2014), 대선 기간의 트위터 데이터를 분석하여 실시간 사회적 이슈를 도출한 연구(Bae et al., 2013), 항공 관련 국내 학술 논문 데이터에 대한 토픽 모델링 분석을 통해 항공 분야의 연구 동향을 분석하고 미래 유망 분야를 제시한 연구(Kim et al., 2015), 무기 기술 관련 단어가 포함된 논문 텍스트 데이터를 활용하여 미래 무인 전투기술에 대한 지식 네트워크 구조와 국방 유망기술에 대한 식별 방안을 제시한 연구(Lee and Lee, 2010), 그리고 텍스트 분석을 통해 건설 분야의 트렌드를 분석한 연구(Jeong and Kim, 2012) 등을 들 수 있다.

하지만 텍스트 분석을 활용한 기존 과학기술 정책 연구의 대부분은 실제 분석 과정에 비해 분석 대상 문서의 수집 과정을 충분히 자세하게 다루지 않았다는 공통점을 갖는다. 분석 대상 문서의 선정이 올바르게 이루어지지 않으면 최종 분석 결과의 품질을 보장할 수 없기 때문에, 분석 대상 문서의 선정 및 수집 과정은 보다 구체적으로 설명될 필요가 있다. 또한 기존의 과학기술 정책 연구들은 과학기술과 사회이슈 간의 다대다(M : N) 대응관계를 충분히 고려했다기 보다는 특정 과학기술과 특정 사회이슈의 조합에 집중에서 수행되어 왔다. 즉 하나의 과학기술이 다양한 사회이슈에 적용될 수 있고 반대로 하나의 사회이슈에 다양한 과학기술이 적용될 수 있기 때문에, 과학기술과 사회이슈를 분리하여 분석하는 방향으로 과학기술 정책 연구가 이루어질 필요가 있다.

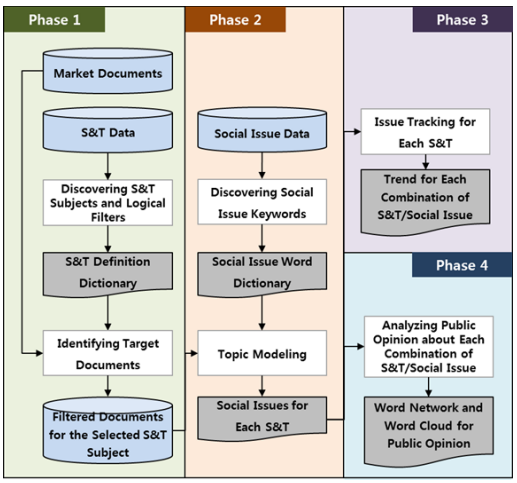
3. 과학기술이슈 여론 분석 방법론

3.1 연구 개요 및 범위

본 절에서는 과학기술이슈에 대한 여론 분석 방법론의 개요를 <Figure 2>를 통해 제시하며, 각 모듈에 대한 상세한 설명은 이후 절에서 다룬다.

우선 <Figure 2>의 Phase 1은 뉴스, 블로그, 트위터, 그리고 토론 등 여론이 담긴 문서 집합에서 특정 과학기술과 관련된 문서만을 추출하는 과정

을 나타낸다. 이를 위해 우선 각종 과학기술 문서로부터 분석 대상 과학기술 주제를 선정하고, 각 주제를 정의할 수 있는 논리 필터를 도출한다. 논리 필터는 각 주제를 키워드의 논리 조합으로 나타낸 것으로, 본 연구에서 새롭게 제안되는 개념이다. 이렇게 생성된 논리 필터는 과학기술 정의사전에 수록되며, 여론 데이터에 해당 논리 필터를 적용함으로써 특정 과학기술과 관련된 문서를 추출할 수 있다.



〈Figure 2〉 Research Overview

한편 Phase 2는 Phase 1에서 추출한 특정 과학기술 관련 문서로부터 해당 과학기술과 관련된 주요 사회이슈를 도출하는 과정을 나타낸다. 이를 위해 각종 사회이슈 자료로부터 주요 키워드를 도출하여 이를 사회이슈 용어 사전에 수록하고, 이 사전을 추후 분석에서 용어 사전으로 사용한다. 즉 특정 과학기술 관련 문서에 대해 토픽 모델링을 수행하는 과정에서 사회이슈 용어 사전에 수록된 어휘만 주요 이슈를 구성하는 키워드로 노출되게 함으로써 충분히 정제된 형태의 사회이슈 리스트를 도출할 수 있다.

다음으로 Phase 3에서는 이슈 트래킹을 통해 과학기술별 주요 사회이슈의 기간별 흐름을 분석하는 작업을 수행한다. 마지막으로 Phase 4에서는 특정 과학기술과 관련된 사회이슈 각각에 대한 여론 분석이 이루어진다. 구체적으로 여론 분석은 대상 문서에 빈번하게 나타나는 어휘 및 동시출현 어휘 패턴을 각각 워드 클라우드와 워드 네트워크(Word Network)로 나타내는 방식으로 수행되며, 이 과정에서는 다양한 여론을 분석 결과에 포함시키기 위해 용어 사전이 아닌 불용어 사전을 적용한다. 이상의 과정을 통해 도출된 결과물인 각 과학기술별 사회이슈와 이들 이슈의 흐름, 선정된 사회이슈에 대한 워드 클라우드 및 워드 네트워크 분석을 통해 각 과학기술과 특정 사회이슈의 조합에 대한 여론을 파악할 수 있다.

3.2 논리 필터를 활용한 특정 과학기술 관련 문서 식별

본 절에서는 방대한 양의 문서로부터 특정 과학기술과 관련된 문서만을 식별하기 위해 새로 도입한 논리 필터의 개념을 소개한다. 텍스트 분석을 통해 여론을 파악하는 기존 대부분의 연구들은 특정 키워드로 검색을 수행한 후 그 결과로 제시된 전체 문서를 분석 대상 문서로 선정한다. 하지만 이러한 방식은 문서가 포함해야 할 키워드를 지나치게 많이 제시하는 경우 분석 대상이 되어야 할 문서가 분석 대상에서 누락될 수 있으며, 반대로 포함해야 할 키워드를 너무 적게 제시하는 경우 주제와 큰 관련이 없는 문서가 우연히 분석에 포함될 수 있다는 한계를 갖는다.

본 연구에서는 분석 대상 문서의 보다 정확한 식별을 위해 키워드의 단순 나열이 아닌 키워드의 논리 조합으로 구성되는 논리 필터의 개념을 새롭게 제안한다. 예를 들어 “감염병 대응 기술”에 관한 문서를 식별하는 경우를 살펴보자. 해당 기술을 기술하기 위해서는 최소 두 가지 개념이 필요한데, 한 가지는 “감염”이고 다른 한 가지는 이에 대한 “대응”이다. 이 때 “감염”은 “전염”, “확산” 등의 용어에 의해서도 표현될 수 있으며, “대응”은 “방역”, “접종”, “격리” 등의 용어로도 표현될 수 있다. 따

라서 “감염병 대응 기술”에 해당되는 문서는 “감염”, “전염”, “확산” 중 최소한 하나의 용어와 “대응”, “방역”, “접종”, “격리” 중 최소한 하나의 용어를 동시에 포함하고 있어야 함을 알 수 있다. 이 경우 “감염병 대응 기술”은 다음의 식 (1)과 같은 논리 필터에 의해 정의할 수 있다. 단, 아래 식에서 ‘*’ 기호는 논리 연산 AND를, ‘+’ 기호는 논리 연산 OR를 나타낸다.

감염병 대응 기술 (1)
← (감염+전염+확산)×(대응+방역+접종+격리)

동일한 방식에 의해 최근 “드론”으로 널리 알려진 “지능형 무인 비행체 기술”은 식 (2)와 같은 논리 필터로 정의될 수 있다.

지능형 무인 비행체 기술 (2)
← (드론)+(무인+지능형)×(비행체+탑재+촬영)

식 (2)의 논리 필터를 사용하여 “지능형 무인 비행체 기술”에 대한 문서만을 식별하는 예가 <Table 1>과 <Table 2>에 나타나있다. <Table 1>은 가상의 문서 Doc.1~Doc.3이 제시된 6개의 용어 각각을 포함하고 있는지 여부를 나타내는 행렬이며, 특정 문서가 특정 용어를 포함한 경우 ‘1’의 값으로, 그렇지 않은 경우 ‘0’의 값으로 표시한다. 한편 <Table 2>는 <Table 1>에서 유도된 것으로, 각 문서가 식 (2)의 논리 필터를 만족시키는지 여부를 판별하기 위한 중간 과정과 최종 결과를 나타낸다. 예를 들어 Doc.1의 경우 용어 “드론”을 포함하고 있기 때문에 다른 용어의 포함 여부와 관계없이 식 (2)를 만족시키며, Doc.3의 경우 “Term 2”와 “Term 3”를 모두 만족시키기 때문에 (Term 2)×(Term 3)의 값도 참이 되어 식 (2)를 만족시키게 된다.

〈Table 1〉 Word Presence(Zero/One) Matrix

Doc.ID	Dron	Unmanned	Intellecral	Aerial Vehicle	Load	Record
Doc.1	1	0	0	1	0	0
Doc.2	0	1	1	0	0	0
Doc.3	0	1	0	1	0	1

〈Table 2〉 Example : Logical Filtering

Doc.ID	Term1 (Dron)	Term2 (Unmanned+ Intellecral)	Term3 (Aerial Vehicle+ Load+Record)	ACCEPT
Doc.1	1	0	1	YES
Doc.2	0	1	0	NO
Doc.3	0	1	1	YES

이상의 과정을 통해 여론 데이터 중 특정 과학기술에 해당하는 문서를 식별할 수 있으며, 이렇게 식별된 문서 집합이 포함하고 있는 주요 사회이슈를 도출하는 과정은 다음 절에서 소개한다.

3.3 과학기술별 주요 사회이슈 흐름 도출

본 절에서는 <Figure 2>의 Phase 2와 Phase 3에서 수행되는 작업을 소개한다. 우선 Phase 2의 토픽 모델링을 위해 각종 사회이슈 자료로부터 빈발 용어를 추출하고, 이에 대한 전문가의 검토를 거쳐서 사회이슈 용어 사전을 구축한다. 이후 Phase 1에서 추출된 특정 과학기술 관련 문서에 대한 토픽 모델링을 수행하여, 해당 과학기술과 관련된 주요 사회이슈를 도출할 수 있다. 이 때 각 사회이슈를 설명하는 이슈 키워드는 모두 사회이슈 용어사전에 수록된 용어들로만 구성된다.

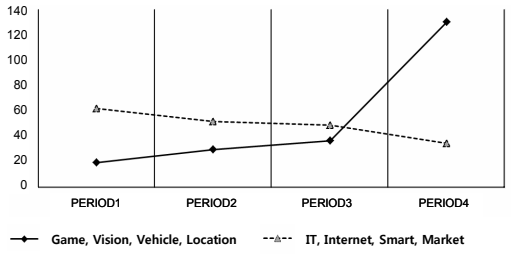
〈Table 3〉 Example : Topic Modeling for “Intellectual Unmanned Aerial Vehicle”

Issue ID	Keywords	Num. Terms	Num. Docs
Issue 1	Game, Vision, Vehicle, Location	217	156
Issue 2	IT, Internet, Smart, Marker	198	130
…	…	…	…

<Table 3>은 과학기술 “지능형 무인 비행체 기술”에 대해 토픽 모델링을 수행한 가상의 결과를 나타내며, 각 이슈별 주요 키워드와 이슈에 포함된 용어의 수, 그리고 이슈에 포함된 문서의 수를 보

이고 있다. 토픽 모델링은 이미 기존의 많은 연구에서 다루어졌을 뿐 아니라 다양한 상용 프로그램을 통해 기계적으로 수행이 가능하기 때문에 토픽 모델링의 자세한 과정에 대한 소개는 생략하도록 한다.

토픽 모델링을 통해 도출된 결과는 특정 과학기술과 관련된 사회이슈를 정적인 형태로 나타내기 때문에, 끊임없이 변화하는 동적인 사회이슈에 대한 통찰을 제공하기에는 한계가 있다. 따라서 Phase 3에서는 이렇게 도출된 과학기술별 주요 사회이슈들에 대한 이슈 트래킹을 통해 이슈들의 기간별 분포를 분석함으로써, 각 이슈의 성장, 지속 및 소멸 추이를 도식화하여 제공한다. 예를 들어 <Table 3>의 이슈에 대한 기간별 분석의 결과는 <Figure 3>과 같은 형태로 나타낼 수 있다. <Figure 3>에서 가로축은 세분화된 기간을 나타내며, 세로축은 특정 기간 각 이슈에 속한 문서의 수를 나타낸다. 이러한 이슈 트래킹을 통해 특정 과학기술에 대해 최근 관심이 증가하고 있는 사회이슈를 식별할 수 있다.



〈Figure 3〉 Example : Issue Tracking for "Intellectual Unmanned Aerial Vehicle"

3.4 특정 과학기술 대응 사회이슈 여론 분석

본 절에서는 <Figure 2>의 Phase 4에 해당하는 부분인 워드 클라우드와 워드 네트워크 도출 과정을 예를 통해 소개한다. 구체적으로는 Phase 2의 토픽 모델링과 Phase 3의 이슈 트래킹 결과를 근거로 세부 여론 분석이 필요한 사회이슈를 선정하고, 선정된 이슈에 속하는 문서 집합에 대해 키워드의 출현 빈도 및 키워드간 연관관계를 각각 워드 클라우드와 워드 네트워크로 나타내게 된다.

<Table 4>는 과학기술 "지능형 무인 비행체" 관련 사회이슈로 도출된 <Table 3>의 이슈 중 "Issue1"에 속하는 문서 156건에 출현하는 주요 어휘의 출현 빈도를 나타낸 가상 예이며, <Figure 4>는 <Table 4>의 값을 워드 클라우드로 도식화한 예이다. 워드 클라우드는 키워드의 빈도수를 시각적 요소인 글자 크기를 이용하여 표현하므로, 각 키워드의 상대적 출현 빈도를 직관적으로 나타낼 수 있다.

〈Table 4〉 Example : Frequent Terms of "Issue1" of "Intellectual Unmanned Aerial Vehicle"

Term	Freq.	Term	Freq.
드론	150	제작	50
가격	100	인권	50
미국	100	사생활	50
정찰	90	cctv	45
감시	85	불법	45
군사	80	청부살인	40
무기	80	게임	40
전쟁	75	비용	40
안보	70	살상	38
택배	60	파손	30

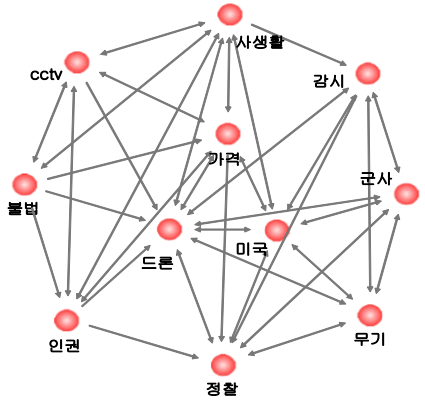
〈Figure 4〉 Example : Word Cloud of "Issue1" of "Intellectual Unmanned Aerial Vehicle"

워드 클라우드는 키워드의 상대적 출현 빈도를 직관적으로 나타낸다는 장점이 있지만, 이는 개별 키워드의 중요도를 나타낼 뿐 키워드간의 관련성은

표현할 수 없다는 한계를 가진다. 일반적으로 워드 클라우드에서 중요하게 부각된 용어는 워드 네트워크에서도 중요하게 나타나지만, 전자는 용어 자체의 빈도에 근거하여 생성되는 반면 후자는 용어 간 동시출현 빈도에 근거하여 생성되기 때문에 두 가지 분석의 결과가 상이하게 나타나는 경우도 비일비재하다.

〈Table 5〉 Example : Association Rules between Terms of "Issue1" of "Intellectual Unmanned Aerial Vehicle"

Conf.	Sup.	Lift	Count	Lhand Term	Hhand Term
100	45	1.11	9	정찰	드론
100	40	1.11	8	군사	드론
100	35	1.11	7	무기	드론
100	25	1.11	5	인권	드론
100	20	3.33	4	cctv	사생활
100	20	1.67	4	cctv	가격
100	15	1.67	3	불법	가격
92	55	1.02	11	미국	드론
92	55	1.02	11	가격	드론
83	25	1.39	5	사생활	가격
…					



〈Figure 5〉 Example : Word Network of "Issue1" of "Intellectual Unmanned Aerial Vehicle"

따라서 본 실험에서는 워드 네트워크를 분석을 수행하여 키워드간 동시출현 패턴을 도식화하고, 이를 통해 특정 키워드와 관련이 있는 키워드 집합을 식별하고자 한다. <Table 5>는 과학기술 "지능형 무인 비행체" 관련 사회이슈로 도출된 <Table 3>의 이슈 중 "Issue1"에 속하는 문서에 출현하는 주요 어휘의 동시출현 패턴을 분석한 연관관계 분석의 가상 결과로, 지지도(Support) 10% 이상의 규칙 중 신뢰도(Confidence) 상위 10개의 규칙을 보여주고 있다. 이들 규칙을 포함하고 지지도 10% 이상, 신뢰도 30% 이상의 규칙을 모두 도식화한 결과가 <Figure 5>에 네트워크 형태로 제시되어 있다.

4. 실험

앞서 제 3장에서는 논리 필터를 사용하여 특정 과학기술 주제와 관련된 문서를 식별하고, 이 문서들에 대한 토픽 모델링을 통해 해당 과학기술과 관련된 사회이슈 및 이들의 추이를 분석하고, 마지막으로 각 과학기술/사회이슈 조합에 대한 여론을 워드 클라우드 및 워드 네트워크를 통해 파악하는 방안을 소개하였다. 본 장에서는 다양한 유형의 시장 데이터에 대해 제안 방법론을 적용한 실험 결과를 소개한다.

4.1 실험 데이터

<Figure 2>의 연구 개요에서 소개한 바와 같이, 제안 방법론을 적용하여 여론을 분석하는 과정에는 크게 과학기술 자료, 사회이슈 자료, 그리고 여론 데이터의 세 가지 데이터가 사용된다.

우선 과학기술 자료의 경우 각 과학기술에 대한 논리 필터를 정의하는 과정에 참고되며, 본 실험에서는 미래창조과학부와 한국과학기술기획평가원(KISTEP)의 "2014년 기술수준 평가"를 사용하였다. 구체적으로는 국가전략기술리스트 120개 가운데 일반 국민들의 관심도가 높을 뿐 아니라 미래창조과학부의 주요 과학기술 육성계획에서 중요하게 다루어진 2개의 전략기술을 분석 대상으로 선정하였으며, 과학기술 자료 중 각 기술에 대한 상세 소개 부분으로부터 주요 어휘를 추출하여 각 전략기술

의 논리 필터를 작성하였다.

다음으로 사회이슈 자료는 토픽 모델링의 결과로 도출되는 이슈 키워드의 정제를 위한 용어 사전 구축에 사용된다. 본 실험에서는 국가과학기술심의회의 “과학기술기반 사회문제 해결 종합실천계획안”으로부터 건강, 환경, 문화, 생활안전, 재난재해, 에너지, 주거교통, 가족, 교육, 사회통합 등 30개 주요 사회 문제의 기술에 사용된 용어를 추출하여 용어 사전을 구축하였다.

마지막으로 여론 데이터는 실제 분석 대상이 되는 문서를 의미하며, 다양한 매체를 통해 수집된 방대한 양의 최근 문서의 집합으로 구성된다. 본 실험에서는 데이터 수집 시점 기준으로부터 최근 1년인 2014년 6월 11일부터 2015년 6월 10일까지를 분석 기간으로 설정하였으나, 일부 데이터의 경우 문서 수집의 한계로 인해 특정 기간의 문서가 누락되기도 하였다. 구체적으로 본 실험에서는 트위터, 네이버 블로그, 다음 아고라, KBS 뉴스, 그리고 연합뉴스로부터 총 1,700,886건의 텍스트 문서를 수집하여 분석하였으며, 각 매체별 수집 문서의 수는 <Table 6>에 요약되어 있다.

〈Table 6〉 Experimental Data Description

	Twitter	NAVER Blog	DAUM Agora	KBS News	YONHAP News
Volume	931,000	365,000	77,486	128,055	199,345
Date	2014. 6. 11. ~ 2015. 6. 10.	2014. 6. 11. ~ 2015. 6. 10.	2014. 7. 6. ~ 2015. 6. 10.	2014. 6. 11. ~ 2015. 6. 10.	2014. 6. 11. ~ 2015. 6. 10.

4.2 특정 과학기술 관련 문서 식별 실험 결과

본 절에서는 제 3.2절에서 소개된 논리 필터의 개념을 적용하여 총 1,700,886건의 시장 데이터로부터 “감염병 대응 기술” 및 “지능형 무인 비행체 기술”과 관련된 문서를 식별한 결과를 소개한다. 우선 과학기술 자료인 “2014년 기술수준평가”를 참고하여 상기 두 가지 과학기술에 대한 논리 필터를 각각 식 (3)과 식 (4)로 정의하였다.

감염병 대응 기술 (3)

← (감염+전염+바이러스)×(방역+면역+항생제+백신+접종+격리+대응+역학조사)

지능형 무인 비행체 기술 (4)

← (드론)+(무인+지능형)×(비행체+탑재+촬영)

한편 위의 두 주제 각각에 대한 논리 필터의 적용을 통해 다양한 매체로부터 추출한 문서의 수가 <Table 7>에 요약되어 있다. 두 주제 모두 뉴스로부터 추출된 문서의 수가 가장 많으며, 트위터에서 도출된 문서의 수는 상대적으로 매우 적은 것으로 나타났다. 이는 트위터의 경우 뉴스 기사에 비해 길이가 상대적으로 매우 짧기 때문에, 논리 필터를 구성하는 다양한 용어를 글 안에 포함하고 있기 어렵다는 점에서 그 원인을 찾을 수 있다. 또한 트위터의 경우 뉴스에 비해 비속어, 은어, 그리고 신조어를 많이 포함하고 있다는 점도 트위터 글이 논리 필터의 식을 만족시키기 어렵게 만드는 또 다른 원인이 될 수 있다. 주제별 문서 수에서는 “감염병 대응 기술”에 대한 문서가 총 3,404건, 그리고 “지능형 무인 비행체 기술”에 대한 문서가 총 450건으로, 동일 기간 각 주제를 다룬 문서의 수는 약 8배의 차이를 보였다.

〈Table 7〉 Number of Logically Filtered Documents

	Twitter	NAVER Blog	DAUM Agora	KBS News	YONHAP News
Total Documents	931,000	365,000	77,486	128,055	199,345
Infectious Disease Control	77	335	217	1,786	989
Intelligent Unmanned Flight Vehicle	38	163	16	109	124

4.3 과학기술별 주요 사회이슈 분석 결과

본 절에서는 <Table 7>에 요약된 “감염병 대응 기술”에 대한 문서 3,404건과 “지능형 무인 비행체 기술”에 대한 문서 450건에 대해 각각 토픽 모델

링을 수행하고, 그 결과로 도출된 주요 사회이슈의 추이를 분석한 결과를 소개한다.

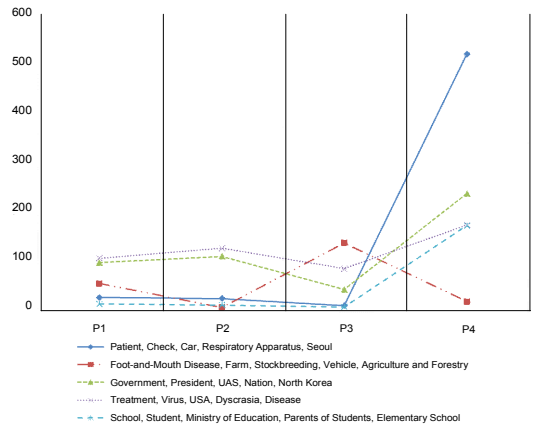
토픽 모델링은 SAS Enterprise Miner 13.1에서 제공하는 Text Miner의 Topic Analysis 모듈을 사용하여 수행하였으며, 토픽, 즉 이슈의 수는 각 주제별 5개씩으로 설정하였다. 각 주제별로 도출된 5개의 이슈에 대한 키워드, 해당 용어의 수 및 문서의 수가 <Table 8>에 제시되어 있다.

이렇게 도출된 사회이슈들의 기간별 추이를 살피기 위해 이슈 트래킹을 수행한 결과가 <Figure 6>과 <Figure 7>에 나타나있다. <Figure 6>은 "감염병 대응 기술"과 관련된 이슈의 추이를, <Figure 7>은 "지능형 무인 비행체 기술"과 관련된 이슈의 추이를 나타낸다. 한편 두 그래프에서 가로축은 기간을 나타내며 2014년 6월 11일부터 2015년 6월 10일까지 1년의 기간이 Period 1~Period 4로 구분되어 있다. 또한 세로축은 해당 기간 각 이슈에 속하는 문서의 수를 나타낸다.

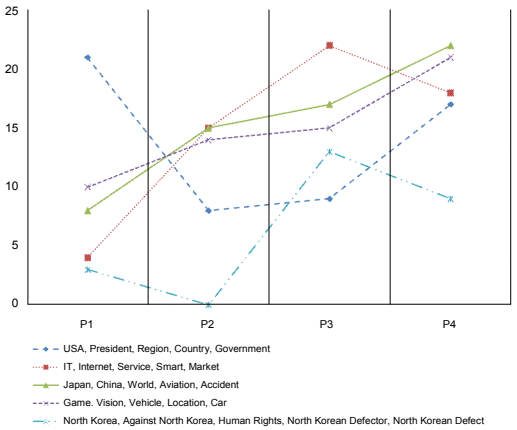
상기 두 주제에 대한 이슈 트래킹 분석 결과, "감염병 대응 기술" 관련 사회이슈인 "Issue1"의 경우 다른 이슈에 비해 최근 관심이 급격히 증가한 것을 알 수 있다. 해당 이슈는 "환자", "검사",

〈Table 8〉 Results of Topic Modeling for Two Subjects

Subject	Issue ID	Keywords	Num. Terms	Num. Docs
	Issue 1	Patient, Check, Car, Respiratory Apparatus, Seoul	81	588
	Issue 2	Foot-and-Mouth Disease, Farm, Stockbreeding, Vehicle, agriculture and Forestry	36	226
	Issue 3	Government, President, USA, Nation, North Korea	100	494
	Issue 4	Treatment, Virus, USA, Dyscrasia, Disease	98	498
	Issue 5	School, Student, Ministry of Education, Parents of Students, Elementary School	46	214
	Issue 6	USA, President, Ragion, Country, Government	35	55
	Issue 7	IT, Internet, Service, Smart, Market	51	59
	Issue 8	Japan, China, World, Aviation, Accident	52	62
	Issue 9	Game, Vision, Vehicle, Location, Car	53	60
	Issue 10	North Korea, Against North korea, Human Rights, North korean Defector, North Korean Defect	21	25



〈Figure 6〉 A Result of Issue Tracking for "Infectious Disease Control"



〈Figure 7〉 A Result of Issue Tracking for "Intellectual Unmanned Aerial Vehicle"

"차", "호흡기", 그리고 "서울"을 이슈 키워드로 가지며, Period 1~Period 3의 기간 동안 거의 주목을 받지 못하다가 "Period 4"에 관련 문서의 수가 급증한 것으로 나타났다. 따라서 본 실험에서는 해당 이슈에 대해 관심이 급증한 원인 및 여론을 분석하기 위해 세부 분석 대상으로 이 이슈를 선정하였으며, 이에 대한 분석 결과는 다음 절인 제 4.4절에서 소개한다.

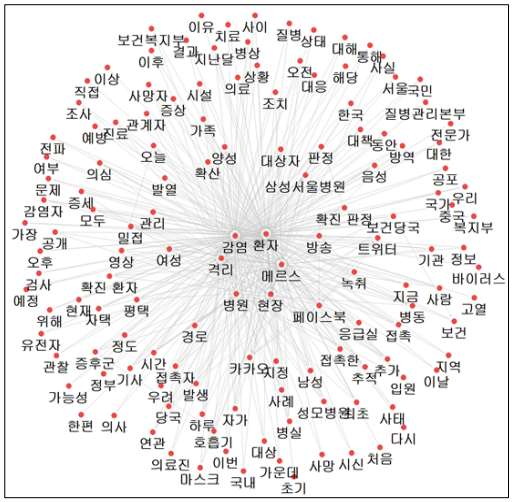
4.4 특정 과학기술 대응 사회이슈 여론 분석 결과

본 절에서는 앞 단계의 이슈 트래킹에서 매우 흥미로운 패턴을 보인 이슈인 "감염병 대응 기술"의 "환자", "검사", "차", "호흡기", "서울" 이슈에 대해 워드 클라우드 및 워드 네트워크 분석을 수행한 결과를 제시한다. 해당 이슈에 속한 문서의 수는 총 588건이며, 워드 클라우드 분석에는 R 3.1.2버전의 wordcloud 패키지를, 워드 네트워크 분석에는 UCINET 6.576버전의 시각화 도구인 NetDraw를 사용하였다.

〈Figure 8〉 A Result of Word Cloud Analysis for the Selected Issue

우선 대상 문서의 빈출 어휘를 워드 클라우드로 나타낸 결과가 <Figure 8>에 제시되어 있다. 그림에서 해당 이슈를 구성하는 가장 중요한 용어는 "메르스", "환자", "병원", "감염", "격리", "정부" 등으로, "메르스"라는 감염병의 발생 및 이에 대한 정부의 대응이 해당 이슈를 형성하였음을 알 수 있다. 본 분석에서는 가급적 다양한 여론을 수렴하기 위해 최소 수준의 불용어 사전을 사용하였으며, 이로 인해 "이날", "가장", "해당" 등 무의미한 단어들도 분석 결과에 일부 포함되어 있음을 확인할 수 있다.

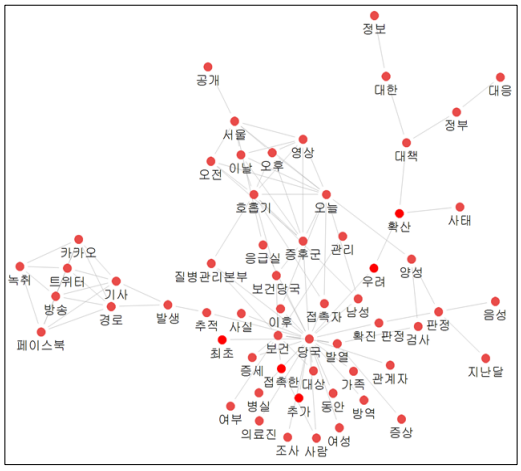
한편, 대상 문서의 용어간 동시출현 패턴을 분석하여 도식화한 결과가 <Figure 9>의 워드 네트워크에 나타나있다.



〈Figure 9〉 A Result of Word Network Analysis for the Selected Issue(conf. ≥ 0.9)

하지만 <Figure 9>는 연결 신뢰도의 임계값을 0.9로 높게 설정하였음에도 불구하고 결과가 매우 복잡하게 나타나므로, 이 그림을 통해 용어간 연결 관계를 파악하기는 매우 어렵다. 따라서 거의 모든 용어와 직접적으로 연결이 되어있는 핵심 용어인 "메르스", "환자", "병원", "감염", "격리", 그리고 "현장"을 제외한 상태에서 연결 관계를 다시 도식화한 결과가 <Figure 10>에 나타나있다. 그 결과 연결 신뢰도의 임계값을 0.7로 하향 조정하였음에도 불구하고, 충분히 해석 가능한 수준의 네트워크가 도출됨을 알 수 있다.

<Figure 10>에서 해당 주제에 대해 나타난 여론은 “정부”, “대응”, “대책”, “보건”, “당국” 등 정부의 대응에 대한 관심과 “확산”, “사태”, “우려” 등 감염병 확산에 대한 두려움이 주를 이루고 있음을 알 수 있다. 이러한 결과는 상기 핵심 용어 6개를 제외하고 다시 워드 클라우드 분석을 수행한 결과인 <Figure 11>에서도 동일하게 확인할 수 있다.



〈Figure 10〉 Filtered Word Network Excluding Six Core Words(conf. ≥ 0.7)

〈Figure 11〉 Filtered Word Cloud Excluding Six Core Words

본 실험에서는 제안 방법론을 실제 시장 데이터 분석에 적용하여 특정 과학기술과 관련된 특정 사회이슈에 대한 여론을 분석하는 과정을 소개하였다. 구체적으로는 트위터, 네이버 블로그, 다음 아고라, KBS 뉴스, 그리고 연합뉴스로부터 수집한 총 1,700,886건의 텍스트 문서로부터 “감염병 대응 기술” 및 “지능형 무인 비행체 기술”과 관련된 문서만을 각각 추출하고, 이에 대한 토픽 모델링을 통해 각 과학기술과 관련된 최근 1년간의 주요 이슈를 5개씩 발굴하였다. 또한 주요 이슈의 기간별 추이를 분석하여 “감염병 대응 기술”과 관련된 “환자”, “검사”, “차”, “호흡기”, “서울” 이슈를 여론 분석 대상 이슈로 선정하였다. 이렇게 선정된 이슈에 대한 워드 클라우드 및 워드 네트워크 분석을 통해 해당 이슈에 대한 주요 여론을 확인할 수 있었다.

5. 결 론

최근 다양한 매체들을 통해 특정 주제에 대해 유사한 의견을 가진 개개인이 모여 여론을 쉽게 형성할 수 있게 되었으며, 정부는 여러 매체에 이미 표출되어 있는 다양한 여론을 충분히 수렴하여 정책을 수립하기 위한 노력을 다각적으로 기울이고 있다. 특히 최근에는 국가적 차원에서 과학기술에 대한 국민의 여론을 반영하여 정책을 수립 또는 개선하기 위한 다양한 시도가 이루어지고 있다. 본 연구는 뉴스, 트위터, 블로그, 토론 등 다양한 시장 데이터로부터 과학기술 관련 사회이슈를 발굴하고, 주요 사회이슈의 기간별 추이 및 특정 과학기술과 사회이슈의 조합에 대한 상세 여론을 살펴볼 수 있는 빅데이터 분석 기반 과학기술 여론 분석 방안을 제시하였다. 또한 제안 방법론을 실제 시장 데이터 1,700,886건 분석에 적용함으로써 특정 과학기술 분야에서 흥미로운 추세를 보인 사회이슈를 발견하고, 이에 대한 여론 분석 결과를 도식화하여 제시하였다.

본 연구의 기여는 크게 학술적 측면과 실무적

측면에서 찾을 수 있다. 우선 학술적 측면에서 본 연구는 유사 연구에서 간과되어 온 구체적 문서 수집 기준 설정의 중요성을 지적하고, 주제에 부합하는 문서의 식별을 위한 방안으로 논리 필터에 의한 정의식 개념을 새롭게 제시하였다. 또한 특정 과학기술에 대한 여론은 과학기술 자체가 아닌 특정 사회 이슈와의 조합을 통해 형성될 수 있음에 착안하여, 과학기술 분석 단계와 사회 이슈 분석 단계를 구분한 2단계 분석 방법론을 새롭게 제시하였다. 본 연구에서 제시한 논리 필터의 개념 및 2단계 분석 방법론은 향후 관련 분야의 유사 연구에서도 충분히 활용되고 더욱 발전할 수 있을 것으로 기대한다. 한편 제안 방법론은 용어 정제 수준 설정의 딜레마 해결을 위해 2단계 분석 과정에서 용어 사전과 불용어 사전을 순차적으로 모두 활용한다. 즉 충분히 정제된 용어로 구성된 이슈 키워드를 제시하기 위해 토픽 모델링 과정에서 용어 사전을 적용하고, 다양한 용어로 표출되는 여론을 수렴하기 위해 워드 클라우드 및 워드 네트워크 분석 과정에서 최소 수준의 불용어 사전을 적용하였다. 이러한 측면에서 제안 방법론은 정제된 형태의 이슈 키워드와 다양한 용어로 표출된 여론을 모두 파악하고자 하는 실무의 수요를 만족시킬 수 있을 것으로 기대한다.

이상의 기여에도 불구하고 본 연구는 다음과 같은 측면에서 더욱 보완되어야 한다. 우선 본 연구의 실험에 소요된 시간과 노력의 대부분이 논리 필터를 통한 문서 식별 과정에 집중되었다. 따라서 향후 논리 필터에 의한 문서 식별이 여러 분야에서 활발하게 이루어지기 위해서는, 방대한 문서와 논리 필터 정의식을 입력으로 받아서 주제에 해당되는 문서만을 출력하는 전체 과정의 자동화 수준이 더욱 높아져야 한다. 또한 제안 방법론의 최종 결과물인 워드 클라우드와 워드 네트워크는 매우 복잡한 형태로 구성되어 제시되므로, 다양한 소셜 네트워크 분석 기법을 활용하여 최종 결과물을 보다 정량적으로 요약할 필요가 있다.

References

Bae, J.H., J.E. Son, and M. Song, “Analysis of Twitter for 2012 South Korea Presidential Election by Text Mining Techniques”, *Journal of Intelligence and Information Systems*, Vol.19, No.3, 2013, 141-156.

(배정환, 손지은, 송민, “텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석”, *지능정보연구*, 제19권, 제3호, 2013, 141-156.)

Gartner, “2012 Hype Cycle for Emerging Technologies”, *Gartner Inc.*, Stamford, 2012.

Huang, S., W. Peng, J. Li, and D. Lee, “Sentiment and Topic Analysis on Social Media : a Multi-task Multi-label Classification Approach”, *Proceedings of the 5th Annual ACM Web Science Conference*, 2013, 172-181.

Hyun, Y.J., N.K. Kim, and Y.H. Cho, “A Multi-Dimensional Issue Clustering from the Perspective Consumers’ Interests and R&D”, *Journal of Information Technology Services*, Vol.14, No.1, 2015, 237-249.

(현윤진, 김남규, 조윤호, “소비자 선호 이슈 및 R&D 관점에서의 다차원 이슈 클러스터링”, *한국IT서비스학회지*, 제14권, 제1호, 2014, 237-249.)

Jeong, C.W. and J.J. Kim, “Analysis of Trend in Construction Using Textmining method”, *Journal of The Korean Digital Architecture・Interior Association*, Vol.12, No.2, 2012, 53-60.

(정철우, 김재준, “텍스트마이닝을 활용한 건설분야 트랜드 분석”, *한국디지털건축인테리어학회논문집*, 제12권, 제2호, 2012, 53-60.)

Kim, H.J., N.O. Jo, and K.S. Shin, “Text Mining-Based Emerging Trend Analysis for the Aviation Industry”, *Journal of Intelligence and Information Systems*, Vol.21, No.1, 2015, 65-82.

(김현정, 조남옥, 신경식, "항공산업 미래유망 분야 선정을 위한 텍스트 마이닝 기반의 트렌드 분석", *지능정보연구*, 제21권, 제1호, 2015, 65-82.)

Kim, J.E., N.K Kim, and Y.H. Cho, "User-Perspective Issue Clustering Using Multi-Layered Two-Mode Network Analysis", *Journal of Intelligence and Information Systems*, Vol.20, No.2, 2014, 93-107.

(김지은, 김남규, 조윤호, "다계층 이원 네트워크를 활용한 사용자 관점의 이슈 클러스터링", *지능정보연구*, 제20권, 제2호, 2014, 93-107.)

Korea Internet and Security Agency, "2014 Korea Internet White Paper", *Korea Internet and Security Agency*, 2014.

Kwahk, K.Y., "Social Network Analysis", *Cheongram*, Seoul, 2014.

Lee, T.B. and C.J. Lee, "A Study on the Identifying Emerging Defense Technology using S&T Text Mining", *Journal of the Military Operations Research Society of Korea*, Vol.36, No.1, 2010, 39-49.

(이태봉, 이춘주, "S&T Text Mining을 이용한 국방 유망기술 식별에 관한 연구", *한국국방경영분석학회지*, 제36권, 제1호, 2010, 39-49.)

Lim, M.S. and N.K. Kim, "Analyzing the Issue Life Cycle by Mapping Inter-Period Issues", *Journal of Intelligence and Information Systems*, Vol.20, No.4, 2014, 25-41.

(임명수, 김남규, "기간별 이슈 매핑을 통한 이슈 생명주기 분석 방법론", *지능정보연구*, 제20권, 제4호, 2014, 25-41.)

McKinsey Global Institute, "Big Data : The next Frontier for Innovation, Competition, and Productivity", *McKinsey and Company*, 2011.

Min, K.Y., H.T. Kim, and Y.G. Ji, "A Pilot Study on Applying Text Mining Tools to Analyzing Steel Industry Trends : A Case Study of the Steel Industry for the Company 'P'", *Journal of Society for e-Business Studies*, Vol.19, No.3, 2014, 51-64.

(민기영, 김훈태, 지용구, "철강산업 트렌드 분석을 위한 텍스트 마이닝 도입 연구-P社 사례를 중심으로", *한국전자거래학회지*, 제19권, 제3호, 2014, 51-64.)

Son, Y.H., I.K. Kim, and N.G. Kim, "Automated Conceptual Data Modeling Using Association Rule Mining", *The Journal of Information Systems*, Vol.18, No.4, 2009, 59-86.

(손윤호, 김인규, 김남규, "연관규칙 마이닝을 활용한 개념적 데이터베이스 설계 자동화 기법", *정보시스템연구*, 제18권, 제4호, 2009, 59-86.)

Witten, I.H., "Text Mining : Practical Handbook of Internet Computing", *CRC Press*, Florida, 2005.

About the Authors

Dasom Kim (dskim1225@kookmin.ac.kr)

Dasom Kim received the B.A. degree in Business Administration from National Institute for Lifelong Education in 2015. She is a Master Candidate in Business IT at Kookmin University. Her research interests include data mining and database.

William Xiu Shun Wong (williamwong@kookmin.ac.kr)

William Xiu Shun Wong received the B.S. degree in Computer Science from Universiti Sains Malaysia in 2011 and Master degree in Business IT from Kookmin University in 2015. He is a Ph.D. candidate in Business IT at Kookmin University. His current research interests include text mining, data mining, and opinion mining.

Myungsu Lim (amr2001@kookmin.ac.kr)

Myunsu Lim received the B.A. degree in Management Information Systems from Wonkwang University in 2014. He is a Master Candidate in Business IT at Kookmin University. His research interests include text mining, social media mining, and data mining.

Chen Liu (liuchen@kookmin.ac.kr)

Chen Liu received the B.A. degree in Management Information Systems from Chungbuk University in 2013. He is a Master Candidate in Business IT at Kookmin University. His research interests include text mining and data mining.

About the Authors

Namgyu Kim (ngkim@kookmin.ac.kr)

Professor Namgyu Kim received the B.S. degree in Computer Engineering from Seoul National University in 1998 and Ph.D. degree in Management Engineering from Korea Advanced Institute of Science and Technology (KAIST) in 2007. He has been working for Kookmin University since then. His current research interests include text mining, data mining, and data modeling.

Junhyung Park (aldoset@naver.com)

Research Engineer Junhyung Park received the B.S. degree in digital media engineering from Anyang University in 2011 and Master. degree in Business IT from Kookmin University in 2013. He has been working for Actural Strategy Research Institute since then. His current research interests include social network analysis and data mining.

Wooyeong Kil (zillian4004@naver.com)

Senior Research Engineer Wooyeong Kil received the B.S. degree in Journalism from Kyunghee University in 2008 and Master. degree in Journalism from Kyunghee University in 2011. He has been working for Actural Strategy Research Institute since then. His current research interests include journalism analysis

Yoonhan Sool (hsool28@naver.com)

Representative Research Engineer Yoonhan Sool received the B.S. degree in business administration from Hanyang University in 1984 and Ph.D. degree in business administration from SungKyunKwan University in 2007. He has been working for Actural Strategy Research Institute since then. His current research interests include business strategy and six sigma.