KIPS Tr. Software and Data Eng.
Vol.5, No.6 pp.295~306 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2016.5.6.295

Trend of Research and Industry-Related Analysis in Data Quality Using Time Series Network Analysis

Kyoung-Ae Jang† · Kwang-Suk Lee†† · Woo-Je Kim†††

ABSTRACT

The purpose of this paper is both to analyze research trends and to predict industrial flows using the meta-data from the previous studies on data quality. There have been many attempts to analyze the research trends in various fields till lately. However, analysis of previous studies on data quality has produced poor results because of its vast scope and data. Therefore, in this paper, we used a text mining, social network analysis for time series network analysis to analyze the vast scope and data of data quality collected from a Web of Science index database of papers published in the international data quality-field journals for 10 years. The analysis results are as follows: Decreases in Mathematical & Computational Biology, Chemistry, Health Care Sciences & Services, Biochemistry & Molecular Biology, Biochemistry & Molecular Biology, and Medical Information Science. Increases, on the contrary, in Environmental Sciences, Water Resources, Geology, and Instruments & Instrumentation. In addition, the social network analysis results show that the subjects which have the high centrality are analysis, algorithm, and network, and also, image, model, sensor, and optimization are increasing subjects in the data quality field. Furthermore, the industrial connection analysis result on data quality shows that there is high correlation between technique, industry, health, infrastructure, and customer service. And it predicted that the Environmental Sciences, Biotechnology, and Health Industry will be continuously developed. This paper will be useful for people, not only who are in the data quality industry field, but also the researchers who analyze research patterns and find out the industry connection on data quality.

Keywords : Data Quality, Social Network Analysis, Time Series Network Analysis, Textmining, Metadata

시계열 네트워크분석을 통한 데이터품질 연구경향 및 산업연관 분석

장 경 애† · 이 광 석†† · 김 우 제†††

요 약

본 연구는 데이터품질과 관련된 선행연구의 메타정보를 활용하여 연구경향을 분석하고 이를 통해서 산업계의 흐름을 예측하기 위한 목적의 연구이다. 다양한 분야에서 연구경향을 분석하려는 시도는 이어져 왔으나, 데이터품질 영역은 그 범위가 방대하여 선행 연구자료에 대한 분석을 수행하기 어려웠다. 본 연구는 Web of Science 색인DB에 수록된 최근 10년간의 연구 메타데이터를 수집하여 텍스트 마이닝, 사회연결망 분석기법을 활용한 시계열 네트워크 분석을 수행하였다. 연구주제 분석 결과, 수학 및 전산 생물학, 화학, 건강관리 과학 및 서비스, 생화학 및 분자 생물학, 운영 연구 및 경영 과학, 의료정보학은 연구비율이 감소하고 있었고, 환경, 수자원, 지질학, 계측기 및 계측의 연구비율은 증가하고 있었다. 또한 사회연결망 분석 결과 데이터품질 연구에서는 분석, 알고리즘, 네트워크의 주제가 중앙성이 높은 중요한 주제로 나타났으며, 이미지와 모델, 센서, 최적화가 데이터품질에서 중요한 주제로 등장하는 추세를 보였다. 데이터품질의 산업과 연관관계 분석 결과는 기술, 산업, 건강, 유틸리티, 고객서비스가 연관성이 높은 산업으로 나타났다. 본 연구의 결과는 데이터품질 연구의 패턴을 분석하고 산업과 연관관계를 찾는 데이터품질 관련 연구자 뿐아니라 산업계에도 유용한 자료로 활용되리라 판단된다.

키워드 : 데이터품질, 사회연결망 분석, 시계열 분석, 텍스트 마이닝, 메타데이터

1. 서 론

1.1 연구배경 및 목적

정보화와 네트워크 기술이 급속히 발전하면서 현대인들은 자동화를 통해 데이터를 수집되고 가공하고 누적해 정리하고 있다. 오늘날 조직은 정보시스템이 업무의 자동화와 효

† 준 회 원 : 서울과학기술대학교 IT정책전문대학원 산업정보시스템전공 박사과정
†† 비 회 원 : 서울과학기술대학교 디지털문화정책학과 교수
††† 비 회 원 : 서울과학기술대학교 글로벌융합산업공학과 교수
Manuscript Received : April 15, 2016
First Revision : May 10, 2016
Accepted : May 10, 2016
* Corresponding Author : Woo-Je Kim(wjkim@seoultech.ac.kr)

율화를 통해 제공해줄 수 있는 기술적 만족에 그치지 않고, 데이터를 통해 또 다른 다양한 가치와 품질을 확보하고자 한다. 데이터 산업 시장에서 빅데이터의 트렌드가 이어지면서, 사람들은 점차 정형 데이터를 넘어서서 반정형이나 비정형 데이터에 관심을 갖고 그 속에서 차별화된 정보 가치를 이끌어 내고자 한다. 하지만, 보다 근원적인 조직의 가치 창출을 위해서는 '데이터품질'이 확보되어야 한다.

데이터의 품질이 저하되면 의사결정에 오랜 시간이 걸리고, 기존 데이터를 활용한 설계가 어려워지며 고객은 조직을 불신하게 된다. 조직 운영 측면에서는 오류 데이터로 인하여 고객의 만족도가 떨어지고 불만이 증대된다. 전술적 측면에서는 조직의 비전 달성을 위한 정확한 의사결정을 수행하기 어렵게 되며, 전략적 측면에서는 계획을 수시로 수정해야 하므로 전략 실행에서 고객의 신뢰를 얻을 수 없게 된다[1]. 이러한 다양한 문제를 통해서 기업의 신뢰도 및 이미지 실추에 따라 자산의 손실, 보수비용 증대 등으로 사업 존폐 위기로 이어질 수 있다.

의료, 통신, 건축, 토목, 보안, 금융, 예술, IT 등 다양한 영역들에서 데이터품질에 대한 연구의 중요성을 인지하고 지속적인 연구가 이루어지고 있다[2]. 그러나 데이터품질에 대한 기대와 연구의 다양성에 비해 데이터품질의 연구경향이나 동향을 통한 산업트렌드 분석은 찾아보기 어려운 실정이다. 그 이유는 데이터품질 관련 연구가 그 도메인이 방대하여 데이터를 분석하기에는 최적화된 분석기술과 자원 및 시간이 크게 소모되기 때문이다. 그럼에도 불구하고 데이터품질에 관한 연구경향을 거시적으로 분석하는 것은 데이터품질의 시계열적 연구흐름을 이해하는데 필요하다. 더불어 데이터품질 분석을 통해 연구에 대한 중복투자를 방지할 수 있고 연구의 방향을 정립할 수 있으며 신규 연구 혹은 깊이 있는 연구의 기초자료로 활용할 수도 있다. 이러한 학계 연구의 흐름을 읽는 일은 관련 산업의 데이터품질에 대한 연구 흐름을 분석하고 산업 트렌드를 예측하는 것과 직접적 연계를 갖는다.

본 연구에서는 최근 10년 동안 국내·외 '데이터품질'과 관련된 학술 논문 및 연구결과물의 메타정보를 웹 크롤링하여 계량서지학 방법으로 정량적인 분석을 수행하였다. 선행연구의 전통적인 계량서지학 방법은 키워드, 단어의 빈도수를 통계적으로 분석하여[3] 개별적 조사를 하는데 그쳐 종합적인 관계나 연구 경향과 방향을 보여주기에는 부족하였다. 또한 분석한 데이터양이 일반화하기에는 불충분 하거나[4] 유사어 중복 데이터의 표준화에 대한 구체적인 방법을 제시하지 않거나 임의 분류에 의한 방법을 수행하였다[3-8]. 따라서 본 연구는 선행연구와 차별화된 방법을 제시하고 있다. 즉 메타데이터를 기반으로 계량서지학 방법에서 도출된 메타데이터의 연관도 분석, 네트워크 분석을 실시하여 데이터품질 연구의 구조와 방향을 제시하였다. 구체적으로 데이터 분석의 신뢰도를 높이기 위하여 동의어 사전을 통한 중복 데이터 통합 과정과 표준산업분류코드를 활용한 산업동향 분석 단계를 추가하여 실험의 완성도를 높였으며, 연구결과를 시계열적으로 분석하여 연구경향 파악에 용이하도록 하였다.

본 연구는 연구문헌의 메타정보를 활용하여 데이터품질의 거시적 연구 흐름을 처음으로 분석한 연구라는 점에서 그 의의가 크다고 하겠다. 또한 연구 정보 속성간의 연관 관계와 네트워크 분석을 시계열적으로 제시하여 데이터품질 연구의 구조를 파악할 수 있고, 선행연구가 특정 도메인이 깊게 연구되고 있는지 혹은 다양한 연구도메인으로 확대되고 있는지를 알 수 있다. 이를 통해서 학계에서는 거시적인 데이터품질 연구의 상태진단과 연구동향 예측이 가능하고 산업계 흐름을 파악할 수 있다. 산업계에서는 IT정책 방향과 시장의 대응전략 구상에 참고할 수 있다.

본 연구의 구성은 다음과 같다. 2장에서는 연구의 이론적 배경과 관련된 선행연구를 소개하고, 3장에서는 연구의 설계 방법을 설명하고 실험결과를 분석한다. 선행 연구데이터의 연관관계를 네트워크 분석과 표준산업분류코드의 매핑으로 산업의 흐름을 예측한다. 마지막 4장에서는 결론과 향후 추가적으로 수행되어야 할 연구과제에 대해서 논의한다.

1.2 연구의 문제

본 연구는 데이터품질에 관련된 선행연구를 통해서 연구경향을 분석하고 이를 통해서 산업계의 흐름을 예측하기 위한 목적을 지니고 있다. 데이터품질 영역은 그 범위가 방대하고 대량의 데이터로 인하여 선행 연구자료에 대한 분석을 수행하기 어려웠다. 본 연구는 대량의 데이터를 수집하기 위하여 세계적인 연구결과물의 색인을 최다 보유하고 있는 Web of Science의 연구 데이터를 수집하고 분석하였다. 이를 통하여 데이터품질 연구의 패턴을 분석하고 산업과 연관관계를 찾아 데이터품질 관련 연구자 뿐만 아니라 산업계에도 유효한 자료로 활용되리라 판단된다. 이를 위해서 본 연구는 다음과 같은 연구문제 해결의 목적으로 진행되었다.

첫째, 최근 데이터품질 연구의 경향과 진행하는 패턴은 어떠한가? 둘째, 데이터품질 연구를 통한 데이터품질 산업은 최근 10년간은 어떠한 거시적 흐름으로 이어졌으며, 향후 트렌드는 무엇인가?

첫번째 연구문제는 데이터품질 연구 문헌을 수집하고 메타정보 분석을 통하여 데이터품질 분야의 연구가 추구하는 방향성을 확인하고자 하는 목적을 갖는다. 데이터품질에 관한 연구는 다양한 분야에서 수행되었으므로 시계열 분석을 네트워크 기반으로 수행하면 연구의 상관관계에서 어떻게 진행되었는지에 대한 패턴이나 경향성을 발견할 수 있을 것이다. 의미있는 데이터품질 연구의 진행 패턴 분석과 결과는 학계 및 산업계에서 중요한 자료로 활용되리라 사료된다. 구체적으로 데이터품질 연구 경향이 특정 지향점을 갖고 움직이는지 혹은 변화 없이 동일 영역에서 한정돼 연구되어 왔는지 등을 보고 문제점이나 방향을 제언하고자 한다. 두 번째 연구문제는 데이터품질 연구와 산업 트렌드의 상관관계를 찾고자 하는 목적을 지닌다. 최근 10여년간의 데이터품질 연구 자료에서 추출한 키워드를 산업분류코드와 함께 매핑하여 시계열적으로 분석하면 산업 트렌드를 확인할 수 있을 것이다. 이를 통해서 과거 데이터품질 산업계를

파악하고 향후 미래 트렌드를 예측하는 효과를 얻고자 한다. 결국 이는 데이터품질 관련 학계 자료가 산업계에 유효하게 활용될 수 있음을 확인하는데 있다.

2. 선행연구 고찰 및 이론적 배경

2.1 이론적 배경

1) 사회연결망 중앙성

사회연결망 이론은 사람의 행위를 이해하기 위해 인간 개개인의 속성과 그들이 맺고 있는 관계를 통해 인간행위와 사회구조를 설명하는 이론이다[9]. 이러한 사회연결망에서 '중앙성'은 한 인간이 다른 인간과의 관계에서 중심이 됨으로써 얻는 효과를 의미하며, 집단 내 영향력이 증대되고, 개인의 목표를 달성하기 용이한 위치라고 볼 수 있다. 이러한 중앙성은 연결정도(degree)와 밀도(density)로 분석할 수 있다. 연결정도는 한 접점(node)에서 맺고 있는 다른 접점의 숫자의 합으로 중앙성이 높은 경우 연결정도가 높다[10]. 또한 밀도는 가능한 총관계수 중에서 실제로 맺어진 관계수를 의미하며 밀도가 높을수록 접점은 중앙성이 높다. 중앙성은 권력과 영향력의 의미를 갖고 있으며 중앙성이 높으면 사회에서 경제적 위치를 갖는 사람이고, 기업의 승진기회가 높고 생존률이 높은 위치이다[10]. 따라서 본 연구에서는 사회연결망 분석의 중앙성을 이용하여 접점을 연구논문으로 두고 각 메타데이터와 어떠한 관계로 구성되어있는지 분석하는 연구를 수행하였다. 이를 통해서 중앙성이 높은 연구분야, 키워드, 산업분야를 시계열별로 분석하였다. 중앙성에는 연결정도 중앙성(degree centrality), 인접 중앙성(Closeness centrality), 사이 중앙성(Betweenness centrality)이 존재한다. 연결정도 중앙성(degree centrality)은 다른 접점과 연결된 정도를 의미하며, 연결망 내에서 한 접점에 연결된 노드들의 합으로 연결정도 중앙성을 판단한다. 연결정도 중앙성은 중앙으로 들어오는 연결정도인 내향중앙성과 중앙에서 밖으로 나가는 연결정도인 외향중앙성이 존재한다. 연결정도 중앙성이 높으면 마당발 효과를 얻을 수 있고, 원하는 정보를 얻을 확률이 높고 권력에 커진다[10]. 본 연구에서는 연결정도 중앙성을 도출하여 연구분야의 영향정도를 파악하는데 활용하였다.

$$D_r = \frac{D_a}{n-1},\ D_a : \text{연결관계수},\ n : \text{총노드수}$$

인접 중앙성(Closeness centrality)은 한 접점이 다른 접점에 얼마나 가까운지 정도를 설명하는 중앙성이다. 연결망 내에서 접점간의 최단거리에 위치한 접점이 인접 중앙성이 높다. 인접 중앙성이 커지면 연결망의 핵심인물과 가까운 곳에 위치하여 주요 권력 확보가 쉬워진다[10]. 본 연구에서는 인접 중앙성을 도출하여 신규 연구분야를 파악하는데 활용하였다.

$$C_r = \frac{n-1}{C_a},\ C_a : \text{해당노드의 연결거리합},\ n : \text{총노드수}$$

사이 중앙성(Betweenness centrality)은 연결망 내에서 한 접점이 다른 접점으로 이동할 경우 거쳐 가는 사이에 위치하는 정도를 나타내는 중앙성이다. 사이 중앙성이 커지면 중재자로서 다른 행위자들이 의존하는 정도가 커지므로 의존 영향력이 커진다[10]. 본 연구에서는 사이 중앙성을 데이터품질의 융합연구 연구주제 영역을 파악하는데 활용하였다.

$$B_r(i) = \sum_{j<k} \frac{N_{jk}(i)}{N_{jk}}$$
$$,\ N_{jk} : \text{임의의 } j\text{와 } k\text{간의 } i\text{노드를 포함한 최단경로수}$$

2) 텍스트 마이닝

텍스트 마이닝은 대량의 텍스트 데이터에서 유용한 패턴 및 관계를 발견하고 추출하는 과정으로 자연어처리 기술, 문서처리 기술 및 데이터마이닝 알고리즘을 적용하여 효율적인 텍스트를 분석하는 연구분야이다. 텍스트 마이닝은 텍스트를 추출하여 데이터의 클러스터링, 문서 검색, 이메일 감시 및 필터링, 이상징후 검출[11-13]등 다양한 영역에서 사용되고 있다. 텍스트 마이닝은 데이터의 수집, 전처리를 통한 데이터 정제, 변수 선택 및 추출, 알고리즘 학습 및 평가의 단계를 거친다. 수집된 데이터는 전처리 과정을 거치는데, 전처리 과정은 데이터를 파싱하고 불용어를 제거하고 단어표준화를 통해서 대표단어를 변환하는 등의 작업을 수행한다. 정제된 데이터는 분석하고자 하는 목적에 따라 변수를 선택하고 데이터 알고리즘을 적용해서 그 결과를 분석한다. 텍스트 마이닝은 데이터마이닝의 분야이며 수집하고 전처리하는 대상의 데이터가 텍스트에 기반한 분석을 의미한다. 본 연구에서는 색인논문 DB에서 메타데이터를 수집하고, 데이터 항목에 따라 데이터를 파싱하여 분류하고 동의어사전, OECD학문분류에 준하여 표준화를 실시하고 데이터의 출판년도 기준으로 TF(Term Frequency)와 분류화(Classification) 분석을 실시하였다.

2.2 선행연구의 고찰

연구논문의 메타 데이터를 활용하여 연구동향을 분석하는 연구는 다양한 영역에서 진행되고 있었다. 특히 키워드를 기반으로 한 연구[3-6]와 키워드를 포함한 여러 메타데이터의 조합에 의한 연구가 수행되었다[3-8]. [3]의 연구에서 동물행동에 대한 연구동향을 분석하기 위하여 키워드와 어휘간의 시멘틱 분석을 진행하였다. 그러나 개별적인 정량적 분석에 그치면서 종합적인 연관성 분석에는 한계가 존재하였다. [4]의 연구에서 최근 10년간의 기술경영분야 해외저널 3개에서 2,611개의 논문에서 키워드를 수집하여 네트워크 중앙성 분석을 실시하였다. 이 연구에서는 수집된 키워드를 그룹화하여 키워드의 논문별 분포도를 조사하여 90%의 키워드가 10년 동안 1번 이하로 사용되어 '좁은세상' 네트워크

의 특징을 따르고 있다는 확인을 할 수 있었다. 그러나 이 연구에서는 논문 수집 대상을 넓히지 못한 점과 용어 표준화의 한계점이 존재한다. [5]연구에서는 컴퓨터 공학분야 논문에서 키워드를 수집하여 연구하였다. 연구자는 키워드에 의한 연구경향 분석에 그치지 않고 시간의 흐름에 따른 키워드의 생성과 소멸에 흐름을 분석하였다. 이 연구를 통해서 년차별로 존재하는 키워드가 소멸하거나 생성되는 키워드에 비해 많다는 것을 알 수 있었다. 이는 연구패턴 분석에는 유용한 자료로 활용될 수 있을 것으로 보이나 산업과 연계된 예측분석은 수행하지 않았다. [6]의 연구에서 캐나다 커뮤니티 임업 연구를 계량서지학 방법으로 수행하였다. 캐나다 지역사회의 임업 연구자료를 기준으로 출판년도, 작가수, 성별, 제목 등 다양한 메타정보를 추출하여 연구동향을 분석하였으며 향후 연구의 기초자료 활용방안을 제시하였다. 또한 도자기 분야의 연구경향을 분석한 [7]연구와 갬블과 로토에 관한 연구를 수행한 연구[8]도 존재하였다.

이들 선행연구는 개별적인 통계분석의 수행에 그쳤거나 수집된 데이터량이 부족하여 해당 산업에 대한 경향분석으로 일반화하기엔 불충분한 경우가 존재하였다. 또한 텍스트 기반의 데이터분석 연구에서 가장 중요한 부분인, 중복 데이터 제거 및 표준화를 통한 오류 없는 데이터로 신뢰성을 기반으로 한 데이터분석을 간과한 아쉬움이 발견되었다. 따라서 본 연구에서는 동의어사전과 표준산업분류코드를 기반으로 데이터의 신뢰성과 최적화를 확보한다. 또한 데이터의 통계정보 분석에 그치지 않고 키워드 네트워크분석과 메타데이터 분석으로 산·학계에서 실효성 있는 연구를 수행하고자 한다.

3. 연구 설계 및 결과

3.1 연구모델 설계

본 연구는 데이터품질 관련 국내·외 논문 및 학술지에 사용된 키워드와 서지정보를 포함한 메타데이터를 활용하여 구조적 패턴을 분석하는 연구이다. 본 연구의 구성은 Fig. 1과 같이 데이터 수집과정, 데이터 분석과정, 분석결과 도출과정으로 3단계를 거친다.

먼저 데이터 수집과정에서는 분석대상 연구자료를 수집하고 해당 데이터를 실험에 적합하게 전처리하는 단계를 거친다. 대상 메타데이터시스템은 전세계 연구자료의 메타정보를 제공하는 'Web of Science'로 선정하였다. Web of Science는 국내뿐 아니라 해외의 각종 메타데이터시스템에서 보유하는 정보를 포괄하여 자연과학(SCI/SCIE), 사회과학(SSCI), 인문예술(AHCI) 분야 등의 핵심저널 12,000여종, 10억개 이상의 전세계 학술문헌을 보유하고 있기 때문이다. 본 연구에서는 최근 데이터품질의 연구경향 분석을 위하여 2005년부터 2015년까지 데이터를 수집하여 분석한다. 수집하는 데이터는 연구의 키워드와 연구 제목, 저자, 학술지명, 연구분야, 출판년도, 출판사명, 언어 등의 메타데이터이다.

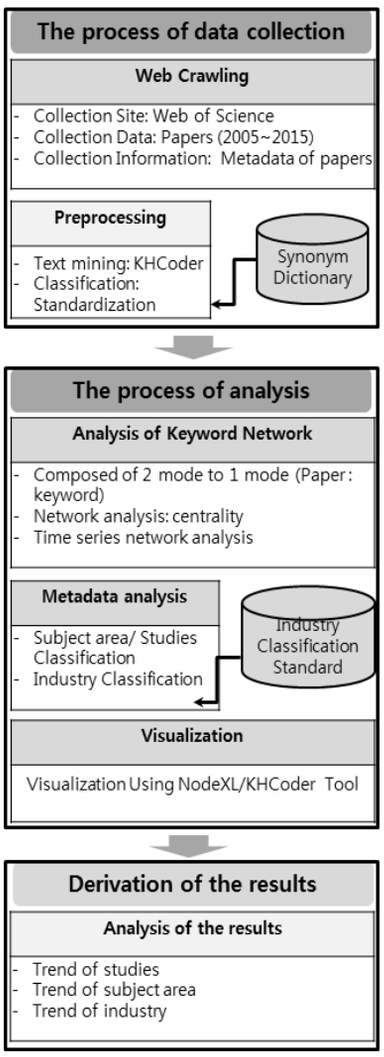

Fig. 1. Research Framework

그 다음 단계는 데이터 전처리 단계이다. 수집된 데이터는 실험에 적합하도록 파싱하고 불용어 및 중복데이터를 제거하는 작업을 거친다. 의미 있는 단어로 파싱된 단어는 메타데이터 분류항목에 따라 분류하여 데이터 분석을 준비한다. 이렇게 정제된 데이터는 저자마다 동일한 의미를 다른 단어로 사용한 경우가 존재하므로, 추가적으로 데이터를 표준화 작업을 수행한다. 데이터 표준화는 동의어 사전을 기준으로 1차적으로 그룹핑하고, 동일한 단어가 들어있는 키워드를 동일 단어군으로 그룹핑하여 중복을 제거하였다.

두 번째는 데이터 분석단계이다. 이 단계에서는 키워드 네트워크 분석과 서지 통계정보 분석을 수행한다. 먼저 키워드 네트워크 분석은 사회연결망 기술을 활용하며, 준연결망 네트워크 분석을 위하여 논문과 키워드를 2 mode to 1 mode방식으로 전환하여 매핑맵을 구성하고, 구성된 매핑맵을 통해서 중앙성 분석을 한다. 키워드의 밀집 정도와 키워드 중심에서 연결되는 노드의 정도가 큰 연결정도 중앙성(degree centrality), 인접 중앙성(Closeness centrality), 사이 중앙성(Betweenness centrality)을 분석하여 연구경향을 파

악한다. 또한 이 단계에서는 출현빈도가 큰 키워드가 시계열적으로 어떠한 변화가 있는지를 분석하기 위하여, 출판년도와 키워드의 준연결망 분석을 수행하여 데이터품질의 연구 흐름을 도출한다. 다음은 수집된 서지 정보를 기준으로 주제영역별, 국가별, 산업분류별 연구 추이를 통계적으로 분석한다. 이때 연구와 산업분류의 매핑분석을 위해서 표준산업분류코드를 활용하여 분류한다. 그리고 마지막으로 가시화된 결과와 통계정보를 통해서 데이터품질의 현 시계열적 연구경향과 산업계의 현 상태를 파악하고 향후 연구방향을 제시하고자 한다.

3.2 연구결과 및 토의

1) 키워드 빈도 분석

키워드 분석을 위하여 Web of Science에서 데이터품질에 관한 연구문헌의 메타데이터를 웹크롤링하여 Table 1과 같이 2005년부터 2015년까지 데이터를 수집하였다. 데이터품질에 관한 연구는 2005년 835건에서 2014년 4,350건, 2015년 4,466건으로 데이터품질에 관한 연구는 지속적으로 늘어나고 있었다.

Table 1. Total Dataset for Analysis

Year	2005	2006	2007	2008	2009	2010
Count	835	1,030	745	882	2,287	3,033
Year	2011	2012	2013	2014	2015	Total
Count	3,455	3,835	4,015	4,370	4,466	28,953

연구자가 명시한 연구문헌의 저자키워드의 빈도와 연구제목의 키워드를 파싱하여 도출된 저자키워드의 빈도를 분석하여 연구경향을 분석하였다. 저자키워드는 연구제목을 형태소 분석하고 불용어를 제거하는 작업을 거쳤다. 그 결과 Table 2, Table 3과 같이 저자 키워드는 network, analysis, model, management 순으로 빈도수가 높게 나타났으며, 제목 키워드에서는 network, analysis, model, image, water 순으로 키워드 빈도수가 높게 나타나 저자 키워드와 제목 키워드는 유사한 결과를 보였다.

Table 2. The results of author keyword analysis (Top 20)

Rank	keyword	Count	Rank	Keyword	Count
1	network	14,972	11	optimization	5,218
2	analysis	13,768	12	modeling	5,158
3	model	13,176	13	method	5,140
4	management	8,569	14	sensor	5,096
5	water	8,161	15	mining	4,635
6	control	8,019	16	performance	4,623
7	algorithm	7,429	17	software	4,600
8	service	6,476	18	power	4,423
9	process	6,197	19	learning	4,129
10	design	5,330	20	energy	4,056

Table 3. The results of title keyword analysis (Top 20)

Rank	keyword	Count	Rank	Keyword	Count
1	network	2,815	11	process	1,163
2	analysis	2,537	12	optimization	1,004
3	model	2,400	13	modeling	941
4	image	1,687	14	design	938
5	water	1,461	15	method	931
6	control	1,451	16	sensor	929
7	management	1,447	17	mining	893
8	algorithm	1,325	18	performance	816
9	information	1,276	19	software	786
10	service	1,180	20	learning	775

또한 키워드의 구조적 연관성을 분석한 결과 9개의 단어가 전체 키워드 단어의 87.24%를 차지하고 나머지가 12.76%에 집중되어 있는 구조를 확인할 수 있었다. 이 구조를 그래프로 옮기면 Fig. 2, Table 4와 같이 멱함수 분포를 보여주고 있다는 것을 확인할 수 있고, 이것은 [14]의 연구에서 언급된 좁은 세상 네트워크(Small world net)구조가 연구키워드에도 존재함을 확인할 수 있었다.

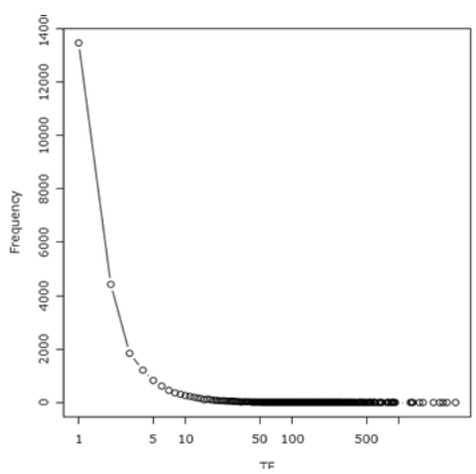

Fig. 2. A distribution curve of power laws about keywords

Table 4. The results of keyword structure analysis (Top 10)

TF	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	13463	49.96	13463	49.96
2	4424	16.42	17887	66.38
3	1843	6.84	19730	73.22
4	1211	4.49	20941	77.71
5	826	3.07	21767	80.78
6	617	2.29	22384	83.07
7	456	1.69	22840	84.76
8	360	1.34	23200	86.1
9	308	1.14	23508	87.24
10	261	0.97	23769	88.21

또한 상위 키워드는 유사한 데이터를 통합하기 위하여 http://www.thesaurus.com/에서 동의어(Synonyms) 사전의 동의어를 기준으로 1차적으로 그룹핑하고, Table 5와 같이 동일한 단어가 들어있는 키워드를 동일 단어군으로 그룹핑하여 중복을 제거하였다.

Table 5. Examples of keyword standardization using synonyms

Standard Keyword	Keywords
network	network, chain, networks, networking, in-network, networked 등
model	model,exemplary, models, e-model, multi-model, q-model, gis-model 등
analysis	analysis, analysis, inquiry 등
estimation	estimation, evaluation, m-estimation, appraisal, self-assessment 등
management	management, administration 등
control	control, authority 등
service	service, data-as-a-service, serviceability, e-service, supply 등
algorithm	algorithm, algorithms, m-algorithm 등

2) 연구 도메인 분석

수집된 연구문헌의 메타정보에서 연구주제를 도출하여 데이터품질 관련된 연구문헌의 연구 도메인을 시계열적으로 분석하였다. 연구 도메인의 학문분류는 OECD에서 제공하는 학문분류기준에 우선하여 분류하고 미 존재할 경우는 WOS (Web of Science)분류에 근거하여 수집된 연구문헌의 연구주제와 매칭하여 분석하였다. 이는 연구 도메인의 학문에서 상위 20개를 기준으로 년차별 평균수치를 기준으로 분석되었다. 그리고 Fig. 3과 같이 연구분야의 시계열적인 분석을 통하여 학문분야 발전과 쇠퇴 양상을 확인 할 수 있었다. 상위 20개의 연구 학문분야는 방사선학, 핵의학 및 의료 영상(Radiology, Nuclear Medicine & Medical Imaging), 수학 및 전산 생물학(Mathematical & Computational Biology), 운영 연구 및 경영 과학(Operations Research & Management Science), 의료 정보학(Medical Informatics), 이미징 과학 & PHOTOGRAPHIC 기술(Imaging Science & Photographic Technology), 자동화 및 제어 시스템(Automation & Control Systems), 정보 통신(Telecommunications), 컴퓨터공학(Computer Science), 화학(Chemistry), 건강관리 과학 및 서비스(Health Care Sciences & Services), 공학일반(Engineering), 생화학 및 분자 생물학(Biochemistry & Molecular Biology), 정보학 및 도서관학(Information Science & Library Science), 환경공학(Environmental Sciences), 계측기 및 계측(Instruments & Instrumentation), 생명 공학 및 응용 미생물학(Biotechnology & Applied Microbiology), 지질학(Geology), 경제(Economy), 수자원(Water Resources), 약리학 및 약국(Pharmacology & Pharmacy) 으로 나타났다.

연구 학문분야의 시계열 분석에서 컴퓨터공학(42%)과 공학일반(16%)이 가장 높게 나타났고, 10년간 지속적으로 많은 연구가 이루어졌으나 2014년부터는 다른 학문분야가 늘

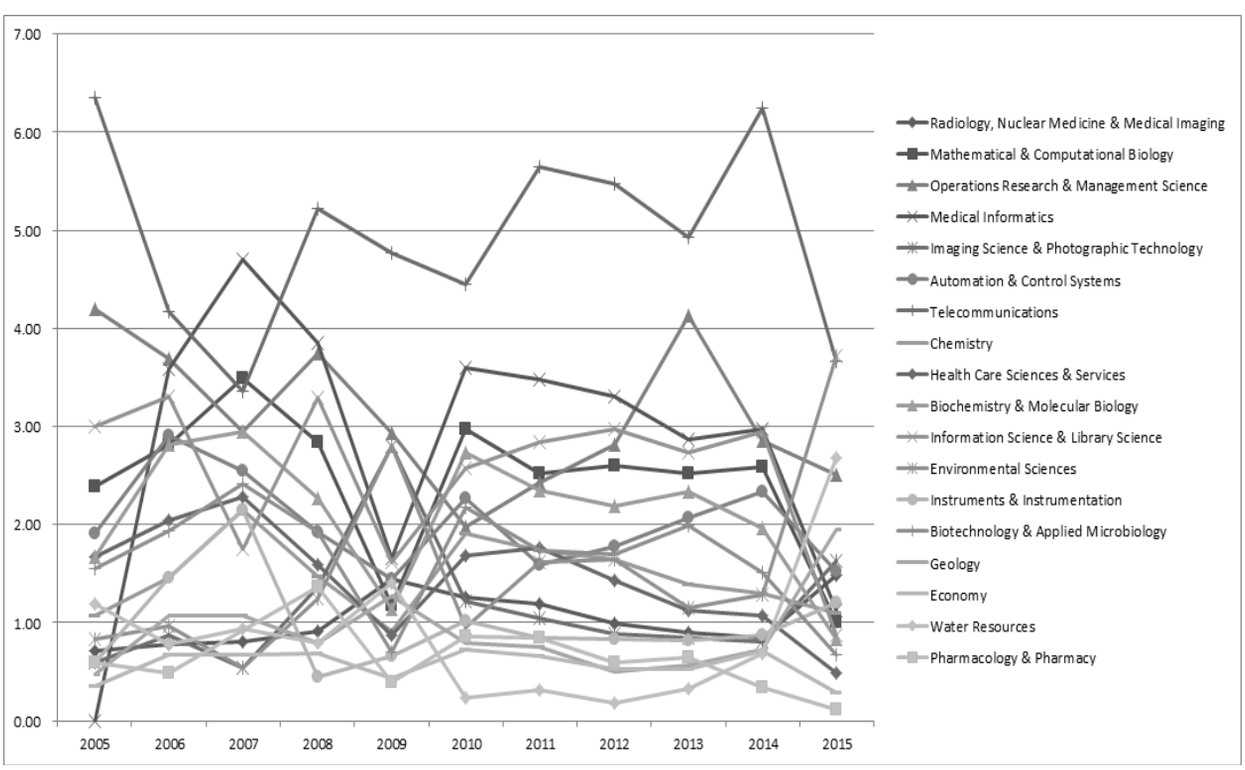

Fig. 3. The results of time series analysis of studies

어나면서 이들 연구 비율이 감소되고 있는 추세를 보였다. 컴퓨터공학과 공학일반을 제외한 상위 10위권의 학문분야의 추세를 분석해 보면, 2005년에는 정보통신과 운영연구 및 경영과학, 정보과학, 의료정보학, 수학, 지리학 등의 연구가 진행되었으며, 10년이 지난 후 2015년의 연구는 환경, 정보통신, 수자원, 운영 연구 및 경영 과학, 에너지 등의 연구 형태로 변화되고 있었다. 정보통신분야의 연구는 지속적으로 높게 나타났으나 2015년 환경관련 연구가 활발해지면서 낮아지는 비율로 나타났다. 2년 연속적으로 증가 혹은 감소 경향을 기준으로 연구 학문분야의 추이를 분석해보면, 수학 및 전산 생물학, 화학, 건강관리 과학 및 서비스, 생화학 및 분자 생물학, 운영 연구 및 경영 과학, 의료정보학은 감소비율로 나타났고, 환경, 수자원, 지질학, 계측기 및 계측은 증가비율로 나타났다.

3) 연구 네트워크 중앙성 분석

연구문헌의 저자 키워드를 중심으로 2005년부터 2015년까지 11년간에 데이터품질에 관한 연구의 연관관계 정도를 통하여 연구의 경향분석을 수행하였다. 키워드 간의 연관관계 분석을 위하여 연결정도 중앙성(degree centrality), 인접 중앙성(Closeness centrality), 사이 중앙성(Betweenness centrality)을 시계열 분석하였다. 10년간 키워드 중앙성이 높은 상위 10위 키워드는 Table 6과 같이 나타났으며 연결정도 중앙성을 NodeXL로 도식화 한 결과 Fig. 4와 Fig. 5로 강결합 구조를 갖고 있었다.

Table 6. The results of centrality analysis (Top 10)

Rank	Betweenness Centrality	Closeness Centrality	Degree Centrality
1	analysis	analysis	analysis
2	network	algorithm	network
3	algorithm	network	algorithm
4	model	process	model
5	control	system	process
6	process	management	control
7	system	mining	system
8	image	optimization	image
9	optimization	clustering	management
10	management	control	optimization

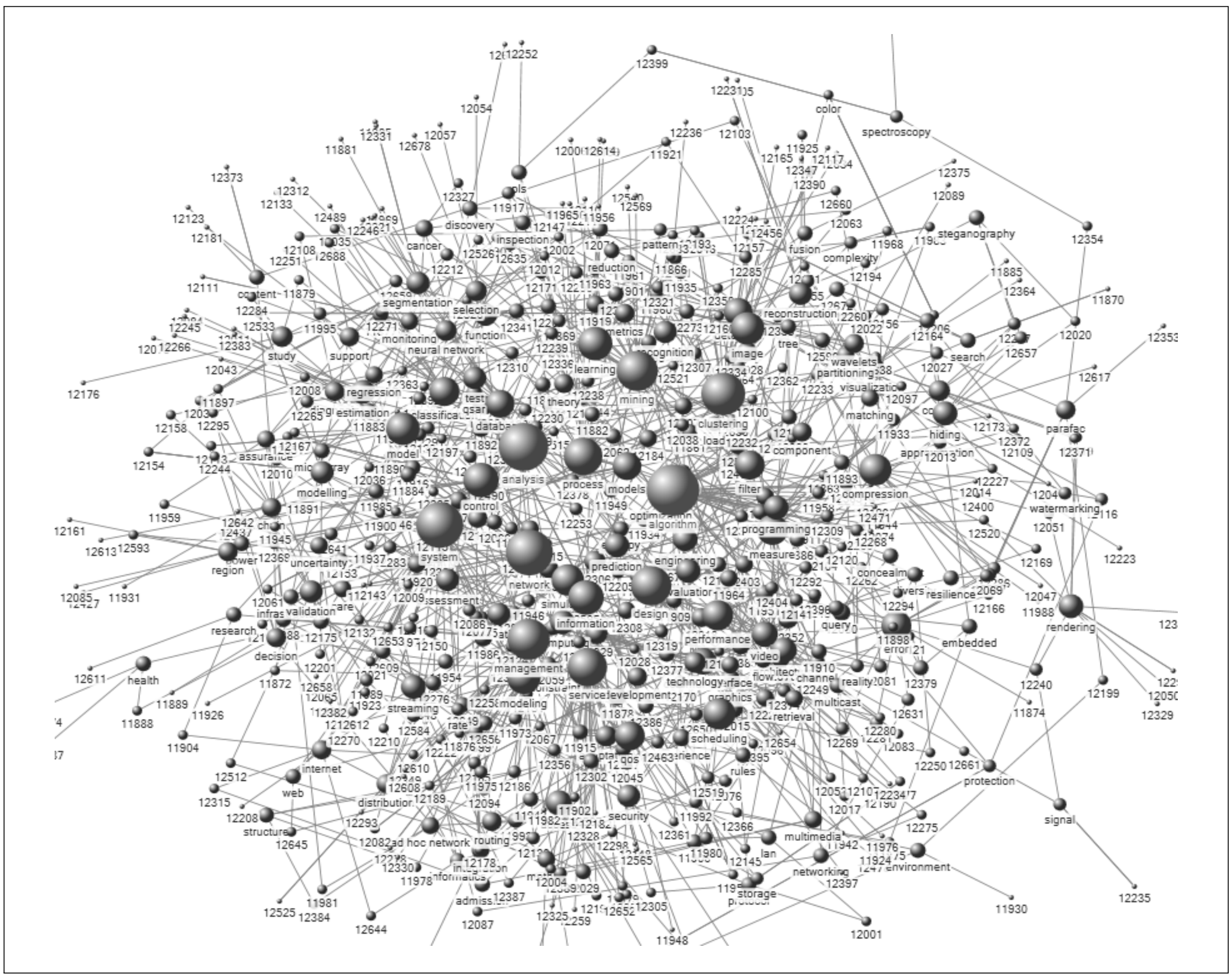

Fig. 4. The results of betweenness centrality analysis(2005)

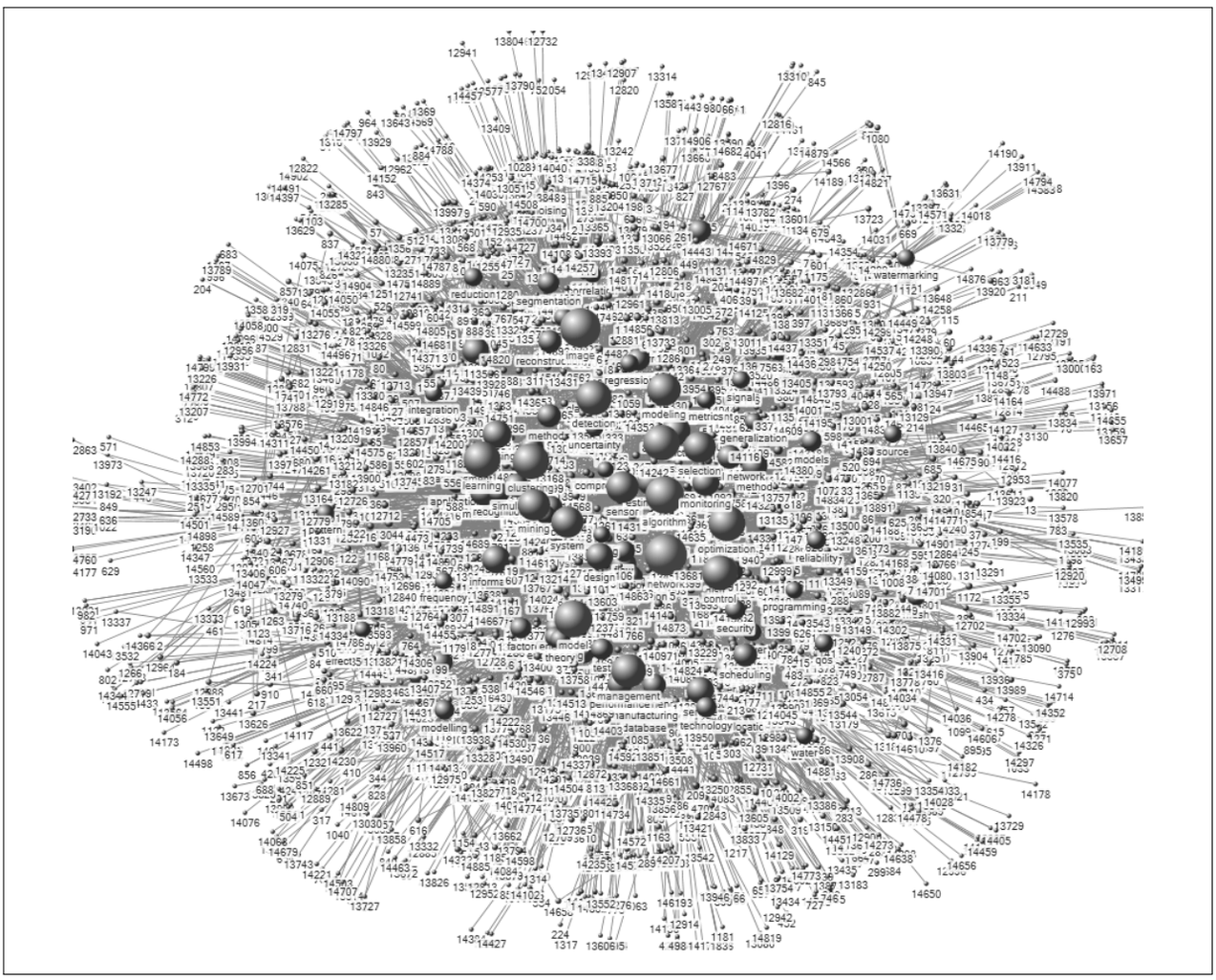

Fig. 5. The results of betweenness centrality analysis(2015년)

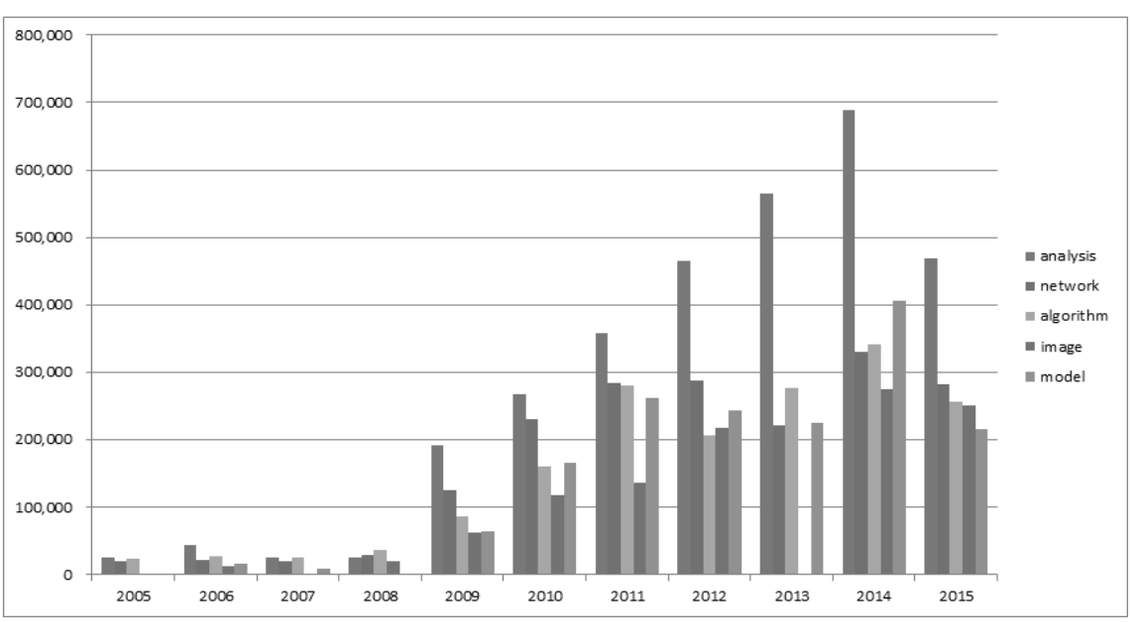

Fig. 6. The results of time series network analysis of betweenness centrality

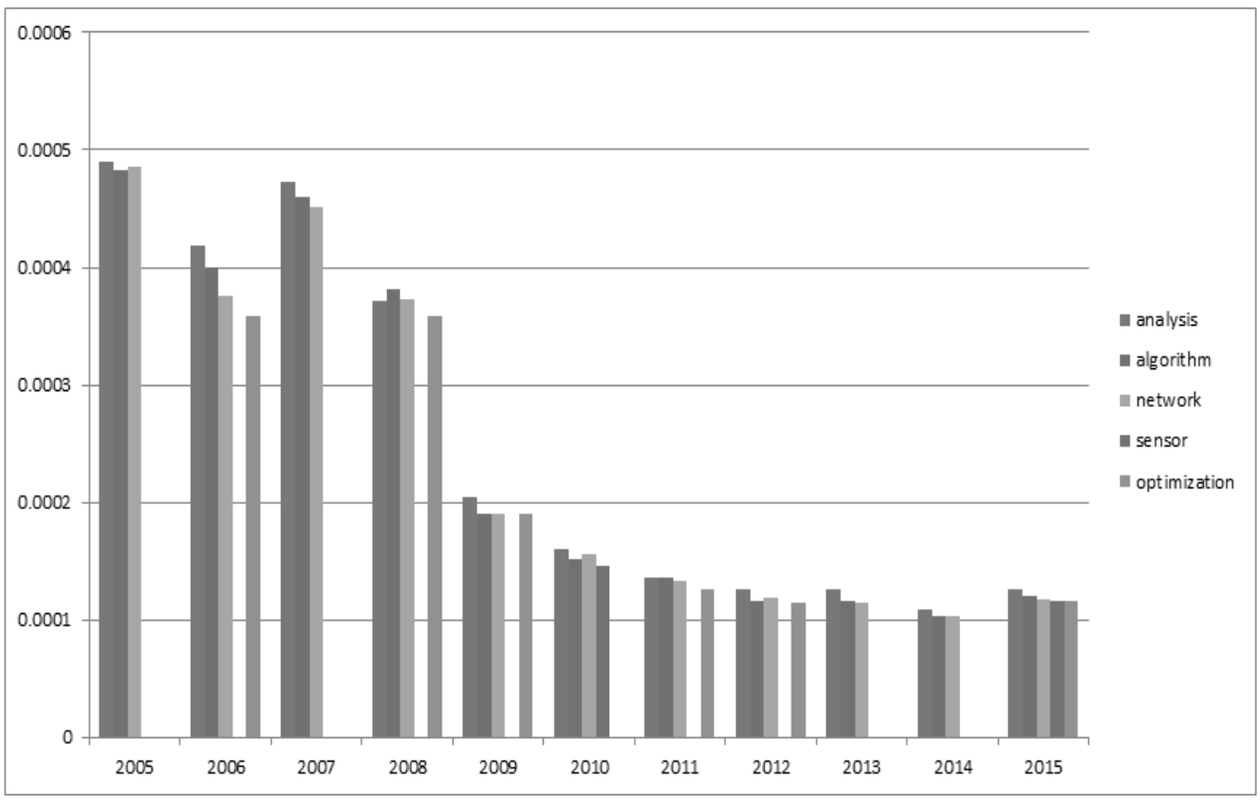

Fig. 7. The results of time series network analysis by closeness centrality

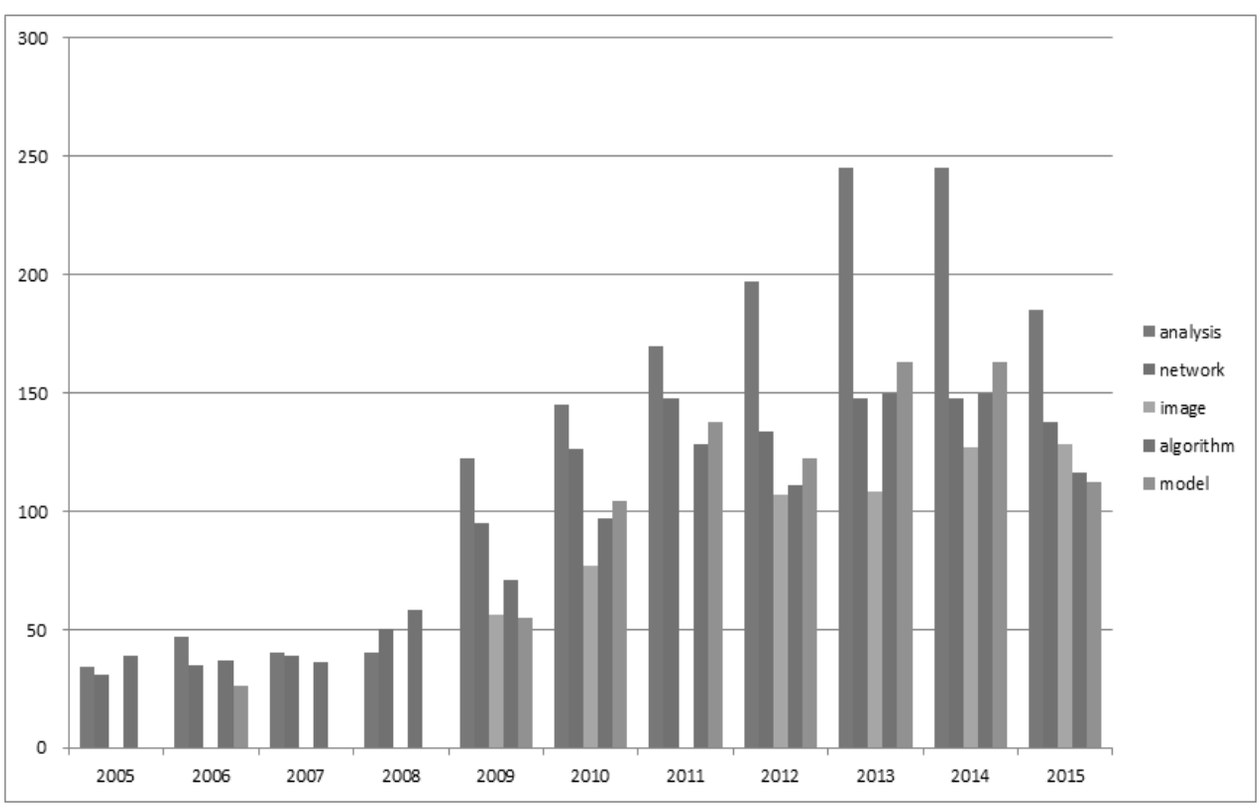

Fig. 8. The results of time series network analysis by degree centrality

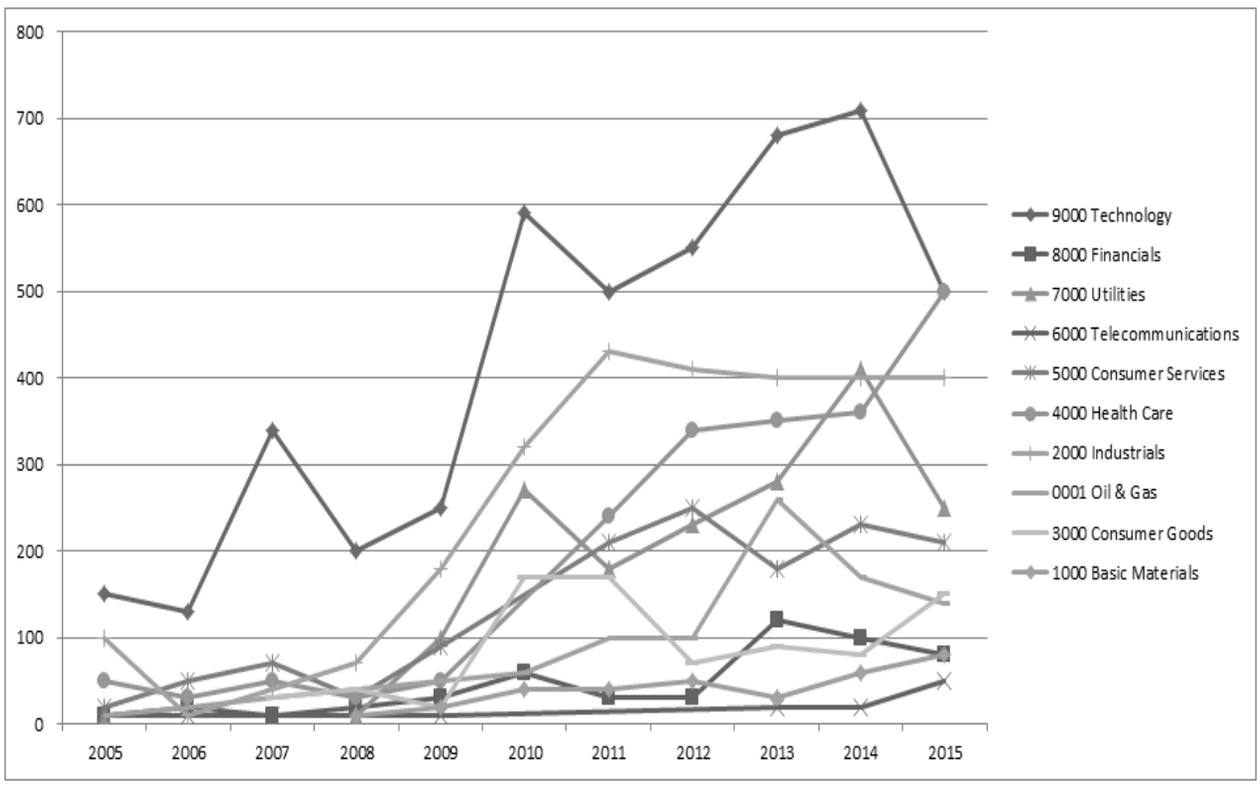

Fig. 9. The results of time series analysis by industry area

또한 중앙성을 기준으로 시계열 분석한 결과 Fig. 6, Fig. 7, Fig. 8과 같이 나타났다. 데이터품질을 주제로한 연구문헌의 사이 중앙성은 Fig. 6과 같이 2005년부터 10여년간 분석(Analysis), 네트워크(network), 알고리즘(algorithm) 순으로 지속적으로 높게 나타났다. 또한 2009년 이후부터 이미지(image)부분의 연구와 모델(model)의 연구가 늘어나고 있었다. 또한 상위 5위 이하에는 제어(Control)과 프로세스(Process)의 연구가 늘어나고 있었다. 이 결과에서 데이터품질 관련한 연구는 분석, 네트워크, 알고리즘, 제어, 프로세스에 관한 융합형 연구가 많이 이루어져 왔고, 새로운 연구주제로는 이미지, 모델에 관한 융합연구의 시도가 증대되고 있는 추세임을 확인할 수 있었다.

데이터품질 연구문헌의 인접 중앙성을 분석한 결과 분석인 Fig. 7이 보여주는 것처럼, 알고리즘, 네트워크 순으로 지속적인 감소세로 나타났다. 또한 상위 5위 이하에서는 시스템과 프로세스 부분도 감소세로 나타났다. 그러나 센서와 최적화 부분은 신규 등장하면서 지속적으로 증가세를 보여줬다. 이 결과를 통해서 분석, 알고리즘, 네트워크, 시스템, 프로세스의 연구가 지속적으로 증대되면서 최신성이 저하되는 주제가 되었으며, 센서와 최적화 부분이 신규 연구 주제로 등장하는 추세임을 확인할 수 있었다.

연결정도를 이용하여 연결정도 중앙성을 분석한 결과, Fig. 8과 같이 분석, 알고리즘, 네트워크가 연결정도가 높게 나타났으며 모델, 프로세스 또한 중앙성이 높은 중요한 주제로 나타났다. 이미지와 모델을 주제로 한 연구의 연결정도가 높아지면서 데이터품질 연구의 중요한 주제로 등장하는 추세를 보였다.

4) 데이터품질의 산업과 연관관계 분석

연구문헌의 분석데이터를 기준으로 산업과 매핑하여 데이터 품질과 산업의 연관관계를 분석하였다. 산업분류체계는 국제적인 분류기준으로 국제산업분류체계인 GICS(Global Industry Classification Standard)를 기준으로 분류하였다. 분류 결과 데이터품질 연구를 통해서 상위 10위로 연관된 산업은 기술(Technology), 산업(Industrials), 건강(Health Care), 유틸리티(Utilities), 고객서비스(Consumer Services), 오일/가스(Oil & Gas), 소비재(Consumer Goods), 금융(Financials), 기반제품(Basic Materials), 이동통신(Telecommunications)의 순으로 Fig. 9와 같이 나타났다. 또한 해당 산업의 데이터품질 연구는 2005년부터 2015년까지 지속적으로 증가하고 있었다.

4. 결론 및 향후 연구

본 연구는 2005년부터 2015년까지 데이터품질에 관한 국·내외 연구문헌을 수집하여 연구경향을 시계열적으로 분석하고 이를 통하여 산업계의 흐름을 예측하고자 하는 목적으로 수행되었다. 데이터품질의 도메인은 그 영역이 방대하여 선행연구 메타분석을 통한 연구가 거의 시도되지 않았으며 또한 다른 도메인에서 진행된 연구경향 분석에서는 데이터양

과 표준화에 따른 신뢰성의 한계점이 존재하였다. 따라서 본 연구에서는 최대의 연구문헌을 보유하고 있는 Web of Science의 메타데이터를 수집하였고, 데이터 표준화의 신뢰성 확보를 위하여 동의어 사전을 활용하여 데이터를 통합하여 분석하였다. 또한 키워드 간 연관관계 분석을 위하여 사회연결망의 중앙성 분석을 수행하고, 이를 학문과 산업부문으로 확대하면서 시계열적으로 연구를 수행하였다.

데이터품질 관련한 연구는 2005년부터 지속적으로 증가하여 2015년 연구문헌 수는 5.3배로 증가되고 있었다. 키워드 빈도 분석결과 저자 키워드 기준으로 network, analysis, model, management 순으로 빈도수가 높게 나타났으며, 제목 키워드에서는 network, analysis, model, image 순으로 높게 나타났다.

연구 도메인 분석결과는 컴퓨터공학과 공학일반에서 높게 나타났으며 2014년부터는 다른 학문분야의 연구비율이 증대되면서 상대적으로 연구비율이 감소되는 추세를 보였다. 2005년에는 정보통신과 운영연구 및 경영과학, 정보과학, 의료정보학, 수학, 지리학 등의 연구 빈도가 높게 나타났고, 10년이 지난 후 2015년의 연구는 환경, 정보통신, 수자원, 운영연구 및 경영 과학, 에너지 등의 연구 빈도가 높게 나타났다. 또한 시계열적인 분석결과 수학 및 전산 생물학, 화학, 건강관리 과학 및 서비스, 생화학 및 분자 생물학, 운영 연구 및 경영 과학, 의료정보학은 감소비율을, 환경, 수자원, 지질학, 계측기 및 계측은 증가비율을 보여줬다.

중앙성 분석을 통하여 연구경향을 파악하고 연구주제 선정에 활용 가능한 결과를 도출하였다. 사이 중앙성을 통해서 데이터품질 관련한 연구는 분석, 네트워크, 알고리즘, 제어, 프로세스에 관한 융합형 연구가 많이 이루어져 왔고, 새로운 연구주제로는 이미지, 모델에 관한 융합연구의 시도가 증대되고 있는 추세임을 확인할 수 있었다. 사이 중앙성이 높은 주제는 상이한 키워드 사이를 연결하는 연구 키워드를 의미하므로 다른 연구주제와 융합 가능한 연구주제 선정에 용이하다고 볼 수 있다. 인접 중앙성 분석결과 분석, 알고리즘, 네트워크, 시스템, 프로세스의 연구가 지속적으로 증대되면서 최신성이 저하되는 주제가 되었으며, 센서와 최적화 부분이 신규 연구 주제로 등장하는 추세임을 확인할 수 있었다. 인접 중앙성이 높은 주제는 중요한 키워드에 인접하여 처음 연구하는 주제로 위험부담이 적은 연구 주제이다. 연결정도를 이용하여 연결정도 중앙성을 분석한 결과, 분석, 알고리즘, 네트워크의 주제가 중앙성이 높은 중요한 주제로 나타났으며, 이미지와 모델 또한 데이터품질에서 중요한 주제로 등장하는 추세를 보였다. 산업분류 트렌드 분석을 통해서 기술, 산업, 건강, 기반, 고객서비스가 데이터품질과 연관성이 높은 산업으로 나타났다.

본 연구는 연구문헌의 메타정보를 활용하여 데이터품질의 거시적 연구 흐름을 시계열적으로 분석하고 데이터품질 연구의 구조를 파악하여 학문 및 산업과 연관관계를 도출하였다. 또한 데이터품질 연구 상태 진단과 연구동향 예측이 가능하고 산업계 흐름을 파악할 수 있다. 이를 통해 연구주제 선정 및 IT정책 방향과 시장의 대응전략 구상에 중요한 참고조점이 되리라 본다. 데이터품질의 중요성을 인식하면서 다양한 분야에서 연구를 진행하고 있었으며, 연구주제 및 학문은 순수학문에서 응용학문으로 시계열적으로 변화되는 추이를 보였다. 본 연구는 선행 연구자료의 연구 키워드를 기반으로 산업분류코드와 매칭하여 산업의 트렌드를 예측하였다.

본 연구에서는 연구 주제키워드와 산업분류코드를 매칭하여 연구와 연관된 산업 트렌드를 확인할 수 있었다. 그러나 연구에서 중요하게 고려되는 키워드가 바로 산업발전과 연관되는지 여부는 찾을 수 없었다. 산업 트렌드의 결과는 연구자들이 해당 시대에 관심을 갖는 산업의 트렌드를 확인할 수 있는 정도였다. 따라서 후속 연구 과제에서는 연구와 산업의 연관관계를 실증적인 연구 작업을 통해서 구체화 할 필요가 있겠다.

References

[1] Thomas C. Redman, “The impact of poor data quality on the typical enterprise,” *Communications of the ACM*, Vol.41, No.2, pp.79-82, 1998.

[2] Web of Science [Internet], http://apps.webofknowledge.com.

[3] Ord. T. J., E. P. Martins, S. Thakur, K. K. Mane, and K. Börner, “Trends in animal behaviour research (1968.2002): ethoinformatics and the mining of library databases,” *Animal Behaviour*, Vol.69, No.6, pp.1399-1413, 2005.

[4] Jaechang Kho, Kuentae Cho, and Yoonho Cho, “A Study on Recent Research Trend in Management of Technology Using Keywords Network Analysis,” *Korea Intelligent Information Systems Society*, Vol.19, No.2, pp.101-123, 2013.

[5] Yung-Keun Kwon, “Understanding of Structural Changes of Keyword Networks in the Computer Engineering Field,” *Korea Information Processing Society*, Vol.2, No.3, pp.187-194, 2013.

[6] Ryan Bullock and Julia Lawler, “Community forestry research in Canada: A bibliometric perspective,” *Forest Policy and Economics*, Vol.59, pp.47-55, 2015.

[7] Hwasuk Cha, “Meta Analysis on Objectives and Types showed on Ceramic Studies: Based on academic journals and dissertations published in 2001-2009,” *Korea Ceramic Studies*, Vol.7, No.1, pp.57-76, 2010.

[8] Hyejung Moon, Seongjhin Choi, and Junho Han, “Trend of Studies Gambling & Lotteries using Big Data Analysis,” 2014.

[9] YungHak Kim, “Social Networking Theory,” Seoul: Parkyoungsa, 2013.

[10] YungHak Kim, “Social Networking Analysis,” Seoul: Parkyoungsa, 2013.

[11] Michael W. Berry, “Survey of text mining,” *Computing Reviews*, Vol.45, No.9, p.548, 2004.

[12] J. D. Kim, T. Ohta, Y. Tateisi, and J. I. Tsujii, “GENIA corpus－a semantically annotated corpus for bio-textmining,” *Bioinformatics*, Vol.19, No.suppl 1, pp.i180-i182, 2003.
[13] Kyoungae Jang, Seong Yong Jang, and Woo-Je Kim, “Project Failure Main Factors Analysis using Text Mining in Audit Evaluation,” *Journal of KIISE*, Vol.42, No.4, pp.468-474, 2015.
[14] Duncan J. Watts and Steven H. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature*, Vol.393, No.6684, pp.440-442, 1998.

장 경 애
e-mail : jkalove@hanmail.net
1996년 대구대학교 문헌정보학과(학사)
2014년 연세대학교 컴퓨터공학과(석사)
2014년～현 재 서울과학기술대학교 IT정책전문대학원 산업정보시스템전공 박사과정
관심분야 : 데이터 품질/분석, 최적화 등

이 광 석
e-mail : kslee@seoultech.ac.kr
1994년 중앙대학교 회계학과(학사)
2001년 텍사스(오스틴)주립대학교 Radio-TV-Film학과(석사)
2008년 텍사스(오스틴)주립대학교 Radio-TV-Film학과(박사)
2011년～현 재 서울과학기술대학교 디지털문화정책학과 교수
관심분야 : 인터넷문화, 모바일노동, 공유경제, 빅데이터와 IT정책

김 우 제
e-mail : wjkim@seoultech.ac.kr
1986년 서울대학교 산업공학과(학사)
1988년 서울대학교 산업공학과(석사)
1994년 서울대학교 산업공학과(박사)
1988년～1991년 동양경제연구소 연구원
1999년～2001년 University of Michigan Visiting Scholar
2003년～현 재 서울과학기술대학교 글로벌융합산업공학과 교수
관심분야 : IT서비스, 소프트웨어 공학, 최적화, 스마트그리드 등