Journal of the Korean Institute of Industrial Engineers
Vol. 41, No. 1, pp. 64-73, February 2015.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2015.41.1.064

<Application Research>

공간빅데이터 연구 동향 파악을 위한 토픽모형 분석

이원상 · 손소영†

연세대학교 공과대학 정보산업공학과

Topic Model Analysis of Research Trend on Spatial Big Data

Won Sang Lee · So Young Sohn

Department of Information and Industrial Engineering, Yonsei University

Recent emergence of spatial big data attracts the attention of various research groups. This paper analyzes the research trend on spatial big data by text mining the related Scopus DB. We apply topic model and network analysis to the extracted abstracts of articles related to spatial big data. It was observed that optics, astronomy, and computer science are the major areas of spatial big data analysis. The major topics discovered from the articles are related to mobile/cloud/smart service of spatial big data in urban setting. Trends of discovered topics are provided over periods along with the results of topic network. We expect that uncovered areas of spatial big data research can be further explored.

Keywords: Spatial Big Data, Literature Survey, Topic Model, Text Mining, Topic Network

1. 서 론

최근 이슈가 되는 빅데이터는 다양한 형태의 데이터가 빠른 속도로 실시간 생성되는 것을 의미하며(Yang, 2012), 빅데이터의 관리 및 분석을 통해 가치를 창출하려는 시도가 광범위하게 일어나고 있다(Yoon, 2013; Kim *et al*., 2014). 그 중에서도, 현존하는 모든 정보의 80% 정도는 공간과 관련된 정보라는 점에서, 공간빅데이터(Spatial Big data)에 대한 실질적인 활용이 주목을 끌고 있다. 공간빅데이터는 6V(Volume, Variety, Velocity, Value, Veracity, Visualization)에 의해 정의될 수 있는데, 근접성 및 규모, 위도, 경도, 고도 등의 공간적 속성과 일반적인 빅데이터의 특성이 결합되어 있다(Ahn *et al*., 2013). 공간빅데이터의 처리와 활용은 국가나 기업의 경쟁력 제고를 위해 중요할 것으로 보이며. 최근 사물인터넷 및 센서 데이터 등의 급증 등으로 인하여 중요성이 배가되고 있다. 이에 공간빅데이터에 대한 연구 동향을 파악하여 그간의 연구 흐름에 대한 이해와 함께 향후 방향을 전망하는 것이 필요한 상황이다.

본 연구에서는 SCOPUS DB에서 “Spatial Big Data”라는 키워드로 검색된 전체문헌의 초록을 분석하여 공간빅데이터의 글로벌 연구 동향을 파악하고자 한다. 수집된 문헌의 초록으로부터 유망 연구 동향을 추출하기 위해서 텍스트마이닝과 토픽모델링을 적용하였다. 특히, 토픽모델링에는 LDA(Latent Dirichlet Allocation) 기법을 사용하였으며, 문헌 별로 발생확률이 높은 유망 주제의 관계를 바탕으로, 유망 주제 간의 네트워크도 제시하였다. 본 연구를 통하여 발견된 글로벌 연구 동향을 바탕으로, 공간빅데이터 분야 연구자들의 향후 연구 방향 수립에 기여할 것이 기대된다.

본 논문의 구성은 다음과 같다. 제 2장에서는 공간정보분야에서의 최신 동향과 관련된 기존 연구를 고찰하며 이 논문의 연구 이슈를 살펴보았다. 제 3장에서는 논문에 사용된 데이터의 수집과 처리, 그리고 사용한 방법론을 소개하였으며, 제 4장에서는 분석 결과를 제시하였다. 마지막으로, 제 5장에서는

본 연구는 ‘국토교통부 국토공간정보연구사업 국토공간정보의 빅데이터 관리, 분석 및 서비스 플랫폼 기술개발(14NSIP-B081011-01)과제’의 연구비 지원에 의해 연구되었음.
† 연락저자 : 손소영 교수, 120-749 서울특별시 서대문구 신촌동 연세대학교 공과대학 정보산업공학과, Tel : 02-2123-4014, Fax : 02-364-7807, E-mail : sohns@yonsei.ac.kr
2014년 9월 25일 접수; 2014년 11월 17일 수정본 접수; 2014년 12월 8일 게재 확정.

결과에 대한 해석과 시사점, 그리고 결론을 제시하였다.

2. 문헌고찰 및 연구이슈

2.1 공간빅데이터 최근 동향 관련 연구

그 동안 공간데이터는 재해, 교통, 날씨, 자원, 범죄 등의 다양한 분야에서 활용이 되어왔다(Messner *et al*., 1999; Schmidt *et al*., 2002; Westen *et al*., 2008; Kearney and Porter, 2009; Wang and Wang, 2011; Yavuz and Erdoğan, 2012). 이러한 공간데이터의 규모가 급증하면서 나타난 공간빅데이터는 향후 국가와 기업의 경쟁력 강화를 위한 중요한 요인으로 고려되면서, 다양한 분야에서 공간빅데이터 관련 정책들이 수립되고 실행되고 있다. 이 분야에서 앞서 나가고 있는 미국의 경우, 당면한 여러 과제 해결과 경쟁력의 강화를 위해 정부주도로 공간빅데이터 활성화를 위한 여러 전략이 수립되었다. 특히 공간정보를 유통하던 GOS(Geospatial One-Stop)와 빅데이터 공공정보를 제공하는 플랫폼인 Data.gov를 통합한 Geo-data.gov 플랫폼을 구축하여, 공간빅데이터 개방과 이용의 활성화를 위한 인프라를 제공하고 있다(Kim *et al*., 2013; Lee, 2013). 유럽에서는 국가 간 도로, 교통, 항만, 범죄, 재난재해, 주택, 환경, 의료 등에 공간 관련 정보의 효과적인 활용을 목적으로 개방적이고 협력적인 인프라 구축을 하여 공공기관, 기업체 및 일반인들에게 제공하고 있다(Kim, 2014). 한국도 2011년 수립된 '빅데이터를 활용한 스마트 정부 구현' 계획을 바탕으로, 국토교통부에서 공간빅데이터 기반 스마트 행정과 맞춤형 국민 서비스를 제공하기 위한 사업을 추진 중에 있으며, 공간빅데이터 관련 기술 경쟁력 확보에도 노력을 기울이고 있다(Kim, 2014). 하지만, 국내의 공간빅데이터 관련 동향은, 국외 사례와 비교하면, 아직은 기획과 추진 단계에 있는 점에서 차이가 있다.

정부뿐 만 아니라, 민간 및 연구자들에 의해서도 공간빅데이터 관련 연구가 다양하게 이뤄지고 있다(Chang *et al*., 2009; Kim *et al*., 2009). Choi *et al*.(2012)은 공간정보 플랫폼 구축과 관련하여 공간정보 생태계 구축을 위한 플랫폼의 필요성을 강조하고, 그에 관한 전략을 제시하였다. 또한, Amirian *et al*. (2014)은 공간빅데이터의 관리, 분석 및 공유를 위한 특화된 기술과 알고리즘의 필요성을 언급하며 공간빅데이터 관리 시스템을 제시하였다. 공간빅데이터 분야 연구 동향과 관련하여, Jian-ya (2002)는 GIS 기술의 최근까지 동향 및 향후 전망을 소개하며 공간 데이터 처리 기술의 중요성을 강조하였다. Zhang(2014)은 Geo-information 분야에서, 공간빅데이터의 분석과 활용을 향후 5년간 이 분야의 주요 이슈로 언급하였고, Goldberg *et al*. (2014)는 공간빅데이터를 효과적으로 처리하기 위한 지도 및 GIS 기술 현황을 제시하였다.

2.2 토픽 모델링

이와 같이 공간빅데이터와 관련하여 다양한 분야의 많은 연구들이 진행되는 가운데, 연구 동향의 파악은 현재까지의 연구에 대한 이해와 함께 향후 연구 방향 수립 등에 기여할 수 있다. 무엇보다도, 연구동향 및 유망 연구주제 파악은 전문가에 의해 이뤄지는 것이 가장 이상적일 것이다. 그러나 이는 한정된 시간과 비용 등에 영향을 받게 되며, 이종 분야에 대한 주제 파악에도 한계가 있을 수 있다. 특히, 공간빅데이터와 같이 다양한 분야가 혼재하는 경우에는 그런 한계는 더 부각될 수 있다. 그렇기 때문에, 연구의 결과물인 다량의 관련 문헌들을 대상으로 한 연구동향 파악이 더 효과적일 수 있다. 이때 많이 적용되는 토픽모델링에 대해 살펴보았다.

토픽모델링은 문헌에서 주제를 찾아내기 위해 사용되며, LSI (Latent Semantic Indexing), pLSI(probabilistic Latent Sematic Indexing), LDA(Latent Dirichlet Allocation) 등이 있다. 그 중에서도 최근에는 확률모형을 바탕으로 문헌에서 주제를 찾아내는 LDA가 많이 사용되고 있다(Blei, 2003). 특히, LDA는 LSI 대비 주제에 대한 해석이 용이하고, pLSI 대비 베이지안 통계 기반의 유연한 모델링이 가능하다. 텍스트마이닝을 통한 연구동향 파악은 많이 시도되어져 왔으며(Cho and Kim, 2012; Cho *et al*., 2014), 더 나아가 토픽모델링에서는 문헌은 주제들의 집합이고, 문헌은 단어들로 구성된다고 가정하여, 문헌 별 주제나 단어의 출현확률을 바탕으로 각 주제들을 파악할 수 있게 된다(Blei, 2003). 이러한 토픽모델링은 다양한 분야에 활용되고 있는데, 트위터에 적용되어 사람들이 관심을 갖는 주제 파악에 활용되거나(Bae *et al*., 2014), 바이오 데이터 분석에 사용되거나(Wu *et al*., 2014), 정치 텍스트 분석에도 적용되고 있다(Masumura *et al*., 2014).

이러한 토픽모델링은 연구 동향 파악에도 유용하게 적용할 수 있다. Jeong and Song(2012)은 Computer Science와 Medical Science 분야에서 논문, 특허, 웹 뉴스 등 다양한 소스로부터의 동향을 토픽모델링을 이용하여 분석하였다. Guo *et al*.(2014)는 과학 문헌에 대한 인용네트워크로부터 지식을 추출하기 위한 기존 토픽모델링을 개선한 Bernoulli Process 토픽모델링을 제안하였다. 이외에도 많은 연구에서는 토픽모델링을 활용하여 다양한 출처의 텍스트 자원을 분석하고 동향을 파악하였다. 위에서 살펴본 토픽모델링 관련 선행연구처럼, 본 연구에서는 토픽모델링의 여러 방법 중에서 주제 해석이 용이하고 모델링이 유연한 LDA를 문헌들의 초록에 적용하여 공간빅데이터에 대한 연구 동향을 파악하였다.

2.3 연구이슈

고찰 된 바와 같이, 공간빅데이터와 관련하여, 거시적인 동향을 파악하거나 어떤 기술들이 있는지에 대한 분석은 이뤄져 왔지만, 구체적인 연구 주제에 대한 파악은 충분하지 않은 상황이다. 특히, 기존 연구에서 효과적으로 사용되어오는 토픽모델링 기법을 공간빅데이터와 관련된 문헌들을 대상으로 적용하여 연구 주제를 파악하려는 시도는 거의 없는 상황이다. 그러므로 본 연구에서는 공간빅데이터 관련 문헌들에 토픽모

델링을 적용하여, 지금까지의 글로벌 연구동향을 파악하고자 한다. 발견된 연구주제들간의 관계를 파악하여, 핵심적인 주제를 발견하며, 각 연구주제들이 시간에 따라 어떻게 증감하는지를 바탕으로 향후 많이 연구될 주제도 살펴보고자 한다. 그 결과를 바탕으로, 공간빅데이터 관련해서 향후 어떤 방향의 연구가 필요할 것인가에 대해 전망하고자 한다.

3. 데이터 및 방법론

이 논문에서는 Elsevier의 Scopus DB에서 “Spatial Big Data”라는 키워드의 조합으로 문헌명/초록/주제어 필드에 검색을 한 후 얻어진 문헌의 초록을 대상으로 분석을 하였다. 2014년 8월까지 총 1,621개의 문헌들이 검색되었으며, DB에서 제공하는 반출 기능을 사용하여 해당 문헌들의 초록을 수집하였다. 초록들을 처리하기 위해 R프로그램 기반의 텍스트마이닝 기법을 적용하여 LDA 적용을 위한 문서-단어 행렬(Document-Term Matrix)를 생성하였다. 이 행렬의 값은 단어의 출현빈도이고 중요한 단어를 발견하기 위하여 TF-IDF를 가중치로 사용하였다. TF-IDF값을 통해서는 특정 문서에 출현하는 단어에 더 높은 가중치를 주고, 모든 문헌에 나오는 단어에는 낮은 가중치를 주어 각 문헌을 잘 나타내는 단어를 찾을 수 있다. 구체적으로는 특정 문헌에서 특정 단어의 출현빈도가 높을수록 중요한 단어로 볼 수 있다. 하지만, 전체 문헌들에서 빈번하게 나타나는 단어라면, 특정 문헌을 구별하는데 잘 활용되지 않을 수 있다. 그래서 문헌빈도의 역수인 역문헌 빈도와 단어의 출현빈도를 반영한 TF-IDF 가중치를 사용하게 된다.

텍스트마이닝으로 처리된 데이터에, Latent Dirichlet Allocation(LDA) 기법으로 토픽모델링을 적용하였다(Blei, 2003). 문헌에서의 주제 발견과 관련하여, 어떤 두 문서가 주제는 유사해도 각 문서에 등장하는 단어의 종류나 빈도는 다를 수 있는데, 기존의 텍스트마이닝 기법인 키워드 기반의 모형으로는 유사도를 계산하거나 주제 분류를 하는 데에는 한계가 발생할 수 있다. 반면 문서를 대상으로 주제를 발견하는 확률 모형인 LDA를 사용하면, 문서의 주제 분포와 주제별 단어의 통계적 분포를 바탕으로 특정 문서가 만들어질 확률을 파악할 수 있다. 다만, LDA로 발견할 주제의 수 K는 사전에 결정되어야 하는 전제조건이 있다. K의 값은 LDA 모형의 Perplexity(또는 Log- likelihood) 등을 고려하여 결정되어야 한다. 다양한 값을 대입하여 LDA를 수행한 후 Perplexity를 관찰하여, 최적의 K를 찾도록 하였다. 데이터의 Perplexity를 최소화하는 Parameter를 추정해야 하는데, 이 논문에서는 Variational Expectation Maximization Algorithm을 구현한 R의 topicmodels 패키지를 사용하여 추정하였다(Hornik and Grün, 2012).

각 주제의 의미는, 주제별로 가장 출현확률이 높은 단어들을 통해 해석할 수 있기 때문에, 이 논문에서는 사후발생확률이 높은 상위 용어들을 바탕으로 해석되었다. 또한, 토픽모델링의 결과 중 문헌-주제 관계가 주어지는데, 이 관계를 사용하여 주제 네트워크를 생성하도록 하였으며, 각 문헌에서 주제가 발생할 확률을 구하였다. 구체적으로는 각 문헌 별 사후출현확률이 높은 주제의 관계는 문헌과 주제의 2-모드 네트워크(또는 Affiliation network)로 볼 수 있다. 2-모드 네트워크를 행렬로 표현하여 주제-문헌 행렬을 구한 후, 이 행렬의 전치행렬인 문헌-주제 행렬을 곱하게 되면 주제-주제 행렬을 구할 수 있다. 이 과정에서 Wesserman(1994)에서 제시된 방법을 참고하였다. 이 행렬은, 방향성이 없는 그래프로 표현되고, 각 주제가 다른 주제에 얼마나 연관있는지(Rate of Participation)를 구하게 된다. 특정 주제가 갖는 문헌의 수는 주제-문헌 행렬에서는 행의 합이고, 주제-주제 행렬에서는 그 대각의 값이 되고, 이 값으로 대각이 아닌 다른 행의 값들을 나눠주었다. 행의 각 값을 0부터 1 사이의 값으로 표준화시킬 수 있었고, 그 값은 어떤 주제가 다른 주제와 연관이 있는지를 알려주는 가중치로 사용할 수 있다(Wasserman, 1994). 결과적으로는 주제 간의 표준화되고 방향성 있는 그래프를 만들 수 있어 연구 동향을 파악할 수 있게 하였다.

4. 분석

본 논문에서는 공간빅데이터 연구 동향 분석을 위하여 수집된 1,621개의 문헌에 대한 토픽모델링 적용에 앞서 해당 문헌들의 출현빈도 추세, 게재된 저널, 연구 분야, 연구기관, 국가, 그리고 피인용 상위 논문 등을 다음과 같이 살펴보았다. 우선, <Figure 1>에서와 같이 최근 공간빅데이터에 대한 연구가 급증하는 것을 볼 수 있었다.

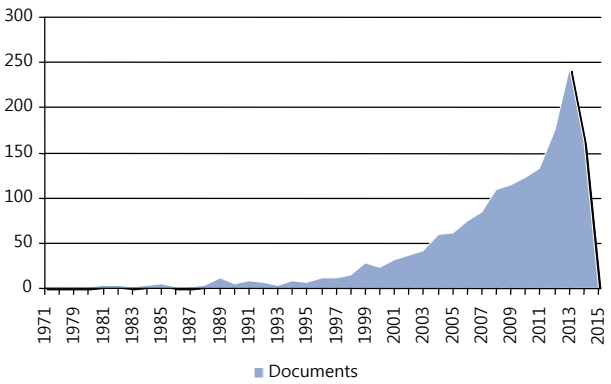

Figure 1. Annual frequency of related article

또한, 1,621개의 수집된 문헌은 총 132개의 다양한 저널에 게재되었으며, 해당 문헌 빈도0..를 기준으로 상위 10개 저널의 연도별 문헌 빈도는 <Figure 2>와 같았다.

<Figure 2>에서 볼 수 있듯이, Remote Sensing, Astronomical Telescopes, Instruments, and Systems, Biomedical Optics를 주로 다루는 SPIE: the International Society for Optical Engineering 저널에 가장 많은 83개의 문헌이 게재되었다. 특히 이 저널에서

는 Astronomy 분야에서의 공간빅데이터의 활용 연구가 많이 나타나고 있었다. 최근에는 Lecture Notes in Computer Science 에 관련문헌이 2010년을 지나면서 급증한 것을 볼 수 있었으며, Computer Science 분야에서 공간빅데이터를 처리하는 많은 시도가 있음을 볼 수 있었다.

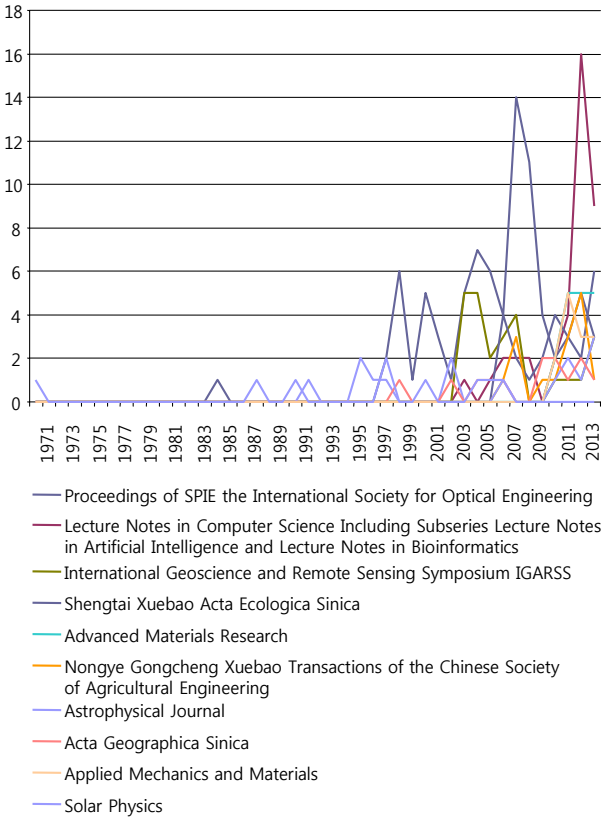

Figure 2. Annual frequency of article from top 10 related journals

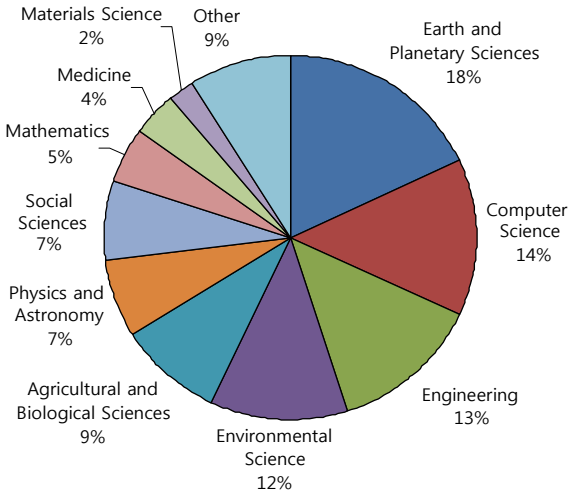

Figure 3. Domain of spatial big data

이러한 공간빅데이터 연구들이 어느 분야에서 주로 발생하는지에 대한 추세는 <Figure 3>과 같이 나타났다. 주로 Earth and Planetary Sciences, Computer Sciences, Engineering, Environmental 등의 분야에서 공간빅데이터 연구가 나타나고 있었으며, 그 외에도 Materials Science, Energy, Neuroscience, Economics, Medicine 등의 다양한 분야도 나타나고 있었다.

공간빅데이터 연구가 어느 기관에서 주로 수행되는지를 파악하기 위해 공간빅데이터 관련 문헌의 저자가 소속된 기관을 살펴보았으며, 그 중 상위 10개 기관은 아래 <Figure 4>와 같이 나타났다.

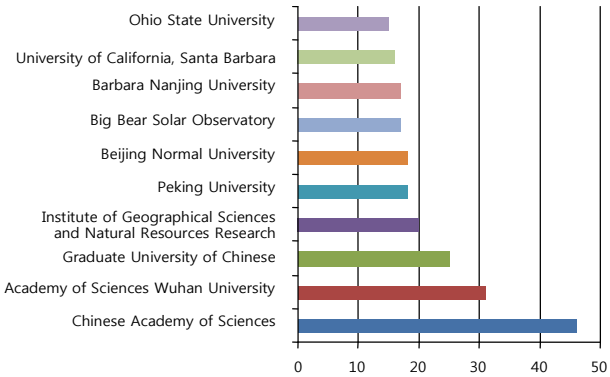

Figure 4. Top 10 Organizations that Publish Spatial Big Data Literature

전체 160여 개의 기관 중 97개가 대학에 속해있는 연구기관인 점을 볼 때, 현재까지 공간빅데이터 관련 연구는 기업보다는 대학이 주도하는 것으로 나타났으며, 흥미롭게도, 현재까지 중국의 대학 및 연구기관에서 공간빅데이터 연구가 활발히 진행되고 있었다.

또한, 국가별 공간빅데이터 연구 동향을 파악하기 위해 국가별 문헌빈도를 살펴보았다. <Figure 5>에서와 같이 중국, 미국, 독일, 영국 순으로 공간빅데이터에 대한 연구가 활발한 것으로 나타났다. 한국의 경우 27개의 문헌이 확인되어 총 93개국가 중 13위 수준으로 나타났다.

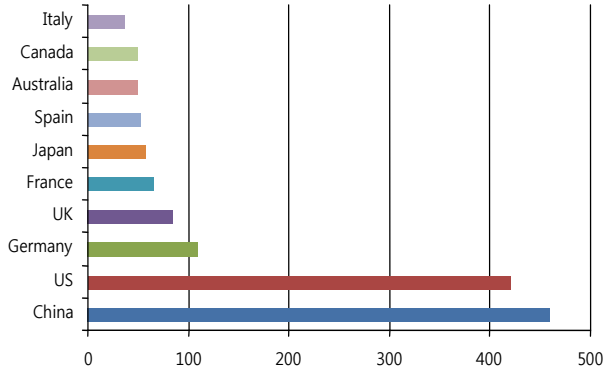

Figure 5. Top 10 Countries that that Publish Spatial Big Data Literature

그리고 공간빅데이터 문헌이 갖는 피인용에 대해서 살펴보

았다. <Figure 6>에서 볼 수 있듯이, 대부분의 문헌은 피인용 횟수가 50 미만인 것을 알 수 있었다. 그러나 몇몇 문헌은 300 이상의 피인용 횟수를 갖는 것으로 나타났다.

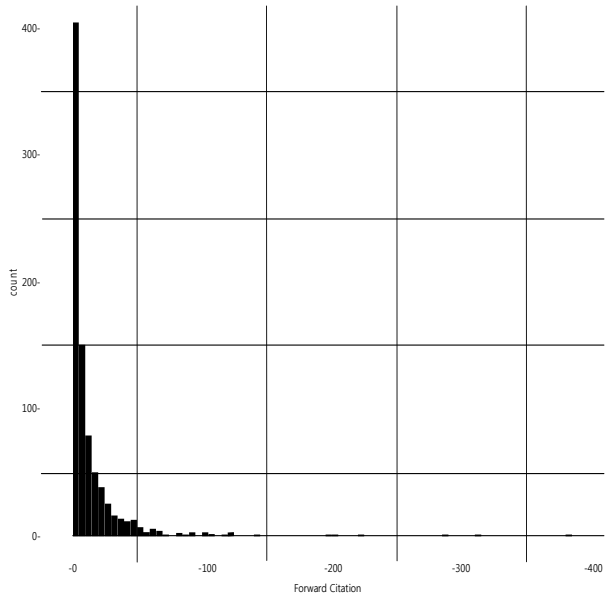

Figure 6. Forward citation distribution of spatial big data literature

보다 구체적인 연구 동향 파악을 위해 본 연구에서 앞서 소개한 토픽모델링을 적용하고자 대상 데이터에 대해 텍스트마이닝을 적용하였다. 초록에 대한 형태소 분석, 3글자 미만 단어 제외, 불용어 처리 등을 통해, 1,621개의 문서와, 19,814개의 용어로 이뤄진 문서-용어 행렬을 만들었으며, 이때 출현 빈도 기준 상위 용어들은 <Figure 7>과 같이 나타났다.

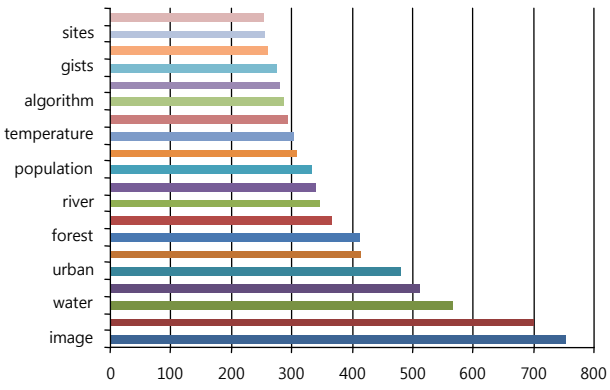

Figure 7. Most Frequent Terms

즉, 출현 빈도 기준으로, 공간빅데이터 연구에서 Land, Image 용어가 가장 많이 나타나며, 이어서 urban, soil, network, forest 등의 용어가 나타나고, 계속해서 algorithm, temperature, species, vegetation, river, population 등의 용어들이 많이 나타난 것을 볼 수 있었다.

하지만, 위와 같은 용어 중에는 공간빅데이터 관련 문헌에 공통적으로 나타나는 용어와 개별 문헌의 특징을 잘 나타내는 용어가 혼재되어 있다. 그렇기 때문에 보다 중요한 용어의 선택이 필요하고, 이 연구에서는 용어의 가중치로 TF-IDF를 계산하여 적용하였다. 문서-용어 행렬의 모든 값에 대해 TF-IDF를 구한 후, 전체 TF-IDF값 중 하위 20% 미만의 값을 갖는 단어는 중요도가 낮다고 고려하여 문서-용어 행렬에서 제거하였고, 문서-용어 행렬의 차원을 1,621개의 문서와 15,861개의 용어로 축소하였다.

이렇게 처리된 문서-용어 행렬에 대해 토픽모델링인 LDA를 적용하였다. LDA는 최적의 주제 수에서 낮은 Perplexity가 낮은 값을 갖기 때문에, 다양한 K값에 대한 LDA 결과 중 Perplexity가 최소가 되는 K를 발견하여 사용하고자 하였다. K의 값은 10부터 순차적으로 하나씩 늘려갔으며, 이때 K와 Perplexity의 변화 추이는 다음과 같았다.

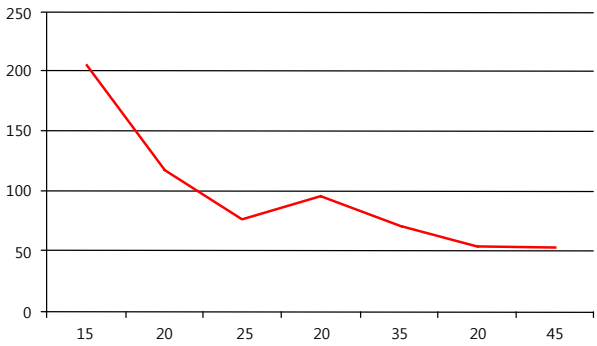

Figure 8. Changes in perplexity by number of topics

<Figure 8>과 같이 Perplexity는 K가 크면 클수록 최소값을 갖지만, 주제의 수가 너무 많아지면 분석 시 어려움이 발생할 수 있다. 그래서 분석의 간결함과 직관성을 위해, 본 논문에서는 Perplexity 감소폭의 변화가 일정 수준까지 내려간 후 계속 유지되기 시작하는 40을 주제의 수로 고려한 후, LDA를 적용하였다. 그리고 결과를 바탕으로 주제별 사후 출현확률이 높은 단어로 주제를 설명하고자 하였으며, 발견된 단어를 다시 Topic을 대표하는 3~4개의 단어, 분석방법이나 기술을 나타내는 단어, 그리고 분석이 적용된 대상 분야에 대한 단어로 구분하여 제시하였다.

다음의 <Table 1>에서는 주제별 출현확률이 높은 단어를 주제 대표 단어, 주제에서 사용된 방법이나 기술, 주제와 관련된 대상 분야 등을 제시하였다. 또한, 이러한 주제별 대표 단어를 참고하여 각 주제를 분류해보았고, 총 19개로 분류가 되었다. 각 주제의 출현빈도를 바탕으로 전체 문헌집합에서 차지하는 비율을 제시하였으며, 대부분의 주제는 비교적 균등하게 분포하는 것을 볼 수 있었다. 주제별 중요도를 좀 더 구체화하기 위해 이러한 주제별 출현빈도에 주제에 속한 문헌들이 갖는 피인용 빈도를 가중치로 반영해보았다. 각 문헌에서 해당 주제가 발생할 사후확률에 그 문헌이 갖는 피인용빈도를 곱한 값을 구하여 가중치로 사용하였으며, 그 값의 전체 대비 비율을

Table 1. Most probable terms by topics

#	Most Probable Terms			Topic Category and Occurrence Weight		
	Representative Terms	Analysis/Technology related Terms	Domain	Category	Topic Weight based on Frequency(Freq.)	Topic Weight based on Forward citation(Forward Citation Weight)
1	Image, coastal/river, vegetation	Image/Spectral	Coastal	Ecology	0.0136(160)	0.9654(138.3234)
2	Disaster, urban, trauma	Sonar/Echolocation	City	Disaster in urban	0.0250(295)	0.0003(0.0444)
3	Software, database, network	DB/GPS/MWL/Mining	Rhine	IT infrastructure	0.0278(328)	0.0004(0.0613)
4	Metereology, precipitation, athmosphere	Interpolation	island, river	Metereology	0.0171(202)	0.0002(0.0319)
5	Endemic goiter, nuclear, geohazard,	Spacetime/Spectral/xray	Hubei/mountain	Astronomy	0.0175(207)	0.0003(0.0428)
6	Earthquake, urban netwrok, algorithm	Smart WSNs/Probabilistic model	Bengal	Disaster in urban	0.0183(216)	0.0002(0.0301)
7	Agriculture, gis, network	GIS	Farm	Agriculture and GIS	0.0064(76)	0.0001(0.0119)
8	Deforest, fire, dieback	Landsat/Image	Forest	Forestry	0.0187(221)	0.0002(0.0306)
9	Urban, socioeconomics, agglomeration	Census/Economic model	Panjin/India	Urban economy	0.0300(354)	0.0005(0.0658)
10	Seismic, ablation, image	Grid/Interpolation/Non-linear model/Realtime/Classification	Sea	Geology	0.0362(427)	0.0005(0.0676)
11	Wetland, methane emission, biomass	Spectral/Image	Yancheng	Ecology	0.0223(263)	0.0003(0.0439)
12	Mobile, cloud, geospatial	Cloud/Virtual/Smart computing, Spatiotemporal mining	-	IT infrastructure	0.0097(115)	0.0130(1.8688)
13	Soil, erosion, runoff	GIS, kriging	Poyang	Geology	0.0412(486)	0.0006(0.0901)
14	Tornado, phyplankton, Image	Image/Flux/Sensor/robot	Sea	Oceanography	0.0511(603)	0.0008(0.1157)
15	Biomass, forecast, geostatistics	Semivariogram/Kriging/entropy	Poland	Geo-statistics	0.0314(371)	0.0005(0.0769)
16	Species, mammal, game	Game theory	-	Ecology	0.04300(507)	0.0007(0.0996)
17	Landscape, classification, image, galaxy	Classification/Hyperspectral/wavelet	-	Geology	0.0422(498)	0.0005(0.0787)
18	River, groundwater, aerosol	Galaxy satellite	Texas, Mexico	Ecology	0.0181(214)	0.0003(0.0366)
19	Geospatial, gis, software	GIS, algorithm, spatiotemporal, image	-	IT infrastructure	0.0219(259)	0.0003(0.0484)
20	Vegetation, image, parallel	Image, multicore, parallel, classification	Qinghai, Wujiang	Agriculture and Parallel processing	0.0252(297)	0.0004(0.0525)
21	Radar, afforest, cypress	Bernoulli distribution/Logistic model	-	Forestry	0.0097(115)	0.0002(0.0278)
22	Image, Scan, Disease	Albedo/Tomography/Fourier	-	Image Analysis(MRI)	0.0105(124)	0.0001(0.0214)
23	Lidar, flood, typhoon	Radar/Forecast	-	Climate	0.0337(398)	0.0003(0.0486)
24	Habitat, species, image	Sagebrush/pygmy, rabbit, hippocamp, panda	Mountain and coast	Ecology	0.0160(189)	0.0002(0.0316)
25	Urban, aerosol, cloud	image/spectral/classification/algorithm	City	Climate	0.0452(534)	0.0006(0.0837)
26	Spatiotemporal, grid, diffusion	Spatiotemporal/sensor/raster/algorithm/tree/prune	-	Geo-statistics	0.0279(329)	0.0005(0.0660)
27	Solar, earthquake, telescope	Image/grid/wavelength	Wenchuan	Astronomy	0.0462(545)	0.0070(1.0072)
28	Ozone, orbit, Image	Image/wave/Bayesian	-	Climate	0.0312(368)	0.0004(0.0626)
29	Species, forest, population	Plot/anova	-	Forestry	0.0073(86)	0.0001(0.0156)
30	Fish, City, transport	Probabilistic model/forecast/biodiversity	City	Ecology	0.0197(232)	0.0003(0.0460)
31	Neuron, cell, auditory	Spectral/Image	-	Image Analysis(MRI)	0.0481(568)	0.0007(0.1061)
32	Urban, lake, landsat	Landsat/classification/algorithm	Shanghai, Poyang, Beijing	Urban and Ecological	0.0169(200)	0.0003(0.0428)
33	Lation(racial), Urban, volcano	Variogram/extrema	Allentown, Finland	Disaster in urban	0.0238(281)	0.0004(0.0613)
34	GIS, agriculture, IOT	Socio-economics	-	Agriculture and GIS	0.0270(319)	0.0006(0.0821)
35	Habitat, marine, biodiversity	Biodiversity	Mediterranean, Istanbul	Ecology	0.0138(163)	0.0008(0.1156)
36	Pollution, risk, contamination	Concentration/exposure/camera	-	Pollution	0.0220(260)	0.0003(0.0459)
37	Video, tumor, surveillance, network	Tweet/surveillance/graph	Tokyo, Canada	Etc	0.0229(270)	0.0003(0.0418)
38	Urban, GIS, economy	Autocorrelation(Moran's I)/image/cluster/socio-economics	Shanghai, Beijing, Poznan	Urban and GIS	0.0162(191)	0.0002(0.0346)
39	Cluster, algorithm, image	Cluster/image/r-tree/trajectory	-	Spatial Analysis	0.0167(197)	0.0003(0.0367)
40	Markov, bayesian, gaussian	MCMC/Bayesian/gaussian/image/interpolation/nonlinear	-	Spatial Analysis	0.0285(336)	0.0006(0.0805)

통하여, 각 주제가 갖는 가중치의 비중도 구해보았다. 흥미롭게도 단순 출현빈도를 기반으로 주제들의 중요도를 살펴보았을 때에는 균등한 양상을 보이는 반면, 피인용빈도를 반영한 가중치 기반 중요도를 볼 경우, 1번 주제의 중요도가 절대적인 것을 볼 수 있었다. 이러한 결과는, 공간분석 관련 생태학 분야의 연구에서 많은 피인용이 발생했기 때문으로 보인다.

공간빅데이터와 관련된 각 주제를 19개의 범주로 분류하였는데, 생태학 관련 주제가 17.5%로 가장 많이 발생한 것으로 나타났다. 그 다음으로는 기상, 도시 환경에서의 재해, 삼림관리, 지리학, 공간데이터 처리를 위한 IT에 대한 내용들이 7.5%로 출현하는 것으로 나타났다. 생태학과 관련된 주제들 중에서 특히 1번 주제는 해안이나 강가의 경작에 대한 주제로 보이며, 주로 이미지 분석 기법이 사용된 것으로 나타났다. 12번 주제는 중국의 Yancheng 지역에 이미지 분석 기법을 사용해 습지를 분석한 것으로 나타났다. 생태학에 속하는 17번 주제는, 동물의 종에 게임이론을 적용하였으며, 19번 주제는 지상관측 위성을 통해 텍사스와 멕시코 일대 지표면의 수분 및 에어로졸을 분석하였다. 그 외에도 생태학에 속하는 주제에서는 종의 다양성 등에 대한 이미지 분석이 이뤄졌다. 기상, 삼림관리, 지리학에 속하는 주제들 역시 이미지 분석 및 지리통계학의 주요 방법들인 Kriging, Variogram 등 Spatial autocorrelation을 사용하여 분석이 이뤄졌다. 이러한 주제들은 공간데이터를 다루는 전통적인 분야들로 고려할 수 있다.

또한, 도시 환경에서의 공간 데이터에 대한 분석과 활용의 주제들도 제시되고 있다. 우선, 도시에서의 재난 상황과 관련한 주제가 있는데, 지진이나 화산 발생 시 공간데이터를 활용하여 어떻게 대처할 수 있는지에 대한 내용이 다뤄지고 있다. 특히, 재난 발생 시 대응할 수 있는 스마트 WSNs에 대한 구축 등의 IT와의 융합에 대한 주제도 나타났다. 또한, 도시에서의 문제를 사회경제학적인 방법으로 해결하는 주제도 나타났는데, 이때에도 GIS를 사용하는 등 공간데이터 관련 IT를 융합해 나가는 주제들이 발견되었다.

그 외에도, 여러 오염물질에 대한 노출로 인한 위험도를 지역별로 측정을 하고 이미지 분석을 통해 처리하는 환경오염 관련 주제, 해양에서의 공간 분석에 로봇을 사용하는 것에 대한 주제, 의학분야에서의 MRI 이미지도 공간으로 고려하여 이에 대한 분석의 주제 등도 발견되었다. 처리 및 분석 기법과 관련된 구체적인 주제도 발견되었다. 우선 공간빅데이터를 처리하기 위한 병렬처리기법이 나타났는데, 이는 경작 관련 공간분석에 사용되는 대용량의 이미지를 분석해내기 위해 나타난 것으로 보인다. 또한, 경작에서의 GIS 및 IoT의 활용에 대한 주제도 나타나서, 경작 분야에서의 빅데이터 관련 기술이나 새로운 기술이 접목되는 것을 볼 수 있었다. 분석 기법과 관련해서는, 분석 및 기술 관련 단어 중에 Spatial Autocorrelation에 대한 autocorrelation, kriging, variogram 등의 방법론이 GIS, entropy, extrema, socio-economics 등과 같이 나오는 양상도 발견이 되었다. 공간 분석의 기존 공간통계의 기법들이 계속 사용되는 것을 볼 수 있으며, 단순히 공간에 대한 분석에서 공간과 시간을 고려한 분석으로 확장해가는 것을 볼 수 있었다. 그 외에도 트위터와 비디오를 같이 사용한 감시체제 등에 대한 주제와 같이, 공간 데이터와 소셜미디어 관련 내용이 함께 나타나는 주제도 발견되었다.

이러한 공간빅데이터 주제들은 기간에 따라서 각각 증가하거나 감소하는 패턴을 보일 수 있으며, 이 패턴을 이해하는 것은 향후 어떤 주제가 많이 발생할지를 파악하는데 도움이 될 수 있다. 최근의 주제 발생 추세를 중심으로, 기간별 주제 발생 분포를 살펴보기 위하여, 대상 논문을 5년 단위로 묶은 후에, 해당 논문에서 발생 사후확률이 높은 주제들의 기간별 분포를 분석하였다.

<Figure 9>에서 제시된 것과 같이, 공간빅데이터 관련 연구들은 2000년 이후부터 본격화되는 것으로 나타났다. 그리고 연구에서 나타나는 주제들의 발생 빈도는, 대체적으로 기간에 따라 점증하는 형태를 보이고 있다. 하지만, 그 중에서 몇몇 주제는 최근 발생빈도가 급증하거나, 또는 많이 발생하다가 최근에 다소 감소하는 것을 볼 수 있었다. 예를 들어, 공간빅데이터를 위한 IT 및 시스템, 관측위성을 통한 공간 분석 및 지리통계학 분야에 대한 12, 18, 24, 26번의 주제는 최근에 그 빈도가 급증하였다. 특히 12번 주제는 클라우드, 가상컴퓨팅, 스마트 컴퓨팅 등 최신의 IT가 시간/공간 분석(Spatiotemporal mining)에 적용되는 내용, 18번과 24번은 생태학 관련 위성 사진 분석에 대한 내용, 그리고 26번은 센서 기반 시간/공간 분석(Spatio-temporal analysis)에서의 Tree 및 Pruning에 대한 내용이어서, 공간빅데이터 분석에 대한 알고리즘과 컴퓨팅에 대한 관심이 크게 증가한 것을 볼 수 있다. 반면에 삼림 관련한 위성 이미지 분석, 생태계 종 간의 게임이론 적용, 위성이미지를 통한 지형 분류, 기상태풍 예측, 에어로졸 분석 등에 대한 8, 11, 16, 17, 23, 25, 27번의 주제는 2000년 후반에 많이 발생을 했고, 최근에는 오히려 빈도가 다소 감소하는 것으로 나타났다.

마지막으로, LDA의 결과로 얻어진 각 주제 간의 관계는 어떻게 되며, 어떤 주제가 중심적인지를 파악하기 위해 주제 네트워크를 분석하였다. 이때, 네트워크의 노드는 위에서 발견된 주제가 되며, 노드 간의 연결은 주제 간의 관련성을 나타낸다. LDA 결과 중 주제-문헌의 관계를 기반으로, 주제-주제 행렬을 얻었으며, 이때, 행렬의 원소는 서로 다른 두 주제에 공통으로 속하는 문헌의 빈도를 사용하였다. 이 행렬을 다시 그래프로 표현하여 방향성이 없는 40개의 노드와 780개의 엣지로 구성된 그래프를 얻을 수 있었다. 또한, 각 행의 대각의 값으로 각 행의 값들을 표준화하고, 노드 자신에 대한 Loop와 노드간 복수의 엣지를 정리하였으며, 가중치가 특정 값(이 논문에서는 0.2를 기준값으로 사용)보다 작은 엣지를 정리하여, 결과적으로는 다음 그림과 같은 40개 노드, 13개 엣지로 구성된 그래프를 얻을 수 있었다.

연구동향 네트워크에서는 크게 세 그룹의 연결된 노드들을 찾을 수 있었다. 우선, 좌측 상단의 그룹은 4, 9, 13, 15, 23, 32,

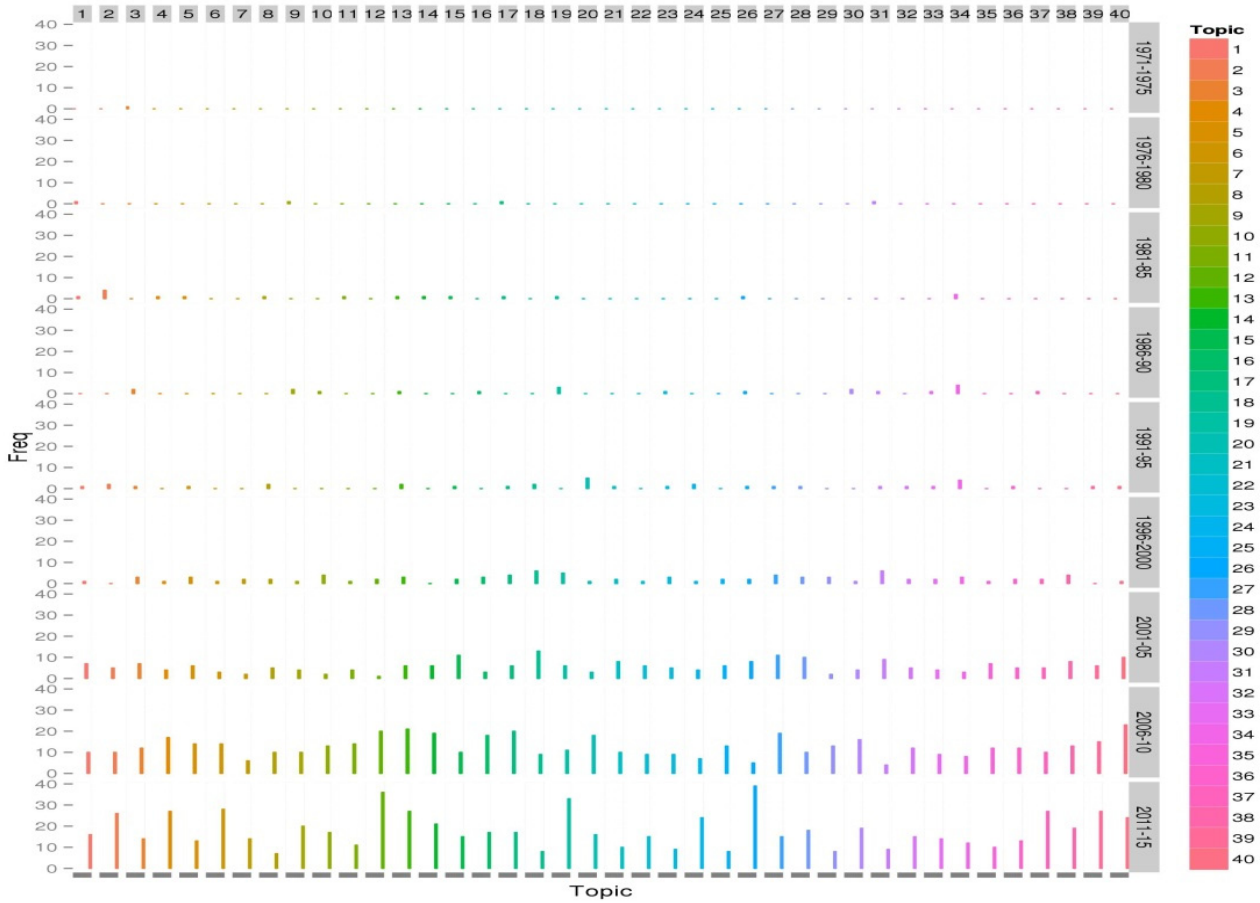

Figure 9. Distribution of topics on spatial big data over different period

38의 주제로 구성되었고(그룹 1), 좌측 하단의 그룹은 14, 22, 40의 주제로(그룹 2), 그리고 우측 상단의 그룹은 2, 3, 6, 12, 19의 주제로(그룹 3) 구성되었다. 이 중 첫 번째 그룹은 다시 4, 13, 15, 23의 주제로 이루어진 하위 그룹(그룹 1-1)과 9, 32, 38의 주제로 이뤄진 하위 그룹으로(그룹 1-2) 구분하였다. 그룹1-1에서는 자연환경과 대기, 기상 관측, 바이오매스 등의 내용이 다뤄지며, 특히 <Figure 9>에서와 같이 지리학 및 Kriging, Semi-variogram 등 공간 통계 기법에 대한 13번과 15번 주제가 최근에 크게 증가하였다. 그룹 1-2는 도시에 대한 지리 연구와 도시의 공해와 경제에 대한 연구가 농작물의 가격과 산업/경제적인 주제로의 연관을 보여주고 있는데, 도시에서의 GIS 분석을 통한 사회경제학적 접근에 대한 38번 주제가 최근 많이 증가한 것으로 나타났다. 그리고 그룹 1-1은 다시 그룹 1-2와 연관을 갖고 있는 것으로 나타나서, 그룹 1에서는 자연환경을 바탕으로 한 산업/경제적 주제로의 연결을 볼 수 있었다. 그룹2는 센서 및 로봇 관련 공간데이터 및 이미지분석을 통한 질병 예측, 그리고 베이지안 MCMC를 이용한 공간 분석 등의 주제로 구성되었으며, 최근에 빈도가 증가하였는 것으로 나타났다. 최근에 출현이 급증한 그룹3에서는 공간빅데이터의 효과적인 처리를 위한 클라우드/가상화/스마트 컴퓨팅과 같은 IT 인프라스트럭쳐 기술, 센서 기술 및 알고리즘, 시간/공간 분석기법 등의 내용이 도시에서의 재난, 교통, 전력망 등에 적용되는 주제들이었으며, 최근 가장 많은 관심을 받는 내용들로 나타났다.

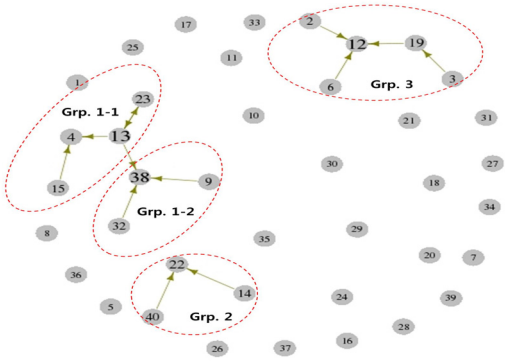

Figure 10. Research trend network on spatial big data

5. 결 론

본 연구에서는 공간빅데이터 분야의 연구동향 파악을 위해 국제저널에 실린 문헌들의 초록을 텍스트마이닝과 LDA를 사용하여 분석하였다. 그리고 이 연구의 결과를 바탕으로 공간빅

데이터의 연구 동향을 다음과 같이 정리해보았다. 첫째, 공간빅데이터 연구가 최근 빅데이터에 대한 관심과 함께 급증하고 있으며, 현재까지 대학 및 연구기관에서의 연구가 주로 이뤄지고 있다. 그리고 국가 중에서는 중국, 미국, 독일, 영국 등이 주도하고 있는 것으로 나타났다. 또한, Optics, Astronomy, Computer Science 분야 등 다양한 분야에서 공간빅데이터에 대한 접근이 이뤄지는 것을 볼 수 있었다. 둘째, 공간빅데이터 관련 문헌을 대상으로 우리는 텍스트마이닝과 토픽모델링을 적용하여 40여개의 주제를 발견할 수 있었다. 우선, 기존에 공간데이터를 많이 분석해온 생태학 관련 주제가 가장 많이 발생하였고, 기상, 도시 환경에서의 재해, 삼림관리, 지리학, 공간데이터 처리를 위한 IT에 대한 내용들이 많이 발견되었다. 흥미롭게도, 효과적이고 효율적인 공간빅데이터 분석을 위한 IT인프라 관련 주제, 클라우드/가상화/스마트 컴퓨팅 같은 최신 IT 트렌드의 공간빅데이터에의 적용, 그리고 해당 내용들의 도시 재난 상황에 대한 적용과 통신망이나 전력망과 같은 공간에서의 IT 및 서비스에 대한 활용이 중요한 주제로 부상하고 있었다. 추가적으로, 경작과 IT의 접목, 경작에서의 병렬처리, IoT 융합 등의 이종 분야간 주제가 같이 결합되는 양상도 발견할 수 있었다.

이러한 연구 동향을 바탕으로, 향후 공간빅데이터 연구는 공간 통계 기법들 중에서도 시간/공간 분석(Spatiotemporal mining)이 더 활발하게 적용될 것이 예측된다. 또한, 대용량의 데이터를 처리하기 위한 클라우드/가상화 컴퓨팅 기반의 인프라스트럭쳐 및 데이터를 관리/저장하는 자료 구조와 처리할 수 있는 알고리즘에 대한 관심이 더 급증할 것으로 기대된다. 그리고 이러한 내용들이 현대인의 대다수가 거주하는 도시 공간에 적용되는 추세가 심화될 것이며, 이 과정에서 기존 공간데이터 영역과 소셜미디어나 재난 데이터, 교통망과 전력망 등의 영역들이 융합되어 나타날 것으로 보인다. 그러므로 향후 공간빅데이터 분야에서의 연구는, 도시 환경에서 개개인의 사회적 관계 및 행동에 대해 공간데이터분석과 융합하여 이뤄지는 방향으로 수행될 것이 기대된다. 셋째, 주제 간의 관계를 주제 네트워크로 분석한 결과, 아직까지 주제 간의 연결이 많이 나타나지는 않았지만 그럼에도 기상 및 환경에 대한 연구, 경작 및 도시 관련 연구, 이미지 처리를 주로 다루는 방법론에 대한 연구, 그리고 소셜 및 도시 환경에서의 서비스에 대한 연구 등 크게 네 가지 방향으로 주제 간 연결이 발생함을 발견할 수 있었다. 이 결과를 바탕으로, 현재 주로 연구되는 주제 분야를 향후 연구 분야로 참고할 수 있을 것이다. 이 연구에서 제시된 경작과 IoT, 병렬처리와의 융합 주제 및 공간에 대한 이미지 분석과 소셜미디어 분석의 융합처럼, 아직 연결되지 않은 이종의 주제들을 기회영역으로 고려하여 연구 가능성을 타진해 나가면서 새로운 영역을 개척하는 것도 필요할 것으로 보인다.

이 연구는 다양한 분야에서 부상 중인 공간빅데이터 분야의 문헌을 수집하기 위해, 서명/초록/키워드로 제한된 필드에서 검색어 “spatial”, “big”, “data”가 같이 검색되는 키워드 검색을 사용했다. 하지만, 검색어 “spatial big data”는, 그 사용이 연구 목적에 더 적합함에도 아직까지는 검색된 문헌의 수가 매우 적었기 때문에 사용되지 못했다. 이 부분은 향후 공간빅데이터에 대한 연구가 보다 활발해지면서 보완될 수 있을 것으로 보인다. 또한, 논문 초록에 대해 기계적인 방법으로 연구 동향을 분석하였기 때문에, 언어처리와 해석, 주제도출 에서 일정 부분 한계가 있을 수 있으며, 보다 정교한 유망 주제 도출을 위해서는 전문가와의 협조가 중요할 것으로 보인다.

그럼에도 불구하고 아직까지 공간빅데이터 분야의 문헌을 바탕으로 한 연구동향 분석이 미미한 가운데, 다수의 문헌에 대해서 연구 동향을 제공한다는 점과, 이런 결과가 향후 추가적이고 구체적인 연구동향 분석에 기여할 수 있다는 점에 이 연구의 의의가 있다. 또한, 토픽모델링 기반으로, 문헌 분석으로 공간빅데이터를 다루기 시작한 연구로써, 발견된 연구 동향들이 후속 연구에 활용될 수 있다면 이 분야 발전에 기여할 것으로 기대된다. 특히, 후속연구에서는 공간빅데이터 관련 문헌들의 초록만이 아닌, 문헌의 전문을 대상으로 연구 동향 분석을 통한 보다 구체적인 주제 도출이 필요할 것으로 보인다. 또한, 시기별 연구동향을 구분하여, 향후 연구 동향의 예측도 시도할 수 있을 것으로 기대된다.

참고문헌

Ahn, J. W., Yi, M. S., and Shin, D. B. (2013), Study for Spatial Big Data Concept and System Building. *Journal of Korea Spatial Information Society*, **21**(5), 43-51.

Amirian, P., Basiri, A., and Winstanley, A. (2014), Evaluation of Data Management Systems for Geospatial Big Data. *In Computational Science and Its Applications-ICCSA 2014*, Springer International Publishing, 678-690.

Bae, J. H., Han, N. G., and Song, M. (2014), Twitter Issue Tracking System by Topic Modeling Techniques. *Journal of Intelligence and Information Systems*, **20**(2), 109-122.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), Latent dirichlet allocation, *Journal of Machine Learning Research*, **3**, 993-1022.

Chang, Y. S., Kim, J. C., Choi, W. G., and Kim, K. O. (2009), Study on the Development of Open Interfaced Geospatial Information Service Platform, *Journal of Korea Spatial Information Society*, **11**(1), 17-24.

Cho, G. H., Lim, S. Y., and Hur, S. (2014), An Analysis of the Research Methodologies and Techniques in the Industrial Engineering Using Text Mining, *Journal of the Korean Institute of Industrial Engineers*, **40**(1), 52-59.

Cho, S. G. and Kim, S. B. (2012), Finding Meaningful Pattern of Key Words in IIE Transactions Using Text Mining. *Journal of the Korean Institute of Industrial Engineers*, **38**(1), 67-73.

Choi, W. W., Hong, S. K., Shin, D. B., and Ahn, J. W. (2012), Concept of Spatial Information Social Platform and Role of Government as a Platformer, *Journal of Korea Spatial Information Society*, **20**(4), 37-45.

Goldberg, D., Olivares, M., Li, Z., and Klein, A. G. (2014), Maps and GIS Data Libraries in the Era of Big Data and Cloud Computing, *Journal of Map and Geography Libraries*, **10**(1), 100-122.

Guo, Z., Zhang, Z., Zhu, S., Chi, Y., and Gong, Y. (2014), A Two-Level Topic Model Towards Knowledge Discovery from Citation Networks, *Knowledge and Data Engineering, IEEE Transactions on KDa*, **26**(4), 780-794.

Hornik, K. and Grün, B. (2011), topicmodels: An R package for fitting topic models, *Journal of Statistical Software*, **40**(13), 1-30.

Jeong, D. H. and Song, M. (2014), Time gap analysis by the topic model-based temporal technique, *Journal of Informetrics,* **8**(3), 776-790.

Jian-ya, G. O. N. G. (2002), The Development Trends of the Contemporary GIS[J], *Northeast Surveying and Mapping*, **4**(3).

Kearney, M. and Porter, W. (2009), Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges, *Ecology Letters*, **12**(4), 334-350.

Kim, H. S. (2014), Policy Implication for implementing and utilizing Spatial Big data system, *Planning and Policy*, **389**, 6-11.

Kim, J. O., Huh, Y., Lee, W. H., and Yu, K. Y. (2009), Matching Method of Digital Map and POI for Geospatial Web Platform, *Journal of the Korean Society for Geospatial Information System*, **17**(4), 23-29.

Kim, J. J., Han, S. G., Shin, I. S., and Han, K. J. (2013), Spatial HBase: An Extension of HBase for Spatial Big Data, *Korea Information Science Society. Database*, **40**(5), 295-304.

Kim, M. S. (2014), Domestic and Overseas Policy on Spatial Big data and Technology Trend, *Planning and Policy*, **389**, 30-39.

Kim, S. W., Kim, G. G., and Yoon, B. K. (2014), A Study on a Way to Utilize Big Data Analytics in the Defense Area, *Journal of the Korean Operations Research and Management Science Society*, **39**(2), 1-20.

Lee, S. H. (2013), The Roles and Challenges of National Geospatial Data Infrastructure for Data Sharing and Openness in Big Data Era, *The Magazine of the Korean Society of Civil Engineers*, **61**(10), 37-42.

Masumura, R., Oba, T., Masataki, H., Yoshioka, O., and Takahashi, S. (2014), Role play dialogue topic model for language model adaptation in multi-party conversation speech recognition, *2014 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)*, IEEE, 4873-4877.

Messner, S. F., Anselin, L., Baller, R. D., Hawkins, D. F., Deane, G., and Tolnay, S. E. (1999), The spatial patterning of county homicide rates: An application of exploratory spatial data analysis, *Journal of Quantitative Criminology*, **15**(4), 423-450.

Schmidt, K. M., Menakis, J. P., Hardy, C. C., Hann, W. J., and Bunnell, D. L. (2002), *Development of coarse-scale spatial data for wildland fire and fuel management*, USDA Forest Service General Technical Report RMRS-GTR-62, USA.

Van Westen, C. J., Castellanos, E., and Kuriakose, S. L. (2008), Spatial data for landslide susceptibility, hazard, and vulnerability assessment: an overview, *Engineering Geology*, **102**(3), 112-131.

Wang, J. and Wang, X. (2011), An ontology-based traffic accident risk mapping framework, *In Advances in Spatial and Temporal Databases*, Springer Berlin Heidelberg, 21-38.

Wasserman, S. (1994), Social network analysis : Methods and applications, *Cambridge University Press, UK*, **8**.

Wu, Q., Zhang, C., Hong, Q., and Chen, L. (2014), Topic evolution based on LDA and HMM and its application in stem cell research. *Journal of Information Science*, 0165551514540565.

Yavuz, H. and Erdoğan, S. (2012), Spatial analysis of monthly and annual precipitation trends in Turkey, *Water Resources Management*, **26**(3), 609-621.

Yang, H. Y. (2012), Technology planning methodology using Big data, *KISTEP Issue Paper, Seoul, Korea*.

Yoon, M. Y. (2013), Analysis on Big Data Policy of Major country and Implication, *Science and Technology Policy*, **23**(3), 31-43.

Zhang, J. (2014). Trends in geo-information fusion for the next 5 years, *International Journal of Image and Data Fusion*, **5**(1), 1-1.