ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2016 December: 22(4): 45~67 http://dx.doi.org/10.13088/jiis.2016.22.4.045

텍스트 분석을 활용한 정보의 수요 공급 기반 뉴스 가치 평가 방안

이동훈
국민대학교 비즈니스IT전문대학원
(*donghoonlee@kookmin.ac.kr*)

최호창
국민대학교 경영학부 경영학전공
(*choi3684@kookmin.ac.kr*)

김남규
국민대학교 비즈니스IT전문대학원
(*ngkim@kookmin.ac.kr*)

최근 정보 유통의 주요 매체인 인터넷 뉴스와 SNS의 매체 간 특성 차이를 주목한 많은 연구가 있었음에도 불구하고, 양 매체의 차이를 정보의 수요 및 공급 관점에서 파악한 연구는 상대적으로 매우 부족하다. 일반적으로 새로운 정보는 언론사의 뉴스 기사를 통해 대중에게 노출되고, 대중은 이러한 기사에 대한 의견 또는 추가 정보를 SNS를 통해 공유함으로써 해당 정보를 수용함과 동시에 확산시킨다. 이러한 측면에서 언론사가 뉴스를 제공하는 행위를 정보의 공급으로 파악할 수 있으며, 대중은 SNS를 통해 이에 대한 관심을 능동적으로 나타냄으로써 해당 정보에 대한 소비 수요를 표출하는 것으로 이해할 수 있다. 이는 상품 및 서비스의 가격이 수요와 공급의 관계에 의해 결정되는 것과 유사한 원리로, 정보의 가치를 정보 수요와 정보 공급의 관계에 기반을 두어 측정할 수 있음을 시사한다. 본 연구에서는 정보 공급의 대표 매체로 인터넷 뉴스 기사를, 정보 수요를 나타내는 대표 매체로 트위터를 선정하고, 특정 이슈에 대한 뉴스의 정보로서의 가치를 이와 관련된 트위터의 양으로 평가하는 뉴스가치지수(NVI, News Value Index)를 고안하여 제시한다. 구체적으로 제안 방법론은 각 이슈별로 NVI를 도출하고 이를 통해 시간의 흐름에 따른 정보 가치의 변화를 시각화하여 나타낸다. 또한 본 연구에서는 제안 방법론의 실무 적용 가능성을 평가하기 위해 인터넷 뉴스 387,018건과 트윗 31,674,795건에 대한 실험을 수행하였다. 그 결과 대부분의 이슈가 전체 정보 시장의 평균 가치에 수렴하는 형태로 변화함을 알 수 있었으며, 꾸준히 평균 이상의 가치를 가지며 정보 시장을 장악하는 등 특이한 양상을 보이는 흥미로운 이슈도 존재함을 파악할 수 있었다.

주제어 : Big Data, News Value Index, SNS, Text Mining, Topic Modeling

논문접수일 : 2016년 11월 19일 논문수정일 : 2016년 12월 13일 게재확정일 : 2016년 12월 18일
원고유형 : 일반논문 교신저자 : 김남규

1. 서론

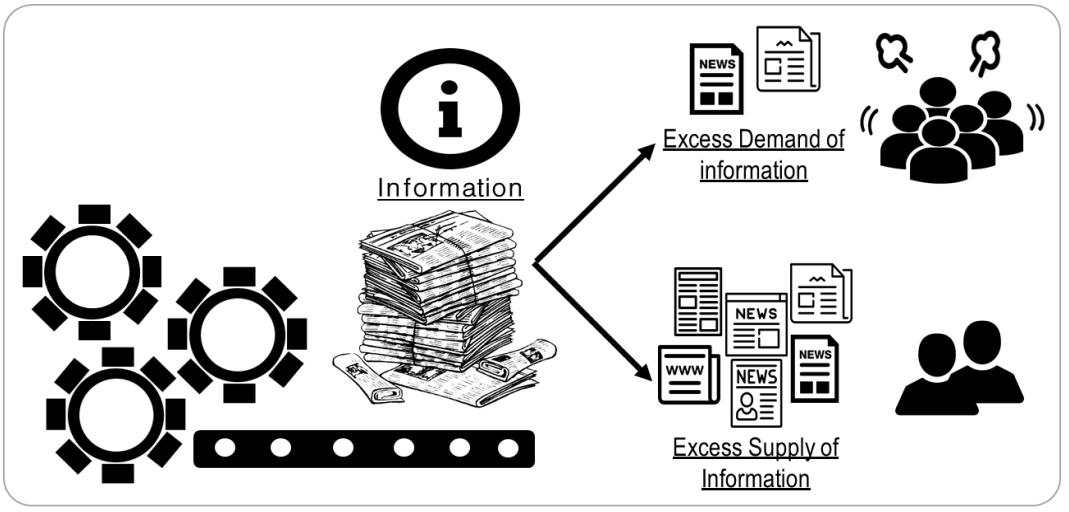
최근 스마트 기기의 발달로 인해 대중들은 인터넷, 소셜 네트워크 서비스(SNS, Social Network Service) 등을 사용하여 다양한 정보를 생산, 공유, 획득하고 있다. 각 개인은 각자의 목적 및 성향에 따라 트위터(Twitter), 페이스북(Facebook) 등 여러 매체를 동시에 사용하고 있으며, 국내 소셜 네트워크 서비스 사용자는 평균 2.09개의 매체를 동시에 이용하고 있는 것으로 나타났다. 이러한 매체를 통한 정보의 표현은 대부분 텍스트 형태로 이루어지므로, 최근 이런 텍스트 분석을 통해 사용자를 보다 깊게 이해하기 위한 연구가 매우 활발하게 수행되고 있다. 텍스트 분석의 초기 연구는 문서의 출처 및 유형에 따른 특성을 간과한 채 문서의 내용 분석에만 초점을 두고 이

루어졌지만, 최근에는 문서의 특성에 따라 이를 분석하고 해석하는 방법이 서로 다르게 적용되어야 한다는 주장도 제기되고 있다.

문서는 관점에 따라 다음과 같이 여러 형태로 구분될 수 있다. 우선 문서는 그 기능에 따라 사실 및 정보 전달을 위한 정보적(Informative) 문서, 감정 및 미학을 표현하기 위한 표현적(Expressive) 문서, 수신자의 행동을 유도하기 위한 작용적(Operative) 문서, 그리고 이미지, 음악 등을 통해 위의 세 가지 기능을 보완하는 시청각 미디어 문서(Jung and Nam, 2006; Munday, 2016; Reiss, 1981)로 구분될 수 있다. 한편 문서는 담고 있는 내용에 따라 구분될 수 있으며(Kim and Jung, 2014), 이들 내용은 사실(Facts), 개념(Concepts), 절차(Procedures), 원리(Principles), 원칙(Rule), 이야기(Story), 의견(Opinion), 그리고 묘사(Description) 등으로 구성된다. 또한 문서는 유통되는 매체에 따라 그 특성이 상이하게 나타난다. 예를 들어 가장 대표적인 전통 언론인 신문의 경우 고도의 훈련을 받은 한정된 사람만이 글을 쓸 수 있고 명확한 정보를 지면 크기에 맞추어 유통하는 반면, SNS의 경우 다양한 수준의 모든 참여자가 수시로 글을 쓸 수 있으며 다양한 성격의 글들이 다소 무질서하게 유통된다는 차이가 있다. 또한 전통적인 신문의 경우 각 기사가 독립적으로 존재하여 기사 간의 관계를 특정하기 어려운 반면, 대표적인 SNS인 트위터(Twitter)의 경우 원문(Tweet), 답글(Reply), 리트윗(Retweet)(Hur and Choi, 2012) 등 각 글들이 서로 관계를 맺고 존재하게 된다.

이처럼 신문과 SNS의 매체 간 특성 차이에 주목한 많은 연구가 있었음에도 불구하고, 양 매체의 차이를 정보의 수요 및 공급 관점에서 파악한 연구는 찾아보기 어렵다. 온라인 뉴스의 등장은 뉴스의 생산량을 급속도로 증가시켰고, 이어 SNS가 등장하면서 뉴스의 전파 속도 또한 급격히 증가하게 되었다(Park, 2012). 즉 대부분의 경우 새로운 정보는 언론사의 온라인 또는 오프라인 뉴스 기사를 통해 대중에게 노출되고, 대중은 이러한 기사에 대한 의견 또는 추가 정보를 SNS를 통해 표현함으로써 해당 정보를 확산시킨다. 이러한 측면에서 언론사에 의해 뉴스가 제공되는 행위를 정보의 공급으로 파악할 수 있으며, 대중은 이에 대한 관심을 SNS를 통해 능동적으로 나타냄으로써 해당 정보에 대한 소비 수요를 표출하는 것으로 이해할 수 있다. 물론 일부 정보의 경우 언론사의 뉴스 기사보다 오히려 SNS를 통해 먼저 소개되는 경우도 있다. 하지만 SNS 상의 정보는 사실관계 확인을 거친 뒤 뉴스에 기사화되어 본격적인 확산이 이루어짐을 감안할 때, 정보의 생성과 소비는 언론사의 뉴스를 통한 공급과 대중의 SNS 활동을 통한 수요의 틀에서 파악하는 것이 타당하다고 할 수 있다.

이처럼 전통적인 뉴스와 SNS 활동을 정보의 공급과 수요의 관점에서 파악함으로써, 정보량에 대한 딜레마, 즉 정보의 양이 넘쳐날수록 오히려 원하는 정보를 찾기가 더욱 어려워지는 현상을 보다 명확하게 설명할 수 있다. 예를 들어 한 이해 집단이 특정 목적 하에 대량으로 생산하는 뉴스의 경우, 온라인 및 오프라인 뉴스를 통해 기사로 게재되는 건수는 많음에도 불구하고 대부분의 대중에게는 외면당할 가능성이 있다. 이러한 정보의 생산은 매체 자원의 낭비를 초래할 뿐 아니라, 실제로 수요가 많은 정보를 더욱 찾기 어렵게 만드는 부작용을 초래한다. 한편 또 다른 정보는 아주 소수의 뉴스 기사를 통해서만 제공되었지만 대중이 이에 대한 관심을 SNS를 통해 폭발적으로 나타냄으로써, 해당 정보에 대



〈Figure 1〉 Excess supply and demand of information

한 일시적인 품귀 현상이 발생하게 되고 결과적으로 이 정보를 공급한 해당 기사, 기자, 그리고 언론사가 동시에 유명세를 얻게 되기도 한다 (Figure 1).

정보의 수급 비대칭 현상으로 인한 부작용을 완화하기 위해서는 각 정보별 수요 및 공급 현황을 분석하고 이에 따라 탄력적으로 정보를 제공하는 노력이 필요하다. 즉 상품 및 서비스의 가격이 수요와 공급의 관계에 의해 결정되는 것과 유사한 원리로, 정보의 가치를 정보 수요와 정보 공급의 관계에 기반하여 측정하는 것이 필요하다는 것이다. 이를 통해 언론사는 정보의 가치가 높은 이슈, 즉 수요 대비 공급이 현저하게 부족한 이슈에 대한 정보의 생산에 집중함으로써 대중의 관심을 유도할 수 있으며, 궁극적으로 정보의 수요자인 대중은 관심 대상이 되는 정보를 더욱 풍족하게 접할 수 있을 것이다. 이를 위해 본 연구에서는 정보 공급의 대표 매체로 인터넷 뉴스 기사를, 정보 수요를 나타내는 대표 매체로 트위터를 선정하고, 특정 이슈에 대한 뉴스의 정보로서의 가치를 이와 관련된 트위터의 양으로 평가하는 뉴스가치지수(NVI, News Value Index)를 고안하여 제시한다. 또한 NVI를 통해 시간의 흐름에 따른 정보 가치의 변화를 시각화하여 나타냄으로써, 전체 정보 시장의 평균 가치에 수렴하는 형태로 변화해가는 이슈, 꾸준히 평균 이상의 가치를 가지며 정보 시장을 장악하는 이슈 등의 양상을 판별할 수 있는 방법론을 제안하고자 한다.

본 연구의 이후 구성은 다음과 같다. 2장에서는 본 연구와 관련된 선행 연구의 성과를 요약하고, 3장에서는 본 연구에서 제안하는 방법론을 간단한 예를 통해 소개한다. 4장에서는 제안 방법론을 실제 인터넷 뉴스 387,018건과 트윗 31,674,795건에 적용한 실험 결과를 분석하고, 마지막 장인 5장에서는 본 연구의 기여 및 한계, 그리고 후속 연구의 방향을 제시한다.

2. 관련 연구

최근 스마트폰을 비롯한 여러 정보통신 기기의 발달로 인해 동영상, 이미지를 포함하는 다양한 유형의 정보가 유통되고 있지만, 그럼에도 불구하고 텍스트는 현실 세계에서 정보를 표현하고 교환하는 가장 대표적인 수단임에 틀림이 없다(Witten, 2004). 따라서 지금도 방대한 양의 지식이 텍스트 형태로 축적되고 있으며, 대량의 텍스트 데이터를 분석해 의미 있는 정보를 추출하는 텍스트 마이닝에 대한 관심은 점점 높아지고 있다.

텍스트 마이닝은 데이터 마이닝을 포함한 다양한 분야의 기술을 포괄적으로 활용한다(Mooney and Bunescu, 2006; Rijsbergen, 1979; Sebastiani 2006). 그 중에서도 자연어(Natural Language) 형태로 작성되는 텍스트의 특성상 이를 정형화하기 위한 자연어 처리(Weiss et al., 2015) 기술은 텍스트 마이닝 분석의 핵심 요소라 할 수 있다. 텍스트 마이닝은 일반적으로 비정형 텍스트 문서를 정형화한 뒤, 기존의 다양한 마이닝 기법을 변형하여 활용하는 형태로 이루어진다. 텍스트의 정형화 과정에는 기본적으로 각 문서에 사용된 용어의 빈도에 따라 문서의 주제 및 특성을 요약하는 벡터공간모델(Vector Space Model)(Albright, 2006; Salton et al., 1975)이 사용된다. 이에 따라 각 용어의 빈도수를 기입하기 위한 다양한 지표들이 개발되어오고 있으며, 그 중 특히 용어의 절대 빈도수 및 상대적 빈도수를 동시에 고려한 TF-IDF(Term Frequency-Inverse Document Frequency)(Han and Kamber, 2011)가 가장 대표적인 지표로 활용되고 있다.

빈도수 기반 분석에서 각 문서는 “(문서 수)×(용어 수)”의 행렬로 표현된다. 이 때, 각 문서를 임의의 이슈들의 집합으로 가정하고 해당 문서를 구성하는 용어들의 중요도를 확률적으로 계산한 값을 통해 키워드의 집합을 도출하는 일련의 알고리즘을 토픽 분석(Topic Analysis)(Bae et al., 2014) 또는 토픽 모델링(Topic Modeling)이라 한다. 최근 국내에서도 이슈의 생명주기 분석(Lim and Kim, 2014; Lim and Kim, 2016), 과학기술이슈 여론을 분석(Kim et al., 2015), 사용자 관심 이슈 분석을 통한 추천시스템 성능 향상(Choi et al., 2015) 등 토픽 모델링을 활용하여 다양한 분야의 문제를 해결하기 위한 시도가 활발하게 이루어지고 있다.

텍스트 마이닝 기법을 활용한 많은 연구의 대부분은 뉴스 또는 트위터 데이터를 대상으로 수행되었다. 우선 뉴스 데이터를 이용한 연구로는 토픽 모델링의 한 방법인 LDA(Latent Dirichlet Allocation)를 사용하여 구제역으로 인한 여러 파급효과를 분석할 수 있는 방법론을 제안한 연구(Noh et al., 2016), 오피니언 마이닝(Opinion Mining)을 활용하여 한국어 역접관계를 고려한 반의법 규칙 알고리즘을 제안한 연구(Jo et al., 2015), 감성분석(Sentiment Analysis)을 활용하여 뉴스 기사 제목을 형태소 분석한 후 특정 장소에서의 이벤트 성과를 사전에 예측하는 방법을 제안한 연구(Choi et al., 2016), 감성분석을 통해 주가지수가 상승할 때의 어휘를 도출, 사전을 구축하고 이를 통해 주가지수의 상승을 예측한 연구(Yu et al., 2013), 토픽분석과 소셜네트워크분석(Social Network Analysis)을 사용해 해당 사용자의 실제 관심분야를 파악하고 동시 방문자 관점에서 군집함으로써 상위수준의 새로운 테마를 발굴하기 위한 연구(Kim et al,. 2014), word2vec을 활용하여 주가의 방향성을 예측하기 위한 연구 등이 있다. 다음으로 트위터 데이터를 이용한

연구로는 토픽분석과 감성분석을 적용하여 스트레스 토픽을 추출하고, 로지스틱회귀모델을 통해 사용자의 거주지역 정보를 유추하여 스트레스 감성과 토픽의 지역차를 확인한 연구(Kang, 2015), word2vec 기법을 사용하여 '세월호사건'에 대한 트위터 이용자의 정치적 의견 차이를 분석한 연구(Jung et al., 2016), 오피니언 마이닝을 활용해 한국어 문법에 적합한 연관규칙 기반의 감성사전을 구축하고 이를 통해 패션 트랜드 분석에서의 향상된 정확도를 가지는 모델을 구축하기 위한 알고리즘을 제안한 연구(Lee et al., 2014), 주성분 분석과 오피니언 마이닝을 활용해 사회감정을 수치화하여 나타낸 연구, TF-IDF와 벡터공간모델을 적용해 범죄발생 위험요소 검색에서의 정보 정확도를 향상시키고, 해당 요소를 효율적으로 추출할 수 있는 방안을 제시한 연구(Lee et al., 2015) 등이 있다. 또한 실시간으로 발생하는 트위터 데이터를 사용해 특정 이슈를 중심으로 발생하는 사회적 네트워크의 특성을 규명하기 위한 연구(Bae et al., 2013), 트위터의 댓글 그래프를 사용해 토픽 모델링의 결과를 향상시키기 위한 연구(Lee and Lee, 2014), 특정 상품명을 포함하는 트윗을 대상으로 토픽 변화를 추적하여 토픽 변화의 시점 및 패턴을 파악한 연구(Jin et al., 2013), 트위터 데이터를 활용해 이슈를 추출하고 이를 웹상에서 시각화하는 시스템을 설계 및 구축하기 위한 연구(Bae et al., 2014)도 진행 되었다.

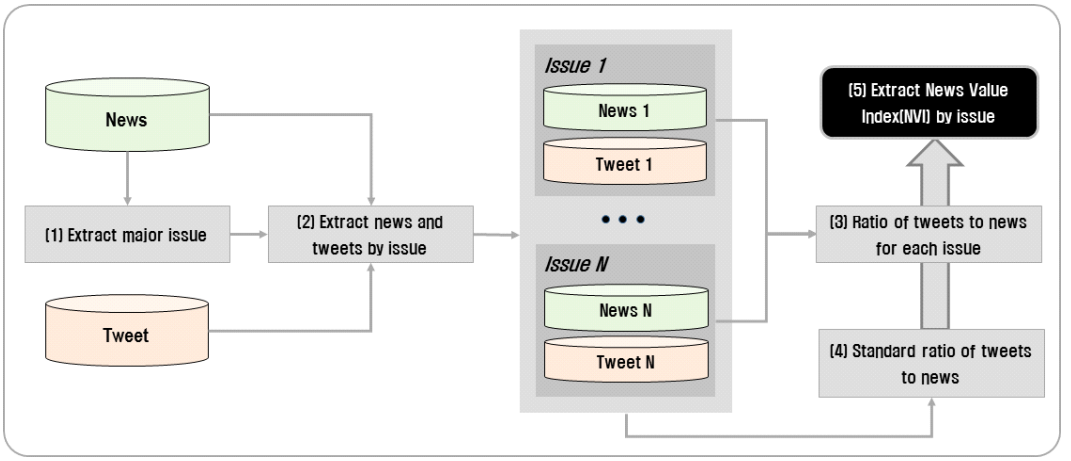
위에서 소개한 대부분의 연구가 단일 매체의 텍스트 데이터를 실험 대상으로 사용한 것과 달리, 다양한 매체로부터 데이터를 수집하여 통합적 분석을 수행하거나 매체 간 데이터의 특성을 비교한 연구도 다수 수행되었다. 대표적인 최근 연구의 사례로는 트위터와 뉴스 기사를 통해 구축한 주가예측 모형의 성능을 비교한 연구(Kim et al., 2014), 뉴스 기사, 블로그, 트위터로부터 수집한 데이터를 이용해 감정유발요인을 도출한 연구(An et al., 2015), 뉴스 기사와 트위터에서 수집한 자료에 대한 연관분석을 통해 특정 행정구역과 관련된 기사 및 월별 이슈를 추출한 연구(Lim and Park, 2015), 특정 드라마의 시청률과 점유율, 뉴스 기사, 트위터 메시지 등을 분석하여 시청률을 예측한 연구(Kim, 2016), 뉴스 기사, 블로그, SNS 등에서 수집한 데이터를 분석하여 데이터의 수요와 공급에 대한 예측모형을 개발한 연구 등을 들 수 있다. 이들 연구는 단일 매체가 아닌 다양한 매체의 데이터를 복합적으로 사용했다는 점에서 의의가 있으며, 나아가 일부 연구는 각 매체별 특성의 차이를 규명하였다는 점에서 그 기여를 찾을 수 있다. 하지만 특정 매체에서 유통되는 정보의 가치 평가를 위해 다른 매체를 활용한 연구는 찾아보기 어려우며, 이러한 측면에서 본 연구는 기존의 연구들과 접근 방향에서 크게 차이가 있다고 할 수 있다.

3. 제안 방법론

3.1 연구 모형

본 절에서는 뉴스 토픽의 가치를 정보의 수급 관점에서 평가하는 뉴스가치지수(NVI, News Value Index) 도출 방법론의 개념과 주요과정을 소개한다. 전체적인 방법론의 개요는 본 절의 <Figure 2>를 통해 제시하며, 각 주요 모듈에 대한 자세한 설명은 이후 절에서 다루도록 한다.

우선 (1)주요 이슈 도출은 뉴스 기사에 대한 토픽 모델링을 통해 분석 대상이 되는 주요 이슈



〈Figure 2〉 Research Overview

를 도출하는 과정을 나타내며, 다음 절인 3.2절에서 소개한다. 다음 모듈인 (2)이슈별 뉴스 및 트윗 추출 모듈에서는 특정 이슈에 대응되는 뉴스 기사와 트윗의 매핑이 이루어진다. 요약하면 각 이슈를 구성하고 있는 키워드 중 특정 임계값 이상의 키워드만을 선별하여 이를 관련 뉴스 및 트위터 추출에 사용하게 되며, 자세한 내용은 3.3절에서 다룬다. 마지막으로 (3) ~ (5)는 각 이슈별 뉴스당 트윗의 대응도를 산출하고 이를 각 일자별 뉴스당 트윗 기준 대응도로 나누어 이슈별 NVI를 도출하는 과정으로, 본 장의 마지막 절인 3.4절에서 자세히 다룬다.

본 장에서 제시되는 모든 예는 방법론의 이해를 돕기 위한 가상 예이며, 실제 데이터를 분석한 결과는 다음 장인 4장에서 제시한다.

3.2 주요 이슈 추출

본 절에서는 토픽 모델링을 통해 뉴스의 주요 이슈를 도출하는 과정을 소개한다. 토픽 모델링은 각 문서에 수록된 용어를 분석하여 유사도에 따라 문서를 그룹화한 뒤, 각 그룹의 주요 용어를 토픽 키워드로 제시하는 기법이다. 토픽 모델링은 이미 많은 연구를 통해 그 원리 및 과정이 충분히 소개되었으므로, 본 연구에서는 이에 대한 자세한 과정은 생략하고 주요 원리만을 요약한다.

일반적으로 각 문서는 상당히 많은 양의 용어를 포함하고 있기 때문에 우선 전체 용어에 대한 차원 축소가 이루어지며, 이 과정에서 사용된 차원의 수가 곧 토픽의 수를 나타내게 된다. 분석 결과 각 용어의 각 토픽에 대한 대응도가 도출되며, 이를 용어 가중치(Term Topic Weight)라고 한다. 이 때 주어진 용어 임계값(Term Cutoff) 이상의 용어 가중치를 갖는 용어는 해당 토픽을 기술하는 용어로 분류된다. 또한 각 문서에 대해서도 유사한 방식으로 문서 가중치(Document Topic Weight)가 계산되며, 주어진 문서 임계값(Document Cutoff) 이상의 문서 가중치를 갖는 문서는 해당 토픽을 포함한 문서로 인정된다. 이러한 과정을 통해 방대한 양의 문서로부터 주요

〈Table 1〉 An example of core issues and related terms

Issue 1: 클린턴, 트럼프, 대선		Issue 2: 최순실, JTBC, 배후		Issue 3: 건강, 위염, 내시경	
Terms	*Weight*	*Terms*	*Weight*	*Terms*	*Weight*
클린턴	0.248	최순실	0.214	건강	0.226
트럼프	0.203	JTBC	0.189	위염	0.162
대선	0.186	배후	0.162	내시경	0.146
스캔들	0.159	국정농단	0.162	염증	0.152
이메일	0.149	대통령	0.155	소화	0.150
힐러리	0.147	문화	0.153	점막	0.132
오바마	0.146	녹취	0.152	습관	0.126
미국대선	0.130	융합	0.150	증상	0.123
민주당	0.129	국가	0.135	음식	0.107
대통령	0.101	국회	0.103	원인	0.106
. . .		. . .		. . .	

토픽을 추출할 수 있으며, 전체 뉴스로부터 주요 이슈를 도출한 가상 예가 <Table 1>에 제시되어 있다.

<Table 1>은 토픽 모델링을 활용하여 뉴스 기사로부터 세 개의 주요 이슈 (클린턴, 트럼프, 대선), (최순실, JTBC, 배후), 그리고 (건강, 위염, 내시경)를 도출한 예를 보이고 있다. 또한 각 이슈를 나타내는 주요 용어를 각 용어의 가중치와 함께 이슈별로 10개씩 제시하고 있다. 이렇게 도출된 이슈 키워드를 활용하여 관련 뉴스 및 트윗을 매핑하는 과정은 다음 절에서 소개한다.

3.3 이슈별 뉴스 및 트윗 도출

본 절에서는 앞에서 도출된 이슈별 키워드 집합을 활용하여, 각 이슈에 대응되는 뉴스 및 트윗을 추출하는 과정을 소개한다. 전술한 바와 같이 용어 임계값 이상의 용어 가중치를 갖는 용어는 해당 토픽의 키워드로 지정되며, 일반적으로 용어 임계값으로는 주로 각 토픽의 모든 용어 가중치의 "평균 + 1 σ (Sigma, 표준편차)"가 사용된다. 이 경우 지나치게 많은 용어가 각 토픽의 키워드로 지정되어, 토픽과 명확한 관계가 없는 일반적인 용어들도 토픽 키워드에 다수 포함되게 된다는 한계가 있다. 따라서 본 연구에서는 토픽 모델링을 통해 도출된 토픽 키워드 전체를 사용하여 이슈별 뉴스 및 트윗을 식별하지 않고, 보다 강화된 기준으로 토픽 키워드를 재선별하여 문서 식별을 수행한다.

구체적으로는 각 토픽별로 용어 가중치에 따라 상위 N개의 용어를 구분한 뒤, 이들 전체 용어의 용어 가중치, 예를 들어 이슈가 10개이고 N=5인 경우 용어 50개의 용어 가중치의 평균을 산출한다. 이렇게 산출된 평균값을 새로운 용어 임계값으로 지정하여 각 토픽의 키워드를 도출함으로써, 각 토픽과 직접적인 관계가 있는 핵심 키워드만을 식별할 수 있다. 예를 들어 이렇게 산출된 값이 0.15일 때, <Table 1>의 이슈별 핵심 용어 중 일부는 제거되고 <Table 2>와 같이 최상위 용어만 이슈의 핵심 용어로 잔류하게 된다. <Table 2>에서 어둡게 표시된 용어는 이 과정에서 제거된 용어를 나타내며, 잔류한 핵심 용어의 수는 각 이슈별로 상이하게 나타날 수 있다.

이렇게 도출된 <Table 2>의 이슈별 핵심 용어는 해당 이슈에 대응되는 뉴스 및 트윗 식별에 사용된다. 이를 위해 다양한 방법이 모색될 수 있으나, 크게 전체 키워드를 'AND' 조건으로 조합하는 방법과 'OR' 조건으로 조합하는 방법을 살펴볼 수 있다. 전자의 경우 해당 이슈와 명확한 관련이 있는 문서를 안전하게 추출할 수 있다는 장점이 있지만, 키워드의 수가 많아질수록 이 조건을 만족시키는 문서의 수는 급격히 줄어드는 단점이 있다. 반대로 후자의 경우 조건을 만족시키는 문서의 수가 충분하게 유지된다는 장점이 있지만, 해당 이슈와 명확한 관련이 없는 문서도 대상으로 포함될 위험이 있다는 단점이 있다.

본 연구에서는 위의 두 가지 방안 중 키워드를 'OR' 조건으로 조합하는 방안을 채택하며, 그 이유는 크게 다음의 두 가지 측면에서 살펴볼 수 있다. 우선 'AND' 조건을 사용하여 문서를 식별

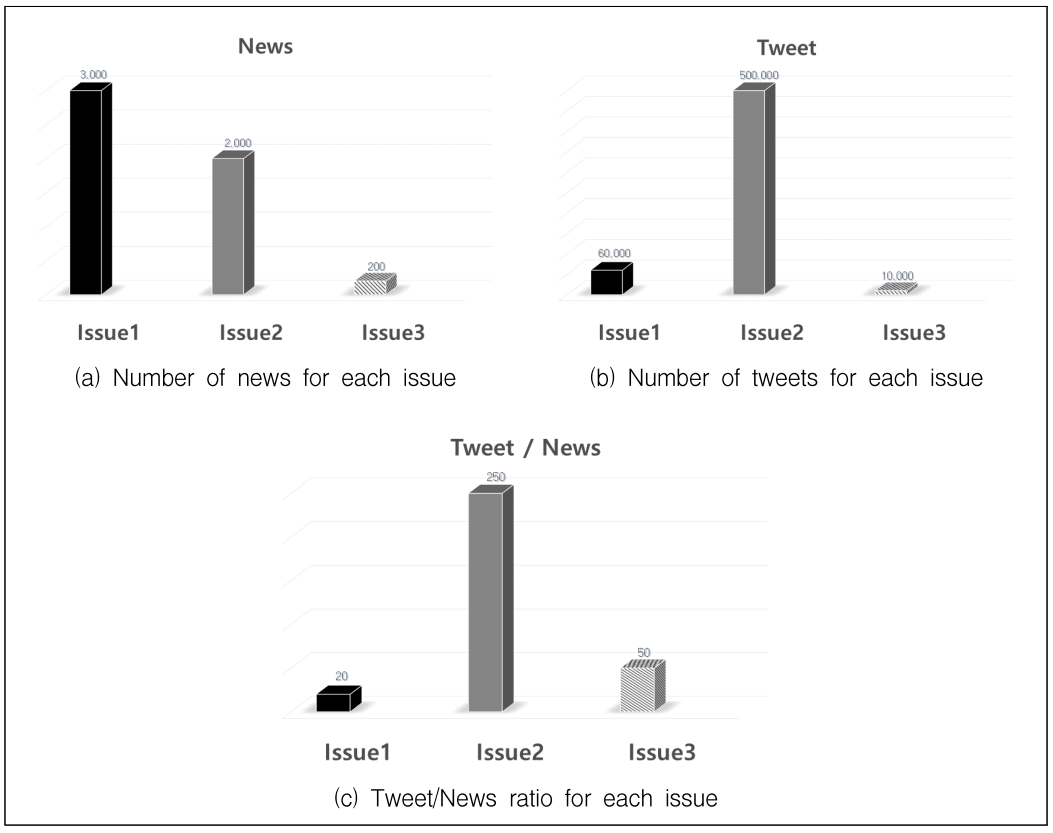
〈Table 2〉 An example of core issues and related terms with strict cutoff

Issue 1: 클린턴, 트럼프, 대선		**Issue 2: 최순실, JTBC, 배후**		**Issue 3: 건강, 위염, 내시경**	
Terms	***Weight***	***Terms***	***Weight***	***Terms***	***Weight***
클린턴	0.248	최순실	0.214	건강	0.226
트럼프	0.203	JTBC	0.189	위염	0.162
대선	0.186	배후	0.162	내시경	0.146
스캔들	0.159	국정농단	0.162	염증	0.152
이메일	0.149	대통령	0.155	소화	0.150
힐러리	0.147	문화	0.153	천마	0.132
오바마	0.146	녹취	0.152	술관	0.126
미국대선	0.130	융합	0.150	증상	0.123
민주당	0.129	국가	0.135	음식	0.107
대통령	0.101	고최	0.103	원인	0.106
. . .		. . .		. . .	

하는 경우 문서의 길이가 상대적으로 긴 뉴스 기사에서는 충분한 양의 문서가 이 조건을 만족시킬 수 있지만, 길이가 매우 짧은 트위터 데이터의 경우 이러한 조건을 만족시키는 트윗의 수가 매우 적을 것으로 예상된다. 또한 이미 전술한 방식에 따라 강화된 용어 임계값을 적용하여 이슈별 핵심 용어를 도출하였기 때문에, 'OR' 조건을 사용하더라도 해당 이슈와 전혀 무관한 문서가 식별될 가능성은 매우 낮을 것으로 판단하였다. 따라서 본 과정에서는 <Table 2>에 제시된 바와 같이 이슈별 핵심 키워드를 'OR' 조건을 사용하여 이슈별 뉴스 및 트윗으로 식별한다.

3.4 이슈별 뉴스가치지수(News Value Index) 도출

본 절에서는 앞서 도출한 이슈별 뉴스 및 트윗의 수로부터 뉴스당 트윗 대응도를 산출하고, 이를 사용하여 뉴스가치지수를 도출하는 과정을 소개한다. 토픽 모델링을 활용하여 여론을 분석한 대부분의 연구에서는 특정 이슈에 대한 뉴스 기사의 수가 많은 경우 해당 이슈를 대중의 관심이 높은 이슈로 파악한다. 하지만 본 연구에서는 뉴스 기사의 수가 많은 이슈는 정보의 공급이 충분히 이루어진 이슈일 뿐, 사용자가 이 이슈에 대해 관심이 많음을 의미하지는 않는 것으로 파



(a) Number of news for each issue

(b) Number of tweets for each issue

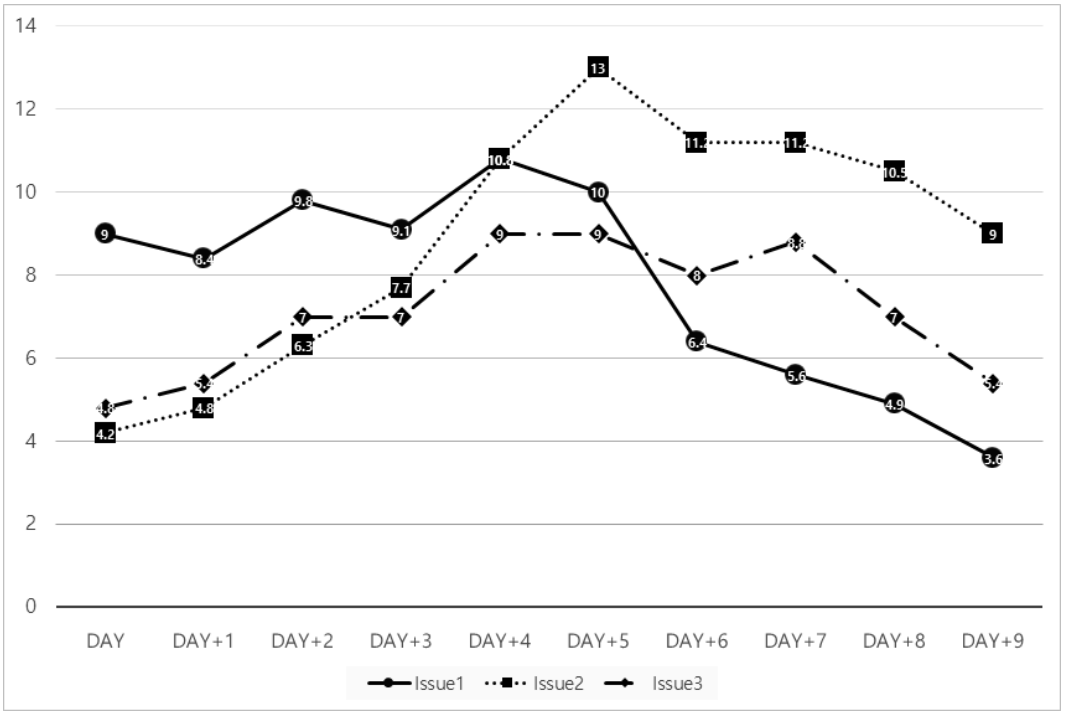
(c) Tweet/News ratio for each issue

〈Figure 3〉 Number of news, tweets, and tweet/news ratio

악한다. 오히려 사용자의 관심, 즉 정보에 대한 수요는 해당 이슈와 관련된 트윗의 수로부터 파악하며, 공급된 정보의 가치는 뉴스 기사 하나당 트윗의 수로 산출한다. 따라서 특정 이슈와 관련된 뉴스 기사의 수와 트윗의 수는 전혀 다른 양상으로 나타날 수 있으며, 이를 설명하기 위한 가상 예가 <Figure 3>에 나타나있다.

<Figure 3>을 살펴보면 뉴스에서 가장 많이 언급된 이슈는 Issue1(클린턴, 트럼프, 대선)이며, Issue2(최순실, JTBC, 배후)이 그 다음, 그리고 Issue3(건강, 위염, 내시경)은 가장 적게 언급되었음을 알 수 있다(Figure 3(a)). 한편 트윗 수 측면에서는 Issue2가 가장 많이 언급되었으며, 그 다음 Issue1, Issue3 순서로 언급되었음을 알 수 있다(Figure 3(b)). 공급 대비 수요의 비율, 즉 뉴스당 트윗의 대응도는 Issue2가 가장 높게 나타나서, 해당 이슈에 대한 대중의 관심에 비해 뉴스가 적게 제공되고 있음을 나타내고 있다(Figure 3(c)). Issue1의 경우 뉴스당 트윗 대응도가 Issue3보다도 낮게 나타났으며, 이는 대중의 관심에 비해 뉴스가 지나치게 많이 공급되고 있음을 나타낸다. 따라서 이 경우 개별 뉴스의 가치는 Issue2 > Issue3 > Issue1의 순서로 높게 나타나는 것으로 해석할 수 있다.

본 연구의 궁극적인 목적 중 하나는 위에서 소개한 뉴스의 가치, 즉 뉴스당 트윗 대응도가 시간의 흐름에 따라 변화하며, 이러한 변화의 패턴이 이슈별로 다르게 나타나는 현상을 파악하는 것이다. 예를 들어 <Figure 4>는 Issue1 ~ Issue3의 뉴스당 트윗 대응도가 DAY ~ DAY+9 까지의

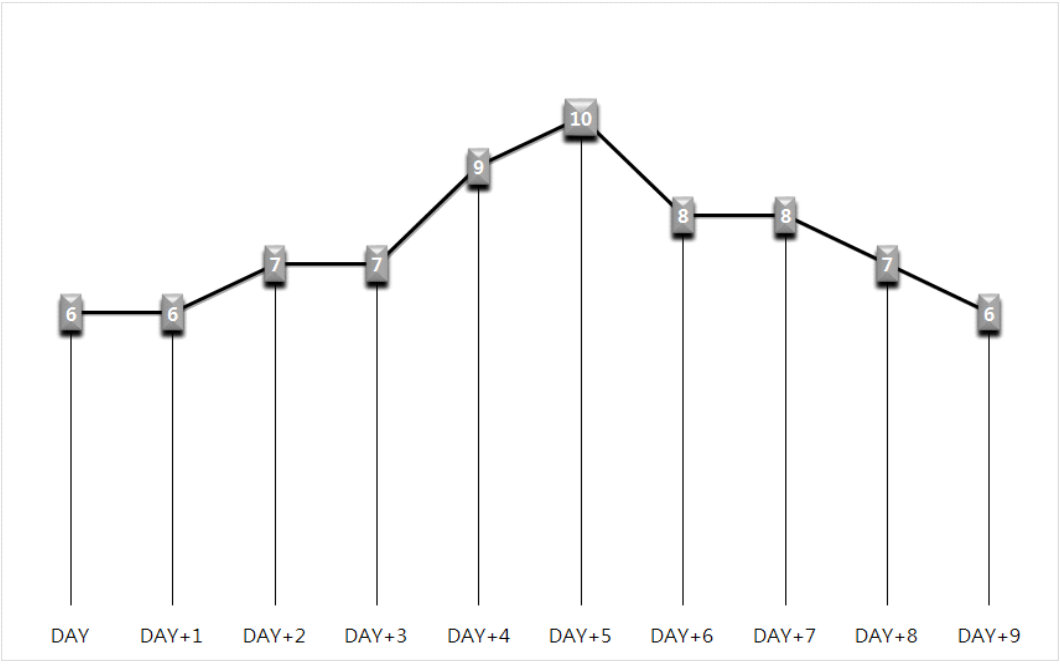


〈Figure 4〉 Changes in the ratio of tweets to news

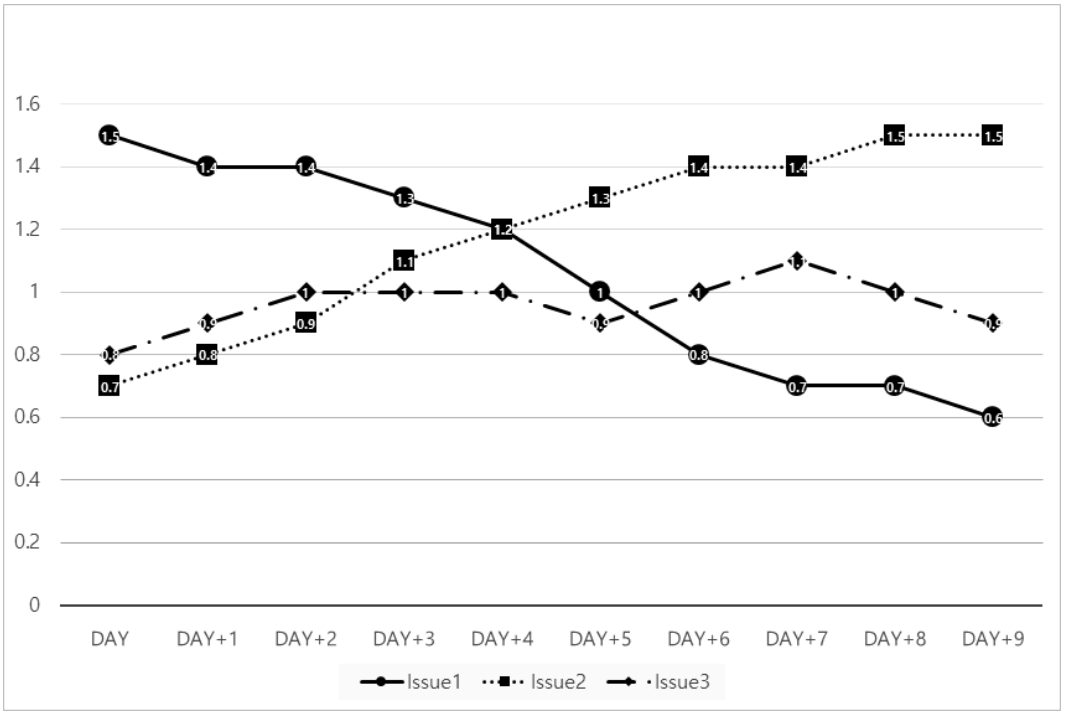
10일 간 변화하는 추이를 나타내는 가상 예이다.

<Figure 4>에서 DAY부터 DAY+4까지는 세 이슈 모두 뉴스당 트윗 대응도가 증가한다. 반면 DAY+5부터 DAY+9까지는 세 이슈 모두 뉴스당 트윗 대응도가 대체로 감소하는 추세를 보인다. 이러한 패턴은 실제 데이터를 분석한 4장의 실험 부분에서도 발견된다. 여러 이슈의 뉴스당 트윗 대응도가 비슷한 추이로 변화하는 현상은 뉴스당 트윗 대응도가 각 이슈별 내용의 영향을 받기도 하지만 정보가 제공되는 매체 특성의 영향을 받는 것에 기인한다. 예를 들면 주말, 휴일, 명절 등에는 평일에 비해 뉴스의 수가 일반적으로 적게 나타나고 트윗의 수는 평일에 비해 오히려 많게 나타나는 경향이 있다. 따라서 주말, 휴일, 명절 등에는 뉴스당 트윗 대응도가 이슈와 무관하게 높게 나타나게 되며 평일에는 이 값이 다소 낮게 나타나게 된다. 따라서 뉴스당 트윗 대응도의 일자별 편차를 보정하여 이슈 자체의 순수 뉴스당 트윗 대응도를 비교하기 위해, 본 연구에서는 각 일자의 전체 뉴스 대비 전체 트윗의 비율을 기준 대응도로 설정하여 이 값을 뉴스당 트윗 대응도의 보정에 적용한다. <Figure 5>는 세 개의 이슈 전체에 대하여 일별로 공급되는 뉴스의 기준 대응도를 나타낸 가상 예이며, 그림에서 DAY+4, DAY+5는 당일의 기준 대응도가 높게 나타난 것으로 보아 주말 또는 휴일일 것으로 추측된다.

<Figure 4>에 나타난 각 이슈의 일별 뉴스당 트윗 대응도를 <Figure 5>의 일별 기준 대응도로 나눈 값, 즉 기준 대응도 대비 이슈의 뉴스당 트윗 대응도가 <Figure 6>에 제시되어 있으며, 본 연구에서는 이 값을 각 이슈의 뉴스가치지수



〈Figure 5〉 Standard ratio of tweets to news



〈Figure 6〉 NVI trends of the three issues

(NVI)라고 정의하였다. 즉 특정 이슈의 NVI값이 높을수록 해당 일자의 전체 뉴스당 트윗 대응도에 비해 해당 이슈의 뉴스당 트윗 대응도가 높음을 의미한다.

<Figure 6>에 제시된 가상 예에서 DAY ~ DAY+9의 전체 10일 동안 Issue3은 기준 대응도와 유사한 NVI를 나타냄을 알 수 있다. 한편 Issue1은 초기에 매우 높은 NVI를 나타냈으나 시간의 흐름에 따라 그 가치가 감소하였으며, Issue2는 이와 반대로 시간이 갈수록 NVI값이 점차 증가하는 것으로 나타났다.

본 장에서는 뉴스가치지수의 개념과 도출 방법을 가상의 예를 통하여 소개하였다. 다음 장인 4장에서는 실제 뉴스 및 트윗 데이터를 분석하여 본 연구에서 제안하는 방법론을 적용한 결과를 소개한다.

4. 실험

4.1 실험 개요

본 절에서는 제안 방법론을 실제 데이터의 분석에 적용하기 위한 실험의 개요를 소개한다. 본 실험에 필요한 데이터 수집 및 분석을 위해 <Table 3>과 같은 시스템을 구축하였다. 이 시스템은 총 21대의 서버로 구성되어 있으며, 1대의 분석 및 데이터베이스 서버와 20대의 수집기 서버로 구분된다.

〈Table 3〉 System environment

Classification	Items	Remarks
HW	Crawler Server	OS: Windows7(64bit) CPU : Intel I5-3470 3.20 GHz RAM : 4GB HDD : 500GB Server : 20 Unit
HW	DB & Analysis Server	OS: Windows7(64bit) CPU : Intel I7-4770 3.40 GHz RAM : 12GB HDD : 4TB Server : 1 Unit
SW	Web Crawler	Java-based news and Twitter Crawler (self-produced)
	DBMS	MySQL 5.7 version
	Java / JVM	Java 1.8 version
	SAS Enterprise Miner	SAS 9.4 version

<Table 3>의 시스템을 통해 2014년 6월 22일부터 2014년 7월 5일까지의 국내 한 뉴스 포털 사이트의 뉴스 기사 387,014건과 트위터 데이터 31,674,795건을 수집하였으며, 각 일자별 수집 데이터 건수는 <Table 4>와 같다. 데이터의 요일별 분포를 살펴보면, 뉴스의 경우 데이터의 공급량이 평일에 비해 주말(토요일, 일요일)에 확연히 적가 나타나며, 트위터의 경우 주말의 데이터가 다른 요일과 비슷하거나 근소하게 많은 것으로 확인되었다.

〈Table 4〉 Number of news and tweets for experiments

Date	News	Twitter
20140622 (일)	15,108	2,621,685
20140623 (월)	33,028	2,350,846
20140624 (화)	32,316	2,113,294
20140625 (수)	34,250	2,118,664
20140626 (목)	36,128	2,149,820
20140627 (금)	33,018	2,248,906
20140628 (토)	10,311	2,210,231

Date	News	Twitter
20140629 (일)	15,566	2,515,511
20140630 (월)	31,957	2,183,429
20140701 (화)	37,897	2,232,870
20140702 (수)	34,977	2,248,862
20140703 (목)	31,832	2,280,943
20140704 (금)	27,892	2,298,972
20140705 (토)	12,738	2,100,762
합계	387,018	31,674,795

4.2 주요 이슈 추출

본 절에서는 토픽 모델링을 통해 주요 이슈를 도출하고, 이슈별 뉴스 기사 및 트윗을 추출한 실험 결과를 소개한다. 이슈 도출에는 2014년 6월 22일자 뉴스 기사 15,108건을 사용하였으며, 이를 통해 도출된 주요 이슈 및 각 이슈의 용어 가중치를 산출하였다. 전체 이슈 10개에 대해 각 이슈로부터 가중치가 높은 20개의 용어, 즉 전체 200개의 용어를 도출하였으며, 이들 200개 용어 가중치의 평균을 핵심용어 식별을 위한 임계값으로 사용하였다. 이를 통해 각 이슈 10개에 대한 핵심 용어를 도출하였으며, 그 결과가 <Table 5>에 요약되어 있다. <Table 5>에 제시된 핵심 용어 중 한개라도 본문에 포함하고 있는 뉴스 또는 트윗을 추출하기 위해 SQL의 Like 검색시 핵심 키워드에 'OR' 조건을 적용하였으며, 이렇게 추출된 뉴스 및 트윗의 수는 다음 절의 이슈별 NVI 도출에 사용된다.

4.3 이슈별 뉴스의 트윗 대응도 및 NVI 산출

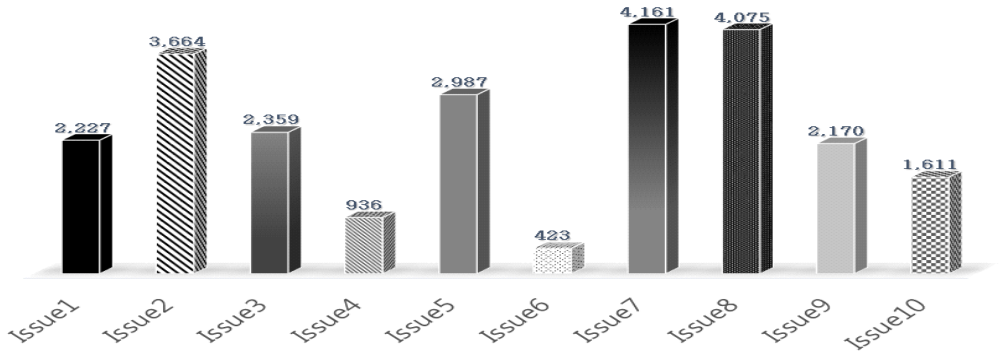
본 절에서는 이슈별 뉴스당 트윗 대응도를 산출한 결과를 소개한다. 이슈별 뉴스당 트윗 대응도의 도출을 위해 2014년 6월 22일 ~ 2014년 7월 5일의 총 2주간 뉴스당 트윗 대응도를 분석하였으며, 한 예로 2014년 6월 22일의 이슈별 뉴스 수, 이슈별 트윗 수, 그리고 이슈별 뉴스당 트윗 대응도를 분석한 결과가 각각 <Figure 7>, <Figure 8>, 그리고 <Figure 9>에 나타나있다.

<Figure 7> ~ <Figure 9>를 통해 뉴스를 통해 가장 많이 공급된 상위 이슈는 Issue7과 Issue8로 나타났다(Figure 7). 또한 트위터에서는 Issue8과 Issue3에 대중의 관심이 집중되었음을 알 수 있

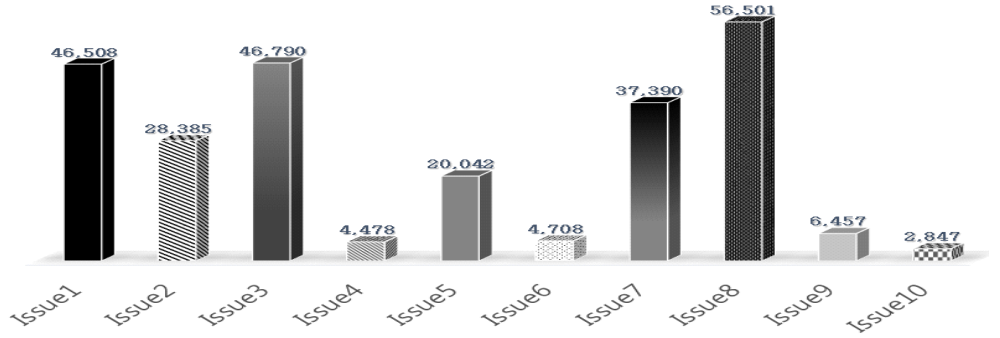
〈Table 5〉 List of issues and related terms

Issue 1: 병장, 충기, 난사		*Issue 2: 후반, 클로, 조별*		*Issue 3: 청라, 골프, 인천*		*Issue 4: 시장*		*Issue 5: 홈런, 안타, 텍사스*	
Terms	*Weight*	*Terms*	*Weight*	*Terms*	*Weight*	*Terms*	*Weight*	*Terms*	*Weight*
병장	0.283	후반	0.209	청라	0.337	시장	0.146	홈런	0.24
충기	0.265	클로	0.194	골프	0.289			안타	0.232
난사	0.202	조별	0.173	인천	0.269			텍사스	0.213
임 병장	0.188	월드컵	0.162	여자	0.247			선발	0.206
사고	0.185	리그	0.159	28회 한국	0.22			에인절	0.186
병사	0.161	슈팅	0.157	오픈	0.197			추신	0.168
사단	0.146			청라골프클럽	0.174			타자	0.157
				최종라운드	0.167			시즌	0.154
				클럽	0.161				
				...					

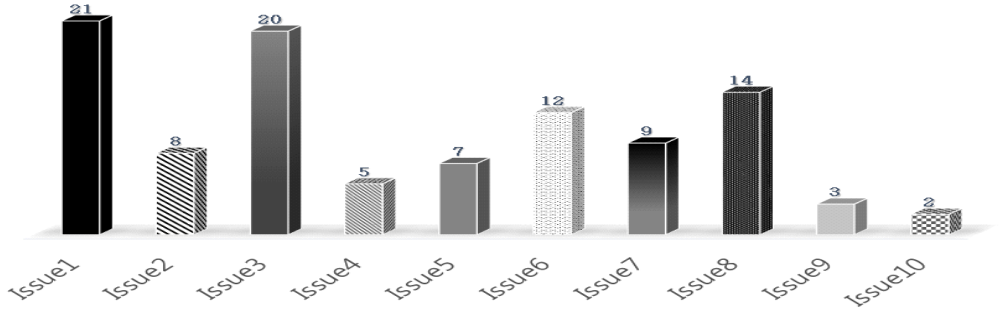
Issue 6: 허스키, 시베, 리안		*Issue 7: 그레, 브라질 포르투 알레그레, 국가대표팀*		*Issue 8: 오픈, us 여자오픈, 금융*		*Issue 9: 데일리, 모습, 룸메이트*		*Issue 10: 감독, 포르투 알레그리, 선수*	
Terms	*Weight*	*Terms*	*Weight*	*Terms*	*Weight*	*Terms*	*Weight*	*Terms*	*Weight*
허스키	0.37	그레	0.284	오픈	0.169	데일리	0.173	감독	0.226
시베	0.368	브라질 포르투 알레그레	0.226	us 여자오픈	0.147	모습	0.162	선수	0.17
리안	0.362	국가대표팀	0.216	금융	0.146	룸메이트	0.141	홍명보 감독	0.17
시베리안 허스키	0.26	스타디움	0.207	us오픈	0.143				
밴드	0.214	축구	0.197	여자	0.142				
보컬	0.184	포르투 알레그레	0.179						
사망	0.164	차전	0.173						
		...							



〈Figure 7〉 Number of news for each news (June 22, 2014)



〈Figure 8〉 Number of tweets for each issue (June 22, 2014)



〈Figure 9〉 Ratio of tweets to news for each issue (June 22, 2014)

었다(Figure 8). 한편 <Figure 7>과 <Figure 8>을 통해 도출한 이슈별 뉴스당 트윗 대응도는 Issue1(병장, 총기, 난사)과 Issue3(청라, 골프, 인천)이 가장 높은 것으로 나타났다(Figure 9). 특히

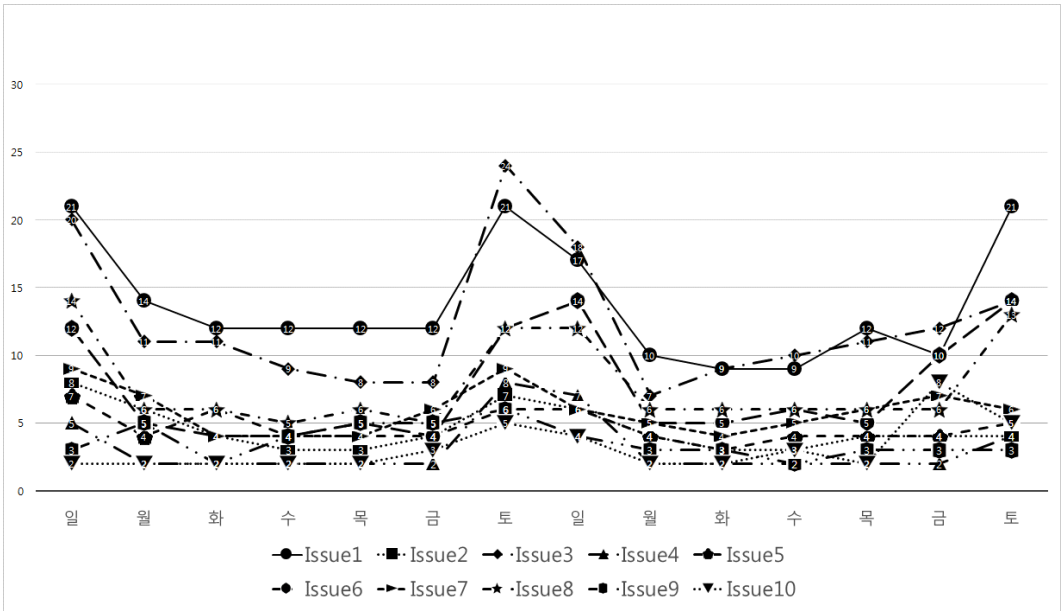
Issue1의 경우 뉴스의 공급량 기준 10개 이슈 가운데 6위에 그쳐 크게 주목받지 못했으나 트위터에서는 10개 이슈 중 3위를 차지할 정도로 비교적 활발하게 언급되었으며, 그 결과 공급 대비 수요가 가장 높은 것으로 나타나게 되었음을 알 수 있다. 한편 뉴스의 공급 수량이 가장 많았던 Issue7의 경우 트위터에서 언급된 빈도수는 크게 많지 않으므로, 공급은 많지만 사용자들이 흥미를 보이지 않는 초과 공급의 정보로 나타났다.

<Figure 9>와 같은 이슈별 뉴스당 트윗 대응도를 2주의 기간에 걸쳐 분석한 결과가 <Figure 10>에 나타나있다.

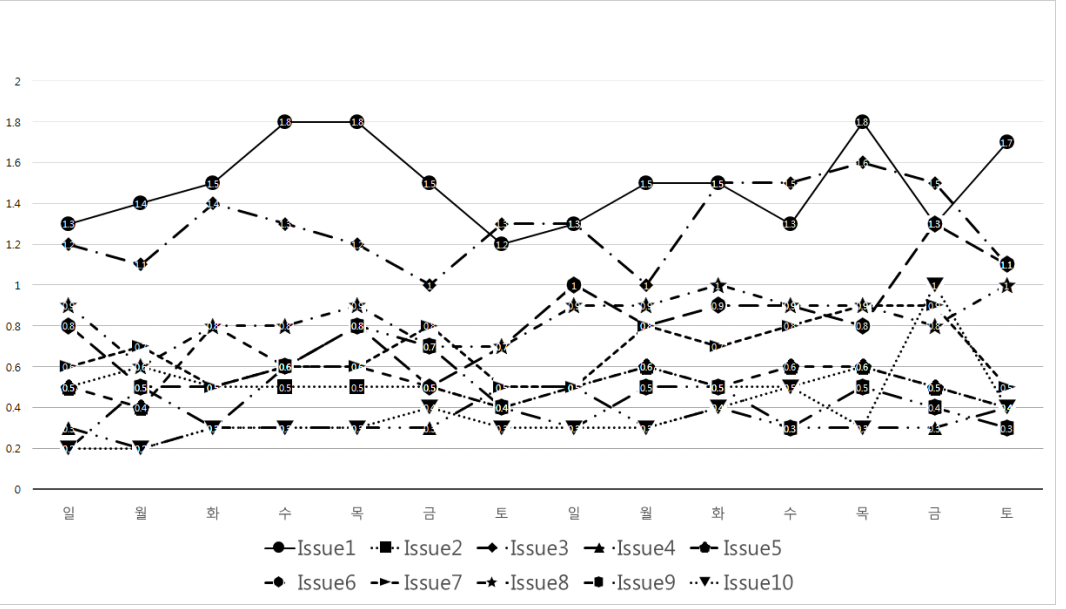
앞서 3장에서 언급한 바와 같이 <Figure 10>에서 토요일과 일요일의 경우 평일에 비해 뉴스당 트윗 대응도가 대부분의 이슈에 대해 높게 나타나는 현상을 발견하였다. 따라서 뉴스당 트윗 대응도가 일자별 특성에 의해 왜곡되는 현상을 방지하기 위해 <Figure 10>의 그래프를 일자별 기준 대응도에 대한 상대적 비율을 나타내게끔 보정하였으며, 이를 통해 도출된 2주간의 이슈별 NVI 값이 <Figure 11>에 나타나 있다.

<Figure 11>에 따르면 전체 10개의 이슈 중 8개의 이슈가 분석 기간 거의 전체에 걸쳐 기준 대응도에 미치지 못하는 NVI를 나타냄을 알 수 있었다. 이와 달리 두 개의 이슈(Issue1, Issue3)은 전체 기간 동안 기준 대응도를 훨씬 상회하며, 정보 시장의 전체 수요를 견인한 것으로 나타났다. 즉 10개 중 8개 이슈에 대한 뉴스는 대중의 관심에 비해 과잉 공급되고 있으며, 2개 이슈는 대중의 높은 관심에 비해 뉴스의 공급이 상대적으로 매우 부족한 것으로 나타났다.

본 연구에서는 정보의 수급 현황에 따라 정보



〈Figure 10〉 Changes in the ratio of tweets to news for each issue (June 22, 2014 to July 5, 2014)



〈Figure 11〉 NVI trend of issues (June 22, 2014 to July 5, 2014)

의 가치를 평가할 수 있는 방법론을 제시하였다. 제안 방법론을 통해 정보의 가치를 평가함으로써 언론사는 치가 높은 이슈, 즉 수요 대비 공급이 현저하게 부족한 이슈에 대한 정보의 생산에 집중하여 대중의 관심을 유도할 수 있으며, 궁극적으로 정보의 수요자인 대중은 관심 대상이 되는 정보를 더욱 풍족하게 접할 수 있을 것으로 기대한다.

5. 결론

최근 텍스트 분석을 통해 정보 유통의 주요 매체인 인터넷 뉴스와 SNS의 매체 간 특성 차이에 주목한 많은 연구가 발표되고 있다. 하지만 뉴스의 가치를 정보의 수요 및 공급 관점에서 파악한 연구는 상대적으로 매우 부족한 실정이며, 다중매체간에 유통되는 정보의 가치평가를 위해 다른 매체를 활용한 연구 역시 찾아보기 어렵다. 이에 본 연구에서는 특정 이슈에 대한 뉴스 정보로서의 가치를 이와 관련된 트위터의 양으로 평가하는 뉴스가치지수를 고안하여 제시하였다. 또한 실제 데이터에 대한 실험을 통해 NVI를 도출하고, 이를 시각화함으로써 시간의 흐름에 따른 뉴스 가치의 변화를 살펴보았다.

본 연구의 학술적, 실무적 기여는 다음의 측면에서 찾을 수 있을 것으로 기대한다. 우선 본 연구는 뉴스 기사를 정보의 공급 측면에서, 트위터 데이터를 정보의 수요를 나타내는 측면에서 파악하여 정보의 수급에 따른 가치를 측정하였으며, 이는 정보의 가치 평가를 위한 새로운 기준을 제시하였다는 점에서 학술적 기여가 인정될 수 있다. 또한 실무적 측면에서는 제안 방법론을 통해 여러 언론사는 가치가 높은 정보의 생산에

집중함으로써 수요에 부응하는 양질의 정보를 제공할 수 있고, 대중들은 관심 대상이 되는 정보를 더욱 풍족하고 쉽게 접할 수 있을 것으로 기대된다.

본 연구에서 제안하는 방법론은 향후 다음과 같은 측면에서 보완될 필요가 있다. 가장 시급한 보완점으로 각 이슈에 대응하는 뉴스 및 트윗을 식별하는 부분의 정교화가 필요하다. 물론 강화된 용어 임계값을 사용하여 핵심 키워드만을 대상 문서 식별에 적용했지만, 그럼에도 불구하고 단 하나의 핵심 키워드만을 포함한 문서를 해당 이슈에 대응하는 문서로 간주한다는 것은 해당 이슈와 무관한 문서가 분석에 포함될 위험성을 내포한다. 따라서 향후 각 이슈에 대응하는 문서를 보다 정확히 식별하기 위한 방법론의 정교화가 필요하다. 또한 본 연구의 실험에서는 이슈의 개수를 임의로 10개로 지정하였으나, 향후 연구에서는 이슈 개수의 설정 단계에서도 체계적인 기준이 적용되어야 한다. 마지막으로 현재는 본 방법론의 최종 결과물인 NVI 그래프를 통해 각 이슈별 가치의 변화를 직관적으로 해석하는 것에 그치고 있지만, 향후 NVI의 추이를 패턴화하여 다양한 이슈의 NVI 변화를 유형화하고 이를 통해 NVI의 활용도를 더욱 높일 필요가 있다.

참고문헌(References)

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

An, J. Y., J. H. Bae, N. G. Han and M. Song, "A Study of "Emotion Trigger' by Text Mining Techniques," *Journal of Intelligence and Information Systems*, Vol.21, No.2(2015), 69~92.

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.

Choi, S. I., Y. J. Hyun and N. Kim, "Improving Performance of Recommendation Systems Using Topic Modeling," *Journal of Intelligence and Information Systems*, Vol.21, No.3(2015), 103~118.

Choi, S. J., J. W. Lee and O. B. Kown, "A Morphological Analysis Method of Predicting Place-Event Performance by Online News Titles," *The Journal of Society for e-Business Studies*, Vol.21, No.1(2016), 15~32.

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

Hur, S. H., K. S. Choi, "A Study on Characteristics and Types of Tweet in Twitter," *Hanminjok Emunhak*, Vol.61(2012), 455~494.

Jin, S. A., G. E. Heo, Y. K. Jeong and M. Song, "Topic-Network based Topic Shift Detection on Twitter," *Journal of the Korean Society for Information Management*, Vol.30, No.1(2013), 285~302.

Jo, H. J., J. H. Seo and J. T. Choi, "OAR Algorithm Technology based on Opinion Mining Utilizing Stock News Contents," *Journal of KIIT*, Vol.13, No.3(2015),

111~119.

Jung, Y. I., W. J. Nam, *Introducing Translation Studies(Theories and Applications)*, Hankuk University of Foreign Studies Knowledge Press, 2006.

Jung, H. J., J. H. Bae, S. L. Hong, C. U. Park and M. Song, "Analysis of Twitter Public Opinion in Different Political Views : A Case Study of Sewol Ferry Accident," *Korean Journal of Journalism and Communication Studies*, Vol.60, No.2(2016), 269~302.

Kang, A. T., *A Study on Regional Characteristics on The Stress Sentiment and Topics Extracted from Tweet Data*, The Graduate School of Ewha Womans University, 2016.

Kim, D. S., W. X. S. Wong, M. S. Lim, C. Liu, N. Kim, J. H. Park, W. Y. Kil and H. S. Yoon, "A Methodology for Analyzing Public Opinion about Science and Technology Issues Using Text Analysis," *Journal of Information Technology Services*, Vol.14, No.3(2015), 33~48.

Kim, D. Y., *Study of TV Ration Prediction through Analysis of On-Line Bigdata : Case of The Drama in My Love from The Star*, The Graduate School of Chungbuk National University, 2016.

Kim, D. Y., J. W. Park and J. H. Choi, "A Comparative Study between Stock Price Prediction Models Using Sentiment Analysis and Machine Learning based on SNS and News Articles," *Journal of Information Technology Services*, Vol.13, No.3(2014), 221~233.

Kim, J. E., N. 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.

Kim, M. J., H. J. Jung, "A Case Study on Visual Expression through Interaction with Information Types - Focusing on Interactive Infographic in The New York Times -," *Journal of the Korean Society of Design Culture*, Vol.20, No.1(2014), 146~158.

Lee, J. H., K. S. Song, J. A. Kang and J. R. Hwang, "A Study on The Efficient Extraction Method of SNS Data related to Crime Risk Factor," *Journal of The Korea Society of Computer and Information*, Vol.20, No.1(2015), 255~263.

Lee, S. Y., K. M. Lee, "A Reply Graph-based Social Mining Method with Topic Modeling," *Journal of Korean Institute of Intelligent Systems*, Vol.24, No.6(2014), 640~645.

Lee, Y. J., J. H. Seo and J. T. Choi, "Fashion Trend Marketing Prediction Analysis based on Opinion Mining Applying SNS Text Contents," *Journal of KIIT*, Vol.12, No.12 (2014), 163~170.

Lim, H. J., S. H. Park, "A Tentative Approach for Regional Futures Strategy with Big Data - Through The Analysis Using The Data of SNS and Newpaper," *Journal of The Korean Cadastre Information Association*, Vol.17, No.1(2015), 75~90.

Lim, M. S. 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.

Lim, M. S. and N. Kim, "Investigating Dynamic Mutation Process of Issues Using Unstructured Text Analysis," *Journal of*

Intelligence and Information Systems, Vol.22, No.1(2016), 1~18.

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

Munday, J., *Introducing Translation Studies: Theories and Applications, 4th Edition*, Routledge, New York, 2016.

Noh, B. J., Z. S. Xu, J. U. Lee, D. H. Park and Y. H. Chung, "Keyword Network based Repercussion Effect Analysis of Foot-and-Mouth Disease Using Online News," *Journal of KIIT*, Vol.14, No.9(2016), 143~152.

Park, S. H., "SNS News Communication - Multiplicity and Orality," *Journal of communication research*, Vol.49, No.2(2012), 37~73.

Reiss, K., "Type, Kind and Individuality of Text : Decision Making in Translation," *Translation Theory and Intercultural Relations*, Vol.2, No.4(1981), 121~131.

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.

Weiss, S. M., N. Indurkhya, and T. Zhang, *Fundamentals of Predictive Text Mining, 2nd Edition*, Springer, 2015.

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

Yu, E. J., Y. S. 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

A Method for Evaluating News Value based on Supply and Demand of Information Using Text Analysis

Donghoon Lee* · Hochang Choi** · Namgyu Kim***

Given the recent development of smart devices, users are producing, sharing, and acquiring a variety of information via the Internet and social network services (SNSs). Because users tend to use multiple media simultaneously according to their goals and preferences, domestic SNS users use around 2.09 media concurrently on average. Since the information provided by such media is usually textually represented, recent studies have been actively conducting textual analysis in order to understand users more deeply. Earlier studies using textual analysis focused on analyzing a document's contents without substantive consideration of the diverse characteristics of the source medium. However, current studies argue that analytical and interpretive approaches should be applied differently according to the characteristics of a document's source.

Documents can be classified into the following types: informative documents for delivering information, expressive documents for expressing emotions and aesthetics, operational documents for inducing the recipient's behavior, and audiovisual media documents for supplementing the above three functions through images and music. Further, documents can be classified according to their contents, which comprise facts, concepts, procedures, principles, rules, stories, opinions, and descriptions.

Documents have unique characteristics according to the source media by which they are distributed. In terms of newspapers, only highly trained people tend to write articles for public dissemination. In contrast, with SNSs, various types of users can freely write any message and such messages are distributed in an unpredictable way. Again, in the case of newspapers, each article exists independently and does not

* Graduate School of Business IT, Kookmin University
** School of Business Administration, Kookmin University
*** Corresponding Author: Namgyu Kim
School of Management Information Systems, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 136-702, Korea
Tel: +82-2-910-5425, Fax: +82-2-910-4017, E-mail: ngkim@kookmin.ac.kr

tend to have any relation to other articles. However, messages (original tweets) on Twitter, for example, are highly organized and regularly duplicated and repeated through replies and retweets.

There have been many studies focusing on the different characteristics between newspapers and SNSs. However, it is difficult to find a study that focuses on the difference between the two media from the perspective of supply and demand. We can regard the articles of newspapers as a kind of information supply, whereas messages on various SNSs represent a demand for information. By investigating traditional newspapers and SNSs from the perspective of supply and demand of information, we can explore and explain the information dilemma more clearly. For example, there may be superfluous issues that are heavily reported in newspaper articles despite the fact that users seldom have much interest in these issues. Such overproduced information is not only a waste of media resources but also makes it difficult to find valuable, in-demand information. Further, some issues that are covered by only a few newspapers may be of high interest to SNS users.

To alleviate the deleterious effects of information asymmetries, it is necessary to analyze the supply and demand of each information source and, accordingly, provide information flexibly. Such an approach would allow the value of information to be explored and approximated on the basis of the supply-demand balance. Conceptually, this is very similar to the price of goods or services being determined by the supply-demand relationship. Adopting this concept, media companies could focus on the production of highly in-demand issues that are in short supply.

In this study, we selected Internet news sites and Twitter as representative media for investigating information supply and demand, respectively. We present the notion of News Value Index (NVI), which evaluates the value of news information in terms of the magnitude of Twitter messages associated with it. In addition, we visualize the change of information value over time using the NVI. We conducted an analysis using 387,014 news articles and 31,674,795 Twitter messages. The analysis results revealed interesting patterns: most issues show lower NVI than average of the whole issue, whereas a few issues show steadily higher NVI than the average.

Key Words : Big Data, News Value Index, SNS, Text Mining, Topic Modeling

Received : November 19, 2016 Revised : December 13, 2016 Accepted : December 18, 2016
Publication Type : Regular Paper Corresponding Author : Namgyu Kim

저 자 소 개

이 동 훈
현재 국민대학교 비즈니스IT전문대학원 박사과정에 재학 중이다.
한국방송통신대학교 컴퓨터과학과에서 학사 학위를 취득하고, 국민대학교 비즈니스IT전문대학원에서 비즈니스IT를 전공하여 석사 학위를 취득하였다. 주요 관심분야는 텍스트 마이닝, 데이터 마이닝 및 온톨로지 등이다.

최 호 창
현재 국민대학교 경영학부 학사과정에 재학 중이며, 2017년도 3월 비즈니스IT전문대학원 석사과정 입학 예정이다.
주요 관심분야는 데이터 마이닝 및 텍스트 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 자문위원으로 활동 중이다. 주요 관심분야는 텍스트 마이닝, 데이터 마이닝 및 데이터베이스 설계 등이다.