정책자료 2018-08

2018년 소셜 빅데이터 기반 보건복지 이슈 동향 분석

오미애 · 송규원 · 전종준 · 천미경

KIHASA
한국보건사회연구원
Korea Institute for Health and Social Affairs

【책임연구자】
오미애 한국보건사회연구원 연구위원

【주요 저서】
기계학습(Machine Learning)기반 사회보장 빅데이터 분석 및 예측모형 연구
한국보건사회연구원, 2017(공저)
기계학습(Machine Learning)기반 이상 탐지(Anomaly Detection) 기법 연구
-보건사회 분야를 중심으로
한국보건사회연구원, 2018(공저)

【공동연구진】
송규원 한국과학기술연구원 영상미디어연구단 Post-doc
전종준 서울시립대학교 통계학과 교수
천미경 한국보건사회연구원 연구원

정책자료 2018-08

2018년 소셜 빅데이터 기반
보건복지 이슈 동향 분석

발 행 일 2018년 12월
저 자 오 미 애
발 행 인 조 흥 식
발 행 처 한국보건사회연구원
주 소 [30147]세종특별자치시 시청대로 370
세종국책연구단지 사회정책동(1층~5층)
전 화 대표전화: 044)287-8000
홈페이지 http://www.kihasa.re.kr
등 록 1994년 7월 1일(제8-142호)
인 쇄 처 경성문화사

ISBN 978-89-6827-571-5 93510

발간사

보건복지 분야는 공급자 중심에서 수요자 중심의 맞춤형 서비스 체계로 변화하고 있다. 이러한 여건 변화를 빠르게 인지하고 보건복지 분야의 이슈를 파악하기 위해서는 소셜 빅데이터 활용이 중요하다. 빅데이터 분석 환경이 갖추어지고 기하급수적으로 늘어나는 비정형 빅데이터를 수집·분석할 수 있게 되면서 소셜 빅데이터 활용 수요는 증가하고 있다.

비정형 빅데이터는 모바일, 소셜네트워크서비스(SNS), 센서 등의 결합을 통한 새로운 형태의 데이터 생성으로 기존의 정형 데이터로는 파악하기 어려운 변화를 감지하고 정책 욕구를 즉시 확인하고 활용할 수 있는 근거를 제공한다. 비정형 빅데이터는 데이터에서 얼마나 많은 부가가치를 창출할 수 있는가의 관점에서 소셜 빅데이터 분석으로부터 새롭게 얻을 수 있는 지식 또는 부가가치의 양과 차이는 크지 않지만, '왜', '어떻게'에 대한 방향성을 제시해 줄 수 있다는 면에서 가치가 있다.

이 연구에서는 4차 산업혁명 시대의 신기술이라고 할 수 있는 블록체인과 관련하여 보건복지 분야의 이슈를 살펴보고자 하였다. 블록체인은 암호화폐와 함께 이슈화되었지만, 블록체인 기술 자체로 해마다 발전하고 있으며, 민간 및 정부에서의 활용 사례도 증가하고 있다. 여기에서는 보건·복지 분야에서 블록체인 기술이 어떻게 이슈화되고 있는지를 알아보기 위해 블록체인 키워드와 보건 및 복지 키워드 관련 문서를 함께 수집하였다. 그리고 한글 문서에서의 이슈와 영문 문서에서의 이슈의 차이도 살펴보기 위해 한글 문서 및 영문 문서를 각각 분석하여 결과를 제시하였다. 본 연구를 위해 조언을 해 주신 많은 전문가들과 원고 집필에 참여해 주신 교수님들께 감사드린다.

끝으로 본 보고서에 수록된 모든 내용은 우리 연구원의 공식적인 견해가 아니며 연구에 참여한 연구진의 의견임을 밝힌다.

2018년 12월

한국보건사회연구원 원장

조 흥 식

목차

표 목차

그림 목차

Abstract

Social big data trend analysis based on health and welfare issues in 2018

Project Head: Oh, Miae

The health and social welfare sector is changing from a provider-oriented to a consumer-oriented customized service system. It is important to use social big data to readily recognize these changes and to identify issues in the health and welfare sector. The use of social big data is on increasing demand as big data analytics has improved rapidly and an ever-growing amount of unstructured big data is available for collection and analysis.

In this study, we examined health and welfare issues related to blockchain technology, which is widely regarded as an essential feature of the fourth industrial revolution era. The issue of blockchain first came to the fore of social attention in connection with crypto-currency. But the blockchain technology is evolving rapidly and its use is increasing across both the private and public sectors. Here, blockchain keywords and documents related to health and welfare keywords are collected together to see how block-chain technology is being used in health and social welfare. In order to investigate differences in

Co-Researchers: Song, Gyuwon · Jeon, Jongjun · Cheon, Mikyung

priority issues in blockchain technology between Korea and other countries, we collected relevant between issues in Korean documents and issues in English documents, Korean and English documents were analyzed and presented.

The collected documents on blockchain technology were mostly on hthe documents written in Korean were those published mostly in 2018; the documents written in English were mostly from 2017 and 2018, mostly addressing health issues.

In the topic model, it can be judged that the topic of the English document is more appropriately classified than the Korean document.

요약

1. 연구의 배경 및 목적

보건복지 분야는 공급자 중심에서 수요자 중심의 맞춤형 서비스 체계로 변화하고 있다. 이러한 여건의 변화를 빠르게 인지하고 보건복지 분야의 이슈를 파악하기 위해서는 소셜 빅데이터 활용이 중요하다. 빅데이터 분석 환경이 갖추어지고 기하급수적으로 늘어나는 비정형 빅데이터를 수집·분석할 수 있게 되면서 소셜 빅데이터 활용 수요는 증가하고 있다.

이 연구에서는 4차 산업혁명 시대의 신기술이라고 할 수 있는 블록체인과 관련하여 보건복지 분야의 이슈를 살펴보고자 하였다. 블록체인은 암호화폐와 함께 이슈화되었지만, 블록체인 기술 자체로 해마다 발전하고 있으며, 민간 및 정부에서의 활용 사례도 증가하는 상황이다. 여기에서는 보건·복지 분야에서 블록체인 기술이 어떻게 이슈화되고 있는지 알아보기 위해 블록체인 키워드와 보건 및 복지 키워드 관련 문서를 함께 수집하고, 소셜 빅데이터 분석 방법론을 적용하여 다양한 분석결과를 제시하였다.

2. 주요 연구 결과

2장에서는 블록체인의 기본 개념 및 기술적인 요소를 기술하고, 블록체인을 분류하고 블록체인 활용 사례를 살펴보았다.

3장에서는 소셜 빅데이터 분석 방법론으로 텍스트마이닝에 대해 전반적으로 설명하고 단어/문장의 특징 추출 방법, 데이터 분석 방법을 정리하였다.

4장에서는 블록체인과 보건복지 이슈를 함께 살펴볼 수 있는 텍스트 수집, 자료 정제, 시각화 방법을 제시하고, 한글 문서에서의 이슈와 영문 문서에서의 이슈의 차이도 살펴보기 위해 한글 문서 및 영문 문서를 각각 분석하였다.

3. 결론 및 시사점

한글 문서에서는 2015~2017년도에 비해 2018년도에 블록체인과 관련된 문서가 상대적으로 많이 수집되었고, 영문 문서의 경우 2015~2016년도에 비해 2017~2018년도에 블록체인과 관련된 보건 이슈가 증가하였음을 알 수 있다. 이는 해외에 비해 우리나라에서 블록체인과 관련된 상황이 상대적으로 늦게 이슈화된 것으로 해석할 수 있다.

수집된 문서에서 복지 쪽 이슈는 거의 없었으며, 대부분 보건 쪽 이슈와 관련이 있었다.

한글 문서의 5개 토픽을 살펴보면 첫 번째 토픽은 블록체인과 관련된 국내외 투자, 개발에 대한 전망이고, 두 번째 토픽은 블록체인과 관련된 정부의 정책 이슈이다. 세 번째 토픽은 블록체인 기술 기반의 의료 관련 플랫폼 이슈, 네 번째 토픽은 블록체인과 관련한 복지, 기업 이슈, 다섯 번째 토픽은 기타로 분류할 수 있다. 영문 문서의 5개 토픽의 경우 첫 번째 토픽은 블록체인과 관련된 기업이고, 두 번째 토픽은 블록체인과 관련된 의료, 보건 이슈이다. 세 번째 토픽은 블록체인의 클라우드, 사물인터넷(IoT) 기술 이슈, 네 번째 토픽은 블록체인과 관련한 가상화폐 이슈, 다섯 번째 토픽은 블록체인 자체의 기술 이슈로 분류할 수 있다. 토픽 모형 관점에서는 한글 문서보다 영문 문서의 토픽이 더 적절히 분류되었다고 판단할 수 있다.

블록체인은 다른 분야에 비해 공공 분야에서 실행 가능성과 영향력이 꽤 큰 편에 속한다. 이는 과학기술정보통신부가 블록체인의 보안성, 투명성 측면에서 블록체인이 산업과 사회를 혁신하는 기반 기술로 4차 산업혁명의 핵심 산업이 될 가능성이 충분하다고 판단한 것과 관련이 있다.

이 연구의 한계점은 과학기술정보통신부가 '블록체인 기술 발전전략'을 발표(2018. 6. 21.)하기 전에 문서를 수집한 결과라는 것이다. 정부의 블록체인과 관련된 정책이 이슈화된 문서를 수집하여 위의 분석을 한다면 다른 결과가 나올 수 있다.

그럼에도 소셜 빅데이터 분석은 보건복지 정책 영역에서 국가적·사회적으로 관심이 있는 이슈의 현 상황을 파악하는 데 중요한 경쟁력으로 작용할 수 있으며, 앞으로 정책과 관련한 이슈를 도출하고 연구 전략을 세우는 데 근거 자료로 활용될 수 있다. 다양한 소셜 빅데이터 분석 기술을 바탕으로 주요 보건복지 정책에 관한 사회적 관심도, 영향력 등을 분석하고 그 변화 과정을 살펴본다면 시의성 높은 보건복지 정책 연구의 기반을 마련할 수 있을 것이다.

*주요 용어: 소셜 빅데이터, 블록체인, 보건복지, 네트워크 분석, 토픽 모형

제 1 장 서론

1 서론

빅데이터 시대에 이메일, 신문, 블로그 등의 데이터뿐만 아니라 소셜 네트워크 사이트 등에서 문서 데이터가 축적되는 속도는 기하급수적으로 증가하고 있다. 이에 따라 비정형 데이터인 텍스트 데이터를 분석하려는 텍스트마이닝에 대한 연구가 활발히 진행되고 있으며 이는 비정형 데이터 분석의 중요한 분야로 여겨지고 있다.

텍스트마이닝은 데이터마이닝의 한 분야로 비정형 데이터인 텍스트 데이터를 이용하여 특정한 정보를 추출하는 작업을 말한다. 문장과 문서에 대한 정성적 분석의 대안으로 활용할 수 있어 경영학, 경제학, 정책 선정 등 다양한 분야에서 활용되고 있다. 예를 들어 키워드 빈도 분석, 연관검색어 및 문서 추천, 키워드 트렌드 분석, 주제어 분석, 문서 요약 등의 분석 방법은 텍스트마이닝의 한 종류이다.

텍스트마이닝 분석 절차는 일반적인 자료 분석 절차와 유사하다. 텍스트마이닝 분석에서 가장 먼저 해야 할 일은 해결해야 할 문제 혹은 목적으로 하는 문제를 정의하고 분석에 필요한 개념을 정립하는 일이다. 예를 들어 보건 · 복지 분야에서 텍스트마이닝을 이용한 분석을 하겠다고 하면 하위 개념의 세부 주제인 '보건 · 복지 분야에서의 신기술(블록체인) 활용 사례'가 분석 목적이 될 수 있다. 일반적으로 텍스트들의 사용 빈도, 시간에 따른 텍스트 출현의 추세, 연관 키워드 탐색으로 어떤 가설을 입증하거나 새로운 발견을 하려는 것이 텍스트 분석 목적이 된다. 텍스트 수집이 이런 텍스트 분석이 입증하려는 가설에 대한 통계적 증거가 되거나 새로운 사실 발견의 수단이 될 수 있어야 할 것이다. 즉 분석 목적을 구체화

하는 과정에서 그 목적을 달성하기 위한 수단으로 텍스트를 이용한 정보 수집이 적절한지 면밀한 검토가 필요하다. 어떤 경우에는 대용량 텍스트 자료원에서 정보를 수집하는 것보다 과거의 문헌 조사가 적합할 수도 있다.

다음으로 분석 목적에 맞는 자료를 수집한다. 자료 수집 단계는 자료원 선정, 자료 수집, 자료의 특성 검토, 환류 및 자료원 보정 과정을 거친다. 텍스트 자료는 일반적으로 방대한 자료원에서 수집되기는 하지만 자료원의 특성에 크게 의존하는 비동질적 자료의 특성을 가지고 있다. 따라서 잘 설계되지 않은 자료 수집 계획은 수집된 자료의 왜곡을 초래할 수 있으며, 분석 목적에 맞지 않는 자료를 얻게 될 가능성이 커진다. 예를 들어 블록체인과 보건이슈를 살펴보고자 한다면 네이버 블로그에서 '블록체인'과 '건강'과 관련된 단어이지만 '비트코인'과는 관련 없는 문서만 수집하는 것이 자료원 선정의 올바른 예라고 할 수 있다. 그뿐 아니라 자료원 수집에 대한 기술적 제한 사항을 검토하여 자료 수집 비용과 편익을 함께 고려해야 한다.

자료 수집이 끝난 후에는 텍스트를 분석 방법에 맞도록 변환하고 정제해야 한다. 수집된 텍스트의 대부분은 의미 없는 단어이거나 분석에서 배제할 단어인 경우가 많다. 이를 불용어라고 부르는데, 많은 경우 불용어를 정제하고 분석 결과를 도출하면 결과를 해석하는 데 많은 도움이 된다. 자료수집 이후에 텍스트 데이터에 대한 탐색적 자료 분석으로 단어의 분포를 파악하고, 불용어를 신중하게 결정하여 텍스트 데이터를 정제한다. 한편 기초적인 탐색적 자료 분석으로 수집된 텍스트 간의 관련성을 파악하고 분석에 필요한 키워드를 추가, 수집하는 환류 과정을 거쳐 텍스트 데이터를 추가한다.

다음으로 수집·정제된 텍스트를 분석 방법에 맞도록 변환한다. 텍스트

수집 과정에서 이루어지는 탐색적 분석 과정에서 필요한 사항이기도 하다. 일반적으로 텍스트의 양적 분석을 위해 벡터 공간으로 변환하고 수리적 분석 방법을 적용한다. 선택된 양적 분석 방법론으로 수집된 데이터의 분석 결과를 도출하고 해석, 시각화하는 것으로 텍스트마이닝 분석이 이루어진다.

앞서 살펴보았듯이 텍스트마이닝 분석은 전통적인 자료 분석 방법과 절차적으로 유사하다. 반면 자료 수집의 기술적 측면에서 텍스트마이닝 분석은 기존에 사용하지 않았던 자료 수집 및 정제 기술을 요구한다. 먼저 데이터를 수집하기 위한 프로그래밍 도구가 필요하다. 잘 설계된 프로그램은 특정한 문서 도메인에서 자동적으로 텍스트 문서를 수집하고, 필요한 텍스트 부분을 추출한다. 주로 웹 문서를 수집하는 방법이 가장 많이 이용되지만, 경우에 따라서는 이미 수집된 PDF 파일에서의 정보 추출, 인트라넷 문서의 추출도 사용된다. 자료 정제와 관련하여 문서를 단어의 집합으로 바꾸는 자연어 처리 기법과 변경된 단어를 특징으로 추출하는 기술이 필요하다. 문서 내의 단어는 수천 가지에서 수백만 가지의 인자를 가지는 범주형 자료이며, 문장은 이런 단어의 열로 이루어진 집합이다. 즉 텍스트는 수집된 자료의 양에 비해 단어와 문서의 차원의 크기가 매우 큰 고차원 데이터이며, 이런 고차원 자료를 다루기 위해 정보를 효과적으로 저차원 공간으로 요약하고 추출하는 방법이 중요하게 다루어지고 있다.

이 연구에서는 '블록체인'이라는 키워드를 주제로 하고 '보건복지' 분야에서의 이슈를 주제로 텍스트를 수집하는 과정에 대해 시작부터 시각화에 이르기까지의 각 과정을 서술하였다.

블록체인은 서울을 비롯한 지방자치단체별로 국가연구개발과제로 수행될 만큼 4차 산업혁명 시대의 신기술 중 하나이다.

〈표 1-1〉 지역별 블록체인 관련 국가연구개발과제의 주요 키워드 및 연구 분야

순위	한글 키워드[1)]	과학기술표준분류 (소분류 기준)	6T 관련 기술 분야
1	· 비트코인	· 인터넷 S/W	· 정보보안 및 암호 기술
2	· 보안 · 스마트컨트랙트 · 핀테크	· 서비스/응용보안	· 기타 정보처리시스템 및 S/W 기술
3	· 사물인터넷	· 공통 보안 기술	· 고속인터넷 네트워킹 기술 · 기타 정보 기술 · 전자상거래 기술 · 정보 검색 및 DB 기술
4	· 데이터마이닝	· 네트워크시스템 보안 · S/W솔루션	· 기타 네트워크 기술
5	· 분산네트워크 · 암호화 · 암호화폐 · 이더리움 · 인증 · 합의 알고리즘	· 산업보안/융합보안	-

주: 1) 블록체인 제외.
자료: 황혜란(2018), 대전의 블록체인 기술관련 R&D 및 산업육성, 대전세종연구원, 46쪽 〈표 3-15〉 재인용

블록체인 기술은 비트코인과 함께 많이 알려졌지만, 실제로 활용할 수 있는 분야가 많다. 보건복지 분야에서도 신기술의 하나인 블록체인 기술을 활용하여 정책 효과의 효율성을 높일 수 있다는 측면에서 기술 발전 가능성이 큰 블록체인 기술과 보건복지 키워드를 소셜 빅데이터로 분석해 보고자 하였다.

2장에서는 블록체인 기술 및 활용 사례를 살펴보고, 3장에서는 문장에서 텍스트로부터 특징을 추출하는 방법론을 중심으로 문서 기반 분석 방법을 전반적으로 다루고자 한다. 4장에서는 블록체인 키워드와 보건복지 키워드를 연계하여 문서를 수집하고, 소셜 빅데이터를 기반으로 보건·복지 분야에서 활용 가능한 블록체인 기반의 신기술 이슈 동향 분석 결과를 살펴보도록 한다. 마지막으로 본 연구의 한계점과 소셜 빅데이터의 활용 가능성을 논의하고자 한다.

제 2 장

블록체인 기술 및 활용 사례

2 블록체인 기술 및 활용 사례

제1절 블록체인의 이해

최근 4차 산업혁명의 주요 기술 중 하나인 블록체인에 대한 관심이 국내외적으로 급속히 확산되고 있다. 블록체인 기술은 현재의 암호화폐 중심의 금융 분야뿐만 아니라 유통, 물류, 에너지, 공공, 복지, 보건의료 등 다양한 분야에서 활용하려는 시도가 더욱 가속화될 것으로 기대된다. 본 절에서는 블록체인 기술을 소개하고 블록체인의 동작 원리를 알아본다.

1. 블록체인의 기본 개념

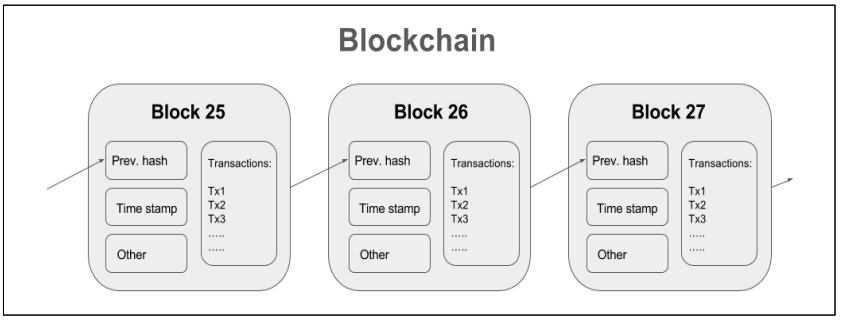
블록체인 기술은 비트코인이라는 분산 컴퓨팅 기반의 전자화폐 시스템(Peer-to-Peer Electronic Cash System)을 구현하기 위하여 사토시 나카모토라는 익명의 개발자가 고안[1]하였다. 해시캐시(Hashcash)와 같은 다양한 선행 기술이 소개되었지만 실제로 사용 가능한 수준까지 도달한 것은 비트코인이 최초라고 할 수 있다. 이후 선구적 웹브라우저 개발자인 마크 앤드리슨은 "비트코인은 정보처리 분야에서 오랜 세월 우리를 괴롭혀 온 비잔티움 장군 문제(Byzantine Generals Problem)를 (Lamport et al., 1982) 해결한 혁신적인 기술"(Marc, 2014. 1. 21.)이라고 평가하는 등 블록체인 기술 자체에 세계적 이목이 집중되고 있으며

1) https://bitcoin.org/bitcoin.pdf, 논문은 2008년 11월에 최초로 게시되었으며, 이를 개발자들이 구현하여 2009년 1월 비트코인 소프트웨어가 처음으로 배포되고 가동을 시작하여 지금까지 끊임없이 이어지고 있다.

다양한 형태로 발전하고 있다.

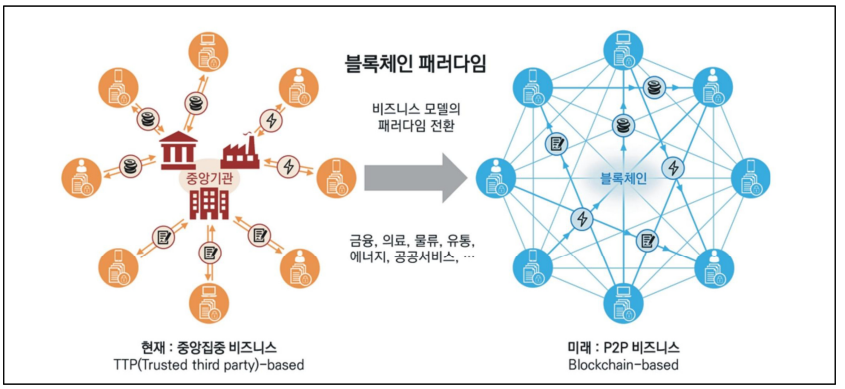
블록체인은 분산원장(distributed ledger) 기술로 거래, 계약 등의 정보가 분산원장에 암호학적 기법으로 연결되어 저장된 데이터 사슬(chain)을 뜻한다. 보다 폭넓게는 데이터를 특정한 중앙 서버 장치 없이 P2P(Peer-to-Peer) 방식으로 체인 형태로 연결(linked list)한 분산 데이터 네트워크 및 그에 수반되는 기술이라고 할 수 있다. 블록체인은 일정한 시간 동안 발생한 거래 내역들이 기록된 하나의 블록을 생성해 모든 구성원에게 전파한다. 전송된 블록의 유효성이 확인될 경우 기존의 블록체인에 연결하는 방식으로 구현된다.

〔그림 2-1〕 블록체인 예시



자료: www.bitcoinforbeginners.io

[그림 2-2] 중앙 집중 처리 방식과 탈중앙화 처리 방식의 비교



자료: 유거송, 김경훈. (2018). 기술동향브리프 블록체인 2018-01호, 한국과학기술기획평가원.

일반적으로 블록체인과 함께 필수적으로 따라다니는 수식어는 '탈중앙화(decentralized)'와 '신뢰성'이라고 할 수 있다. 기존의 중앙 집중화된 거래 시스템은 은행과 같은 중앙기관에 완전히 의존적인 시스템으로 모든 거래의 승인과 거래 결과의 내역 보존 등을 중앙기관이 홀로 책임진다. 그러나 이러한 중앙 집중식 시스템은 시스템 장애 발생 시 전체 시스템을 중단해야 하는 문제(Single Point of Failure, SPoF), 시스템 확장성 문제가 있고 시스템 내외부의 각종 해킹 공격에 대비하기 위하여 막대한 예산을 쏟아부어야 한다. 그럼에도 불구하고 중앙기관의 시스템에 대한 각종 보안 사고가 지속적으로 발생하고 있다. 반면 블록체인은 기존 중앙기관의 역할을 블록체인 네트워크에 참여하는 모든 노드[2]가 나누어 갖는다. 노드들은 합의(consensus)에 따라 거래 승인이 이루어지고 그 결과는 모든 노드들이 함께 보관한다.[3] 이러한 환경에서는 블록체인 네트워크의 대다수의 노드가 동시에 정지하지 않는 이상 해당 서비스는 영

2) 인터넷에 연결된 컴퓨팅 장치를 의미함.
3) 이러한 처리 방식을 분산 컴퓨팅(distributed computing)이라고 하며 특히 Peer-to-Peer(P2P) 컴퓨팅이라고 함.

구적으로 지속 가능하다. 또한 모든 거래 내역을 모든 노드들이 동일하게 유지함으로써 정보 위변조가 불가능에 가까우며, 거래 내역을 투명하게 공유함으로써 참여자 간의 신뢰를 확보할 수 있다. 다시 말해 블록체인에 한 번 기록된 내용은 체인 형태로 연결 및 분산 저장되어 누구도 임의로 수정할 수 없는 강력한 데이터 위변조 방지 기술을 제공할 수 있음을 의미한다. 이는 데이터 유출을 방지하는 정보 보안 기술의 개념이라기보다는 네트워크에 참여하는 모든 참여자가 데이터를 공유하여 데이터 위변조를 방지한다는 개념[4]이라고 할 수 있다.

이해를 높이기 위하여 학생들 간의 간단한 '돈 빌리기' 상황을 가정해 볼 수 있다. 초등학교에 다니는 철수는 영희에게 1만 원을 빌리고자 한다. 영희는 철수를 완전히 신뢰할 수 없어 선뜻 돈을 빌려주기가 불안하다. 따라서 영희는 철수에게 돈을 빌려주면서 자신의 노트에 'a월 b일 철수에게 1만 원을 빌려줌, c월 d일 갚기로 약속함'이라고 기록해 두었다. 그러나 약속한 시일이 되었지만 철수는 돈을 마련하지 못해 나쁜 마음을 먹기로 하였다. 영희의 노트를 몰래 훔쳐서 해당 기록을 없애 버리거나, 오히려 자신이 돈을 빌려주었다고 교묘히 수정한 후 적반하장으로 영희에게 돈을 요구할 수도 있다. 이러한 황당한 상황을 미연에 방지하기 위하여 영희는 새로운 방법을 사용할 수 있다. 즉 영희가 철수에게 돈을 빌려주었다는 사실을 자신의 노트에만 기록하는 것이 아니라, 영희가 다니는 학교의 전체 학생에게 해당 사실을 알리면서 각자의 노트에 모두 기록하도록 하는 것이다. 이에 더해 기록한 내용을 추후에 철수가 수정할 수 없도록 해당 기록을 복사하여 노트의 다음 쪽에 붙여 준다. 거래 내용의 기록과 보존에 대한 대가로 영희는 학생들에게 사탕을 제공할 수도 있다. 이

4) 미리 데이터를 암호화한 뒤 블록체인에 저장함으로써 내용을 참여자에게 공개하지 않도록 구성할 수도 있음.

러한 상황에서 철수가 증거를 인멸할 수 있는 방법은 전교생의 노트를 일시에 모두 폐기하는 방법 외에는 없다고 할 수 있다. 물론 일부 학생이 둘의 채무관계가 청산되기 전에 자신의 노트를 들고 이사를 가거나, 철수와 동조하여 악의적으로 노트 내용을 조작해 놓을 수 있다. 그러나 과반수의 노트가 온전히 보전되어 있다면 학생들은 '영희가 철수에게 1만 원을 빌려주었다'는 것을 사실로 인정할 수밖에 없다.

이처럼 서로를 신뢰하기 힘든 상황에서 저장하려는 정보를 참여자에게 모두 알리고 그것을 안전하게 보관하는 방법이 블록체인 기술의 개념이다. 위의 예에서 등장한 노트는 거래 내용을 기록할 수 있는 원장에 해당하며, 전교생이 해당 내용을 동일하게 기록하고 안전하게 보관하는 것은 분산원장 기술이다. 또한 기록한 내용의 위변조를 막기 위해 복사를 해두고 변경 사항을 쉽게 알아차릴 수 있도록 하는 것은 암호기법의 하나인 해시 함수(hash function) 기술이다. 이와 더불어 전자 서명 기술, P2P 네트워크 기술 등이 블록체인의 기반 기술을 이루는데, 이는 다음 장에서 보다 자세히 설명하도록 한다.

2. 블록체인의 동작 원리

앞에서 기술한 블록체인의 개념적 이해에 기초하여, 블록체인의 동작 원리에 보다 기술적으로 접근하도록 한다. 본 절에서는 다른 언급이 없는 한 비트코인에 기반을 두어 설명하도록 한다. 다른 블록체인의 경우 비트코인을 개선 및 확장한 것으로 핵심이 되는 기술은 모두 유사하다고 볼 수 있기 때문이다.

가. 블록체인의 기술 구성 요소

블록체인은 P2P 네트워크를 기반으로 하고 해시 함수, 전자 서명, 합의 알고리즘 등의 기술 조합으로 구성된다.

1) P2P 네트워크

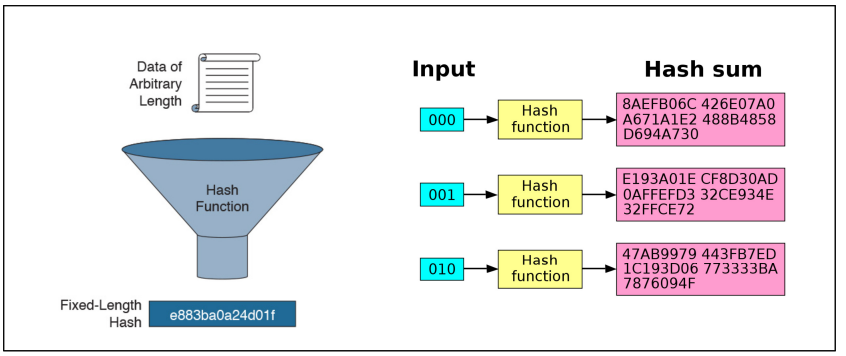
P2P 네트워크는 블록체인이 탈중앙화된 형태로 서비스를 제공하기 위한 일련의 플랫폼 기술이다. 네트워크에 참여하는 모든 노드가 서버와 클라이언트의 역할을 동시에 수행하는 컴퓨팅 형태를 말한다. 과반수의 노드가 한꺼번에 고장 나지 않는 이상 끊임없는 서비스 제공이 가능하나, 서버/클라이언트 컴퓨팅 구조에 비해 시스템 설계 및 운용이 복잡하다는 단점이 있다. 간단히 P2P 네트워크는 블록체인을 구동하는 분산된 컴퓨터 네트워크 정도로 이해하면 된다.

2) 해시 함수

해시 함수는 블록체인 기반 기술에서 매우 중요한 부분을 차지한다. 해시 함수는 임의의 길이의 데이터를 정해진 계산 과정을 거쳐 고정된 길이의 데이터로 매핑하는 함수이다. 해시 함수의 결과값은 해시 코드, 해시 체크섬 또는 간단하게 해시라고 한다. 해시 함수는 암호학적으로 자주 활용되며 SHA(Secure Hash Algorithms)라고 불린다. 만약 해싱 결과 값만 알고 있다면 원래 입력값을 알아내기 힘들다는 사실에 따라 인터넷 사이트의 로그인 시스템에 사용될 수 있다. 또한 전송된 데이터의 무결성(integrity)을 확인해 주는 데 사용되기도 하는데, 이는 해시 함수의 입력

데이터 중 한 비트만 바뀌어도 해시 결과 값이 크게 달라지는 특징을 활용한다. 이러한 성질을 바탕으로 블록체인은 거래 내역이 담긴 블록의 정보가 위변조되었는지 쉽게 확인하기 위하여 해시 함수를 활용한다. 이에 더하여 해시 함수는 한 방향 계산은 쉬우나 역으로 계산하기는 매우 어려운 계산식으로 새로운 블록을 생성하는 규칙에도 활용된다. 이와 관련된 내용은 합의 알고리즘으로 다음 절에서 다루도록 한다.

[그림 2-3] 해시 함수의 개념



자료: www.wikipedia.org

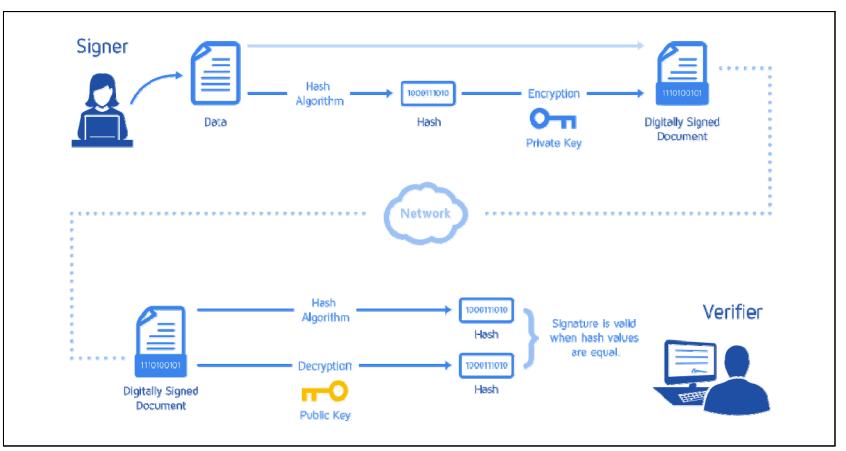
3) 전자 서명

블록체인에서 전자 서명은 일반적으로 공개 키 기반 구조(Public Key Infrastructure)를 바탕으로 네트워크를 통해 전송하려는 정보의 무결성 검증, 서명자의 부인 방지(non-repudiation)를 위한 장치로 사용된다.

전자 서명은 보통 세 단계(비대칭 키 생성, 서명 생성, 서명 검증)로 구성된다. 전자 서명을 하려는 자는 비대칭 키를 미리 발급받아야 한다. 발급 과정에서 공개 키와 비밀 키 세트를 받게 되며, 공개 키[5]는 다른 사람에게 공개되어도 무관하지만 비밀 키[6]는 절대로 외부에 공개되어서는 안

된다. 전송할 데이터에 대한 서명이 생성되는 과정은 다음과 같다. 전송할 데이터에 대해 우선 해시 함수를 통해 해시값을 생성한다. 그리고 생성된 해시값을 서명자의 비밀 키로 암호화한다. 이 과정에서 전자 서명(digital signature)이 생성된다. 전송하려는 데이터의 원본과 생성된 전자 서명을 함께 네트워크로 전송한다. 전송받은 전자 서명을 검증하기 위해서 검증 노드는 우선 원본 데이터를 해시 함수에 적용하여 해시값을 얻는다. 또한 전송받은 전자 서명을 서명자의 공개 키로 복호화한다. 계산한 해시값과 복호화된 결과값을 비교하여 두 값이 서로 일치하면 검증에 성공하며 그 외의 경우는 모두 실패하게 된다.

[그림 2-4] 전자 서명의 동작 예시



자료: www.medium.com

5) 블록체인에서는 주로 타원 곡선 암호(ECDSA)를 사용하여 비대칭 키를 생성하고 공개 키를 가공하여 지갑(은행의 계좌번호 개념)의 주소로 사용함.
6) 만약 비밀 키를 잃어버리게 되면 그와 연관된 지갑에 포함된 모든 자산을 잃게 됨.

블록체인에서는 이러한 전자 서명 과정이 거래(transaction)마다 부여된다. 블록체인 네트워크 참여자는 모든 거래 기록을 순차적으로 검증할 수 있다. 이를 통하여 정당한 서명자가 제대로 거래를 생성했는지와 해커에 의해 어느 거래 내용이 위변조되었는지를 쉽게 확인할 수 있다.

4) 합의(consensus) 알고리즘

블록체인은 중앙기관과 같은 하나의 컨트롤타워 없이 탈중앙화된 형태로 운영되며, 블록체인 네트워크에 참여하는 모든 노드들이 동일한 역할을 한다. 이때 거래의 유효성 검증, 새로운 블록의 생성과 연결, 그리고 생성된 블록체인을 전체 노드가 동일하게 유지하기 위해 필요한 것이 바로 합의 알고리즘이다. 기본적으로 블록체인 네트워크, 즉 P2P 네트워크에 참여하는 노드는 지리적으로 전 세계 분산되어 있으며 다양한 기종의 컴퓨팅 머신으로 구성될 수 있다. 이러한 환경에서는 생성된 거래 내역이 전체 노드에 전달되기까지 소요되는 시간이 각기 다를 수 있으며, 심지어 해당 정보를 수신하지 못하는 노드가 발생할 수도 있다. 따라서 악의적 목적의 데이터 위변조뿐만 아니라 중복된 거래 내역 수신이나 미수신된 정보로 인한 오작동의 가능성을 피하는 것이 합의 알고리즘의 목적이라 할 수 있다.

비트코인은 작업 증명(Proof-of-Work, PoW)이라는 합의 알고리즘을 사용하여 최초의 순수 P2P 네트워크 기반의 전자 지불 시스템을 실현하였다. 작업 증명은 비트코인을 시작으로 많은 블록체인 진영에서 채택하고 있는 합의 알고리즘이다. 작업 증명은 막대한 전기 자원을 낭비한다는 비판을 받고 있으며, 이를 개선하기 위하여 지분 증명(Proof-of-Stake, PoS), PBFT(Practical Byzantine Fault

Tolerance) 알고리즘 등 다양한 합의 알고리즘이 등장하고 있다.

〈표 2-1〉 대표적인 합의 알고리즘과 채택 시스템

구분	채택 시스템
Proof-of-Work	Bitcoin, Ethereum 등
Proof-of-Stake	Ethereum(예정), EOS 등
PBFT	Hyperledger Fabric 등
Paxos	Google Chubby 등
Raft	IPFS, RAMCloud 등

자료: 아카히네 요시하루, 아이케이 마나부 지음, 양현 옮김. (2017). 블록체인 구조와 이론. 위키북스를 바탕으로 내용 재구성

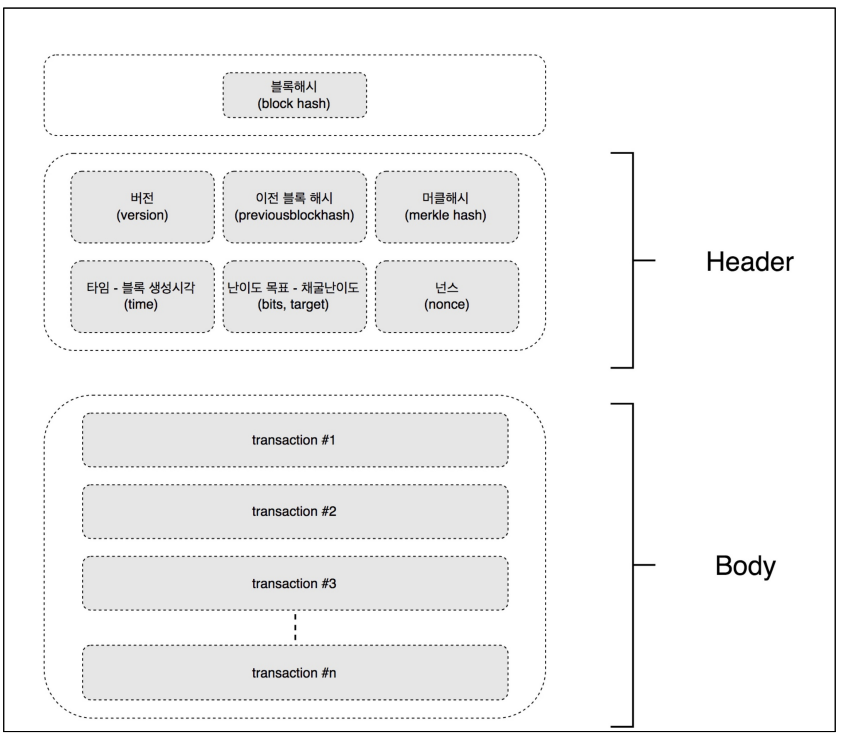
나. 블록체인의 동작

블록체인을 구성하는 기반 기술을 활용하여 실제로 블록체인이 어떻게 동작하는지 그 원리를 설명한다.

1) 블록의 구성

일반적으로 블록은 블록의 메타정보를 수록하는 블록 해시, 블록 헤더, 그리고 거래 내역을 수록하는 블록 보디로 구성된다. 블록 해시는 각 블록을 구분하는 고유의 식별자로 해당 블록 헤더에 포함된 모든 값을 해시 함수의 입력값으로 사용하여 나온 결과값이다. *i*-블록의 블록 해시값은 *i+1*블록의 블록 헤더에 포함된다. 이렇게 블록 해시값으로 연결되어 체인 형태를 갖추게 된다.

〔그림 2-5〕 블록의 구조



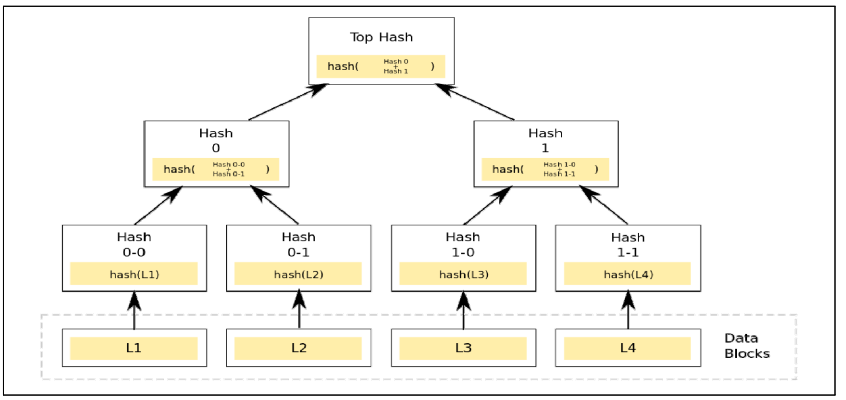
자료: brownbears.tistory.com

블록 헤더는 버전, 이전 블록 해시, 머클 해시, 타임스탬프, 난이도 목표, 그리고 넌스(nonce)값으로 구성된다. 버전은 말 그대로 현재 블록체인의 합의 알고리즘(소프트웨어) 버전을 의미한다. 이전 블록 해시는 바로 직전 블록의 식별자(블록 해시) 정보를 의미하며, 머클 해시는 해당 블록에 포함된 모든 거래 내역을 머클 트리의 형태로 구성한 뒤 그 트리의 최상단에 위치하는 해시값을 말한다. 머클 트리는 블록 보디 부분에서 다시 설명한다. 타임스탬프는 해당 블록이 생성된 시각, 난이도 목표는 블록의 생성 시간을 일정하게 유지하기 위한 수치이며 해당 목표를 만족시키는 값이 넌스에 기록된다. 난이도 목표와 넌스는 블록 생성(채굴) 과정

에서 자세히 다루도록 한다.

블록의 보디에는 거래 내역이 저장되며, 각 거래 내역은 머클 트리(merkle tree) 형태로 구성된다. 거래 내역들을 이진 트리(binary tree)의 가장 아래에 위치시키고 두 개씩 쌍을 이뤄 해시값을 상위 노드[7]에 기록한다. 이 과정을 반복하여 최상위 노드의 결과값이 머클 해시값으로 기록된다. 머클 해시값으로 단일 블록에 존재하는 모든 거래의 무결성을 검증할 수 있다. 또 이 값이 블록 해시값을 생성하는 데 입력으로 사용되며 그 결과값은 다음 블록에 기록됨으로써 검증이 완료된 거래[8]의 변경이 사실상 불가능하게 되는 것이다. 즉 노드 검증이 완료된 특정 거래 내역을 위변조하려면 그 거래 내역이 포함된 특정 블록부터 그 이후의 모든 블록을 혼자서 수정 및 생성해야 하는데 이는 전 세계의 슈퍼컴퓨터를 계속해서 동원해야 할 만큼 어렵다고 볼 수 있다.

[그림 2-6] 머클 트리(merkle tree) 예시



자료: www.wikipedia.org

7) P2P 네트워크에서 노드는 컴퓨터를 의미하나, 트리 자료 구조에서는 데이터가 저장되는 하나의 공간을 의미함.

8) 거래 내역이 블록에 기록되고 그 블록이 기존의 블록체인으로 연결된 후 이어서 일정한 수 이상의 블록이 더 연결된 상태를 의미함.

2) 블록의 생성

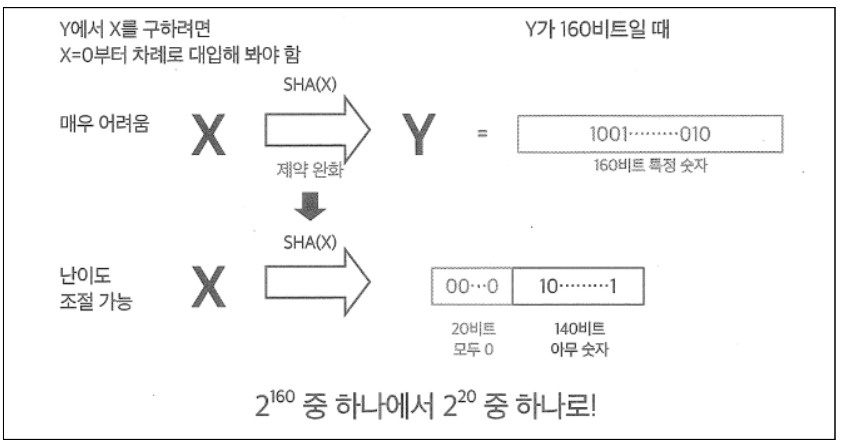
블록 생성 과정은 마치 금광에서 금을 캐는 과정과 유사하다고 하여 채굴(mining)이라고 일컫는다. 블록체인에 참여하는 노드는 어떤 노드가 생성하여 전파한 블록을 전송받아서 그 블록에 포함된 거래 내역의 유효성을 검증한다. 이때 '어떤 노드가 새로운 블록을 생성하여 전파'하는 과정을 채굴이라 한다. 블록 생성은 정해진 합의 알고리즘에 따라 진행된다. 가장 대표적인 합의 알고리즘인 작업 증명(PoW)에 따른 블록 생성 방법을 설명하도록 한다.

누군가 새로운 거래를 생성(예시: 비트코인을 다른 사람에게 지불)했다면 이를 블록체인에 기록하기 위해 그 거래 내역을 모든 노드에 전파하게 된다. 각 노드는 아직 블록체인에 포함되지 않은 거래 내역들을 모아 두었다가 자기 자신만의 새로운 블록(정확히는 블록 보디)에 포함시킨다. 새로운 블록에 일정량의 거래 기록들이 모이면 위변조를 막기 위한 머클 해시값을 계산하여 블록 헤더에 기록한다. 노드가 현재 가지고 있는 블록체인의 마지막 블록의 블록 해시값을 새로운 블록의 블록 헤더에 기록함으로써 기존의 블록체인과 연결한다. 새로운 블록의 헤더의 넌스값에 여러 가지 값을 무작위로 대입해 보면서 난이도 목표값(발견해야 할 숫자 비트의 연속된 0의 개수)을 만족시키는 해시값을 계속해서 계산[9]한다. 만약 다른 노드들보다 먼저 해당 넌스값을 찾았다면 새로운 블록을 제일 먼저 생성한 것이며, 이렇게 새로 만든 블록을 다른 노드로부터 검증받기 위하여 전체 노드에 채굴한 블록을 전파한다. 새로 발견된 블록을 전송받은 다른 노드들은 해당 블록의 넌스값을 검증한 후 유효하다면 자신의 블록체인에 저장하고 유효하지 않다면 무시한다. 이렇게 일정 시간마다 새

9) 마치 원하는 숫자의 조합이 나올 때까지 주사위를 반복해서 던지는 과정과 유사함.

로운 블록이 생성되고 검증을 거친 후 전체 블록체인 네트워크로 퍼져 나간다. 새로운 블록을 생성하는 데 비트코인의 경우 평균 10분 정도가 소요되며, 이를 일정하게 유지하기 위하여 난이도 목표값이 일정 주기마다 적절히 수정된다. 따라서 비트코인의 경우 하나의 거래가 검증받기 위해서는 새로운 블록에 포함되어야 하므로 최소 10분 이상이 소요되며, 그것이 전체 노드에 완전히 인정받기 위해서는 그 블록 이후 최소 5~6개 이상의 블록이 추가되어야 한다. 즉 약 1시간 정도가 필요하다.[10)]

[그림 2-7] 채굴 과정에서 넌스(nonce)값을 찾아내는 과정



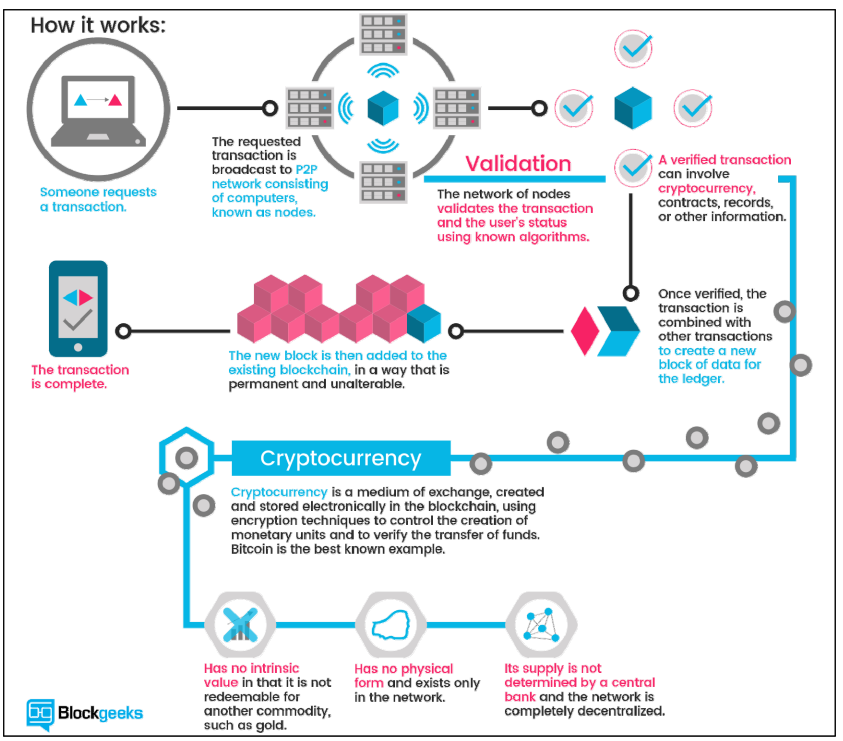
주: 위의 그림은 난이도 목표를 조절함으로써 넌스값을 발견할 확률이 $1/2^{160}$에서 $1/2^{20}$으로 대폭 상승하는 것을 보여 줌.
자료: 김석원 지음. (2017). 블록체인 펼쳐보기(4차 산업혁명을 이끌 또 하나의 기술), BJ퍼블릭

블록 생성 과정이 금광에서 금을 캐는 과정과 유사하여 이를 채굴이라고 표현하였다. 금광에서 채굴에 성공하면 금을 얻게 되는 것처럼 블록체

10) 이러한 비트코인의 긴 블록 생성 시간으로 인해 화폐의 수단으로 사용하기에는 부적합하다는 비판을 받고 있으며, 이를 수초 이내로 개선하려는 라이트닝 네트워크(Lightning Network) 같은 기술이 개발되고 있음.

인에서도 채굴에 성공하면 채굴 보상금을 받는다. 은행에서 새 화폐를 발행하는 것처럼 비트코인에서는 채굴의 대가로 새 비트코인을 만들어서 보상금으로 준다. 이에 더하여 새로운 블록에 포함된 거래들의 수수료[11]를 채굴자가 받게 된다.

[그림 2-8] 블록체인의 새로운 블록 생성(채굴) 과정 개념도



자료: blockgeeks.com

11) 비트코인에서 거래를 할 때, 즉 코인을 다른 사람에게 보낼 때는 수수료를 지불해야 함. 채굴 노드는 많은 수수료를 받기 위해 높은 수수료를 제공하는 거래를 우선적으로 처리하려고 함. 따라서 거래가 순간적으로 폭증한 상황에서는 낮은 수수료를 제시한 거래의 처리가 매우 늦어질 수 있음.

3) 블록체인의 관리

앞서 기존의 블록체인에 새로 생성된 블록을 연결하는 방법을 알아보았다. 새로운 블록을 빨리 생성하기 위해 여러 노드들이 경쟁한다고 하였는데, 만약 두 노드가 동시에 넌스값을 찾아냈다면 어떻게 처리해야 할 것인가 하는 문제가 남아 있다.[12)]

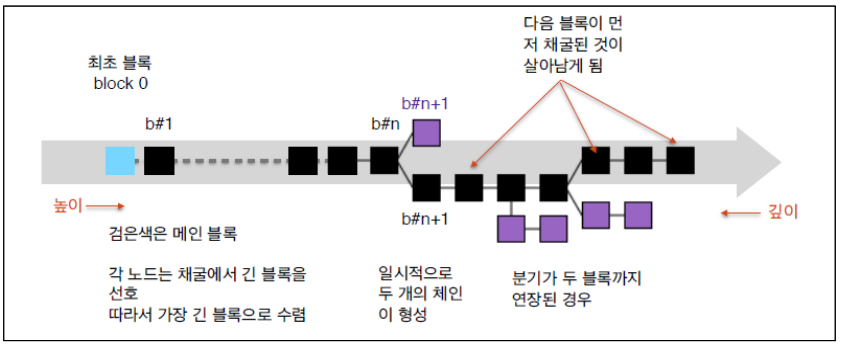
다행히 이러한 상황 또한 합의 알고리즘에 반영되어 있다. 이러한 상황을 블록체인의 분기라고 하며, 이는 긴 블록체인 선호 정책에 따라 해결한다. 즉 만약 동시에 새로운 블록이 생성되었다면 일시적으로 블록체인은 두 가닥으로 분기된다. 이후 노드들은 분기된 두 마지막 블록 중 하나의 블록을 선택하여 자신의 새로운 블록을 생성하여 연결한다. 이렇게 되면 두 분기의 길이가 달라지기 시작한다. 다음에 생성된 블록은 정책에 따라 길이가 더 긴 분기에 연결된다. 이러한 과정이 반복되면 결국에는 분기가 사라지고 가장 긴 분기를 기준으로 하나의 블록체인이 존재하게 된다. 노드가 길이가 긴 분기를 선택할 수밖에 없는 것은 만약 선택한 분기가 더 이상 연결되지 못하면 해당 분기는 모두 무효화되고 채굴에 대한 보상을 받을 수 없기 때문이다.

무효화된 블록에 포함된 거래 내역은 노드들이 다시 수집하여 새로운 블록 생성 시 다시 처리하게 된다. 이때 경쟁하던 두 분기의 블록에 모두 들어 있던 거래 내역은 그대로 두고 무효화된 분기의 블록에 있으면서 승리한 분기의 블록에는 없는 거래 내역만 다음번 채굴을 기다리게 된다.[13)]

---

12) 아주 낮은 확률이지만 이러한 상황이 비트코인 네트워크에서 실제로 발생하고 있음.
13) 비트코인의 경우 보통 5~6개의 블록이 추가로 연결된 경우 분기가 무효화되는 경우가 거의 발생하지 않게 됨. 따라서 거래 승인이 완전히 이뤄지는 데 약 한 시간이 소요됨.

[그림 2-9] 긴 블록체인 선호 정책의 개념



자료: 김석원(2016. 1. 20.) 비트코인의 기반 기술 블록체인의 원리. 소프트웨어정책연구소(SPRi)

제2절 블록체인의 분류 및 활용

앞서 살펴본 것처럼 블록체인 기술은 비트코인과 같은 암호화폐를 중심으로 발전해 왔다. 그러나 이더리움(Ethereum)의 등장과 함께 블록체인상에 탈중앙화된 응용 소프트웨어를 구현하여 서비스할 수 있는 스마트 컨트랙트(Smart Contract)가 도입됨으로써 다양한 분야로의 응용이 급속히 확산되고 있다. 이 장에서는 블록체인 기술을 활용한 몇 가지 사례를 알아본다. 또한 블록체인의 확산을 가로막고 있는 기술적 한계점을 논의한다.

1. 블록체인의 분류

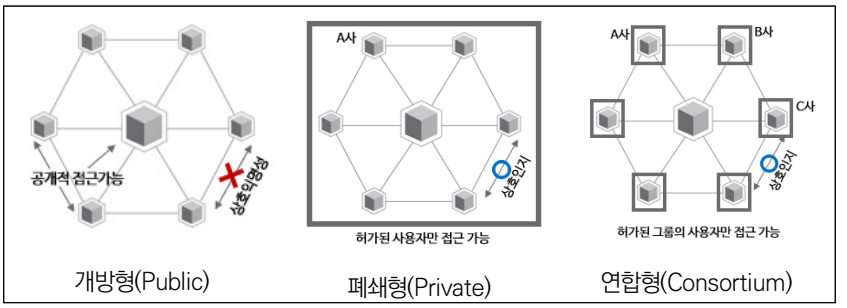
본격적으로 블록체인의 응용 분야를 알아보기 전에 블록체인의 용도나 블록체인이 적용되는 네트워크의 운영 형태에 따른 구분을 할 필요가 있다. 블록체인은 네트워크 참여에 사전 승인이 필요한지에 따라 크게 허가

형과 비허가형 블록체인으로 구분할 수 있다. 좀 더 세분화한다면 개방형(Public), 폐쇄형(Private) 혹은 연합형(Consortium)으로 나눌 수 있다. 허가형 블록체인에는 폐쇄형과 연합형이, 비허가형에는 개방형이 속한다.

비허가형 블록체인은 해당 네트워크에 참여하는 데 아무런 제한이 없어 누구나 참여할 수 있다. 비허가형 블록체인에 참여하는 노드는 합의 과정에 기여한 보상으로 암호화폐를 받는다. 비허가형 블록체인에 이러한 인센티브가 없다면 참여 노드가 사라지게 되고 결국 운영이 중단될 수밖에 없으므로 비허가형 블록체인과 암호화폐는 분리할 수 없는 개념이라고 할 수 있다. 비허가형 블록체인은 탈중앙화된 형태로 운영되므로 정보의 투명성과 무결성이 보장되나 다른 형태에 비해 거래 처리 속도가 느리다. 비트코인이 대표적인 비허가형 블록체인에 해당한다.

허가형 블록체인은 네트워크에 참여하려면 운영 기관에서 사전 승인을 받아야 하며 참여자 개개인을 지정하는 폐쇄형 블록체인(주로 하나의 기업 혹은 단체가 참여)과 동일한 목적으로 생성된 연합 조직 내에서 합의에 따라 권한을 가지는 연합형 블록체인(주로 다수의 기업 혹은 단체가 참여)으로 분류된다. 허가형 블록체인의 경우 네트워크 참여자를 제한하고 검증 과정을 단순화하여 성능을 극대화하고, 데이터 통제권을 유지하며, 기밀 정보 유출을 방지한다. 비록 허가형 블록체인은 '탈중앙화'라는 가치를 일부 포기해야 하지만, 처리 속도와 정보 보안을 중요시하는 기업을 중심으로 개발이 진행되고 있으며, 정부의 경우 권한을 부여받은 기관만이 데이터 통제 권한을 갖는 허가형 블록체인이 보다 적합한 형태라고 할 수 있다.

〔그림 2-10〕 블록체인 유형별 구성 예시



자료: 박종대 (2018). 블록체인 기술 전망. TePRI 2018-83호, 한국과학기술연구원.

〈표 2-2〉 블록체인 분류표

유형	구분	참여 노드 (정보 접근)	채굴 보상	속도	예시
비허가형	개방형 (Public)	제한 없음 (완전 개방)	보상 필요 (암호 화폐)	느림	비트코인, 이더리움 등
허가형	폐쇄형 (Private)	사전 승인 (제한된 접근)	임의 선택	빠름	링크 등
	연합형 (Consortium)	사전 승인 (제한된 접근)	임의 선택	빠름	하이퍼레저 패브릭, 이더리움 EEA 등

2. 블록체인의 활용

대부분의 나라에서 블록체인 규제를 완화하고, 공공 및 민간 영역에 블록체인 연구 및 개발을 장려하고 있다.

유럽연합(EU)은 '허라이즌(Horizon)2020'으로 다양한 블록체인 관련 프로젝트를 지원해 왔으며 2020년까지 최대 3억 4000만 유로를 블록체인 기술 프로젝트에 지원할 계획이다(정성교, 2018 재인용). 또한 2018년 2월 블록체인 정보와 전문성을 종합적으로 이해하고 블록체인의 도입과 나아갈 방향을 모색하기 위해 'EU Blockchain Obervatory and Forum'을 발족했으며 같은 해 4월 영국, 프랑스, 네덜란드, 독일, 스페

인, 노르웨이를 포함한 22개국이 블록체인 기술을 통한 정보 교환을 목적으로 'European Blockchain Partnershop'을 발표하였다. Boucher 등(2017)은 블록체인의 가장 큰 특징인 탈중앙화와 투명성으로 인해 활용 분야가 금융 분야를 넘어 다양한 분야로 확대될 것으로 예상하였으며, 대표적인 사례로 온라인 전자 투표, 공문서의 진위 판별, 제품(다이아몬드, 샴페인 등)의 이력 추적, 디지털 미디어의 저작권 관리, 지식 재산권 보호, 복지 및 의료 서비스 등을 언급하였다.

네덜란드는 다른 국가보다 먼저 블록체인을 미래를 위한 핵심 기술로 판단하여, 공공-민간 협력 프로젝트에 투자하는 등 블록체인 기술의 응용 활성화에 앞장서고 있다. 대표적으로 "The Dutch Blockchain Coalition, Dutch Digital Delta(Dutch Digital Delta)"은 네덜란드 정부의 주도로 다양한 연구기관, 기업 등이 참여하고 있는 공공-민간 협력 블록체인 프로젝트이다(이소정, 2018. 1. 24.).

미국은 전 세계적으로 블록체인 특허 출원이 가장 많은 국가이며, 주 정부 차원에서 블록체인 기술 및 활용에 관한 다양한 법안이 마련되고 있다. 콜로라도주는 정부 기록 보존 및 사이버보안을 위해 블록체인 기술을 사용하는 법안을 통과시켰고, 애리조나주도 마찬가지로 블록체인에 데이터를 보관하고 공유할 수 있는 법과 세금 지불 및 공과금 결제에 비트코인 사용을 허가하는 법안, 테네시주에서는 블록체인 기술을 통한 스마트 계약을 법적으로 인정하는 법안을 통과시켰고, 델라웨어주는 주식 거래 명부에 블록체인 사용을 허용했다(과학기술정보통신부, 2018. 6.).

영국 과학기술부에서 발간한 2016년 보고서에서는 블록체인 기술의 효용성 평가 및 실증 사업의 추진, 규제, 개선, 기술력 확보 등을 추진할 것을 권고 하고, 공공 서비스 추진 전략을 밝혔다(UK Government office for Science, 2016).

에스토니아는 전 세계 국가 중에서도 상당 부분이 디지털화된 국가로 블록체인과 빅데이터를 활용한 전자정부 시스템 분야의 선두 국가이다. 태어나자마자 전자신분증(e-ID)을 발급받으며, 신분증을 컴퓨터에 꽂고 본인 인증만 하면 본인의 부동산 관련 내역, 차량번호, 의료기록, 건강보험 기록을 확인할 수 있으며 납세, 투표 등 약 2600가지 정부서비스를 온라인으로 대부분 이용할 수 있다(신흥지역정보 종합지식포털, 2017).

다른 나라와 마찬가지로 우리나라에서도 공공과 민간이 협력하여 블록체인 선도 국가로 나아갈 수 있도록 '블록체인 기술 발전전략(2018. 6. 22.)'을 수립하였다. 세부 추진 전략은 공공, 민간의 다양한 사업으로 블록체인에 대한 국민체감도를 높이고, 블록체인의 기술경쟁력을 확보하며 블록체인 산업의 활성화를 위한 기반을 조성하는 것 등이며, 국내 산업을 혁신하고 디지털 신뢰사회를 구축해 블록체인 선도 국가로 도약하는 것이 목표이다. 2018년 6개의 블록체인 시범사업을 추진하고 2019년부터는 시범사업을 확대하며 상용서비스를 확산시킬 예정이다.

〈표 2-3〉 2018년 6대 블록체인 시범사업

구분	사업명	설명	협업부처
축산물 이력 관리	안심하고 먹을 수 있는 소고기 이력 관리	- 사육에서 판매에 이르기까지 모든 정보를 블록체인으로 공유하여 추적 - 최대 6일 걸렸던 추적이 10분 이내로 단축	농식품부
개인통관	신속하게 처리하고 허위 신고도 예방하는 개인통관	- 12시간 이상 소요되던 통관 처리의 주문에서 선적, 배송, 통관까지의 전 과정을 블록체인에 기록 - 실시간 수입 신고로 시간 단축 및 물류비용 절감, 저가 신고 사례도 예방	관세청
부동산 거래	간편한 부동산 거래	- 부동산 담보 대출 시 서류 제출을 위해 여러 관계기관(주민센터, 국세청 등)에 방문해야 했으나, 블록체인 기반 부동산 거래 플랫폼을 통해 은행 또는 관계기관 중 한 곳만 방문하여 업무 처리	국토부
온라인 투표	편리하고 믿을 수 있는 온라인 투표	- 전에는 선관위에서만 투표 내역을 소유하였으나, 온라인 투표 정보를 블록체인에 기록하여 선거 후보자, 참관인, 선관위 모두 투개표 과정 및 결과 공유	선관위

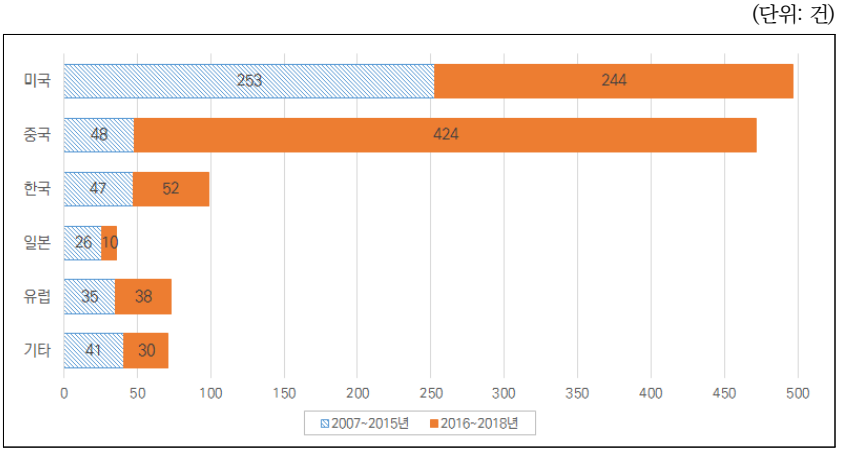
구분	사업명	설명	협업부처
전자문서 유통	외국기관에 공문서 제출 편리	- 외국기관에 공문서 제출 시 이전에 최소 14일 소요되던 업무를 블록체인에 공문서와 외교부 인증서를 저장하여 종이문서 대신 전자문서 형태로 외국기관과 공유하고 실시간으로 처리	외교부
해운물류	터미널 간 환적 컨테이너 운송 효율화	- 컨테이너 터널마다 별도의 앱을 이용하는 것에서 컨테이너 반출입증을 블록체인 기반으로 공유하여 운송 프로세스 개선	해수부

자료: 과학기술정보통신부(2018. 6. 22.) 신뢰할 수 있는 4차 산업혁명을 구현하는 블록체인 기술 발전전략. 과학기술정보통신부

블록체인에 대한 관심이 증가함에 따라 블록체인과 관련한 특허 출원도 전 세계적으로 폭발적으로 증가하고 있다. 2018년 1월 말 기준으로 공개된 블록체인 관련 특허 출원은 총 1248건이며, 누적 건수로는 미국이 497건으로 가장 많으나 2016년 이후 연간 출원 건수는 중국이 가장 많다(특허청 보도자료, 2018. 3. 22.). 국내에서 2016~2018년까지 52건 특허를 출원하여 2007년부터 2018년까지 총 99건 특허 출원 하였다.

[그림 2-11] 국가별 블록체인 특허 출원 현황

(단위: 건)

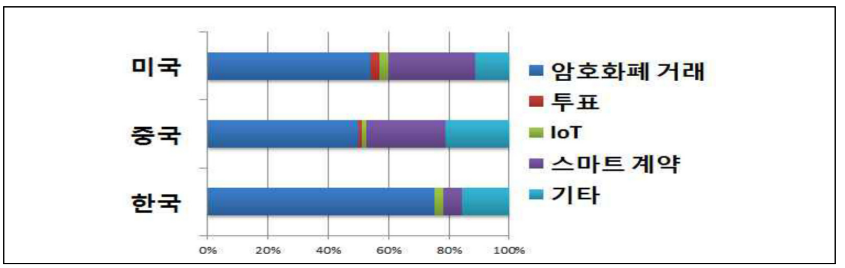


자료: 특허청 보도자료 2018. 3. 22. 재구성

특허 출원은 주로 보안, 운용, 활용 등을 중심으로 진행되며, 출원된 특허가 현재는 암호화폐 거래에 집중되어 있으나 미국과 중국은 점차 물류, 의료, 공공서비스 등 서비스 분야로 확대되고 있다. 미국의 경우 암호화폐에는 58건, 서비스 분야에는 31건 활용되었고, 중국의 경우 암호화폐에는 78건, 서비스 분야에는 41건 활용되었다(특허청 보도자료, 2018. 3. 22.).

[그림 2-12] 블록체인 활용 분야의 국가별 특허 현황

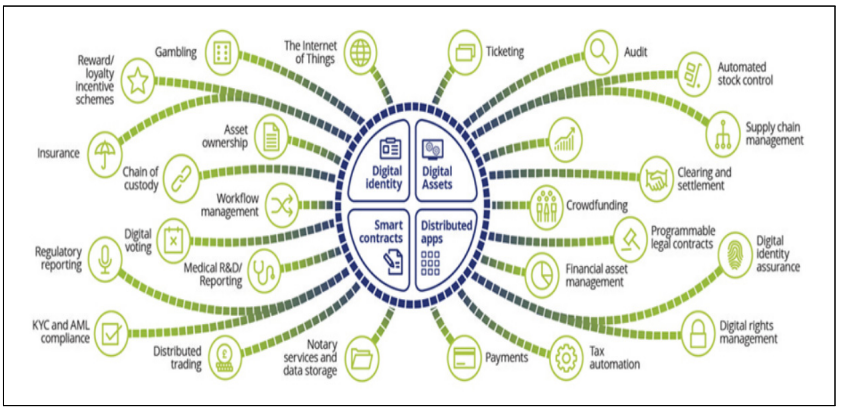
(단위: %)



자료: 특허청 보도자료 2018. 3. 22. 재인용

앞서 블록체인의 기술적 특징을 알아보고 용도에 따라 분류해 보았다. 이러한 블록체인에 기반한 다양한 적용 사례가 존재하며 현재도 새로운 사업이 생겨나고 있다.

〔그림 2-13〕 블록체인 활용 분야



주: KYC Know Your Customer /AML Anti Money Laundering
자료: Deloitte 홈페이지
(https://www2.deloitte.com/uk/en/pages/innovation/solutions/deloitte-blockchain-practice.html). 2018. 8. 27. 인출

보건의료 및 복지 분야 그리고 그 외에 대표적인 금융, 물류, 에너지, 사물인터넷 등에서의 블록체인의 국내외 활용 사례를 좀 더 자세히 살펴보고자 한다.

가. 보건의료

보건의료 분야에서는 정밀의료와 맞춤형 의료가 확대됨에 따라 수면, 운동, 영양과 관련한 정보뿐만 아니라 유전체정보, 혈액, 혈압, 신체검사 등에 많은 양의 정보 처리가 필요하고, 기관별로 분산되어 있는 데이터의 상호 운용과 접근성 및 기록의 보관·교환 등에 드는 비용의 문제 등이 블록체인 기술을 도입함으로써 많은 부분 해결될 수 있을 것이다(박순영, 2018).

〈표 2-4〉 보건의료산업의 요구와 블록체인 기술의 기회

보건의료산업의 요구	블록체인 기술의 기회
분절화된 데이터	- 컴퓨터 네트워크를 활용한 환자데이터 분산 저장 - 네트워크와 노드를 통한 데이터 공유 - IoT 데이터의 분산 소싱
환자데이터에 대한 실시간 접근	- 분산원장으로 환자의 건강데이터에 안전하게 접근 - 공유데이터의 네트워크상 실시간(real-time) 업데이트 가능
시스템 상호운용성	- 전 지역에 걸쳐 분산된 인터넷과 컴퓨터 네트워크 - 신빙성 보장(enables authenticity)
데이터 보안성	- 디지털화된 데이터 보안성으로 환자의 개인정보 보호
환자생성데이터	- 웨어러블 디바이스(IoT)로 집적된 데이터가 전체론적 환자 치료(holistic patient care)에 제공
접근성과 데이터 불일치	- 스마트계약은 특정 의료기관에서만 승인받은(permissioned) 환자 데이터에 대한 접근과 분석에 일관되고 규칙에 기반한 방법론 제시
비용 효과성	- 거래비용 감소와 실시간 처리 - 제3자를 거치지 않은 데이터 접근으로 시간 단축

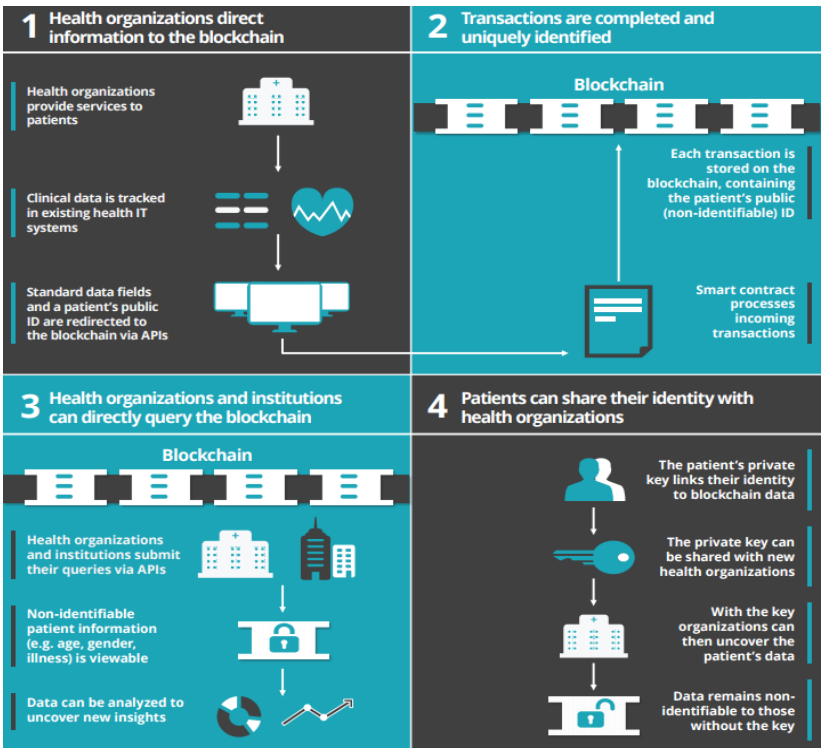
자료: Dash et al. (2016). Blockchain: A Healthcare Industry View. p.6 내용 재구성

개인의 건강 및 의료기록 관리에 블록체인 기술을 적용하여, 효율적으로 의료 정보를 관리하도록 하는 프로젝트가 추진되고 있다. IBM 왓슨(Watson) 헬스는 미국 질병관리예방센터와 협력하여 기존에 병원이 보유한 진료 정보를 블록체인 네트워크에 저장하고 관리하는 기술을 개발하고 있으며 에스토니아 정부는 국민의 의료기록을 블록체인에서 관리할 수 있도록 시스템을 구축하고 있다. 구글 딥마인드(DeepMind) 헬스는 영국 국가보건서비스와 함께 환자가 자신의 의료기록을 실시간으로 확인할 수 있는 블록체인 기술을 개발하고 있다. 국내의 경우 메디블록, 써트온, 포씨게이트 등의 업체가 의료기록 관리를 위한 블록체인 기술을 개발하고 있다.

미국 보건복지부(HHS) 산하 기관인 ONC HIT(The Office of the National Coordinator for Health Information Technology)는 '블록체인 챌린지(Blockchain Challenge)'를 진행하였다. 이 공모전에서

블록체인으로 의료 정보를 보호하고 관리하며 교환할 수 있는 방법에 대한 다양한 아이디어를 공모하였다. 특히 건강데이터는 암호화되어 건강기록이 사용자의 고유 식별표식과 함께 데이터 호수(data lake)에 저장되어 자신의 데이터에 대한 접근 권한을 가진 사용자가 데이터를 공유하는 방법을 관리하도록 구축하는 아이디어가 확인되었다(김경민, 2018. 2. 16. 재인용). 이 외에도 ONC HIT에서는 2017년 3월 오픈소스 분산원장 기술과 건강 관련 표준을 활용하여 세 가지 트랙 중 하나를 해결하는 'Blockchain in Healthcare Code-A-Thon'을 진행하였다. 또한 미국 식품의약국(FDA)은 의료 분야에서 건강관리 효과를 극대화하기 위해서는 병원마다 분산된 정보를 통합하고 데이터를 상호 연계하는 빅데이터를 구성할 필요가 있다고 보고 미국인의 의료정보 전체를 블록체인으로 만들 계획이다(김경민, 2018. 2. 16. 재인용).

〔그림 2-14〕 보건의료 블록체인 생태계

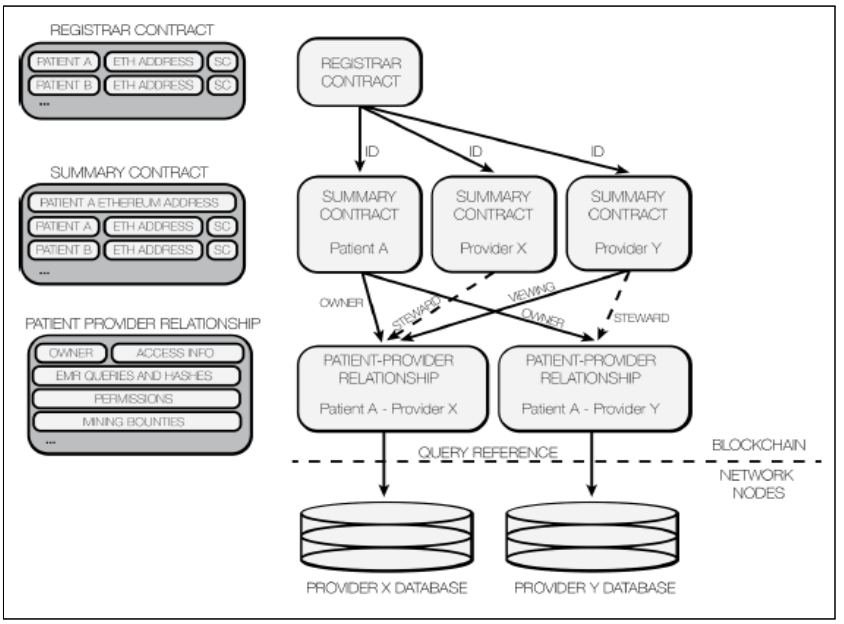


자료: Deloitte (2016) Blockchain: Opportunities for Health Care. 재인용

MedRec[14]은 미국의 MIT Media Lab과 이스라엘의 Beth Israel Deaconess Medical Cetert에서 블록체인 원장으로 환자 치료에 관한 정보를 공유하는 테스트를 진행한 것이다. 이는 의사(제공자)와 환자가 전자건강기록(Electronic Health Record, EMR)을 공유하는 것을 목적으로 하며, 환자의 기록을 직접 저장하지 않고, 환자가 데이터에 안전하게 접근할 수 있도록 메타 데이터를 인코딩하여 서로 다른 공급자(의사, 약사 등) 간의 데이터를 통합한다. 현재 MedRec 2.0이 개발되고 있다.

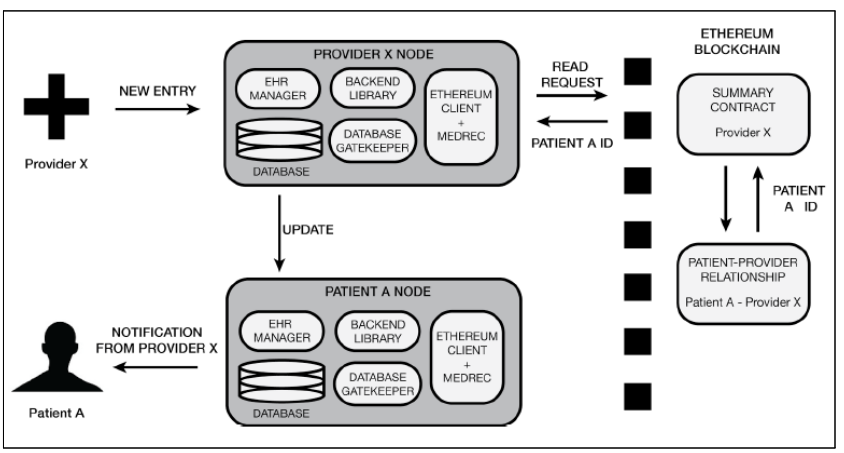
14) https://medrec.media.mit.edu의 기술 문서(Techincal Documentation)를 재정리함.

〔그림 2-15〕 MedRec의 스마트계약 프로세스



자료: https://medrec.media.mit.edu의 Techincal Documentation 재인용

〔그림 2-16〕 MedRec 2.0의 구조



자료: https://medrec.media.mit.edu의 Techincal Documentation 재인용

나. 복지 및 공공서비스

정보의 투명성 제공이라는 블록체인의 특성을 활용하여 공공 및 복지 분야에서 블록체인 기술의 활용이 적극적으로 고려되고 있다. 미국, 스페인, 호주 등의 정치권에서는 블록체인 기반의 전자 투표 방식을 활용하고 있다. 국내의 경우 경기도가 전자 투표에 블록체인 기술을 적용하였다. 우리나라를 비롯하여 미국, 스웨덴, 조지아공화국 등은 부동산 등기와 거래에 블록체인을 적용하는 시범사업을 추진하고 있다. 복지예산 집행의 투명성 제고를 위해 블록체인 도입이 활발히 진행되고 있다. 영국은 노동연금에 블록체인을 결합하는 시스템을 개발하고 있으며 우리나라의 경우 서울시가 청년수당 지급 과정에 블록체인 기술을 적용하고 있다.

유엔 세계식량계획(World Food Programme)에서는 블록체인 기반의 난민 식량 원조를 위한 '빌딩블록스(Building Blocks)' 프로젝트 개발을 완료했다. 이 프로젝트는 이더리움 블록체인 기반의 네트워크에 가족계좌를 개설해 재정 지원을 받음으로써 보다 효율적이고 투명한 식량 원조를 지원하게 되었다. 전에는 자금 지원을 위한 국경 간 자금 이동에 과도한 수수료가 들었으나 블록체인 기법으로 수수료를 거의 지불하지 않게 되었다. 난민들은 생체정보(예: 홍채 스캐너)로 신원 확인을 하고 블록체인 저장소에 저장된 식량 바우처를 이용해 가게에서 현금 없이 식품을 구입한다(유엔 세계식량계획 홈페이지).

중국 정부는 블록체인 기술의 장점인 정보의 위변조 불가성 및 추적성 등을 이용해 빈민구제 플랫폼을 구축할 예정이다. 2016년 10월 블록체인 전자지갑을 개발한 중국은행은 빈민구제 공유 플랫폼에 '중국공익'을 탑재하여 기부금 정보를 블록체인상에서 운용하면서 신뢰도를 높였다(KDB 미래전략연구소, 2018. 4. 30.; 지디넷 보도자료, 2018. 8. 24.).

대표적 복지국가인 영국은 매년 160억 파운드를 취약계층에게 복지수당으로 지급하지만, 이 중 고의적인 부정수급 12억 파운드, 수급자의 실수로 인한 과잉지급 15억 파운드, 공무원의 행정 실수로 인한 손실로 7억 파운드를 추산하고 있다. 이에 고브코인(GovCoin)과 협업하여 복지수급 대상자에게 블록체인을 기반으로 수당을 지급하고 사용하도록 실험을 하였다. 이 실험으로 수당 지출 내역을 암호화해 여러 서버에 분산 저장했고, 저장된 정보를 활용해 노동연금부(DWP)는 부정수급자를 발견하고 신청이나 지급 과정에서 벌어지는 실수를 찾아낼 수 있었으나, 2018년 현재까지 이 시스템을 적용하지는 않았다(한국일보 보도자료, 2018. 8. 20.).

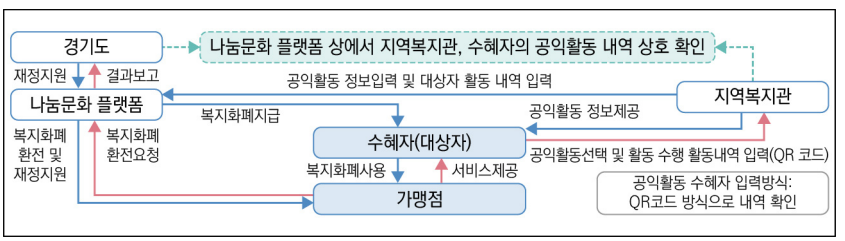
행정안전부는 종이증명서 발급에 따른 국민·기업의 불편과 사회적 비용을 획기적으로 줄이기 위하여 4차 산업혁명의 핵심 기술인 블록체인 기반의 전자증명서 발급·유통 플랫폼 구축에 나선다. 행정안전부는 2019년에는 전자증명서 발급·유통 플랫폼을 구축하고, 2019년 말에 시범서비스를 거쳐 2020년부터 전자증명서를 발급·유통할 계획이다(행정안전부 보도자료, 2018. 5. 16.). 이를 위해 2018년에는 블록체인 기반의 전자증명서 발급 및 유통센터, 전자문서지갑, 전자증명서의 진본성 확인 및 사용자 인증시스템 등에 대한 정보전략계획 등 전자증명서 발급·유통에 대한 청사진을 제시할 계획이다(행정안전부 보도자료, 2018. 5. 16.).

각 지방자치단체는 지역화폐를 발행하여 지역 경제를 부흥시키기 위해 노력하고 있다. 대전의 '두루', 과천의 '품앗이', 성남의 '성남누리' 등이 이에 해당하며 행정구역 밖으로 자금이 유출되는 것을 막고 외부 자금이 행정구역 안으로 유입되는 것을 목적으로 사용된다. 지역화폐 활성화로 최근 블록체인을 기반으로 한 전자화폐를 준비하는 지자체들이 생겨나고 있다. 서울시는 현금이나 카드 없이 QR코드로 결제가 가능한 S코인 발행을 준비하고 있으며 블록체인을 활용하여 청년수당 신청 서류 발급을 간

소화하고 장한평 중고차매매의 자동차 주행거리, 사고 사실들의 투명성과 신뢰성 제고를 위한 정보화전략계획(ISP)을 수립한다(서울시 보도자료, 2017. 9. 19.).

경기복지재단은 경기복지자원 공유 플랫폼에 다양한 하위 플랫폼의 하나로 복지화폐 플랫폼을 운영하는 방향성을 제시하였다. 복지화폐 수혜 대상자의 공익활동 수행을 기본으로 하고, 기부자, 서비스 제공자 등 특정인을 대상으로 운영되는 컨소시엄 블록체인의 형태가 적합하도록 설계되었다(경기복지재단, 2017. 3. 14.).

[그림 2-17] 경기도의 블록체인 기반 복지화폐 운영 체계(안)



자료: 경기복지재단(2017. 3. 14.) 블록체인 기반 복지화폐 활용방안. G-Welfare Brief. vol 4, p.1 재인용

블록체인을 활용한 온라인 투표는 기존 투표에 비해 시간과 비용이 절약되며, 많은 사람들이 정책 결정 과정에 참여할 수 있고 탈중앙화된 정보공유 시스템으로 무결성 및 보안성 확보가 가능해진다(홍승필 등, 2017). 2016년 미국 텍사스주에서 미국 사상 처음으로 블록체인 방식의 인터넷 투표가 실시되었고, 유타주에서는 공화당 대선 후보 선출 과정에 활용되었다. 스페인의 '아고라 투표(Agora voting, 현재는 nVotes로 불림)', 에스토니아의 '아이 보팅(I-voting)' 등이 대표적인 블록체인 기반의 온라인 투표 시스템이다. 우리나라 또한 경기도에서 처음으로 '따복공동체 주민제안 공모사업 평가'를 블록체인 시스템으로 실시하였다.

〔그림 2-18〕 2017 따복공동체 주민제안 공모사업의 온라인 투표 방법

자료: 경기도 따복 공동체지원센터. 2017년 따복공동체 주민제안 공모사업 블록체인 도입 인포그래픽스. (ddabok.go.kr 2018. 9. 5. 인출) 재인용

다. 금융

뱅크오브아메리카, 씨티그룹, 골드만삭스, SBI홀딩스 등 세계적인 금융기관들은 블록체인을 통한 시스템 구축 및 운영 비용 절감과 효율성 제고를 목표로 자체 폐쇄형 블록체인의 개발 또는 금융 연합체 참여로 생태계 구축과 서비스 개발에 적극적으로 나서고 있다.

R3CEV(Crypto, Exchanges and Venture practice)는 금융서비스와 관련된 블록체인 기술을 개발하기 위해 2015년 9월 9개의 금융기관과 컨소시엄을 결성했으며 최근 43개의 금융회사, 청산소, 거래소, 기술업체 등 80개 이상이 가입하였다. 2016년에는 국내 은행인 국민은행, 신한은행, 하나은행, 우리은행, 기업은행이 가입했다. R3CEV에서는 블록체인 기반의 송금·결제, 계약 체결, 자금세탁 방지 등 다양한 프로젝트가 진행되고 있다(백웅조, 2017). 국내의 경우 공인인증서를 대체하기 위한 기술로 블록체인이 적극 검토되고 있다.

라. 무역·물류·유통

금융과 더불어 블록체인 기술이 가장 적극적으로 적용될 분야로 무역·물류·유통 분야를 꼽을 수 있다. 정부, 기업 등 다양한 이해 당사자가 관여해야 하는 이러한 분야에 블록체인 기술은 매우 효율적으로 활용될 수 있다. 복잡한 계약 단계를 블록체인의 스마트컨트랙트로 간소화하고 계약서의 위변조 가능성을 원천 차단할 수 있기 때문이다. 월마트, 네슬레, 돌 등 주요 기업들이 IBM과의 협력으로 물류·유통에 블록체인을 도입하는 작업을 시작하여 개념 증명(PoC: Proof-of-Concept)을 완료하였다. 국내의 경우 관세청에서 민관 합동 해운물류 블록체인 컨소시엄 시범

사업을 통하여 수출통관 업무에 블록체인을 적용하는 개념 증명 단계를 통과하였으며, 올해 블록체인 기반 기술을 수출 통관, 물류 서비스에 시범 적용할 계획이다. 또한 농식품에 대한 생산 및 유통 이력 추적을 위한 블록체인 기술 접목도 시도되고 있다.

[그림 2-19] 수출 통관, 물류 서비스의 블록체인망 개념도

자료: 관세청 보도자료(2017. 12. 21.) 관세청, 세계최초 블록체인 기반 수출 통관 서비스 기술검증 완료. 관세청 보도자료. 재인용

마. 사물인터넷(Internet of Things, IoT)

사물인터넷과 관련해서는 사물인터넷-블록체인 컨소시엄이 결성되어 사물인터넷 운영의 효율성을 높일 수 있도록 하는 기술 개발이 진행되고 있다. 특히 사물인터넷은 의료나 물류·유통 분야에서도 필수적으로 활용되는 주요 기술로, 블록체인 기반 사물인터넷의 표준화를 위한 노력도 함께 진행되고 있다.

바. 에너지

에너지 생산 및 거래를 위한 스마트 마이크로그리드 기술에 블록체인을 접목하여 생산자와 소비자가 실시간으로 전력을 거래할 수 있도록 하는 기술의 개발이 활발히 진행되고 있다. 솔라 코인(Solar Coin)은 태양광 발전량에 따라 암호화폐를 지급하여 신재생에너지 활용을 장려하고 있으며, 블록 차지(Block Charge)는 블록체인 기반의 전기차 충전을 위한 지불 시스템으로 개발되고 있다.

블록체인에 에너지 데이터를 공유함으로써 새로운 비즈니스 모델을 구축하기도 하였으며, 에너지 공급이 필요한 개발도상국에 신재생에너지 설비를 공유하는 데도 블록체인을 활용하였다. IBM과 중국의 스마트 블록체인 랩(Smart Blockchain Lab)에서는 탄소할당량을 관리하는 비용을 줄이기 위해 블록체인 기술을 이용하는 탄소거래시장 감독 관련 방안도 모색하고 있다(우청원, 2018).

국내에서는 한국전력이 블록체인 기반의 이웃 간 전력 거래 및 전기차 충전 서비스를 구축하는 등 개념 증명 단계를 진행하고 있다. 이웃 간 전력 거래는 지붕 위 태양광 등으로 전기를 생산소비하는 사람(프로슈머, Prosumer, ‘producer’와 ‘consumer’의 합성어)이 남는 전기를 한국전력의 중개를 통해 이웃에게 판매하는 전력 거래 방법으로 프로슈머와 구매자를 매칭하여 에너지 포인트로 직접 거래할 수 있게 하였다(과학기술정보통신부 보도자료, 2017. 12. 6.).

〈표 2-5〉 블록체인 기술의 적용 분야 예시

적용 분야	활용 방안
금융서비스	- 블록체인 환경에서 주식, 채권, 예금증서, 기업어음 등의 금융 수단을 보다 빠른 결제 처리와 보다 낮은 거래 비용으로 이전시킴으로써 현재의 서비스를 보다 효율적으로 제공할 수 있음 - 장기적인 블록체인 금융의 비전은 시장에서 발생하는 자연적인 행위들에 직접적으로 임베딩됨으로써 그 자체로 작동하는 금융시장이 되는 것
블록체인 금융과 결합한 자동차 생태계	- 자동차 산업 생태계의 모든 것이 담긴 기록을 호스팅하는 블록체인을 운영함으로써 소유권, 금융, 등록, 보험과 서비스 거래 등을 모두 추적 - 무인자동차 제조업체는 택시 운송 사업자들에게 차량을 공급한 후, 탑승객이 요금을 지불할 때마다 파이낸싱이 임베딩된 블록체인의 스마트계약으로 제조업체와 수익을 분배할 수 있음 - 제조업체는 소비자에게 차량을 판매할 필요가 전혀 없으며, 자신들이 상품에 의해 수행되는 모든 거래에서 나오는 분배금으로 현금 흐름을 확보할 수 있어 은행 금융도 필요하지 않게 됨
의료정보 생태계	- 전체 의료정보의 생태계가 보험사, 의료기관, 환자를 연결하는 블록체인의 형태로 이뤄질 수 있음
디지털 저작권 보호	- 음악 파일 사용 기록을 공공 블록체인에 기록할 수 있음 - 아티스트는 자신들의 음악을 블록체인 기반의 음악 생태계에 출시하고, 데이터 이용 조건을 관리할 수 있으며, 로열티는 스마트계약을 통해 실시간으로 분배할 수 있음
저가 자산을 위한 새로운 신용 시장	- 거래비용으로 인해 대출할 때 담보로 사용할 수 있는 자산에 대해서는 실질적인 하한선이 설정되어 있음 - 블록체인은 거래비용의 상당 부분을 제거할 수 있기 때문에 저가 자산을 기반으로 한 새로운 등급의 대출을 현실화할 수 있음
성과 기반 과금	- 스마트계약을 활용해 성과 정도에 따른 지급 계약을 실행할 수 있음 - 피자가 30분 이내에 도착하지 않으면 무료로 해 주던 서비스가 블록체인에 의해 부활할 가능성이 있음. 또는 차등제도를 활성화할 수 있음
세금 과세	- 실시간으로 기록되는 디지털 거래의 세계에서 불법 거래를 은폐하기는 매우 어려워짐
산업 매시업 (Industrial Mash-ups)	- 블록체인은 기업 간에 유동적으로 협력하는 완전히 새로운 세계를 실현할 수 있는데, 이를 '산업 매시업'이라 부름 - 산업 매시업 형태의 제휴에서 한쪽 기업은 상대방이 그들의 고유의 비즈니스 목적을 위해 자산과 역량을 지속적으로 사용하는 것에 영향을 미치지 않으면서 상대방이 가진 자산과 기능을 활용하여 새로운 비즈니스 가치를 창출할 수 있음
산업용 IoT	- 블록체인 기술을 이용하여 산업 매시업과 IoT가 융합함으로써 기업이 소유한 고가의 산업용 자산의 활용도를 높일 수 있음 - 운송 컨테이너, MRI 장비, 건설 장비 등 모든 자산을 실시간으로 디지털 마켓플레이스로 연결함으로써, 기업들은 장비를 가동하지 않는 유휴 시간을 기업 간에 거래할 수 있음

원자료: Cudahy, G. (2016). Blockchain reaction techcompanies plan for critical mass. Ernst & Young, 재정리
자료: 박종훈(2016. 10. 12.) 최신 ICT 이슈: 금융을 넘어 모든 비즈니스에 파괴적 영향을 미칠 블록체인. IITP, 주간기술동향 재인용

제3절 블록체인의 한계점

지금까지 블록체인 기술 및 활용사례에 대해 살펴보았다. 블록체인 기술은 기존과 다른 탈중앙화와 신뢰성이라는 새로운 패러다임을 제시하면서, 다양한 응용 분야에 활용될 것으로 예상된다. 그러나 현재 블록체인 기술의 완성 수준은 초기 단계에 불과하며 기존 시스템과 비교할 때 뚜렷한 장단점을 가지고 있다. 따라서 블록체인 기술이 모든 분야에 적합한 것은 아니라고 할 수 있다. 다만 블록체인 기술의 한계를 극복한다면 응용 분야는 점차 확대될 것으로 기대된다. 현재의 블록체인 기술이 직면한 몇몇 한계점에 대한 논의를 끝으로 블록체인 기술에 대한 소개를 마치도록 한다.

1. 성능(performance)·확장성(scalability) 문제

개방형 블록체인의 경우 초당 처리할 수 있는 거래가 수십 개에 불과하다. 폐쇄형 블록체인의 경우에도 수천 개에 불과한 상황이다. VISA 신용카드 거래의 경우 최대 초당 5만 건[15]에 달하는 것으로 알려져 있는데, 블록체인이 이를 대체하기에는 아직까지 성능이 턱없이 부족하다.

확장성 문제는 더 많은 자원을 투입하더라도 거래 처리 속도가 증가하지 않는 것을 말한다. 이는 블록 생성 시간과 블록 크기의 제한과 연관되어 있다. 거래 처리 속도를 높이기 위하여 블록의 생성 시간을 빠르게[16] 하면 블록체인의 분기가 발생할 확률이 높아지며 이는 블록체인의 보안

15) https://usa.visa.com/dam/VCOM/download/corporate/media/visa-fact-sheet-Jun2015.pdf

16) 넌스값을 쉽게 찾도록 난이도를 낮춤.

을 취약하게 만든다. 또 한 번에 많은 양의 거래를 담기 위하여 블록의 크기를 늘리면 네트워크로 주고받아야 하는 데이터의 양이 급속히 증가하며, 이는 더 고사양의 컴퓨팅 장비가 동원되어야 함을 의미한다. 참고로 블록 하나의 크기가 1MB였던 비트코인의 경우, 지난 십여 년간의 거래 내역이 모여 있는 블록체인의 용량이 현재 약 180GB[17]에 달하고 있다. 블록체인의 용량은 시간이 지남에 따라 계속 증가할 것이므로 노드는 지속적으로 저장 장치를 추가해야 한다. 저장 문제는 블록체인의 확산에서 최대의 장애다. 이러한 확장성의 한계를 극복하기 위하여 최근 사이드 체인, 샤딩(sharding), 라이트닝 네트워크 기술 등이 개발되고 있다.

폐쇄형 블록체인의 경우 합의 알고리즘의 복잡성이 노드 수에 비례하여 증가하게 되는데, 실제로 대표적인 폐쇄형 블록체인인 패브릭 블록체인의 경우 수십 노드 이상이 블록체인 네트워크에 참여하게 되면 성능이 나빠지는 것으로 보고되고 있다(Scherer, 2017).

2. 개인정보(privacy) 침해

블록체인에 기록된 장부는 네트워크에 참여하는 모두가 볼 수 있다. 단지 참여자는 거래지갑 주소에 따라 익명화될 뿐 거래 정보는 모두에 의해 추적 및 분석이 가능하다. 이를 원천적으로 차단하기 위해 Z캐시와 같이 영지식 증명(Zero-knowledge Proof)에 기반하여 익명성이 특별히 보장되는 블록체인이 존재하지만 이를 일반화하기에는 많은 어려움이 따른다.

17) https://www.blockchain.com/en/charts/blocks-size 2018. 8. 31. 인출.

3. 자원 낭비

블록체인 네트워크를 유지하기 위해 모든 노드는 합의 알고리즘을 준수해야 한다. 현재 채굴 시장에서 가장 점유율이 높은 비트코인 및 이더리움은 작업 증명(PoW) 방식을 따르고 있다. 작업 증명은 막대한 전기에너지를 필요로 하며, 이미 채굴에 소비되는 전기에너지가 작은 나라 전체가 소비하는 에너지양을 넘어서고 있다. 더욱 심각한 것은 채굴에 소비되는 전기는 우리 사회에 전혀 도움이 되지 않는 단순 숫자 찾기 계산에 불과하다. 자원 낭비를 최소화하는 혁신적인 합의 알고리즘 개발이 절실하다.

4. 보안 위협

개방형 블록체인의 경우 누구나 네트워크에 참여하여 채굴에 도전할 수 있다. 그러나 더 고성능의 컴퓨터를 갖춘 노드가 작업 증명에 유리할 수밖에 없다. 성능을 높이기 위해서는 더 많은 자본을 투자해야 함을 의미하고, 이는 블록체인 네트워크에 참여하려는 자의 진입 장벽을 높게 만든다. 이러한 상황에서 개인이 채굴에 도전하여 새로운 블록을 생성할 확률은 극히 낮다. 이를 타개하기 위해 채굴 연합을 형성한 후, 공동으로 채굴에 참여한다. 만약 해당 연합에서 새로운 블록을 발견하였을 경우 보상을 기여도에 따라 나누어 갖는 형태를 공동 채굴(mining pool)이라고 하며, 현재 이러한 채굴 방식이 주를 이루고 있다. 그러나 하나의 공동 채굴 단체가 블록체인 네트워크의 과반수의 컴퓨팅 파워를 지속적으로 확보하게 된다면 이는 블록체인의 보안에 심각한 위협이 된다.[18]

---

18) 새로운 블록을 계속해서 만들어 낼 수 있으므로 마음대로 거래 내역을 조작할 수 있음.

5. 기타

블록체인 기술은 초기 단계로 숙련된 전문 인력이 부족하여 기술 확산과 생태계 조성에 한계가 있으며, 아직 먼 미래의 이야기지만 양자 컴퓨팅(quantum computing)이 현실화된다면 현재의 블록체인 암호 체계는 쉽게 뚫릴 수 있다. 따라서 보다 강력한 암호 체계를 개발하기 위한 기술도 함께 발전해 나가야 한다.

〈표 2-6〉 블록체인의 장단점

구분	장점	단점
익명성	개인정보를 요구하지 않음 은행계좌, 신용카드 등 기존의 지급수단에 비해 높은 익명성 제공	불법 거래대금 결제, 비자금 조성, 탈세를 가능하게 함
P2P	공인된 제3자 없이 P2P 거래 가능 불필요한 수수료 절감	문제 발생 시 책임 소재가 모호
확장성	공개된 소스에 따라 쉽게 구축·연결 확장 가능 IT 구축비용 절감	결제 처리 가능한 거래 건수가 실제 경제 내 거래 규모 대비 미미
투명성	모든 거래 기록에 공개적 접근 가능 거래 양성화 및 규제 비용 절감	거래 내역이 공개되어 있어 원칙적으로는 모든 거래 추적 가능 완벽한 익명성 보장이 어려울 수 있으며 조합에 의한 재식별이 가능
보안성	장부를 공동으로 소유(무결성) 보안 관련 비용 절감	개인 키 해킹, 분실 등의 경우 일반적으로 해결 방법 없음 기밀성 제공되지 않음
시스템 안정성	단일 실패점이 존재하지 않음 일부 참가 시스템에 오류 또는 성능 저하 발생 시 전체 네트워크에 미치는 영향 미미	채굴이 대형 마이닝 풀에 집중 실시간 대용량 처리의 어려움

자료: 강승준 (2018). 블록체인 기술의 이해와 개발 현황 및 시사점, 정보통신산업진흥원, 제4차 산업혁명과 소프트파워, 이슈리포트 2018-제13호. p 4 재인용

제 3 장 문서 기반 분석 방법

3 문서 기반 분석 방법

제1절 불용어 처리 및 형태소 분석

텍스트마이닝 분석에서 불용어 처리 및 형태소 분석은 분석 결과의 질을 높이는 데 매우 중요한 과정이다. 불용어를 처리하기 위해서는 다양한 방법을 사용할 수 있다. 주제별로 특이성을 갖는 단어에 관심이 있다면, 특정 분야에서 많이 관찰되는 단어가 중요한 의미를 가질 가능성이 크다. 반대로 모든 주제에서 같은 비율로 관찰되는 단어는 주제의 특징을 나타내는 데 적합하지 않거나 분석 단계에서 사전적으로 알고 있는 정보일 가능성이 크다. 예를 들어 '블록체인'에 대해 보건 · 복지 분야 외의 다른 분야에서 함께 데이터를 수집했다고 가정해 보자. 이 경우 주제 분야와 관계없이 '암호', '최신', '기술' 등의 키워드가 도출되며, 이 단어들은 '블록체인'이 가지는 대표 속성을 나타낼 뿐 보건 · 복지 관련 분야임을 나타내는 직접적인 특징은 아닐 것이다. 특히 이러한 단어들은 전체 검색된 텍스트 집합에서 매우 높은 빈도를 차지하기 때문에 상대적으로 빈도가 낮은 단어들이 가지는 특징들을 파악하기 힘들게 만든다.

만약 분야별로 신기술의 경향을 파악하려는 연구에서 이러한 단어를 제외하려 한다면, 다음과 같은 방법의 적용을 고려해 볼 수 있다. 먼저 분야별로 단어의 출현 빈도의 분포를 계산하고 분포들의 Hellinger distance를 기준으로 단어를 선택할 수 있다(Beran, 1977). 여기서 단어의 출현 분포란 '하나의 정해진 텍스트를 추출하였을 때, 특정 분야에서 해당 텍스트가 관찰된 확률'을 의미한다. 이 분포의 추정량을 Multinomial

분포의 모수 추정치로 사용할 수 있다. 어떤 단어의 출현 분포가 평균적인 분포에서 멀리 떨어지지 않았다면, 그 단어는 주제의 특징을 잘 반영하지 못한다고 판단하고 불용어로 처리할 수 있을 것이다.

두 개의 분포가 주어진 경우 분포의 거리는 다양한 방법으로 정할 수 있다. 많이 사용되는 분포 간의 거리 측도로 Hellinger distance가 있다. 두 확률 분포 P, Q를 Hellinger distance는 다음과 같이 정의한다.

$$H^2(P,Q) = \frac{1}{2}\int (\sqrt{\frac{dP}{d\lambda}} - \sqrt{\frac{dQ}{d\lambda}})^2 d\lambda$$

Hellinger distance는 단어 출현 빈도의 유사성을 나타내는 측도로, 분야별로 상대 빈도가 유사할수록 Hellinger distance가 작은 값을 가지게 된다. 즉 분야별로 출현 빈도가 유사한 단어는 해당 분야의 특징을 나타낼 수 없기 때문에 Hellinger distance가 작을수록 일반적으로 사용되는 단어이며 distance의 크기가 클수록 특정한 분야에서 사용되는 단어라고 해석할 수 있다. 이 외에도 분포의 특성을 반영하여 목적하는 단어의 특징을 잘 구분해 낼 수 있는 다양한 분포의 거리 측도를 도입할 수도 있다.

불용어를 제거한 뒤에는 문서 데이터에서 어절을 최소의 의미 단위인 '형태소'로 분해한 뒤, 각 '형태소'에 적합한 형태의 품사를 부여하는 자연어 처리 과정을 이용하여 문서를 정형 데이터로 변형한다. 이때 자연어 처리 과정에는 분해와 품사 부여를 위해 형태소 사전의 이용이 필요하다. 한글의 경우 R의 'KoNLP' 패키지 내부의 사전을 이용할 수 있다(Jeon and Kim, 2016). 영문 문서의 경우 'tm' 패키지의 stopword를 이용하여 불용어를 제거할 수 있다(Feinerer, 2018). 또한 영문 문서의 경우 띄어쓰기 단위로 문자를 나누고 주변 단어를 고려하여 단어를 확장할 수 있

는 'n-그램(n-gram)' 모형을 사용할 수 있다. n-gram 모형은 해당 단어를 기준으로 양옆 n개까지의 단어를 하나의 단어로 고려하는 모형으로 예를 들어 'I am a boy'라는 문장은 2-gram 모형을 사용하면 'I', 'am', 'a', 'boy', 'I am', 'am a', 'a boy'로 문장을 7개의 단어로 변환할 수 있다. 문서가 m개의 단어로 이루어져 있을 때 n-gram을 사용하면 mn-(n-1)n/2개의 복합어가 생성됨을 알 수 있다. 이처럼 단어 수집과 형태소 분석을 통해 문서 데이터를 단어의 집합으로 저장할 수 있다.

제2절 텍스트의 정량화(text quantification)

텍스트를 계량적 모형에 입력변수로 사용하기 위해 숫자로 변환하는 과정을 거친다. 즉 단어를 벡터로 변환하는 과정이다. 본 조사에서는 단어를 벡터로 변환하는 방법을 크게 두 가지로 구분하여 다룬다. 두 가지 방법은 One hot encoding으로 단어를 변환하는 방법과 Distributional hypothesis 가정하에 단어들의 분포를 기반으로 단어를 벡터화하는 Semantic embedding 방법이다.

1. One hot encoding

단어를 One hot encoding을 통해 벡터로 변환하는 방법은 다음과 같다. 전체 단어의 개수의 길이를 갖는 0 벡터를 생성하고 해당하는 단어의 위치에만 1을 할당하여 표현하는 방법이다. 각 단어를 위의 벡터로 변환할 수 있으므로 같은 방식을 문서에 적용하면 문서 또한 마찬가지로 벡터로 변환할 수 있으며, 이렇게 생성된 문서별 단어 벡터는 문서-단어 행렬

(Document-Term matrix) 혹은 단어-문서 행렬(Term-Document matrix)로 표현할 수 있다. 문서-단어 행렬에서 행은 문서를 나타내며 열은 단어의 등장 빈도 혹은 출현 여부를 표현한다. 이를 통해 문서별로 어떤 단어가 몇 번 등장했는지를 쉽게 확인할 수 있다. 단어-문서 행렬은 문서-단어 행렬의 전치 행렬로 단어가 문서별로 몇 번씩 등장했는지 확인하기 용이한 행렬이다. [그림 3-1]은 문서-단어 행렬의 예시이다. 첫 번째 문서에서는 'mining'이 5번, 'blockchain'이 2번 등장했으며, 두 번째 문서에서는 'blcokchain'이 1번 등장했음을 행렬의 각 원소를 통해 쉽게 확인할 수 있다. 이러한 방식을 사용하는 대표적인 분석 방법은 Topic model이다. Topic model은 분석자가 토픽이 몇 개일지 미리 정한 뒤, 문서를 토픽의 분포로 표현하거나 토픽을 구성하는 단어의 분포를 파악하는 모형이다. Topic model에는 대표적으로 잠재 디리클레 할당(LDA: Latent Dirichlet Allocation)이 있다(Blei, Andrew and Michael, 2003).

[그림 3-1] 문서-단어 행렬의 예시

mining	foundation	blockchain	based	cryptocurrencies	miner
5	1	2	2	1	1
0	0	1	0	0	0
0	0	3	1	0	0
0	0	2	1	0	0
0	0	4	0	0	0
0	0	2	1	0	0

문서-단어 행렬의 각 원소는 단어의 중요도를 평가하기 위해 가중치를 사용하여 변형할 수 있다. 가중치로는 주로 TF(Term Frequency)-IDF(Inverse

Document Frequency)가 사용된다(Leskovex, Rajaraman and Ullman, 2014). TF-IDF는 TF 가중치와 IDF 가중치의 곱으로 표현할 수 있다. 예를 들어 전체 문서 D 중 문서 d에서 발생한 단어 t의 가중치는 다음의 식으로 표현할 수 있다.

$$TFIDF(t,d,D) = TF(t,d)IDF(t,D)$$

TF-IDF 가중치 중 앞의 TF는 단어가 문서에서 발생하는 빈도에 대한 가중치를 의미한다. 문서에서 특정한 단어가 많이 등장할수록 해당 문서에서 단어를 중요하게 고려해야 한다는 아이디어로 이해할 수 있다. 문서 d에서 발생한 단어 t의 빈도를 $f_{t,d}$라고 할 때, TF에 대해서는 아래와 같은 여러 측도가 개발되어 있다.

binary: $TF(t,d) = I(f_{t,d} \geq 1)$

raw frequency: $TF(t,d) = f_{t,d}$

log normalization: $TF(t,d) = 1 + \log(f_{t,d})$

double normalization 0.5: $TF(t,d) = 0.5 + 0.5\frac{f_{t,d}}{\max_{t' \in d} f_{t',d}}$

double normalization K: $TF(t,d) = K + K\frac{f_{t,d}}{\max_{t' \in d} f_{t',d}}$

반대로 IDF는 모든 문서에서 등장하는 단어의 빈도에 대한 가중치를 의미한다. 전체 문서에서 많이 등장하는 단어일수록 중요한 단어가 아니라는 아이디어로 이해할 수 있다. 단어 t를 포함하는 문서의 개수를 $n_t = |\{d \in D : t \in d\}|$, 문서의 개수를 $N = |D|$라고 할 때 IDF에 대해

서는 아래와 같은 여러 측도가 개발되어 있다.

unary: $IDF(t, D) = 1$

inverse document frequency: $IDF(t, D) = \log \frac{N}{n_t}$

inverse document frequency smooth: $IDF(t, D) = \log \frac{N}{1 + n_t}$

inverse document frequency max: $IDF(t, D) = \log \frac{\max_{t' \in d} n_{t'}}{1 + n_t}$

probabilistic inverse document frequency:

$IDF(t, D) = \log \frac{N - n_t}{n_t}$

형태소 분석을 거친 문서는 TF-IDF 등의 가중치를 통해 단어의 중요도에 따라 가중치가 반영된 행렬로 변환되어 데이터 분석에 사용할 수 있다. 이때 행렬에서 한 셀(cell)은 문서에 포함된 키워드들의 빈도 혹은 중요도가 반영된 점수 측도이다.

2. Semantic embedding

앞서 소개했던 문서-단어 행렬을 생성하기 위해서는 전체 단어의 개수에 비례하여 데이터의 차원이 증가하며, 단어의 개수가 많은 경우 분석에 활용하기 어렵다는 단점이 있다. 또한 각 문서를 벡터로 표현했으나, 벡터가 가지는 의미가 없음을 알 수 있다. 예를 들어 '프랑스' 벡터와 '수도' 벡터를 더한다고 해도 '파리'라는 의미를 가지는 숫자 벡터를 구하는 것이 어렵다. 이러한 단점을 보완하기 위해 벡터를 숫자로 변형했을 때 그

의미를 반영하도록 하는 다양한 Semantic word embedding 방법이 제안되었으며 'Word2vec', 'GloVe' 등의 모형이 사용되고 있다(Mikolov, et al., 2013a, Pennington, Socher and Manning, 2014).

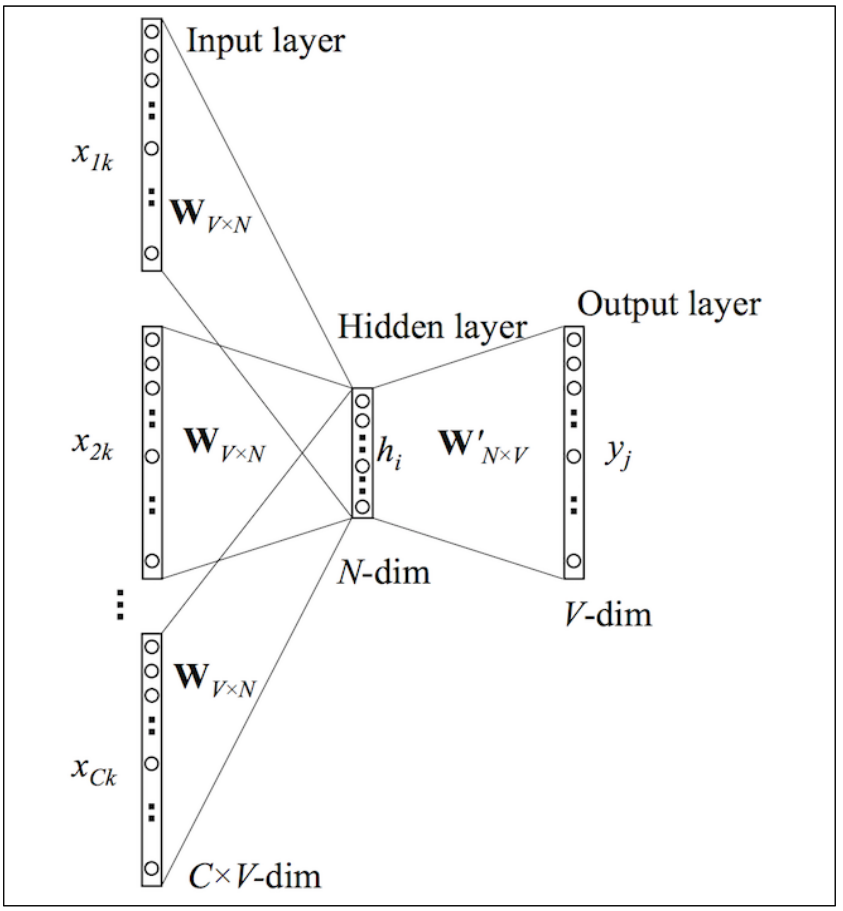
대표적인 Word embedding 방법인 Word2vec에는 크게 Continuous Bag of Words(CBOW)와 Skip-gram이 있다. CBOW는 주변에 있는 단어로 중심에 있는 단어를 예측하는 방식으로 특징 벡터를 도출하는 방식이며, Skip-gram은 중심 단어로 주변 단어와의 관계를 고려하여 임베딩하는 방법이다. 일반적으로 Skip-gram 모델의 성능이 CBOW 모형보다 더 나은 것으로 알려져 있다. CBOW 모형은 본문의 [그림 3-2]를 통해 도식화할 수 있으며 Skip-gram 모형은 본문의 [그림 3-3]을 통해 도식화할 수 있다. Skip-gram 모형은 은닉층이 하나인 간단한 신경망 모형으로 주변 단어와의 관계를 통해 특징 벡터를 추출할 수 있다. 그림에서 V는 전체 단어의 개수와 같고, N은 특징 벡터의 차원과 같다. Skip-gram은 중심 단어와 주변 단어의 확률을 최대화하는 방향으로 학습하며, 단어의 개수에 비례하여 추정해야 할 모수의 개수가 크게 증가한다. 이를 보완하기 위해 자주 등장하는 단어는 확률적으로 제외하고 모형을 학습하는 서브샘플링(subsampling) 방법이나 확률을 구할 때 전체 단어를 구하지 않고 일부 단어만 뽑아서 계산하는 네거티브샘플링(negative sampling) 등의 방법으로 모형 적합의 계산량을 축소했다(Mikolov, et al., 2013b). 위의 두 가지 방법은 패스트텍스트(fasttext)를 이용하여 쉽게 접근할 수 있으며, 파이썬과 R에서 모두 활용할 수 있도록 라이브러리를 제공하고 있다.

Skip-gram 모형을 수식적으로 표현하기 위해 다음과 같은 표기를 사용한다. 입력 단어 데이터는 One hot encoding을 이용하며 이때 One hot encoding 벡터를 $(x_1,\cdots,x_V)^T$를 통해 표현한다. 은닉층의 값은

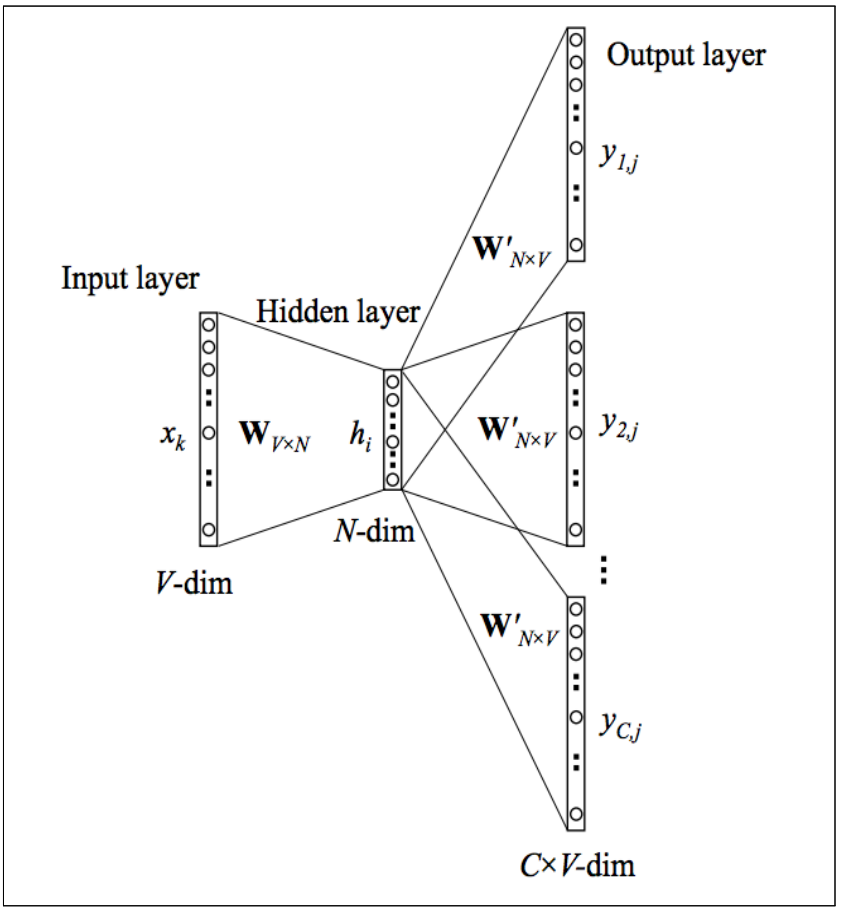
$h = W^T x = (v_1, \cdots, v_V)\begin{pmatrix} x_1 \\ \vdots \\ x_V \end{pmatrix} = v_k = W^T_{(k,\,\cdot\,)}$ 이고 마지막 출력층의 값은

$u = (u_1, \cdots, u_V)^T = W^{'T} h = \begin{pmatrix} v'^T_1 \\ \vdots \\ v'^T_V \end{pmatrix} h = \begin{pmatrix} v'^T_1 h \\ \vdots \\ v'^T_V h \end{pmatrix}$이며 최종적으로 출력되는 값은 활성함수를 이용하여 $p(w_j|w_k) = f(u_j)$라고 정의한다. Skip-gram 모형은 출력값을 최대화하는 모수를 추정하는 것이 목적이다. 즉 $\frac{1}{V}\sum_{v=1}^{V}\sum_{-c \le j \le c, j \neq 0} logp(w_{v+j}|w_v)$을 목적함수로 하여 이를 최대화하는 모수를 추정하는 것이 Skip-gram 모형의 목적이라고 할 수 있다. 여기에서는 앞서 소개한 단어 특징 추출 방법 중 CBOW와 Skip-gram의 아이디어를 설명하고자 한다.

〔그림 3-2〕 CBOW의 도식화



〔그림 3-3〕 Skip-gram의 도식화



제3절 단어/문장 특징 추출 방법론(word embedding method)

이 절에서는 2절에서 개략적으로 설명한 word embedding을 자세히 다루고자 한다. 4장의 분석 결과는 word embedding 방법론을 기반으로 분석되었기 때문에 분석에 사용된 방법론을 좀 더 이론적으로 살펴볼 필요가 있다.

단어들의 집합을 V라고 한다면 단어의 집합 위에서 직접 어떤 연산을 정의하기 어렵다. 의미가 가까운 단어를 찾는다거나 단어들 간의 위계를 탐색하는 등 단어를 이용한 분석을 위해서는 단어들이 연산이 쉬운 공간 위에 정의되어 있으면 편리할 것이다. 그중 가장 간단하면서도 많이 사용되는 방법이 One hot encoding이다.

One hot encoding은 하나의 단어를 벡터 공간에 대응시키는 간단한 방법이다. 일반적으로 단어의 집합 V에 총 n개의 단어가 있다고 하면, One hot encoding 방법은 하나의 단어를 R^n 위의 한 벡터 원소에 대응시킨다. 특별히 One hot encoding은 하나의 원소만 1이고 나머지는 모두 0인 elementary unit vector 위에 대응시켜서 하나의 단어를 하나의 벡터로 표시한다. 즉 One hot encoding된 벡터가 있다면 1의 값을 갖는 원소의 위치가 벡터가 대응시키고 있는 단어를 나타내며, 이 방법은 회귀분석에서 범주형 자료를 변환하는 더미코딩(dummy encoding) 방법과 같다. 만약 우리가 분석할 대상의 전체 단어 수가 1000개라면, One hot encoding으로 변환된 단어 벡터는 R^{1000} 위에 있는 하나의 벡터가 될 것이다.

단어를 분석하는 많은 경우 단어들의 집합은 매우 크며, 결과적으로 One hot encoding이 주는 단어들의 특징 벡터들이 매우 높은 차원의 공간 위에서 정의된다. 데이터의 차원이 높은 경우 계산상의 어려움이 있

을 뿐만 아니라, 더욱 중요한 것은 제한된 자료의 고차원 자료에서 추론이 매우 어렵다는 데 문제가 있다. 따라서 고차원 공간에서 표시된 단어들을 저차원으로 표현해 주는 것이 필요하고, 이에 대한 방법론을 단어 임베딩(word embedding)이라고 한다. 단어 임베딩에서 우리가 관심 있게 보아야 할 것은 '단어들 간의 어떤 정보의 손실을 가장 작게 하면서 단어 벡터의 차원을 줄여 나가는가' 하는 것이다. 단어가 가진 정보를 개략적으로 설명하자면 '임의의 문장에서 단어들의 출현 패턴'이라고 할 수 있다. 구체적인 출현 패턴은 단어 임베딩마다 다르게 정의된다. 여기서 소개할 단어 임베딩 방법인 주성분 분석, CBOW, Skip-gram은 각각 다른 방법으로 단어들의 출현 패턴을 정의하고 있다. 이것은 여기서 소개하는 단어의 특징 추출 방법들이 범용성을 가졌다기보다는 사용 목적에 따라 제한적으로 사용될 수 있음을 의미한다. 물론 최근에 개발된 단어 임베딩 방법이 기존의 방법에 비해 일반성을 가지고 있지만, 목적에 따라 잘 개발된 단어 특징 추출 방법을 사용하는 것이 바람직하다는 것을 언급해 둔다.

1. 주성분 분석을 이용한 단어 임베딩

주성분 분석에 대해 언급하기 전에 인코딩과 디코딩의 개념부터 소개한다. 인코딩은 고차원 벡터를 저차원 벡터로 변환하는 함수를 의미하며, 디코딩은 변환된 저차원 벡터를 고차원 벡터로 변환하는 함수를 의미한다. 일상적인 용어로는 인코딩이 데이터 압축, 디코딩이 압축 해제 혹은 압축 풀기라고 볼 수 있다. 동영상을 MPEG-4와 같은 형식으로 압축해서 저장하는 과정을 인코딩, 동영상 플레이어에 MPEG-4 디코더를 설치해 동영상을 재생하는 과정을 디코딩이라고 할 수 있다.

일반적으로 디코딩이 매우 빠르게 이루어지면서 동시에 압축된 데이터를 풀었을 때 원본의 정보 손실이 일어나지 않기를 바랄 것이다. 디코딩 과정을 선형 변환이라고 가정하면 원본의 복원 제곱 오차를 최소화하는 인코딩 역시 선형 변환이라는 것을 보일 수 있다. 이는 주성분 분석을 이용한 단어 임베딩의 기본 원리가 된다.

앞서 정의한 One hot encoding으로 변환된 단어를 $x \in R^n$이라고 한다. 이 단어를 보다 낮은 차원으로 변환하는 함수 인코더 f와 변환된 단어를 다시 복원하는 디코더 g를 생각한다.

- 인코더 $f : R^n \mapsto R^m (m < n)$
- 디코더 $g : R^m \mapsto R^n$

여기서 특별히 $g(z) = Dz$(단 $z \in R^m$, D는 n행 m열 행렬, D의 각 열의 길이는 1)이라고 가정한다. 여기서 정의한 디코더 g는 선형 변환함수다. z는 R^m 위의 원소로 원래 단어가 있던 차원보다 낮은 차원에 있으며, 단어들에서 추출된 특징 벡터로 볼 수 있다. 만약 이런 특징 벡터가 잘 추출된 것이라면 다음과 같은 복원 오차 $|x - g(f(x))|^2$가 매우 작을 것이다. 앞서 언급한 바와 같이 이 복원 오차를 최소화하는 f는 선형 변환으로 주어지며 구체적으로 최적의 인코더 $f(x) = D^T x$로 주어짐을 보일 수 있다. 한편 인코더와 디코더를 결정하는 행렬 D는 One hot encoding 벡터들로 만든 문장의 데이터 행렬에 Singular Value Decomposition을 적용하여 구할 수 있다(Jolliffe, 2011). 즉 전통적으로 사용하였던 주성분 분석이 단어의 특징 추출에 사용된다.

정리하면 주성분 분석을 이용한 단어 임베딩 방법은 문장 내에서 단어들의 동시 출현 빈도를 요약하여 특징을 추출하는 방법이라고 할 수 있다.

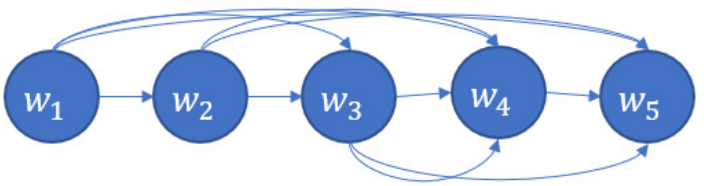
2. 신경망 모형을 이용한 주성분 분석의 확장

CBOW 방법은 단어가 가지고 있는 정보를 임의의 문장에서 다른 단어들이 주어졌을 때 한 단어의 출현 빈도의 확률로 정의하였다. 예를 들면 "나는 어제 미국에 가기 위해서 ○○공항으로 이동하였다"라는 문장에서 빈 곳에 들어갈 단어를 우리는 쉽게 유추할 수 있다. 왜냐하면 미국이라는 나라 이름이 나오고, 이동을 나타내는 동사, 그리고 공항이라는 단어 바로 옆에 붙어 있는 단어는 제한적이기 때문이다. 단어의 속성이 이 조건부 확률의 우도를 최대화하는 방향으로 속성이 요약되도록 인코딩 함수를 만드는 것이 CBOW 방법이다.

그렇다면 신경망 모형에서 다루고 있는 조건부 확률을 조금 더 자세히 살펴보자. 단어의 특성은 단어들 간의 연관성이며 이 연관성은 한 문장 내에서 출현 빈도의 조건부 확률로 정의된다. 어떤 문장은 단어들의 열로 표현되고 다음과 같이 $w_{1:T} = (w_{1,} \cdots, w_T)'(w_t \in V$는 문장의 t번째 위치에 나타난 단어)라 하겠다. 여기서 문장 내에서 i번째부터 j번째 위치에 있는 단어들의 부분열이다. 이 문장 $w_{1:T}$에 대한 생성모형을 단어들에 대한 조건부 확률을 이용하여 다음과 같이 모형화한다.

$$\Pr(w_{1:T}) = \prod_{t=2}^{T} \Pr(w_t | w_{1:(t-1)})$$

이 모형에서 문장 내 단어들의 분포는 바로 '시간에 흐름에 따라' 직전에 출현한 단어들의 영향을 받는 모형으로 가정하고 있다. 그래프에서 이를 표현하면 다음과 같을 것이다.



한 문장이 5개의 단어로 이루어져 있는 예이며, 시간상 후에 나타난 단어들은 과거에 출현한 단어들에 모두 영향을 받는 모형이다. 이 모형에서 단어들 간의 의존성에 대한 가정을 통해 단순한 다음과 같은 모형을 생각해볼 수 있다.

$$\Pr(w_{1:T}) = \prod_{t=2}^{T} \Pr(w_t | w_{(t-n+1):(t-1)})$$

즉 문장에서 한 단어의 분포는 주변 단어들에 의해서만 영향을 받는 모형이다. 가장 간단한 모형으로 마코프 체인을 생각해 볼 수 있다.

CBOW는 문장 생성모형이 아니라 주변 정보를 활용한 조건부 확률로 단어의 특징을 추출한다. CBOW 모형은 한 단어의 출현이 시간에 흐름에 따라 이전에 출현한 단어가 아니라 단지 주변 단어에 의해 영향을 받는다고 가정하였다. 즉 문장의 생성에서 시계열 모형으로부터 벗어난 것이 큰 특징이다. 반대로 Skip-gram 방법은 단어가 하나 주어졌을 때 문장 내 주변 단어들의 출현 확률을 이용하여 단어의 특징을 추출하는 방법이다. 기본적인 아이디어는 CBOW와 유사하지만 단어의 특징을 추출하는 방법에서 차이가 있다.

신경망 모형을 이용한 단어의 특징 추출 모형이 주성분 분석과 다른 점은 디코딩 단계에서 주성분 분석은 선형 변환을 그대로 이용하는 반면 신

경망 모형 방법은 단어 벡터를 선형 변환한 이후 결과값을 합이 1이 되는 확률값의 형태(표준 심플렉스 위의 원소)로 변환한다는 것이다. 이는 softmax function의 출력값으로 정의하고 이 값을 이용하여 multinomial의 우도를 통해 복원 오차를 정의한다.

주성분 분석과 신경망 모형을 이용한 단어 특징 추출의 차이점을 살펴보면 다음과 같다.

(1) 주성분 분석은 선형 디코더를 사용하는 반면 신경망 모형은 softmax function을 이용한 비선형 디코더 함수를 사용한다.
(2) 주성분 분석의 복원 오차는 L2 손실함수로 정의하지만, 신경망 모형은 multinomial 분포의 음의우도로 정의한다.
(3) 주성분 분석은 동시 출현 빈도에 기반하여 단어의 특징을 추출하지만, 신경망 모형은 조건부 확률을 통해 단어의 특징을 추출한다.

제4절 데이터 분석 방법

데이터를 문서-단어 행렬 혹은 단어의 특징 행렬로 변환한 뒤에 여러 분석 방법으로 결과를 도출할 수 있다. 텍스트 데이터와 같은 대규모의 데이터베이스에서 단어 간의 의미 있는 관계를 발견하기 위해 연관성 분석 방법, 단어 사이의 상관성을 파악하기 위해 Graphical lasso 모형을 사용할 수 있다.

1. 연관성 분석

연관성은 단어의 출현 빈도를 이용하여 분석한다. 연관성 분석은 변수 간의 관계를 발견하기 위한 규칙 기반의 기계 학습 방법으로 하나의 단어에 대해 다른 단어가 등장하는 사건을 여러 측도로 수치화하는 방법이다 (Agrawal, Imielinski, and Swami, 1993). 측도는 크게 지지도(support), 신뢰도(confidence), 향상도(lift)로 나눌 수 있다. 지지도는 특정 단어가 얼마나 많은 문서에 포함되어 있는지를 나타내는 측도이다. 신뢰도는 검색어를 이용하여 검색했을 때 특정한 단어가 등장할 확률을 나타내는 측도이며 조건부 확률과 같은 개념으로 단어들 간의 연관성을 나타낸다. 향상도는 대분류 키워드로 검색된 단어의 빈도와 일반적으로 등장할 단어의 빈도의 비율을 나타내는 측도이다.

연관성 분석에서 사용되는 측도를 수식으로 표현하기 위해서 다음의 표기를 사용한다. 모든 단어의 집합을 $W=\{w_1,\cdots,w_p\}$으로 표기하고, 총문서의 개수가 n개라고 하면, 하나의 문서는 단어의 집합의 부분집합이라고 할 수 있다. 즉 하나의 문서 s_i는 $s_i \subset W, i=1,\cdots,n$이다. 또한 모든 문서의 집합을 $S=\{s_i : 1 \le i \le n\}$로 표기하고 단어 $X \subset W$가 등장했을 때, $Y \subset W$가 등장하는 규칙을 $X \Rightarrow Y$으로 표기한다.

지지도는 특정 단어가 얼마나 많은 문서에 포함되어 있는지를 나타내는 측도로 특정 단어 $X \subset W$가 전체 문서에서 차지하는 비율이다.

$$supp(X)=\frac{|\{s_i \in S: X \subset s_i\}|}{|S|}$$

신뢰도는 검색어를 이용하여 검색했을 때 특정한 단어가 등장할 확률을 말하며 특정 단어 $X \subset W$로 검색했을 때 $Y \subset W$가 등장할 비율을 말한다. 즉 X라는 검색어로 검색했을 때 Y라는 단어가 검색될 확률이며 다음과 같이 표현할 수 있다.

$$conf(X \Rightarrow Y) = \frac{|\{s_i \in S : X \subset s_i, Y \subset s_i\}|}{|\{s_i \in S : X \subset s_i\}|}$$

마지막으로 향상도는 검색어를 통해 검색된 단어의 빈도와 일반적으로 등장할 단어의 빈도의 비를 의미한다. 다시 말해 향상도는 키워드의 일반적인 출현을 고려하여 동시 출현 확률의 특이성을 나타내는 측도라고 할 수 있다. $X \subset W$라는 검색어와 $Y \subset W$라는 단어에 대한 향상도는 다음과 같다.

$$lift(X \Rightarrow Y) = \frac{conf(X \Rightarrow Y)}{supp(Y)}$$

만약 Y가 흔히 사용되는 단어일 경우, 예를 들어 X와 Y가 독립사건일 경우 향상도의 값은 1에 가깝다. 만약 향상도가 큰 값을 가진다면 일반적으로 등장할 확률에 비해 X라는 검색어를 통해 검색될 확률이 크다는 것을 의미한다.

실제 데이터 분석에서는 생성된 문서-단어 행렬에서 각 단어의 신뢰도, 향상도를 계산한 후 신뢰도가 증가하며 향상도가 높은 단어를 선택하여 신기술과 관련된 단어라고 판단하는 방법으로 연관성 분석을 활용할 수 있다. 왜냐하면 시간이 흐름에 따라 신뢰도가 증가하는 단어는 해당 검색어와의 관련성이 높아지는 단어라고 생각할 수 있으며, 높은 향상도를 가지는 단어는 특정 단어가 일반적으로 검색될 빈도에 비해 해당 검색어를 입력했을 때 등장할 빈도의 비가 큰 단어라는 것을 의미하기 때문이다.

2. Graphical lasso 모형

다른 방법으로 Graphical lasso 모형을 사용하여 단어 간의 연관성을 분석할 수 있다. 이 모형에서는 단어의 출현 빈도뿐만 아니라 단어에서 추출된 특징을 이용해서 분석을 할 수 있다. Graphical lasso 모형은 Gaussian graphical 모형의 확장으로 Gaussian graphical 모형은 하나의 단어가 주어졌을 때 단어별로 다른 단어와의 상관성을 추정하는 모형이다(Friedman, Hastie and Tibshirani, 2008). 상관성은 여러 변수 중에서 두 변수에 대한 관계를 파악하는 분석 방법인 데 반해, Graphical lasso 모형은 다른 변수의 효과를 제거한 후 정의된 상관성인 편상관계수를 추정하는 방법이다. 예를 들어 단어 A-B-C는 서로 관계가 있는 단어임에도 A, C의 동시 출현은 오직 B에 의해 결정되는 경우를 쉽게 발견할 수 있다. 예를 들어서 '메디블록', '헬스메디코인', '블록체인' 세 단어는 관계가 큰 단어임에도 '블록체인'이라는 단어에 의해 생성되는 것으로 생각할 수 있다. Graphical lasso 모형은 하나의 단어에 대해 유의미한 단어를 찾아내는 것과 동시에 해당 단어 사이의 상관성도 파악할 수 있기 때문에 편리하게 사용할 수 있다.

분석을 수식으로 표현하기 위해 다음의 표기를 사용한다. 앞의 표기와 마찬가지로 모든 단어의 집합을 $W=\{w_1,\cdots,w_p\}$이라고 한다. 이때 각 단어 $w_1,\cdots,w_p$를 하나의 꼭짓점(vertex)으로, 단어 사이의 연결 관계를 변(link)으로 정의한다. 단어 사이의 연결 관계는 꼭짓점 W와 변 ε의 그래프 (W,ε)로 표현할 수 있다. Gaussian graphical 모형은 $(w_1,\cdots,w_p)^T \sim N(\mu,\Omega^{-1})$이라는 가정하에 반응변수를 w_j, 설명변수를 w_i, for $i \in \{1,\cdots,p\}-\{j\}$인 회귀 모형을 적용하고, w_k에 대응

되는 최소제곱추정량(Least square estimator) β_k^{-j}를 계산한다. 추정된 회귀계수에 대해 $\beta_k^{-j} = 0 \Leftrightarrow (\Omega)_{jk} = 0$임이 알려져 있다. 즉 Gaussian graphical 모형은 단어 간의 상관관계를 회귀모형의 계수가 0인지 0이 아닌지를 이용하여 추정하는 모형이라고 할 수 있다.

Graphical lasso 모형은 Gaussian graphical 모형에서 회귀계수를 정확하게 0으로 줄 수 있는 분석 방법이다. Lasso 벌점 함수를 이용하여 w_k에 대응되는 회귀계수 β_k^{-j}를 추정할 때 계수를 0으로 추정할 수 있다. 회귀계수는 고정된 $\lambda > 0$과 모든 j에 대해 아래의 식을 이용하여 추정한다.

$$\widehat{\beta^{-j}}(\lambda) = \arg min_{\beta^{-j}} \sum_{i=1}^{n} (w_{ij} - \beta_0^{-j} - \sum_{k \neq j} \beta_k^{-j} w_{ik})^2 + \lambda \sum_{k \neq j} |\beta_k^{-j}|$$

3. 두 기법의 비교분석

Graphical lasso 모형을 이용한 기법은 단어 사이의 편상관계수를 추정하여 계수를 이용하는 방법이며 연관성 분석 방법은 단어 사이의 상관계수를 이용하는 방법이다. 또한 Graphical lasso 기법은 모형의 회귀계수를 활용하여 단어를 도출하는 반면 연관성 분석 기법은 단어의 빈도를 이용하여 직접적으로 단어를 도출한다. 그리고 연관성 분석은 관련된 측도의 기준에 따라 단어를 선택하는 반면 Graphical lasso 기법은 벌점 상수에 따라 회귀계수의 크기가 변한다는 차이점이 있다.

제 4 장

보건복지와 블록체인 키워드 분석

4 보건복지와 블록체인 키워드 분석

제1절 텍스트 수집 및 자료 처리

전체적인 분석 과정에서 텍스트를 수집하는 과정은 양질의 분석을 위해 중요하다. 여기서 웹에 게시되어 있는 모든 데이터를 웹 데이터라고 정의한다. 예를 들어 분석에서는 보건·복지 분야의 블록체인 키워드를 포함하는 뉴스, 블로그, 카페 등의 글을 웹 데이터라고 정의할 수 있다. 웹 데이터를 수집하는 과정은 웹 스크래핑(web scraping) 혹은 웹 크롤링(web crawling)이라고 부르며, 여러 프로그램을 사용하여 웹 스크래핑 도구를 개발할 수 있다(Munzert, et al., 2014).

웹 스크래핑 도구는 웹 관련 언어인 XML 혹은 html 등의 언어로 작성된 문서를 정해진 규칙에 따라 수집하는 기술이다. 문서 내의 구조적인 정보를 이용하여 텍스트를 추출할 수 있으며 파이썬(Python) 혹은 R 기반의 오픈소스 패키지를 이용하여 수행할 수 있다. 수집하려는 웹페이지의 특성에 맞도록 세부적인 프로그램 작성이 필요하기 때문에 다양한 페이지에서 웹 데이터를 수집하기 위해 코드 작성에 충분한 시간이 필요하다. 예를 들면 네이버에서 제공하는 뉴스들 중 네이버에서 편집되어 나온 텍스트들은 몇 개의 템플릿을 따라 변환되어 있기 때문에 통일된 방법으로 수집이 가능하다. 반면 각 인터넷 신문사들이 웹에 게시한 뉴스들은 각기 다른 템플릿으로 작성되었으며, 그 문서에 접근하는 방법 역시 다르기 때문에 문서가 저장된 자료원에 따라 맞춤형 프로그램 작성이 필요하다. 이는 자료원 수집 비용을 증가시키게 된다.

한편 스크래핑이 가능한 웹페이지는 오픈 API(Application Programming Interface) 제공 여부에 따라 텍스트 수집 프로그래밍 방법이 달라진다. API는 웹페이지의 핵심 기능을 외부 사이트에서 활용할 수 있도록 공개된 인터페이스를 말한다. 일반적으로 API를 제공하는 웹페이지인 경우 프로그래밍 작성에 필요한 비용이 줄어든다. 예를 들어 IEEE에서 제공하는 논문 정보 검색 API는 200여 개의 저널에 대한 검색 기능을 제공하여, 검색 프로그램을 작성해야 하는 수고를 줄일 수 있다. 그러나 API를 제공하는 웹페이지더라도 서비스별로 요구되는 변수가 다르기 때문에 각각의 페이지에 대한 맞춤형 프로그래밍은 여전히 필요하다.

이 연구에서는 R의 httr과 rvest 패키지를 활용하여 데이터를 수집하였다. 여기서 선정한 자료원은 네이버 카페, 블로그, 뉴스, 학술정보와 arxiv에서 검색된 논문의 초록, BBC, CNN, ITNews, RISS 홈페이지다. 웹사이트별로 다른 웹 스크래핑 프로그래밍을 하였으며, 사이트별로 다른 검색어를 입력하여 데이터를 수집했다.

분석된 문서의 개수와 이용한 검색어는 〈표 4-1〉과 같다. 자료 수집 기간은 2015년 1월 1일부터 2018년 5월 31일이며, 입력한 검색어와 관련된 전체 1만 801건의 문서 중 보안과 코드 오류 등의 이유로 1291건의 문서는 내용을 가져오지 못했으며 한글 문서는 8360건, 영문 문서는 1150건으로 나뉘며 한글, 영어에 따라 다른 과정을 거쳐 문서 분석을 진행했다.

〈표 4-1〉 자료 출처별 개수 및 검색어

(단위: 건)

출처	문서 개수	검색어
네이버 학술검색	377	blockchain
네이버 블로그	742	블록체인+건강-비트코인
	683	블록체인+보건-비트코인
	818	블록체인+복지-비트코인
	747	블록체인+의료-비트코인
	747	블록체인+헬스-비트코인
네이버 카페	920	블록체인+건강-비트코인
	355	블록체인+보건-비트코인
	517	블록체인+복지-비트코인
	970	블록체인+의료-비트코인
	1120	블록체인+헬스-비트코인
네이버 뉴스	782	블록체인+건강-비트코인
	357	블록체인+보건-비트코인
	571	블록체인+복지-비트코인
	812	블록체인+의료-비트코인
	842	블록체인+헬스-비트코인
arxiv	372	blockchain
BBC	123	blockchain
CNN	106	blockchain
ITNews	143	blockchain
RISS	1680	블록체인

주: 네이버 검색어에서 블록체인 +건강 -비트코인 등의 검색어는 블록체인과 건강을 포함한 문서 중에서 비트코인이라는 단어를 포함한 문서를 제외한 결과를 의미함.

제2절 텍스트 정제 및 결과 시각화 방법

1. 불용어 처리

수집한 문서 데이터를 분석 목적에 맞는 형태로 변형하기 위해 정제 과정을 거친다. 정제 과정은 일반적으로 수집된 문서 데이터에서 자주 사용하는 단어 혹은 분석 결과에 의미가 없는 단어인 불용어를 처리하는 것이다. 예를 들어 '해서, 은, 는, 에, 에서, 를, 으로, 입니다, 합니다, 하기, 위한, 한다, 하다, 것인가, 겠다, 냐, 되, 됨, 할, 한, 을, 까, 어찌, 있, 려면, 뭔지, 씩' 등의 단어는 분석에서 필요 없는 단어이며 실제 분석에서도 이러한 패턴을 가진 단어는 제거하고 사용하였다.

2. 단어 변환

분석에서는 한글 문서의 경우 'KoNLP' 패키지의 'Woorimalsam', 'Insighter', 'Sejong'의 107만 개의 명사와 추가로 네이버 백과사전에서 블록체인과 관련된 245개의 단어 및 사전에 정의되어 있지 않은 단어, '헬스메디블록', '헬스메디', '메디블록', '헬스코인', '헬스케어', '메디토큰', '암호화폐', '법정화폐', '텔레헬스', '예방보건', '텐센트', '헬스메디코인', '공공의료', '의료비', '전자화폐'를 추가하여 새로운 사전을 정의했다. 영문 문서의 경우 띄어쓰기 기준으로 단어를 나누어 해당 단어를 기준으로 앞의 2개 단어, 뒤의 2개 단어 총 5개까지의 단어를 하나의 단어로 여기는 5-gram 모형을 사용했다.

3. 결과 시각화 방법

앞서 소개한 텍스트마이닝 분석은 단어와 단어 사이의 상관관계를 파악하는 데 집중한다. 즉 분석 결과를 통해 단어와 단어 사이의 특정한 관계 유무를 파악할 수 있다. 따라서 이러한 분석 결과를 시각화하기 위해서는 단어와 단어 사이의 관계를 쉽게 도식화할 수 있는 방법이 필요하다. 변수와 변수 사이의 관계를 시각화할 때는 네트워크 그림이 많이 활용되며 텍스트마이닝 분석 결과에서도 마찬가지로 네트워크 그림으로 단어와 단어 사이의 관계를 표현해줄 수 있다. 예를 들어 네트워크 그림으로 만약 '블록체인'과 '헬스케어' 단어 사이에 연관성이 있다면 두 단어를 이어줄 수 있으며 전체 네트워크 그림으로 중심이 되는 단어를 추론해볼 수 있다는 장점이 있다.

분석에서는 계산된 문서-단어 행렬에서 단어가 출현했으면 1, 아니면 0의 값을 할당하고 문서-단어 행렬의 내적을 통해 단어와 단어 사이의 관계를 빈도로 확인할 수 있는 co-occurrence 행렬을 생성했다. Co-occurrence 행렬의 행과 열은 모두 단어로 구성되어 있으며 행렬 내부의 (i,j) 원소는 I번째 단어와 j번째 단어가 함께 등장한 빈도를, 대각 원소는 각 단어의 빈도수를 나타내는 행렬이다. Co-occurrence 행렬을 R 프로그램의 네트워크(Network)D3 패키지를 이용하여 결과를 시각화했으며, 그중 코드네트워크(chordNetwork)와 생키네트워크(sankeyNetwork)를 활용하였다(Gandrud, et al., 2015). 그리고 방대한 문서 자료에서 맥락과 관련된 단서들을 이용하여 해석 가능성이 높은 주제들을 추출하기 위해 토픽 모형 분석을 실시하였다. 모형과 관련한 설명은 오미애 등(2017)을 참고하기 바란다.

제3절 한글 문서 분석 결과

먼저 한글 문서에서 등장 빈도가 높은 20개의 단어는 〈표 4-2〉와 같다. 등장 빈도가 높은 단어에는 분석에 필요한 특정 기술이 포함되어 있지 않으며 일반적으로 사용되는 단어가 포함되어 있기 때문에 키워드를 포함하는 단어를 중심으로 그 빈도수를 파악할 수 있다.

〈표 4-2〉 한글 문서에서 등장 빈도가 높은 20개의 단어

단어	빈도수
블록체인	10,162
기술	7,329
서비스	5,441
개발	5,332
관련	5,164
기반	5,132
기업	5,095
정보	4,729
시장	4,432
분야	4,263
플랫폼	4,056
사업	3,991
시스템	3,960
이상	3,812
기능	3,734
기자	3,655
제공	3,595
국내	3,568
활용	3,496
계획	3,447

블록체인 키워드는 기술, 서비스, 개발, 관련, 기반, 기업, 정보, 시장, 분야, 플랫폼, 사업, 시스템 등 블록체인 인프라와 서비스 관련 이슈와 밀접한 관련이 있다. 여기에서는 블록체인과 관련된 보건 및 복지 이슈를 세부적으로 살펴보기 위해 특정 키워드가 포함된 단어들을 따로 산출하였다. 키워드는 '보건', '헬스', '건강', '의료', '메디', '복지'로 키워드 별로 등장하는 단어의 빈도는 〈표 4-3〉~〈표 4-7〉에 제시하였다.

블록체인과 관련된 이슈에서 '보건'이 포함된 단어를 연도별로 살펴보면 '보건의료' 키워드가 2018년도에 상대적으로 많이 언급되었음을 알 수 있다. 전반적으로 2015~2016년도에 비해 2017년도에 블록체인 활용에 대한 이슈화가 되었고 2018년도에 본격적인 논의가 시작되었다고 볼 수 있다.

〈표 4-3〉 연도별 '보건' 키워드를 포함하고 있는 단어

단어	2015년	2016년	2017년	2018년	연도 불명	전체
보건의료	6	7	52	150	0	215
보건소	12	8	8	73	0	101
보건	9	0	2	36	0	47
보건산업	0	2	14	21	0	37
안전보건	17	3	1	14	0	35
보건당국	4	2	3	23	0	32
안전보건공단	1	2	5	13	0	21
보건의료산업	0	1	1	18	0	20
보건행정학과	0	0	3	17	0	20
산업보건	4	14	0	2	0	20
보건대학원	0	0	7	12	0	19
의료보건	1	0	3	14	0	18
보건의료 분야	0	0	1	17	0	18
보건교사	2	0	0	2	12	16
보건교육	9	0	0	6	0	15

단어	2015년	2016년	2017년	2018년	연도 불명	전체
보건과	8	0	3	3	0	14
보건부	2	0	4	6	0	12
보건용	0	0	0	11	0	11

블록체인과 관련된 이슈에서 '헬스'가 들어간 단어를 살펴보면 연도별 경향성은 '보건' 키워드와 동일하다. 다른 연도에 비해 2018년도에 '헬스'를 포함한 키워드가 전반적으로 많이 언급되었으며 블록체인과 함께 '헬스케어' 키워드가 가장 많이 언급되었음을 알 수 있다. '헬스메디' 키워드는 헬스케어를 기반으로 블록체인 플랫폼을 서비스하는 곳으로 비즈니스 산업과 관련된 키워드도 많이 언급되고 있음을 확인하였다.

〈표 4-4〉 연도별 '헬스' 키워드를 포함하고 있는 단어

단어	2015년	2016년	2017년	2018년	연도 불명	전체
헬스케어	69	59	367	1,767	1	2,263
헬스	80	12	82	303	9	486
헬스메디	0	0	0	123	0	123
바이오헬스	1	1	22	71	0	95
헬스클럽	31	1	0	22	0	54
헬스장	18	0	1	18	0	37
헬스혈맹원	23	0	0	0	0	23
헬스산업	0	0	2	18	0	20
헬스커넥트	0	0	5	7	0	12
헬스메디코인	0	0	0	10	0	10
헬스메디달러	0	0	0	9	0	9
헬스테크	0	0	0	9	0	9

블록체인과 관련된 이슈에서 '건강'이 포함된 키워드로는 '건강', '건강관리', '건강보험'의 빈도가 높게 나타났다.

〈표 4-5〉 연도별 '건강' 키워드를 포함하고 있는 단어

단어	2015년	2016년	2017년	2018년	연도 불명	전체
건강	146	31	279	1,944	5	2,405
건강관리	6	3	81	345	0	435
건강보험	5	2	40	195	0	242
건강정보	0	3	13	197	0	213
생활건강	14	3	11	119	0	147
건강검진	0	1	16	110	0	127
건강상태	9	2	10	76	0	97
정신건강	7	1	32	31	0	71
건강식품	1	0	16	50	0	67
건강보험료	0	0	7	59	0	66
국민건강	5	3	0	47	0	55
건강증진	7	0	7	35	0	49
건강상담	0	0	0	42	0	42
건강도	42	0	0	0	0	42
건강권	0	0	1	33	0	34
건강증진형	0	0	0	29	0	29
건강공동체	0	0	0	24	0	24
국민건강보험	0	1	4	16	0	21
건강음료	14	0	0	5	2	21
건강상	6	1	0	14	0	21

블록체인과 관련된 이슈에서 '메디'가 포함된 키워드로는 '메디블록', '헬스메디', '메디컬', '메디톡스'의 빈도가 상대적으로 높았으며, 보건과 관련된 블록체인 비즈니스와 밀접한 키워드가 많이 도출되었다.

〈표 4-6〉 연도별 '메디' 키워드를 포함하고 있는 단어

단어	2015년	2016년	2017년	2018년	연도 불명	전체
메디블록	0	0	62	276	0	338
메디컬	6	2	17	108	0	133
헬스메디	0	0	0	123	0	123
메디톡스	0	1	10	69	0	80
메디포스트	1	0	7	31	0	39
메디프론	0	0	3	34	0	37
메디토큰	0	0	6	13	0	19
세종메디칼	0	0	0	17	0	17
보령메디앙스	2	1	3	11	0	17
메디컬그룹	0	0	0	16	0	16
메디슨	1	0	2	12	0	15
세운메디칼	0	0	2	12	0	14
메디칼	4	0	3	6	0	13
메디플란트	0	0	0	12	0	12
메디칼업저버	0	0	0	11	0	11
메디파나뉴스	0	0	0	11	0	11
헬스메디코인	0	0	0	10	0	10

블록체인과 관련된 이슈에서 '복지' 키워드가 포함된 문서는 보건 쪽 관련 문서에 비해 상대적으로 적었다. 전체적으로 빈도수가 높은 키워드는 '복지', '사회복지', '사회복지사', '복지관', '공공복지' 순이었다. 복지 역시 보건과 마찬가지로 연도별로 비교해 보았을 때, 2018년도에 관련 문서가 월등히 많음을 볼 수 있다.

〈표 4-7〉 연도별 '복지' 키워드를 포함하고 있는 단어

단어	2015년	2016년	2017년	2018년	연도 불명	전체
복지	41	25	185	1,453	0	1,704
사회복지	11	0	9	73	0	93
사회복지사	22	0	1	13	0	36
복지관	2	1	0	23	0	26
공공복지	9	0	2	15	0	26
사회복지과	0	5	0	7	0	12
행복지수	0	0	1	7	0	8
복지법	0	1	0	7	0	8

한글 문서에서 전체적으로 빈도수가 높은 키워드를 20여 개 살펴본 결과 블록체인과 관련된 키워드로 '헬스케어', '암호화폐'가 가장 연관성이 높았으며, '클라우드' 등의 인프라, '메디블록'과 같은 비즈니스 산업, '환자정보', '스마트 헬스'와 같은 보건 관련 이슈들이 많이 언급되었다.

〈표 4-8〉 한글 문서에서 도출된 단어

단어	2015년	2016년	2017년	2018년	연도 불명	전체
헬스케어	69	59	367	1,767	1	2,263
암호화폐	0	2	68	1,619	0	1,689
클라우드	20	32	253	872	0	1,177
사물인터넷	4	46	188	486	0	724
의료정보	0	0	117	497	0	614
메디블록	0	0	62	276	0	338
의료비	23	2	18	292	0	335
스마트계약	0	5	21	120	0	146
헬스메디	0	0	0	123	0	123
휴먼스케이프	0	0	0	100	0	100
알파콘	0	0	0	96	0	96
원격의료	0	0	8	73	0	81

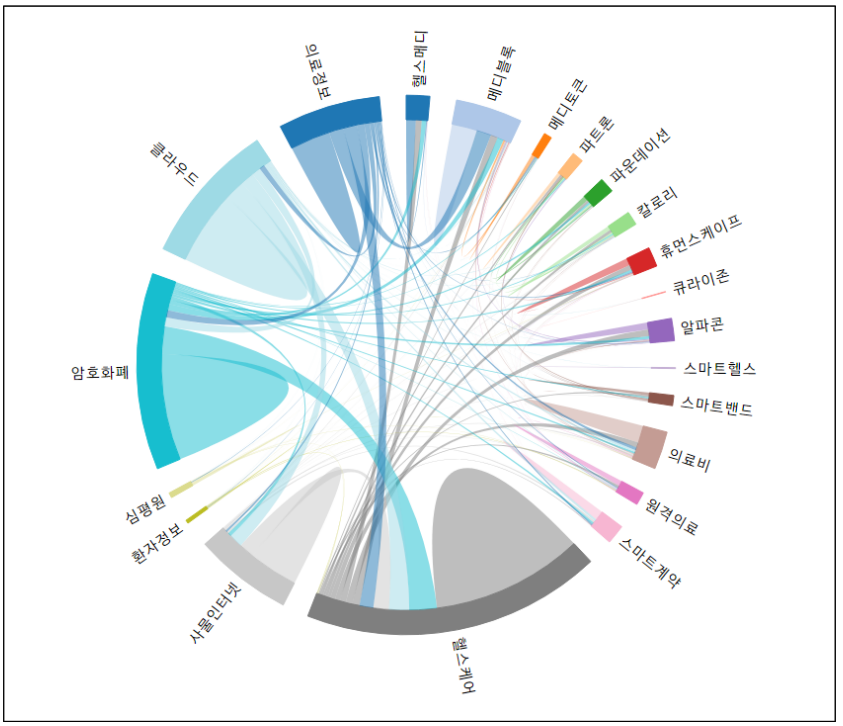
단어	2015년	2016년	2017년	2018년	연도 불명	전체
칼로리	9	11	4	56	0	80
파운데이션	6	0	1	63	4	74
파트론	2	0	1	39	0	42
심평원	0	0	9	33	0	42
스마트밴드	0	0	5	34	0	39
메디토큰	0	0	6	13	0	19
환자정보	0	0	1	18	0	19
큐라이존	0	0	0	8	0	8
스마트헬스	0	1	5	2	0	8

한글 문서에서 위의 특정 단어들을 포함하고 있는 문서들을 가지고 상위 20여 개 키워드의 네크워크 구조를 시각화한 결과는 다음과 같다.

Chord network에서 각 선의 굵기는 연결 강도를, 선의 개수는 얼마나 많은 단어들과 연결되어 있는가를 보여준다.

'헬스케어' 키워드는 가장 많이 언급되기는 했지만 헬스와 관련된 키워드에 국한하여 연결되어 있다. 반면 '암호화폐' 키워드는 모든 단어들과 함께 언급되었음을 알 수 있다.

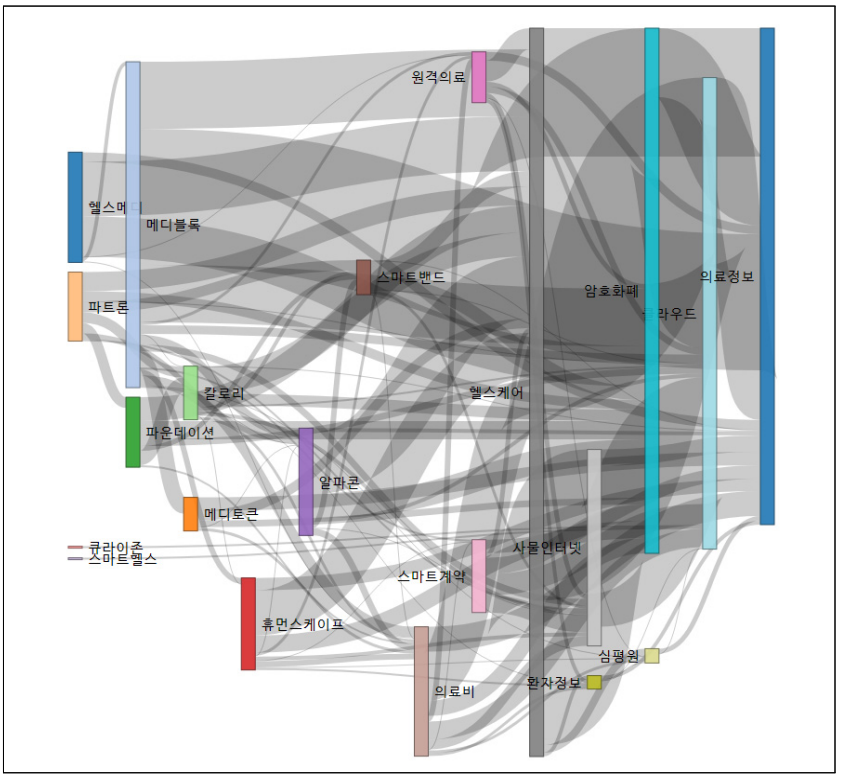
〔그림 4-1〕 한글 문서의 네트워크 구조(Chord network)



Sankey network의 경우 막대의 길이는 단어 빈도를, 검은 선들의 흐름은 연결의 흐름과 중요도를 표현하는 것으로 검은 선들의 굵기가 연결의 중요도를 의미한다.

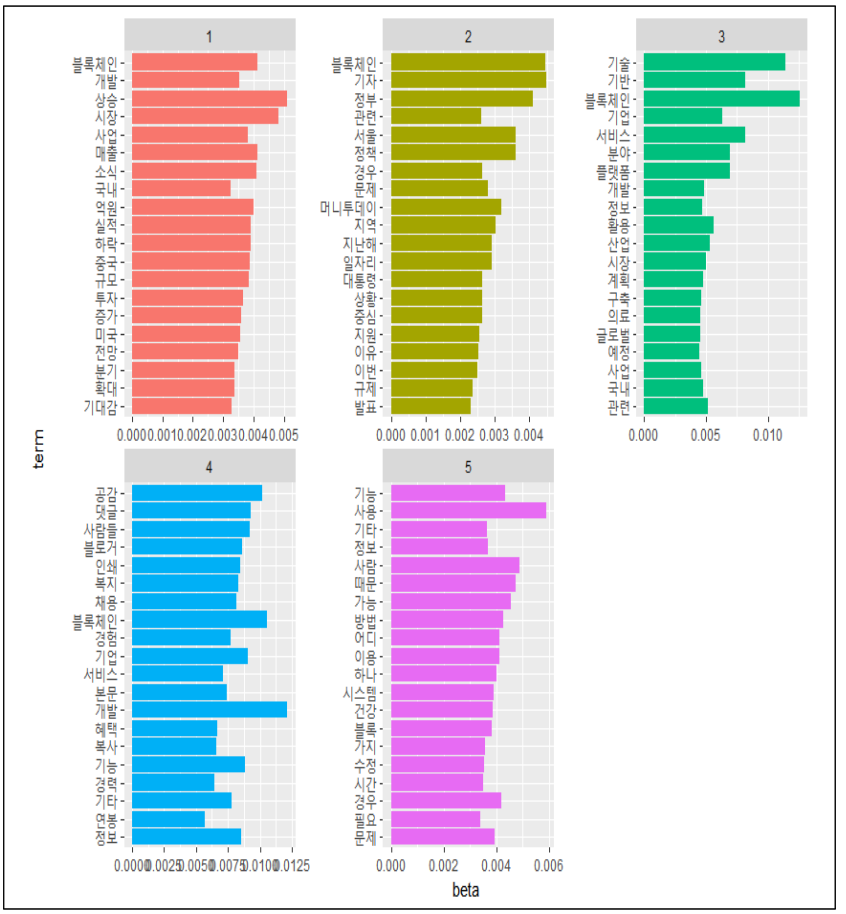
의료와 관련된 키워드는 서로 밀접한 연관성이 있음을 아래의 그림으로 확인할 수 있다.

[그림 4-2] 한글 문서의 네트워크 구조(Sankey network)



토픽 모형은 비정형화된 문서 집단에서 관측할 수 없는 토픽을 문서 내 단어의 분포를 통해 찾으려는 통계 모형이다. 한글 문서의 dtm으로 토픽의 개수를 4, 5, 6개로 적용해 보았고, 그중 가장 적합하다고 판단되는 5개의 토픽 결과를 제시하였다. 각 토픽에서 상위 20개 단어의 분포를 보면 첫 번째 토픽은 블록체인과 관련된 국내외 투자, 개발과 관련한 전망이고, 두 번째 토픽은 블록체인과 관련된 정부 정책 이슈이다. 세 번째 토픽은 블록체인 기술 기반의 의료 관련 플랫폼 이슈, 네 번째 토픽은 블록체인과 관련한 복지, 기업 이슈, 다섯 번째 토픽은 기타로 분류할 수 있다.

[그림 4-3] 한글 문서의 토픽 모형



제4절 영문 문서 분석 결과

영문 문서에서 등장 빈도가 높은 20개의 단어는 다음과 같다. 영문 문서에서는 '비트코인' 키워드를 제거하지 않았기 때문에 빈도 2순위로 많이 언급되었고 데이터, 기술, 디지털, 시스템, 보안, 사업 등의 단어가 블

록체인 키워드와 함께 많이 언급되었다.

〈표 4-9〉 영문 문서에서 등장 빈도가 높은 20개의 단어

단어	빈도수
blockchain	3,484
bitcoin	1,877
data	1,550
technology	1,461
based	1,007
digital	945
system	917
security	781
business	714
network	689
time	655
currency	614
transactions	590
information	563
people	558
company	541
financial	532
world	524
companies	513
distributed	508

한글 문서와 마찬가지로 특정 키워드를 포함하는 단어의 빈도수는 별도로 산출하였다. 이때 사용된 키워드는 이 연구의 관심 분야인 'health', 'welfare', 'medical'이다. 각각의 키워드에 대한 단어의 출현 빈도는 〈표 4-9〉~〈표 4-11〉에 제시하였다.

연도별로 특정 키워드를 포함하는 단어의 빈도수를 비교해 보면, 2017년도와 2018년도에 언급된 정도가 거의 비슷함을 알 수 있다. 이는 한글

문서가 2018년도에 집중된 것과는 다른 결과이며, 블록체인과 관련된 이슈는 해외에서 먼저 많이 언급되었음을 알 수 있다.

〈표 4-10〉 연도별 'health' 키워드를 포함하고 있는 단어

단어	2015년	2016년	2017년	2018년	연도 불명	전체
healthcare	0	48	48	58	4	158
health	17	3	63	50	1	134
health care	2	1	16	5	0	24
healthcare data	0	18	2	2	0	22
health records	0	0	9	7	0	16
healthcare providers	0	1	10	2	0	13
biomedical health care	0	0	12	0	0	12
biomedical health	0	0	12	0	0	12
health data	0	0	6	4	0	10
healthcare system	0	4	2	1	1	8
health checkup	0	0	8	0	0	8
health information	0	0	3	4	0	7
healthcare industry	0	1	2	2	2	7
healthy	2	1	1	3	0	7
ibm watson health	0	0	4	3	0	7
watson health	0	0	4	3	0	7
electronic health	0	0	1	5	0	6
biomedical health care domains	0	0	6	0	0	6
health care domains	0	0	6	0	0	6
healthcare systems	0	6	0	0	0	6
electronic health records	0	0	1	4	0	5
blockchain healthcare	0	1	3	1	0	5
public health	2	0	2	1	0	5
healthcare orgaizations	0	0	0	5	0	5
personal health	0	0	2	2	1	5

〈표 4-11〉 연도별 'welfare' 키워드를 포함하고 있는 단어

단어	2015년	2016년	2017년	2018년	연도 불명	전체
welfare	9	3	4	4	0	20
social welfare	4	0	4	3	0	11

블록체인과 관련된 복지(welfare) 키워드가 포함된 문서는 많지 않음을 확인할 수 있다.

〈표 4-12〉 연도별 'medical' 키워드를 포함하고 있는 단어

단어	2015년	2016년	2017년	2018년	연도 불명	전체
medical	63	2	45	46	0	156
biomedical	1	0	13	5	0	19
medical records	1	1	6	10	0	18
medical service	17	0	1	0	0	18
medical data	0	0	6	7	0	13
biomedical health care	0	0	12	0	0	12
biomedical health	0	0	12	0	0	12
medical inforamtion	2	0	4	3	0	9
electronic medical records	0	0	3	4	0	7
electronic medical	0	0	3	4	0	7
medical industry	5	0	2	0	0	7
biomeical health care domains	0	0	6	0	0	6
biomedical research	0	0	1	4	0	5
electronic medical records EMRs	0	0	3	2	0	5
medicarl records EMRs	0	0	3	2	0	5
medical center	1	0	0	4	0	5

한글 문서의 분석과 마찬가지로 선택된 특정 단어의 빈도 및 단어 간의 네트워크 구조를 시각화한 결과는 다음과 같다. 여기에서도 보건 이슈인 'healthcare', 'medical' 키워드가 많이 언급되었으며, 복지와 관련된 키워드는 거의 없음을 볼 수 있다.

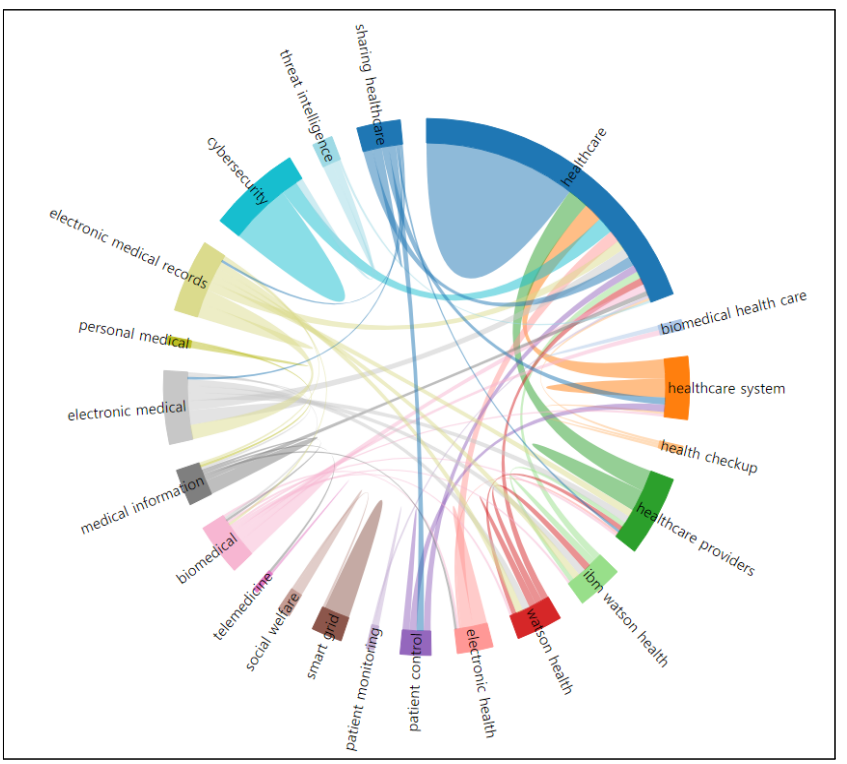
〈표 4-13〉 영문 문서에서 도출된 단어

단어	2015년	2016년	2017년	2018년	연도 불명	전체
healthcare	0	48	48	58	4	158
cybersecurity	1	14	9	34	0	58
biomedical	1	0	13	5	0	19
smart grid	1	0	5	10	0	16
healthcare providers	0	1	10	2	0	13
biomedical health care	0	0	12	0	0	12
social welfare	4	0	4	3	0	11
medical information	2	0	4	3	0	9
healthcare system	0	4	2	1	1	8
health checkup	0	0	8	0	0	8
ibm watson health	0	0	4	3	0	7
watson health	0	0	4	3	0	7
electronic medical	0	0	3	4	0	7
electronic medical records	0	0	3	4	0	7
electronic health	0	0	1	5	0	6
patient control	0	6	0	0	0	6
threat intelligence	3	2	0	1	0	6
patient monitoring	0	0	0	4	0	4
telemedicine	4	0	0	0	0	4
sharing healthcare	0	3	1	0	0	4
personal medical	0	0	2	1	0	3

영문 문서에서 Chord network로 단어의 연결 강도 및 연결 정도를 살펴보았다.

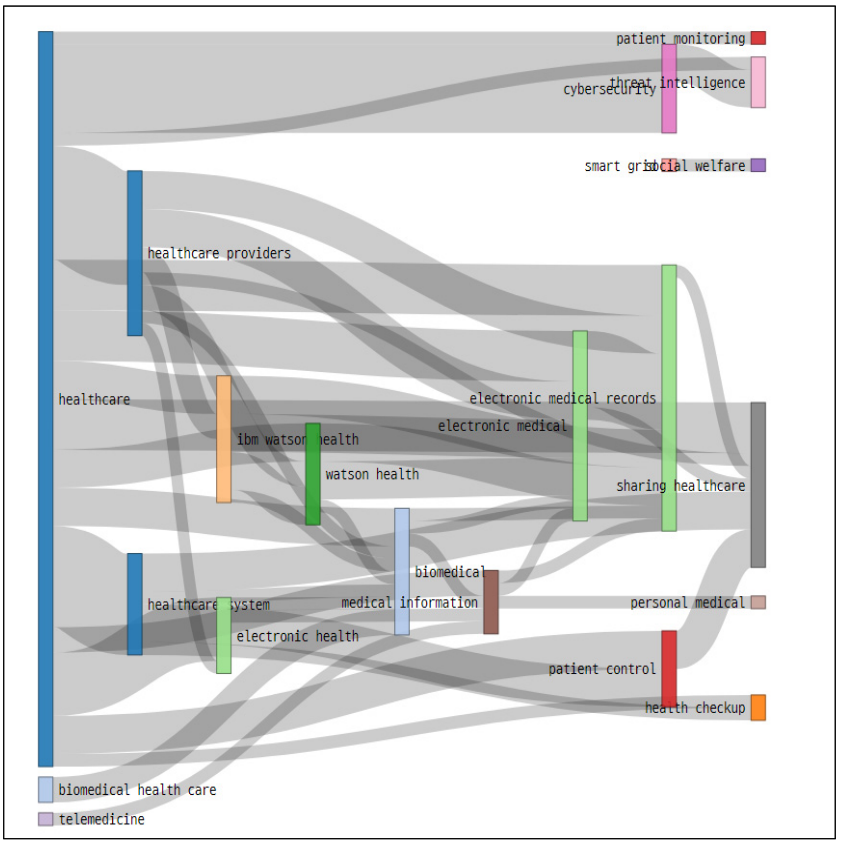
'healthcare' 키워드가 가장 많이 언급되었고, 상위 빈도의 단어들과 함께 언급된 정도도 높음을 알 수 있다. 반면 'social welfare' 키워드는 'smart grid'를 제외하고 다른 단어와의 연결성이 없었다.

[그림 4-4] 영문 문서의 네트워크 구조(Chord network)



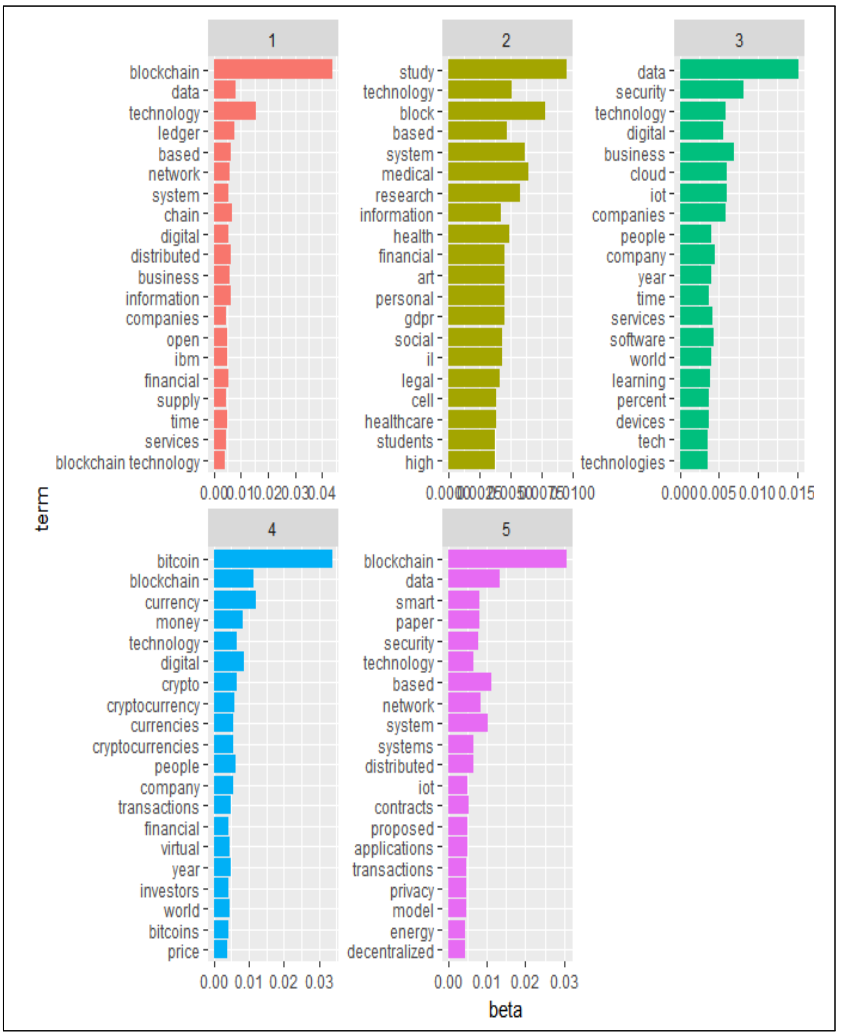
영문 문서에서 Sankey network로 단어 빈도 및 단어 간 연결의 중요도를 살펴보면 Chord network 결과와 마찬가지로 'healthcare' 키워드가 다른 키워드들과 서로 밀접하게 연결되어 있음을 알 수 있다.

〔그림 4-5〕 영문 문서의 네트워크 구조(Sankey network)



토픽 모형은 한글 문서와 마찬가지로 영문 문서의 dtm으로 토픽 개수를 4, 5, 6개로 적용해보았고, 그중 가장 적합하다고 판단되는 5개의 토픽 결과를 제시하였다. 각 토픽에서 상위 20개 단어의 분포를 보면 첫 번째 토픽은 블록체인과 관련된 기업이고, 두 번째 토픽은 블록체인과 관련된 의료, 보건 이슈이다. 세 번째 토픽은 블록체인의 클라우드, IoT 기술 이슈, 네 번째 토픽은 블록체인과 관련한 가상화폐 이슈, 다섯 번째 토픽은 블록체인 자체의 기술 이슈로 분류할 수 있다.

〔그림 4-6〕 영문 문서의 토픽 모형



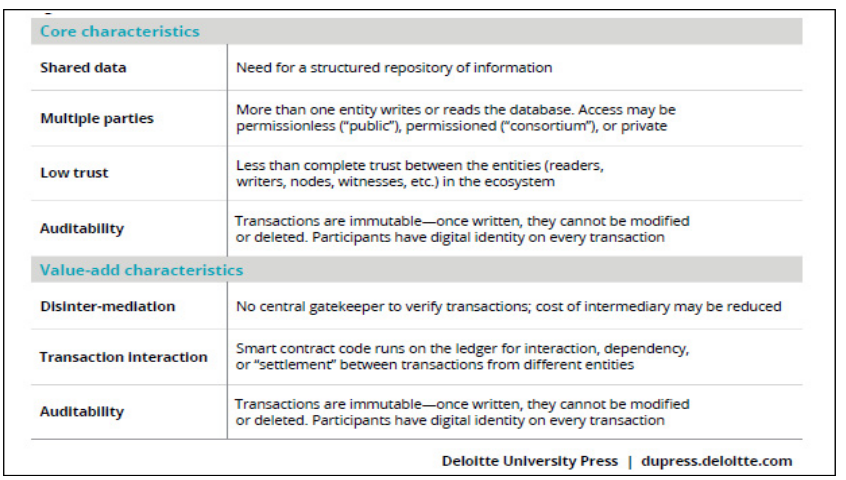
제 5 장 결론

5 결론

이 연구에서는 4차 산업혁명 시대의 신기술이라고 할 수 있는 블록체인과 관련하여 보건복지 분야의 이슈를 살펴보고자 하였다.

블록체인 기술이 공공 부문에서 유용할 수 있는 상황은 공유데이터, 다양한 이해관계자, 낮은 신뢰도, 감사 기능의 특성이 있는 경우이다.

〔그림 5-1〕 블록체인 기술이 유용한 경우



Core characteristics	
Shared data	Need for a structured repository of information
Multiple parties	More than one entity writes or reads the database. Access may be permissionless ("public"), permissioned ("consortium"), or private
Low trust	Less than complete trust between the entities (readers, writers, nodes, witnesses, etc.) in the ecosystem
Auditability	Transactions are immutable—once written, they cannot be modified or deleted. Participants have digital identity on every transaction
Value-add characteristics	
Disinter-mediation	No central gatekeeper to verify transactions; cost of intermediary may be reduced
Transaction interaction	Smart contract code runs on the ledger for interaction, dependency, or "settlement" between transactions from different entities
Auditability	Transactions are immutable—once written, they cannot be modified or deleted. Participants have digital identity on every transaction

Deloitte University Press | dupress.deloitte.com

자료: Will blockchain transform the public sector? Blockchain basics for government, deloitte.com, 11page

블록체인은 암호화폐와 함께 이슈화되었지만, 블록체인 기술 자체로 해마다 발전하고 있으며, 민간 및 정부에서의 활용 사례도 증가하고 있다. 여기에서는 보건·복지 분야에서 블록체인 기술이 어떻게 활용되고 있는지 알아보기 위해 블록체인 키워드와 보건 및 복지 키워드와 관련된 문

서를 함께 수집하였다. 그리고 한글 문서에서의 이슈와 영문 문서에서의 이슈 차이도 살펴보기 위해 한글 문서 및 영문 문서를 각각 분석하였다. 그 결과 한글 문서에서는 2015~2017년도에 비해 2018년도에 블록체인과 관련된 문서가 상대적으로 많이 수집되었고, 영문 문서의 경우 2015~2016년도에 비해 2017~2018년도에 블록체인과 관련된 보건 이슈가 증가하였음을 알 수 있다. 이는 해외에 비해 우리나라에서 블록체인과 관련된 상황이 상대적으로 늦게 이슈화되었다고 해석할 수 있다.

수집된 문서에서 복지 쪽 이슈는 거의 없었으며, 대부분이 보건 쪽 이슈와 관련이 있었다.

한글 문서의 5개 토픽을 살펴보면 첫 번째 토픽은 블록체인과 관련된 국내외 투자, 개발 관련 전망이고, 두 번째 토픽은 블록체인과 관련된 정부 정책 이슈이다. 세 번째 토픽은 블록체인 기술 기반의 의료 관련 플랫폼 이슈, 네 번째 토픽은 블록체인과 관련한 복지, 기업 이슈, 다섯 번째 토픽은 기타로 분류할 수 있다. 영문 문서의 5개 토픽의 경우 첫 번째 토픽은 블록체인과 관련된 기업이고, 두 번째 토픽은 블록체인과 관련된 의료, 보건 이슈이다. 세 번째 토픽은 블록체인의 클라우드, IoT 기술 이슈, 네 번째 토픽은 블록체인과 관련한 가상화폐 이슈, 다섯 번째 토픽은 블록체인 자체의 기술 이슈로 분류할 수 있다. 토픽 모형의 관점에서는 한글 문서보다 영문 문서의 토픽이 더 적절히 분류되었다고 판단할 수 있다.

블록체인은 다른 분야에 비해 공공 분야에서 실행 가능성과 영향력이 꽤 큰 편에 속한다. 이는 과학기술정보통신부가 블록체인의 보안성, 투명성 측면에서 블록체인이 산업과 사회를 혁신하는 기반 기술로 4차 산업혁명의 핵심 산업이 될 가능성이 충분하다고 판단한 것과 관련이 있다.

이 연구의 한계점은 과학기술정보통신부가 '블록체인 기술 발전전략'을 발표(2018. 6. 21.)하기 전에 문서를 수집한 결과라는 것이다. 정부의

블록체인과 관련된 정책이 이슈화된 문서를 수집하여 위의 분석을 실시한다면 다른 결과가 나올 수 있다.

〔그림 5-2〕 산업 분야별 블록체인의 활용 가능성

자료: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/blockchain-beyond-the-hype-what-is-the-strategic-business-value?reload

그럼에도 소셜 빅데이터 분석은 보건복지 정책 영역에서 국가적·사회적으로 관심이 있는 이슈에의 현 상황을 파악하는 데 중요한 경쟁력으로 작용할 수 있으며, 앞으로 정책 관련 이슈를 도출하고 연구 전략을 세우는 데 근거 자료로 활용될 수 있다. 다양한 소셜 빅데이터 분석 기술을 바탕으로 주요 보건복지 정책에 관한 사회적 관심도, 영향력 등을 분석하고 그 변화 과정을 살펴본다면 시의성 높은 보건복지 정책 연구의 기반을 마련할 수 있을 것이다.

참고문헌

경기도 따복 공동체지원센터. (2018). *2017년 따복공동체 주민제안 공모사업 블록체인 도입 인포그래픽스.* Retrieved from ddabok.go.kr

경기복지재단. (2017). 블록체인 기반 복지화폐 활용방안. *G-Welfare Brief,* (4)

과학기술정보통신부. (2017). 블록체인으로 이웃 간 전력 거래 한다. *과학기술정보통신부 보도자료,*

과학기술정보통신부.(2018). *신뢰할 수 있는 4차 산업혁명을 구현하는 블록체인 기술 발전전략* 과학기술정보통신부.

관세청. (2017, december 21,). 관세청, 세계최초 블록체인 기반 수출 통관 서비스 기술검증 완료. *관세청 보도자료*

김경민. (2018, february 16,). 美 블록체인 기술의 응용현황과 전망. *Kotra 해외시장 뉴스*

김석원. (2016). 비트코인의 기반 기술 블록체인의 원리. *소프트웨어정책연구소 (SPRi),*

김석원.(2017). *블록체인 펼쳐보기* 비제이퍼블릭.

박순영. (2018). *보건의료산업에서의 블록체인 기술 활용* 융합연구정책센터.

박종대. (2018). 블록체인 기술 전망. *한국과학기술연구원,* (83)

박종훈. (2016). 최신 ICT 이슈: 금융을 넘어 모든 비즈니스에 파괴적 영향을 미칠 블록체인. *Iitp, 주간기술동향,*

백웅조. (2017). R3cev(글로벌 블록체인 컨소시엄) 최근 동향. *이슈브리프. Weekly KDB Report,*

서울시. (2017, September 19,). 서울시, 4차산업 '블록체인' 기술 공공서비스에 첫 도입. *서울시 보도자료*

신흥지역정보종합지식포털. (2017). *에스토니아의 전자 정부 구축 성과* EMERiCs 이슈분석.

아카바네 요시하루, 아이케이 마나부, &양현(옮김).(2017). *블록체인 구조와 이론* 위키북스.

오미애, 최현수, 송태민, 이상인, &천미경.(2017). *2017년 소셜 빅데이터 기반 보건복지 이슈 동향 분석*. 세종: 한국보건사회연구원.

우청원. (2018). *에너지 블록체인 도입방원 연구* STEPI Insight 과학기술정책연구원.

유거송, &김경훈. (2018). 기술동향브리프 블록체인. *한국과학기술기획 평가원,* (1)

유엔세계식량계획 홈페이지. from https://innovation.wfp.org/project/building-blocks

이소정. (2018, January 24,). 블록체인으로 미래를 설계하는 네덜란드. *Kotra 해외시장 뉴스*

정성교. (2018). 블록체인 기술 및 연구동향 분석. *Kerc,*

지디넷. (2018, August 24,). 중국 정부, 빈민구제 활동에 블록체인 기술 적용. *지디넷*

특허청. (2018). 블록체인, 핵심·표준 특허 확보 서둘러야. *특허청 보도자료,*

한국일보. (2018, August 20,). 영국, 복지수당에도 블록체인 적용… 연간 5조원대 부정수급 차단. *한국일보*

행정안전부. (2018, May 16,). 2020년부터 종이 대신 모바일로 전자증명서 발급·제출. *행정안전부 보도자료*

홍승필, 민경식, &김혜리. (2017). 블록체인 방식을 활용한 온라인 투표시스템 적용 가능성연구. *한국인터넷정보학회,*

황혜란.(2018). *대전의 블록체인 기술관련 R&D 및 산업육성* 대전세종연구원.

Agrawal, R., Imieli 'nski, T., &Swami, A. (1993). Mining association rules between sets of items in large databases. *SIGMOD Rec., 22*(2), 207-216. doi:10.1145/170036.170072

Beran, R. (1977). Minimum hellinger distance estimates for parametric models. *Ann.Statist., 5*(3), 445-463. doi:10.1214/aos/1176343842

Bitcoin for beginners. Retrieved from www.bitcoinforbeginners.io 2018 july 26, 인출.

Blei, D. M., Ng, A. Y., &Jordan, M. I. (2003). Latent dirichlet allocation. *J.Mach.Learn.Res., 3*, 993-1022.

Block geeks. Retrieved from www.blockgeeks.com 2018 july 26, 인출.

Boucher, P.(2017). *How blockchain technology could change our lives: In-depth analysis* European Parliament.

Dash, S., Majumdar, A., &Gunjikar, P. (2016). Blockchain: A healthcare industry view.

Deloitte.(2016). *Blockchain: Opportunities for health care* Deloitte.

Deloitte 홈페이지. Retrieved from https://www2.deloitte.com/uk/en/pages/innovation/solutions/deloitte-blockchain-practice.html 2018 August 27, 인출.

Feinerer, I. (2018). Introduction to the tm package text mining in R.

Friedman, J., Hastie, T., &Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, 9*(3), 432-441. doi:10.1093/biostatistics/kxm045 [doi]

Gandrud, C., Allaire, J. J., Russell, K., &Yetman, C. J. (2015). *network D3: D3 JavaScript network graphs from R*

Jeon, H., &Kim, T. (2016). *Package 'KoNLP'*

Jolliffe, I. (2011). Principal component analysis. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 1094-1096). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-04898-2_455 Retrieved from https://doi.org/10.1007/978-3-642-04898-2_455

KDB 미래전략연구소. (2018). 중국 블록체인 발전 현황 및 시사점. *China Focus,*

Lamport, L., Shostak, R., &Pease, M. (1982). The byzantine generals problem ACM transactions on progamming languages and syetems, vol. 4 no. 3 pp382-401.

Marc, A. (2014, January 21,). Why bitcoin matters. *Dealbook, the New york Times* Retrieved from https://dealbook.nytimes.com/2014/01/21/why-bitcoin-matters/

Mckinsey digital. Retrieved from www.mckinsey.com/business-functions/digital-mckinsey/our-insights/blockchain-beyond-the-hype-what-is-the-strategic-business-value?reload 2018 july 26, 인출.

Medium. Retrieved from www.medium.com 2018 july 26, 인출.

MedRec. *Techinal documentation* MedRec.

Mikolov, T., Chen, K., Corrado, G., &Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR, abs/1301.3781* Retrieved from http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., &Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *CoRR, abs/1310.4546* Retrieved from http://arxiv.org/abs/1310.4546

Munzert, S., Rubba, C., Meissner, P., &Nyhuis, D. (2014). Automated data collection with R: A practical guide to web scraping and text mining.

Rajaraman, A., &Ullman, J. D.(2011). *Mining of massive datasets* Cambridge University Press.

Scherer, M. (2017). Performance and scalability of blockchain networks and smart contracts.

UK Government office for Science. (2016). *Distributed ledger technology: Beyond block chain*

Wikipedia. Retrieved from www.wikipedia.org 2018 july 26, 인출.

부록

이더리움(Ethereum)

블록체인 기술은 2008년 비트코인과 함께 처음 등장한 이후 몇 년 동안 큰 반향을 일으키지는 못했다. 그러나 '이더리움'이라는 새로운 블록체인의 등장으로 비로소 블록체인 기술이 주목받기 시작했다. 이더리움은 비탈리크 부테린에 의해 2013년 기술 백서(White paper) 공개를 시작으로, 이더리움재단에서 개발이 진행되고 있는 오픈소스 프로젝트[19]이다. 이더리움은 블록체인 기술을 기반으로 스마트 컨트랙트(Smart Contract) 기능을 구현하기 위한 분산 컴퓨팅 플랫폼으로 정의할 수 있다. 즉 블록체인으로 화폐 거래뿐만 아니라 프로그램 코드를 실행할 수 있도록 하여, 전 세계의 수많은 컴퓨팅 자원을 활용해 분산 컴퓨팅 환경을 구성하고, 이 플랫폼을 이용하여 다양한 서비스를 제공할 수 있도록 고안한 시스템이다.

이더리움은 비트코인과 마찬가지로 P2P 네트워크상에서 거래 이력을 블록체인에 기록하는 한편, 탈중앙화 응용 소프트웨어(Decentralized Application, dApp)의 프로그램 코드 자체나 그 실행 이력도 기록할 수 있다는 특징이 있다. 기본적인 블록체인의 동작은 비트코인과 거의 동일하나 블록 생성 주기가 짧고, 블록체인 관리에서 일부 분기를 허용하며, 합의 알고리즘이 곧 PoS로 변경될 예정이다. 그 외에도 개방형 블록체인의 한계를 극복하기 위한 개발이 활발히 진행되고 있다.

19) https://github.com/ethereum