ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org

J Intell Inform Syst 2016 March: 22(1): 1~18 http://dx.doi.org/10.13088/jiis.2016.22.1.01

비정형 텍스트 분석을 활용한 이슈의 동적 변이과정 고찰*

임명수
국민대학교 비즈니스IT전문대학원
(amr2001@kookmin.ac.kr)

김남규
국민대학교 경영대학 경영정보학부
(ngkim@kookmin.ac.kr)

최근 가용한 텍스트 데이터 자원이 증가함에 따라 방대한 텍스트 분석을 통해 새로운 가치를 창출하고자 하는 수요가 증가하고 있다. 특히 뉴스, 민원, 블로그, SNS 등을 통해 유통되는 글로부터 다양한 이슈를 발굴해내고 이들 이슈의 추이를 분석하는 이슈 트래킹에 대한 연구가 활발하게 이루어지고 있다. 전통적인 이슈 트래킹은 토픽 모델링을 통해 오랜 기간에 걸쳐 지속된 주요 이슈를 발굴한 후, 각 이슈를 구성하는 문서 수의 세부 기간별 분포를 분석하는 방식으로 이루어진다. 하지만 전통적 이슈 트래킹은 각 이슈를 구성하는 내용이 전체 기간에 걸쳐 변화 없이 유지된다는 가정 하에 수행되기 때문에, 다양한 세부 이슈가 서로 영향을 주며 생성, 병합, 분화, 소멸하는 이슈의 동적 변이과정을 나타내지 못한다. 또한 전체 기간에 걸쳐 지속적으로 출현한 키워드만이 이슈 키워드로 도출되기 때문에, 핵실험, 이산가족 등 세부 기간의 분석에서는 매우 상이한 맥락으로 파악되는 구체적인 이슈가 오랜 기간의 분석에서는 북한이라는 큰 이슈에 함몰되어 가려지는 현상이 발생할 수 있다. 본 연구에서는 이러한 한계를 극복하기 위해 각 세부 기간의 문서에 대한 독립적인 분석을 통해 세부 기간별 주요 이슈를 도출한 후, 각 이슈의 유사도에 기반하여 이슈 흐름도를 도출하고자 한다. 또한 각 문서의 카테고리 정보를 활용하여 카테고리간의 이슈 전이 패턴을 분석하고자 한다. 본 논문에서는 총 53,739건의 신문 기사에 제안 방법론을 적용한 실험을 수행하였으며, 이를 통해 전통적인 이슈 트래킹을 통해 발굴한 주요 이슈의 세부 기간별 구성 내용을 살펴볼 수 있을 뿐 아니라, 특정 이슈의 선행 이슈와 후행 이슈를 파악할 수 있음을 확인하였다. 또한 카테고리간 분석을 통해 단방향 전이와 양방향 전이의 흥미로운 패턴을 발견하였다.

주제어 : 빅데이터, 데이터 마이닝, 이슈 트래킹, 텍스트 마이닝, 토픽 모델링, 트랜드 분석

논문접수일 : 2015년 11월 25일 논문수정일 : 2016년 1월 19일 게재확정일 : 2016년 2월 9일
원고유형 : 일반논문 교신저자 : 김남규

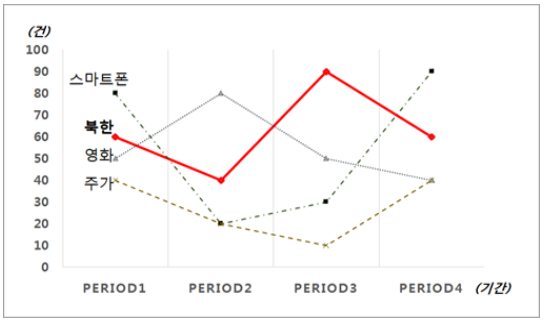
1. 개요

최근 IT 기술의 발달과 다양한 웹 미디어의 보급으로 인해 방대한 양의 데이터가 생성, 공유, 저장되고 있다. 이러한 웹 미디어 상에는 이미지, 영상, 텍스트 등 여러 유형의 비정형 데이터가 유통되고 있으며, 최근 이러한 비정형 데이터에 대한 분석을 통해 새로운 가치를 창출하기 위한 빅데이터 분석에 대한 수요가 급증하고 있다. 특히 웹 상에서 개인의 의견을 표출하고 공유하는 주요 수단으로 사용되는 텍스트 데이터의 경우, 텍스트 마이닝(Text Mining) 분야의 발전으로 인해 다양한 목적으로 활용되고 있다.

특히 방대한 양의 문서로부터 주요 이슈를 발굴하는 토픽 모델링(Topic Modeling)에 대한 연구가 학계와 산업계에서 매우 활발하게 이루어

* 이 논문은 2015년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2015S1A5A2A01011926)

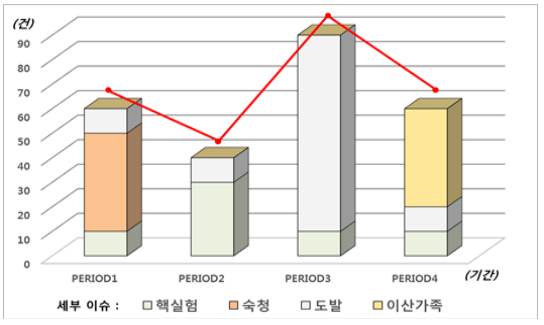
지고 있으며, 최근에는 주요 이슈의 기간별 추이를 분석하는 이슈 트래킹(Issue Tracking) 또는 트랜드 분석(Trend Analysis)에 대한 관심이 급증하고 있다. 전통적인 이슈 트래킹은 토픽 모델링을 통해 오랜 기간에 걸쳐 지속된 주요 이슈를 발굴한 후, 각 이슈를 구성하는 문서 수의 세부 기간별 분포를 분석하는 방식으로 이루어진다<Figure 1>.



〈Figure 1〉 Example of traditional issue tracking

하지만 전통적 이슈 트래킹은 각 이슈를 구성하는 내용이 전체 기간에 걸쳐 변화 없이 유지된다는 가정 하에 수행되기 때문에, 다양한 세부 이슈가 서로 영향을 주며 생성, 병합, 분화, 소멸하는 이슈의 동적 변이과정을 나타내지 못한다. 또한 전체 기간에 걸쳐 지속적으로 출현한 키워드만이 이슈 키워드로 도출되기 때문에, 세부 기간의 분석에서는 매우 상이한 맥락으로 파악되는 이슈가 오랜 기간의 분석에서는 장기 이슈에 함몰되어 가려지는 현상이 발생할 수 있다.

예를 들어 <Figure 1>의 "북한" 이슈를 구성하고 있는 세부 이슈의 가상 분포 예가 <Figure 2>에 나타나 있다. 전통적인 이슈 트래킹의 경우 장기적으로 지속되는 "북한"이라는 이슈가 PERIOD1 ~ PERIOD4에 걸쳐 각각 60건, 40건, 90건, 그리고 60건 발생했다는 사실은 알려주지만, PERIOD1 ~ PERIOD3까지는 "핵실험", "숙청", "도발" 등 부정적인 이슈가 "북한" 이슈를 구성하는 주요 내용인 반면 PERIOD4에서는 "이산가족"이라는 긍정적인 이슈가 해당 이슈를 구성하는 주요 내용임은 설명하지 못한다.



〈Figure 2〉 Detailed issues composing long period issues

<Figure 2>는 장기간에 걸친 이슈 트래킹 분석이 수행되는 경우 전체 기간에 걸쳐 지속적으로 출현한 용어들이 이슈를 구성하는 키워드로 채택될 가능성이 높기 때문에 나타나는 현상을 보이고 있다. 즉 각 이슈의 세부 내용을 묘사하는 "지뢰", "싸이", "손흥민" 등의 구체적 용어보다는 전체 기간에 걸쳐 꾸준히 나타나는 "북한", "한류", "월드컵" 등의 다소 일반적인 용어들이 이슈 키워드를 구성하게 된다. 하지만 이슈 분석을 통해 향후 사회적 변화에 선제적으로 대응한다는 측면에서 볼 때, 오랜 기간을 관통하는 일반적인 용어뿐 아니라 특정 기간에 활발하게 출현했다가 이후 기간에 빠른 속도로 소멸되는 구체적인 용어의 활용가치도 매우 높다고 할 수 있다.

이러한 한계를 보완하기 위해 본 연구에서는 각 세부 기간의 문서에 대한 독립적인 분석을 통

해 세부 기간별 주요 이슈를 도출한 후, 각 이슈의 유사도에 기반하여 이슈 흐름도(Issue Flow Diagram)를 도출하고자 한다. 또한 각 문서의 카테고리 정보를 활용하여 카테고리간의 이슈 전이 패턴을 분석하고자 한다.

본 논문의 이후 구성은 다음과 같다. 우선 다음 장에서는 텍스트 마이닝 및 이슈 트래킹에 대한 선행 연구의 성과를 요약하고, 3장에서는 제안 방법론 및 분석 시나리오를 예를 통해 소개한다. 제안 방법론을 실제 데이터에 적용한 실험 결과는 4장에서 소개하고, 마지막 장인 5장에서는 본 연구의 기여 및 한계를 요약한다.

2. 관련 연구

텍스트는 현실 세계에서 정보를 교환하거나 표현하기 위한 가장 대표적인 수단으로 사용된다(Witten, 2004). 최근 웹 미디어의 발달로 인해 대량의 텍스트 정보가 유통되고 있으며, 이러한 다량의 텍스트에 대한 분석을 통해 의미 있는 가치를 창출하고자 하는 텍스트 마이닝 기반의 시도가 다양한 분야에서 이루어지고 있다(Hearst, 1999; Mooney and Bunescu, 2006; Sebastiani, 2002; Sebastiani, 2006). 텍스트 마이닝은 분석 목적에 따라 행렬, 계층, 벡터 등의 다양한 형태로 표현될 수 있지만, 일반적으 벡터공간모델(Vector Space Model)(Albright, 2006; Salton et al., 1975; Stanvrianou et al., 2007)을 사용하여 구조화되어 나타난다. 즉 각 문서에 포함된 용어의 빈도에 따라 해당 문서의 주제 및 특성이 요약되며, 빈도 표현에는 TF-IDF(Han et al., 2011; Provost and Fawcett, 2013) 값이 주로 사용된다. 빈도의 벡터로 구조화된 문서 집합에 대해 군집분석(Clustering Analysis)을 수행하여 유사 문서 집합을 도출할 수 있으며, 각 군집의 주제는 해당 군집에 속한 문서에서 빈번하게 나타나는 용어들의 집합으로 설명될 수 있다.

개별 문서로부터 토픽을 추출하고 평가하는 일부 연구(Haribhakta et al., 2012)를 제외하면, 대부분의 연구에서 토픽은 다량의 문서 집합으로부터 추출된다. 주로 오류 보고서(Bug Report)(Aggarwal et al., 2014), 헬프데스크 자료(Morinaga and Yamanishi, 2004) 등의 문서로부터 고객의 주요 불만 사항을 식별하기 위한 분석이 활발하게 수행되어 왔으며, 최근에는 특허 키워드 분석을 통해 유망기술을 예측하는 연구(Kim et al., 2014), 트위터(Twitter) 상의 이슈를 추출하고 시각화한 연구(Bae et al., 2014), 문헌정보학 분야의 학술지 분석을 통해 해당 분야의 연구 동향을 파악한 연구(Park and Song, 2013), 인터넷 뉴스 기사의 조회 기록 분석을 통해 사용자 관심의 흐름을 분석한 연구(Liu and Kim, 2014), 특허, 뉴스, 학술지 등 이질적 문서로부터 도출되는 토픽들간의 시간 간격(Time Gap)을 분석한 연구(Jeong and Song, 2014) 등이 수행된 바 있다. 또한 특정 분야의 이슈 발견이 아닌 토픽 분석 기법 자체의 정확도 향상을 다루는 연구도 수행되어 왔다(Rajaraman and Tan, 2001; Wang and McCallum, 2006).

하지만 위의 연구들을 포함한 대부분의 이슈 분석 연구는 각 이슈를 구성하는 내용이 전체 기간에 걸쳐 변화 없이 유지된다는 가정 하에 수행되기 때문에, 다양한 세부 이슈가 서로 영향을 주며 생성, 병합, 분화, 소멸하는 이슈의 동적 변이과정을 나타내지 못한다. 또한 전통적인 이슈 트래킹은 앞에서 제기한 문제, 즉 특정 기간에 활발하게 출현했다가 빠른 속도로 소멸된 흥미로운 세부 이슈가 오랜 기간 지속된 일반적인 이슈에 함몰되는 현상을 방지하지 못한다는 한계

를 갖는다.

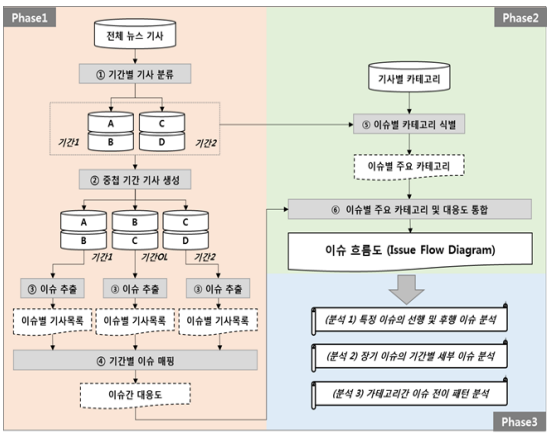
전체 기간으로부터 이슈를 도출하는 방식과 달리 새로 추가되는 기간의 문서에 대해서만 점진적 추가 분석을 통해 효율적으로 이슈를 도출하기 위한 연구(Alsumait et al., 2008)도 수행되었지만, 이 경우 서로 다른 기간에 속하는 이슈간의 관계를 밝혀내기 어렵기 때문에 대부분의 이슈 트래킹 연구들은 전체 기간으로부터 이슈를 도출하는 방식을 사용하였다. 또한 최근 온톨로지(Ontology)에 기반하여 이슈간의 의미적 관계를 분석하기 위한 방안(Ma et al., 2014)이 제시된 바 있지만 이는 의미적 모호성이 해소된 온톨로지의 구축이 선행되어야 한다는 점에서 실제 적용에 큰 어려움이 있다. 또한 사용자의 문서 조회 기록을 활용하여 상이한 시기의 이슈간 관계를 도출하기 위한 시도가 최근 이루어졌지만, 이러한 방식은 문서뿐 아니라 사용자의 문서 조회 기록 정보를 분석해야 한다는 측면에서 활용의 제약을 받는다.

한편 본 연구에서는 상이한 시기의 이슈간 매핑을 위한 방안으로 기간의 인위적 중첩(Overlap)을 사용하고자 한다. 기간 중첩의 개념은 인터넷 뉴스 기사의 기간별 이슈 매핑을 다룬 이전 연구(Lim and Kim, 2014)에서 처음 소개되었다. 하지만 이 연구는 실제 데이터에 대한 충분한 분석 결과를 제시하지 못했을 뿐 아니라, 전통적인 이슈 트래킹을 통해 도출된 이슈와 세부 기간별 분석을 통해 도출된 이슈들간의 관계를 명시적으로 보이지 못했다는 한계를 갖는다. 본 연구에서는 실제 데이터에 대한 충분한 분석을 통해 실제 이슈 흐름도를 도출하고, 전체 기간의 이슈와 세부 기간별 이슈의 관계를 명시적으로 보이고자 한다. 또한 각 이슈의 카테고리 정보를 활용하여 카테고리간 이슈의 전이 패턴을 살펴보고자 한다.

3. 이슈 변이과정 고찰 방법론

3.1 연구 범위

본 연구의 전체 개요는 <Figure 3>과 같다. Phase1은 세부 기간별 이슈를 도출하고 기간별 이슈를 매핑하는 과정을 나타낸다. 본 과정의 기본 개념은 [5]에 소개되어 있으며, 본 연구에서는 3.2절에서 핵심 내용을 요약한다. Phase 2는 카테고리간 이슈의 전이 현상을 규명하기 위해 각 이슈별 주요 카테고리를 식별하고 이슈 흐름도를 도출하는 과정이며 3.3절에서 자세히 다룬다. 마지막으로 Phase3은 도출된 이슈 흐름도의 분석 시나리오를 나타내며, 세 가지 주요 시나리오가 간단한 예와 함께 3.4절에서 소개된다. 본 장에서 제시되는 모든 예는 방법론의 이해를 돕기 위한 가상 예이며, 실제 데이터를 분석한 결과는 다음 장인 4장에서 제시된다.

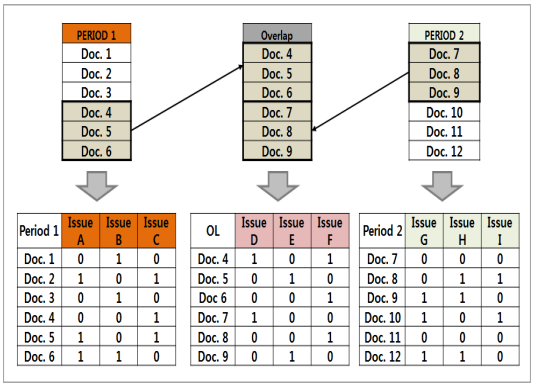


〈Figure 3〉 Research overview

3.2 기간별 이슈 매핑

<Figure 3>의 Phase 1에서는 세부 기간별 이슈를 도출하고 기간별 이슈를 매핑하는 작업을 수

행한다. 우선 수집된 뉴스 기사를 각 기간별로 분류한다(①). 본 개요도에서는 두 기간의 이슈를 매핑하는 경우만을 표현하였으나, 이를 반복적으로 확장하여 셋 이상 기간에 대한 매핑도 동일한 방식으로 수행할 수 있다. 이후 두 기간의 인위적 중첩을 통해 중첩 기간을 설정한 뒤(②), 두 기간과 중첩 기간 각각의 문서 집합에 대한 토픽 모델링을 통해 기간별 주요 이슈를 도출한다(③). 이 과정의 수행 예가 <Figure 4>에 제시되어 있다.



Period 1	Issue A	Issue B	Issue C
Doc. 1	0	1	0
Doc. 2	1	0	1
Doc. 3	0	1	0
Doc. 4	0	0	1
Doc. 5	1	0	1
Doc. 6	1	1	0

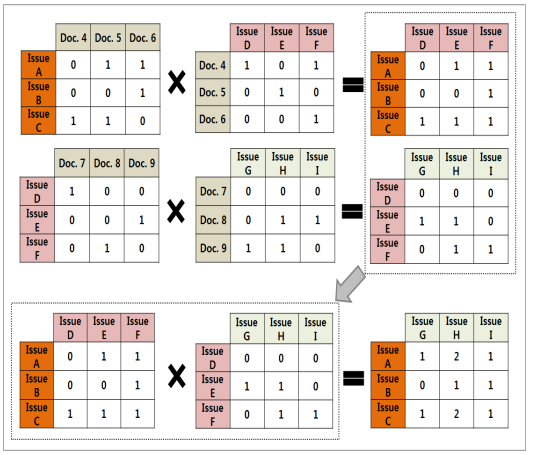
OL	Issue D	Issue E	Issue F
Doc. 4	1	0	1
Doc. 5	0	1	0
Doc 6	0	0	1
Doc. 7	1	0	0
Doc. 8	0	0	1
Doc. 9	0	1	0

Period 2	Issue G	Issue H	Issue I
Doc. 7	0	0	0
Doc. 8	0	1	1
Doc. 9	1	1	0
Doc. 10	1	0	1
Doc. 11	0	0	0
Doc. 12	1	1	0

〈Figure 4〉 Discovering issues of each period

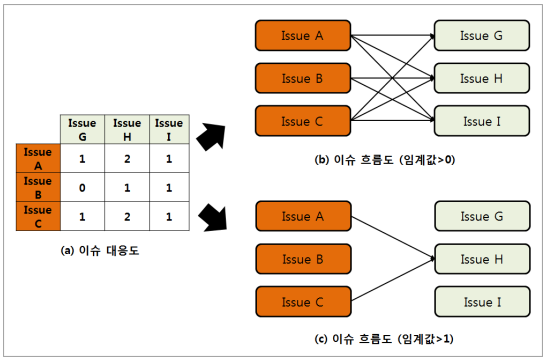
<Figure 4>에서 Period1의 Doc.4 ~ Doc.6과 Period2의 Doc.7 ~ Doc.9의 문서를 통합하여 중첩 기간(Overlap)의 문서 집합을 구성하였으며, 이 집합은 Period1과 Period2의 연결고리 역할을 수행한다. 또한 <Figure 4>에서 Period 1, 중첩 기간, Period2의 주요 이슈로 각각 Issue A ~ C, Issue D ~ F, Issue G ~ I가 도출된 것을 알 수 있다. <Figure 4>의 하단 표는 각 이슈와 문서의 대응 관계를 보여준다.

다음으로 기간별 이슈의 매핑이 수행된다(④). Period1과 Period2의 매핑은 Period1과 Overlap의 매핑, Overlap과 Period2의 매핑, 그리고 두 결과의 최종 매핑의 순서로 이루어지며, 이 과정은 <Figure 5>의 예를 통해 설명된다. <Figure 5>의 첫 행과 둘째 행은 각각 Period1과 Overlap의 매핑, 그리고 Overlap과 Period2의 매핑을 나타내며, 마지막 행은 위의 두 결과의 최종 매핑을 나타낸다. 각 행에서 매핑은 행렬 곱(Matrix Multiplication) 연산에 의해 이루어진다. 예를 들어 첫 행의 우측 테이블은 Issue A ~ C와 Issue D ~ F에 공통으로 포함된 문서의 수를 나타내며, 공통 문서 수가 많을수록 해당 이슈 조합은 유사도가 높은 것으로 해석된다. 예를 들어 마지막 행의 우측 테이블에서 Issue H는 Issue A 및 Issue C와 각각 2개의 문서를 공유하고 있으므로 이슈간 유사도가 높다고 할 수 있다.



〈Figure 5〉 Mapping issues of different periods

이렇게 도출된 이슈간 대응도를 도식화하여 이슈 흐름도를 생성할 수 있다. 이 때 이슈간 대응도의 임계값(Threshold)을 높게 설정하면 이슈간 주요 흐름만 결과로 나타나며, 임계값을 낮게 설정하면 이슈간 대부분의 흐름이 결과로 나타난다. <Figure 6>은 임계값을 0 초과 또는 1 초과로 설정한 경우의 이슈 흐름도 도출 예를 보여준다.



〈Figure 6〉 Generating issue flow diagram

3.3 이슈별 카테고리 식별

본 절에서는 3.2절에서 도출한 기간별 이슈 각각에 대해 주요 카테고리를 식별하는 과정(Phase 2)을 소개한다. 이슈의 연결, 즉 이슈의 흐름은 동일 카테고리 내에서 빈번하게 발생될 것으로 예상되지만 경우에 따라서는 특정 카테고리의 이슈가 다른 카테고리의 이슈로 성질의 변형이 이루어지는 경우도 존재할 수 있다. 이처럼 카테고리간 이슈의 전이가 발생하는 패턴을 분석함으로써 카테고리간의 영향 관계를 파악할 수 있으며, 이를 위해 각 이슈의 주요 카테고리가 무엇인지 식별하는 과정이 필요하다.

각 이슈의 주요 카테고리는 이슈의 카테고리 지수(Category Index, CI)에 의해 식별되며, 각 이슈의 카테고리 지수는 <Figure 4>의 하단 테이블의 "문서/이슈" 대응표를 통해 산출된다. 예를 들어 "Issue A"의 카테고리 "Entertainment"에 대한 카테고리 지수는 식 (1)에 의해 계산된다. 아래 식에서 n은 "Issue A"의 기사의 수를 나타내며 $Count_i(A)$는 i번째 기사의 "Issue A"에 대한 대응 여부를 나타낸다.

$$CI_A^{Entertainment} = \sum_{i=1}^{n} Count_i(A) \qquad (1)$$

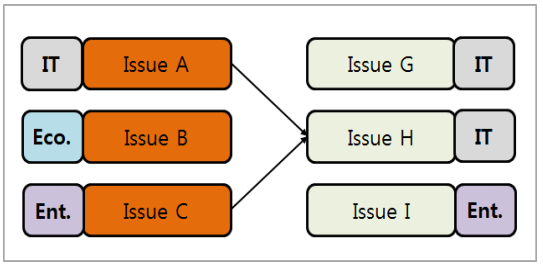
이와 동일한 방식으로 각 이슈의 모든 카테고리에 대한 카테고리 지수를 산출할 수 있으며 가장 지수가 높은 카테고리를 해당 이슈의 카테고리로 식별한다(⑤). 예를 들어 <Figure 7>에서 Issue A를 구성하는 문서 중 IT/Science 카테고리에 속하는 문서는 Doc.2와 Doc.5의 두 가지이며, 따라서 Issue A의 IT/Science 카테고리에 대한 지수는 2로 계산된다. 이와 같은 방식에 따라 각 이슈의 주요 카테고리를 도출하고, 이를 <Figure 6(c)>와 통합하여 도식화한 예가 <Figure 8>에 나타나 있다(⑥). <Figure 8>은 IT 카테고리의 Issue A와 Entertainment 카테고리의 Issue C가 병합하여 IT 카테고리의 Issue H를 형성하였음을 나타낸다.

Period 1	Category	Issue A	Issue B	Issue C
Doc.1	Entertainment	0	1	0
Doc.2	IT/Science	1	0	1
Doc.3	Economy	0	1	0
Doc.4	IT/Science	0	0	1
Doc.5	IT/Science	1	0	1
Doc.6	Economy	1	1	0

Period 2	Category	Issue G	Issue H	Issue I
Doc.7	Economy	0	0	0
Doc.8	Economy	0	1	1
Doc.9	IT/Science	1	1	0
Doc.10	Entertainment	1	0	1
Doc.11	Entertainment	0	0	1
Doc.12	IT/Science	1	1	0

	Period 1			Period 2		
	Issue A	Issue B	Issue C	Issue G	Issue H	Issue I
Entertainment	0	1	0	1	0	2
IT/Science	2	0	3	2	2	0
Economy	1	2	0	0	1	1
Category	*IT/Science*	*Economy*	*IT/Science*	*IT/Science*	*IT/Science*	*Entertainment*

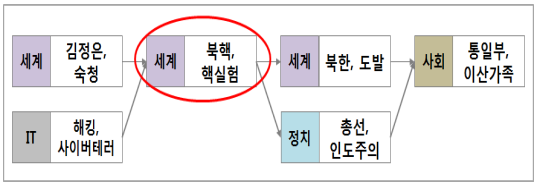
〈Figure 7〉 Calculating index of issue category



〈Figure 8〉 Issue flow diagram with category information

3.2 이슈 흐름도 기반의 주요 분석 시나리오

본 절에서는 앞의 과정을 통해 생성된 이슈 흐름도를 활용한 주요 분석 시나리오 세 가지를 간단한 예를 통해 소개한다. 우선 가장 직관적인 활용으로 특정 이슈의 선행 및 후행 이슈를 분석할 수 있다. 예를 들어<Figure 9>는 (북핵, 핵실험)이라는 이슈의 형성에 영향을 준 선행 이슈, 그리고 해당 이슈의 영향을 받아 형성된 후행 이슈를 보이고 있다. 이러한 분석을 통해 이슈간의 흐름을 이해함으로써 이슈가 생성, 병합, 분화, 소멸되는 동적인 양상을 파악할 수 있다.



〈Figure 9〉 Analyzing preceeding and following issues

두 번째 주요 활용 시나리오로는 전통적 이슈 트래킹을 통해 도출된 장기 이슈의 기간별 세부 이슈 분석을 들 수 있다. 이 분석은 해당 장기 이슈와 제안 방법론을 통해 도출된 각 기간의 모든 세부 이슈와의 대응도 분석을 통해 수행되며, <Figure 10>의 예를 통해 설명될 수 있다.

<Figure 10(a)>는 장기 이슈인 "북한"의 Period1 ~ Period4에 걸친 기간별 문서 수의 분포를 보이고 있다. 한편 <Figure 10(b)>는 Period1의 주요 이슈 세 가지와 이들 이슈를 구성하는 문서 수, 이들 문서 중 <Figure 10(a)>의 문서와 일치하는 문서 수, 그리고 대응도, 즉 (일치 문서 수 / 문서 수)의 비율을 보이고 있다. 특정 임계값 이상의 대응도를 갖는 세부 이슈를 해당 장기 이슈의 기간별 세부 이슈로 식별하게 되며, 식별

Period 1	60
Period 2	40
Period 3	90
Period 4	60

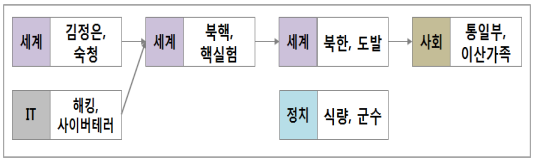
(a) "북한" 이슈의 기간별 문서 분포

	김정은, 숙청	해킹, 사이버테러	전염병
문서 수	50	42	80
일치 문서 수	48	25	8
대응도	96%	59.5%	10%

(b) "북한" 이슈와 PERIOD 1의 세부 이슈간 대응도

〈Figure 10〉 Analyzing correspondence between a selected long term issue and detailed issues

결과는 <Figure 11>과 같은 형태로 나타낼 수 있다. <Figure 11>에 나타난 6개 각각의 이슈는 장기 이슈인 "북한"과의 대응도가 임계값 이상임을 나타내며, 이들 이슈간 매핑 관계가 존재하는 경우 해당 관계도 함께 표시된다.



〈Figure 11〉 Analyzing detailed issues of the selected long term issue

<Figure 11>의 경우 그 형태가 <Figure 9>와 매우 유사하게 나타나지만, 분석 기준 및 목적 측면에서 차이가 있다. <Figure 9>의 경우 특정 세부 이슈와 선행 또는 후행 관계를 갖는 모든 이슈를 표시한 것이고, <Figure 11>의 경우는 세부 이슈가 아닌 특정 장기 이슈와의 대응도가 높은 이슈들을 표시한 것이다. 예를 들어 <Figure 9>에서 (총선, 인도주의) 이슈는 "북한" 이슈와 직접적인 대응도가 낮게 나타나기 때문에 <Figure 11>에서 제외되었고, <Figure 11>의 (식

량, 군수) 이슈는 “북한”과의 대응도는 높지만 (북핵, 핵실험)과의 관계가 형성되지 않았기 때문에 <Figure 9>에서 제외된 것으로 해석될 수 있다.

세 번째 주요 활용 시나리오는 각 이슈에 대한 부가 정보를 활용하여 추가 지식을 창출하는 과정을 보인다. 예를 들어 각 이슈를 구성하는 문서의 조회 수, 댓글 수 등을 추가로 분석하여 해당 이슈에 대한 입체적인 분석을 수행할 수 있으며, 본 연구에서는 부가 정보로 각 이슈를 구성하는 문서의 카테고리 분류 정보를 활용한다. 본 분석의 목적은 특정 카테고리에서 다른 카테고리로 이슈의 전이가 빈번하게 발생하는 현상을 파악하고, 이를 통해 각 카테고리간의 영향 관계를 규명하는 것이다. 카테고리간 이슈의 전이 양상을 정량적으로 측정하기 위해 연관관계 분석(Association Rule Mining)에서 주로 사용되는 신뢰도(Confidence)와 지지도(Support)를 변형하여 사용하고자 한다. 본 분석에서 Period n의 카테고리 Ci와 Period (n+1)의 카테고리 Cj간의 전이 신뢰도(Transition Confidence, T-Conf.)와 전이 지지도(Transition Support, T-Sup.)는 식 (2), (3)과 같이 정의된다. 예를 들어 <Figure 12>에서 P1의 “사회” 카테고리에서 P2의 “IT/과학” 카테고리로의 전이에 대한 전이 신뢰도 및 전이 지지도는 각각 0.2 (= 30/150)와 0.0625 (= 30/480)로 계산된다.

$$T-Conf(C_i,C_j)=\frac{\#\ of\ Links\ from\ C_i\ to\ C_j}{\#\ of\ Links\ from\ C_i} \quad (2)$$

$$T-Sup(C_i,C_j)=\frac{\#\ of\ Links\ from\ C_i\ to\ C_j}{\#\ of\ Links\ from\ All\ Categories\ in\ Period\ n} \quad (3)$$

		P2				
		IT/과학	정치	사회	연예	총합
P1	IT/과학	60	10	20	5	95
	정치	5	70	30	10	115
	사회	***30***	45	55	20	***150***
	연예	5	5	40	70	120
	총합	...				***480***

〈Figure 12〉 Category transition confidence and support

이상 본 장에서는 제안 방법론의 원리 및 주요 분석 시나리오를 간단한 예를 통해 소개하였다. 실제 데이터에 대해 제안 방법론을 적용한 실험 결과는 다음 장에서 소개한다.

4. 실험

4.1 실험 개요

본 장에서는 앞서 제안한 방법론을 실제 텍스트 분석에 적용한 실험 결과를 소개한다. 다양한 유형의 텍스트 문서 중 뉴스 기사는 다른 매체에 비해 비교적 정제가 잘 되어있고 구조화 수준이 높아서 분석에 적합하다는 특징을 갖는다. 따라서 본 실험에서는 2012년 7월부터 2013년 6월의 1년 동안 국내 한 뉴스 포털 사이트에 게시된 문서 중 53,739건을 표본으로 추출하여 분석에 사용하였다. 또한 전체 기간을 3개월씩 분할하여 2012년 7월~9월, 10월~12월, 2013년 1월~3월, 4월~6월의 총 4개의 기간을 설정하고, 각 기간별 이슈의 흐름을 분석하였다.

4.2 이슈 흐름도 도출

앞서 설명한 바와 같이 제안 방법론은 전체 기간이 아닌 세부 기간별로 주요 이슈를 도출한다.

3개월씩 4개로 구분된 각 기간 Q1 ~ Q4 각각에서 도출한 주요 이슈는 <Figure 13>과 같다. 각 기간별로 주요 이슈 25개씩을 도출하였으며, 각 이슈는 5개의 이슈 키워드로 정의된다.

Q1	Q2
가입자,요금,텔레콤,sk텔레콤,서비스	무한,무한도전,도전,특집,예능
드라마,각시탈,시청률,시청자,아랑	전 후보,정치,단일화,캠프,대선
강남,스타일,강남스타일,뮤직비디오,차트	스마트폰,제품,전자,디스플레이,시장
태풍,기상청,강풍,초속,지방	강남스타일,강남,스타일,공연,춤
대선,출마,경선,지지율,박 후보	경찰,형의,사건,조사,흉기
경찰,형의,범행,사건,조사	드라마,연기,시청률,시청자,남자
화영,코어,콘텐츠,미디어,왕따설	주택,대출,가구,은행,금융
무한,무한도전,도전,슈퍼,콘서트	슈퍼스타,심사위원,오디션,생방송,노래
아파트,주택,대출,부동산,은행	박 후보,문 후보,여론조사,오차,조사
올림픽,선수,경기,런던,금메달	보조금,요금,텔레콤,sk텔레콤,가입자
특허,소송,법원,배심원,평결	영화,감독,부산국제영화제,부산,관객
영화,감독,피에타,관객,베니스	결혼,친구,결혼식,여자,열애
온라인,몸매,반응,누리꾼,모습	기온,눈,영하,지방,중부
의원,검찰,대표,원내,국회	도사,무릎팍,무릎팍도사,녹화,황금어장
독도,대통령,정부,방문,이 대통령	선수,구단,시즌,감독,한화
스마트폰,기능,제품,전자,화면	건강,운동,음식,환자,질환
가수,음악,노래,곡,슈퍼스타	몸매,온라인,모습,반응,수지
결혼,열애,친구,교제,남자	음원,팝스타,차트,k팝스타,데뷔
시장,매출,판매,차량,업계	법원,재판,형의,소송,판결
환자,학교,음식,여성,병원	개그,개그콘서트,콘서트,개그맨,상
위원,변호사,택시,불출마,협박	대통령,당선인,의원,위원장,박 당선인
방송,프로그램,예능,제작진,힐링	차량,수입,자동차,엔진,판매
은정,손가락,드라마,다섯손가락,하차	친구,여자,남자,여성,게임
고인,병원,사고,빈소,유족	아이유,팬,주니어,소속사,슈퍼
개그,콘서트,개그콘서트,수지,개그맨	발사,로켓,나로호,장거리,위성

Q3	Q4
후보자,당선인,대통령,박 당선인,장관	주택,부동산,대출,아파트,대책
드라마,하류,수애,도훈,연기	예능,프로그램,사나이,아빠,일밤
스마트폰,제품,디스플레이,화면,전자	대변인,윤,전 대변인,청와대,성추행
경찰,사건,성폭행,경찰서,서부	경찰,수사,사건,경찰서,용의자
연금,소득,국민연금,상품,은행	스마트폰,제품,기능,시장,디스플레이
팝스타,k팝스타,악동,뮤지션,심사위원	젠틀맨,뮤직비디오,강남스타일,신곡,강남
보조금,텔레콤,가입자,요금,sk텔레콤	미사일,무수단,발사,동해안,스커드
아빠,아들,윤후,일밤,지아	결혼,열애,결혼식,커플,연인
기온,영하,지방,눈,기상청	의원,국정원,국회,대선,선거
오연,이준,열애,커플,가상	경기,안타,시즌,투수,선발
고인,빈소,사망,병원,유족	건강,환자,음식,운동,몸
건강,운동,여성,교수,음식	대리점,영업,본사,사원,밀어내기
정글,법칙,프로그램,대표,촬영	드라마,연기,구가,작품,시청률
무한도전,음원,무한,도전,음악	요금,보조금,가입자,무제한,텔레콤
아파트,부동산,주택,전세,가구	힐링,힐링캠프,캠프,어머니,이혼
무릎팍,도사,무릎팍도사,게스트,방송	형의,재판,검찰,공판,소송
남자,결혼,여자,걸그룹,소녀시대	회담,남북,정부,대화,북측
형의,검찰,재판,경찰,징역	은행,시장,상품,금리,금융
핵실험,훈련,도발,정부,대북	지방,기온,중부,지역,기상청
영화,감독,관객,선수,경기	뮤지션,음악,악동,곡,노래
그 겨울,오수,겨울,오영,바람	학교,학생,교사,교육,부모
카드,서비스,신용카드,결제,고객	차량,경찰,자살,고인,여자
제품,시장,매출,판매,소비자	연예병사,안마,시술,세븐,상추
휴가,복무,외박,사병,특혜	대통령,전 대통령,박 대통령,박근혜 대통령,정와대
의원,캠프,힐링,힐링캠프,대선	회장,회사,검찰,기업,사업

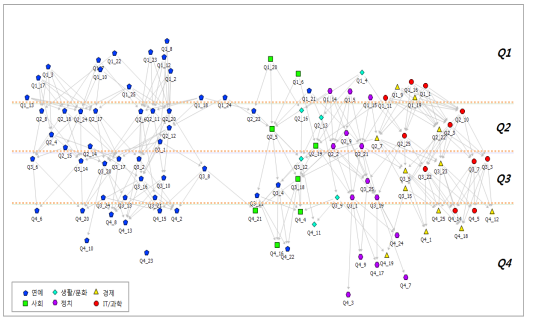
〈Figure 13〉 Results of detailed issues

다음으로 기간별 이슈간 매핑이 이루어진다. 이슈 매핑은 3.2절에서 소개한 기간 중첩에 기반하여 수행된다. 두 기간의 이슈 대응도는 25 X 25의 행렬로 나타나며, 대응도가 높을수록 해당 두 이슈는 관련이 높다고 할 수 있다. 복잡한 행렬에서 의미있는 대응도에 집중하기 위해, 특정 임계값 미만의 값을 갖는 대응도의 값을 '0'으로 재설정하였다. 이렇게 도출된 이슈 대응도 행렬의 일부가 <Figure 14>에 나타나 있다.

Q1 \ Q2	무한, 무한도전, 도전, 특집, 예능	전 후보, 정치, 단일화, 캠프, 대선	스마트폰, 제품, 전자, 디스플레이, 시장	강남스타일, 강남, 스타일, 공연, 춤	경찰, 형의, 사건, 조사, 흉기	...
가입자,요금,텔레콤, sk텔레콤,서비스	0	0	0.181417	0	0	
드라마,각시탈,시청률, 시청자,아랑	0	0	0	0	0	
강남,스타일,강남스타일, 뮤직비디오,차트	0	0	0	0.212739	0	...
태풍,기상청,강풍, 초속,지방	0	0	0	0	0	
대선,출마,경선,지지율, 박 후보	0	0.185364	0	0	0	
...	...					

〈Figure 14〉 Results of issue correspondence matrix (part)

<Figure 14>는 Q1과 Q2간의 이슈 대응도의 일부를 나타낸다. 이러한 방식으로 Q2와 Q3 그리고 Q3와 Q4간의 이슈 대응도를 도출할 수 있으며, 이상 세 개의 이슈 대응도 행렬을 도식화하여 <Figure 15>의 이슈 흐름도를 도출할 수 있다.



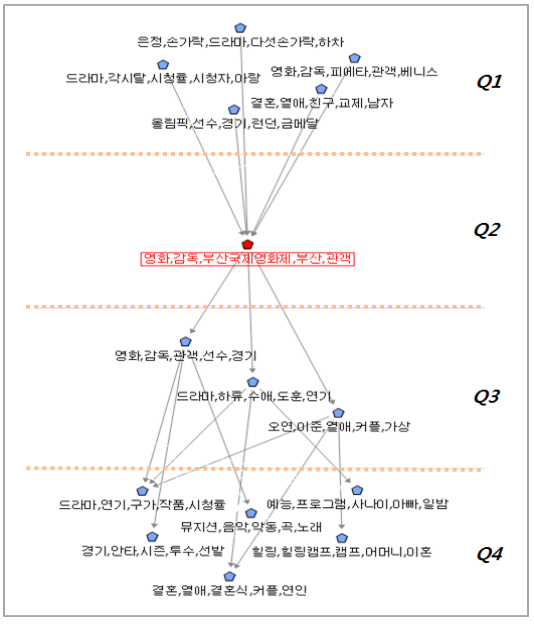
〈Figure 15〉 Results of issue flow diagram

4.3 이슈 흐름도 기반 주요 분석 결과

<Figure 15>의 전체 이슈 흐름도는 제안 방법론의 주요 최종 산출물 중 하나이지만, 많은 이슈와 관계를 포함하고 있어서 전체 내용을 하나의 그림을 통해 설명하기엔 어려움이 있다. 따라

서 <Figure 15>의 이슈 흐름도 자체를 설명하는 대신, 이슈 흐름도를 기반으로 수행되는 주요 추가 분석 결과를 통해 실제 분석 내용을 소개한다.

우선 특정 세부 이슈에 대한 선행 및 후행 이슈 분석 결과를 살펴보자. 예를 들어 Q2의 이슈 중 (영화, 감독, 부산국제영화제, 부산, 관객)의 이슈를 기준 이슈로 선택하고, 이 이슈의 선행 이슈 및 후행 이슈를 추출한 결과가 <Figure 16>에 나타나 있다. 기준 이슈는 영화, 드라마 등 기준 이슈와 직결된 이슈뿐 아니라 올림픽, 결혼 등 다양한 이슈와도 선행 및 후행 관계를 가짐을 알 수 있다.



〈Figure 16〉 Results of preceding and following issue analysis

다음으로 전체 기간에 대한 토픽 모델링을 통해도출된 장기 이슈의 기간별 세부 이슈를 분석하였다. 53,739건의 문서 전체에 대해 토픽 모델링을 수행하여 주요 이슈를 추출한 결과는 <Figure 17>과 같다. 총 25개의 이슈를 도출하였으며, 이들 이슈 각각에 대한 기간별 분포는 본 실험에서는 파악하지 않았다.

전체 기간 이슈	
드라마,시청률,연기,시청자,사랑	힐링,캠프,힐링캠프,소송,이혼
스마트폰,제품,기능,디스플레이,전자	오디션,팝스타,k팝스타,노래,심사위원
대선,문 후보,박 후보,단일화,여론조사	화영,소속사,은정,코어,논란
혐의,검찰,재판,법원,소송	온라인,몸매,반응,모습,눈길
강남스타일,강남,스타일,뮤직비디오,차트	매출,시장,판매,제품,업계
카드,은행,금융,서비스,대출	영화,감독,관객,작품,개봉
태풍,지방,기상청,기온,중부	대변인,윤,청와대,전 대변인,대통령
무한,무한도전,도전,예능,특집	프로그램,예능,방송,무릎팍,도사
보조금,요금,가입자,텔레콤,sk텔레콤	미사일,정부,발사,도발,당국
결혼,열애,결혼식,커플,친구	학교,학생,교사,친구,교육
의원,대통령,국회,당선인,후보자	경찰,차량,사고,경찰서,택시
건강,환자,운동,음식,질환	주택,아파트,부동산,대출,가구
선수,경기,시즌,감독,올림픽	

〈Figure 17〉 Extracting long term issues for the entire period

다음으로 장기 이슈 중 두 가지를 임의로 선정하여 해당 이슈들이 각 기간의 세부 이슈들과 갖는 대응도를 분석하였다. 우선 장기 이슈 (영화, 감독, 관객, 작품, 개봉)이 각 기간 Q1 ~ Q4에 걸친 세부 이슈 총 100개와 갖는 대응도를 분석하였으며, 그 결과의 일부가 <Figure 18>에 나타나 있다. 분석 결과 중 대응도가 15%이상인 이슈들만 추출하고, 이들간의 관계를 도식화한 결과가 <Figure 19>에 나타나 있다. <Figure 19>에서 붉은색으로 진하게 표시된 이슈는 대응도가 70% 이상인 핵심 이슈를 의미한다. 핵심 이슈간의 관계를 보면 장기 이슈인 (영화, 감독, 관객, 작품, 개봉)은 Q1 ~ Q3에 걸쳐 (영화, 감독, 피에타, 관객, 베니스), (영화, 감독, 부산국제영화제, 부산, 관객), 그리고 (영화, 감독, 관객, 선수, 경기)의 세부 이슈로 변화해 왔음을 알 수 있다.

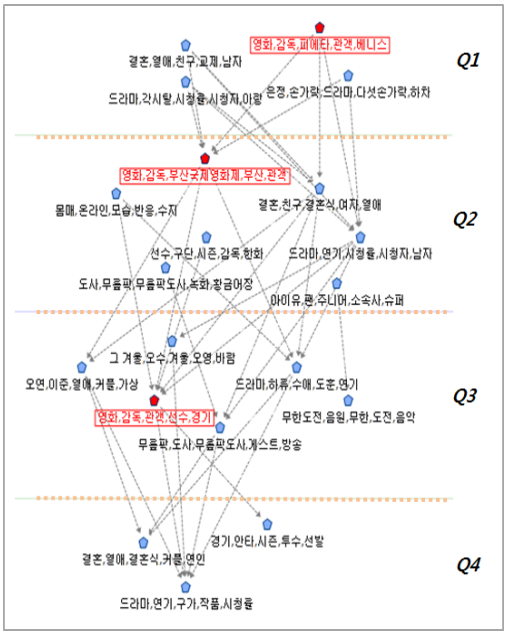
Q1	가입자,요금,텔레콤,sk텔레콤,서비스	드라마,각시탈,시청률,시청자,아랑	강남,스타일,강남스타일,뮤직비디오,차트	태풍,기상청,강풍,조속,지방	대선,출마,경선,지지율,박 후보	경찰,혐의,범행,사건,조사	화영,코어,콘텐츠,미디어,왕따설	무한,무한도전,도전,슈퍼,콘서트	아파트,주택,대출,부동산,은행	올림픽,선수,경기,런던,금메달	특허,소송,법원,배심원,평결	영화,감독,피에타,관객,베니스	...
문서 수	798	1211	648	622	813	1219	413	455	799	764	575	797	
일치 문서 수	6	299	31	1	8	19	9	35	5	77	10	775	
대응도	0.8%	24.7%	4.8%	0.2%	1%	1.6%	2.2%	7.7%	0.6%	10.1%	1.7%	97.2%	

Q2	무한,무한도전,도전,특집,예능	전 후보,정치,단일화,캠프,대선	스마트폰,제품,전자,디스플레이,시장	강남스타일,강남,스타일,공연,춤	경찰,혐의,사건,조사,음기	드라마,연기,시청률,시청자,남자	주택,대출,가구,은행,금융	슈퍼스타,심사위원,오디션,생방송,노래	박 후보,문 후보,여론조사,오차,조사	보조금,요금,텔레콤,sk텔레콤,가입자	영화,감독,부산국제영화제,부산,관객	결혼,친구,결혼식,여자,열애	...
문서 수	509	854	888	642	1154	1137	982	746	596	638	929	1363	
일치 문서 수	58	3	19	41	12	413	9	41	1	3	850	272	
대응도	11.4%	0.4%	2.1%	6.4%	1%	36.3%	0.9%	5.5%	0.2%	0.5%	91.5%	20%	

Q3	후보자,당선인,대통령,박 당선인,장관	드라마,하류,수애,도훈,연기	스마트폰,제품,디스플레이,화면,전자	경찰,사건,성폭행,경찰서,서부	연금,소득,국민연금,상품,은행	팝스타,팝스타,악동,뮤지션,신사위원	보조금,텔레콤,가입자,요금,sk텔레콤	아빠,아들,윤후,일밤,지아	기온,영하,지방,눈,기상청	오연,이준,열애,커플,가상	고인,빈소,사망,병원,유족	건강,운동,여성,교수,음식	...
문서 수	807	970	679	615	793	542	520	562	278	623	451	1168	
일치 문서 수	1	363	8	6	4	41	9	38	12	129	31	59	
대응도	0.1%	37.4%	1.2%	1%	0.5%	7.6%	1.7%	6.8%	4.3%	20.7%	6.9%	5.1%	

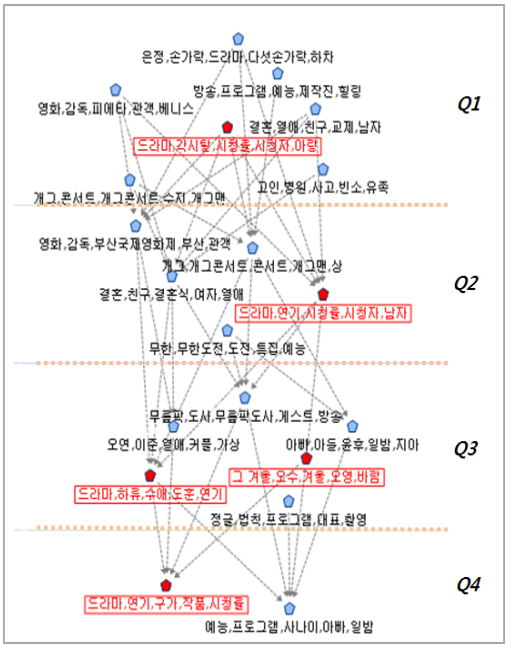
Q4	주택,부동산,대출,아파트,대책	예능,프로그램,사나이,아빠,일밤	대변인,윤,전 대변인,청와대,성추행	경찰,수사,사건,경찰서,용의자	스마트폰,제품,기능,시장,디스플레이	잰틀맨,뮤직비디오,강남스타일,신곡,강남	미사일,무수단,발사,동해안,스커드	결혼,열애,결혼식,커플,연인	의원,국정원,국회,대선,선거	경기,안타,시즌,투수,선발	건강,환자,음식,운동,몸	대리정,영업,본사,사원,떨어내기	...
문서 수	709	1411	475	1727	1017	610	482	1378	1115	434	1153	610	
일치 문서 수	3	191	3	15	34	49	1	268	12	73	47	4	
대응도	0.4%	13.5%	0.6%	0.9%	3.3%	8%	0.2%	19.4%	1.1%	16.8%	4.1%	0.7%	

〈Figure 18〉 Correspondence between the long term issue and detailed issues (movie, director, audiences, product, opening)



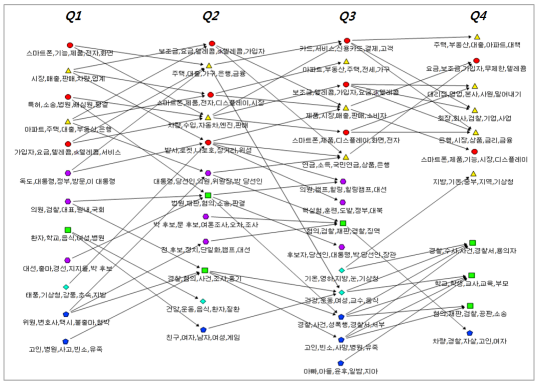
〈Figure 19〉 Results of detailed issue flow diagram (movie, director, audiences, product, opening)

유사 분석을 또 다른 장기 이슈인 (드라마, 시청률, 연기, 시청자, 사랑)에 대해 수행한 결과가 <Figure 20>에 나타나 있다. <Figure 20> 역시 대응도 15% 이상인 이슈들만 나타내고 있으며, 대응도가 70% 이상인 핵심 이슈는 붉은색으로 진하게 나타나 있다. 장기 이슈를 통해 드라마의 시청률, 시청자의 사랑 등이 이슈가 되었음은 알 수 있지만, 구체적으로 어떤 드라마의 어떤 배우에 대한 내용이 이슈가 되었는지는 파악할 수 없다. 하지만 <Figure 20>을 보면 해당 이슈는 Q1 ~ Q4의 기간에 걸쳐 (드라마, 각시탈, 시청률, 시청자, 아랑), (드라마, 연기, 시청률, 시청자, 남자), (그 겨울, 오수, 겨울, 오영, 바람), (드라마, 하류, 수애, 도훈, 연기), 그리고 (드라마, 연기, 구가, 작품, 시청률)의 세부 이슈로 형성된 것임을 파악할 수 있다.



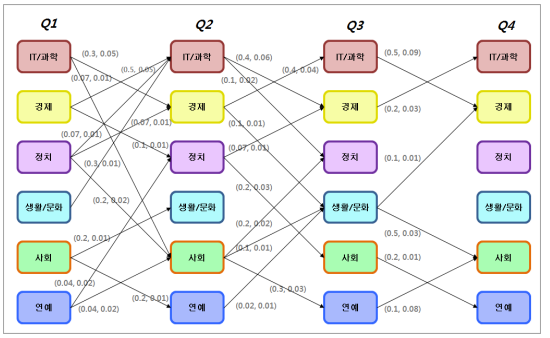
〈Figure 20〉 Results of detailed issue flow diagram (drama, ratings, performance, viewer, love)

마지막으로 이슈 흐름도를 활용하여 카테고리간 이슈 전이 현상을 파악하는 실험을 수행하였다. <Figure 15>의 전체 이슈 흐름도는 모든 이슈의 흐름을 나타내고 있기 때문에 복잡도가 높아 판독이 어려운 측면이 있다. 한편 본 분석의 목표는 상이한 카테고리간의 이슈 전이 현상을 규명하는 것이므로, <Figure 15>에서 동일 카테고리간의 이슈 흐름을 나타내는 관계를 모두 제거하였다. 이를 통해 서로 다른 카테고리간의 이슈 흐름만을 도식화한 결과가 <Figure 21>에 나타나 있다.



〈Figure 21〉 Inter-category issue flows

<Figure 21>을 통해 주로 “IT/과학” 카테고리와과 “경제” 카테고리의 이슈간의 전이 현상이 빈번하게 발생함을 파악할 수 있다. 예를 들어 Q2의 “IT/과학” 이슈인 (스마트폰, 제품, 전자, 디스플레이, 시장)이 Q3에서는 “경제” 이슈인 (아파트, 부동산, 주택, 전세, 가구)와 (제품, 시장, 매출, 판매, 소비자)로 대응됨을 확인할 수 있다. <Figure 21>을 통해 이처럼 개별 이슈의 카테고리 전이 현상을 발견할 수 있지만, 이를 통해 카테고리 차원에서의 이슈 전이 패턴을 발견하기에는 어려움이 있다. 따라서 <Figure 21>의 이슈를 카테고리 기준으로 그룹화하여 추상화하였으며, 그 결과가 <Figure 22>에 나타나 있다. <Figure 22>에서 괄호 내의 두 숫자는 각각 T-Conf.와 T-Sup.를 나타낸다.

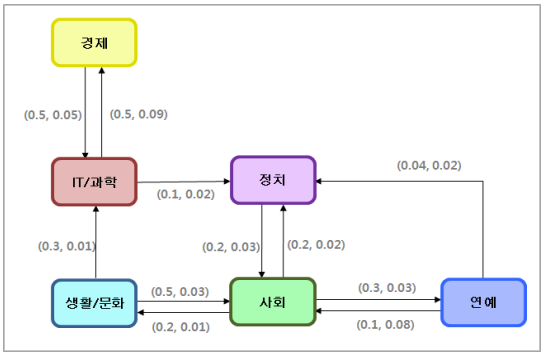


〈Figure 22〉 Patterns of inter-category issue flows

<Figure 22>에서 카테고리간 관계가 강하지 않은 연결, 즉 T-Conf.가 0.1 이하이면서 T-Sup.가 0.01인 연결을 제거하고, 남은 연결 중 각 카테고리 조합에 대해 가장 높은 값의 T-Conf.와 T-Sup.만 표시한 결과가 <Figure 23>에 나타나 있다. <Figure 23>에서 파악되는 이슈 전이는 양방향 전이와 단방향 전이로 구분될 수 있다. 양방향 전이의 경우 양 카테고리에 속하는 이슈의 유사성이 매우 높고, 상호 영향력이 매우 높음을 의미한다. 구체적으로 양방향 전이는 “IT/과학”과 “경제”, “정치”와 “사회”, “생활/문화”와 “사회”, 그리고 “사회”와 “연예” 간에 발생하였음을 확인할 수 있다.

한편 <Figure 23>은 위에서 언급한 양방향 전이 외에 단방향 전이도 포함하고 있다. 단방향 전이의 경우 양방향 전이와 달리 두 카테고리의 유사성이 두드러지게 나타나지 않으며, 한 카테고리에서 다른 카테고리로의 이슈의 흐름이 단방향으로 나타남을 의미한다. 구체적으로는 “생활/문화”에서 “IT/과학”으로, “IT/과학”에서 “정

치"로, 그리고 "연예"에서 "정치"로 단방향의 이슈 흐름이 나타난 것으로 파악되었다.



〈Figure 23〉 Conceptual diagram of inter-category issue flows

5. 결론

본 연구에서는 토픽 모델링을 통해 각 기간별 주요 이슈를 도출하고 기간 중첩을 통해 각 이슈 간 관계를 파악한 후, 이를 이슈 흐름도로 도식화하여 나타내는 방안을 제시하였다. 제안 방법론은 전통적인 이슈 트래킹 분석이 갖는 한계, 즉 이슈가 끊임없이 생성, 소멸, 분화, 병합하는 동적인 변이 과정을 설명하지 못한다는 점과 세부 기간의 구체적인 이슈가 오랜 기간 지속되는 일반적인 이슈를 구성하는 각 기간별 세부 이슈를 설명하지 못한다는 한계를 극복할 수 있을 것으로 기대한다.

또한 본 연구에서는 제안하는 이슈 흐름도를 활용한 세 가지 주요 분석 시나리오를 제시하였다. 즉 이슈 흐름도를 통해 특정 이슈의 선행 및 후행 이슈를 파악할 수 있고, 장기 이슈의 기간별 세부 이슈를 분석할 수 있다. 또한 카테고리 간 이슈가 전이하는 패턴을 발견할 수 있다. 실제 인터넷 뉴스 기사 53,739건에 대해 제안 시나리오에 따른 분석을 수행하였으며, 실험 결과 제안 방법론에 따른 분석을 통해 전통적인 이슈 트래킹의 한계를 보완할 수 있음을 확인하였다.

하지만 본 연구에서 제안하는 방법론이 실무적인 성과로 연결되기 위해서는 다음과 같은 후속 연구가 반드시 필요하다. 우선 대부분의 텍스트 분석을 다루는 연구와 마찬가지로, 정제된 결과를 도출하기 위해서는 양질의 용어사전 및 불용어 사전이 반드시 구축되어야 하며, 문서의 수뿐 아니라 문서의 조회 수, 문서의 댓글 수 등을 활용하여 보다 다면적인 분석을 수행할 필요가 있다. 또한 제안 모형을 검증하기 위한 시도의 일환으로 본 연구에서 장기 이슈와 세부 이슈간의 대응도를 분석하여 제시하였지만, 추후 제안 방법론의 활용 성과에 대한 보다 엄밀한 검증이 이루어질 필요가 있다.

참고문헌(References)

Aggarwal, A., G. Waghmare, and A. Sureka, "Mining Issue Tracking Systems Using Topic Models for Trend Analysis, Corpus Exploration and Understanding Evolution," *Proceedings of the 3rd International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering*, (2014), 52~58.

Albright, R., *Taming Text with the SVD*, SAS Institute Inc, 2006.

Alsumait, L., D. Barbara, and C. Domeniconi, "On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking," *Proceedings of the 8th IEEE International Conference on Data Mining in Data Mining*, (2008), 3~12.

Bae, J. H., N. G. Han, and M. Song, "Twitter

Issue Tracking System by Topic Modeling Techniques," *Journal of Intelligence and Information Systems*, Vol.20, No.2(2014), 109~122.

Han, J., M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd Edition, Morgan Kaufmann Publishers, 2011.

Haribhakta, Y., A. Malgaonkar, and P. Kulkarni, "Unsupervised Topic Detection Model and Its Application in Text Categorization," *Proceedings of the CUBE International Information Technology Conference*, (2012), 314~319.

Hearst, M. A., "Untangling Text Data Mining," *Proceedings of the 37th ACL*, (1999), 3~10.

Jeong, D. H. and M. Song, "Time Gap Analysis by the Topic Model-Based Temporal Technique," *Journal of Informetrics*, Vol.8, No.3(2014), 776~790.

Kim, J., N. Kim, and Y. 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.

Kim, J. C., J. H. Lee, G. J. Kim, S. S. Park, and D. S. Jang, "Data Engineering : Time Series Analysis of Patent Keywords for Forecasting Emerging Technology," *KIPS Transactions on Software and Data Engineering*, Vol.3, No.9(2014), 355~360.

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

Liu, C., and N. Kim, "Individual Interests Tracking : Beyond Macro-level Issue Tracking," *Journal of The Korea Society of IT Services*, Vol.13, No.4(2014), 275~287.

Ma, J., Y. Wang, H. Zhu, and Y. Shen, "Research on Method of Adaptive Topic Tracking Based on Evolution of Public Opinion Ontology," *ACEEE International Journal on Information Technology*, Vol.4, No.1(2014), 1~10.

Mooney, R. J. and R. Bunescu, "Mining Knowledge from Text using Information Extraction," *ACM SIGKDD Explorations*, Vol.7, No.1 (2006), 3~10.

Morinaga, S. and K. Yamanishi, "Tracking Dynamics of Topic Trends Using a Finite Mixture Model," *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2004), 811~816.

Park, J. H. and M. Song, "A Study on the Research Trends in Library & Information Science in Korea using Topic Modeling," *Journal of the Korean Society for Information Management*, Vol.30, No.1(2013), 7~32.

Provost, F. and T. Fawcett, *Data Science for Business*, O'Reilly, 2013.

Rajaraman, K. and A. H. Tan, "Topic Detection, Tracking, and Trend Analysis Using Self-Organizing Neural Networks," *Proceedings of Advances in Knowledge Discovery and Data Mining*, (2001), 102~107.

Salton, G., A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, Vol.18, No.11 (1975), 613~620.

Sebastiani, F., "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, Vol.34, No.1(2002), 1~47.

Sebastiani, F., "Classification of Text, Automatic," *The Encyclopedia of Language and Linguistics,* Vol.14, 2nd Edition, Elsevier Science Pub, 2006.

Stanvrianou, A., P. Andritsos, and N. Nicoloyannis, "Overview and Semantic Issues of Text Mining," *ACM SIGMOD Record*, Vol.36, No.3(2007), 23~34.

Wang, X. and A. McCallum, "Topics Over Time: a Non-Markov Continuous-Time Model of Topical Trends," *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2006), 424~433.

Witten, I. H., *Text Mining, Practical Handbook of Internet Computing*, CRC Press, 2004.

Yu, E., Y. Kim, N. Kim, and S. R. Jeong, "Predicting the Direction of the Stock Index by Using a Domain-Specific Sentiment Dictionary," *Journal of Intelligence and Information Systems*, Vol.19, No.1(2013), 95~110.

Abstract

Investigating Dynamic Mutation Process of Issues Using Unstructured Text Analysis

Myungsu Lim* · Namgyu Kim**

Owing to the extensive use of Web media and the development of the IT industry, a large amount of data has been generated, shared, and stored. Nowadays, various types of unstructured data such as image, sound, video, and text are distributed through Web media. Therefore, many attempts have been made in recent years to discover new value through an analysis of these unstructured data. Among these types of unstructured data, text is recognized as the most representative method for users to express and share their opinions on the Web. In this sense, demand for obtaining new insights through text analysis is steadily increasing. Accordingly, text mining is increasingly being used for different purposes in various fields. In particular, issue tracking is being widely studied not only in the academic world but also in industries because it can be used to extract various issues from text such as news, (SocialNetworkServices) to analyze the trends of these issues. Conventionally, issue tracking is used to identify major issues sustained over a long period of time through topic modeling and to analyze the detailed distribution of documents involved in each issue.

However, because conventional issue tracking assumes that the content composing each issue does not change throughout the entire tracking period, it cannot represent the dynamic mutation process of detailed issues that can be created, merged, divided, and deleted between these periods. Moreover, because only keywords that appear consistently throughout the entire period can be derived as issue keywords, concrete issue keywords such as "nuclear test" and "separated families" may be concealed by more general issue keywords such as "North Korea" in an analysis over a long period of time. This implies that many meaningful but short-lived issues cannot be discovered by conventional issue tracking. Note that detailed keywords are preferable to general keywords because the former can be clues for providing actionable strategies.

To overcome these limitations, we performed an independent analysis on the documents of each detailed period. We generated an issue flow diagram based on the similarity of each issue between two

* Graduate School of Business IT, Kookmin University
** Corresponding Author: Namgyu Kim
School of MIS, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 136-702, Korea
Tel: +82-2-910-5425, Fax: +82-2-910-5209, E-mail: ngkim@kookmin.ac.kr

consecutive periods. The issue transition pattern among categories was analyzed by using the category information of each document.

In this study, we then applied the proposed methodology to a real case of 53,739 news articles. We derived an issue flow diagram from the articles. We then proposed the following useful application scenarios for the issue flow diagram presented in the experiment section. First, we can identify an issue that actively appears during a certain period and promptly disappears in the next period. Second, the preceding and following issues of a particular issue can be easily discovered from the issue flow diagram. This implies that our methodology can be used to discover the association between inter-period issues. Finally, an interesting pattern of one-way and two-way transitions was discovered by analyzing the transition patterns of issues through category analysis. Thus, we discovered that a pair of mutually similar categories induces two-way transitions. In contrast, one-way transitions can be recognized as an indicator that issues in a certain category tend to be influenced by other issues in another category.

For practical application of the proposed methodology, high-quality word and stop word dictionaries need to be constructed. In addition, not only the number of documents but also additional meta-information such as the read counts, written time, and comments of documents should be analyzed. A rigorous performance evaluation or validation of the proposed methodology should be performed in future works.

Key Words : Big Data, Data Mining, Issue Tracking, Text Mining, Topic Modeling, Trend Analysis

Received : November 25, 2015 Revised : January 19, 2016 Accepted : February 9, 2016
Publication Type : Regular Paper Corresponding Author : Namgyu Kim

저 자 소 개

임 명 수
원광대학교 정보·전자상거래학부에서 학사 학위를 취득하였으며, 현재 국민대학교 비즈니스IT전문대학원 비즈니스IT전공 석사과정에 재학 중이다. 주요 관심 분야는 데이터 마이닝, 텍스트 마이닝, 소셜 미디어 마이닝 등이다.

김 남 규
현재 국민대학교 경영정보학부에서 부교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국지능정보시스템학회 부회장, 한국정보기술응용학회 부회장, 한국경영정보학회 이사, 한국CRM학회 이사, 한국IT서비스학회지 편집위원, JITAM 편집위원, 한국인터넷정보학회논문지 편집위원을 역임하였으며, 한국생산성본부 TOPCIT 개발사업 출제 및 자문위원으로 활동하였다. 주요 관심분야는 텍스트 마이닝, 데이터 마이닝 및 데이터베이스 설계 등이다.