ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2014 December: 20(4): 25~41 http://dx.doi.org/10.13088/jiis.2014.20.4.25

기간별 이슈 매핑을 통한 이슈 생명주기 분석 방법론

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

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

최근 스마트 기기를 통해 소셜미디어에 참여하는 사용자가 급격히 증가하고 있다. 이에 따라 빅데이터 분석에 대한 관심이 높아지고 있으며 최근 포털 사이트에서 검색어로 자주 입력되거나 다양한 소셜미디어에서 자주 언급되는 단어에 대한 분석을 통해 사회적 이슈를 파악하기 위한 시도가 이루어 지고 있다. 이처럼 다량의 텍스트를 통해 도출된 사회적 이슈의 기간별 추이를 비교하는 분석을 이슈 트래킹이라 한다. 하지만 기존의 이슈 트래킹은 두 가지 한계를 가지고 있다. 첫째, 전통적 방식의 이슈 트래킹은 전체 기간의 문서에 대해 일괄 토픽 분석을 실시하고 각 토픽의 기간별 분포를 파악하는 방식으로 이루어지므로, 새로운 기간의 문서가 추가되었을 때 추가된 문서에 대해서만 분석을 추가 실시하는 것이 아니라 전체 기간의 문서에 대한 분석을 다시 실시해야 한다는 실용성 측면의 한계를 갖고 있다. 둘째, 이슈는 끊임 없이 생성되고 소멸될 뿐 아니라, 때로는 하나의 이슈가 둘 이상의 이슈로 분화하고 둘 이상의 이슈가 하나로 통합되기도 한다. 즉, 이슈는 생성, 변화(병합, 분화), 그리고 소멸의 생명주기를 갖게 되는데, 전통적 이슈 트래킹은 이러한 이슈의 가변성을 다루지 않았다는 한계를 갖는다. 본 연구에서는 이러한 한계를 극복하기 위해 대상 기간 전체의 문서를 한꺼번에 분석하는 방식이 아닌 세부 기간별 문서에 대해 독립적인 분석을 수행하고 이를 통합할 수 있는 방안을 제시하였으며, 이를 통해 새로운 이슈가 생성되고 변화하며 소멸되는 전체 과정을 규명하였다. 또한 실제 인터넷 뉴스에 대해 제안 방법론을 적용함으로써, 제안 방법론의 실무 적용 가능성을 분석하였다.

주제어 : 이슈 생명주기, 이슈 매핑, 이슈 트래킹, 텍스트 마이닝, 토픽 분석

논문접수일 : 2014년 11월 12일 논문수정일 : 2014년 12월 6일 게재확정일 : 2014년 12월 7일
투고유형 : 국문급행 교신저자 : 김남규

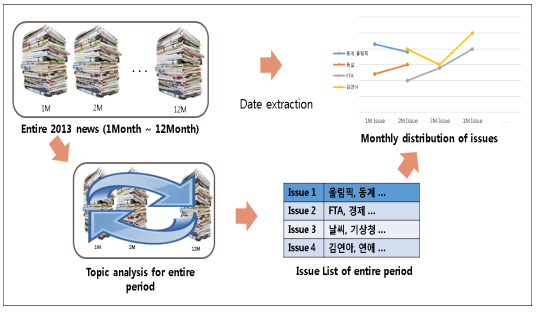
1. 서론

최근 스마트 기기를 통해 소셜미디어에 참여하는 사용자가 급격히 증가함에 따라 실시간으로 생성, 저장, 공유되는 정보의 양이 기하급수적으로 증가하고 있다. 또한 단순히 정보의 양만 증가한 것이 아니라, 기존의 전통적인 방법으로는 분석이 불가능했던 비정형 데이터에 대한 분석이 가능해짐에 따라, 최근 다양한 유형의 비정형 데이터에 대한 분석을 통해 새로운 부가가치를 창출하고자 하는 관심, 즉 빅데이터(Big Data) 분석에 대한 관심이 꾸준히 높아지고 있다. 빅데이터 관련 기술은 향후 수년 내에 IT분야에서 주요 기술로 자리잡을 것으로 예상되어 왔으며 (Gartner, 2012), 이러한 예상은 학계와 산업계 전반에서 이미 실현되고 있다.

이처럼 빅데이터 분석에 대한 관심이 급증한 원인 중 하나는 다양한 소셜미디어를 통해 유통

되는 텍스트 데이터의 양이 급격히 증가한 것에서 찾을 수 있다. 다양한 비정형 데이터 중 특히 텍스트 데이터는 소셜미디어, 인터넷 뉴스, Wikipedia 등에서 정보의 표현 수단으로 널리 사용되고 있다. 즉, 대부분의 사람은 텍스트를 통해 가장 많은 지식을 표현하고 전달하고 습득하기 때문에, 텍스트 분석을 통해 새로운 부가가치를 창출하고자 하는 시도가 증가하는 것은 매우 당연하다고 할 수 있다. 이와 같이 텍스트 형태의 비정형 데이터에 대한 분석을 다루는 분야를 텍스트 마이닝(Text Mining)이라고 하며, 빅데이터 분석의 다양한 분야 중 가장 관심이 많은 한 축을 이루고 있다. 텍스트 마이닝은 오피니언 마이닝(Opinion Mining), 소셜 네트워크 분석(Social Network Analysis) 등 다양한 분석 기법들과 연결되어 사용되기도 한다. 특히 최근에는 포털 사이트에서 검색어로 자주 입력되거나 소셜미디어에서 자주 언급되는 단어에 대한 분석을 통해 사회적 이슈를 분석하여 제시하기 위한 시도가 활발하게 이루어지고 있다. 이러한 서비스들은 텍스트 분석을 통해 사람들의 관심사를 발견하고 이를 시각화하여 제시하는 서비스를 제공하고 있으며, 대표적 예로는 네이버사의 '네이버 트렌드(http://trend.naver.com)', 구글사의 '구글 트렌드(http://www.google.com/trends)' 등이 있다. 이처럼 다량의 텍스트로부터 이슈, 주제 등의 토픽을 도출하는 기법을 토픽 분석(Topic Analysis) 이라 하며, 토픽 분석을 통해 도출한 사회적 이슈의 기간별 추이를 비교하는 분석을 이슈 트래킹(Issue Tracking)이라 한다.

특정 기간 동안의 검색 빈도에 기반하여 특정 토픽의 생성과 소멸을 보여주는 이슈 트래킹은 구글이나 네이버 등 많은 검색 포털 사이트들을 중심으로 서비스되고 있으며, 이러한 전통적 방식의 이슈 트래킹은 전체 기간의 문서에 대한 토픽 분석을 실시하여 주요 토픽을 식별하고, 각 토픽에 해당되는 문서들의 기간별 출현 빈도를 비교 분석하는 방식으로 수행된다. 하지만 이러한 방식은 새로운 기간의 문서가 추가되었을 때 추가된 문서에 대해서만 분석을 추가 실시하는 것이 아니라, 전체 기간의 문서에 대한 분석을 다시 실시해야 하므로 시간 및 비용의 부담이 크다는 한계를 갖고 있다. 따라서 이러한 방식은 지속적으로 추가 기간에 대한 분석을 수행해야 하는 대부분의 응용에 적용되기 어렵다<Figure 1>.



〈Figure 1〉 Traditional Issue Tracking Process

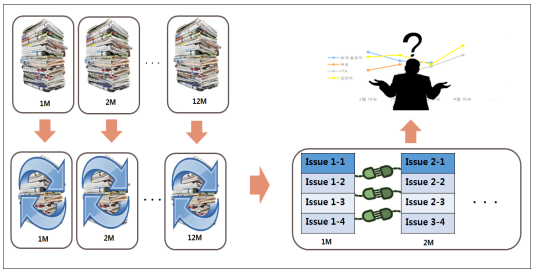
<Figure 1>은 전통적인 이슈 트래킹의 과정을 도식화하여 보이고 있다. 즉, 2013년 1월부터 2013년 12월 사이 1년 간의 전체 뉴스에 대한 토픽 분석을 통해 1년 간의 주요 이슈를 찾아내고, 각 이슈에 대응되는 기사의 기간별 빈도수를 찾아 이슈의 월별 분포를 분석한다. 하지만 만약 분석 기간을 2014년 1월까지로 확장하고자 하는 경우, 2013년 1월부터 2014년 1월까지의 13개월에 걸친 전체 기간의 뉴스에 대한 토픽 분석을 새로 실시해야 한다는 부담이 있다.

전통적 이슈 트래킹의 한계는 위와 같은 분석 방법 측면에서뿐 아니라 이슈의 일시성을 충분히 반영하지 못했다는 측면에서도 찾을 수 있다.

즉, <Figure 1>을 예로 들면, 모든 이슈가 2013년 1월부터 2013년 12월까지 동일 기간 존재한 것으로 가정할 수 없다는 것이다. 어떤 이슈는 2013년 12월에 새로 생성되었을 것이며, 이 이슈는 12월 한 달 동안은 매우 큰 이슈로 인식되었을 수 있지만 1년이라는 전체 기간으로 보면 큰 이슈에 속하지 못했을 수도 있다. 이 경우 해당 이슈는 <Figure 1>의 이슈의 월별 분포 그래프에서 아예 나타나지 않게 된다.

위에서 설명한 이슈의 일시성은 이슈의 가변성의 한 가지 특수한 형태로 설명될 수 있다. 즉, 이슈는 끊임없이 생성되고 소멸될 뿐 아니라, 때로는 하나의 이슈가 둘 이상의 이슈로 분화하고 둘 이상의 이슈가 하나로 통합되기도 한다. 따라서 이슈는 생성, 변화(병합, 분화), 그리고 소멸의 생명주기를 갖게 되는데, 전통적인 이슈 트래킹은 이러한 이슈의 가변성을 다루지 않았다는 한계를 갖는다.

본 연구의 목적은 이처럼 전통적인 이슈 트래킹의 두 가지 한계, 즉 분석 방법 측면의 한계와 이슈의 가변성을 고려하지 않았다는 한계를 극복하는 데 있다. 이를 위한 대안으로 우선 기간을 작게 나누어서 기간별로 토픽 분석을 실시하고, 그 결과를 통합하여 전체 기간에 대한 이슈 흐름을 보이는 방법을 고려할 수 있다. 하지만 이러한 방법은 각 기간별로 도출된 이슈들을 서로 연결할 수 없기 때문에, 전체 기간에 걸친 이슈의 흐름을 살펴보기 어렵다는 한계를 갖는다 <Figure 2>.



〈Figure2〉 Problem of Issue Mapping

<Figure 2>는 2013년 1월부터 2013년 12월까지의 뉴스에 대해 각 월별로 토픽 분석을 실시한 결과를 보여주고 있다. 이슈의 추이를 보기 위해서는 1월의 이슈들(이슈1-1 ~ 이슈 1-3)과 2월의 이슈들(이슈 2-1 ~ 이슈 2-3) 간의 매핑이 이루어져야 한다. 하지만 각기 독립적으로 수행한 토픽 분석의 결과로 도출된 이슈들간의 비교를 위한 직접적인 방법은 존재하지 않으므로, 이를 간접적으로 수행하기 위한 새로운 기법이 고안되어야 한다.

본 연구에서는 이를 위해 서로 다른 기간에 대해 독립적으로 수행된 토픽 분석의 결과를 통합하기 위한 방안을 제시하고자 한다. 또한 이를 통해 각 기간별 이슈의 생성, 변화(병합, 분화), 그리고 소멸을 도식화하여 나타냄으로써 이슈 생명주기를 분석할 수 있는 방법론을 제안하고자 한다. 제안 방법론은 새로 추가되는 기간에 대한 문서만을 분석하여 기존 결과에 통합하기 때문에 시간과 비용 측면에서 매우 효율적일 뿐 아니라, 이슈의 가변성을 충분히 고려하기 때문에 현실 상황에 보다 부합된 결과를 보여줄 수 있을 것으로 기대한다.

본 연구의 이후 구성은 다음과 같다. 다음 장인 2장에서는 본 연구와 관련된 텍스트 마이닝, 토픽 분석, 이슈 트래킹, 이슈 매핑에 대한 선행연구들을 요약하고, 3장에서는 본 연구의 핵심인 이슈 매핑 방안과 이슈 생명주기 분석 방법론을 보다 구체적으로 소개한다. 또한 4장에서는 제안 방법론을 실제 데이터에 적용한 실험 결과를 살펴보고, 마지막 5장에서는 본 연구의 결론과 기여 및 한계, 그리고 후속 연구 방향을 제시한다.

2. 관련 연구

2.1 텍스트 마이닝 및 토픽분석

텍스트는 현실에서 정보를 표현하고 교환하는 가장 대표적인 수단으로(Witten, 2004), 방대한 양의 지식이 텍스트 형태로 저장되어 왔다. 이에 따라 학계와 산업계 전반에서 텍스트 분석을 통해 의미 있는 정보를 발굴하기 위한 노력을 끊임없이 기울여왔다. 텍스트 마이닝은 이러한 노력의 대표적인 성과 중 하나로, 대량의 텍스트 데이터를 분석해 의미 있는 정보를 추출하는 과정으로 정의된다. 특히 최근에는 대량의 데이터에 대한 텍스트 마이닝 분석을 통해 새로운 가치를 창출하기 위한 시도가 다양한 분야(Hong et al., 2014; Mooney and Bunescu, 2006; Song et al., 2009; Yu et al., 2012)에서 이루어지고 있다. 구체적으로 특정 기사의 원문을 파악하기 위한 연구(Metzler et al., 2005), 특정 범죄와 다른 범죄들 간의 유사성 측정을 통해 새로운 범죄를 발견하기 위한 연구(Fan et al., 2006), 텍스트 범주화를 통해 비구조적 저장소를 구조화하기 위한 연구(Salton et al., 1975), 2012년 대선 당시의 트위터 데이터를 수집하여 각 후보 별 이슈를 분석 연구(Bae et al., 2013), 토픽 분석을 활용하여 하나의 카테고리만 부여되어 있는 문서에 둘 이상의 카테고리를 자동으로 부여하는 연구(Hong et al., 2014), 사용자 관점에서 이슈를 클러스터링하여 상위 이슈를 도출하기 위한 연구(Kim et al., 2014) 등을 텍스트 마이닝 관련 연구의 대표적 예로 들 수 있다.

텍스트 마이닝은 데이터 마이닝을 포함한 자연어 처리, 정보 검색, 전산 언어학, 토픽 추적 등 분야의 기술을 포괄적으로 활용한다(Mooney and Bunescu, 2006; Rijsbergen, 1979; Sebastiani, 2006). 특히 비정형 데이터를 행렬, 계층, 벡터 등의 형식으로 정형화하기 위한 기술인 자연어 처리(Weiss et al., 2010) 는 텍스트 마이닝 분석 결과의 성패를 결정 짓는 중요한 핵심 기술 이라고 할 수 있다. 텍스트 데이터의 정형화 과정에서는 기본적으로 각 문서에 사용된 용어의 빈도에 따라 문서의 주제 및 특성을 요약하는 벡터공간모델(Vector Space Model)(Albright, 2006; Salton et. al., 1975)이 사용된다. 용어의 빈도수 측정을 위한 다양한 지표가 개발되어 왔으며, 이들 중 특히 TF-IDF(Term Frequency-Inverse Document Frequency)(Han and Kamber, 2011)가 토픽분석을 비롯한 다양한 응용에서 널리 사용된다. TF-IDF는 어떤 문서 A 에서 용어 X와 Y가 동일한 빈도수로 출현했을 때, Y는 문서 A에만 출현하는 반면 X는 다른 문서들에도 자주 출현했다면 A 문서에서 용어 Y는 용어 X보다 더 중요한 키워드로 인정받아야 한다는 인식을 반영한 척도이다. 빈도수 기반 분석에서 각 문서는 용어 수만큼 차원을 갖게 되며 “(문서 수) × (용어 수)”로 표현된 행렬의 각 셀에 각 문서에서 해당 용어가 나타난 빈도수를 기재함으로써 모든 문서를 행렬화할 수 있다. 즉, 토픽분석은 각 문서를 어떤 이슈들의 집합으로 가정하고, 그 문서를 구성하는 용어들의 중요도를 확률적으로 계산한 뒤, 그 결과 값을 키워드의 집합으로 도출하는 일련의 알고리즘(Bae et al., 2014)이라 할 수 있다.

2.2 이슈 트래킹

소셜미디어의 보급으로 방대한 양의 텍스트 데이터가 생성되면서 텍스트 데이터의 활용에 대한 관심이 급증하였다. 특히 빅데이터 안에서

사람들이 자주 언급하는 경제, 정치, 스포츠 등 사회 전반에 걸친 이슈들을 찾아내려는 연구(Ding and Chen, 2014)가 활발하게 이루어지고 있다. 이슈 트래킹은 토픽분석을 기반으로 기간별 사회적 이슈를 추적하는 방법론으로, 일반적으로 그 결과로 세부 기간별 주요 이슈를 시각화하여 나타낸다. 대부분의 이슈 트래킹 관련 연구는 주로 뉴스 데이터(Ma et al., 2014; Jin et al., 1999) 또는 트위터 데이터(Bae et al., 2014)를 사용하여 이루어지고 있으며, 국내에서도 대선 후보별 이슈를 분석하는 연구(Bae et al., 2013), 트위터 데이터를 사용하여 이슈를 트래킹 하고 이를 웹 상에서 시각화하는 TITS(Twitter Issue Tracking System)을 설계하는 연구(Bae et al., 2014), 학술지 및 도서관 논문의 초록 데이터를 분석하여 문헌정보학 분야의 연구동향을 규명하는 연구(Park and Song, 2013) 등의 연구가 활발하게 이루어지고 있다.

이슈 트래킹에 대한 다양한 시도는 이슈가 고정되어 있지 않고 시간의 흐름에 따라 생성되고 변화하기 때문에 존재 가치가 있다. 예를 들어 여름에 생성된 '폭염' 이슈는 장마 기간이 되면 '장마'라는 이슈로 변할 수도 있고, '패션'이라는 이슈와 결합하여 '여름 패션'이라는 새로운 이슈를 만들어 낼 수도 있다. 이러한 변화를 거친 '폭염' 이슈는 가을이 되면 자연스럽게 소멸된다. 이처럼 이슈는 고정적인 것이 아니라, 시간의 흐름에 따라 생성, 변화(분화, 병합), 소멸하는 생명주기(Life Cycle)를 가지고 있다. 하지만 기존의 이슈 트래킹 연구는 분석 대상 기간 전체의 문서로부터 한꺼번에 주요 이슈를 추출하고, 이 이슈의 세부 기간별 분포를 파악하고 비교하는 방식으로 이루어진다. 이러한 방식은 전체 기간에 대한 정적인 토픽 분석의 결과를 세부 기간별로 나누어 정리한 것으로, 이슈의 생명 주기를 따라 지속적으로 추적하는 진정한 트래킹으로 보기 어렵다. 즉, 대상 기간 전체의 문서를 한꺼번에 분석하는 방식이 아니라 새로 확장된 기간의 문서만을 추가적으로 분석할 수 있는 방안이 필요하며, 이를 통해 새로운 이슈가 생성되고 변화하며 소멸되는 전체 과정을 규명할 수 있는 연구가 반드시 필요하다.

3. 기간별 이슈 매핑을 이용한 이슈 생명주기 분석 방법론

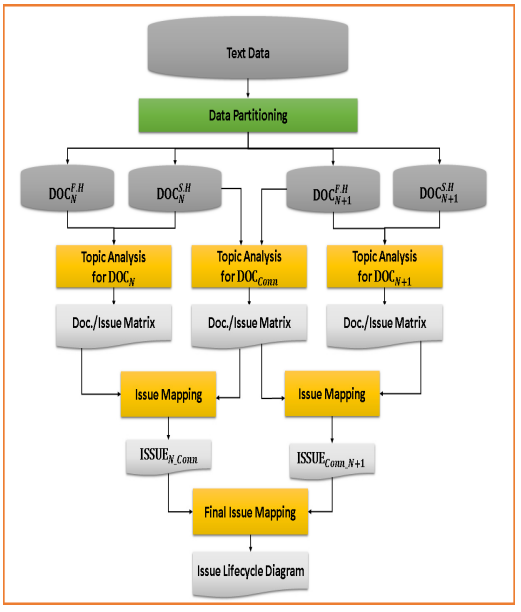
3.1 연구 모형

본 장에서는 기간별 이슈의 추출과 매핑 그리고 이슈의 생명주기를 분석하는 이슈 생명주기 분석(Issue Lifecycle Analysis, ILCA) 방법론의 개념 및 주요 과정을 소개한다. 전체적인 방법론의 개요를 본 절에서 <Figure 3>을 통해 제시하며, 각 주요 모듈에 대한 상세한 설명은 이후 절에서 각각 다루도록 한다.

우선 분석 대상 기간에 속한 전체 문서를 세부 기간별로 분할한다. <Figure 3>에서 세부 기간 N에 속한 문서의 집합을 DOC_N으로 표시하며, 세부 기간 (N+1)에 속한 문서의 집합을 DOC_{N+1}로 표시한다. 물론 더 많은 수의 세부 기간으로 분할할 수 있지만, 본 장에서는 두 기간의 이슈를 매핑하는 과정만을 예로 들어 방법론을 소개한다. 이슈의 흐름을 파악하기 위해서는 DOC_N에서 도출된 이슈와 DOC_{N+1}에서 도출된 이슈를 서로 매핑하는 과정이 반드시 필요하며, 이는 제안 방법론의 핵심 과정이다. 하지만 두 문서 집합 간에는 서로 교집합이 존재하

지 않으므로, 두 기간의 이슈 간 유사성을 자동으로 식별할 수 있는 방안이 존재하지 않는다.

따라서 두 기간의 이슈 간 유사성을 비교하기 위한 연결고리가 필요하며, 이를 위해 제안 방법론에서는 각 기간의 문서를 일부 발췌하고 통합하여 연결고리 문서 집합을 생성한다. 즉, DOC_N을 기간에 따라 전반기 문서($DOC_N^{F.H}$)와 후반기 문서($DOC_N^{S.H}$)로 구분하고 DOC_{N+1}을 기간에 따라 전반기 문서($DOC_{N+1}^{F.H}$)와 후반기 문서($DOC_{N+1}^{S.H}$)로 구분한 뒤, $DOC_N^{S.H}$와 $DOC_{N+1}^{F.H}$를 통합하여 DOC_N과 DOC_{N+1}의 연결고리 문서인 DOC_{Conn}을 임시로 생성한다. 그 결과로 생성된 3개의 문서 집합은 약 50%의 문서를 다른 문서 집합과 공유하게 되며, 이를 근거로 각 문서집합에서 도출된 이슈의 유사도 측정이 이루어지게 된다.



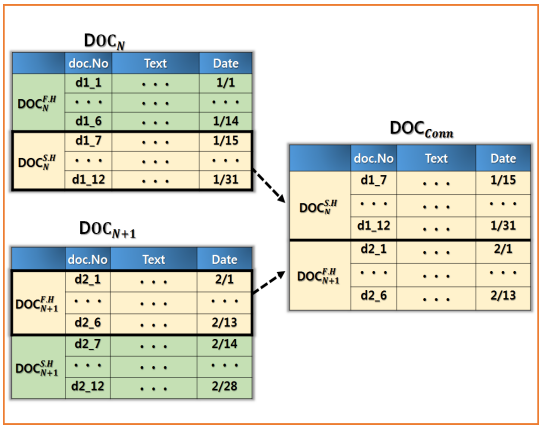
〈Figure 3〉 Research overview

다음으로 DOC_N과 DOC_{Conn} 각각의 토픽분석 결과로 도출된 이슈 집합을 통합하여 $ISSUE_{N\,Conn}$을 생성하고, DOC_{N+1}과 DOC_{Conn}에서 도출된 이슈 집합을 통합하여 $ISSUE_{N+1\,Conn}$을 생성한다. 마지막으로 $ISSUE_{N\,Conn}$과 $ISSUE_{N+1\,Conn}$의 매핑을 통해 최종 산출물인 이슈 생명주기도(Issue Lifecycle Diagram)을 도출한다.

3.2 세부 기간별 데이터 분할

전통적 이슈 트래킹과 달리 본 연구는 기간별 문서에 대해 독립적으로 토픽 분석을 실시하므로, 분석 대상 기간의 전체 문서를 여러 세부 기간으로 분할하는 작업을 분석 첫 단계에서 실시한다. 본 연구에서는 세부 기간 간격을 한 달로 설정하였으나, 이 기간 간격은 분석 목적에 따라 유연하게 설정할 수 있다. 이후 과정에서 DOC_N은 1월달의 문서 집합을, DOC_{N+1}은 2월달의 문서 집합을 의미하는 것으로 가정한다. 두 기간의 연결고리 문서 집합을 생성하는 과정이 <Figure 4>에 나타나있다.

<Figure4>는 문서 집합 DOC_N의 후반인



〈Figure 4〉 generating station Document DOC_{Conn}

$DOC_N^{S.H}$와 문서 집합 DOC_{N+1}의 전반인 $DOC_{N+1}^{F.H}$을 합성하여 연결고리 문서 집합 DOC_{Conn}을 생성하는 예이다. <Figure 4>에 나타난 세 개의 문서 집합 각각에 대해 독립적으로 토픽분석을 수행하게 되며, 이 과정은 다음 절에서 소개한다.

3.3 각 문서 집합별 토픽분석

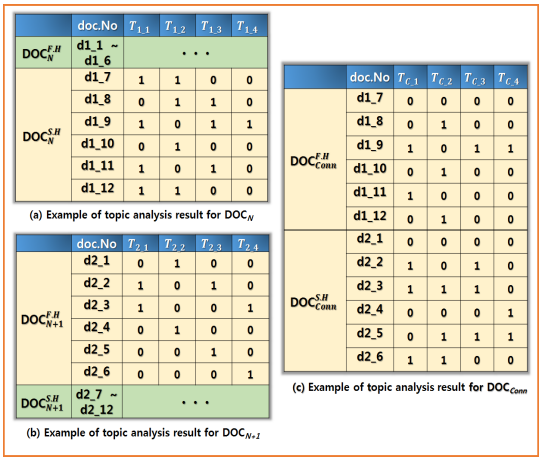
본 절에서는 토픽분석을 통해 주요 이슈를 도출하고, 각 이슈와 문서들간의 대응 관계를 매트릭스로 나타내는 과정을 소개한다. 토픽분석은 많은 연구 및 서적에서 이미 소개되었을 뿐 아니라 상용 분석 도구를 통해 쉽게 수행 가능하기 때문에 본 연구에서는 자세한 과정은 다루지 않는다.

토픽분석을 통해 도출되는 문서/이슈 대응 매트릭스는 0/1의 이진 값 또는 0에서 1사이의 연속 값으로 표현될 수 있다. 연속 값으로 나타내는 경우 각 값은 해당 문서가 해당 이슈에 대응되는 정도를 나타내며, 이진 값으로 나타내는 경우는 해당 문서가 해당 이슈에 포함되는지의 여부를 나타낸다. <Figure 5>는 <Figure 4>의 세 문서 집합에 대해 가상의 토픽분석을 수행한 결과를 이진 값 표기법으로 나타낸 예를 보여준다. 예를 들어 <Figure 5(a)>에서 문서 d1_7은 이슈 T_{1_1}과 T_{1_2}의 이슈에 포함됨을 알 수 있다.

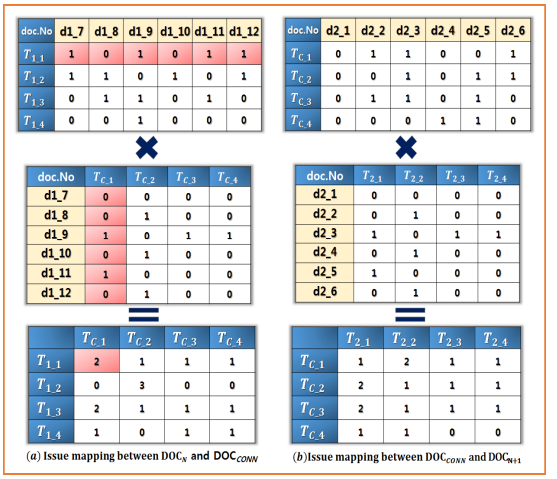
<Figure 5(a)>는 DOC_N의 총 12개 문서에 대한 이슈 대응 관계를 나타내지만 $DOC_N^{F.H}$에 속하는 6개의 문서는 이후 매핑 과정에서 사용되지 않으므로 문서의 이슈 대응 값 표기를 생략하였으며, <Figure 5(b)>의 경우는 $DOC_{N+1}^{S.H}$에 속하는 6개 문서의 이슈 대응 값이 동일한 이유로 생략되었다. 한편 <Figure 5(c)>에서는 DOC_{Conn}에 속한 12개 문서 모두의 이슈 대응 값이 명시되어 있다. 이들 세 매트릭스를 사용한 이슈 매핑 과정은 다음 절에 소개되어 있다.

3.4 문서 집합 간 이슈 매핑

본 절에서 다룰 문서 집합 간 이슈 매핑은 세



〈Figure 5〉 Example of Generating Document/Issue Correspondence Matrix



〈Figure 6〉 Issue Mapping Using DOCConn

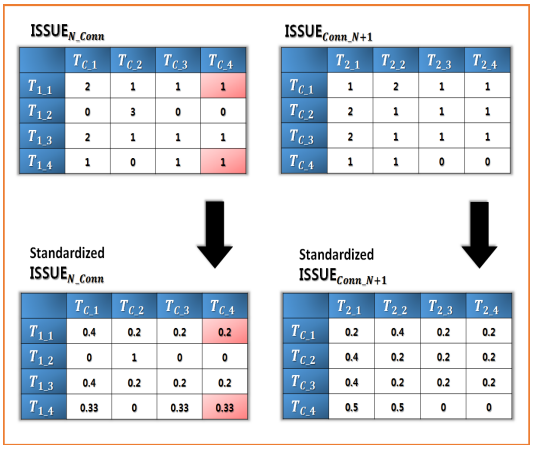
부 기간/연결고리 간 이슈 매핑, 이슈 대응도 표준화, 그리고 최종 이슈 매핑의 세 단계로 구성되어 있다. <Figure 6>은 <Figure 5>의 세 개의 매트릭스를 이용하여 기간/연결고리 간 이슈 매핑을 수행하는 과정을 보여준다.

두 이슈 간 대응도는 각 이슈에 속한 모든 문서에 대해 각 문서의 두 이슈에 대한 대응도의 가중합으로 계산된다. 즉, 식 (1)에서 Match(T_a, T_b)는 이슈 T_a와 이슈 T_b의 대응도를 나타내며, n은 공통 문서의 수를, C_i^a와 C_i^b는 각각 문서 i가 이슈 T_a와 이슈 T_b에 대해 갖는 대응도를 의미한다.

$$Match(T_a, T_b) \sum_{i=1}^{n} (C_i^a \times C_i^b) \quad (1)$$

위의 식에 의하면 두 이슈 T_a와 이슈 T_b에 동시에 높은 대응도를 가진 문서가 많을수록 두 이슈의 대응도인 Match(T_a, T_b)가 높게 나타남을 알 수 있다. 이와 같은 방식으로 두 세부 기간의 모든 이슈 간 대응도를 위 식에 의해 산출해야 하며, 이 과정은 행렬 곱에 의해 쉽게 구현될 수 있다. <Figure 6(a)>는 DOC_N과 DOC_{Conn}간의 이슈 매핑 과정을, <Figure 6(b)>는 DOC_{Conn}과 DOC_{N+1}간 이슈 매핑 과정을 보여준다. <Figure 6(a)>의 최상단 매트릭스는 <Figure 5(a)>의 매트릭스의 전치 행렬(Transposed Matrix)이며, <Figure 6(a)>의 중간 매트릭스는 <Figure 5(c)>의 상단 6개 문서에 대한 매트릭스이다. 한편 <Figure 6(b)>의 최상단 매트릭스는 <Figure 5(b)>의 매트릭스의 전치 행렬이며, <Figure 6(b)>의 중간 매트릭스는 <Figure 5(c)>의 하단 6개 문서에 대한 매트릭스이다.

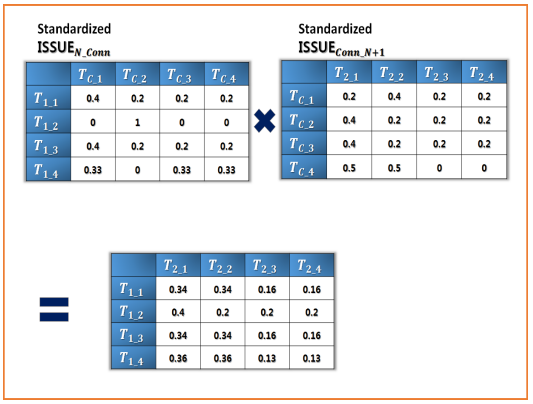
<Figure 6(a)>는 DOC_N의 각 이슈가 DOC_{Conn}의 각 이슈에 대응되는 정도를 보이고 있다. 여기서 Match(T_{1_1}, T_{C_4})과 Match(T_{1_4}, T_{C_4})은 모두 1로 동일하게 나타난다. 하지만 T_{1_1}의 경우 Match (T_{1_1}, T_{C_4})은 T_{1_1}이 DOC_{Conn}의 모든 이슈와의 대응도 총 합인 5에 대해 20%의 비중을 갖는 반면, T_{1_4}의 경우 Match(T_{1_4}, T_{C_4})은 T_{1_4}이 DOC_{Conn}의 모든 이슈와의 대응도 총 합인 3에 대해 약 33%의 비중을 갖는다. 이는 이슈의 매핑 과정에서 이슈 간 대응도의 원 값이 아닌 표준화된 값을 사용할 필요가 있음을 의미한다. 즉, <Figure 6(a)>와 <Figure 6(b)>의 이슈 매핑 매트릭스는 각각 행으로 나타난 이전 시점 기준으로 표준화되어 이후 과정에 사용된다<Figure 7>.



〈Figure 7〉 Standardizing Temporary Issue Correspondence Matrix

마지막으로 DOC_N과 DOC_{N+1}의 이슈 간 대응 관계를 파악하기 위해, <Figure 6>에 소개된 방법과 동일한 방법으로 표준화된 $ISSUE_{N\,Conn}$과 표준화된 $ISSUE_{N+1\,Conn}$을 통합한다. 이를 통

해 양 기간의 이슈 대응 매트릭스를 생성하는 과정이 <Figure 8>에 나타나있다.

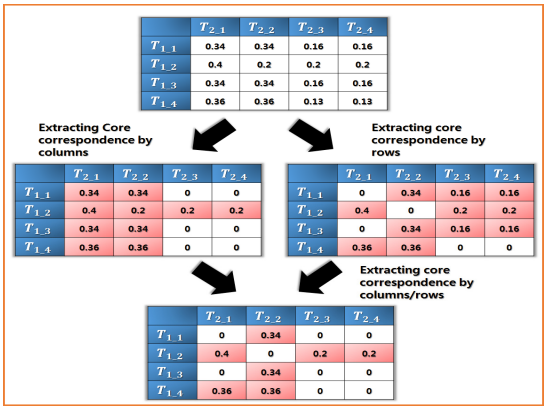


〈Figure 8〉 Integrated Issue Correspondence Matrix between Two Periods

3.5 이슈 생명주기도 생성

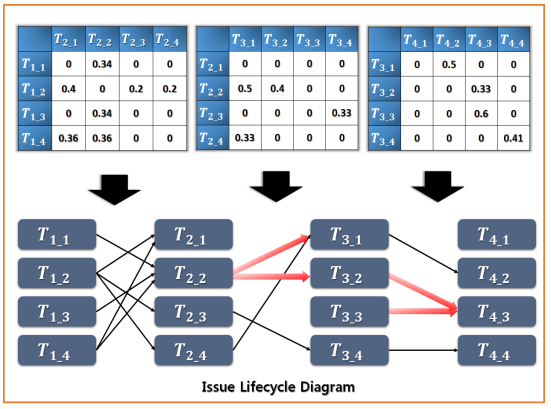
본 절에서는 <Figure 8>에서 도출된 양 기간의 이슈 대응 매트릭스를 토대로 이슈 생명주기도를 생성하는 과정을 소개한다. 생명주기도는 여러 기간에 걸친 이슈의 대응 관계를 도식화하여 보여주는 것으로, 이를 통해 이슈의 생성, 변화(분화, 병합), 소멸을 직관적으로 파악할 수 있다. 하지만 모든 이슈들의 대응 관계를 나타내는 경우 지나치게 복잡한 구조로 인해 해석의 어려움이 따르기 때문에, 임의의 임계값을 적용하여 임계값 이상의 대응도를 갖는 대응 관계만을 이슈 생명주기도에 나타낸다. 본 연구에서는 이를 위해 절대 임계값이 아닌 순위에 기반한 상대 임계값을 사용하였다. 즉, 대응 매트릭스 원본에서 각 행(이전 기간의 이슈)을 기준으로 상위 2순위 이내의 대응 관계만을 추출하고, 다시 대응 매트릭스 원본에서 각 열(이후 기간의 이슈) 기준으로 상위 2순위 이내의 대응 관계만을 추출하였다. 이렇게 도출된 두 개의 매트릭스에 모두에서 '0'이 아닌 값을 갖는 대응 관계만을 추출하여 이를 양 기간의 주요 대응 관계로 정의하였다(그림 9). 이를 통해 주요 대응 관계만을 이슈 생명주기도에 도식화함으로써 이슈 간 대응 및 변화 흐름을 보다 명확하게 파악할 수 있다.

<Figure 9>의 최종 테이블에서 각 이슈를 노드로, '0'이 아닌 값을 간선으로 표시함으로써 이슈 생명주기도를 도출할 수 있다. <Figure 9>의 최종 테이블이 <Figure 10> 상단의 좌측에 나타나 있으며, 이는 <Figure 10> 하단 이슈 생명주기도의 좌측 두 열로 도식화된다. 한편 <Figure 10> 상단의 나머지 두 매트릭스와 하단 이슈 생명주기도의 나머지 부분은 이슈 생명주기도의 개념을 설명하기 위해 임의로 가정하여 삽입되었다.



〈Figure 9〉 Extracting Core Correspondence between Issues

<Figure 10>의 이슈 생명주기도는 각 기간별 이슈의 생성, 변화, 그리고 소멸과정을 보여준다. 예를 들어 이슈 T_{2_2}는 이슈 T_{3_1}과 T_{3_2}로 분화되었으며, T_{3_2}는 새로 생성된 이슈 T_{3_3}과 함께 T_{4_3}으로 병합되었다.



〈Figure 10〉 Issue Life Cycle Diagram

본 장에서는 가상 시나리오를 통해 제안 방법론을 간략하게 소개하였다. 하지만 본 장에서 사용된 예는 매우 단순할 뿐 아니라 설명을 위해 임의로 만들어진 것이므로, 본 장의 내용을 통해 제안 방법론의 실제 적용 가능성을 가늠하기에는 무리가 따른다. 따라서 충분한 양의 실제 데이터에 대한 추가 실험이 반드시 필요하며, 이는 다음 장인 4장에서 다룬다.

4. 실험

본 장에서는 실제 텍스트 문서를 이용하여 제안 방법론을 적용한 실험 과정 및 분석 결과를 소개한다.

4.1 데이터 소개

본 실험에서는 이슈 도출을 위해 인터넷 포털 사이트에 게시된 뉴스 기사를 사용하였다. 뉴스 기사는 소셜미디어 상의 글, 댓글, 또는 일반 게시물 등에 비해 비속어, 은어의 사용이 적고 문장 구조가 명확하여 보다 정확한 분석이 가능하다는 특징을 갖고 있다. 뉴스 기사의 수집을 위해 크롤러를 제작하여, 2012년 7월부터 2013년 6월까지의 뉴스 기사 중 234,776건을 수집하였다. 3장에서 소개한 바와 같이 세부 기간의 기준은 분석 목적에 따라 유연하게 설정할 수 있으나, 본 실험에서는 기존의 일반적인 이슈 트래킹 연구와의 일관성 유지를 위해 1개월을 분석 단위로 설정하였다. 구체적으로는 2012년 7월의 기사 18,781건과 2012년 8월의 기사 20,846건에 대해 각각 토픽 분석을 실시하고, 이를 통해 도출된 7월의 이슈와 8월의 이슈 간 대응 관계를 파악하였다.

4.2 데이터 분할 및 기간별 토픽분석

본 절에서는 3.2절과 3.3절에 소개된 데이터

Document/Issue matrix of 7Month document

	화영,티아라,멤버,코어콘텐츠미디어,왕따설	대선,원장,위원장,경선,안철수	스타,무한,무한도전,멤버,도전	애플,갤럭시,삼성전자,스마트폰,특허	사람,한다,것이,있는,여성
doc_no	TOPIC1	TOPIC2	TOPIC3	TOPIC4	TOPIC5
109036	0	0	1	0	0
113871	0	0	0	0	0
113961	0	0	0	0	0
1781	0	0	0	0	0
291708	0	1	1	0	0
150701	0	0	0	0	0

Document/Issue matrix of 7.5Month document

	화영,티아라,멤버,코어콘텐츠미디어,소속사	올림픽,런던,선수,금메달,한국	드라마,시청자,신사,품격,연기	원장,안철수,대선,박근혜,캠프	경찰,제주,여성,을례,범행
doc_no	TOPIC1	TOPIC2	TOPIC3	TOPIC4	TOPIC5
1781	0	0	0	0	0
291708	0	0	0	1	0
150701	0	0	0	0	0
2719	0	0	0	0	0
82740	1	1	0	0	0
10186	0	1	0	0	0

Document/Issue matrix of 8Month document

	애플,특허,삼성,삼성전자,소송	태풍,볼라,볼라벤,피해,강풍	대선,원장,경선,박근혜,새누리당	태풍,기상청,지방,조속,볼라벤	일본,독도,대통령,정부,방문
doc_no	TOPIC1	TOPIC2	TOPIC3	TOPIC4	TOPIC5
2719	0	0	0	0	0
82740	0	0	0	0	1
10186	0	0	0	0	0
221779	0	0	0	0	0
221900	0	1	0	1	0
222184	0	0	0	0	0

〈Figure 11〉 Document/Issue Matrix (Part)

분할 및 세부 기간별 토픽분석 수행 결과를 소개한다. 우선 2012년 7월의 기사 18,781건과 2012년 8월의 기사 20,846건 중 7월 16일부터 8월 15일까지의 기사를 사용해 연결고리 문서 집합을 생성하였다. 본 문서 집합은 7월과 8월의 연결고리 역할을 수행하므로 이후 과정에서 7.5월 문서 집합으로 부르기로 한다. 이렇게 준비된 7월, 7.5월, 8월의 문서 집합에 대해 각각 독립적인 토픽분석을 수행하고, 이를 통해 기간별 주요 이슈 및 각 이슈에 대응되는 문서를 파악하였다.

토픽분석은 데이터 마이닝 상용도구인 SAS Enterprise Miner 12.1의 Text Miner 모듈을 사용하여 각각의 문서에 대해 파싱, 필터링, 토픽 분석의 순으로 수행하였으며, 각 분석에서 토픽의 수는 25개로 한정하였다. 토픽분석의 결과로 나타난 각 문서 집합별 문서/이슈 매트릭스 일부에 대한 스냅샷이 <Figure 11>에 제시되어 있으며, 각 매트릭스에서 최상단 행은 각 토픽별 주요 키워드, 다음 행은 이슈의 번호를 나타낸다. 맨 좌측 열은 해당 문서의 번호를 나타내며, 특정 문서가 특정 토픽에 해당되는 경우 해당 문서 번호와 토픽 번호가 교차하는 셀의 값이 '1'로 나타난다. 그림에서 점선으로 표시된 문서는 7월 16일부터 8월 15일까지의 기사를 의미한다.

4.3 이슈 간 주요 대응관계 도출

본 절에서는 7.5월의 문서/이슈 매트릭스를 이용하여 7월과 8월의 주요 이슈에 대한 대응 관계를 도출한 결과를 제시한다. <Figure 12>에서 T7은 7월, T7.5는 7.5월 그리고 T8은 8월의 이슈를 나타낸다. 두 매트릭스는 각각 7월과 8월의 이슈들이 7.5월의 이슈들과 대응되는 정도를 나타내고 있다. <Figure 13>은 <Figure 12> 의 두 매트릭스를 사용하여 7월과 8월의 이슈 간 대응도를 도출한 최종 매트릭스를 나타낸다. 이 결과물을 이용하여 이슈 생명주기도를 도출한 결과는 다음 절에서 소개한다.

	T7.5_1	T7.5_2	T7.5_3	T7.5_4	T7.5_5	T7.5_6	T7.5_7	T7.5_8	T7.5_9	T7.5_10
T7_1	0.3	0.04	0.02	0	0	0	0	0.01	0	0
T7_2	0	0.01	0.01	0.25	0.04	0.03	0.01	0.03	0.16	0.02
T7_3	0.14	0.05	0.13	0.02	0.01	0	0	0.05	0.01	0.01
T7_4	0	0	0	0	0.01	0.04	0.24	0.02	0.01	0.09
T7_5	0.05	0.03	0.07	0.05	0.06	0.07	0.05	0.24	0.03	0.06
T7_6	0.04	0.02	0.21	0.01	0	0	0	0.05	0	0
T7_7	0.14	0.06	0.12	0.01	0.01	0	0	0.04	0	0.01
T7_8	0	0.01	0	0.03	0.02	0.23	0.05	0.05	0.02	0.12
T7_9	0	0.01	0	0.15	0.05	0.02	0	0.02	0.23	0.01
T7_10	0.01	0	0.01	0.09	0.13	0.03	0.02	0.03	0.17	0.01

Correspondence matrix of 7month/7.5month document (part)

	T8_1	T8_2	T8_3	T8_4	T8_5	T8_6	T8_7	T8_8	T8_9	T8_10
T7.5_1	0	0	0	0	0	0.01	0.03	0	0.01	0
T7.5_2	0	0	0	0	0.02	0.01	0.01	0	0.37	0
T7.5_3	0	0	0	0	0	0.01	0.02	0	0.03	0
T7.5_4	0.01	0	0.32	0	0.09	0.05	0.01	0.21	0.01	0.02
T7.5_5	0	0	0.04	0	0.02	0.02	0	0.1	0.01	0.02
T7.5_6	0.02	0	0.03	0.01	0.04	0.33	0	0.02	0.01	0.12
T7.5_7	0.19	0	0.01	0	0.02	0.1	0	0.02	0.01	0.24
T7.5_8	0.01	0	0.02	0	0.03	0.08	0.02	0.03	0.03	0.04
T7.5_9	0.01	0	0.25	0	0.07	0.04	0	0.26	0.01	0.02
T7.5_10	0.05	0	0.01	0	0.03	0.22	0	0.01	0.01	0.16

Correspondence matrix of 7.5month/8month document (part)

〈Figure 12〉 Issue Correspondence Using Connecting Period (Part)

	T8_1	T8_2	T8_3	T8_4	T8_5	T8_6	T8_7	T8_8	T8_9	T8_10
T7_1	0.0011	0	0.0026	0	0.0065	0.0088	0.0282	0.0035	0.0404	0.0039
T7_2	0.016	0.0008	0.2075	0.0043	0.0898	0.0819	0.0091	0.1818	0.0265	0.0444
T7_3	0.0049	0	0.0227	0.0001	0.0244	0.0293	0.0566	0.0229	0.0847	0.0187
T7_4	0.0672	0.0004	0.0181	0.002	0.0238	0.1065	0.0033	0.0204	0.0099	0.1257
T7_5	0.0301	0.0036	0.0683	0.0139	0.0569	0.1256	0.0298	0.0678	0.0684	0.0847
T7_6	0.0017	0	0.0086	0.0001	0.0134	0.0135	0.0353	0.0085	0.0554	0.007
T7_7	0.0042	0	0.0115	0	0.0195	0.0242	0.0659	0.0137	0.0977	0.0178
T7_8	0.0386	0.0028	0.0554	0.0128	0.053	0.1876	0.0054	0.0492	0.0224	0.1215
T7_9	0.0116	0.0008	0.1938	0.0039	0.0774	0.0635	0.0052	0.1823	0.0189	0.0321
T7_10	0.0154	0.0012	0.1414	0.0052	0.0625	0.0662	0.007	0.1426	0.0192	0.0402

〈Figure 13〉 Issue Correspondence between July and August (Part)

4.4 이슈 생명주기도를 통한 이슈 트래킹

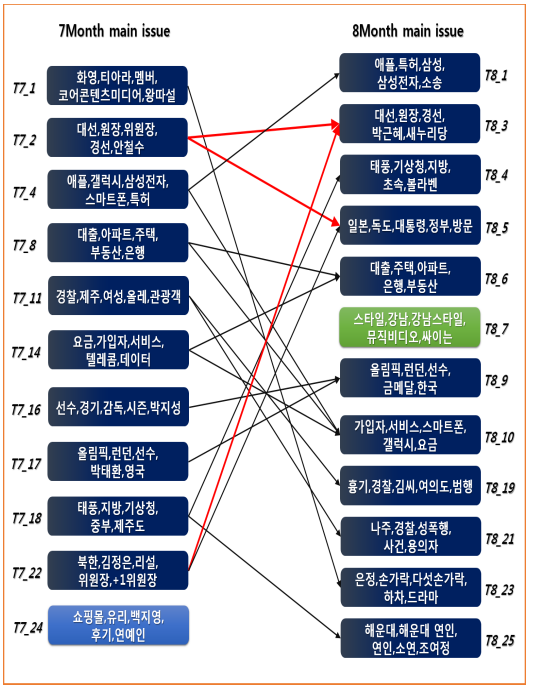
<Figure 13>은 7월의 이슈 25개와 8월의 이슈

25개로 구성된 25x25 매트릭스의 일부를 보여주고 있다. 전체 매트릭스를 이슈 생명주기도로 도식화하여 7월과 8월의 이슈 간 대응 관계를 파악할 수 있다. 하지만 3.5절에서 소개한 바와 같이 모든 이슈들의 대응 관계를 나타내는 경우 지나치게 복잡한 구조로 인해 해석의 어려움이 따르기 때문에, 임계값을 적용하여 임계값 이상의 대응도를 갖는 대응 관계만을 이슈 생명주기도에 나타낸다. 본 실험에서는 각 행과 열에서 각각 5개의 상위 값들을 추출하였다. 이렇게 생성된 두 매트릭스의 교집합을 취하여 주요 이슈 대응 매트릭스를 도출하였으며, 그 결과가 <Figure 14>에 나타나있다. 단 동일 기간 내의 25개의 이슈들 중에서도 서로 중복된 이슈가 다수 존재하므로, 이들 중 대표적 이슈만을 선택하여 <Figure 14>에 나타냈다. 또한 <Figure 14>의 매트릭스를 이슈 생명주기도로 도식화 한 결과가 <Figure 15>에 제시되어 있다.

	T8_1	T8_3	T8_4	T8_5	T8_6	T8_7	T8_9	T8_10	T8_19	T8_21	T8_23	T8_25
T7_1	0	0	0	0	0	0	0	0	0	0	0.0642	0
T7_2	0	0.2075	0	0.0898	0	0	0	0	0	0	0	0
T7_4	0.0672	0	0	0	0	0	0	0.1257	0	0	0	0
T7_8	0	0	0	0	0.1876	0	0	0.1215	0	0	0	0
T7_11	0	0	0	0	0	0	0	0	0.0909	0.121	0	0
T7_14	0	0	0	0	0.1507	0	0	0.1482	0	0	0	0
T7_16	0	0	0	0	0	0	0.1717	0	0	0	0	0
T7_17	0	0	0	0	0	0	0.2207	0	0	0	0	0
T7_18	0	0	0.055	0	0	0	0	0	0	0	0	0.0712
T7_22	0	0.1247	0	0.083	0	0	0	0	0	0	0	0
T7_24	0	0	0	0	0	0	0	0	0	0	0	0

〈Figher 14〉 Core Correspondence between Issues in July and August

<Figure 15>는 7월과 8월 간 주요 이슈의 흐름을 도식화하여 나타낸 이슈 생명주기도이다. 그림에서 7월의 일부 이슈는 소멸되었으며, 일부 이슈는 8월 이슈로 지속되었음을 알 수 있다. 또한 어떤 이슈는 8월에 새로 생성되기도 하였다. 예를 들면 이슈 T7_24는 8월의 어떤 이슈로도 연결되지 않으므로 소멸된 것으로 이해할 수 있으며, 이슈 T7_2는 8월에는 이슈 T8_3과 이슈 T8_5로 분화되었으며, 7월의 이슈 T7_2와 이슈 T7_22는 8월에는 이슈 T8_3으로 병합되어 지속된 것을 알 수 있다. 이슈 T8_7은 7월에는 존재하지 않았으나 8월에 새로 생성된 이슈의 예이다.



〈Figure 15〉 Issue Flow Diagram for July and August

본 장에서는 3장에서 제안한 방법론을 실제 뉴스 기사에 적용하는 실험의 과정과 결과를 소개하였다. 실험 결과 각기 독립적으로 수행된 두 기간의 토픽분석 결과로부터 주요 이슈를 추출하고, 각 기간별 주요 이슈간의 대응 관계를 도

식화하여 제공한 이슈 생명주기도를 통해 이슈의 생성, 분화, 병합, 소멸을 파악하는 이슈 트래킹을 수행할 수 있음을 알 수 있었다. 향후 더욱 많은 기간을 포함한 분석을 통해 다양한 이슈의 생명주기를 분석하여 의미 있는 결과를 창출할 수 있을 것으로 기대한다.

5. 결론

최근 다량의 텍스트 문서로부터 주요 이슈를 추출하고, 여러 기간에 걸친 이슈의 흐름을 파악하는 이슈 트래킹에 대한 관심이 높아지고 있다. 하지만 전통적인 이슈 트래킹은 분석 기간의 추가가 어렵다는 확장성 측면과, 이슈의 생성, 변화, 소멸 과정을 충분히 다루지 못한다는 측면에서 한계를 갖고 있다. 이에 본 연구에서는 전통적인 이슈 트래킹의 한계를 극복하기 위한 ILCA 방법론, 즉 이슈 생명주기 분석 방법론을 제안하였다. 또한 실제 뉴스 기사에 대한 실험을 통해 제안 방법론이 이슈가 생성, 변화, 소멸되는 과정을 추적할 수 있으며, 분석 대상 기간을 지속적으로 확장할 수 있음을 보였다. 따라서 제안하는 ILCA 방법론은 일회성 이슈 트래킹이 아닌 지속적인 분석 기간 추가가 요구되는 상시적 이슈 트래킹에 매우 적합한 특성을 갖고 있는 것으로 판단되며, 본 방법론을 통해 이슈들간의 연결 및 영향 관계를 파악함으로써 복잡다단한 사회 현상을 더욱 명확하게 이해할 수 있을 것으로 기대된다. 구체적으로는 장기간 지속되는 이슈 또는 다른 이슈로의 파급 효과가 큰 이슈의 특성을 파악하여, 신규 상품 및 서비스 개발, 대선 전략 수립 등에 활용할 수 있을 것으로 기대한다.

하지만 본 연구가 제안하는 ILCA 방법론이 실무적 성과로 연결되기 위해서는 다음과 같은 후속 연구가 반드시 필요하다. 우선 대부분의 텍스트 분석을 다루는 연구와 마찬가지로, 정제된 결과를 도출하기 위해서는 양질의 용어 사전 및 불용어 사전의 구축이 반드시 선행되어야 한다. 또한 본 연구의 실험에서는 세부 기간을 1달 단위로 설정하였지만, 기간의 기준을 더욱 다양하게 설정함으로써 기간 간격이 제안 방법론의 결과에 미치는 영향을 보다 엄밀하게 분석할 필요가 있다. 마지막으로 2달 간의 단기간이 아닌 훨씬 더 오랜 기간의 문서에 대한 분석을 통해 ILCA 방법론의 성과를 보다 극대화할 수 있을 것으로 기대한다.

참고문헌(References)

Albright, R., *Taming Text with the SVD*, SAS Institute Inc., Cary, NC, 2004.

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.

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.

Ding, W. and C. Chen, “Dynamic topic detection and tracking: A comparison of HDP, C-word, and cocitation methods,” *Journal of the Association for Information Science and Technology*, Vol.65, No.10(2014), 2084~2097.

Fan, W., L. Wallace, S. Rich, and Z. Zhang,

"Tapping the Power of Text Mining," *Communications of the ACM*, Vol.49, No.9 2006), 76~82.

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

Han, J. and M. Kamber, *Data Mining: Concepts and Techniques*, 3nd, Morgan Kaufmann Publishers, San Francisco, 2011.

Hong, J. S., H. S. Choi, H. J. Han, J. S. Kim, E. J. Yu, S. R. Lim, and N. Kim, "A Data Analysis-based Hybrid Methodology for Selecting Pending National Issue Keywords," *Entrue Journal of Information Technology*, Vol.13, No.1(2014), 97~111.

Hong, J.-S., N. Kim, and S. Lee, "A Methodology for Automatic Multi - Categorization of Single - Categorized Documents,"*Journal of Intelligence and Information Systems*, Vol.20, No.3(2014), 77~92.

Jin, H., R. Schwarts, S. Sista, and F. Walls, "Topic Tracking for Radio, TV Broadcast and Newswire," *Proceedings of DARPA Broadcast News Workshop*, (1999).

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

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.

Metzler, D., Y. Bernstein, W. B. Croft, A. Moffat, and J. Zobel, "Similarity Measures for Tracking Information Flow," *Proceedings of the 14th ACM international conference on Information and knowledge management,* (2005), 517~524.

Mooney, R. J. and R. Bunescu, "Mining Knowledge from Text using Information Extraction," *ACM SIGKDD Explorations Newsletter*, Vol.7, No.1(2005), 3~10.

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.

Rijsbergen, C. J. V., *Information Retrieval*, 2nd edition, Butterworths, London, 1979.

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., *Classification of Text, Automatic, The Encyclopedia of Language and Linguistics 14*, 2nd edition, Elsevier Science Pub, 2006.

Song, S. M., J. S. Yu, and E. M. Kim, "Offering system for major article Using Text Mining and Data Mining," *Proceedings of the 32th annual conference on Korea Information Processing Society*, (2009), 733~734.

Weiss, S. M., N. Indurkhya, and T. Zhang, *Fundamentals of Predictive Text Mining*, Springer, 2010.

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

Yu, E.-J., J.-C. Kim, C.-Y. Lee, and N.-G. Kim, "Using Ontologies for Semantic Text Mining," *The Journal of Information System*, Vol.21, No.3(2012), 137~161.

Abstract

Analyzing the Issue Life Cycle by Mapping Inter-Period Issues

Myungsu Lim* · Namgyu Kim**

Recently, the number of social media users has increased rapidly because of the prevalence of smart devices. As a result, the amount of real-time data has been increasing exponentially, which, in turn, is generating more interest in using such data to create added value. For instance, several attempts are being made to analyze the relevant search keywords that are frequently used on new portal sites and the words that are regularly mentioned on various social media in order to identify social issues. The technique of "topic analysis" is employed in order to identify topics and themes from a large amount of text documents. As one of the most prevalent applications of topic analysis, the technique of issue tracking investigates changes in the social issues that are identified through topic analysis. Currently, traditional issue tracking is conducted by identifying the main topics of documents that cover an entire period at the same time and analyzing the occurrence of each topic by the period of occurrence.

However, this traditional issue tracking approach has two limitations. First, when a new period is included, topic analysis must be repeated for all the documents of the entire period, rather than being conducted only on the new documents of the added period. This creates practical limitations in the form of significant time and cost burdens. Therefore, this traditional approach is difficult to apply in most applications that need to perform an analysis on the additional period. Second, the issue is not only generated and terminated constantly, but also one issue can sometimes be distributed into several issues or multiple issues can be integrated into one single issue. In other words, each issue is characterized by a life cycle that consists of the stages of creation, transition (merging and segmentation), and termination. The existing issue tracking methods do not address the connection and effect relationship between these issues.

The purpose of this study is to overcome the two limitations of the existing issue tracking method,

* 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

one being the limitation regarding the analysis method and the other being the limitation involving the lack of consideration of the changeability of the issues. Let us assume that we perform multiple topic analysis for each multiple period. Then it is essential to map issues of different periods in order to trace trend of issues. However, it is not easy to discover connection between issues of different periods because the issues derived for each period mutually contain heterogeneity.

In this study, to overcome these limitations without having to analyze the entire period's documents simultaneously, the analysis can be performed independently for each period. In addition, we performed issue mapping to link the identified issues of each period. An integrated approach on each details period was presented, and the issue flow of the entire integrated period was depicted in this study. Thus, as the entire process of the issue life cycle, including the stages of creation, transition (merging and segmentation), and extinction, is identified and examined systematically, the changeability of the issues was analyzed in this study. The proposed methodology is highly efficient in terms of time and cost, as it sufficiently considered the changeability of the issues. Further, the results of this study can be used to adapt the methodology to a practical situation. By applying the proposed methodology to actual Internet news, the potential practical applications of the proposed methodology are analyzed. Consequently, the proposed methodology was able to extend the period of the analysis and it could follow the course of progress of each issue's life cycle. Further, this methodology can facilitate a clearer understanding of complex social phenomena using topic analysis.

Key Words : Issue Life Cycle, Issue Mapping, Issue Tracking, Text Mining, Topic Analysis

Received : November 12, 2014 Revised : December 6, 2014 Accepted : December 7, 2014
Type of Submission : Outstanding Conference 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. 한국정보기술응용학회 부회장, 한국경영정보학회 이사, 한국지능정보시스템학회 이사, 한국CRM학회 이사, 한국IT서비스학회지 편집위원, JITAM 편집위원, 한국인터넷정보학회논문지 편집위원을 역임하였으며, 한국생산성본부 TOPCIT 개발사업 자문위원으로 활동 중이다. 주요 관심분야는 텍스트 마이닝, 데이터 마이닝 및 시맨틱 데이터 관리 등이다.