ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2017 September: 23(3): 69~94 http://dx.doi.org/10.13088/jiis.2017.23.3.069

전역 토픽의 지역 매핑을 통한 효율적 토픽 모델링 방안

최호창
국민대학교 비즈니스IT전문대학원
(*hochangchoi@kookmin.ac.kr*)

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

최근 빅데이터 분석 수요의 지속적 증가와 함께 관련 기법 및 도구의 비약적 발전이 이루어지고 있으며, 이에 따라 빅데이터 분석은 소수 전문가에 의한 독점이 아닌 개별 사용자의 자가 수행 형태로 변모하고 있다. 또한 전통적 방법으로는 분석이 어려웠던 비정형 데이터의 활용 방안에 대한 관심이 증가하고 있으며, 대표적으로 방대한 양의 텍스트에서 주제를 도출해내는 토픽 모델링(Topic Modeling)에 대한 연구가 활발히 진행되고 있다.

전통적인 토픽 모델링은 전체 문서에 걸친 주요 용어의 분포에 기반을 두고 수행되기 때문에, 각 문서의 토픽 식별에는 전체 문서에 대한 일괄 분석이 필요하다. 이로 인해 대용량 문서의 토픽 모델링에는 오랜 시간이 소요되며, 이 문제는 특히 분석 대상 문서가 복수의 시스템 또는 지역에 분산 저장되어 있는 경우 더욱 크게 작용한다. 따라서 이를 극복하기 위해 대량의 문서를 하위 군집으로 분할하고, 각 군집별 분석을 통해 토픽을 도출하는 방법을 생각할 수 있다. 하지만 이 경우 각 군집에서 도출한 지역 토픽은 전체 문서로부터 도출한 전역 토픽과 상이하게 나타나므로, 각 문서와 전역 토픽의 대응 관계를 식별할 수 없다.

따라서 본 연구에서는 전체 문서를 하위 군집으로 분할하고, 각 하위 군집에서 대표 문서를 추출하여 축소된 전역 문서 집합을 구성하고, 대표 문서를 매개로 하위 군집에서 도출한 지역 토픽으로부터 전역 토픽의 성분을 도출하는 방안을 제시한다. 또한 뉴스 기사 24,000건에 대한 실험을 통해 제안 방법론의 실무 적용 가능성을 평가하였으며, 이와 함께 제안 방법론에 따른 분할 정복(Divide and Conquer) 방식과 전체 문서에 대한 일괄 수행 방식의 토픽 분석 결과를 비교하였다.

주제어 : 분할 정복 접근법, 빅데이터, 텍스트 마이닝, 토픽 모델링

논문접수일 : 2017년 7월 21일 논문수정일 : 2017년 9월 14일 게재확정일 : 2017년 9월 19일
원고유형 : 일반논문 교신저자 : 김남규

1. 서론

최근 빅데이터 분석 수요의 지속적 증가와 함께 관련 기법 및 도구의 비약적 발전이 이루어지고 있다. 또한 IT 기술의 발달 및 스마트기기의 보급률 증가로 인해 많은 양의 데이터가 생성되고 있으며, 관련 분석 기술 또한 빠른 속도로 대중화됨에 따라 이를 통해 새로운 통찰(Insight)을 얻고자 하는 시도가 지속적으로 증가하고 있다. IDC(International Data Corporation)의 최근 보고는 세계 빅데이터 시장이 매년 11.9%의 지속적 성장을 할 것으로 예상하고 있으며, 2020년에는 해당 시장의 규모가 2,100억 달러에 이를 것으로 전망하고 있다(IDC, 2017). 이렇듯 빅데이터 분

석은 가까운 미래에 다양한 산업 분야의 핵심 기술로 자리잡으며 그 중요성이 더욱 강조될 것으로 예상된다.

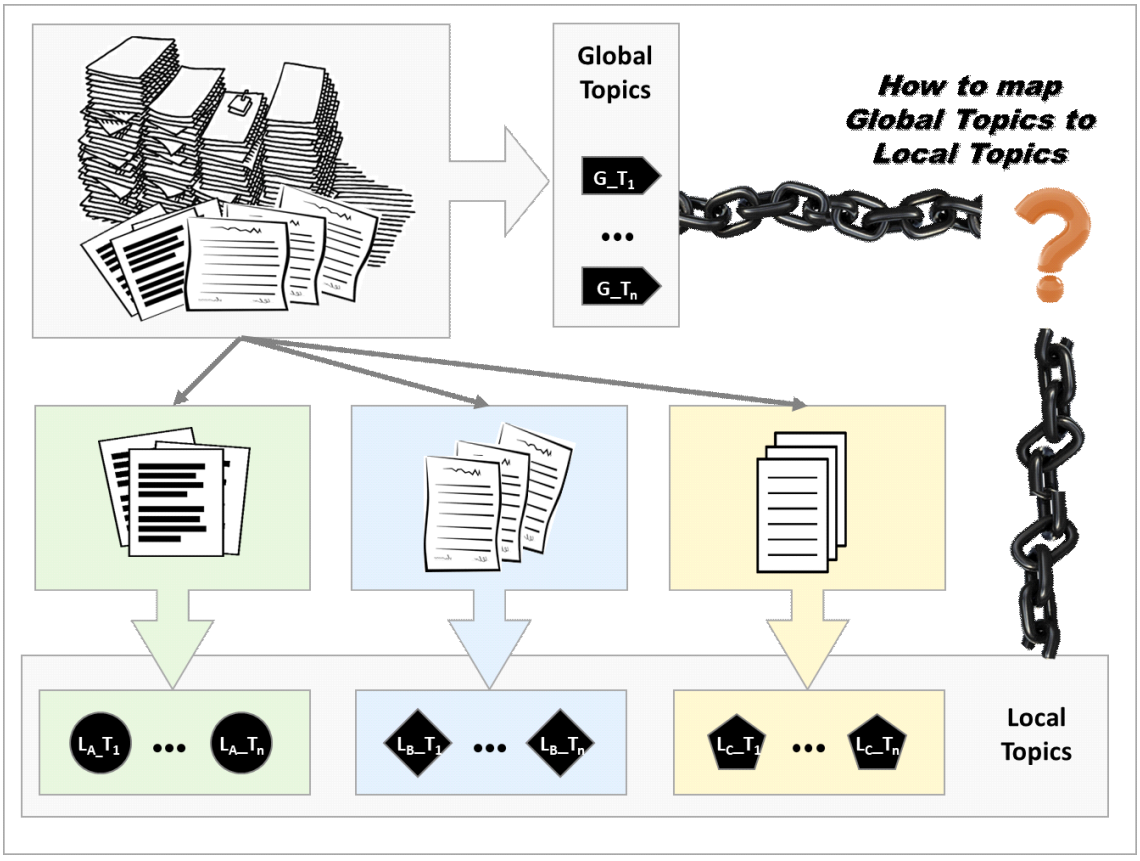
기존의 빅데이터 분석은 대부분 전문적 지식을 갖춘 소수의 사람들에 의해 수행되어 각 수요자에게 전파되었다. 하지만 최근 컴퓨터 프로그래밍 교육의 활성화와 다양한 상용 프로그램의 개발은 분석에 대한 진입장벽을 점차 낮추고 있으며, 최신 기술 동향에서는 빅데이터 분석을 시민 데이터 과학(Citizen Data Science) 및 자가 서비스 전달을 수반한 고급 분석(Advanced Analytics with Self-Service Delivery)과 같은 새로운 용어로 표현하고 있다(Gartner, 2015). 이에 따라 향후 빅데이터 분석은 전문가 중심의 수행에 머물지 않고 사용자의 자가 분석에 의한 맞춤형 분석 서비스 형태로 제공되며, 수요자의 필요에 의해 특정 분석 결과만을 선별 및 통합하는 방향으로 진화할 것으로 예상된다.

이러한 현상과 함께 비정형 데이터의 분석 기술이 대중화됨에 따라 다양한 비정형 데이터의 수집, 분석, 그리고 활용에 대한 관심이 증가하고 있다. 특히 의사소통의 매체로 가장 널리 사용되는 텍스트 데이터에 대한 분석 시도가 급증하고 있으며, 소셜 네트워크 서비스(Social Network Service, SNS)로 대표되는 웹 플랫폼의 활성화와 바이럴 마케팅(Viral Marketing) 등 웹을 이용한 신종 기법의 등장에 따라 텍스트 데이터의 양은 기하급수적 증가 추세를 보이고 있다. 따라서 텍스트의 분석을 통해 유용한 정보를 도출하고 이를 산업 각 분야에 활용하기 위한 노력이 매우 활발하게 이루어지고 있으며, 이와 더불어 텍스트 분석을 위한 이론 및 기법을 총칭하는 텍스트 마이닝(Text Mining)에 대한 관심 역시 고조되고 있다.

텍스트 마이닝의 다양한 응용 가운데, 특히 방대한 양의 문서로부터 주요 이슈를 추출하고 각 이슈에 해당하는 문서를 식별하여 이들을 군집으로 제공하는 토픽 모델링에 대한 연구(Kim et al., 2017; Steyvers and Griffiths, 2007)가 매우 활발하게 수행되고 있다. 토픽 모델링은 일반적 문서 군집화 기법과는 달리 문서의 의미적 요소를 반영한 결과를 제시하기 때문에 매우 유용한 기법으로 평가받는다. 하지만 전통적인 토픽 모델링은 전체 문서에 걸친 주요 용어의 분포에 기반을 두고 수행되기 때문에, 각 문서의 토픽을 식별하기 위해서는 전체 문서에 대한 일괄 분석이 이루어져야 한다. 이로 인해 대용량 문서의 토픽 모델링에는 매우 오랜 시간이 소요되며, 분석 대상 규모가 증가함에 따라 분석 시간이 지수적으로 증가하는 확장성(Scalability)의 문제가 발생한다. 이러한 현상은 특히 여러 문서들이 복수의 시스템 또는 지역에 분산 저장되어 있는 경우 더욱 큰 부담으로 작용할 수 있다.

이러한 한계를 극복하기 위해 대량의 문서를 하위 군집으로 분할하고 각 군집별 분석을 통해 토픽을 도출하는 방안, 즉 토픽 모델링에 분할정복 접근법을 적용하는 방안을 생각할 수 있다. 이는 대량의 문서에 대한 일괄 토픽 모델링을 수행하는 대신 분할된 소량의 문서에 대한 토픽 모델링을 반복 수행하는 방법으로, 제한된 시스템 자원으로 대량의 문서에 대한 분석을 수행할 수 있으며 분석 속도 측면에서도 효율성을 기대할 수 있다. 또한 문서가 여러 지역 또는 사이트에 산재해 있는 경우, 이들을 모두 취합하여 분석을 수행할 필요 없이 각 지역 또는 사이트에서 분석을 수행하는 것이 가능하기 때문에 시간과 비용을 매우 절약할 수 있을 것이다.

하지만 이러한 방식, 즉 전체 문서를 하위 군



〈Figure 1〉 Difficulties in Mapping Global and Local Topics

집으로 분할하여 각 군집별 토픽을 도출하는 방식은 크게 두 가지 측면에서 난제를 야기한다. 우선 이렇게 도출된 지역 토픽들과 애초에 도출하고자 했던 전역 토픽들, 즉 전체 문서에 대한 일괄 분석을 통해 도출할 수 있었던 토픽들과의 관계가 불명확하다는 것이다. 예를 들어 <Figure 1>은 전체 문서에 대한 토픽 모델링을 통해 전역 토픽(Global Topics)을 도출하기 어려운 상황에서, 전체 문서를 세 개의 군집(L_A, L_B, L_C)으로 분할하고 각 군집별로 토픽 모델링을 수행하여 지역 토픽(Local Topics)을 도출한 예이다. 이 때 <Figure 1>에서 $L_A\_T_1 \sim L_A\_T_n$, $L_B\_T_1 \sim L_B\_T_n$, $L_C_T_1 \sim L_C_T_n$의 지역 토픽들로부터 $G_T_1 \sim G_T_n$의 전역 토픽을 도출해 낼 수 있는 방법이 없다. 이는 각 문서에 존재하는 지역 토픽을 식별할 수는 있지만, 각 문서가 어떤 전역 토픽을 다루고 있는지 파악할 수 있는 방법이 없음을 의미한다. 또 다른 문제는 방법론의 정확성 측면에서 발생한다. 즉 전체 문서에 대해 일괄적으로 수행한 토픽 모델링 결과를 가장 이상적인 정답이라고 가정했을 때, 분할 정복 접근법을 통해 수행한 토픽 모델링 결과가 이와 얼마나 차이가

있는지 측정하기 위한 방안이 마련되어야 한다. 이러한 어려움으로 인해 분할 정복 접근법에 따라 토픽 모델링을 수행하는 방안에 대한 연구는 토픽 모델링을 다룬 다른 연구들에 비해 상대적으로 충분히 이루어지지 않았다.

본 연구에서는 위의 두 가지 난제를 해결하기 위한 방안을 다음과 같이 제시하고자 한다. 우선 전체 문서를 하위 군집으로 분할하고, 각 하위 군집에서 대표 문서(Delegate Documents)를 추출하여 축소된 전역 문서 집합(RGS, Reduced Global Set)을 구성한 후, 대표 문서를 통해 RGS에 대한 토픽 모델링 결과와 각 지역 토픽 모델링 결과를 대응시킴으로써 첫 번째 난제를 해결하고자 한다. 또한 이상적인 분석에서 동일 토픽으로 식별된 문서들이 분할 정복 접근법에 따른 분석을 통해 여전히 동일 토픽으로 식별되는지 여부를 파악하여 제안 방법론의 정확도를 파악하고자 한다.

본 연구의 이후 구성은 다음과 같다. 우선 2장에서는 텍스트 분석, 토픽 모델링, 그리고 토픽 모델링의 성능 이슈 등 본 연구와 관련한 선행 연구의 성과를 요약하고, 3장에서는 본 연구에서 제안하는 방법론을 간단한 예를 통해 소개한다. 또한 4장에서는 제안 방법론을 실제 뉴스 데이터에 적용한 결과를 분석하고, 마지막 장인 5장에서는 본 연구의 기여 및 한계를 요약한다.

2. 관련 연구

2.1 텍스트 분석

최근 4차 산업혁명에 대한 관심이 증가함과 함께 인공지능(Artificial Intelligence) 및 빅데이터 처리 기술에 대한 수요가 급증하고 있다. 이와 더불어 해당 기술의 적용 대상이 되는 다양한 데이터에 대한 관심 또한 증가하고 있으며, 특히 텍스트로 대표되는 비정형 데이터를 분석에 적용하기 위한 시도가 학계 및 산업계 전반에서 매우 활발하게 시도되고 있다. 텍스트는 현실 세계에서 정보의 표현과 교환을 위해 사용되는 가장 대표적인 의사소통 수단으로(Witten, 2004), 최근 웹 플랫폼의 활성화에 따라 방대한 양의 텍스트 데이터가 생성됨과 함께 텍스트 정보 분석의 중요성이 더욱 강조되고 있다.

이러한 텍스트의 분석을 위해서는 전통적인 데이터 마이닝 기법을 포함하는 여러 분야의 기술들이 포괄적으로 사용되며, 해당 기술들은 텍스트 마이닝이라는 새로운 분야의 기술로 분류되고 있다(Hotho et al., 2005; Mooney and Bunescu, 2006; Sebastiani, 2006). 텍스트 마이닝은 일반적으로 문서 수집, 파싱(Parsing), 필터링(Filtering), 구조화, 빈도 분석 및 유사도 분석의 순서로 수행되며, 특히 자연어 형태로 저장되는 텍스트를 벡터 또는 행렬과 같은 형식으로 정형화하는 기술인 구조화는 텍스트 분석의 핵심기술이라 할 수 있다. 텍스트 구조화에는 각 문서를 용어의 출현 빈도에 따라 벡터로 표현하여 문서의 특성을 나타내는 벡터공간모델(Vector Space Model)이 주로 사용된다(Salton, 1971; Salton et al., 1975). 이에 따라 각 용어의 빈도수를 고려하여 이를 가중치로 활용하기 위한 여러 지표들이 개발되어 왔으며, 이 중 용어의 절대 빈도수 및 상대적 빈도수를 동시에 고려한 지표인 TF-IDF(Term Frequency – Inverse Document Frequency) 및 이를 변형한 지표가 가장 대중적으로 사용되고 있다(Han et al., 2011). 이렇게 구조화된 텍스트 데이터는 (문서 수) × (용어 수)의

행렬로 표현될 수 있으며, 이때 각 문서는 용어 수에 해당하는 만큼의 차원을 갖는다. 그 결과 전체 문서의 행렬은 매우 높은 차원으로 구성되기 때문에, 이 행렬을 변환 과정 없이 분석에 사용하는 것은 매우 비현실적이다. 따라서 이를 분석 가능한 수준의 차원으로 표현하기 위한 차원 축소기법과 함께, 특정 문서들에서 동시에 출현하는 용어들의 의미적 유사성을 활용하기 위한 기법에 대한 연구가 활발하게 수행되고 있다.

2.2 토픽 모델링

토픽 모델링은 각 문서를 임의의 주제들로 구성된 집합으로 간주하고, 각 문서를 구성하는 주제와 각 주제에 해당하는 용어의 중요도를 확률적으로 제시하는 기법으로 정의할 수 있다(Kim et al., 2017; Steyvers and Griffiths, 2007). 이때 문서들 간의 유사도는 기본적으로 코사인 유사도(Cosine Similarity)를 이용하여 측정될 수 있으며, 용어의 의미적 유사도를 반영하기 위한 여러 기법들이 지속적으로 개발되고 있다. 특히 텍스트 분석에 주로 활용되는 상용 프로그램인 SAS Enterprise Miner은 잠재 의미 분석(Latent Semantic Analysis)을 적용하여 토픽 모델링을 수행하고 있으며(Deerwester et al., 1990; Koll, 1979), R을 비롯한 오픈소스 소프트웨어를 사용하는 많은 연구에서는 잠재 디리클레 할당(Latent Dirichlet Allocation)을 적용하여 해당 분석을 수행하고 있다.

토픽 모델링은 기본적으로 유사성에 따라 분석 대상을 군집으로 형성하여 제시하기 때문에 전통적인 군집화의 한 부류로 파악할 수 있다. 하지만 토픽 모델링은 각 개체가 여러 군집에 동시에 포함될 수 있는 연성 군집화(Soft Clustering)을 사용한다는 점과 분석 결과를 단순한 식별자가 아닌 의미를 가진 용어 집합으로 제시한다는 점에서 일반적인 군집화 기법과는 차이가 있다. 이러한 특징으로 인해 토픽 모델링은 현재 여러 분야에서 매우 다양한 목적으로 활용되고 있다(Byun et al., 2016; Kim and Kim, 2016; Lee et al., 2016; Livermore et al., 2016). 하지만 토픽 모델링의 활용이 증가함에 따라 해당 기법의 구조적 한계에 대한 지적도 활발하게 제기되고 있다.

2.3 토픽 모델링의 성능 이슈

전통적 토픽 모델링의 대표적인 구조적 한계는 분석 대상으로 새로운 문서가 추가될 경우, 이를 결과에 반영하기 위해 전체 문서에 대한 분석을 다시 수행해야 한다는 점이다. 이러한 한계는 전체 분석의 시간 및 비용의 막대한 증가를 야기할 뿐 아니라, 분석 대상이 끊임없이 생성되는 현실 상황에서 해당 기법의 적용을 매우 비효율적으로 만든다. 따라서 이러한 한계를 보완함과 함께 새로운 방법론을 통해 알고리즘을 개선시키기 위한 여러 연구들이 지속적으로 수행되고 있다(AlSumait et al., 2008; Blei and Lafferty, 2006). 특히 본 연구에서 주목하고 있는 개념인 분할 정복 접근법은 방대한 양의 문제를 해결이 용이한 단위로 분할하고, 여기서 도출한 결과를 통합하여 전체 문제를 해결할 수 있다는 개념으로 정의될 수 있다. 이 개념은 전통적인 데이터 마이닝 분야에서는 지리적으로 떨어진 위치의 데이터들을 효율적으로 군집화 하기 위한 연구(Forman and Zhang, 2000), 데이터의 밀도에 기반하여 군집화 알고리즘을 효율적으로 개선하기 위한 연구(Liang and Chen, 2016), 다차원의 대규

모 데이터 집합을 효율적으로 군집화하기 위한 캐노피 알고리즘을 소개한 연구(McCallum et al., 2000) 등에서 다양한 문제의 해결을 위해 활용된 바 있다.

한편 토픽 모델링에서의 분할 정복 접근법은 기본적으로 대규모 데이터를 시간에 따른 하위 군집으로 분할하여 각 군집별 이슈를 추출하고, 이렇게 도출된 이슈들 간의 유사성을 파악하여 시간에 따른 이슈들의 흐름을 확인하는 단계로 구성된다(Mei and Zhai, 2005). 또한 토픽 모델링의 성능을 다룬 최근 연구로는 텍스트 데이터에서 주기적, 반복적으로 등장하는 패턴을 효과적으로 추출하기 위한 프레임워크를 제안한 연구(Wang et al., 2015), 방대한 양의 주제 관련 웹 페이지를 효율적으로 군집화 하기 위한 새로운 알고리즘을 제안한 연구(Wang et al., 2009), 캐노피 알고리즘 기반의 텍스트 클러스터링을 위한 새로운 알고리즘을 제안한 연구(Song et al., 2012) 등이 있다. 하지만 토픽 모델링의 성능 향상을 위해 대량의 문서를 분할하여 군집별 토픽을 추출하고, 이들 군집 또는 군집별 토픽의 통합을 통해 다시 전체 문서의 토픽을 도출하는 방안을 제안한 연구는 찾아보기 어렵다.

3. 제안 방법론

3.1 연구 모형

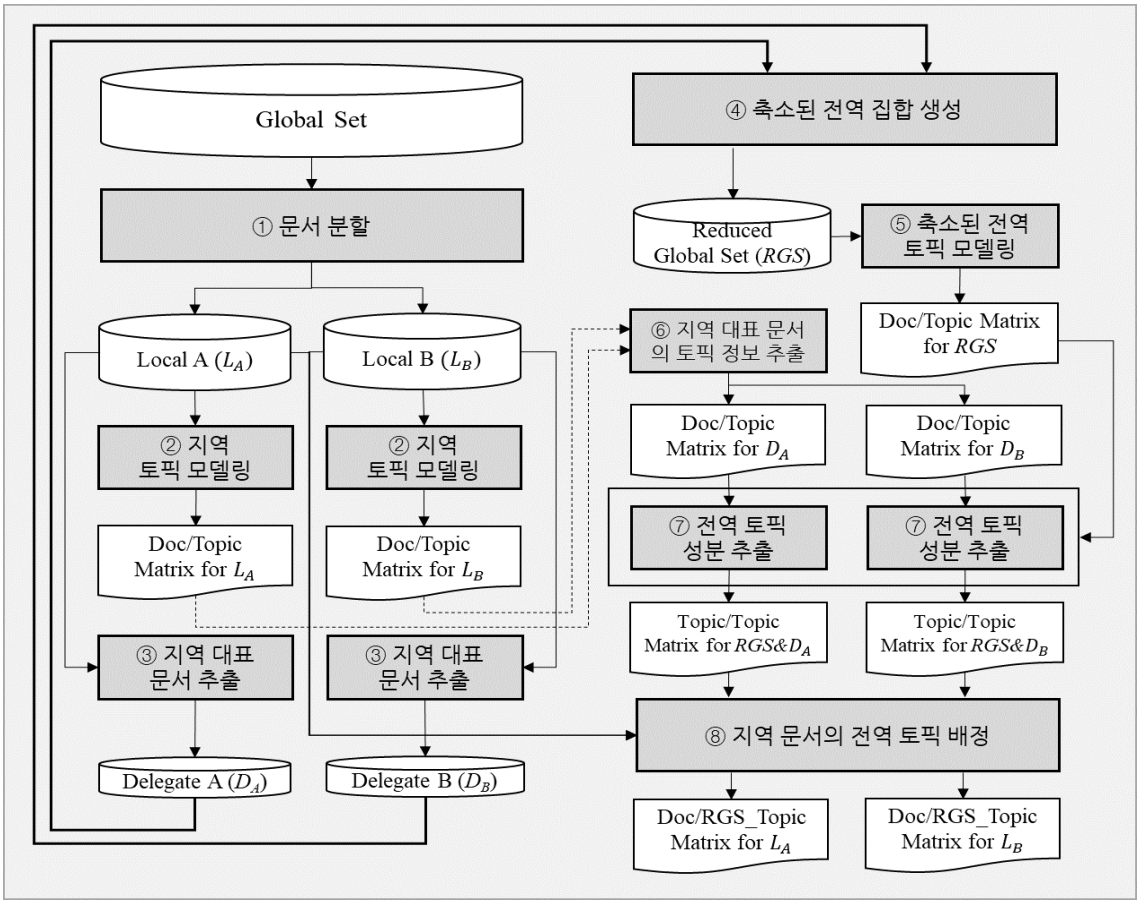
본 장에서는 효율적 토픽 모델링 수행을 위한 방안을 제시한다. 구체적으로 대량의 문서를 하위 군집으로 분할한 후 군집별로 지역 토픽을 도출하고, 군집별 대표 문서를 추출하여 전역 토픽을 도출한 뒤, 전역 토픽과 지역 토픽의 관계를 파악하여 각 문서에 할당하는 방안을 제시한다. 제안 방법론의 전체적인 개요는 <Figure 2>를 통해 제시하며, 분석 단계에 따른 구체적 설명은 이후 절에서 다루도록 한다.

우선 분석을 위해 수집된 문서 집합인 전역 군집(Global Set)을 하위 지역 군집(Local Set)으로 분할한 후, 이들을 대상으로 지역 군집별 주요 토픽을 추출한다. 다음으로 각 지역 군집에서 일부 문서를 임의로 선발하여 지역 대표 문서(Delegate)로 지정하며, 이들을 통합하여 모든 문서의 특질을 대표할 수 있는 축소된 전역 집합을 생성한다. 이후 축소된 전역 집합으로부터 전역 토픽을 추출하고 이를 지역 대표 문서의 지역 토픽 정보와 비교함으로써, 지역 토픽으로부터 전역 토픽의 성분을 추출하는 규칙을 도출한다. 마지막으로 이 규칙을 각 문서의 지역 토픽 가중치에 적용하여 각 문서의 전역 토픽을 배정한다. 본 장에서는 가상 예를 통해 제안 방법론의 각 단계를 소개하며, 실제 데이터를 분석한 결과는 다음 장인 4장에서 제시한다.

3.2 문서 군집 생성 및 토픽 추출

본 절은 <Figure 2>의 ① ~ ⑤에 해당하는 과정, 즉 전체 문서의 분할을 통해 생성한 지역 군집에서 각 지역 토픽을 추출하는 과정과 각 지역 군집에서 추출된 대표 문서로 구성되는 축소된 전역 집합에서 전역 토픽을 추출하는 과정을 설명한다.

우선 전체 문서를 적정 수의 하위 지역 군집으로 분할한다(①). 군집의 수는 문서의 수 또는 문서를 관리하는 기관 및 사이트와 같은 다양한 기준에 의해 결정될 수 있다. 이때 각 군집별 주제 및 시기는 서로 동일하거나 상이할 수 있다.



〈Figure 2〉 Research Overview

<Figure 2>에서는 설명의 편의를 위해 두 개의 군집으로 분할하는 경우를 표현하였으나, 제안 방법론은 군집의 수와 무관하게 동일한 방식으로 동작한다. 이와 더불어 각 지역 군집별로 토픽 모델링을 수행하여 군집을 구성하는 문서들의 주요 지역 토픽을 추출한다(②). 해당 분석의 결과로, 지역 군집에 속한 문서들의 토픽 정보를 식별하는 지역 군집별 문서/토픽 행렬이 <Figure 3>과 같이 생성된다. <Figure 3>은 두 군집의 주제가 상이한 경우를 가정하고 있으며, *Local A*와 *Local B*는 각각 정치 및 생활/문화 관련 문서로 구성되어 있다.

토픽 모델링은 많은 텍스트 분석 연구에서 문서의 구조화 및 주요 토픽 추출을 목적으로 활용되고 있다. 이 기법의 주요 원리를 간략히 요약하면 다음과 같다. 우선 토픽 모델링은 일반적으로 단어 꾸러미(Bag of Words) 개념을 사용하며, 이는 각 문서를 해당 문서에서 등장한 용어들의 집합으로 인식한다. 각 문서는 많은 수의 단어를 포함하기 때문에 차원 축소 과정을 통해 적절한

Local A _ Topic	
1	정부, 문재인, 내각
2	이낙연, 청문회, 아들
3	북핵, 사드, 중국
4	정유라, 기각, 강부영
5	귀국, 특사, j노믹스

Local A	Topic1	Topic2	Topic3	Topic4	Topic5
Doc. 1	0.013	0.009	0.048	0.022	0.021
Doc. 2	0.020	0.062	0.030	0.022	0.045
Doc. 3	0.010	0.490	0.035	0.005	-0.011
Doc. 4	0.027	0.013	0.081	0.020	0.069
Doc. 5	0.021	0.031	0.002	0.005	0.066
…					

(a) Topics and Document/Topic Matrix of *Local A*

Local B _ Topic	
1	초여름, 날씨, 일교차
2	식단, 다이어트, 열량
3	고속도로, 주말, 정체
4	축제, 제주, 숙박
5	여행, 해외, 수면

Local B	Topic1	Topic2	Topic3	Topic4	Topic5
Doc. 10	0.028	0.015	0.026	0.021	0.007
Doc. 11	0.033	0.014	0.040	0.031	0.010
Doc. 12	0.088	0.010	0.103	0.006	0.036
Doc. 13	0.385	0.036	0.018	0.010	-0.007
Doc. 14	0.057	0.010	0.017	0.022	0.007
…					

(b) Topics and Document/Topic Matrix of *Local B*

〈Figure 3〉 Example of Document/Topic Matrix for Local Sets

수의 단어 군집으로 표현되며, 이 과정에서 도출된 차원의 수가 토픽의 수에 해당한다. 이에 따라 각 문서는 개별 토픽에 대한 대응도인 문서 가중치(Document Topic Weight)를 갖게 되며, 일반적으로 문서 가중치의 "평균 + 1σ"을 통해 산출되는 문서 임계값(Document Cutoff)을 통해 각 문서의 토픽 포함 여부를 판단한다. 즉, 문서 임계값 이상의 문서 가중치를 갖는 문서가 해당 토픽을 포함하고 있는 것으로 해석된다. 또한 이 과정을 통해 각 문서와 토픽의 문서 가중치를 2차원 행렬로 나타낸 것을 문서/토픽 행렬(Document/Topic Matrix)이라 한다.

다음으로 전역 토픽 추출을 위해 각 지역 군집에서 단순 무작위 추출법(Simple Random Sampling)을 통해 각 군집을 대표하는 문서를 추출하고(③), 이들을 통합하여 축소된 전역 집합 군집인 RGS를 생성한다(④). RGS로부터 전역 토픽을 추출하는 과정은 일반적인 토픽 모델링과 동일하다(⑤). 예를 들어 <Figure 4>는 지역 군집 *Local A*와 *Local B*의 대표 문서로 각각 *Doc. 1, 3, 5*와 *10, 12, 14*가 선정되어 RGS를 구성하는 경우를 보인다. <Figure 4>의 <Figure 3(a)>와 <Figure 3(b)>에 나타난 토픽과 유사하지만 완전히 동일하지는 않은 전역 토픽이 발견되며, 이 때 문서/토픽 행렬은 대표 문서와 전역 토픽간의 대응도를 나타낸다.

RGS _ Topic	
1	이낙연, 청문회, 김상조
2	초여름, 식단, 다이어트
3	정유라, 기각, 재판
4	축제, 고속도로, 주말
5	귀국, 해외, 수면

RGS	Topic1	Topic2	Topic3	Topic4	Topic5
Doc. 1	0.312	0.024	0.003	0.004	0.050
Doc. 3	0.032	0.032	0.006	0.035	0.020
Doc. 5	0.073	0.007	0.015	0.020	0.013
Doc. 10	0.100	0.011	0.025	0.009	0.036
Doc. 12	0.106	0.019	0.023	0.026	-0.015
Doc. 14	0.218	0.017	0.006	0.020	0.110

〈Figure 4〉 Example of Document/Topic Matrix for RGS

3.3 전역 토픽 성분 도출 및 지역 문서의 전역 토픽 배정

본 절은 <Figure 2>의 ⑥ ~ ⑧에 해당하는 과정, 즉 지역 대표 문서를 활용하여 지역 토픽으로부터 전역 토픽의 성분을 도출하고, 이를 통해 각 문서 모두에 대해 전역 토픽을 할당하는 과정을 소개한다. 이후 설명에서 "Local i"의 "Topic j"를 $L_i_T_j$로, RGS의 "Topic k"를 RGS_T_k로 나타내기로 한다.

지역 대표 문서의 경우 <Figure 3>의 지역 토픽 모델링 뿐 아니라 <Figure 4>의 전역 토픽 모델링에도 참여하기 때문에, 지역 토픽과 전역 토

Local A	Doc. 1	Doc. 3	Doc. 5
$L_A\_T_1$	0.013	0.010	0.021
$L_A\_T_2$	0.009	0.490	0.031
$L_A\_T_3$	0.048	0.035	0.002
$L_A\_T_4$	0.022	0.005	0.005
$L_A\_T_5$	0.021	-0.011	0.066

×

RGS	RGS_T_1	RGS_T_2	RGS_T_3	RGS_T_4	RGS_T_5
Doc. 1	0.312	0.024	0.003	0.004	0.050
Doc. 3	0.032	0.032	0.006	0.035	0.020
Doc. 5	0.073	0.007	0.015	0.020	0.013

=

RGS / Local A	RGS_T_1	RGS_T_2	RGS_T_3	RGS_T_4	RGS_T_5
$L_A\_T_1$	0.006	0.001	0.000	0.001	0.001
$L_A\_T_2$	0.021	0.016	0.003	0.018	0.011
$L_A\_T_3$	0.016	0.002	0.000	0.001	0.003
$L_A\_T_4$	0.007	0.001	0.000	0.000	0.001
$L_A\_T_5$	0.011	0.001	0.001	0.001	0.002

(a) *Local A* to RGS Topic Weight Conversion Matrix

Local B	Doc. 10	Doc. 12	Doc. 14
$L_B\_T_1$	0.028	0.088	0.057
$L_B\_T_2$	0.015	0.010	0.010
$L_B\_T_3$	0.026	0.103	0.017
$L_B\_T_4$	0.021	0.006	0.022
$L_B\_T_5$	0.007	0.036	0.007

×

RGS	RGS_T_1	RGS_T_2	RGS_T_3	RGS_T_4	RGS_T_5
Doc. 10	0.100	0.011	0.025	0.009	0.036
Doc. 12	0.106	0.019	0.023	0.026	-0.015
Doc. 14	0.218	0.017	0.006	0.020	0.110

=

RGS / Local B	RGS_T_1	RGS_T_2	RGS_T_3	RGS_T_4	RGS_T_5
$L_B\_T_1$	0.025	0.003	0.003	0.004	0.006
$L_B\_T_2$	0.005	0.001	0.001	0.001	0.001
$L_B\_T_3$	0.017	0.003	0.003	0.003	0.001
$L_B\_T_4$	0.008	0.001	0.001	0.001	0.003
$L_B\_T_5$	0.006	0.001	0.001	0.001	0.000

(b) *Local B* to RGS Topic Weight Conversion Matrix

〈Figure 5〉 Local to RGS Topic Weight Conversion Matrix

픽의 정보를 모두 갖고 있다. 예를 들어 *Local A*의 *Doc.1*은 지역 토픽 모델링의 결과로 $L_A\_T_1$ ~ $L_A_T_5$의 지역 토픽 5개에 대해 (0.013, 0.009, 0.048, 0.022, 0.021)의 문서 가중치를 가지며, 이와 동시에 전역 토픽 모델링의 결과로 RGS_T_1 ~ RGS_T_5의 전역 토픽 5개에 대해 (0.312, 0.024, 0.003, 0.004, 0.050)의 문서 가중치를 갖는다. 본 절의 목적은 $L_A\_T_1$ ~ $L_A_T_5$으로부터 RGS_T_1, RGS_T_2, RGS_T_3, RGS_T_4, RGS_T_5의 값을 예측하는 규칙 $Rule_A$와 $L_B\_T_1$ ~ $L_B_T_5$으로부터 RGS_T_1, RGS_T_2, RGS_T_3, RGS_T_4, RGS_T_5의 값을 예측하는 규칙 $Rule_B$에 대한 도출 과정을 소

Local A	$L_A\_T_1$	$L_A\_T_2$	$L_A\_T_3$	$L_A\_T_4$	$L_A\_T_5$
Doc. 1	0.013	0.009	0.048	0.022	0.021
Doc. 2	0.020	0.062	0.030	0.022	0.045
Doc. 3	0.010	0.490	0.035	0.005	-0.011
Doc. 4	0.027	0.013	0.081	0.020	0.069
Doc. 5	0.021	0.031	0.002	0.005	0.066
…					

×

RGS / Local A	RGS_T_1	RGS_T_2	RGS_T_3	RGS_T_4	RGS_T_5
$L_A\_T_1$	0.006	0.001	0.000	0.001	0.001
$L_A\_T_2$	0.021	0.016	0.003	0.018	0.011
$L_A\_T_3$	0.016	0.002	0.000	0.001	0.003
$L_A\_T_4$	0.007	0.001	0.000	0.000	0.001
$L_A\_T_5$	0.011	0.001	0.001	0.001	0.002

=

RGS	RGS_T_1	RGS_T_2	RGS_T_3	RGS_T_4	RGS_T_5
Doc. 1	0.001	0.000	0.000	0.000	0.000
Doc. 2	0.003	0.001	0.000	0.001	0.001
Doc. 3	0.011	0.008	0.002	0.009	0.005
Doc. 4	0.003	0.000	0.000	0.000	0.001
Doc. 5	0.002	0.001	0.000	0.001	0.000
…					

(a) Assignment of Global Topic Weights to *Local A*

Local B	$L_B\_T_1$	$L_B\_T_2$	$L_B\_T_3$	$L_B\_T_4$	$L_B\_T_5$
Doc. 10	0.028	0.015	0.026	0.021	0.007
Doc. 11	0.033	0.014	0.040	0.031	0.010
Doc. 12	0.088	0.010	0.103	0.006	0.036
Doc. 13	0.385	0.036	0.018	0.010	-0.007
Doc. 14	0.057	0.010	0.017	0.022	0.007
…					

×

RGS / Local B	RGS_T_1	RGS_T_2	RGS_T_3	RGS_T_4	RGS_T_5
$L_B\_T_1$	0.025	0.003	0.003	0.004	0.006
$L_B\_T_2$	0.005	0.001	0.001	0.001	0.001
$L_B\_T_3$	0.017	0.003	0.003	0.003	0.001
$L_B\_T_4$	0.008	0.001	0.001	0.001	0.003
$L_B\_T_5$	0.006	0.001	0.001	0.001	0.000

=

RGS	RGS_T_1	RGS_T_2	RGS_T_3	RGS_T_4	RGS_T_5
Doc. 10	0.001	0.000	0.000	0.000	0.000
Doc. 11	0.002	0.000	0.000	0.000	0.000
Doc. 12	0.004	0.001	0.001	0.001	0.001
Doc. 13	0.010	0.001	0.001	0.001	0.002
Doc. 14	0.002	0.000	0.000	0.000	0.000
…					

(b) Assignment of Global Topic Weights to *Local B*

〈Figure 6〉 Assignment of Global Topic Weights to Local Documents

개하는 것이다. 이를 위해 제안 방법론은 각 지역 토픽 모델링의 결과로 나타난 문서/토픽 행렬에서 지역 대표 문서에 해당하는 부분만을 발췌하고(⑥), 이를 RGS에 대한 토픽 모델링을 통해 도출된 문서/토픽 행렬과 비교한다. 지역 토픽으로부터 전역 토픽의 성분을 도출하는 구체적인 과정은 <Figure 5>의 예를 통해 설명할 수 있다.

<Figure 5(a)> 및 <Figure 5(b)>는 각각 *Local A*와 *Local B*로부터 전역 토픽의 성분을 도출하는 예를 보이고 있다. 예를 들어 <Figure 5(a)>에서 $L_A_T_1 \sim L_A_T_5$으로부터 RGS_T_1의 가중치를 도출하는 과정을 살펴보자. <Figure 5(a)>의 세 번째 테이블에서 점선 사각형으로 표시된 '0.021'이라는 값은 $L_A\_T_2$ 값의 '1' 증가가 RGS_T_1 값의 '0.021'의 증가를 가져옴을 나타낸다. 즉 RGS_T_1는 다른 지역 토픽들에 비해 $L_A_T_2$의 영향을 많이 받으며, 이러한 결과는 지역 대표 문서 *Doc. 1, Doc. 3, Doc. 5*에서 $L_A_T_2$의 값이 높을수록 RGS_T_1의 값이 높게 나타나는 현상을 반영하고 있다. 제안 방법론에서는 지역 대표 문서를 매개로 하여 지역 토픽 가중치를 전역 토픽 가중치로 변환하는 방법을 제안하며, 실제 변환은 행렬 곱 연산을 활용하여 수행하였다(⑦).

제안 방법론의 마지막 단계는 앞에서 도출한 전역 토픽 성분 도출 결과를 활용하여 모든 지역 문서에 전역 토픽 대응도를 배정하는 과정으로(⑧), 자세한 과정은 <Figure 6>을 통해 설명한다.

<Figure 6(a)>에서 좌측 상단의 문서/토픽 행렬은 <Figure 3(a)>에서 도출된 것이며, 우측 상단의 토픽 가중치 변환 행렬은 <Figure 5(a)>에서 도출한 것이다. 규칙의 적용을 위해 몇 가지 표기법을 정리하여 <Table 1>에 제시하였다.

<Table 1>의 표기법을 사용하여, *Local A*에 속한 문서 *Doc. i*의 지역 토픽 가중치로부터 전역 토픽 가중치를 도출하기 위한 규칙을 정의하면 다음과 같다(단 N은 *Local A*의 전체 지역 토픽 수).

$$d_i(RGS\_T_k) = \sum_{j=1}^{N} w(L_A\_T_j, RGS\_T_k) \times d_i(L_A\_T_j)$$

예를 들어 *Local A*에 속한 *Doc. 2*의 전역 토픽 RGS_T_1에 대한 문서 가중치를 구한 결과는 다음과 같다.

$$d_2(RGS\_T_1) = (0.006 \times 0.020) + (0.021 \times 0.062) + (0.016 \times 0.030) + (0.007 \times 0.022) + (0.011 \times 0.045) = \mathbf{0.003}$$

즉 *Doc. 2*의 전역 토픽 RGS_T_1에 대한 문서 가중치는 0.003으로 계산된다. 이러한 방식으로 지역 대표 문서뿐 아니라 대표에 포함되지 않은 모든 지역 문서에 대해 전역 토픽 문서 가중치를 배정할 수 있다.

〈Table 1〉 Notations for Deriving and Appling Weight Conversion Rules

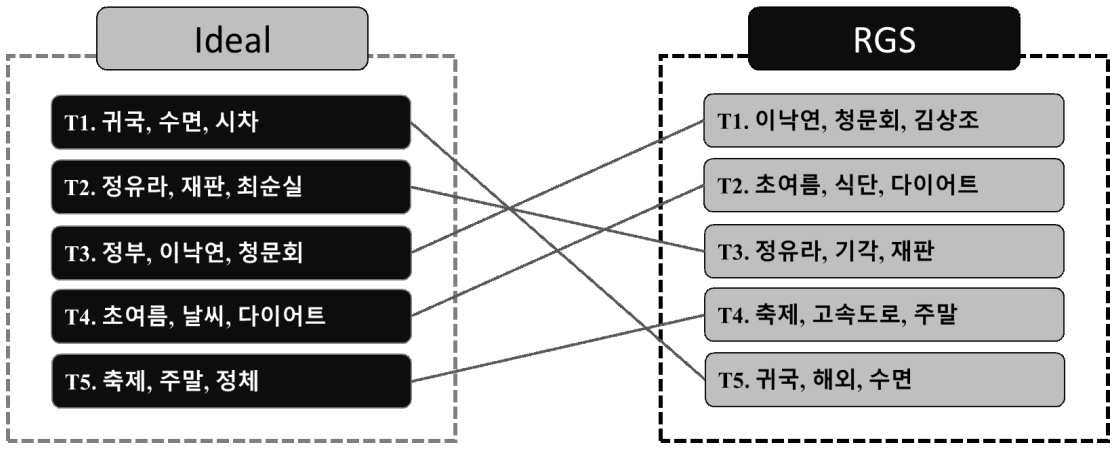
Notation	Description
$d_i(L_A\_T_j)$	문서 *Doc. i*의 *Local A* 지역 토픽 "Topic j"에 대한 문서 가중치
$d_i(RGS\_T_k)$	문서 *Doc. i*의 전역 토픽 "Topic k"에 대한 문서 가중치
$w(L_A\_T_i,\ RGS\_T_k)$	*Local A* 지역 토픽 "Topic i"의 전역 토픽 "Topic k" 변환에 대한 가중치

3.4 RGS의 대표성에 대한 검증 방안

본 절은 제안 방법론의 설명과는 별개로, 제안 방법론의 우수성을 평가하기 위한 검증 과정을 소개한다. 제안 방법론의 우수성은 크게 두 가지 측면에서 검증이 이루어져야 한다. 우선 RGS의 대표성에 대한 분석이 이루어져야 한다. 즉 제안 방법론은 RGS로부터 전역 토픽을 도출하지만, 어디까지나 RGS는 전체 문서가 아닌 전체 문서의 일부이므로 실제 전체 문서로부터 도출한 토픽과는 차이가 있을 수 있다. 따라서 실제 전체 문서에서 도출한 토픽과 제안 방법론의 RGS로부터 도출한 토픽과의 차이에 대한 분석이 이루어져야 한다. 또한 전역 토픽 배정의 정확성에 대한 분석이 수행되어야 한다. 즉 전체 문서에 대한 일괄 분석과 제안 방법론에 의한 분할 분석에 의해, 각 문서의 토픽 배정이 얼마나 상이하

V_Ideal / V_RGS	*Ideal_T1*	*Ideal_T2*	*Ideal_T3*	*Ideal_T4*	*Ideal_T5*
RGS_T1	0.254	0.157	0.857	0.270	0.145
RGS_T2	0.727	0.285	0.126	0.890	0.139
RGS_T3	0.136	0.925	0.275	0.129	0.385
RGS_T4	0.482	0.414	0.387	0.417	0.913
RGS_T5	0.937	0.237	0.216	0.375	0.423
Match_no	**5**	**3**	**1**	**2**	**4**

(a) Cosine Similarity between *Global Topics and RGS Topics*



(b) Topic Mapping between *Global Topics and RGS Topics*

〈Figure 7〉 Comparing Global Topics and RGS Topics

게 나타나는지 파악할 필요가 있다. 본 절에서는 RGS의 대표성에 대한 검증 방안에 대해서만 간략히 소개하며, 전역 토픽 배정의 정확성 분석에 대한 방안은 다음 장에서 실험과 함께 소개한다.

우선 전체 문서 모두에 대해 일괄 토픽 모델링을 수행하여 문서/토픽 행렬을 도출한다. 다음으로 이 행렬에서 RGS에 포함된 문서, 즉 지역 대표 문서들에 대한 부분만을 발췌하면, 각 문서는 실제 전역 토픽 수만큼의 차원을 갖는 벡터(V_Ideal)로 표현될 수 있다. 한편 이들 문서는 이미 <Figure 4>의 과정을 통해 RGS로부터 도출한 토픽에 대해 벡터(V_RGS)로 표현되어 있다. *V_Ideal*과 *V_RGS*는 동일한 문서에 대한 벡터이며, 동일한 수의 차원으로 구성된다. 따라서 이 두 벡터 간의 코사인 유사도를 계산하여 실제 전역 토픽과 RGS 토픽간의 유사도 행렬을 도출할 수 있다(<Figure 7(a)>).

<Figure 7(a)>를 구성하는 각 셀의 값은 해당 토픽 쌍의 코사인 유사도를 나타내며, 각 Ideal 토픽, 즉 실제 전역 토픽에 대해 가장 큰 값을 갖는 RGS 토픽을 찾을 수 있다. 예를 들어 $Ideal_T_1$의 경우 RGS 토픽들 중 RGS_T_5 와 가장 유사한 것으로 나타나며, 유사도는 “0.937”로 매우 높음을 의미한다. 이러한 방식으로 나머지 Ideal 토픽들에 대해서도 대응되는 RGS 토픽들을 식별할 수 있으며, 토픽 쌍의 코사인 유사도 평균이 높을수록 RGS 토픽이 실제 전역 토픽을 잘 대표한다고 볼 수 있다.

이상 본 장에서는 제안 방법론의 개념을 간단한 예를 통해 소개하였다. 다음 장인 4장에서는 제안 방법론을 통해 실제 뉴스 데이터를 분석한 실험 결과를 소개한다.

4. 실험

4.1 실험 개요

본 절에서는 실험 환경 및 데이터에 대해 간략히 소개한다. 실험의 핵심인 토픽 모델링은 SAS Enterprise Miner Workstation 14.1을 사용하여 수행하였으며, 별도의 한글 전용 형태소 분석기는 사용하지 않았다. 실험 데이터의 경우 2012년 7월부터 2013년 6월까지의 기간 동안 국내 한 포털사이트에 게시된 뉴스를 사용하였으며, IT과학, 정치, 경제, 사회, 생활, 세계, 스포츠, 연예의 총 8개의 카테고리에서 3,000건씩, 총 24,000건의 문서를 사용하였다.

4.2 제안 방법론의 적용

본 절에서는 3장에서 소개한 순서에 따라 실제 뉴스를 분석한 결과를 소개한다. 먼저 전체 24,000건의 뉴스를 각 2,400건씩 10개의 지역 군집으로 분할하고, 이들 각 군집에서 10개의 토픽과 토픽별 5개의 핵심 용어를 <Figure 8>과 같이 도출하였다. 이때 각 지역 군집은 토픽 모델링의 결과로 문서/토픽 행렬을 갖게 된다. <Table 2>는 Local 1 지역 군집의 문서/토픽 행렬의 일부를 보여준다.

다음으로 각 지역 군집에 포함된 문서의 일부를 무작위로 추출하여 지역 대표 문서를 선정하고, 이를 통합하여 RGS 데이터를 생성하였다. 본 실험에서는 각 지역 군집에서 문서의 1/10을 대표 문서로 추출하였으며, 그 결과 RGS는 총 2,400개의 문서로 구성되었다. 이후 RGS에 대한 토픽 모델링을 수행하여 10개의 주요 토픽을 추출하고, <Figure 9>과 같이 문서/토픽 행렬을 생성하였다.

Local 1		Local 2		Local 3		Local 4		Local 5	
T1	보조금, 영업정지, 가입자, 만원, 방통	T1	보조금, 요금, 가입자, 탤레콤, 번호	T1	분기, 영업이익, 실적, 전년, 배출	T1	경찰, 혐의, 검찰, 수사, 사건	T1	도, 기온, 눈, 지방, 낮
T2	특허, 소송, 법원, 배심원, 판결	T2	주택, 아파트, 가구, 부동산, 단지	T2	경찰, 혐의, 사건, 경찰서, 수사	T2	태풍, 볼라, 제주, 강풍, 영향	T2	엔진, 모델, 연비, 디자인, 뉴
T3	카메라, 화면, 기능, 스마트폰, 인치	T3	종목, 분기, 외국인, 주가, 저수	T3	주택, 아파트, 가구, 부동산, 전세	T3	모델, 수입, 자동차, 차량, 연비	T3	태풍, 볼라, 피해, 강풍, 초속
T4	요금, 무제한, 데이터, 음성, 통화	T4	특허, 소송, 법원, 판결, 기술	T4	사업, 계발, 드림, 출자, 계발사업	T4	학생, 학교, 폭력, 교사, 대학	T4	연구, 결과, 운동, 건강, 음식
T5	정보, 개인, 악성코드, 번호, 해킹	T5	사업, 계발, 출자, 드림, 허브	T5	카드, 체크, 신용카드, 서비스, 수수료	T5	도, 기온, 눈, 지방, 영하	T5	택시, 버스, 운행, 업계, 요금
T6	분기, 스마트폰, 시장, 점유율, 전자	T6	은행, 대출, 금융, 계속, 주택	T6	제품, 가격, 업계, 마트, 매장	T6	택시, 버스, 운행, 업계, 요금	T6	사진, 여성, 영화, 남성, 온라인
T7	발사, 나로호, 위성, 우주, 발사체	T7	스마트폰, 화면, 제품, 인치, 전자	T7	학생, 학교, 대학, 수능, 교육	T7	의료, 사업, 검찰, 전주, 노조	T7	수입, 판매, 가격, 현대, 국내
T8	서비스, 카카오, 가입자, 모바일, 콘텐츠	T8	회장, 그룹, 계열사, 경영, 지분	T8	회장, 그룹, 검찰, 혐의, 회사	T8	사고, 경찰, 차량, 화재, 분	T8	차량, 버스, 사고, 운행, 구간
T9	행성, 지구, 연구, 우주, 입자	T9	가격, 매출, 제품, 백화점, 매장	T9	연금, 대출, 은행, 국민연금, 상품	T9	아이, 사람, 여성, 환자, 병원	T9	피부, 치료, 환자, 증상, 제품
T10	미니, 태블릿, 윈도, 인치, 달러	T10	카드, 서비스, 신용카드, 정보, 결제	T10	환율, 경제, 달러, 수출, 정부	T10	연금, 국민연금, 기초, 소득, 보험료	T10	여행, 음식, 바음, 항공권, 맛

Local 6		Local 7		Local 8		Local 9		Local 10	
T1	후보, 대통령, 공화당, 대선, 선거인단	T1	감독, 코치, 구단, 사령탑, 한국시리즈	T1	감독, 코치, 야구, 대표팀, 구단	T1	영화, 감독, 관객, 피에타, 도둑	T1	후보, 단일화, 토론, 대선, 포인트
T2	사진, 네티즌, 해외, 화제, 영상	T2	안타, 홈런, 2, 1, 두	T2	영화, 사람, 영애, 친구, 결혼	T2	후보, 대선, 단일화, 정치, 토론	T2	회담, 남북, 대화, 북측, 수석대표
T3	교황, 추기경, 프란치스코, 가톨릭, 콘클라베	T3	올림픽, 금메달, 종목, 레슬링, 양학	T3	안타, 2, 두, 홈런, 1	T3	대통령, 정부, 당선인, 독도, 국민	T3	원내, 대표, 국회, 검찰, 여야
T4	테러, 용의자, 형제, 폭탄, 수사	T4	맨유, 퍼거슨, 레알, 영입, 유나이티드	T4	올림픽, 대회, 금메달, 여자, 포상금	T4	무대, 뮙스타, 악동, 뮤지션, 방에	T4	당선인, 인수위, 인선, 정부, 인사
T5	위안부, 발언, 유신회, 망언, 담화	T5	구단, 연봉, 계약, 달러, 협상	T5	경기, 본, 스완지, 측구, 리그	T5	드라마, 시청률, 본, 연기, 시청자	T5	후보자, 내정자, 청문회, 인사청문회, 의혹
T6	반군, 정부군, 화학무기, 정부, 국가	T6	경기, 본, 스완지, 리그, 후반	T6	구단, 연봉, 시즌, 계약, 아구	T6	화영, 멤버, 사진, 몸매, 네티즌	T6	수사, 검찰, 경찰, 의혹, 사건
T7	지진, 폭발, 사고, 지역, 피해	T7	광고, 뉴스, 경찰, 여성, 사람	T7	강남, 스타일, 뮤직비디오, 차트, 무대	T7	프로그램, 예능, 멤버, 무한, 도전	T7	미사일, 발사, 도발, 핵실험, 로켓
T8	여성, 경찰, 뉴스, 사건, 성폭행	T8	대회, 우승, 여자, 피겨, 연기	T8	맨유, 퍼거슨, 감독, 영입, 유나이티드	T8	결혼, 데일리, 열애, 결혼식, 사람	T8	대통령, 독도, 국민, 경제, 정부
T9	총기, 규제, 사건, 난사, 대통령	T9	선발, 경기, 투수, 자책점, 평균	T9	양학, 금메달, 기술, 제조, 올림픽	T9	의원, 검찰, 수사, 혐의, 사건	T9	원장, 출마, 대선, 경향, 경선
T10	열도, 독도, 시위, 영토, 시위대	T10	선수, 스완지, 아구, 시즌, 트레이드	T10	무한, 시청률, 도전, 방송, 멤버	T10	스타일, 강남, 차트, 뮤직비디오, 빌보드	T10	의원, 캠프, 경선, 공천, 위원장

〈Figure 8〉 Local Topics and Key Terms

〈Table 2〉 Document/Topic Matrix of Local 1 (Part)

	$L_1\_T_1$	$L_1\_T_2$	$L_1\_T_3$	$L_1\_T_4$	$L_1\_T_5$	$L_1\_T_6$	$L_1\_T_7$	$L_1\_T_8$	$L_1\_T_9$	$L_1\_T_{10}$
Doc. 1	0.013	0.009	0.048	0.022	0.021	0.010	0.081	0.012	0.131	0.004
Doc. 2	0.020	0.062	0.030	0.022	0.045	0.092	0.039	0.060	0.039	0.034
Doc. 3	0.010	0.490	0.035	0.005	-0.011	0.037	0.007	-0.020	-0.010	0.055
Doc. 4	0.027	0.013	0.081	0.020	0.069	0.049	0.015	0.109	0.023	0.032
Doc. 5	0.021	0.031	0.002	0.005	0.066	0.061	0.011	0.056	0.014	0.034
Doc. 6	0.015	0.014	-0.108	0.011	0.034	0.299	0.010	0.091	-0.008	0.200
Doc. 7	0.034	0.000	0.328	0.021	0.018	0.037	0.017	-0.018	0.010	0.111
Doc. 8	0.018	0.005	0.015	0.369	0.033	-0.013	0.048	-0.025	-0.001	-0.007
Doc. 9	0.016	0.020	0.256	0.018	0.016	0.082	0.013	-0.025	0.034	0.019
Doc. 10	0.405	0.019	0.127	0.014	0.060	0.191	0.010	-0.053	-0.027	0.036

Reduced Global Set	
T1	스마트폰, 제품, 특허, 인치, 화면
T2	경기, 시즌, 2, 선발, 1
T3	후보, 대선, 단일화, 정치, 선거
T4	경찰, 수사, 검찰, 혐의, 사건
T5	사진, 사람, 영화, 방송, 모습
T6	주택, 사업, 기업, 부동산, 은행
T7	감독, 구단, 코치, 야구, 시즌
T8	보조금, 요금, 휴대폰, 가입자, 이통사
T9	태풍, 도, 기온, 피해, 지역
T10	대통령, 정부, 후보자, 의원, 국회

(a) Topic of *RGS*

	RGS_T_1	RGS_T_2	RGS_T_3	RGS_T_4	RGS_T_5	RGS_T_6	RGS_T_7	RGS_T_8	RGS_T_9	RGS_T_{10}
Doc. 7	0.312	0.024	0.003	0.004	0.050	-0.040	0.009	0.083	0.014	0.000
Doc. 28	0.032	0.032	0.006	0.035	0.020	0.033	0.010	0.021	0.024	0.029
Doc. 32	0.073	0.007	0.015	0.020	0.013	0.031	0.016	0.365	-0.009	0.013
Doc. 40	0.100	0.011	0.025	0.009	0.036	0.061	0.010	0.157	0.031	0.127
Doc. 51	0.106	0.019	0.023	0.026	-0.015	0.174	0.021	0.009	0.021	0.076
Doc. 65	0.218	0.017	0.006	0.020	0.110	-0.015	0.005	0.008	-0.006	0.052
Doc. 84	0.177	0.024	0.021	0.044	0.018	0.011	0.010	0.159	0.009	0.039
Doc. 95	0.108	0.049	0.007	0.020	0.036	0.048	0.000	0.020	0.030	0.010
Doc. 107	0.139	0.032	0.003	0.016	0.014	0.030	0.012	0.020	0.013	0.006
Doc. 120	0.173	0.021	0.009	0.022	0.007	0.073	0.009	0.042	0.016	0.032

(b) Document/Topic Matrix of *RGS*

〈Figure 9〉 Topics and Document/Topic Matrix of RGS (Part)

위의 과정을 통해, RGS에 참여한 지역 군집의 대표 문서들은 RGS의 전역 토픽뿐 아니라 원 소속 지역 군집의 토픽 정보도 함께 가진다. 따라서 이들이 갖는 두 가지 유형의 문서/토픽 행렬에 대한 행렬 곱 연산을 수행하여, 지역 토픽 가중치로부터 전역 토픽 가중치를 도출하기 위한 규칙을 생성하였다. 실제 실험을 통해 도출한 Local 1 토픽과 RGS 토픽간 가중치 변환 행렬은 <Table 3>과 같다.

마지막으로 <Table 3>의 가중치 변환 행렬을 적용하여 <Table 2>에 제시된 Local 1 문서의 지역 토픽 가중치를 전역 토픽 가중치로 변환하였으며, 그 결과의 일부가 <Table 4>에 소개되어 있다. 이와 유사한 방식으로 Local 2 ~ Local 10에 대해서도 전역 토픽 변환 규칙을 도출하였으며, 각 군집의 규칙에 따라 모든 문서에 대해 전역 토픽을 배정하였다.

4.3 제안 방법론의 성능 분석

본 절에서는 제안 방법론의 성능을 RGS의 전체 문서에 대한 대표성과 전역 토픽 배정의 정확성 측면에서 분석한다.

우선 실제 전역 토픽의 도출을 위해 분석 대상 문서 24,000건 전체에 대한 일괄 토픽 모델링을 수행하여 10개의 토픽을 추출하고, 전체 문서에 대한 문서/토픽 행렬을 도출하였다. 다음으로 이 행렬에서 RGS에 포함된 문서 2,400개에 대한 부

〈Table 3〉 Local 1 to RGS Topic Weight Conversion Matrix

	RGS_T_1	RGS_T_2	RGS_T_3	RGS_T_4	RGS_T_5	RGS_T_6	RGS_T_7	RGS_T_8	RGS_T_9	RGS_T_{10}
$L_1\_T_1$	1.420	0.274	0.157	0.309	0.069	0.184	0.190	5.691	-0.028	0.267
$L_1\_T_2$	2.988	0.366	0.057	0.459	-0.258	-0.098	0.136	0.065	0.229	1.178
$L_1\_T_3$	3.683	0.364	0.133	0.276	0.520	0.017	0.181	1.411	0.246	0.334
$L_1\_T_4$	1.044	0.190	0.101	0.163	0.092	0.123	0.123	3.170	0.127	0.250
$L_1\_T_5$	0.744	0.124	0.133	0.311	0.231	0.458	0.099	0.739	0.155	0.347
$L_1\_T_6$	3.390	0.403	0.153	0.329	0.109	1.251	0.227	1.498	0.105	0.157
$L_1\_T_7$	0.622	0.231	0.111	0.217	0.276	0.168	0.073	0.373	0.495	0.360
$L_1\_T_8$	0.938	0.157	0.132	0.217	0.319	0.570	0.112	0.317	0.135	0.285
$L_1\_T_9$	0.380	0.135	0.113	0.176	0.337	0.287	0.075	-0.044	0.190	0.186
$L_1\_T_{10}$	2.241	0.211	0.096	0.197	0.292	0.402	0.129	0.638	0.130	0.159

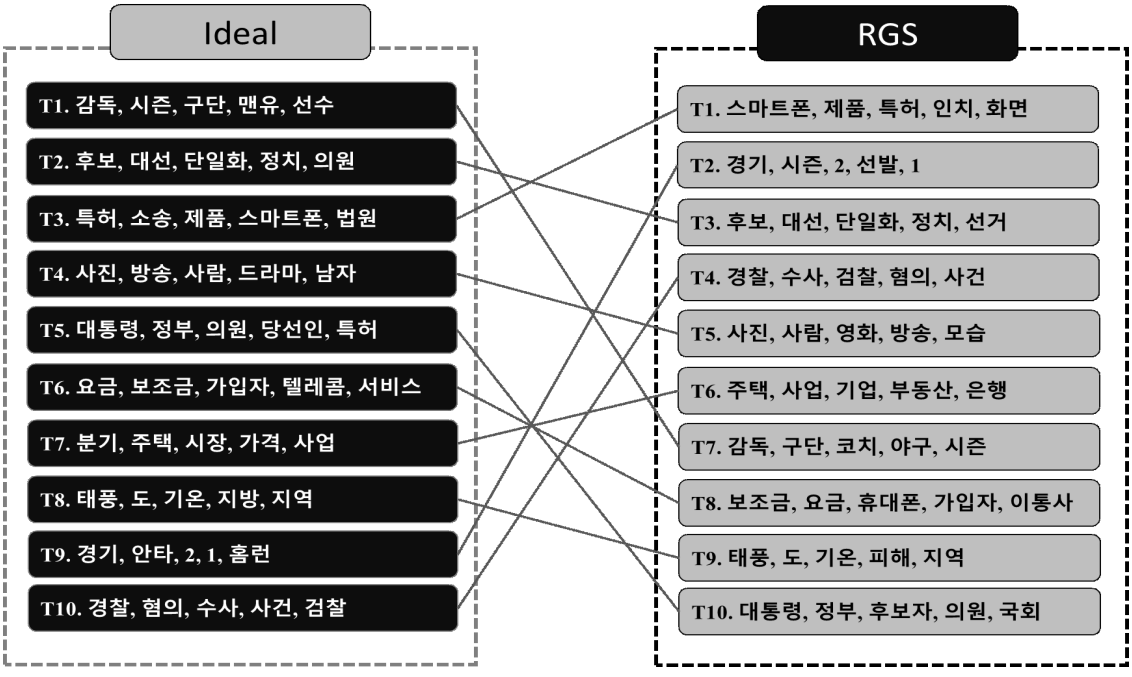
〈Table 4〉 Converted Document/RGS Topic Matrix (Part)

	RGS_T_1	RGS_T_2	RGS_T_3	RGS_T_4	RGS_T_5	RGS_T_6	RGS_T_7	RGS_T_8	RGS_T_9	RGS_T_{10}
Doc. 1	0.415	0.074	0.041	0.079	0.103	0.087	0.037	0.273	0.088	0.102
Doc. 2	0.864	0.117	0.053	0.126	0.076	0.202	0.063	0.454	0.080	0.168
Doc. 3	1.835	0.218	0.041	0.253	-0.097	0.005	0.088	0.233	0.130	0.599
Doc. 4	0.806	0.104	0.056	0.115	0.120	0.185	0.061	0.516	0.071	0.132
Doc. 5	0.532	0.072	0.037	0.086	0.052	0.159	0.042	0.323	0.044	0.104
Doc. 6	1.253	0.155	0.072	0.152	0.070	0.522	0.094	0.602	0.056	0.106
Doc. 7	1.663	0.176	0.070	0.146	0.218	0.109	0.093	0.863	0.112	0.156
Doc. 8	0.452	0.087	0.046	0.082	0.051	0.038	0.052	1.304	0.074	0.126
Doc. 9	1.374	0.151	0.058	0.127	0.156	0.123	0.078	0.653	0.093	0.143
Doc. 10	1.833	0.249	0.115	0.246	0.112	0.321	0.151	2.871	0.053	0.217

분 행렬만을 추출하였다. 또한 이 행렬과 <Figure 9>의 RGS에 대한 문서/토픽 행렬 간에 대해 유사도 분석을 수행하여, 실제 전역 토픽 10개와 RGS 전역 토픽 10개 간의 코사인 유사도를 <Table 5>와 같이 행렬로 정리하였다. <Table 5>에서 각 행은 RGS의 전역 토픽을, 각 열은 실제 전역 토픽을 나타낸다. 이 실험에서 10개의 실제 전역 토픽 각각은 서로 다른 10개의 RGS 전역 토픽에 대응되었으며, 대응 결과는 <Figure 10>에 나타나있다. 또한 대응되는 토픽의 유사도는 평균 "0.8545"로 매우 높게 나타났으며, 이는 RGS를 통해 도출한 전역 토픽이 전체 문서에 대

〈Table 5〉 Cosine Similarity between Global Topics and RGS Topics

	Ideal_T1	*Ideal_T2*	*Ideal_T3*	*Ideal_T4*	*Ideal_T5*	*Ideal_T6*	*Ideal_T7*	*Ideal_T8*	*Ideal_T9*	*Ideal_T10*
RGS_T1	0.206	0.158	0.802	0.270	0.117	0.606	0.434	0.175	0.222	0.223
RGS_T2	0.727	0.137	0.178	0.298	0.151	0.172	0.192	0.146	0.885	0.203
RGS_T3	0.136	0.925	0.115	0.138	0.382	0.119	0.107	0.095	0.138	0.201
RGS_T4	0.366	0.414	0.386	0.414	0.551	0.368	0.427	0.314	0.389	0.820
RGS_T5	0.519	0.260	0.257	0.831	0.312	0.270	0.289	0.233	0.494	0.409
RGS_T6	0.261	0.241	0.389	0.272	0.423	0.453	0.841	0.292	0.318	0.409
RGS_T7	0.919	0.135	0.149	0.251	0.133	0.147	0.157	0.103	0.655	0.175
RGS_T8	0.161	0.139	0.396	0.137	0.125	0.891	0.293	0.129	0.170	0.182
RGS_T9	0.263	0.188	0.275	0.331	0.246	0.248	0.325	0.849	0.265	0.400
RGS_T10	0.249	0.535	0.356	0.297	0.782	0.294	0.337	0.208	0.250	0.407
Match_no	**7**	**3**	**1**	**5**	**10**	**8**	**6**	**9**	**2**	**4**



〈Figure 10〉 Mapping between Global Topics and RGS Topics

한 일괄 분석을 통해 도출한 실제 전역 토픽을 잘 설명하고 있음을 의미한다.

추가로 본 실험에서는 대응되는 토픽간 내용의 부합 정도를 파악하기 위해, 각 토픽을 구성하는 주요 용어의 일치 수준을 확인하였다. 구체적으로 실제 전역 토픽 10개와 RGS 전역 토픽 10개, 총 20개 토픽에 대해 각 토픽별로 가장 높은 용어 가중치를 갖는 주요 용어 10개를 선정하여, 이들간 일치 수준을 확인하였다. 그 결과 대응 토픽간 주요 용어의 일치 수준은 10개 중 평균 "7.1개"로 비교적 높게 나타남을 확인하였다 (<Figure 11>).

Ideal_T1		*Ideal_T2*		*Ideal_T3*		*Ideal_T4*		*Ideal_T5*		*Ideal_T6*		*Ideal_T7*		*Ideal_T8*		*Ideal_T9*		*Ideal_T10*	
term1	감독	term1	후보	term1	특허	term1	사진	term1	대통령	term1	요금	term1	분기	term1	태풍	term1	경기	term1	경찰
term2	시즌	term2	대선	term2	소송	term2	방송	term2	정부	term2	보조금	term2	주택	term2	도	term2	안타	term2	혐의
term3	구단	term3	단일화	term3	제품	term3	사람	term3	의원	term3	가입자	term3	시장	term3	기온	term3	2	term3	수사
term4	맨유	term4	정치	term4	스마트폰	term4	드라마	term4	당선인	term4	텔레콤	term4	가격	term4	지방	term4	1	term4	사건
term5	선수	term5	의원	term5	법원	term5	남자	term5	특허	term5	서비스	term5	사업	term5	지역	term5	홈런	term5	검찰
term6	경기	term6	캠프	term6	디자인	term6	모습	term6	국회	term6	데이터	term6	달러	term6	눈	term6	시즌	term6	조사
term7	리그	term7	토론	term7	판결	term7	스타	term7	후보자	term7	스마트폰	term7	기업	term7	피해	term7	타자	term7	여성
term8	축구	term8	선거	term8	배심원	term8	영화	term8	인수위	term8	무제한	term8	업체	term8	볼라	term8	선발	term8	광고
term9	코치	term9	국민	term9	기술	term9	연예	term9	경제	term9	이통사	term9	국내	term9	제주	term9	루	term9	범행
term10	대표팀	term10	경선	term10	침해	term10	뉴스	term10	국민	term10	음성	term10	아파트	term10	낮	term10	볼넷	term10	경찰서

(a) Key Terms of *Global Topics*

RGS_T1		*RGS_T2*		*RGS_T3*		*RGS_T4*		*RGS_T5*		*RGS_T6*		*RGS_T7*		*RGS_T8*		*RGS_T9*		*RGS_T10*	
term1	스마트폰	term1	경기	term1	후보	term1	경찰	term1	사진	term1	주택	term1	감독	term1	보조금	term1	태풍	term1	대통령
term2	제품	term2	시즌	term2	대선	term2	수사	term2	사람	term2	사업	term2	구단	term2	요금	term2	도	term2	정부
term3	특허	term3	2	term3	단일화	term3	검찰	term3	영화	term3	기업	term3	코치	term3	휴대폰	term3	기온	term3	후보자
term4	인치	term4	선발	term4	정치	term4	혐의	term4	방송	term4	부동산	term4	야구	term4	가입자	term4	피해	term4	의원
term5	화면	term5	1	term5	선거	term5	사건	term5	모습	term5	은행	term5	시즌	term5	이통사	term5	지역	term5	국회
term6	디자인	term6	상대	term6	국민	term6	검사	term6	친구	term6	분기	term6	선수	term6	텔레콤	term6	지방	term6	당선인
term7	기능	term7	홈런	term7	투표	term7	의혹	term7	연예	term7	대출	term7	영입	term7	단말기	term7	눈	term7	인사
term8	전자	term8	타자	term8	캠프	term8	조사	term8	뉴스	term8	정부	term8	계약	term8	가격	term8	오후	term8	특허
term9	태블릿	term9	안타	term9	여론조사	term9	회장	term9	드라마	term9	금융	term9	리그	term9	스마트폰	term9	볼라	term9	인수위
term10	기술	term10	대회	term10	의원	term10	광고	term10	스타	term10	아파트	term10	영화	term10	번호	term10	제주	term10	문제

(b) Key Terms of *RGS Topics*

	Ideal_T1 & RGS_T7	*Ideal_T2 & RGS_T3*	*Ideal_T3 & RGS_T1*	*Ideal_T4 & RGS_T5*	*Ideal_T5 & RGS_T10*	*Ideal_T6 & RGS_T8*	*Ideal_T7 & RGS_T6*	*Ideal_T8 & RGS_T9*	*Ideal_T9 & RGS_T2*	*Ideal_T10 & RGS_T4*	*Total*
Concurrence	6	8	5	9	8	6	5	9	8	7	7.1

(c) Number of Co-occurred Key Terms in the Corresponding Topics

〈Figure 11〉 Analyzing Co-occurred Key Terms

다음으로 본 절에서는 전역 토픽 배정의 정확성 분석 결과를 소개한다. 이를 위해 본 실험에서는 일괄 토픽 모델링에서 동일한 토픽을 갖는 것으로 분류된 문서들이 제안 방법론에 의한 분석을 통해서도 여전히 동일한 토픽으로 분류되는지 여부를 측정하였다.

우선 모든 문서에 대해 각 문서가 특정 RGS 토픽을 포함하는지 여부를 결정하기 위해 문서 임계값을 설정하였으며, 이를 위해 일반적으로 사용되는 기준인 문서 가중치의 "평균 + 1σ"를 적용하였다. <Table 6>은 24,000개 전체 문서에 대한 RGS 토픽 10개의 문서 임계값을 산출한 결과를 나타낸다.

<Table 6>의 기준을 적용했을 때, 24,000개 전체 문서에 대한 일괄 분석으로 도출한 실제 전역 토픽 10개의 토픽별 문서 수는 <Figure 12(a)>에 나타나있다. 한편 <Figure 12(b)>는 각 RGS 토픽을 포함하는 것으로 식별된 문서의 수를 나타낸다. 또한 대응되는 토픽 쌍에 동시에 출현한 문서의 수는 <Figure 12(c)>에 요약되어 있다.

<Figure 12>에 따라 실제 전역 토픽의 문서와 RGS 전역 토픽의 문서 간 일치 수준을 분석하였으며, 평가에 사용한 척도는 <Table 7>과 같다. <Table 7>에서 *Freq(RGS_Ti)*는 RGS 토픽의 i번

〈Table 6〉 Document Cutoff for RGS Topics

	RGS_T_1 Doc.Weight	RGS_T_2 Doc.Weight	RGS_T_3 Doc.Weight	RGS_T_4 Doc.Weight	RGS_T_5 Doc.Weight	RGS_T_6 Doc.Weight	RGS_T_7 Doc.Weight	RGS_T_8 Doc.Weight	RGS_T_9 Doc.Weight	RGS_T_{10} Doc.Weight
Average	0.183	0.226	0.161	0.163	0.180	0.159	0.159	0.140	0.118	0.157
Sigma	0.298	0.407	0.401	0.130	0.167	0.180	0.323	0.308	0.161	0.198
Threshold	**0.481**	**0.633**	**0.562**	**0.292**	**0.347**	**0.339**	**0.482**	**0.448**	**0.280**	**0.355**

$Ideal\_T_1$ Doc.Num	$Ideal\_T_2$ Doc.Num	$Ideal\_T_3$ Doc.Num	$Ideal\_T_4$ Doc.Num	$Ideal\_T_5$ Doc.Num	$Ideal\_T_6$ Doc.Num	$Ideal\_T_7$ Doc.Num	$Ideal\_T_8$ Doc.Num	$Ideal\_T_9$ Doc.Num	$Ideal\_T_{10}$ Doc.Num
1708	1154	1611	3697	2617	1460	3006	1065	1549	2371

(a) Number of Documents in *Global Topics*

RGS_T_1 Doc.Num	RGS_T_2 Doc.Num	RGS_T_3 Doc.Num	RGS_T_4 Doc.Num	RGS_T_5 Doc.Num	RGS_T_6 Doc.Num	RGS_T_7 Doc.Num	RGS_T_8 Doc.Num	RGS_T_9 Doc.Num	RGS_T_{10} Doc.Num
1927	2295	1519	2565	3503	2656	2090	1314	1287	2075

(b) Number of Documents in *RGS Topics*

$Ideal\_T_1$ & RGS_T_7 Doc.Num	$Ideal\_T_2$ & RGS_T_3 Doc.Num	$Ideal\_T_3$ & RGS_T_1 Doc.Num	$Ideal\_T_4$ & RGS_T_5 Doc.Num	$Ideal\_T_5$ & RGS_T_{10} Doc.Num	$Ideal\_T_6$ & RGS_T_8 Doc.Num	$Ideal\_T_7$ & RGS_T_6 Doc.Num	$Ideal\_T_8$ & RGS_T_9 Doc.Num	$Ideal\_T_9$ & RGS_T_2 Doc.Num	$Ideal\_T_{10}$ & RGS_T_4 Doc.Num
1564	941	1141	2391	1394	980	1987	596	1455	1527

(c) Number of Co-occurred Documents in the Corresponding Topics

〈Figure 12〉 Number of Documents in Each Topic

〈Table 7〉 Accuracy Measures

Measure	Definition
Recall(RGS_T_i, Ideal_T_j)	Co_Freq(RGS_T_i, Ideal_T_j) / Freq(Ideal_T_j)
Precision(RGS_T_i, Ideal_T_j)	Co_Freq(RGS_T_i, Ideal_T_j) / Freq(RGS_T_i)
F1 Measure(RGS_T_i, Ideal_T_j)	$\frac{2 \times (Precision) \times (Recall)}{(Precision) + (Recall)}$

〈Table 8〉 Result of Accuracy Analysis

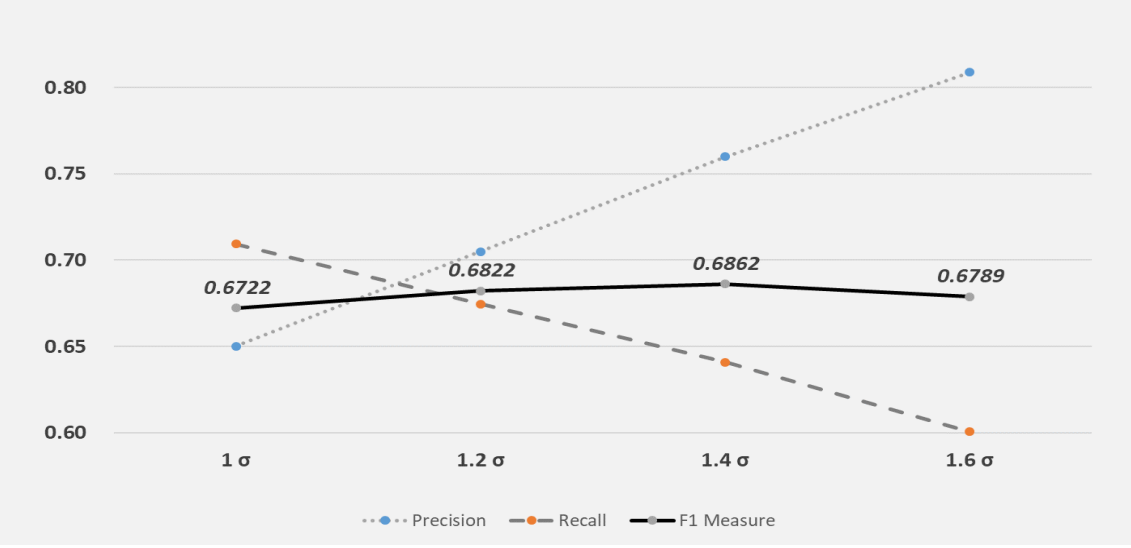
	Ideal_T_1 & RGS_T_7	*Ideal_T_2 & RGS_T_3*	*Ideal_T_3 & RGS_T_1*	*Ideal_T_4 & RGS_T_5*	*Ideal_T_5 & RGS_T_{10}*	*Ideal_T_6 & RGS_T_8*	*Ideal_T_7 & RGS_T_6*	*Ideal_T_8 & RGS_T_9*	*Ideal_T_9 & RGS_T_2*	*Ideal_T_{10} & RGS_T_4*	*Total*
Precision	0.748	0.619	0.592	0.683	0.672	0.746	0.748	0.463	0.634	0.595	0.650
Recall	0.916	0.815	0.708	0.647	0.533	0.671	0.661	0.560	0.939	0.644	0.709
F1 Measure	**0.824**	**0.704**	**0.645**	**0.664**	**0.594**	**0.707**	**0.702**	**0.507**	**0.757**	**0.619**	**0.672**

째 토픽에 속한 문서 수를, *Freq(Ideal_Tj)*는 실제 전역 토픽의 j번째 토픽에 속한 문서의 수를 나타낸다. 한편 *Co_Freq(RGS_Ti, Ideal_Tj)*는 *RGS_Ti* 와 *Ideal_Tj*가 서로 대응 관계에 있을 때, 이들 두 토픽에 동시에 포함된 문서의 수를 나타낸다.

<Table 7>의 정의에 따라 재현율(Recall), 정밀도(Precision), F1 Measure를 계산한 결과는 <Table 8>에 요약되어 있다. 또한 이 실험에서 10개 토픽 전체에 대한 F1 Measure의 평균은 "0.672"로 나타났다.

본 실험에서 문서 임계값 기준의 변경에 따라 각 토픽에 속하는 문서의 수가 변경되며, 재현율과 정밀도, 그리고 F1 Measure의 값 또한 영향을 받게 된다. 임계값 기준이 강화되면 RGS의 각 토픽에 속하는 문서의 수는 감소하므로, 정밀도는 증가하고 재현율은 감소할 것으로 예상된다. 반대로 임계값 기준이 완화되면 정밀도는 감소하고 재현율은 증가할 것으로 예상된다. 하지만 정밀도와 재현율의 조화 평균(Harmonic Mean)으로 산출되는 F1 Measure 값은 임계값 기준의 변화에 따라 어떠한 변화 양상을 보일지 예측하기 어려우므로, 이에 대한 실험을 수행하였다. 실험 결과 <Figure 13>과 같이 정밀도와 재현율은 예상했던 방향의 변화를 보였으며, F1 Measure는 이에 비해 큰 폭의 변화를 보이지 않음을 알 수 있었다. 또한 실험에 따르면 문서 임계값 기준으로 "평균 + 1.4σ"의 값을 적용한 경우 전체 토픽의 F1 Measure 평균이 "0.6862"의 값을 가지며 가장 높게 나타나는 것으로 확인되었다.

본 장에서는 제안 방법론의 실무 적용 가능성을 평가하기 위해 제안 방법론을 실제 뉴스 기사 24,000건의 분석에 적용한 실험 결과를 소개하였다. 이를 통해 24,000건의 문서에 대한 일괄 토픽 모델링을 수행하는 대신, 각 2,400건으로 구성된 소규모 문서 군집 10개에 대한 개별 토픽 모델링



〈Figure 13〉 Changes in F1 Measures with varied to Document Cutoff

을 수행하고 각 군집별로 240건씩 선정된 대표 문서 2,400건에 대한 전역 토픽 모델링을 수행하여, 대규모의 일괄 토픽 모델링 없이도 전체 문서에 대한 전역 토픽 배정이 가능함을 보였다. 또한 추가 실험을 통해 대표 문서 2,400건에서 도출한 전역 토픽이 전체 문서 24,000건에서 도출한 토픽과 매우 유사함을 확인하였으며, 제안 방법론을 통해 도출한 토픽 모델링의 결과를 일괄 방식의 토픽 모델링의 결과와 비교할 수 있는 방안 또한 제시하였다.

5. 결론

최근 대용량 문서로부터 토픽을 도출하는 토픽 모델링의 성능에 관한 이슈가 꾸준히 제기되고 있다. 특히 전체 문서에 대한 일괄 분석을 수행하는 것에 비해 전체를 소규모 군집으로 분할하여 군집별 분석을 수행하고, 이 결과를 취합하는 방안에 대한 모색이 이루어지고 있다. 이에 본 연구에서는 지역 군집의 일부 문서를 대표로 추출하여 축소된 전역 집합 군집을 생성하고, 대표 문서를 매개로 지역 토픽으로부터 전역 토픽의 성분을 도출하는 방법을 통해, 궁극적으로 군집별 토픽 모델링 결과를 전역으로 취합할 수 있는 방안을 제시하였다. 또한 실제 뉴스 24,000건을 활용한 실험을 통해 제안 방법론의 실무 적용 가능성을 파악하였으며, 추가 실험을 통해 제안 방법론에 따른 분할 분석 방식이 대량의 문서에 대한 일괄 분석과 유사한 결과를 효율적으로 도출할 수 있음을 확인하였다.

본 연구의 학술 및 실무적 기여는 다음과 같다. 우선 분할 수행된 토픽 모델링의 결과를 합리적으로 통합하기 위한 새로운 방안을 제안하였다는 점에서 학술적 기여가 인정될 수 있다. 또한 지역 토픽으로부터 전역 토픽의 변환 가중

치를 도출하는 부분은 본 연구에서 채택한 방법 이외에도 기계학습 기반의 모형 등 다양한 모형이 적용될 수 있으므로, 이 부분에 대한 소폭 변형을 통해 분석 정확도를 향상시키는 방식으로 후속 연구가 수행될 수 있을 것이다. 이와 함께 서로 다른 환경에서 수행된 토픽 모델링의 결과를 체계적으로 비교하는 방안을 제시하였다는 점 또한 본 연구의 기여로 인정받을 수 있다. 한편 제안 방법론은 기본적으로 성능의 이슈를 다루고 있다는 점에서 실무적 기여를 더욱 크게 인정받을 수 있다. 즉 시스템 자원의 한계 또는 시간 측면의 비용으로 인해 대용량 문서의 일괄 토픽 모델링 수행이 어려운 경우, 제안 방법론을 통해 추가 설비의 확충 없이 이를 분할하여 소량의 문서에 대한 토픽 모델링을 여러 시스템에서 병렬 수행함으로써 대량의 문서에 대한 일괄 토픽 모델링과 유사한 결과를 얻을 수 있다. 이러한 실무적 기여는 분석 대상 문서가 지역적으로 또는 시스템적으로 여러 곳에 분산되어 있는 경우 더욱 크게 나타날 것으로 예상된다. 이와 더불어 제안 방법론은 분할 분석된 결과를 하나로 취합하는 구조를 갖기 때문에 여러 개의 장비를 통한 분석의 동시 수행이 가능하며, 이에 따라 대용량 문서의 분석에서 기존 방법에 비해 더욱 단축된 분석 시간을 기대할 수 있다.

이러한 기여에도 불구하고 본 연구는 향후 다음의 측면에서 보완이 이루어져야 한다. 우선 분할 정복 접근 방식의 토픽 모델링에 대한 전체적 측면의 보강이 필요하다. 이 방법은 현실적 제한에 의해 수행이 어려운 대규모 문서의 토픽 분석을 효율적으로 가능하게 하지만, 전역 및 지역 토픽의 연결 관계 파악이 어려운 구조적 한계를 갖는다. 따라서 이 분야의 연구는 다른 분야의 토픽 모델링 연구에 비해 상대적으로 더딘 발전을 보이고 있으며, 이론적 배경 및 실증적 검증에 대한 연구가 충분히 수행되지 못했다. 따라서 추후 연구에서는 전통적 토픽 모델링 및 해당 방법의 분석 시간 비교를 포함하는 이론적 배경에 대한 연구와 해당 방법에서 분할 군집 개수에 따른 소요 시간 및 결과 정확도의 변화 양상 파악 등의 실증적 검증이 추가적으로 진행되어야 한다. 이와 함께 제안 방법론에 대한 보강 또한 추가적으로 이루어져야 한다. 가장 시급한 보완점으로 전역 문서 군집의 분할 기준이 체계적으로 마련될 필요가 있다. 본 연구의 실험에서는 연구자의 판단에 따라 전체 문서를 10개의 지역 군집으로 분할하였다. 하지만 제안 방법론의 실무 적용을 위해서는 분할 군집 수에 따른 분석 결과를 살펴보고, 이를 고려하여 최적 지역 군집의 개수를 설정하는 방안이 마련되어야 한다. 다음으로 본 연구에서 사용된 문서의 수보다 훨씬 방대한 양의 데이터에 대한 추가적 실험을 통해, 제안 방법론의 견고성을 다방면에서 확인할 필요가 있다. 이와 함께 각 지역 군집의 주제가 유사하거나 상이할 경우 제안 방법론에 따른 분석 결과가 어떻게 도출되는지, 대표 문서의 선정 과정에 토픽 정보가 활용된다면 제안 방법론의 성능이 개선될 수 있을지 등에 대한 깊은 고찰이 필요하다.

참고문헌(References)

AlSumait, L., D. Barbará and C. Domeniconi, "On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking," *2008 Eighth IEEE International Conference on Data*

Mining, (2008), 1~12.

Blei, D. M. and J. D. Lafferty, "Dynamic Topic Models," *Proceedings of the 23rd International Conference on Machine Learning*, (2006), 113~120.

Byun, S., D. Lee, and N. Kim, "Methodology for Identifying Issues of User Reviews from the Perspective of Evaluation Criteria – Focus on a Hotel Information Site," *Journal of Intelligence and Information Systems*, Vol.22, No.3(2016), 23~43.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science,* Vol.41, No.6(1990), 391~407.

Forman, G. and B. Zhang, "Distributed Data Clustering can be Efficient and Exact," *ACM SIGKDD Explorations Newsletter,* Vol.2, No.2(2000), 34~38.

Gartner, *Gartner's 2015 Hype Cycle for Emerging Technologies Identifies the Computing Innovations that Organizations Should Monitor,* Gartner, 2015. Available at http://www.gartner.com/newsroom/id/3114217 (Accessed 19 June, 2017).

Han, J., J. Pei and M. Kamber, *Data Mining: Concepts and Techniques,* Elsevier, Amsterdam, 2011.

Hotho, A., A. N¨urnberger and G. Paaß, "A Brief Survey of Text Mining," *Ldv Forum,* Vol. 20, No. 1(2005), 1~37.

IDC, *Big Data and Business Analytics Revenues forecast to reach $150.8 Billion this Year, Led by Banking and Manufacturing Investments*, IDC, 2017. Available at http://www.idc.com/getdoc.jsp?containerId=prUS42371417 (Accessed 19 June, 2017).

Kim, D. and N. Kim, "Mapping Categories of Heterogeneous Sources using Text Analytics," *Journal of Intelligence and Information Systems*, Vol.22, No.4(2016), 193~215.

Kim, N., D. Lee, H. Choi and W. X. S. Wong, "Investigations on Techniques and Applications of Text Analytics," *The Journal of The Korean Institute of Communication Sciences,* Vol.42, No.2(2017), 471~492.

Koll, M. B., "WEIRD: An Approach to Concept-Based Information Retrieval," *ACM SIGIR Forum,* Vol.13, No.4(1979), 32~50.

Lee, D., H. Choi and N. Kim, "A Method for Evaluating News Value based on Supply and Demand of Information using Text Analysis," *Journal of Intelligence and Information Systems*, Vol.22, No.4(2016), 45~67.

Liang, Z. and P. Chen, "Delta-Density based Clustering with a Divide-and-Conquer Strategy: 3DC Clustering," *Pattern Recognition Letters,* Vol.73, (2016), 52~59.

Livermore, M. A., A. Riddell and D. Rockmore, "Agenda Formation and the US Supreme Court: A Topic Model Approach," *Arizona Law Review,* (2016), Forthcoming.

McCallum, A., K. Nigam and L. H. Ungar, "Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching," *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2000), 169~178.

Mei, Q. and C. X. Zhai, "Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining,"

Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, (2005), 198~207.

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

Salton, G., *The SMART Retrieval System—Experiments in Automatic Document Processing*, Prentice-Hall, New Jersey, 1971.

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*, Vol.14, (2006), 457~462.

Song, Y., J. Du and L. Hou, "A Topic Detection Approach Based on Multi-Level Clustering," *Proceeding of the 31st Chines Control Conference*, (2012), 3834~3838.

Steyvers, M. and T. Griffiths, *Probabilistic Topic Models : Handbook of Latent Semantic Analysis,* Psychology Press, Oxfordshire, 2007.

Wang, J., H. Deng and J. Han, "Torpedo : Topic Periodicity Discovery from Text Data," *Next-Generation Analyst III,* (2015), 94990A~ 94990A-10.

Wang, L., P. Chen and L. Huang, "An Efficient Clustering Algorithm for Large-Scale Topical Web Pages," *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, (2009), 1851~1854.

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

Abstract

Efficient Topic Modeling by Mapping Global and Local Topics

Hochang Choi* · Namgyu Kim**

Recently, increase of demand for big data analysis has been driving the vigorous development of related technologies and tools. In addition, development of IT and increased penetration rate of smart devices are producing a large amount of data. According to this phenomenon, data analysis technology is rapidly becoming popular. Also, attempts to acquire insights through data analysis have been continuously increasing. It means that the big data analysis will be more important in various industries for the foreseeable future. Big data analysis is generally performed by a small number of experts and delivered to each demander of analysis. However, increase of interest about big data analysis arouses activation of computer programming education and development of many programs for data analysis. Accordingly, the entry barriers of big data analysis are gradually lowering and data analysis technology being spread out. As the result, big data analysis is expected to be performed by demanders of analysis themselves.

Along with this, interest about various unstructured data is continually increasing. Especially, a lot of attention is focused on using text data. Emergence of new platforms and techniques using the web bring about mass production of text data and active attempt to analyze text data. Furthermore, result of text analysis has been utilized in various fields. Text mining is a concept that embraces various theories and techniques for text analysis. Many text mining techniques are utilized in this field for various research purposes, topic modeling is one of the most widely used and studied. Topic modeling is a technique that extracts the major issues from a lot of documents, identifies the documents that correspond to each issue and provides identified documents as a cluster. It is evaluated as a very useful technique in that reflect the semantic elements of the document.

Traditional topic modeling is based on the distribution of key terms across the entire document. Thus, it is essential to analyze the entire document at once to identify topic of each document. This condition

* 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

causes a long time in analysis process when topic modeling is applied to a lot of documents. In addition, it has a scalability problem that is an exponential increase in the processing time with the increase of analysis objects. This problem is particularly noticeable when the documents are distributed across multiple systems or regions. To overcome these problems, divide and conquer approach can be applied to topic modeling. It means dividing a large number of documents into sub-units and deriving topics through repetition of topic modeling to each unit. This method can be used for topic modeling on a large number of documents with limited system resources, and can improve processing speed of topic modeling. It also can significantly reduce analysis time and cost through ability to analyze documents in each location or place without combining analysis object documents.

However, despite many advantages, this method has two major problems. First, the relationship between local topics derived from each unit and global topics derived from entire document is unclear. It means that in each document, local topics can be identified, but global topics cannot be identified. Second, a method for measuring the accuracy of the proposed methodology should be established. That is to say, assuming that global topic is ideal answer, the difference in a local topic on a global topic needs to be measured. By those difficulties, the study in this method is not performed sufficiently, compare with other studies dealing with topic modeling.

In this paper, we propose a topic modeling approach to solve the above two problems. First of all, we divide the entire document cluster(Global set) into sub-clusters(Local set), and generate the reduced entire document cluster(RGS, Reduced global set) that consist of delegated documents extracted from each local set. We try to solve the first problem by mapping RGS topics and local topics. Along with this, we verify the accuracy of the proposed methodology by detecting documents, whether to be discerned as the same topic at result of global and local set. Using 24,000 news articles, we conduct experiments to evaluate practical applicability of the proposed methodology. In addition, through additional experiment, we confirmed that the proposed methodology can provide similar results to the entire topic modeling. We also proposed a reasonable method for comparing the result of both methods.

Key Words : Divide and Conquer, Big Data, Text Mining, Topic Modeling

Received : July 21, 2017 Revised : September 14, 2017 Accepted : September 19, 2017
Publication Type : Regular Paper Corresponding Author : Namgyu Kim

저 자 소 개

최 호 창

현재 국민대학교 비즈니스IT전문대학원 석사과정에 재학 중이다. 국민대학교 경영학부에서 학사 학위를 취득하였으며, 주요 관심분야는 텍스트 마이닝, 데이터 마이닝 및 소셜네트워크분석 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 역임하였다. 주요 관심분야는 텍스트 마이닝, 데이터 마이닝 및 데이터베이스 설계 등이다.