ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2021 December: 27(4): 73~95 http://dx.doi.org/10.13088/jiis.2021.27.4.073

Node2vec 그래프 임베딩과 Light GBM 링크 예측을 활용한 식음료 산업의 수출 후보국가 탐색 연구*

이재성
과학기술연합대학원대학교
과학기술경영정책 박사과정
jslee@kisti.re.kr

전승표
한국과학기술정보연구원 데이터분석본부
책임연구원 / 과학기술연합대학원대학교
과학기술경영정책 교수
spjun@kisti.re.kr

서진이
한국과학기술정보연구원
데이터분석본부 책임연구원
jinny@kisti.re.kr

본 연구는 Node2vec 그래프 임베딩 방법과 Light GBM 링크 예측을 활용해 우리나라 식음료 산업의 미개척 수출 후보국가를 탐색한다. Node2vec은 네트워크의 공통 이웃 개수 등을 기반으로 하는 기존의 링크 예측 방법에 비해 상대적으로 취약하다고 알려져 있던 네트워크의 구조적 등위성 표현의 한계를 개선한 방법이다. 따라서 해당 방법은 네트워크의 커뮤니티 탐지와 구조적 등위성 모두에서 우수한 성능을 나타내는 것으로 알려져 있다. 이에 본 연구는 이상의 방법을 우리나라 식음료 산업의 국제 무역거래 정보에 적용했다. 이를 통해 해당 산업의 글로벌 가치사슬 관계에서 우리나라의 광범위한 마진 다각화 효과를 창출하는데 기여하고자 한다. 본 연구의 결과를 통해 도출된 최적의 예측 모델은 0.95의 정밀도와 0.79의 재현율을 기록하며 0.86의 F1 score를 기록해 우수한 성능을 나타냈다. 이상의 모델을 통해 도출한 우리나라의 잠재적 수출 후보국가들의 결과는 추가 조사를 통해 대부분 적절하게 나타난 것을 알 수 있었다. 이상의 내용을 종합하여 본 연구는 Node2vec과 Light GBM을 응용한 링크 예측 방법의 실무적 활용성에 대해 시사할 수 있었다. 그리고 모델을 학습하며 링크 예측을 보다 잘 수행할 수 있는 가중치 업데이트 전략에 대해서도 유용한 시사점을 도출할 수 있었다. 한편, 본 연구는 그래프 임베딩 기반의 링크 예측 관련 연구에서 아직까지 많이 수행된 적 없는 무역거래에 이를 적용했기에 정책적 활용성도 갖고 있다. 본 연구의 결과는 최근 미중 무역갈등이나 일본 수출 규제 등과 같은 글로벌 가치사슬의 변화에 대한 빠른 대응을 지원하며 정책적 의사결정을 위한 도구로써 충분한 유용성이 있다고 생각한다.

주제어 : Node2vec, 그래프 임베딩, Light GBM, 링크예측, 글로벌 가치사슬

논문접수일 : 2021년 10월 30일 논문수정일 : 2021년 11월 22일 게재확정일 : 2021년 11월 30일
원고유형 : 일반논문 교신저자 : 서진이

1. 서론

본 연구는 미개척 수출 후보 국가의 탐색을 통해 우리나라 식음료 산업의 수출시장 확대 전략을 지원하는 분석 시스템을 제안한다. 이러한 지원을 통해 해당 산업의 글로벌 가치사슬(Global Value Chain, 이하 GVC)을 보다 안정화하는 산업 수출의 다각화 효과를 기대할 수 있다. 본 연구의 이러한 노력은 최근 미국과 중국의 무역갈등이나 일본의 우리나라에 대한 수출규제 등과 같이 갑작스러운 대외무역 관계의 변화에 빠르게 대응할 수 있다는데 의의가 있다.

* 이 연구는 2021년도 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원에 의한 연구임(20009398)

본 연구에서는 미개척 수출 후보 국가를 탐색하기 위해 네트워크 구조를 학습하는 그래프 임베딩 기법인 Node2vec과 이를 통해 도출된 임베딩된 벡터를 활용하는 예측모델로 Logistic Regression과 Light GBM을 사용한다. Node2vec은 고전적인 그래프 임베딩 기법인 임의 보행(Random walk)의 취약점을 개선해 비교적 최근에 학계에 소개된 기술이다(Grover and Leskovec, 2016).기존 임의 보행 기법은 Local neighborhood와 Higher-order neighborhood의 정보를 함께 고려할 수 있도록 노드 유사도를 정의할 수 있다는 표현성(Expressivity)의 장점이 있었다. 하지만 네트워크 노드의 구조적 역할(Structural role) 등을 제대로 표현하지 못한다는 취약점이 존재했다. 이에 Node2vec은 보행을 통해 노드 시퀀스를 탐색하는 전략을 설정할 수 있는 파라미터를 도입함으로써 기존 임의 보행 기반 알고리즘에서 존재하던 약점을 극복했다(Grover and Leskovec, 2016).

본 연구에서는 이러한 방법을 활용하여 양자 무역관계에 있는 누락된 연결의 예측을 수행하고자 한다. 이상의 예측을 수행하기 위해 본 연구에서는 UN 국제 무역 통계 데이터베이스에서 제공하는 UN Comtrade 데이터를 사용하여 모델을 만들고 평가를 진행해서 Node2vec의 활용 가능성에 대해 시사하고자 한다. 이때, 본 연구에서는 분석 범위를 한정하기 위해 우리나라의 식음료 산업에 해당하는 데이터로 예측 모델의 구축을 실험한다. 이상의 연구를 통해 본 연구는 궁극적으로 우리나라 식음료 산업의 광범위한 마진 다각화 효과를 위해 미개척 수출 후보 국가에 대해 시사하고자 한다. 그리고 이러한 후보 국가가 나온 배경에 대해 추가적인 자료조사를 통해 결과가 적절한지를 평가하고자 한다. 이러한 노력을 통해 실무에서 본 연구의 예측 모델의 활용에 대한 깊은 토의를 수행하고자 한다.

본 연구의 구성은 다음과 같다. 이어지는 2장에서는 이론적 배경과 연구 모델에 대해 다룬다. 3장에서는 본 연구의 결과에 대해 살펴보고, 마지막 4장에서는 본 연구의 결론과 시사점에 대해서 논의하고자 한다.

2. 이론적 배경

2.1. GVC의 이론적 배경

세계 경제는 점점 더 GVC를 중심으로 구조화되고 있다. 점점 GVC가 국제 무역, 세계 GDP 및 고용의 증가에 차지하는 비중이 커지고 있기 때문이다(Gereffi and Fernandez-Stark, 2011). 이렇듯 GVC는 상품, 의류, 전자, 관광 및 비즈니스 서비스 아웃소싱과 같은 다양한 부문에서 글로벌 무역, 생산 및 고용, 그리고 개발 도상국의 기업과 생산자 및 근로자가 세계 경제에 통합되는 측면에서 중요한 의미를 갖는다(Gereffi and Fernandez-Stark, 2011).

이러한 GVC의 구체적인 정의를 알아보기 위해서는 VC를 먼저 이해할 필요가 있다. VC는 기업과 근로자가 제품의 개념에서 최종 사용 및 그 이상까지 가져오기 위해 수행하는 전체 범위의 활동을 모두 포함한다. 따라서 최종 소비자에 대한 디자인, 생산, 마케팅, 유통 및 지원과 같은 활동이 모두 VC의 하위 개념을 구성한다(Gereffi and Fernandez-Stark, 2011). 이렇게 VC를 구성하는 다양한 활동은 단일 회사 또는 여러 회사에 분할되어 수행될 수 있다(Global Value Chains Center, 2011).

상기와 같은 일반적인 개념 외에도 VC의 개

념을 다음과 같이 협의로 설정하기도 한다. 보다 좁은 관점에서 VC는 더 적은 비용으로 제품이나 서비스를 제공하기 위해 공급업체 및 고객과 기업의 관계를 설명한다(Christopher, 2005).VC의 개념은 한 단계 더 나아가 기업이 연결되어 경쟁우위의 원천이 되는 가치를 창출할 수 있다고 설명한다(A. Al-Mudimigh et al., 2004; Stabell and Fjeldstad, 1998).이 후자의 개념은 또한 특권적 위치에 있는 고객을 고려하고(Cox, 1999),그들의 요구를 이해하고 가치 창출과 그 포착을 조사함으로써(C. Di Domenico et al., 2007)가치를 제공한다(G. Gereffi and J. Lee, 2012).

이러한 VC는 점차 글로벌 네트워크를 통해 상호 연결된 조직 배열과 기업의 집합체를 포함하는 GVC로 발전했다(de Marchi et al., 2013; Giroud and Mirza, 2015; Mudambi and Puck, 2016).즉, 세계화의 맥락에서 VC를 구성하는 활동은 글로벌 규모의 기업 간 네트워크에서 수행되고 있었다. 따라서 이러한 VC의 범위는 점차 확장되어 개념 및 생산에서 최종 사용에 이르기까지 유형 및 무형 부가가치 활동의 순서에 초점을 맞춤으로써 글로벌 산업에 대한 전체론적 관점을 제공하는 GVC로 확장할 수 있었다(Gereffi and Fernandez-Stark, 2011).

2.2. 산업 수출 다각화 전략의 이론적 배경

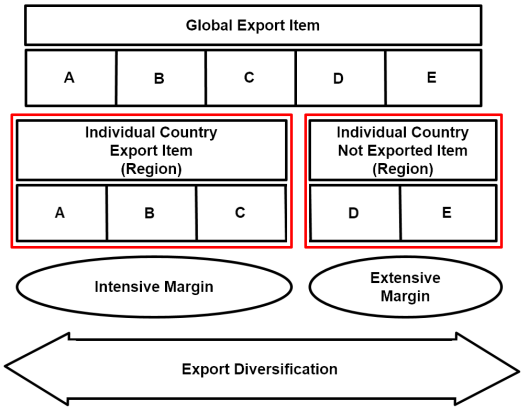
이론적인 관점에서 수출의 다각화가 증가하는 기제는 다음과 같이 크게 두 가지로 구분할 수 있다. 첫 번째 경우는 상대적으로 가치가 낮은 수출 상품이 상대적으로 가치가 높은 상품보다 더 빨리 성장한다는 것이다. 이러한 종류의 다각화는 새로운 제품의 수출을 포함하지 않는다는 점에서 집약적 마진(Intensive margin)에서 발생하는 것으로 생각할 수 있다. 한편, 다각화가 일어날 수 있는 두 번째 방법은 광범위한 마진(Extensive margin)을 통해서 발생할 수 있다. 이는 한 국가의 기존 수출 품목(Export bundle)에 신제품을 도입함으로써 발생하거나(Cadot et al., 2007),새로운 수출 후보 지역의 다변화로부터 발생한다(Abreha et al., 2020; da Costa Neto and Romeu, 2011; El Hag and El Shazly, 2012).

이 중에서 본 연구가 주목하고 있는 광범위한 마진 다각화 효과는 개발 정책의 관점에서 특히 중요하게 다뤄지고 있다. 과거에는 이러한 개발 전략을 후진국들의 경제개발 과정에서 수출활동을 증진시키기 위해 주로 고려되는 활용전략으로 사용했다(Hesse, 2009; Herzer and Nowak-Lehnmann, 2006).이와 관련하여 대표적으로 Dennis & Shepherd (2011)는 광범위한 마진 다각화를 '신제품'의 마진에 독점적으로 초점을 맞추고 연구를 수행한 바 있다(Dennis and Shepherd, 2011)

하지만 최근에는 대외 무역 관계의 급격하게 변하는 GVC의 안정화를 위해 수출 후보 지역의 다변화와 관련된 광범위한 마진 다각화도 중요하게 고려되고 있다(Abreha et al., 2020; da Costa Neto and Romeu, 2011; El Hag and El Shazly, 2012). 이와 관련하여 대표적으로 Collier and Venables (2007)는 광범위한 마진 다각화는 수출 구성 변화와 연결되기 때문에 GVC의 안정화에 기여한다는 연구를 발표된 바 있다.

한편, 국내 연구에서 Min et al. (2011)는 한국 산업의 수출다각화 패턴에 대한 분석을 담은 연구 결과를 발표한 바 있다. 해당 연구에서는 집약적 다각화 마진을 집중영역, 광범위한 다각화 마진을 확장영역으로 설정하며 수출 다각화를 분석하기 위한 체계를 아래 <Figure 1>과 같이 정리했다. 그리고 이러한 체계를 통해 한국 산업

의 수출의 제품 다양화와 지역 다변화의 중요성을 함께 강조했다(Minetal.,2011).그리고 Yeo & Ki(2020)은 우리나라와 중동부 유럽 4개국 간 무역 및 투자 관계와 비교우위를 SWOT 분석을 통해 도출하고, 이를 통해 해당 국가에 대한 우리나라의 무역전략과 유망한 수출상품을 도출했다. 이상의 분석을 통해 저자는 우리나라의 무역확대 전략을 제시할 수 있었다(Yeo & Ki, 2020).



〈Figure 1〉 Framework of Export Diversification (Min et al., 2011)

최근에는 새로운 수출후보의 발굴을 위해 GVC 양자 무역 관계에서 누락된 연결에 주목한 연구들이 수행되고 있다. Tran et al. (2020)은 1997년 동 아시아와 2008년 글로벌 금융 위기 동안 수출 다각화가 실질 환율(Real Exchange Rate, 이하 RER)의 변동성을 줄이는데 얼마나 유의미한 효과가 있었는지 실증적으로 연구했다. 해당 연구에서는 GDP 디플레이터(Deflator)와 같은 거시 경제 지표를 사용해 아시아와 라틴 아메리카 국가들의 RER을 계산하고 수요 측면에서 제품 다양화에 대한 잠재적인 누락 정보를 포착하기 위해 노력했다(Tran et al., 2020).

이에 본 연구도 네트워크 연결 관계에서 누락된 연결에 주목하는 방법을 적용하고자 한다. 본 연구의 차별점은 Tran et al. (2020)가 변동성의 개선에 주목한 것과 달리 누락된 연결의 예측에 대한 정확도에 주목한다는 부분이다. 그리고 이를 위해 Tran et al. (2020)과 달리 예측된 연결의 정확도를 실증적으로 평가하고 검증한다는 부분에 차이가 있다.

2.3. 딥러닝 기반의 GVC 분석

2.3.1 분석 방법론

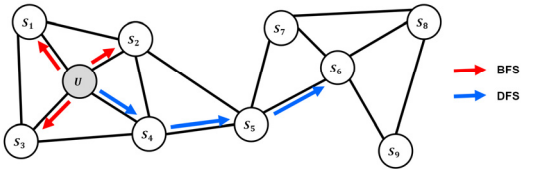
본 연구에서는 국가 단위의 GVC 관계 구조에 딥러닝 기반의 링크 예측 모델을 학습시키고 이와 관련된 실무적 활용 가능성에 대해 논의한다는 연구 목표를 갖는다. 이렇게 학습된 예측 모델을 기반으로 미수출국가에 대한 우리나라 유통 산업의 확장영역을 다각화를 지원하고자 한다. 이때 사용된 연구방법론은 크게 GVC 관계 구조를 임베딩 시키기 위한 Node2vec과 Logistic Regression, 그리고 예측 성능을 극대화하기 위한 Light Gradient Boosting Machine(이하 Light GBM) 방법을 사용하고자 한다.

Node2vec은 대표적인 그래프 임베딩(Graph embedding) 기법으로 인접행렬(Adjacency matrix)를 활용하는 전통적인 방식이 아니라 임의 보행(Random walk)를 기반으로 네트워크 그래프를 임베딩하는 방법을 말한다(Grover and Leskovec, 2016). 이러한 임의 보행 기법은 Deep Walk라는 알고리즘에서 먼저 도입된 방법으로 임의 보행하며 얻은 그래프 노드의 시퀀스를 통해 네트워크의 연결 관계를 임베딩 과정에 활용한다(Perozzi et al.,2014).이렇게 임의 보행을 사용하면

Local neighborhood와 Higher-order neighborhood의 정보를 함께 고려할 수 있도록 노드 유사도를 정의할 수 있다는 표현성(Expressivity)의 장점이 있다. 그리고 노드의 임베딩을 모델이 학습할 때 모든 노드의 쌍을 고려할 필요가 없기 때문에 연산 비용이 보다 효율적(Efficiency)이라는 장점이 있다(Perozzi et al.,2014).그리고 임의 보행을 수행할 때 노드는 임의로 지정된 시작점에서 일정한 길이(Fixed length)의 조건으로 동작한다. 따라서 노드의 시퀀스를 Logistic Regression, Support Vector Machine, Random Forest 등과 같은 다운스트림 작업(Downstream task)에 적용하기 수월하다는 장점이 있다.

하지만 이렇게 고정된 길이의 노드 시퀀스는 다음과 같은 취약점을 가지고 있다. 예컨대, 노드의 구조적 역할(Structural role)은 비슷하나 거리는 멀리 떨어져 있는 경우에 대한 표현이 제대로 수행되지 않는 경우가 있다. 그리고 서브 그래프의 전체 구조를 고려하지 못한다는 단점도 역시 존재한다. 이에 Node2vec은 Breadth-first Sampling(이하 BFS)와 Depth-fist Sampling(이하 DFS)의 개념을 도입함으로써 Deep Walk 알고리즘에서 존재하던 약점을 극복했다(Grover and Leskovec, 2016).

BFS는 임의 보행을 수행할 때 임의로 지정된 시작점으로 다시 되돌아올 확률을 높게, 즉 시작점으로 지정된 노드가 임의 보행에 다시 포함될 확률을 높게 노드 시퀀스를 생성한다. 따라서 이렇게 생성한 노드 시퀀스는 서브 그래프의 구조를 보다 잘 표현할 수 있다(Local expression). 한편, DFS는 임의 보행을 수행할 때 임의 지정 시작 노드로 다시 되돌아올 확률을 낮게 설정하는데, 이는 임의 보행이 점진적으로 보다 넓게 진행되게끔 만든다. 따라서 이상의 방법으로 생성한 노드 시퀀스의 경우에는 멀리 떨어져 있는 노드를 더 잘 나타낼 수 있게 된다(Global expression). 이상의 방법에 대한 도식이 아래 <Figure 2>와 같다(Grover and Leskovec, 2016).



〈Figure 2〉 BFS and DFS search strategies from node (k = 3) (GroverandLeskovec,2016)

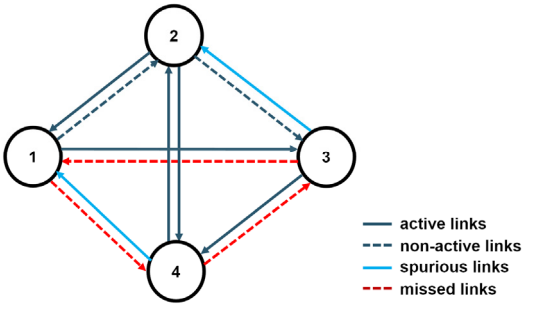
본 연구는 이후 Node2vec으로 임베딩된 벡터를 예측 변수로 설정하고 그래프 노드의 연결을 반응 변수로 설정하여 연결 관계를 예측하는 링크 예측(Link prediction) 모델을 학습한다. 따라서 제안하는 모델에 입력으로 설정된 예측 변수는 임베딩 과정을 통해 임의의 차원 값으로 고정된 길이로 표현되어 있는 개별 국가들의 고유 벡터를 사용한다. 그리고 제안하는 모델의 출력으로 설정된 반응 변수는 개별 국가들의 고유 벡터의 링크 유무를 사용한다. 이때 사용되는 제안 모델은 Logistic Regression을 이용한 이진 분류기(Binary classifier)이며 이러한 분류기의 성능을 보다 극대화하기 위해 Extreme Gradient Boosting (XGBoost) 방법을 개선한 앙상블 부스팅(Ensemble boosting) 기법인 Light GBM(Ke et al.,2017)을 추가로 적용한다. 이때 분류 클래스의 불균형(Class imbalance)이 존재하기 때문에 이진 분류기 학습에는 모두 클래스 균형을 맞춰주는 모델 파라미터를 설정하고 수행해야 한다.

2.3.2 제안하는 방법과 유사한 접근의 기존 연구

본 연구와 같이 무역 네트워크에서 잠재적 거래 대상 국가를 네트워크 노드 개념에서접근한 연구로는 Tuninetti et al. (2017)가 있다. 해당 연구는 식량 생산과 불가분의 관계에 있는 담수 자원 무역 네트워크에서 잠재적 거래 대상 국가의 도출을 위해 링크 예측 방법을 활용했다(Tuninetti et al.,2017).전 세계 담수 사용의 약 70%가 농산물 생산을 위한 것이기 때문에(Falkenmark et al., 2004; Oki and Kanae, 2004) 담수 자원은 식량 가용성을 제어하는 주요한 요인이라고 볼 수 있다(Godfray et al., 2010; Tilman et al., 2011). 이에 농산물의 국제 무역을 통해 생산국에서 물리적으로 사용되는 담수 자원이 소비국으로 가상으로 이전된다. 이렇게 이전되는 물의 양을 가상 담수(virtual water)라고 하며(Allan, 2006; Antonelli and Sartori, 2015), 이러한 무역을 가상 담수 거래(virtual water trade)라고 한다(Tuninetti et al.,2017).이러한 가상 담수 거래는 담수 자원이 부족한 지역에서 식품에 대한 물리적, 경제적 접근을 개선하여 국가가 담수 집약적인 제품의 수입을 통해 국내 담수 자원을 절약할 수 있도록 하는 능력으로 인식되곤 한다(Yang et al., 2006; Chapagain et al., 2006).

상기와 같은 연구배경에서 해당 연구는 네트워크의 노드를 담수 자원 무역을 수행하는 국가로 설정하고, 임의의 두 국가 간의 수입/수출 관계를 링크로 표현했다는 점이다. 해당 연구는 이러한 아이디어를 바탕으로 링크 예측 방법에 대한 분석 틀을 제안했다. 해당 연구가 다루는 문제는 국가 속성(예: 인구, 국내 총생산, 물 수요) 및 링크 특성(예: 지리적 거리)를 가지고 수행된다는 특징이 있다. 이러한 특징에 따라 두 국가 사이에 잠재적인 링크가 있다는 가정에서 담수 자원 거래를 의미하는 네트워크의 링크가 존재할 것으로 예상되는지 식별한다. 그리고 해당 연구에서는 임계값(Threshold)을 사용하여 이상의 기준보다 높은 담수 자원 거래 링크는 활성 링크(Active link)로 간주하고, 다른 링크는 비활성(Non-active link)으로 간주한다. 최종적으로 활성 링크에서 거래되는 담수 자원 거래 정보를 바탕으로 비활성 링크를 다시 추정한 결과를 평가한다는 특징이 있다.

구체적인 성능 평가는 다음과 같다. 링크 예측으로 도출된 결과와 실제 네트워크와 비교했을 때 얼마나 일치하는지 계산한다. 즉, 활성 링크를 실제 네트워크로 제대로 예측한 경우(True Positive)와 그렇지 못한 경우(False Positive)를 구분하고, 비활성 링크를 실제 네트워크로 제대로 예측한 경우(True Negative)와 그렇지 못한 경우(False Negative)로 계산하여 모델의 정확도, 정밀도 및 F1 score를 평가한다. 이상의 내용이 아래 <Figure 2>와 같다.



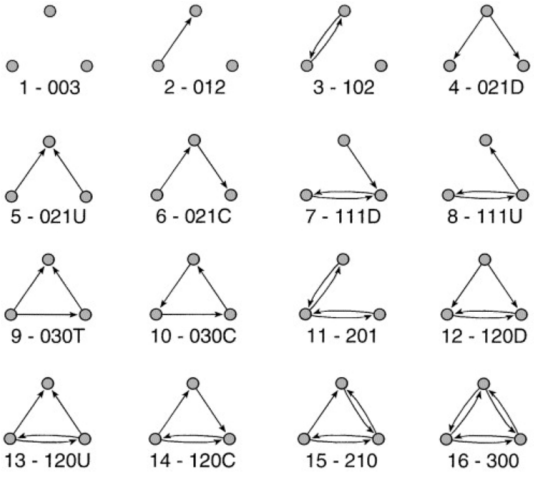
〈Figure 2〉 Link Prediction Framework of Tuninetti et al. (2017)

3. 연구 방법과 활용 데이터

3.1. 네트워크 분석 방법론

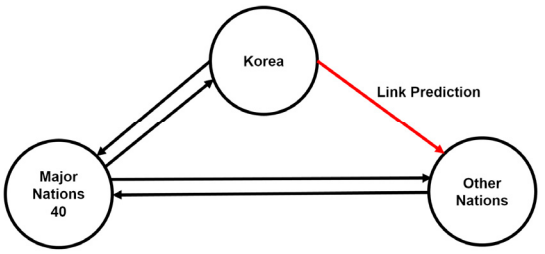
본 연구는 앞선 유사 연구와 다음과 같은 차별점을 갖는다. 앞선 연구는 국가 간의 수출 무역 관계를 양자 무역(Bilateral trade)으로 연결되어 있다고 가정하고 있다. 이러한 부분에서 본 연구는 기존연구의 가정을 충실히 따르고는 있으나, 무역 관계를 형성하는 국가의 숫자에서 차이가 있다. 선행연구의 경우에는 4개의 국가 노드 간의 관계에서 새로운 무역 거래 후보 국가를 탐색하려고 시도한 반면에, 본 연구는 3개의 국가 노드의 관계에서 이를 살펴본다는 부분에서 첫번째 차별점이 있다. 본 연구는 선행연구와 달리 3개의 국가 노드의 관계를 설정함으로써 이론적 관계 틀(framework)에 기초해 새로운 링크에 대한 예측을 수행할 수 있다는 장점을 갖는다.

이러한 이론적 관계 틀을 동형 3요소 집합(The triad isomorphism class)이라고 하며, 기존의 다양한 네트워크 관련 선행연구(Batagelj and Mrvar, 2011; Doroud et al., 2011; Ortmann and Brandes, 2017; Uddin et al., 2018)에서 이론적으로 충분히 검토된 네트워크 관계에 대한 사전 정의를 말한다. 본 연구에서 참고한 선행연구는 4개의 국가 노드 간의 관계에서 새로운 무역 거래 후보를 탐색했지만, 그러한 관계에 대한 구체적인 이론적 가정 등에 대한 설명이 부족했다. 이에 본 연구에서는 이러한 관계 가정에 대한 이론적 기초를 확보하기 위해 동형 3요소 집합에서 제시하고 있는 관계 집합을 잠재적 수출 후보국 도출을 위한 무역 관계 설정에 응용하고자 한다. 이러한 관계의 집합은 아래 <Figure 3>과 같은데, 본 연구에서는 11-201 모형에 주목하고자 한다.



〈Figure 3〉 The Triad isomorphism class (BatageljandMrvar,2016)

본 연구는 11-201 모형에 기초해 아래 <Figure 4>와 같이 우리나라와 주로 무역 거래를 수행하는 주요 40개 국이 다른 국가와 무역 거래를 수행한 정보를 바탕으로 새로운 수출 후보 국을 탐색하고자 한다. 이상의 주요 40개 국은 우리나라 무역 거래의 90%를 차지하고 있는 국가들을 말하며, 이러한 주요 국가들이 거래하는 우리나라의 미개척 수출 지역의 탐색을 통해 무역 확장영역을 다각화하는데 기여하고자 한다. 구체적으로, 주요 40개국을 기준으로 한국과 수행하는 양자 무역 거래 정보와 기타 국가들과 수행하는 양자 무역 거래 정보를 활용하여 링크 예측을 수행하고 최종적으로 한국이 기타 국가로 미개척 수출 지역을 넓힐 기회를 탐색한다.



〈Figure 4〉 Strategies to explore potential export candidates

본 연구의 두번째 차별점은 유사한 선행연구에서 링크 예측에 수행한 전통적인 통계 기반의 방법과 달리 딥러닝 기반 방법을 사용했다는 점이다. 선행연구에서는 링크의 예측을 위해 Common Neighbor index (CN), Salton index (Salton), Preferential Attachment index (PA), Hub Depressed Index (HDI) 등과 같은 방법을 사용한다. 하지만 본 연구는 앞서 설명한 것과 같이 Node2vec을 사용한다. Node2vec은 Grover and Leskovec (2016)에 의해 발표된 이후 현재 5년의 시간이 지남에 따라 학문적 검증이 완료된 상태이며, 비교적 최근에서야 Node2vec을 응용해 채권 거래 시장의 잠재 거래 후보군을 도출하거나(Hao et al., 2019), 인간 표현형과 유전 정보의 링크 예측(Patel and Guo, 2021)과 관련된 연구 보고들이 발표되고 있는 실정이다. 이러한 부분에서 본 연구는 아직 우리나라의 거래 무역에 적용이 미미한 Node2vec을 활용함으로써 연구의 최신성을 확보할 수 있으며, 그래프 임베딩의 유용성에 대해서 논의하며 다양한 시사점을 도출하고자 한다.

3.2. 데이터 수집 및 변수 설정

본 연구에서 사용한 데이터는 UN에서 제공하고 있는 UN 국제 무역 통계 데이터베이스인 UN Comtrade이다. 해당 데이터베이스는 유엔 통계국(UNSD)에서 연간 및 월간으로 제공하고 있으며, 국제 무역 통계 데이터를 상품과 서비스 범주 및 파트너 국가(Partner)별로 자세하게 제공하고 있다는 특징이 있다. UNSD의 데이터 수집에는 170개 이상의 보고 국가(Reporter)와 지역에 대해 다루고 있기 때문에 UN Comtrade는 국제 무역 데이터베이스 중에서 가장 큰 규모를 자랑한다. 해당 데이터베이스가 갖는 장점은 모든 상품에 대한 가치가 리포터 국가에서 제공하는 환율에 기초해 미국 달러로 변환되어 있다는 점이다. 그리고 상품에 대한 분류도 국제 무역 분류 코드인 Harmonized System(HS)로 나눠져 있기 때문에 특정 산업에 대한 분석에 있어서도 굉장히 유용하다는 장점이 있다(UN Trade Statistics, 2006).

본 연구는 이러한 HS 분류를 이용해 식료품 및 담배 등으로 이뤄진 농수산물 산업에 한정해 분석을 수행한다. 한국은 2004년 체결한 최초의 한-칠레 FTA 이후 지금까지 16건의 FTA를 추가로 체결했고 최종적으로 55개국과 연결 고리를 만들었다. 본래 FTA는 농수산물 산업에 위기로 인식된다. 하지만 최근에 보고되는 자료에 따르면 FTA를 통한 무관세 적용 이후 2018년 아세안으로 수출된 농산물이 13억 2,700달러, 2019년에는 13억 7,500만 달러까지 올라간 것으로 나타나며 새로운 기회로 인식이 전환되고 있다. 실제로 2019년 곡물 음료, 차 류, 기능성 음료 등의 다양한 제품의 시장 진출이 활발하게 이뤄지며 베트남에 수출한 음료가 처음으로 3,000만 달러를 넘어서기도 했다. 이러한 상황에서 정부는 농수산물 수출을 위한 새로운 시장 확보에 적극적으로 나서며, 러시아와 몽골 등과 같이 지금까지 교류가 적었던 국가로 수출 시장을 선점하려고 노력

〈Table 1〉 Description of HS Code (UN Trade Statistics, 2006)

HSCODE	Description
0200	Chapter 2 Meat and edible meat offal
0300	Chapter 3 Fish and crustaceans, molluscs and other aquatic invertebrates
0700	Chapter 7 Edible vegetables and certain roots and tubers
0800	Chapter 8 Edible fruit and nuts; peel of citrus fruit or melons
0900	Chapter 9 Coffee, tea, mate and spices
1500	Chapter 15 Animal or vegetable fats and oils and their cleavage products; prepared edible fats; animal or vegetable waxes
1603	Extracts and juices of meat, fish or crustaceans, molluscs or other aquatic invertebrates.
1700	Chapter 17 Sugars and sugar confectionery
1900	Chapter 19 preparations of cereals, flour, starch or milk; pastry cooks′ products
2103	Sauces and preparations therefor; mixed condiments and mixed seasoning; mustard flour and meal and prepared mustard.
2200	Chapter 22 Beverages, spirits and vinegar.
2400	Chapter 24 Tobacco and manufactured tobacco substitutes

〈Table 2〉 Sample Data (27,889 x 3)

Reporter	Partner	Trade Value (US$)
China	Argentina	2,345,674,022
China	Australia	2,201,335,843
China	Austria	119,043,916
China	Belgium	47,316

하고 있다(Etoday, 2020).

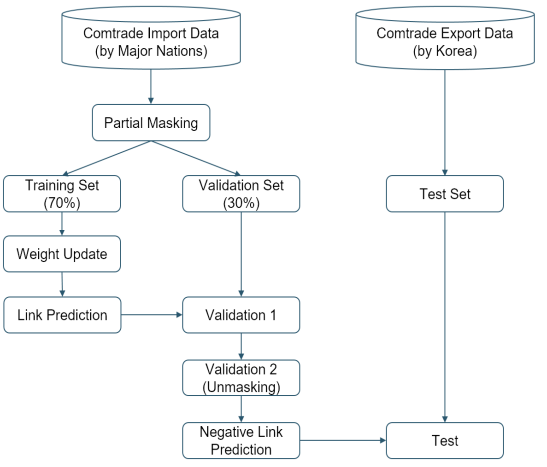
본 연구에서는 이상의 산업을 한정하기 위해 사용한 HS 분류는 0200, 0300, 0700, 0800, 0900, 1500, 1603, 1700, 1900, 2103, 2200, 2400이다. 각 분류에 대한 자세한 설명은 아래 <Table 1>과 같다.

본 연구에서 수집한 UN Comtrade는 가장 최신의 정보인 2020년에 해당하는 국가들에 대한 무역 정보이다. 그리고 앞선 2.3 섹션에서 설명한 것과 같이 우리나라를 제외한 주요 40개국이 수입 거래를 하는 정보를 수집한다. 2020년에 해당하는 주요 40개국의 수입 거래 정보에는 27,889건의 거래내역이 포함되어 있었으며, 이중에서 본 연구에서 사용한 변수는 보고 국가와 파트너 국가, 그리고 무역 거래 규모의 3종의 변수를 사용한다. 이상의 데이터에 대한 샘플이 아래 <Table 2>와 같다.

한편, 주요 40개국의 수입 거래 정보에 기반해

도출하는 우리나라의 잠재적 수출 대상국가에 대한 예측 결과를 평가하기 위해서 2020년에 해당하는 실제 우리나라의 수출 데이터도 함께 수집한다.

종합하면, 본 연구는 상기와 같이 수집한 데이터를 활용해 아래 <Figure 5>와 같이 모델 훈련과 성능 평가 프로세스를 설계할 수 있다. 우리나라를 제외한 주요 40개 국가의 Comtrade 수입 데이터의 일부를 고의로 누락(Masking)시키고, 이상의 데이터를 7:3의 비율로 학습 데이터 셋(Training set)과 검증 데이터 셋(Validation set)으로 나눈다. 이때 학습 데이터 셋은 모델의 가중치를 학습(Weight update)하는 과정에 활용되며 이를 통해 최종적으로 링크 예측 모델링을 수행한다. 링크 예측 모델의 성능은 이전에 나눴던 검증 데이터 셋을 통해 첫번째 검증 단계를 거친다. 이상의 검증 데이터 셋에는 앞서 누락시켰던 값들이 포함되어 있으므로 두번째 검증 단계에서는 누락시켰던 값에 대한 링크 예측과 누락되기 전에 가지고 있던 실제 링크 존재 유무를 비교 평가한다. 이렇게 거친 두번의 검증 단계에서 도출된 최적화된 예측 모델을 사용해서 우리나라와 무역 거래(Link)가 존재하지 않는 국가들에 대한 링크 예측을 수행한다(Negative link prediction). 최종적으로 주요국의 수입 데이터를 기반으로 계산된 우리나라의 수출 미개척 지역에 대한 잠재적 후보군 결과를 실제 우리나라의 Comtrade 수출 데이터로 성능을 최종 평가하고자 한다. 이상의 과정을 통해 주요국의 수입 데이터를 기반으로 도출한 잠재적 후보군 결과가 실제 우리나라의 수출과 얼마나 겹치는지를 평가하고, 이러한 잠재 후보군 중에서도 아직까지 우리나라의 수출 대상으로 미개척된 국가들에 대한 후보를 최종적으로 산출하고자 한다.



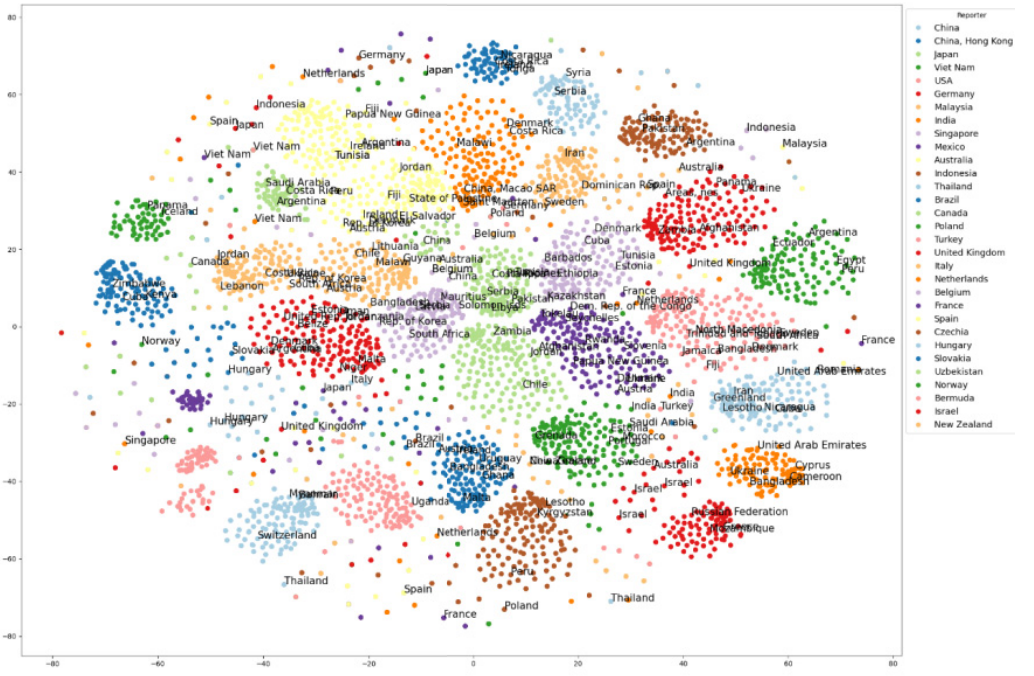
〈Figure 5〉 Model training and performance evaluation process

4. 연구결과

4.1. 임베딩 결과

Node2vec을 수행하여 네트워크 노드를 임베딩하기 위해 본 연구에서는 임의 보행(Random walk) 횟수를 50번으로, 보행 거리(Walk length)를 최대 16까지 설정했다. 임의 보행 탐색 전략은 BFS와 DFS의 사이 중간 정도의 값으로 설정해 구조적 등위성도 적당히 포착하면서도 네트워크 커뮤니티도 적당하게 표현할 수 있도록 임베딩을 수행했다(p=1, q=1). 이때, 임베딩 벡터의 길이는 100차원으로 설정했으며, 주변 노드 벡터가 위치할 확률을 예측하기 위한 Skip-Gram 과정에서 검토한 윈도우 크기는 7로 설정하고 분석을 수행했다.

이상의 방법을 통해 수행한 임베딩 결과를 확인하기 위해서 아래 <Figure 6>과 같이 T-SNE 분석으로 시각화를 진행했다. T-SNE 시각화 결



〈Figure 6〉 T-SNE visualization of embedding results

과에서 노드의 색깔 범례는 보고 국가(Reporter)를 나타냈다. 이때 대부분의 데이터 포인트가 색깔 별로 잘 뭉쳐진 것을 확인할 수 있었다. 이를 통해 국가의 의미론적 속성(Semantic Attributes)이 충분히 표현하도록 임베딩이 수행된 것을 알 수 있었다. 그밖에 각 노드의 주석(Annotation)은 파트너 국가(Partner)를 의미했다. 따라서 T-SNE 시각화 결과는 보고 국가 별로 무역 관계를 맺고 있는 파트너 국가에 대한 위상 정보를 나타내고 있었다. 이때 노드의 주석은 가독성을 위해 일부만 표현했다.

4.2. 모델학습 및 평가 환경설정

Node2vec 임베딩을 통해 얻은 보고 국가와 파트너 국가의 고유 임베딩 벡터 값을 활용해 양자 무역관계의 유무를 예측하는 이진 분류기(Binary classifier)를 학습할 수 있었다. 이상의 이진 분류기 학습에는 앞선 연구방법론에서 설명한 것과 같이 Logistic Regression과 Light GBM 모델을 활용했다. 본 연구에서는 특별히 링크를 예측하기 위한 모델의 학습과정에서는 Masking을 어떤 그룹을 중심으로 수행할지를 설정함으로써 모델의 가중치 학습 전략을 다르게 설정할 수 있었다. 이때 Masking을 세분하는 기준은 보고 국가와 파트너 국가의 무역 거래 Trade Value 규모로 설정했다. 이렇게 나눈 기준을 구분하면 크게 (1) 전체 규모의 파트너 국가를 대상으로 하는 검증 데이터 그룹(All group), (2) 평균 이상의 규모를 갖는 파트너 국가를 대상으로 하는 검증 데이터 그룹(Over Average group), (3) 상위 25% 이상의

〈Table 3〉 Non-Specific Target Group Training Model Performance (All Trade Value)

Threshold	Logistic Regression				Light GBM			
	Prec.	Rec.	F1	Test	Prec.	Rec.	F1	Test
0.50	0.80	0.52	0.63	0.51	0.80	1.00	0.89	1.00
0.55	0.81	0.04	0.07	0.03	0.80	1.00	0.89	1.00
0.60	-	-	-	-	0.80	1.00	0.89	1.00
0.65	-	-	-	-	0.80	0.97	0.88	0.97
0.70	-	-	-	-	0.80	0.47	0.59	0.47
0.75	-	-	-	-	0.80	0.08	0.15	0.08
0.80	-	-	-	-	0.81	0.01	0.01	0.01
0.85	-	-	-	-	1	0.00	0.00	0.00
0.90	-	-	-	-	-	-	-	-

규모를 갖는 파트너 국가를 대상으로 하는 검증 데이터 그룹(Over Q1 group)과 같이 나눌 수 있었다. 따라서 각각의 그룹을 대상으로 총 3개의 모델을 학습할 수 있었다. 그리고 이렇게 학습한 모델이 분류하는 범주의 확률(Probability)에 대한 임계치(Threshold)를 Cut-off 기준으로 설정해 각 모델의 성능을 보다 자세하게 평가했다. 이때 사용된 임계치는 일반적으로 사용되는 0.5 이상의 확률에서 굉장히 엄격한 0.9 이상의 확률까지 0.05씩 증가시켜가며 성능을 평가했다.

4.3. 모델 학습 결과

모델을 학습한 결과는 무역 거래 Trade Value 규모로 구분한 3개의 집단에 대해 각각 도출했다. 먼저 특정한 타겟 그룹을 설정하지 않고 모든 Trade Value 규모의 국가를 대상으로 학습한 모델의 성능은 아래 <Table 3>과 같았다.

상기 그룹에 대해 Logistic Regression 모델의 성능은 0.5의 임계치에서 약 80%의 정밀도와 약 52%의 재현율을 기록했다. 이때, 조화평균인 F1 score는 약 63%로 나타났고, Masking 되어 있던 예측 결과를 실제 결과와 비교해본 Test 성능은 약 51% 수준인 것으로 나타났다. 하지만 0.55의 임계치에서는 재현율과 F1 score, Test 성능 모두 현저하게 떨어져 심각한 과적합이 발생했다. 따라서 더 높은 임계치에 대한 결과를 예측할 수 없었다.

한편, Light GBM 모델의 성능은 Logistic Regression 모델보다 훨씬 높은 0.65의 임계치까지 과적합이 발생하지 않고 링크를 예측하는 것으로 나타났다. 하지만 해당 임계치까지 도달하기 전 모델들의 재현율과 Test 성능이 모두 100%로 기록돼 민감도(sensitivity)가 지나치게 높은 것으로 나타났다.

두 번째로, 평균 이상인 Trade Value 규모를 가진 국가를 타겟 그룹으로 학습한 모델의 성능은 아래 <Table 4>와 같았다. 상기 그룹에 대해

〈Table 4〉 Specific Target Group Training Model Performance (Over Average Trade Value)

Threshold	Logistic Regression				Light GBM			
	Prec.	Rec.	F1	Test	Prec.	Rec.	F1	Test
0.50	0.95	0.73	0.83	0.67	0.95	0.79	0.86	0.73
0.55	0.96	0.68	0.79	0.62	0.95	0.77	0.85	0.71
0.60	0.96	0.62	0.75	0.56	0.96	0.74	0.84	0.68
0.65	0.97	0.55	0.70	0.49	0.96	0.71	0.81	0.64
0.70	0.97	0.47	0.63	0.42	0.97	0.67	0.79	0.61
0.75	0.98	0.39	0.56	0.35	0.97	0.63	0.76	0.57
0.80	0.98	0.30	0.46	0.27	0.97	0.57	0.72	0.51
0.85	0.99	0.21	0.35	0.19	0.98	0.49	0.66	0.44
0.90	0.99	0.12	0.22	0.11	0.98	0.38	0.55	0.34

Logistic Regression 모델의 성능은 0.50의 임계치에서 0.90의 임계치까지 정밀도가 선형적으로 증가하는 것으로 나타났고, 반면 재현율은 선형적으로 감소하는 것으로 나타났다. 이때, 조화평균인 F1 score는 임계치에 따라 약 83%에서 약 22%까지 점진적으로 성능이 떨어지는 것을 알 수 있었다. 이는 Masking 되어 있던 예측 결과를 실제 결과와 비교해본 Test 성능에서도 마찬가지였는데, 임계치에 따라 약 67%에서 약 11%로 성능이 낮아졌다. 해당 모델은 왜곡된 과적합 또는 민감도가 발견되지 않았기 때문에 임계치에 따라 성능의 분포가 이상적으로 나타났다.

한편, Light GBM 모델의 성능은 Logistic Regression 모델보다 전체적으로 향상된 것으로 나타났다. 두 모델 모두 임계치 0.5에서 가장 우수한 성능을 기록했는데, 이때 Light GBM 모델은 Logistic Regression 모델보다 F1 score에서 약 3%, Test 성능에서는 무려 약 6%나 높게 기록됐다. 더군다나 상당히 높게 설정한 0.8의 임계치에서 Logistic Regression 모델은 F1 score가 약 46%, Test 성능이 약 27% 밖에 불과했던 반면에 Light GBM 모델에서는 각각 약 72%, 약 51%로 준수한 성능을 기록했다.

마지막으로, 상위 25% 이상인 Trade Value 규모를 가진 국가를 타겟 그룹으로 학습한 모델의 성능은 아래 <Table 5>와 같았다. 상기 그룹에 대해 Logistic Regression 모델의 성능은 이전 타겟 그룹에 대한 성능과 마찬가지로 임계치에 따라 정밀도는 선형적으로 증가했지만, 재현율과 F1 score 및 Test 성능이 모두 점차 낮아지는 것으로 나타났다. 이러한 성능 감소의 폭은 상위 25% 이상인 Trade Value 규모를 갖는 타겟 그룹의 학습 모델에서 훨씬 큰 폭으로 성능이 저하되는 것을 알 수 있었다.

이러한 현상은 Light GBM 모델의 성능에서도 마찬가지였는데, 비록 Light GBM 모델이 Logistic Regression 모델에 비해 어느정도 성능의 향상을 기록했지만 Logistic Regression 모델과 같이 재현율과 F1 score 및 Test 성능이 <Table 4>에 비해 빠르게 저하되는 것을 알 수 있었다.

〈Table 5〉 Specific Target Group Training Model Performance (Over Q1 Trade Value)

Threshold	Logistic Regression				Light GBM			
	Prec.	Rec.	F1	Test	Prec.	Rec.	F1	Test
0.50	0.88	0.71	0.79	0.61	0.88	0.76	0.82	0.64
0.55	0.89	0.63	0.74	0.53	0.89	0.71	0.79	0.60
0.60	0.90	0.55	0.68	0.46	0.90	0.66	0.76	0.55
0.65	0.91	0.47	0.62	0.38	0.92	0.61	0.73	0.50
0.70	0.92	0.39	0.55	0.32	0.92	0.53	0.67	0.43
0.75	0.94	0.20	0.33	0.16	0.93	0.46	0.61	0.37
0.80	0.94	0.20	0.33	0.16	0.94	0.37	0.54	0.30
0.85	0.94	0.12	0.22	0.10	0.96	0.28	0.43	0.22
0.90	0.96	0.06	0.10	0.04	0.67	0.14	0.25	0.11

〈Table 6〉 Negative Link Prediction Result

Class A (Prob, >= 96%)	Class B (96% > Prob, >= 95%)	Class C (95% > Prob, >= 91%)	Class D (91% > Prob, >= 50%)
Tokelau	Saint Lucia	Saint Pierre and Miquelon	Dem. People's Rep. of Korea
Grenada	Chad	Bosnia Herzegovina	Montenegro
Lesotho	Niue	Saint Kitts and Nevis	Burundi
Bermuda	Saint Barthelemy	Niger	Cabo Verde
Bonaire	San Marino	Pitcairn	-
Sao Tome and Principe	Holy See (Vatican City State)	Nauru	-
Saint Vincent and the Grenadines	North Macedonia	Tuvalu	-
Gambia	Anguilla	Syria	-
Andorra	-	Dominica	-
Greenland	-	Eritrea	-
Antigua and Barbuda	-	Saint Helena	-
Saint Marrten	-	Bouvet Island	-

4.4. 최종 결과

본 연구는 앞서 도출한 링크 예측 모델 결과를 바탕으로 평균 이상인 Trade Value 규모를 가진 국가를 타겟 그룹으로 학습한 Light GBM 모델(Threshold=0.50)이 최적의 모델이라고 판단할 수 있었다.본 연구는 이상의 모델을 사용해 최종적으로 우리나라의 미개척 수출 국가에 대한 예

측(Negative Link Prediction)을 수행했다. 이를 위해서 앞서 Node2vec 그래프 임베딩을 통해 얻은 각 국가의 고유 벡터를 링크 예측모델에 활용했다. 이때 우리나라의 미개척 수출 국가에 대한 예측을 설정하기 위해 보고 국가를 모두 우리나라 고유 벡터로 입력했다. 그리고 파트너 국가에는 우리나라를 제외한 모든 국가들에 대한 고유 벡터를 입력했다. 이와 같이 입력 받은 보고 국가와 파트너 국가의 고유 벡터 쌍에 대한 링크 예측 확률을 구할 수 있었다. 이후, 이상의 예측 결과에서 기존에 우리나라가 수행하고 있던 국가를 제외한 나머지 국가들에 대한 예측 결과를 최종적으로 37개의 미개척 수출 국가에 대한 예측 결과로 도출할 수 있었다. 도출된 수출 후보 국가들은 예측된 확률을 기준으로 4개의 범주(Class)로 분류할 수 있었다. 이상의 범주는 각각 최상위, 차상위, 차하위, 최하위 후보 국가 그룹으로 명명할 수 있었다. 이렇게 도출된 미개척 수출 국가에 대한 결과가 아래 <Table 6>와 같다.

5. 결론 및 시사점

본 연구에서 도출된 결과를 해석하면 다음과 같은 결론을 이끌어낼 수 있다. 최상위 수출 후보 국가를 살펴보면, Tokelau, Grenada, Bonaire, Saint Vincent and the Grenadines, Antigua and Barbuda의 국가들이 모두 카리브해 연안에 위치한 것을 알 수 있다. 그 외에 Bermuda의 경우도 비록 북대서양해에 위치한 국가지만 지리적 위치가 카리브해 연안 국가들과 가까웠다. 즉, 최상위 수출 후보 국가로 전체 12개 국가 중에서 유럽 국가의 구성국 및 아프리카 일부 국가 외에 카리브해 연안 국가로만 6개국이 예측된 것이다.

이에 카리브해 연안 국가에 대해 조사한 결과 우리나라 수출 확대, 특히 식음료 산업에서 후보 수출 국가로 고려되기 충분하다는 판단을 할 수 있었다. 카리브 연안 국가에는 33개국이 밀집해 있다. 보고에 따르면, 2018년 한국과 중국이 체결한 FTA로 인해 음료 등의 가공식품에 대한 관세 철폐가 이뤄졌으며 이를 통해 앞으로 중남미 및 카리브 연안 국가에 음료 등의 수출이 크게 확대될 것으로 예상됐다. 당시 우리나라 對중미 주요 수출 품목이 음료수 등의 가공식품이었으며 수출액이 가장 높은 것은 기타 음료에 속했기 때문이다(Kim et al., 2018).

특히, 이중에서 대표적인 카리브 연안 국가 중 하나인 파나마는 이미 우리나라의 주요한 식음료 수출 대상 국가였다. 해당 국가에서는 건강에 대해 높아진 관심으로 탄산음료를 제치고 생과일 무설탕 등의 과일 및 채소 주스의 수요가 많다. 이와 관련해 KOTRA의 자료에 의하면 2018년 기준 對파나마 수출의 90% 이상이 알로에 음료이며, 파나마를 통해 카리브해 연안 시장으로 진출 계획을 추진할 예정이라고 말했다(Hwang, 2018).

본 연구에서 도출한 수출 후보 국가는 이러한 배경으로 인해 카리브 연안 국가 중에서 아직 미개척 지역에 대한 우선순위가 높게 나타난 것으로 이해할 수 있다. 최근 해당 국가들은 코로나 바이러스로 인해 타격이 심각했다. 국제중남미 카리브경제위원회(이하 CEPAR)에 따르면 코로나 기간인 중남미 및 카리브 지역의 2020년 경제 성장률은 -7.7%로 추정됐다고 한다(KOTRA, 2021).

해당 지역의 국가들은 코로나 기간 수요공급이 전례없이 줄어들어 수입에도 영향을 크게 미

친 것으로 나타났다. 카리브 연안 커뮤니티(Caribbean Community, 이하 CARICOM)는 2019년 전반기(1-6월)에 비교해 2020년 동기간 수입이 23.8% 감소했다(CEPAL, 2020).이를 2018년 1-5월과 2020년 동기간 수입을 비교했을 때는 43.1%의 감소로 기록됐다. 구체적으로 우리나라가 포함된 기타 아시아 국가들(중국 별도)에 대한 CARICOM의 수입은 35.0% 감소했다(CEPAL, 2020).

하지만 2021년에는 CARICOM에서 제한적인 회복이 발생해 3.7%의 경제성장이 예상되는 것으로 나타났다(KOTRA, 2021).최근 2021년 10월에는 중남미 경제회복에 맞춰 우리나라가 수출 확대 전략을 펼치면 큰 효과를 거둘 것이라는 분석이 발표됐다. 실제로 2021년 중남미 및 카리브 연안 국가에서 국제시장에서 중남미 및 카리브 상품 수요 증가에 따른 긍정적인 효과와 1차 상품 가격 상승에 따른 수출 증가, 해당 국가들의 총수요 증가 등이 주요한 요인으로 작용해 코로나 충격을 벗어나 경제회복을 시작했다고 보고됐다. 이에 비록 해당 지역의 국가가 완전한 회복을 이루기까지는 여전히 시간이 필요하겠지만, 전통적으로 2019년까지 매년 우리나라의 대중남미 교역이 흑자였던 점을 볼 때 계속해서 대중남미 수출증대에 노력을 하면 좋은 결과를 얻을 수 있을 것이라고 밝혔다(Global economy, 2021).

한편, 본 연구에서 사용한 Node2vec의 방법 같은 경우에는 각 보고 국가들에 해당하는 무역 거래 파트너 국가들이 군집이 형성되는 것을 통해 임베딩이 우수하게 수행된 것을 알 수 있었다. 이상의 내용을 통해 임의 보행 기반의 알고리즘을 수차례 반복함으로써 추출한 확률적인 네트워크의 노드 시퀀스로 전체 네트워크의 구조를 잘 표현할 수 있는 것으로 평가할 수 있었다. 그리고 링크를 예측하는 방법에 대해서는 로지스틱보다 앙상블 기법이 정밀도뿐만 아니라 재현율에서도 성능이 잘 나오는 것을 확인할 수 있었다. 이는 Light GBM이 의사결정나무 기반 학습 알고리즘에 부스팅 방법을 적용함으로써 입력 데이터의 미세한 변화에도 예측 결과를 강건하게 도출해내는 것으로 이해할 수 있었다. 이상의 방법은 본 연구에서 사용한 것과 같이 10,000건 이상의 다량의 데이터를 학습할 때 유용하며 GPU 학습을 지원하기 때문에, 성능의 측면과 학습속도의 측면 모두에서 우리나라의 수출 후보 지역을 탐색하는데 유용할 것이라고 볼 수 있었다.

본 연구에서는 동일한 예측 모델이라고 해도 어떠한 특징을 가진 학습군을 대상으로 모델의 파라미터를 업데이트하는 전략에 따라서 예측의 성능이 다르게 나타날 것이라고 생각했다. 이러한 이유에서 본 연구는 학습군을 무역 거래 규모로 구분하여, 특정 학습군의 설정 없이 전체 데이터를 대상으로 수행할 때와 상위 25% 이상 규모, 그리고 평균 이상의 규모를 갖는 학습군에 대하여 각각 모델링을 수행했다. 그 결과, 다른 경우에 비해 평균 이상의 규모를 갖는 학습군에 대한 예측 성능이 우수하게 나타나는 것을 알 수 있었다. 이를 통해 어떤 집단의 정보를 학습할 것인가는 예측 모델을 설계하는데 생각보다 훨씬 중요하다는 사실을 시사할 수 있었다. 따라서 본 연구의 결과를 바탕으로 우수한 성능을 나타낼 수 있는 학습군에 대한 탐색 단계를 선행해야 함을 연구자 및 실무자들에게 주장한다.

본 연구의 한계는 다음과 같다. 본 연구에서 수행한 예측 모델은 GVC의 네트워크 구조에 대한 위상 정보만 가지고 수행한 예측이기 때문에

국가 간의 특수한 이벤트 등의 통제는 처리할 수 없다는 한계를 갖고 있다. 즉, 미국과 중국의 수출무역 규제와 같이 어떤 국가의 정치적 또는 사회적 이유로 발생한 국제적 갈등에 대한 우리나라의 반사이익을 탐색하기 위해서 본 연구를 그대로 적용하기는 힘들다는 것이다. 만약 상기와 같은 이유에서 링크 예측 모델을 수행하고자 한다면 별도의 모델 구축이 필요하다. 예를 들어, 미국과 중국의 기존 무역 관계가 끊어져 없는 상태로 상정하고 본 연구에서 설명하는 단계와 같이 모델을 새로 구축해야 할 것이다. 이는 분명 어려운 작업은 아니지만, 본 연구에서 설정한 연구질의는 아니었기에 본 연구에서 다루지는 않았다.

마찬가지의 이유에서 본 연구의 결과는 링크 예측확률만 나타낸 것으로, 이러한 잠재적 관계로 벌어들일 수익은 따로 계산하지 않았다. 즉, 임의의 A라는 국가의 링크 예측확률 결과는 가장 높았던 것에 비해 임의의 B라는 국가의 링크 예측확률 결과는 상대적으로 낮지만 거래 규모가 큰 경우가 존재할 수 있다는 점이다. 이러한 부분은 본 연구와 같이 도출한 예측확률 결과에 해당 임의의 국가가 다른 나라와 갖는 무역관계의 평균값을 가중하여 산출하는 기대 값으로 확인할 수 있을 것이다. 해당 부분 역시 어려운 작업은 아니지만, 마찬가지로 본 연구에서 관심있게 살펴본 것은 아직 무역 관계를 갖고 있지는 않지만, 갖게 될 확률이 높은 국가의 탐색에 있었기 때문에 본 연구에서는 특별히 다루지는 않았다.

마지막으로 본 연구의 결과는 산업 수준에 대한 결과로 우리나라의 식음료 산업을 대상으로 도출한 잠재적 수출 후보 국가에 대한 예측을 다루고 있다. 따라서 기업 관점에서 관심을 가질만한 제품 수준에 대한 결과는 다루고 있지 않다. 이렇게 본 연구가 산업수준에서 진행되었기에 기업 이용자에게 제공할 수 있는 시사점에는 분명 한계가 존재하고 있다. 하지만, 본 연구에서 개별 HS 분류에 기초해 제품별로 무역 거래 대상 국가를 다르게 설정하는 방법으로 이를 극복할 수 있을 것으로 생각된다. 이렇게 제품별로 새로 정의된 관계에서 예측 모델의 가중치를 학습한다면, 기업 이용자들이 자사 제품의 잠재적 수출 대상 국가를 탐색하는데도 본 연구의 모델을 활용할 수 있을 것으로 기대된다.

본 연구에서 도출한 최적 링크 예측 모델은 <Table 4>와 같이 0.79의 True Positive Rate (Recall)을 기록했는데, 이상의 성능은 본 연구가 참고한 Tuninetti et al. (2017)의 제안 모델보다 우수한 성능을 나타냈다. Tuninetti et al. (2017)는 ROC curve를 통해 링크 예측 모델의 성능을 평가했는데, 본 연구와 동일한 약 0.05의 False Positive Rate (FPR)에서 Tuninetti et al. (2017)의 모델은 약 0.75의 True Positive Rate (Recall)만을 기록했다.

참고문헌(References)

A. Al-Mudimigh, et al., “Extending the Concept of Supply Chain: The Effective Management of Value Chains,” *International Journal of Production Economics,* Vol.87, No.3(2004), 309~320.

Abreha, K. G., et al., “Coping with the Crisis and Export Diversification,” *The World Economy,* Vol.43, No.5(2020), 1452~1481.

Allan, J., “Virtual water-the water, food, and trade

nexus, Useful concept or misleading metaphor?," *Water international,* Vol.28, No.1(2003), 106~113.

Antonelli, M., and M. Sartori, "Unfolding the potential of the virtual water concept. What is still under debate?," *Environmental Science & Policy,* Vol.50(2015), 240~251.

Batagelj, Vladimir, and Andrej Mrvar, "A Subquadratic Triad Census Algorithm for Large Sparse Networks with Small Maximum Degree," *Social Networks,* Vol.23, No.3(2001), 237~243.

Cadot, O., et al., *Export Diversification: What's Behind the Hump?,* 2007.

CEPAL, *INTERNATIONAL TRADE IN GOODS IN LATIN AMERICA AND THE CARIBBEAN.* 2020, https://www.cepal.org/sites/default/files/publication/files/46518/Boletin_40_ingles.pdf.

CEPAL, *The Effects of the Coronavirus Disease (COVID-19) Pandemic on International Trade and Logistics.* 2020, https://www.cepal.org/sites/default/files/publication/files/45878/S2000496_en.pdf.

Chapagain, A., Hoekstra, A., and H. Savenije, "Water saving through international trade of agricultural products," *Hydrology and Earth System Sciences,* Vol.10, No.3(2006), 455~468.

Christopher, M., *Logistics and Supply Chain Management: Creating Value-Added Networks.* Pearson Education, 2005.

Collier, Paul, and Anthony J. Venables, "Rethinking Trade Preferences: How Africa Can Diversify Its Exports," *The World Economy,* Vol.30, No.8(2007), https://doi.org/10.1111/j.1467-9701.2007.01042.x.

Cox, Andrew, "Power, Value and Supply Chain Management," *Supply Chain Management: An International Journal,* (1999).

da Costa Neto, M. N. C., and R. Romeu, *Did Export Diversification Soften the Impact of the Global Financial Crisis?,* 2011.

de Marchi, V. D., et al., "Environmental Strategies, Upgrading and Competitive Advantage in Global Value Chains," *Business Strategy and the Environment,* Vol.22, No.1(2013), 62~72.

Dennis, A., and B. Shepherd. "Trade Facilitation and Export Diversification." *The World Economy,* vol. 34, no. 1, 2011, pp. 101–22.

di Domenico, C., et al., "Supply Chain Management Analysis: A Simulation Approach of the Value Chain Operations Reference Model (VCOR)," *Advances in Production Management Systems,* (2007), 257~264.

Doroud, M., et al., "The Evolution of Ego-Centric Triads: A Microscopic Approach toward Predicting Macroscopic Network Properties," *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing,* (2011), 172~179.

El Hag, S., and M. El Shazly, "Oil Dependency, Export Diversification and Economic Growth in the Arab Gulf States," *European Journal of Social Sciences,* Vol.29, No.3(2012), 397~404.

Etoday, "[Planting K-Agriculture in the World ①] Agricultural exports soaring thanks to FTA... 'now pioneering the New Northern Region'", 2020, https://www.etoday.co.kr/news/view/1897990.

Falkenmark, M., Rockstrom, J., and J. Rostström, *Balancing water for humans and nature: the new approach in ecohydrology,* Earthscan, 2004.

Gereffi, G., and K. Fernandez-Stark, *Global Value Chain Analysis: A Primer,* Duke University, 2011.

Gereffi, G., and J. Lee, "Why the World Suddenly Cares about Global Supply Chains," *Journal of Supply Chain Management,* Vol.48, No.3 (2012), 24~32.

Giroud, A., and H. Mirza, "Refining of FDI Motivations by Integrating Global Value Chains' Considerations," *The Multinational Business Review,* (2015).

Global Value Chains Center, (2011), https://Global valuechains.Org.

Global Economy, "Latin America, Economic Recovery," 2021, https://news.g-enews.com/view.php?ud=20210930221224191102_7&ssk=g080000&md=20211007000005_R.

Godfray, H., Beddington, J., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., and C. Toulmin, "Food security: the challenge of feeding 9 billion people," *Science,* Vol.327, No.5967(2010), 812~818.

Grover, Aditya, and Jure Leskovec. "Node2vec: Scalable Feature Learning for Networks," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* (2016), 855~864.

Hao, Wei, et al., "Bond Transaction Link Prediction Based on Dynamic Network Embedding and Time Series Analysis," *2019 6th International Conference on Systems and Informatics, ICSAI 2019,* Vol. Icsai, (2019), 1471~1477, https://doi.org/10.1109/ICSAI48974.2019.9010471.

Hwang Seo-young. "'Aloe Beverage' is attacking Central American markets such as Panama." Food and Beverage Newspaper, 25 Apr. 2018.

Herzer, D., and D. F. Nowak-Lehnmann, "What Does Export Diversification Do for Growth?: An Econometric Analysis," *Applied Economics,* Vol.38, No.15(2006), 1825~1838.

Hesse, H., "Breaking into New Markets: Emerging Lessons for Export Diversification," *Export Diversification and Economic Growth,* (2009), 55~80.

Ke, Guolin, et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Advances in Neural Information Processing Systems,* Vol.30(2017), 3146~3154.

Kim Sung-Hoon, et al., *Analysis of the impact of FTA negotiations on processed food and the food and restaurant industry,* 2018.

Min Sunghwan, et al., Analysis of export diversification patterns of Korean industry, 2011.

Mudambi, R., and J. Puck, "A Global Value Chain Analysis of the 'Regional Strategy' Perspective," *Journal of Management Studies,* Vol.53, No.6(2016), 1076~1093.

Oki, T., and S. Kanae, "Virtual water trade and world water resources," *Water Science and Technology,* Vol.49, No.7(2004), 203~209.

Ortmann, Mark, and Ulrik Brandes, "Efficient Orbit-Aware Triad and Quad Census in Directed and Undirected Graphs," *Applied Network Science,* Vol.2, No.1(2017), 1~17.

Park Kang-wook. "The economic growth of Latin America and the Caribbean is expected to be 3.7%." KOTRA Overseas Market News, 2021, https://news.kotra.or.kr/user/globalAllBbs/kotranews/album/2/globalBbsDataAllView.do?dataIdx=186882.

Patel, Rushabh, and Yanhui Guo, “Graph Based Link Prediction between Human Phenotypes and Genes,” *ArXiv,* (2021), 1~13.

Perozzi, Bryan, et al., “Deepwalk: Online Learning of Social Representations,” *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* (2014), 701~710.

Stabell, C. B., and Ø. D. Fjeldstad, “Configuring Value for Competitive Advantage: On Chains, Shops, and Networks,” *Strategic Management Journal,* Vol.19, No.5(1998), 413~437.

Tilman, D., Balzer, C., Hill, J., and B. L., Befort, “Global food demand and the sustainable intensification of agriculture,” *Proceedings of the national academy of sciences,* (2011), 20260~20264.

Tran, Thi Anh Dao, et al., “Global Value Chains and the Missing Link between Exchange Rates and Export Diversification,” *International Economics,* Vol.164(2020), 194~205, https://doi.org/10.1016/j.inteco.2020.10.001.

Tuninetti, Marta, et al., “To Trade or Not to Trade: Link Prediction in the Virtual Water Network,” *Advances in Water Resources,* Vol. 110(2017), 528~537, https://doi.org/10.1016/j.advwatres.2016.08.013.

Uddin, S., et al., “Triad Census and Subgroup Analysis of Patient-Sharing Physician Collaborations,” *IEEE Access,* Vol.6(2018), 72233~72240.

UN Trade Statistics, “What Is UN Comtrade?,” 2006, https://unstats.un.org/unsd/tradekb/knowledgebase/50075/what-is-un-comtrade.

Yang, H., Wang, L., Abbaspour, K., and A. Zehnder, “Virtual water trade: an assenssment of water use efficiency in the international food trade,” *Hydrology and Earth System Sciences,* Vol.10, No.3(2006), 443~454.

Yeo Take-dong, and Ki Seok-do, “Korea’s Trade Strategy and Promising Export Items in the Visegrad Group Market,” E-Trade Research, Vol.18, No.1(2020), 141~166.

Abstract

A Study on Searching for Export Candidate Countries of the Korean Food and Beverage Industry Using Node2vec Graph Embedding and Light GBM Link Prediction*

Jae-Seong Lee** · Seung-Pyo Jun*** · Jinny Seo****

This study uses Node2vec graph embedding method and Light GBM link prediction to explore undeveloped export candidate countries in Korea's food and beverage industry. Node2vec is the method that improves the limit of the structural equivalence representation of the network, which is known to be relatively weak compared to the existing link prediction method based on the number of common neighbors of the network. Therefore, the method is known to show excellent performance in both community detection and structural equivalence of the network. The vector value obtained by embedding the network in this way operates under the condition of a constant length from an arbitrarily designated starting point node. Therefore, it has the advantage that it is easy to apply the sequence of nodes as an input value to the model for downstream tasks such as Logistic Regression, Support Vector Machine, and Random Forest. Based on these features of the Node2vec graph embedding method, this study applied the above method to the international trade information of the Korean food and beverage industry. Through this, we intend to contribute to creating the effect of extensive margin diversification in Korea in the global value chain relationship of the industry. The optimal predictive model derived from the results of this study recorded a precision of 0.95 and a recall of 0.79, and an F1 score of 0.86, showing excellent performance. This performance was shown to be superior to that of the binary classifier based on Logistic Regression set

* This study was supported by the Ministry of Trade, Industry and Energy and Korea Evaluation Institute of Industrial Technology(KEIT) in 2021(20009398)

** Ph.D student, University of Science & Technology, 217 Gajeong-ro, Yuseong-gu, Daejeon, Korea
Tel: +82-2-3299-6122, Fax: +82-2-3299-6041, E-mail: jslee@kisti.re.kr

*** Principal researcher, Korea Institute of Science and Technology Information, 85 Hoegi-ro, Dongdaemun-gu, Seoul 130-722, Korea / Professor, University of Science & Technology, 217 Gajeong-ro, Yuseong-gu, Daejeon, Korea
Tel: +82-2-3299-6095, Fax: +82-2-3299-6041, E-mail: spjun@kisti.re.kr

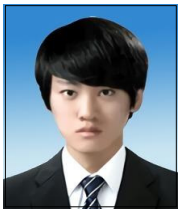
**** Corresponding Author: Jinny Seo
Principal researcher, Korea Institute of Science and Technology Information
85 Hoegi-ro, Dongdaemun-gu, Seoul 130-722, Korea
Tel: +82-2-3299-6056, Fax: +82-2-3299-6041, E-mail: jinny@kisti.re.kr

as the baseline model. In the baseline model, a precision of 0.95 and a recall of 0.73 were recorded, and an F1 score of 0.83 was recorded. In addition, the light GBM-based optimal prediction model derived from this study showed superior performance than the link prediction model of previous studies, which is set as a benchmarking model in this study. The predictive model of the previous study recorded only a recall rate of 0.75, but the proposed model of this study showed better performance which recall rate is 0.79. The difference in the performance of the prediction results between benchmarking model and this study model is due to the model learning strategy. In this study, groups were classified by the trade value scale, and prediction models were trained differently for these groups. Specific methods are (1) a method of randomly masking and learning a model for all trades without setting specific conditions for trade value, (2) arbitrarily masking a part of the trades with an average trade value or higher and using the model method, and (3) a method of arbitrarily masking some of the trades with the top 25% or higher trade value and learning the model. As a result of the experiment, it was confirmed that the performance of the model trained by randomly masking some of the trades with the above-average trade value in this method was the best and appeared stably. It was found that most of the results of potential export candidates for Korea derived through the above model appeared appropriate through additional investigation. Combining the above, this study could suggest the practical utility of the link prediction method applying Node2vec and Light GBM. In addition, useful implications could be derived for weight update strategies that can perform better link prediction while training the model. On the other hand, this study also has policy utility because it is applied to trade transactions that have not been performed much in the research related to link prediction based on graph embedding. The results of this study support a rapid response to changes in the global value chain such as the recent US-China trade conflict or Japan's export regulations, and I think that it has sufficient usefulness as a tool for policy decision-making.

Key Words : Node2vec, Graph Embedding, Light GBM, Link Prediction, Global Value Chain

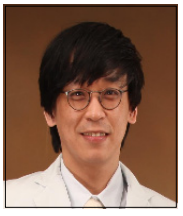
Received : October 30, 2021 Revised : November 22, 2021 Accepted : November 30, 2021
Corresponding Author : Jinny Seo

저 자 소 개



이 재 성

과학기술연합대학원대학교 과학기술경영정책학과에서 석사 및 박사과정을 수료하고 현재 한국과학기술정보연구원 데이터분석본부 학생연구원으로 재직 중이다. Government Information Quarterly, Technological forecasting and social change, 지능정보연구, 한국기술혁신학회, 지식재산연구, 통계연구 등 다수의 논문을 게재하였다. 주요 관심 연구분야로는 인공지능 윤리 실현을 위한 기술적 접근법, 인공지능을 활용한 맞춤형 정부지원 또는 국가 연구개발과제 관리 등으로, 국가 혁신성과 제고를 위한 데이터 기반의 다양한 정책 연구를 수행했다.



전 승 표

KAIST에서 경영학으로 석사학위를 취득하고, 고려대학교에서 과학관리학 전공으로 이학박사를 취득했다. 현재 한국과학기술정보연구원 글로벌R&D분석 센터에 책임연구원으로 재직 중이며, 과학기술연합대학원대학교(UST) 과학기술경영정책과 교수로 재직중이다. Technological forecasting and social change, Government Information Quarterly, Scientometrics, Energy policy, Internet research 등 해외학술지와 지능정보연구, 한국기술혁신학회 등 국내학술지에 주저자로 다수의 논문을 게재했다. 주요 관심분야는 빅데이터를 활용한 수요 예측, 글로벌 R&D 동향 분석, 유망 기술 탐색, 기술가치평가, 중소기업 R&D 정책 등을 위한 지능형 정보 시스템 개발 연구이다.

서 진 이

한국과학기술정보연구원 글로벌R&D분석센터 책임연구원으로 재직중이며, 이화여자대학교 멀티미디어학으로 석사과정을 마쳤다. 데이터베이스 구축, 관리 및 데이터 분석을 통한 산업, 시장분석을 수행하였다. 주요 관심분야는 빅데이터를 통한 산업, 시장분석, 기술혁신 분석 및 기술가치평가, 경제성 평가이다.