KJCEM 20. 6. 107~116
November 30, 2019

www.kicem.or.kr (pISSN 2005-6095 / eISSN 2465-9703)
http://dx.doi.org/10.6106/KJCEM.2019.20.6.107

세계은행 공적개발원조사업의 엔지니어링 기업 간 협력관계 예측모델 개발

유영수[1] · 구본상* · 이관훈[2] · 한승헌[3]
[1]서울과학기술대학교 건설시스템공학과 · [2]고려대학교 컴퓨터학과 · [3]연세대학교 건설환경공학과(現, 한국건설기술연구원장)

Predicting Cooperative Relationships between Engineering Companies in World Bank's ODA Projects

Yu, Youngsu[1], Koo, Bonsang*, Lee, Kwanhoon[2], Han, Seungheon[3]
[1]Department of Civil Engineering, Seoul National University of Science and Technology
[2]Department of Computer Science and Engineering, Korea University
[3]Department of Civil and Environmental Engineering, Yonsei University

Abstract : Korean construction engineering firms want to pave the way for expansion of overseas markets through the World Bank's Official Development Assistance (ODA) projects as a way to improve their overseas project performance. However, since the World Bank project competes with global companies for limited projects, building partnerships with suitable business partners is essential to gain an upper hand in bidding competition and meet the institutional conditions of the recipient country. In this regard, many network studies have been conducted in the past through Social Network Analysis (SNA), but few have been analyzed based on the process of changes in the network. So, This study collected winning data from the three Southeast Asian countries that ended after the World Bank's ODA project performed smoothly, and established a learning-based link prediction model that reflected the dynamic nature of the network. As a result, the 11 main variables acting on building a cooperative relationship between winning companies were derived and the effect of each variables on the probability value of cooperation between individual links was identified.

Keywords : International Cooperative Strategies, World Bank ODA, Link Prediction, XGBoost

1. 서론

1.1 연구의 배경 및 목적

국내 엔지니어링 수주실적은 2018년 7조 4천억 원으로 전년 대비 12% 증가하였다. 특히, 해외사업 수주 금액은 전년도보다 81% 증가하며 2016년 이후 상승하는 추세를 보이고 있다. 하지만 실제로 엔지니어링 사업의 수주 실적 증가는 대부분 비건설(원자력, 설비, 기계, 환경) 부문에 집중되어 있으며, 건설 부문의 경우 국내 실적은 증가하였으나 해외 실적은 감소한 것으로 나타났다(KENCA, 2019).

이에 해외 실적 향상을 위해 건설 엔지니어링 기업은 다방면으로 해외시장 진출 방안을 모색하고 있으며, 그 중 다자간개발은행(Multi-lateral Development Bank)인 세계은행(World Bank)의 공적개발원조(Official Development Assistance; 이하 ODA) 사업을 통한 해외시장 확장의 발판을 마련하고자 한다.

세계은행 ODA 사업은 재정적으로 안정적이며, 수원국에서 경험을 쌓아 향후 해당 국가의 민간 부문으로 진출할 수 있는 기회를 제공한다는 장점이 있다. 하지만 기술 · 가격적 평가요소를 바탕으로 유수의 글로벌 엔지니어링 기업과 한정된 사업 내에서 경쟁하기 때문에(Koo et al., 2017), 기업의 약점을 보완해줄 파트너와의 협력관계를 통한 전략적 입찰참여가 필요하다. 또한 입찰 조건으로 현지기업과의 협력을 입찰 조건으로 제시하거나 낙찰 가점으로 반영되는 국가가 있기 때문에 현지기업과의 협력이 필수적이다(Lee et al., 2018). 이에 국내기업은 사업 파트너로 적합한 글로벌 · 현지기업을 선별하고, 그들과 협력관계를 구축하는 전략적 접근을 통해 입찰 경쟁력 향상을 도모해야 한다.

과거 사업 데이터를 이용한 협력관계 분석을 통해 적합한

* Corresponding author: Koo, Bonsang, Department of Civil Engineering, Seoul National University of Science and Technology, Seoul 01811, Korea
E-mail: bonsang@seoultech.ac.kr
Received September 17, 2019: revised October 25, 2019
accepted November 1, 2019

사업파트너 후보를 추천하거나, 협력 전략 구축을 통한 시장 진출 전략을 제공하는 것이 국내 기업의 해외 시장 입찰경쟁력 향상에 도움이 될 수 있다. 이와 관련하여 사회 네트워크 분석(Social Network Analysis; 이하 SNA)을 중심으로 과거 네트워크의 특성분석 혹은 영향인자 도출과 같은 연구가 다수 진행된 바 있다(Chinowsky et al., 2008; Lee et al, 2018; Liu et al., 2015).

그러나 세계은행 ODA 사업과 같은 건설 사업은 일반적인 기업이 구축하는 장기적 협력관계와 달리, 특정 사업을 위해 맺어진 일시적이고 수명주기가 제한적인 관계이기 때문에 그 관계의 변화 과정에 집중할 필요가 있으나, 이에 필요한 시계열 기반의 분석을 실시한 연구는 드물다(Zheng, X. et al., 2016).

이에 본 연구에서는 세계은행이 발주한 ODA 엔지니어링 사업 낙찰데이터를 기반으로 과거 네트워크의 동적 변화를 반영한 협력관계의 예측모델을 개발하고자 하였다.

1.2 연구의 범위 및 방법

본 연구는 세계은행 ODA 사업이 원활히 진행되었고, 현재 지원 종료단계인 아시아 3개국(베트남, 인도네시아, 방글라데시)을 분석 대상으로 삼았다. 본 연구의 과정은 아래와 같다.

1) 1단계: 세계은행 ODA 엔지니어링 사업 3개국 데이터 취합

분석 데이터 구축을 위해 세계은행이 제공하는 'World Bank Finance'오픈 데이터베이스를 사용하였으며, 국내 건설 엔지니어링 기업이 주로 참여하는 설계 컨설팅 서비스로 데이터를 국한하였다. 해당 데이터에는 2004년부터 2017년까지 발주된 사업에 대한 프로젝트, 사업, 낙찰자 단위의 데이터가 수록되어 있다.

2) 2단계: 기간 분할 및 데이터 변환 후 네트워크 구축

네트워크 변화 과정을 분석에 반영하기 위해 구축할 예측모델의 학습、평가데이터는 시간의 흐름을 기준으로 분류해야한다. 이에 각 국가별 사업 참여 건수를 바탕으로 데이터를 3개 기간으로 분할하였다. 이후 데이터를 기존 노드 단위에서 상호 간 관계를 나타내는 링크 단위로 변환하였다. 이를 통해 3개 기간에 대한 네트워크를 구축하고, 네트워크 중심성 지표를 추출하였다.

3) 3단계: 네트워크 변수 수집과 취합 및 학습데이터 구축

링크예측모델 구축을 위해 변환한 링크 단위 데이터를 대상으로 3개의 기간에 대해 1)데이터 기반(10개)과 2)위상정보 기반(10개) 두 방향으로 총 20개의 변수를 작성하였다. 이후 링크예측 학습모델 구축을 위해 현 단계의 입력변수와 다음 단계의 목적변수를 교차하여 최종 학습 데이터를 구축하였다.

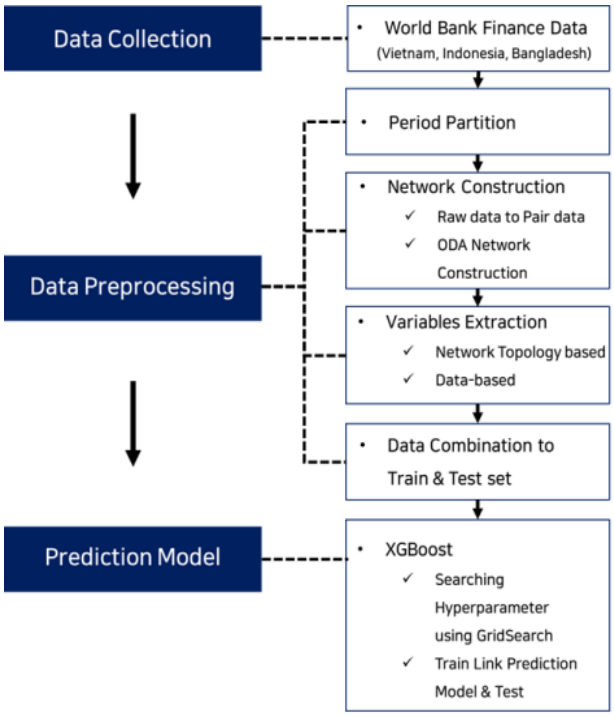

Fig. 1. Research Process

4) 4단계: XGBoost기반 링크예측 모델 구축 및 해석

최종 학습 데이터를 기반으로 XGBoost 알고리즘을 이용하여 학습모델을 구축하였다. 학습모델 구축은 R의 'xgboost'패키지를 사용하였다. 이를 통해 예측모델을 구축하고, 모델의 변수 의존도를 나타내는 'Feature Importance'를 도출하여 협력관계 형성에 유의한 요인을 확인하였다. 또한, R의 'xgboost Explainer'패키지가 제공하는 Waterfall Chart를 통해서 개별 사례의 예측 확률 도출 과정을 해석하였다.

2. 선행 연구 및 이론적 고찰

2.1 선행 연구 고찰

사업 수주 기회 확보와 안정성을 위해 기업들은 그들 간의 제휴를 통해 시장 진출을 도모하고 있다. 이러한 협력관계로부터 시장 진출 전략을 구축하기 위해 기존에는 SNA를 중심으로 다수의 연구가 이루어져왔다.

Liu et al. (2015)은 건설프로젝트 내에서 협력네트워크가 구조적으로 어떻게 변화하는지 분석하였다. 그 결과 네트워크의 크기 증가에 따른 구조적 변화가 발생하며, 이 과정에서 구성원 간 경로 길이(path length)가 감소하며 상호 협력 가능성이 증가하는 것으로 나타났다.

Lee et al. (2018)은 ODA 입 · 낙찰 데이터를 이용하여 SNA 분석을 수행하였다. 이후 SNA 지표 값을 변수로 사용

하여 로그 회귀분석을 수행함으로써 입찰 경쟁력을 향상 시킬 수 있는 공종별 파트너십 전략을 구축하였다.

이와 같이 SNA는 네트워크 구조 내 노드의 위상과 역할 등을 도출하는데 유용하다. 그러나 네트워크의 시간에 따르는 진화 양상과 이를 기반으로 향후 노드 간의 관계를 예측하는데 한계가 존재한다. 이러한 단점을 보완하는 방안으로 최근에는 기계학습 기반으로 노드가 아닌 '링크'를 예측하는 방법과 관련 연구가 등장하고 있다.

일례로 Mori, J. et al. (2012)은 SVM (Support Vector Machine) 알고리즘 기반 링크예측을 통해 잠재적 사업 파트너 추천 시스템을 개발하였다. 이를 통해 기업의 직원 수, 순위, 설립 일자가 상호 간 협력관계를 구축하는데 주요 요인임을 확인하고 향후 형성될 상호관계를 예측하였다.

Seo (2018)는 특허 정보에 링크예측을 적용하여 제조-서비스 융합 현상을 분석하고 향후 새롭게 융합될 기술의 동향을 예측하였다. 그 결과 132쌍의 제조-서비스 기술융합 관계 발생을 예측하였고, 이를 통해 새로운 기술융합 분야에 대한 전략 구축 방안을 제시하였다.

그러나 건설 분야에서는 아직까지 SNA 기반 링크예측을 시도한 연구는 매우 드문 것으로 파악된다. 특히 ODA 사업 데이터를 기반으로 엔지니어링 기업 간 협력 여부를 예측하는 연구는 처음 시도하는 것으로 연구의 필요성에 무게를 두고 있다.

2.2 이론적 고찰

2.2.1 사회 네트워크 링크예측

링크예측은 현재 주어진 네트워크를 기반으로 미래의 네트워크에 새롭게 추가되거나 사라질 링크를 예측하는 것으로 네트워크의 동적변화를 분석하는데 사용된다(Wang et al., 2015). 본 기법은 사회연결망 분석에 주로 사용되는 노드(node)의 네트워크 중심성(network centrality) 지표 대신 그들 간의 관계를 나타내는 엣지(edge) 또는 링크(link)를 기반으로 한다. 분석 방법은 노드 기반, 위상 기반, 학습 기반으로 구분할 수 있으며, 위상 기반과 학습 기반 링크예측이 대표적인 방법이다.

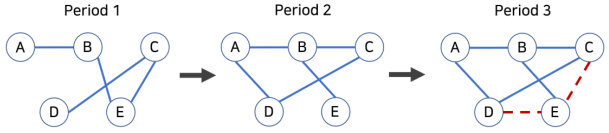

Fig. 2. Link Prediction

위상 기반 링크예측은 위상적 정보(topological information)를 기반으로 노드 사이의 유사도를 측정하여, 유사도의 임계치(cutoff) 이상의 값에서 링크가 연결될 것으로 예측하는 방법이다. 유사도 값은 이웃 노드가 연결된 링크의 수를 바탕으로 다양한 계산 공식을 바탕으로 도출되며, 본 연구에서 산정된 개별 유사도와 그 범위를 〈Table 1〉에 나타내었다. 해당 표에서 $\Gamma(x)$는 노드 x의 이웃들 집합을 나타내며, $|\Gamma(x)|$는 노드 x의 이웃 개수를 의미한다.

Table 1. Network Topology Variables & Definition

Topology	Abbreviation	Definition	Range
Preferential Attachment	PA	$PA(x,y)=\|\Gamma(x)\| \cdot \|\Gamma(y)\|$	0~158
Adamic/Adar	AA	$AA(x,y)=\sum_{Z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{\log\|\Gamma(Z)\|}$	0~10.40
Resource Allocation	RA	$RA(x,y)=\sum_{Z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{\|\Gamma(Z)\|}$	0~1.58
Common Neighbors	CN	$CN(x,y)=\|\Gamma(x)\cap\Gamma(y)\|$	0~9
Sorensen Index	SI	$SI(x,y)=\frac{\|\Gamma(x)\cap\Gamma(y)\|}{\|\Gamma(x)\|+\|\Gamma(y)\|}$	0~0.5
Salton Cosine Similarity	SC	$SC(x,y)=\frac{\|\Gamma(x)\cap\Gamma(y)\|}{\sqrt{\|\Gamma(x)\|\cdot\|\Gamma(y)\|}}$	0~1
Leicht-Holme_Nerman	LHN	$LHN(x,y)=\frac{\|\Gamma(x)\cap\Gamma(y)\|}{\|\Gamma(x)\|\cdot\|\Gamma(y)\|}$	0~1
Hub Promoted	HP	$HP(x,y)=\frac{\|\Gamma(x)\cap\Gamma(y)\|}{\min(\|\Gamma(x)\|,\|\Gamma(y)\|)}$	0~1
Hub Depressed	HD	$HD(x,y)=\frac{\|\Gamma(x)\cap\Gamma(y)\|}{\max(\|\Gamma(x)\|,\|\Gamma(y)\|)}$	0~1
Jaccard's Coefficient	JC	$JC(x,y)=\frac{\|\Gamma(x)\cap\Gamma(y)\|}{\|\Gamma(x)\cup\Gamma(y)\|}$	0~1

두 번째 방법으로 학습 기반 링크예측은 링크 연결 유무에 대한 이진분류(binary classification) 문제로, 노드 간 데이터를 바탕으로 각종 학습 알고리즘을 이용하여 효율적인 분류 예측 모델을 구축하는 방법이다. 이는 각종 데이터를 기반으로 효율적인 예측모델 구축이 가능하여 최근 머신 러닝 또는 딥러닝을 활용하여 예측하는 연구가 등장하고 있다.

2.2.2 XGBoost

XGB (eXtreme Gradient Boosting; 이하 XGB)는 Chen and Guestrin(2016)에 의해 소개된 빠른 연산과 좋은 유연성이 장점인 알고리즘으로, 캐글(Kaggle)[1]의 많은 우승자들이 이 알고리즘을 빠른 속도로 로직을 검증하기 좋은 모델로 평가하였다.

XGB는 여러 개의 CART (Classification and Regression Trees)모델로 구성된 tree ensemble을 기반으로 다수의 분류기(tree)를 만들고, 부스팅(gradient boosting)을 통해 분류기 별 비중(weights)을 최적화하여 최적의 분류 모델을 찾는 알고리즘이다.

XGB에 사용되는 CART 모델은 기존 의사결정나무(decision

1) 통계문제를 놓고 통계분야 종사자들이 경쟁하는 온라인 플랫폼

tree)는 리프(leaf)에 결정값(decision value)만 존재하는 것과 달리 각 리프별로 점수가 할당되어 모델의 최종 결과에 연관되는 방식으로 더 좋은 분류기를 만들 수 있다.

이후 과정인 부스팅은 Gradient Descent 알고리즘을 이용하며, 이전모델의 error를 최소화하는 새로운 모델을 생성 후 조합함으로써 최적의 분류모델을 구축한다. 이 과정이 병렬 처리로 진행되어 XGB의 연산속도를 빠르게 한다.

이 같은 일련의 과정에서 XGB는 자동 가지치기(Greedy-algorithm)를 사용한 과적합(overfitting) 방지, 그리고 다양한 커스텀 최적화 옵션을 통한 예측성능 향상에 유리하여 본 연구에 적합한 기계학습 모델로 선정하였다.

2.2.3 민감도(sensitivity)

학습모델을 평가하는 다수의 방법 중 본 연구와 같은 분류문제는 민감도와 특이도를 채택한다(Kim, 2017). 여기서 민감도는 양성에 얼마나 민감한지를 나타내는 지표로 전체 양성 수 대비 예측을 통해 맞춘 양성의 수로 산출하고, 특이도는 전체 음성 수 대비 맞춘 음성의 수로 산출하는 방식이다.

분류문제 내에서도 연구 목적에 따라 특정 평가 방법을 선정해야한다. 특히 질병 진단이나 보안과 관련된 연구(Frederickson & Laporte, 2002; Klinkman et al., 1998)에서는 음성을 양성으로 판정하는 것(1종 오류)보다 실제 양성 값을 탐지하지 못함(2종 오류)으로 인한 비용이 크기 때문에 정확도가 줄어들더라도 민감도를 주된 평가방법으로 사용한다.

세계은행 ODA와 같은 건설 산업은 특성상 협력 가능 횟수가 한정되어 있기 때문에 비협력 사례 대비 협력 사례가 매우 극소수이다. 이로 인해, ODA 데이터 기반의 학습모델은 과거 사례를 바탕으로 향후 협력할 가능성이 있는 모든 후보자를 검출하지만, 실제로 그 사례는 소수이기 때문에 모델의 정확도가 하락하는 현상이 발생한다. 따라서 본 연구에서는 많은 후보자를 선별하여 정확도가 하락하더라도, 실제 협력사례 만큼은 정확하게 맞추는 민감도를 주된 모델 평가 방법으로 설정하였다.

3. 세계은행 데이터 및 전처리

3.1 세계은행 데이터 개요

세계은행은 Open Financial Data[2]를 통해 금융 조달을 비롯한 프로젝트, 예산, 자기 자본 등의 정보를 공개하고 있으며, 특히 조달(procurement) 항목의 'Major Contract Awards'에 각국에서 시행된 세계은행 주관 ODA 사업 관련 데이터가 수록되어있다.

해당 데이터는 1) 프로젝트 정보(수원국, 프로젝트명, 조달기관), 2) 사업 정보(사업 명, 조달방식, 공종), 3) 낙찰자 정보(낙찰기업, 낙찰기업국적, 계약금액) 등의 세부정보를 포함하고 있다. 본 분석에서는 세계은행 ODA 사업이 종료된 아시아 3개국(베트남, 인도네시아, 방글라데시)을 주 대상으로 삼았으며, 그 외 세부정보 별 분석에 사용한 범위는 〈Table 2〉와 같다. 또한 본 연구는 협력 네트워크의 형성 요인을 주된 분석 목적으로 삼았기 때문에 최종적으로는 협력 형태로 사업에 참여한 경우만을 대상으로 분석을 실시하였다. 최종적으로 분석에 사용된 데이터는 273개의 기업이 협력 형태로 471회 참여한 182건의 사업 데이터이다.

Table 2. Analysis Scope by Category

Category	Analysis Scope
Borrower Country	Vietnam
	Bangladesh
	Indonesia
Procurement Type	Management / Technical Assistant
	Feasibility Studies
	Project Management
	Construction Supervision
	Procurement Technical Assistant
	Design Service
	Geological / Geophysical Services
Procurement Method	Quality and Cost-Based Selection
	Quality Based Selection
Major Sector	Water / Sanit. / Waste
	Transportation
	Energy & Extractives

3.2 기간 분할

본 연구는 과거시점의 요인이 미래의 협력 요인에 미치는 영향에 관한 학습모델 구축을 목적으로 한다. 따라서 학습모델 훈련과 평가에 필요한 학습、평가데이터는 시간의 흐름을 기준으로 분류되며, 이에 각 국가별 사업 참여 건수를 바탕으로 데이터를 3개 기간으로 분할하였다〈Table 3〉.

Table 3. Period Division and Participation Count by Country

Country		Period1	Period2	Period3
Vietnam	Year	2003~.10	2011~.14	2014~.17
	# of Participation	77	76	60
	# of Enterprise	51	46	49
Indonesia	Year	2003~.05	2006~.11	2012~.17
	# of Participation	61	66	59
	# of Enterprise	41	56	42
Bangladesh	Year	2003~.09	2010~.12	2013~.17
	# of Participation	29	20	23
	# of Enterprise	19	17	21

2) https://finances.worldbank.org/

3.3 네트워크 구축

본 분석의 기본 단위는 링크로, 사업을 기준으로 참여한 기업 목록을 나타내는 기존 데이터로부터의 변형이 필요하다. 〈Fig. 3〉에 나타낸 바와 같이 이를 링크 단위로 변경하기 위해 pair data 형태로 변형하였으며, 해당 데이터는 링크가 연결되었는지 나타내는 여부를 이진 분류로 나타낸다. 이 방식으로 데이터는 273개의 기업에 대한 네트워크 조합인 $_{273}C_2$로 각 기간 별 37,128개이다.

이후 사회 네트워크 분석 프로그램인 'Gephi'를 이용하여 네트워크를 구축하고, 그로부터 각 기간의 노드별 4가지 중심성 지표(degree, closeness, betweenness, eigen-vector centrality)를 추출하였다.

3.4 변수 추출

링크예측 학습모델은 〈Fig. 3〉의 pair data에 나타난 개별 링크가 하나의 데이터 단위이다. 학습모델 변수는 각 링크별로 필요하며, 이에 3개의 기간에 대해 1)데이터 기반(10개)과 2)위상정보 기반(10개) 두 방향으로 총 20개의 변수를 작성하였다.

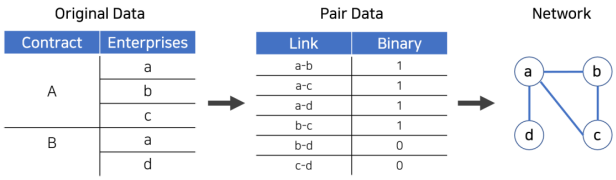

Fig. 3. Link Prediction

데이터 기반 변수는 기업 정보와 사회 네트워크 중심성 지표(social network centrality)로 다시 세분화된다. 기업 정보는 각 기업의 사업수행 경험 실적, 평균 수주금액, Joint Venture 내에서의 위상 등과 같이 세계은행 데이터베이스에 공개된 있는 정보를 개별 노드(기업) 값으로 입력한 후, 두 노드의 차이 값을 링크 변수로 작성하였다(〈Table 4〉 참조). 또한, 3.3절의 네트워크 구축 후 도출 가능한 4가지 사회 네트워크 중심성 지표는 SNA에서 정의된 산정 공식을 활용하여 추출하였다.

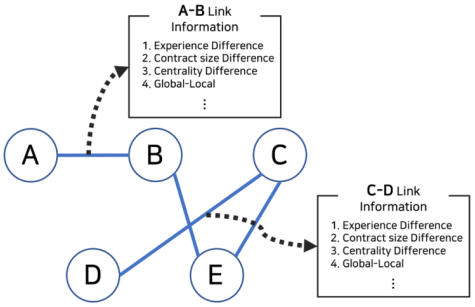

Fig. 4. Link Variables Extraction

위상정보 기반 변수는 〈Table 1〉에서 명시한 개별 링크의 10가지 Network Topology 값을 직접 계산하여 링크 특성값으로 사용하였다.

Table 4. Data-based Variables

No.		Variable	Definition	Class	Range
Difference between Enterprises (from Finance Data)	1	Established Year Difference	Difference in the year of establishment between the two enterprises	Numeric	0~136
	2	Global-Local	Nationality attribute of the link(combination of global&local enterprises)	Factor	0~1
	3	Position Difference	Position in JV (Leader, Sub)	Factor	0~1
	4	Experience Difference	Differences in the number of WB ODA projects	Numeric	0~8
	5	Enterprise Size Difference	Difference in the number of employees	Numeric	0~1
	6	Contract Size Difference	Difference in the size of contract that are primarily involved	Numeric	0~13,118,866
SNA Centrality (from ODA Network) SNA	7	Degree Centrality Difference	Difference in the sum value of the other nodes that one node is connected within the connection	Numeric	0~7
	8	Closeness Centrality Difference	Difference in distance between all indirectly connected nodes	Numeric	0~1
	9	Betweenness Centrality Difference	Difference in the degree to which one node is located between the other nodes	Numeric	0~3,354
	10	Eigen-vector Centrality Difference	Difference in the weighted centrality of the other nodes associated with one node	Numeric	0~1

3.5 학습모델 데이터 구축

링크예측 학습모델의 목적은 현 단계 또는 기간 입력변수로 동일 단계 협력여부를 예측하는 것이 아닌, 그 다음 기간의 협력여부를 예측하는 것이다. 따라서 훈련、평가 데이터 구축 시 현 단계의 입력변수와 그 다음 단계의 목적변수를 교차하여 구축해야 한다.

〈Fig. 5〉와 같이 Period 1에서의 협력 여부인 목적변수(y_1)를 바탕으로 위상정보를 추출하여 해당 기간 입력변수(x_1)로 작용한다. 즉 x_1은 당해기간 실적으로부터 도출되는 값으로, 이 값이 Period 2의 목적변수(y_2)에 미치는 영향을 바탕으로 학습 모델을 생성하는 방식이다.

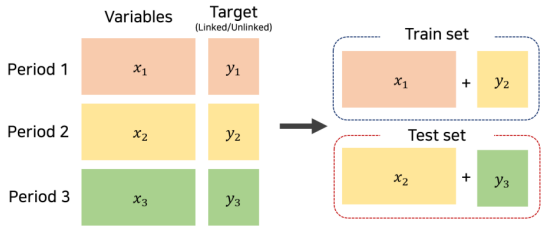

Fig. 5. Data Combination to Train&Test set

따라서 Period 1의 입력변수인 x_1과 Period 2의 목적변수인 y_2로 학습데이터를 구축하여 학습하고, Period 2의 입력변수인 x_2과 Period 3의 목적변수 y_3로 평가데이터로 사용한다.

상기 과정을 앞서 구축된 ODA 기간별 네트워크에 적용하여 〈Table 5〉와 같이 학습 · 평가 데이터를 구분하였으며, 이 때 훈련데이터와 평가데이터의 협력 횟수는 각각 134회, 232회이다.

Table 5. Link Prediction Dataset(Original)

Dataset		Linked	Unlinked	Total
Train	Count	134	36,994	37,128
	%	0.36%	99.64%	100.00%
Test	Count	232	36,896	37,128
	%	0.62%	99.38%	100.00

3.6 데이터 불균형 조정

클래스의 비율이 너무 차이 나면 단순히 우세한 클래스를 택하는 모형의 정확도가 높아지므로 학습모델의 분류능력이 의도한 방향과 달라질 수 있다.

특히 2.3절에 언급한 바와 같이 훈련데이터에서 협력(linked) 사례가 전체의 0.36%(〈Table 5〉 참고)로 탐색하고자 하는 목적함수의 수가 극소수인 경우 정확도가 높아도 민감도가 급격히 작아지는 비대칭 데이터 문제(imbalanced data problem)가 발생한다. 즉, 정확도를 높이는 방향으로 작성되는 학습 모델 특성 상 전체를 비협력(unlinked)으로 선별하는 방향으로 구축될 가능성이 높아 본 연구의 목적인 협력사례를 탐색해내는 모델 형성이 어려워질 수 있다.

이를 해결하는 방식으로는 1)다수의 클래스 중 일부만 추출하여 사용하는 언더 샘플링(under sampling), 2)소수 클래스를 증가시키는 오버 샘플링(over sampling)이 있다. 일반적으로 학습 알고리즘은 데이터 수가 많을수록 효과적이기 때문에 오버샘플링 기법을 택한다(Yap et al., 2014).

본 연구에서는 오버샘플링 기법으로 가장 많이 사용되고 있는 SMOTE (Synthetic Minority Over-sampling Technique)를 사용하였다(Chawla et al., 2002). SMOTE는 최근접 인근 알고리즘을 활용하여 소수 데이터 사이를 보간하여 새로운 데이터를 생성하는 방식이다. 그 결과를 〈Table 6〉에 나타내었으며, 기존 훈련데이터의 협력건수가 134건(0.36%)에서 비협력건수와 동일한 36,994건(50.00%)로 증가되었다.

Table 6. Link Prediction Dataset(Over-sampled)

Dataset		Linked	Unlinked	Total
Train	Count	36,994	36,994	73,988
	%	50.00%	50.00%	100.00%
Test	Count	232	36,896	37,128
	%	0.62%	99.38%	100.00

3.7 입력변수 기술통계 분석

위 과정에서 구축한 학습 모델 데이터의 입력변수에 대한 분포를 살펴보기 위해 기술통계분석을 실시하였다.

〈Fig. 6〉의 box plot을 보면 위상적 정보 기반 변수는 전반적으로 0에 수렴하며 협력(linked)과 비협력(unlinked) 사이에 구별되는 차이가 없는 것을 확인할 수 있었다. 반면 데이터 기반 변수는 뚜렷하진 않지만 협력과 비협력 사례간의 차이가 존재하는 것으로 나타났다. 특히 비협력 대비 협력관계에서 ODA 사업 경험 차이(experience difference)와 네트워크 중심성 차이(degree centrality difference)가 더 작은 양상을 보였다.

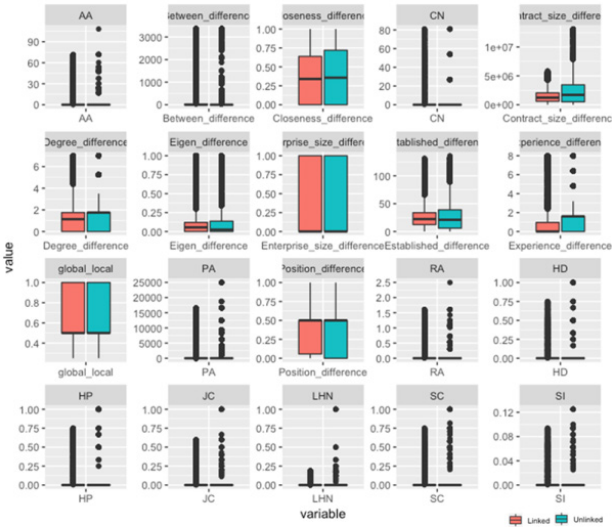

Fig. 6. Data Distribution

4. XGB 모델 훈련 및 평가 결과

4.1 XGB 모델 훈련(training)

〈Table 6〉의 training set와 〈Table 3, 4〉의 변수들 기반으로 XGB 모델을 훈련하였다.

XGB는 최적의 분류 모델을 구축하기 위해서 분석 데이터에 맞는 최적 파라미터(hyperparameter) 탐색이 필요하다. 본 연구에서는 최적 파라미터 탐색을 위해 R의 'mlr'패키지 기반 GridSearch 방법을 사용하였다. GridSearch는 자동으로 다수의 내부 모형을 생성한 후, 이들의 성능을 모두 테스트함으로써 최적의 파라미터를 찾아주는 방법으로, 이를 통해 본 연구에서 구축한 모델의 최적 파라미터는 〈Table 7〉과 같다.

Table 7. XGB Hyperparameter

Parameter	Definition	Hyperparameter Value
nrounds	Number of boosting round	100
eta	Learning rate	0.05
max_depth	Maximum number of boosting iterations	9
min_child_weight	Minimum sum of instance weight in a child	9.88
subsample	Subsample ratio of the training instance	0.786
colsample_bytree	Subsample ratio of columns when constructing each tree	0.955

4.2 XGB 모델 검증(test) 결과

앞서 훈련된 XGB 모델을 〈Table 5〉 테스트 데이터를 기반으로 검증하였다.

결과적으로 정확도(accuracy) 38.93%, 민감도(sensitivity) 70.26%의 결과가 도출되었다. 앞서 2.3절에 언급한대로 본 분석에서는 민감도를 높이는데 주목하였다. 즉, 실제링크(actual 'linked')를 올바로 예측(predicted 'linked')하는 것에 집중하였다. 본 모델의 민감도 70.26%는 232개 중에서 163개를 올바로 예측한 것을 의미한다.

Table 8. XGB model's Predictive Results

Observed	Predicted		Accuracy	Sensitivity
	Unlinked	Linked		
Unlinked	10,939	25,957	29.90%	70.26%
Linked	69	163		

구축한 XGB 모델로부터 낙찰기업의 협력관계 유의변수 및 변수 별 중요도를 도출하였다. 변수 별 중요도는 XGB에서 개별 변수가 가지치기에 얼마나 유용하게 작용하였는가를 나타내는 지표로 의사결정 나무 구축 시 속성이 많이 사용될수록 가중치가 증가하는 방식으로 산정된다. 이를 통해 최종적으로 전체 20개의 변수 중 11개가 낙찰기업의 협력관계 형성 시 유의한 것으로 나타났으며, 중요도는 1) ODA 경험 차이(experience Difference), 2) 일반 건설사업 경력 차이(established Year Difference), 3) 주 참여사업 규모 차이(contract size difference), 4) degree centrality difference 등의 순으로 나타났다〈Fig. 7〉.

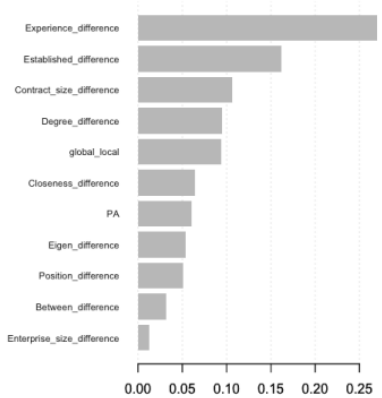

Fig. 7. XGB Feature Importance

5. 예측결과 세부 사례 및 분석

예측모델 구축 과정을 통해 협력관계에 유의미한 변수 별 중요도를 도출할 수 있었다. 하지만 XGB 모델의 특성 상 본 분석에서 사용된 1,323개의 분류기가 개별 예측과정에 관여하기 때문에, 변수 별 영향을 일반화시킬 수 없다. 즉, 하나의 의사결정나무로만 예측하는 CART의 경우 예측결과에 대한 영향을 주는 변수 별 분류 기준 값을 특정할 수 있으나, XGB의 경우 변수 별 영향이 개별 링크예측마다 매번 바뀌기 때문에 기준 값을 특정할 수 없다.

대신, XGB 모델은 예측결과에 대한 개별 변수의 비중을 로그 오즈(log-odds)로 아래 식(2)에 적용함으로써 개별 예측결과에 대한 해석을 할 수 있다.

$$probablity = \frac{1}{1+e^{-(\log - odds)}} \qquad (2)$$

이 과정은 R의 'xgboost Explainer'패키지를 통해서 구현 및 시각화할 수 있으며, 이는 앙상블 과정에서 모든 의사결정 나무에 대한 각 변수의 영향을 로그 오즈(log-odds) 값으로 나타내고, 이를 합산하여 최종 확률 값 도출과정을 전체 데이터별로 Waterfall Chart를 통해 보여준다.

Waterfall Chart는 영향력이 크게 작용한 변수 순으로 로그 오즈 증가는 파란색으로, 감소는 붉은색으로 나타내어 최종 예측 값에 이르는 과정을 보여주며 이를 통해 해당 변수가 특정 사례에서 협력 확률을 증가시켰는지 감소시켰는지 확인할 수 있다.

이를 통해 구축된 XGB 모델이 개별 링크의 협력 확률 도출을 위해 유의변수들을 어떻게 활용했는지 살펴 볼 수 있다. 아래에는 예측에 성공한 두 개의 협력사례를 그 예시로 제시하였다.

5.1. 예측 사례 분석

첫 번째 예측 성공 사례로서 2011년 6월에 발주한 감리사업[3]에서 베트남 Thang Long社와 영국의 WSP社 간의 실제 협력 사례를 들 수 있다.

〈Fig. 8〉에 제시된 바와 같이 XGB 모델이 98.86%(로그 오즈 값: 4.47)의 확률로 협력을 예측하였다. 본 Waterfall

3) World Bank Project ID: P070197 / Contract No.: 1309775

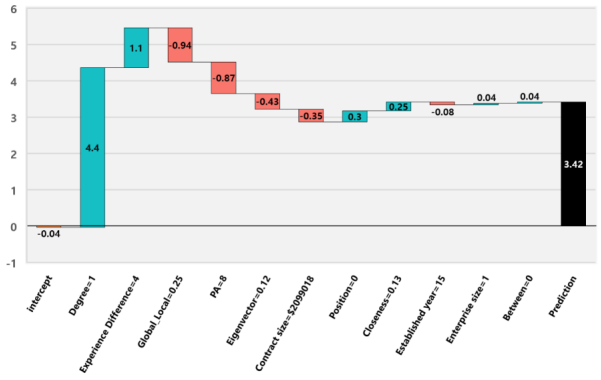

Fig. 8. Thang Long-WSP Cooperative Case

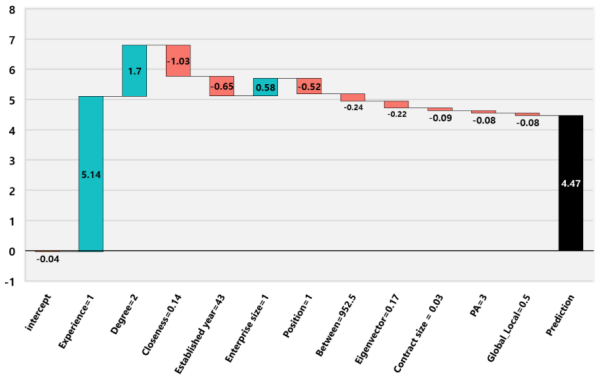

Fig. 9. Nippon Koei-Thai Eng. Consultant Cooperative case

Chart 그림을 통해 협력확률을 구할 때 XGB가 활용한 변수들의 영향 정도와 이들의 로그오즈 값을 살펴 볼 수 있다. 본 사례의 경우, 두 기업 간 ODA 사업 경험이 거의 유사하고(experience difference: 1회; 로그 오즈 값: 5.14), 네트워크 내의 위상 내지 중심성 차이가 적으며(degree centrality difference: 2; 로그 오즈 값: 1.7), 두 회사의 규모가 상이한 점(enterprise size=1, 로그 오즈 값; 0.58)들을 협력 확률을 높이는 요인으로 예측에 활용한 것을 볼 수 있다.

반면, 네크워크 내의 연결 근접성 차이가 작고(closeness centrality difference: 0.14; 로그 오즈 값: −1.03), 일반 경력 차이가 큰 점(established year difference: 43년, 로그 오즈 값: −0.65)은 협력 확률을 줄이는 주요 요인으로 활용되었다.

요약컨대, 이 경우 ODA 경험, 수주 실적 그리고 네트워크 내의 위상이 유사한 두 기업을 협력할 확률이 높은 것으로 예측한 것을 볼 수 있으며, 네트워크내의 근접성이 유사하면서 일반 경력 차이가 큰 것은 그 확률을 줄일 수 있는 것으로 해석할 수 있다.

〈Fig. 9〉은 두 번째 예측 성공 사례로 2011년 11월 설계 사업[4]에서 일본 Nippon Koei社와 태국의 Thai Engineering Consultants社 간의 협력사례이다. 그림에 제시된 바와 같이 XGB 모델이 예측한 협력확률은 96.83%(로그 오즈 값: 3.42)였다.

Waterfall Chart에 나타난 것처럼, 본 사례는 네트워크 중심성 차이가 적으며(degree centrality difference: 1, 로그 오즈 값: 4.4.), 두 기업 간 ODA 사업 경험 차이가 4회인 점(로그 오즈 값: 1.1)이 협력 확률을 높이는 요인으로 작용하였다.

이와 반대로, 글로벌 기업 간 협력인 점(global_local: 0.25, 로그 오즈 값: −0.94), 해당 링크의 선호적 연결[5] 지표(PA)가 낮은 점(PA: 8, 로그 오즈 값: −0.87)은 협력 확률을 낮추는 요인이었다.

본 경우는 첫 번째 사례와 동일하게 네트워크 내 위상이 유사한 것이 두 기업 간 협력 확률을 증가시키는 주요 요인이었으며, 글로벌 기업 간의 협력이면서 개별 기업의 네트워크 위상이 낮아 PA 값이 작은 것은 확률을 감소시키는 요인으로 작용하였다.

5.2. 예측 사례 결과 해석

앞선 두 사례를 비교해보면, 네트워크 내 위상(degree centrality)과 ODA 경험(experience)이 유사한 것은 협력 확률을 높이는 요인이며, 유사한 정도가 증가할수록 더 높은 확률을 가산한다는 것을 알 수 있다.

하지만, 네트워크 내 위상 차이와 선호적 연결 지표(PA)간의 관계를 보면, 위상이 비슷하더라도 해당 링크의 개별 노드에 연결된 다른 링크의 수가 적어 PA값이 낮은 경우 협력 확률이 감소하는 것을 볼 수 있다. 특히 네트워크 위상의 유사성이 높을수록 PA에 의한 확률 감소 정도가 더 크게 나타나는 것을 볼 때, 노드 간 위상이 유사하더라도 그들 자체의 위상이 낮으면 협력할 확률이 감소하며 이는 그들의 위상이 유사할수록 더 크게 작용하는 것을 알 수 있다.

또한 두 사례의 연결 근접성(closeness centrality) 차이가 유사함에도 불구하고 〈Fig. 8〉 사례에서는 확률을 증가시키는 요인으로, 〈Fig. 9〉에서는 감소하는 요인으로 작용하며 서로의 영향이 상이한 것을 볼 수 있다.

이들 두 사례 뿐 아니라 XGB 모델이 올바로 예측한 163건의 waterfall chart를 종합 분석해 본 결과 특정 변수가 같은 값이라도 다른 변수들의 값이 변함에 따라 그 영향이 변화하는 것을 알 수 있었다. 즉, XGB 모델을 통해 개별 사례에 따라 유의 변수에 대한 해석이 가능하지만 변수 별 영향을 일반화 시키는 것에는 어려움이 있다.

4) World Bank Project ID: P106235 / Contract No.: 1314118

5) 선호적 연결(Preferential Attachment, PA): 새로운 링크가 형성될 때, 네트워크 내부의 노드들은 이미 많은 링크를 갖고 있는 노드와 연결되는 것을 선호한다는 네트워크 이론. 링크를 구성하는 노드들이 기존에 다른 링크와 연결되어 있는 정도가 높아질수록 PA값이 증가함.

6. 결론

본 연구는 해외 ODA 사업에서 엔지니어링 기업들의 협력 여부를 네트워크 링크 예측이라는 새로운 기법으로 예측하는 방안을 정립하는 것으로 목적을 두었다.

구축된 예측모델은 민감도(sensitivity) 70.26%로 검증데이터의 실제 협력사례 232건 중 163건을 올바로 예측한 것으로 나타났다.

본 모델을 통해 개별 링크의 협력확률 도출 과정을 확인할 수 있는데, 이 과정에서 변수 별 영향이 단순히 범주화되는 것이 아닌 사례별로 상이하며 구체적인 요인 분석이 가능하다는 점을 알 수 있었다. 이를 통해 국내 기업이 협력 후보를 선정할 때 고려해야 할 구체적 기업 특성 및 실적을 파악할 수 있을 것이다.

더불어 본 예측모델은 생성되거나 소멸될 협력관계를 예측함으로써 국내기업이 향후 참여하고자 하는 특정 국가의 네트워크가 어떻게 변화될지 미리 예측해 볼 수 있다. 특히 네트워크 변화과정을 반영해 낙찰에 유리한 협력후보를 제시하는 만큼, 해당 시장에 진출하고자하는 국내 기업에게 시의 적절한 협력 후보를 맞춤형으로 추천해 주는데 활용될 것으로 기대된다.

그러나 본 연구에서는 분석 대상으로 세계은행 ODA 사업이 종료된 국가를 선정하여 다음 단계에 대한 실질적 분석은 실시하지 않았다. 이에 구축한 예측모델을 국내기업이 진출하고자 하는 신규시장에 적용함으로써 해당 시장의 협력 네트워크 변화를 예측하고, 이를 통해 해당 시장에 전략적 접근이 가능할 것으로 기대한다.

또한 실제 협력사례가 소수인 관계로 예측모델의 민감도를 향상시키는 과정에서 다수의 비협력 사례를 협력 사례로 예측함에 따라 정확도(ACC)가 하락하는 한계점이 존재하였다. 이에 향후 추가 유의변수를 도입하고, 다수의 세계은행 ODA 사업데이터로 부터 협력 사례를 추가로 수집하여 예측모델의 민감도와 정확도를 향상시킬 계획이다.

감사의 글

본 연구는 국토교통부 건설기술연구사업의 연구비지원(19SCIP-C079445-06)에 의해 수행되었습니다.

References

Altman, D.G., and Bland, J.M. (1994). "Diagnostic tests. 1: Sensitivity and specificity." *BMJ: British Medical Journal*, 308(6943), p. 1552.

Bastian M., Heymann S., and Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research*, 16, pp. 321-357.

Chen, T., and Guestrin, C. (2016). "Xgboost: A scalable tree boosting system." In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794.

Fessler, J.A., and Sutton, B.P. (2003). "Nonuniform fast Fourier transforms using min-max interpolation." *IEEE transactions on signal processing*, 51(2), pp. 560-574.

Frederickson, H.G., and LaPorte, T.R. (2002). Airport security, high reliability, and the problem of rationality. Public Administration Review, 62, pp. 33-43.

KENCA (2019). Engineering Insight. Korea Engineering & Consulting Association. pp. 1-23.

Kim, T.Y. (2017). Python Deep Learning Keras with Block. Digital Books, pp. 10-340.

Klinkman, M.S., Coyne, J.C., Gallo, S., and Schwenk, T.L. (1998). "False positives, false negatives, and the validity of the diagnosis of major depression in primary care." *Archives of Family Medicine*, 7(5), pp. 451-461.

KNA (2018). WB Bid Guidelines, Korea Energy Agency, pp. 3-25.

Koo, B.S., Shin, B.J., Yu, Y.S., and Jung, J.W. (2017). "Formulating International Entry Strategies for World Bank Consulting Projects Through Country-level Competitive Analysis: A Vietnam Case Study." *Korean Journal of Construction Engineering and Management*, KICEM, 18(4), pp. 57-66.

Lee, J.S., Lee, J.H., Han, S.H., and Kang, S.Y. (2018). "Partnering Strategy for Bidding Success in World

Bank's Vietnam Consulting Project." *Journal of The Korean Society of Civil Engineers*, 38(6), pp. 1021–1028.

Liu, L., Han, C., and Xu, W. (2015). "Evolutionary analysis of the collaboration networks within National Quality Award Projects of China." *International Journal of Project Management*, 33(3), pp. 599–609.

Mori, J., Kajikawa, Y., Kashima, H., and Sakata, I. (2012). "Machine learning approach for finding business partners and building reciprocal relationships." *Expert Systems with Applications*, 39(12), pp. 10402–10407.

Seo. H.B. (2017). A Deep Learning based Approach to Prediction of Technological Convergence, Seoul National Univ. of Science & Technology Master's Thesis, pp. 1–50.

Seo. H.B., and Lee. H.Y. (2018). "Predicting the Technological Convergence between Manufacturing and Service based on SVM-based Link Prediction." *Journal of the Korean Institute of Industrial Engineers*, 44(2), pp. 141–152.

Wang, P., Xu, B., Wu, Y., and Zhou, X. (2015). "Link prediction in social networks: the state-of-the-art." *Science China Information Sciences*, 58(1), pp. 1–38.

World Bank (2018). IBRD/IDA/IFC/MIGA Guidance Country Engagement. World Bank, pp. 3–46.

Yap, B.W., Rani, K.A., Rahman, H.A.A., Fong, S., Khairudin, Z., and Abdullah, N.N. (2014). "An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets." *In Proceedings of the first international conference on advanced data and information engineering* (DaEng-2013), pp. 13–22. Springer, Singapore.

Zheng, X., Le, Y., Chan, A.P., Hu, Y., and Li, Y. (2016). "Review of the application of social network analysis (SNA) in construction project management research." *International Journal of project management*, 34(7), pp. 1214–1225.

요약 : 국내 건설 엔지니어링 기업은 해외 실적 향상을 위한 방안으로 세계은행의 공적개발원조 사업을 통한 해외시장 확장의 발판을 마련하고자 한다. 하지만 세계은행 사업은 한정된 사업을 두고 다수의 글로벌 기업과 경쟁하기 때문에 입찰경쟁에서 우위를 선점하고, 수원국의 제도적 조건을 충족하기 위해 적합한 사업파트너와의 협력관계 구축이 필수적이다. 이러한 협력관계를 통한 입찰 전략 구축의 일환으로 사회 네트워크 분석을 이용한 다수의 과거 네트워크 분석 연구가 진행된 바 있으나, 네트워크의 변화과정을 기반으로 분석한 연구는 드물다. 이에 본 연구는 세계은행 ODA 사업이 원활히 시행된 후 종료된 아시아 3개국의 낙찰 데이터를 수집하고, 네트워크의 동적 변화를 반영한 학습기반 링크예측 모델을 구축하였다. 그 결과 낙찰기업들 간 협력관계 구축에 작용하는 11가지 주요 요인을 도출하고, 각 변수가 개별 링크의 협력 여부 확률 값에 미치는 영향을 확인하였다.

키워드 : 국제 협력 전략, 세계은행 ODA, 링크예측, XGBoost