ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2020 March: 26(1): 1~21 http://dx.doi.org/10.13088/jiis.2020.26.1.001

Word2Vec을 활용한 제품군별 시장규모 추정 방법에 관한 연구*

정예림
한국과학기술정보연구원
데이터분석본부
(*yelima@kisti.re.kr*)

김지희
한국과학기술정보연구원
데이터분석본부
(*kjh@kisti.re.kr*)

유형선
한국과학기술정보연구원
데이터분석본부
과학기술연합대학원대학교
과학기술경영정책학과
(*hsyoo@kisti.re.kr*)

인공지능 기술의 급속한 발전과 함께 빅데이터의 상당 부분을 차지하는 비정형 텍스트 데이터로부터 의미있는 정보를 추출하기 위한 다양한 연구들이 활발히 진행되고 있다. 비즈니스 인텔리전스 분야에서도 새로운 시장기회를 발굴하거나 기술사업화 주체의 합리적 의사결정을 돕기 위한 많은 연구들이 이뤄져 왔다. 본 연구에서는 기업의 성공적인 사업 추진을 위해 핵심적인 정보 중의 하나인 시장규모 정보를 도출함에 있어 기존에 제공되던 범위보다 세부적인 수준의 제품군별 시장규모 추정이 가능하고 자동화된 방법론을 제안하고자 한다. 이를 위해 신경망 기반의 시멘틱 단어 임베딩 모델인 Word2Vec 알고리즘을 적용하여 개별 기업의 생산제품에 대한 텍스트 데이터를 벡터 공간으로 임베딩하고, 제품명 간 코사인 거리(유사도)를 계산함으로써 특정한 제품명과 유사한 제품들을 추출한 뒤, 이들의 매출액 정보를 연산하여 자동으로 해당 제품군의 시장규모를 산출하는 알고리즘을 구현하였다. 실험 데이터로서 통계청의 경제총조사 마이크로데이터(약 34만 5천 건)를 이용하여 제품명 텍스트 데이터를 벡터화 하고, 한국표준산업분류 해설서의 산업분류 색인어를 기준으로 활용하여 코사인 거리 기반으로 유사한 제품명을 추출하였다. 이후 개별 기업의 제품 데이터에 연결된 매출액 정보를 기초로 추출된 제품들의 매출액을 합산함으로써 11,654개의 상세한 제품군별 시장규모를 추정하였다. 성능 검증을 위해 실제 집계된 통계청의 품목별 시장규모 수치와 비교한 결과 피어슨 상관계수가 0.513 수준으로 나타났다. 본 연구에서 제시한 모형은 의미 기반 임베딩 모델의 정확성 향상 및 제품군 추출 방식의 개선이 필요하나, 표본조사 또는 다수의 가정을 기반으로 하는 전통적인 시장규모 추정 방법의 한계를 뛰어넘어 텍스트 마이닝 및 기계학습 기법을 최초로 적용하여 시장규모 추정 방식을 지능화하였다는 점, 시장규모 산출범위를 사용 목적에 따라 쉽고 빠르게 조절할 수 있다는 점, 이를 통해 다양한 분야에서 수요가 높은 세부적인 제품군별 시장정보 도출이 가능하여 실무적인 활용성이 높다는 점에서 의의가 있다.

주제어 : Word2Vec, 기계학습, 텍스트 마이닝, 시장규모 추정, 시장분석

논문접수일 : 2019년 12월 9일 논문수정일 : 2020년 2월 12일 게재확정일 : 2020년 2월 28일
원고유형 : 일반논문(급행) 교신저자 : 유형선

* 본 연구는 한국과학기술정보연구원 주요사업의 일환으로 수행된 연구로서 사업보고서 내용의 일부를 수정 및 보완한 결과이며, 한국과학기술정보연구원(K-19-L03-C04-S01)과 한국연구재단(No.2019R1C1C1006358)의 지원을 받아 수행되었습니다.

1. 서론

폭발적인 속도로 생산 및 축적되고 있는 방대한 양의 데이터로부터 필요한 정보만을 효율적으로 검색하고 추출하여 활용하는 것은 현대 사회에서의 생존을 위한 필수적인 과정이다. 데이터의 유형은 크게 수치 데이터와 같이 일정한 규격과 형태를 가지며 연산이 가능한 정형 데이터(Structured data)와 텍스트 데이터나 음성, 사진, 영상과 같이 구조화 되어있지 않은 비정형 데이터(Unstructured data)로 나눌 수 있다. 이 중 비정형 데이터는 빅데이터의 상당 부분을 차지할 뿐 아니라 그 비중이 계속적으로 증가하고 있어 활용의 중요성이 더욱 강조되고 있다(Chakraborty and Krishna, 2014). 가장 대표적인 비정형 데이터의 하나인 텍스트 데이터(Text data)로부터 유용한 정보를 효과적으로 추출하기 위해 그동안 수많은 텍스트 마이닝(Text mining) 기법들이 개발되고 적용되어 왔다(Weiss et al., 2015). 실무적인 활용 측면에서 비즈니스 인텔리전스(Business Intelligence) 분야에서도 텍스트 마이닝 기법이 활발하게 적용되어 왔는데, 주로 시장 트렌드를 예측하거나 새로운 시장/기술 기회를 발굴하고 기술사업화 주체의 합리적인 의사결정을 돕기 위해 많은 연구들이 이루어져 왔다(An et al., 2016; Kim et al., 2018; Lee et al., 2018; Lee et al., 2014).

중소기업을 포함한 모든 사업 주체들의 합리적인 의사결정을 위한 가장 기초적이면서도 필수적인 정보 중의 하나는 시장규모와 시장 성장률, 시장점유율과 같은 시장정보라고 할 수 있다. 참여하고자 하는 산업 혹은 시장의 특징은 사업의 성공 여부를 결정지을 수 있는 핵심적인 요소 중의 하나이기 때문이다(Balachandra and Friar, 1997; Jung et al., 2019). 따라서 기업에서는 정확하고 신뢰성 있는 시장정보, 특히 현재 자신들이 판매하고 있는 제품이나 개발 예정인 신제품 등 특정한 제품에 대한 구체적이고 세부적인 시장규모 및 수요정보를 필요로 한다. 그러나 시장정보를 획득하고자 할 때 여러가지 어려움이 따른다. 먼저, 민간 시장조사 기관이나 시장보고서 발간 기업의 자료의 경우 매우 고가일 뿐만 아니라 정보가 제한적이다. 또한 공공기관 등에서 무료로 공개하는 공공 통계데이터의 경우 대부분 범위가 큰 산업 단위로 제공되거나 특정한 분류체계를 기준으로 폭넓은 카테고리별로 제공되는 경우가 많기 때문에 세부적인 제품 단위의 시장규모를 파악하기에는 적합하지 않은 경우가 많다. 시장정보 산출 시 분류체계나 통계기준을 활용하는 경우 도출된 정보의 정확성이 높고 체계적이라는 장점이 있지만, 사전(事前)에 정의되고 자주 갱신되지 않는 분류체계의 특성상 새롭게 개발된 신규 유형의 제품에 대해서는 적합한 품목이 존재하지 않거나 정확한 분류가 어려운 경우도 많다. 또한 동일한 품목 또는 제품군의 범위를 어느 수준까지로 볼 것인지에 대해서도 시장정보를 파악하고자 하는 주체마다 모두 기준이 다르기 때문에, 미리 부여된 기준이나 체계에 한정되지 않고 정보 수요자의 목적이나 요구에 따라 제품군의 수준과 범위를 보다 자유롭게 조절할 수 있는 시장규모 추정 방법이 요구된다.

한편 통계청에서는 다양한 통계조사를 통해 국가 전체 산업에 대한 구조와 분포, 경영실태 등에 관한 사항을 종합적으로 파악하고 경제 및 산업정책 수립 기초자료, 각종 통계의 모집단 자료 등으로 활용하기 위한 노력을 하고 있다. 그 중 핵심적인 예를 들자면, 전국의 모든 사업체를 대상으로 5년마다 경제총조사를 실시하고 있으

며(Statistics Korea, 2017), 매년마다 전국사업체 조사를 실시하여 사업체 단위 각종 통계조사의 표본틀을 제공하고 있다(경제총조사가 시행되는 연도에는 경제총조사로 통합 실시)(Statistics Korea, 2015). 위 조사들은 전수조사이기 때문에 조사 결과의 신뢰성과 정확성이 높으며, 정부의 공공데이터 개방 및 이용 활성화 정책에 따라 다양한 분야의 연구에서 기초 자료로서 활용되고 있다(Kang and Cho, 2018; Nam, 2017; Son et al., 2018).

본 연구에서는 통계청 경제총조사를 기초 데이터로 활용하여 상세한 제품군별 시장규모를 추정하는 방법론을 제안한다. 상세하게는 경제총조사 조사항목 중 사업의 종류(주사업)에 대한 마이크로데이터를 활용하여 시장규모 산출 범위와 기준 설정의 측면에서 세부적인 제품군에 대한 시장규모 추정이 가능하고, 제품군의 범주를 쉽게 조절할 수 있는 자동화된 모델을 구현하고자 한다. 사업의 종류 조사항목에 포함되어 있는 최종생산제품(명칭)에 대한 텍스트 데이터를 텍스트 마이닝을 통해 수치화 또는 벡터화함으로써 연산이 가능한 형태로 변환하여 분석하였다. 이를 위해 단어 임베딩(Word Embedding) 모델의 하나인 워드투벡터(Word2Vec) 알고리즘을 적용하여 제품명에 대한 텍스트 데이터를 벡터 공간에 맵핑하였다. 그리고 제품명간의 거리(유사도) 계산을 통해 특정한 제품명을 기준으로 제품군을 도출하고, 이들의 시장규모를 산출하는 모형을 구축하였다. 또한 일부 품목별 실제 출하규모 데이터와의 비교를 통해 본 방법론의 성능을 평가하였다.

2장에서는 시장규모 추정 방법 및 Word2Vec을 활용한 선행 연구를 소개하였으며, 3장에서는 실험 데이터와 연구 모델에 대한 내용을 설명하였다. 4장에서는 실험 결과와 이에 따른 고찰을, 5장에서는 연구의 시사점과 후속 연구방향을 제시하였다.

2. 선행연구

2.1 시장규모 추정에 관한 연구

시장규모를 추정하는 방법은 크게 필요한 정보를 신규로 수집 및 조사하는 방식인 1차 자료 활용 방법과 기존의 공개된 자료인 2차 자료를 활용하는 방법으로 구분할 수 있다(Yoo et al., 2015). 1차 자료를 활용하는 방법은 다시 해당 시장에 포함되는 사례를 전수 조사하는 방법과 표본 조사를 통해 모집단의 시장을 추정하는 방법으로 나눌 수 있다. 전수 조사 방법은 정확성과 포괄성이 높은 반면 조사 과정에 많은 비용과 시간이 소요되는 단점이 있으며(Shin, 2010), 주로 통계청 등 전수 조사를 수행할 여력을 갖춘 공공기관에서 조사를 수행하여 그 결과를 공표하는 경우가 많다. 표본 조사 방법은 샘플링한 기업의 매출액 정보나 설문 조사를 통한 예측을 바탕으로 전체 시장을 추정하는 방법으로(Jang et al., 2013), 전수 조사에 비해 조사가 수월하다는 장점이 있지만 정확성을 향상시키기 위한 세심한 표본 설계와 추정 방법을 필요로 한다. Chun et al.(2010)은 대표성이 높은 표본 조사대상 선정을 통해 국내 교과서 및 학습서의 시장규모를 추정하는 방법을 제시하였다. Lee et al.(2012)은 유사한 품목들을 그룹화하는 방식을 통한 표본조사로 국내 패션시장의 규모를 추정하는 모형을 제안하였고, Yoon et al.(2016)은 아직 정확히 알려지지 않은 시장인 지능정보산업

에 대해 서비스, 인프라, 플랫폼 분야로 시장을 구분하고 종사기업에 대한 표본조사를 통해 세분시장별 시장규모를 추정하는 방법론을 제안한 바 있다.

2차 자료 활용 방법의 경우 기존에 공개되어 있는 여러 1차 자료를 목적에 맞게 정제, 연계 및 가공하여 활용한다. 그러나 정보를 가공하는 과정에서 다수의 가정이 포함될 수밖에 없는 한계점이 존재한다(Shin, 2010). 2차 자료의 출처는 보통 공신력 있는 공공기관에서 발표하는 공공데이터나 대용량 데이터베이스 구축 기관의 데이터를 많이 활용한다. Choi et al.(2013)은 지자체 및 다수 기관의 데이터를 종합하여 국내 식용 천일염 시장규모를 추정 및 예측하였으며, Park et al.(2014)은 조달청 데이터(나라장터 쇼핑몰 판매실적)를 활용하여 G2B(Government to Business) 전자상거래 부문별 시장규모를 예측하는 방법을 제시하였다. Yoo et al.(2015)은 통계청과 관세청의 데이터를 합리적으로 연계하여 내수 시장규모를 보다 용이하게 추정할 수 있는 방법론을 제안한 바 있으며, Jun et al.(2017)은 최근 10년간 국내에서 창업한 기업의 매출실적 데이터를 이용하여 신규 사업의 매출액을 추정하는 지능형 매출추정 시스템을 제안한 바 있다.

일반적으로 시장규모를 추정하는 이유는 이를 기반으로 미래 시장규모(수요)를 예측하고자 하는 것이기 때문에, 시장규모 추정과 예측이 동시에 이뤄지는 경우가 많다. 수요예측 방법론을 포괄하여 좀 더 자세히 살펴보면, 시장규모 추정·예측방법은 크게 정성적 방법과 정량적 방법으로 구분할 수 있다(Lim and Oh, 1992). 정성적 방법은 아직 시장이 본격적으로 형성되지 않았거나 과거데이터가 부족하여 예측모형을 수리적으로 구현하기 어려운 경우 활용되며, 잠재고객에 대해 선호도 및 구매의사를 직접 조사하는 시장조사법(소비자조사법), 전문가의 의견을 수집·종합하여 추정하는 전문가 의견법(델파이법, 시나리오법), 유사제품 혹은 해외시장에서의 수요패턴을 반영하여 유추하는 자료 유추법(역사적 유추법) 등이 있다. 정량적 방법은 과거데이터를 근거로 시장규모를 추정하는 모형을 만들고 미래 수요를 예측하는 방법으로, 크게 시계열모형, 확산/성장곡선 모형, 인과모형 등으로 구분할 수 있다. 시계열모형은 과거 성장패턴이 미래에도 지속된다는 가정을 바탕으로 하며, 단순히 데이터를 기반으로 하는 추세분석법, 평균성장률법, 이동평균법, 지수평활법 등의 전통적 시계열모형과 확률적 시계열모형인 누적자기회귀이동평균(ARIMA) 모형 등이 널리 쓰이고 있다. 확산 및 성장곡선 모형은 신제품의 초기구매에 대한 누적수요량이 완만한 S자 곡선을 따른다는 경험에 근거하여, 과거에 존재하지 않은 신제품의 수요를 예측함에 있어 Bass 확산모형, Gompertz 모형, Logistic 모형, Probit 모형 등이 개발되어 사용되고 있다. 인과모형은 시장규모에 영향을 미치는 요인변수와의 인과관계를 규명하여 미래를 예측하는 방법으로, 회귀모형, 계량경제모형, 선도지표모형, 투입산출모형 등이 활용되고 있다.

이러한 연구의 흐름 속에서, 본 연구는 2차 자료를 활용하는 새로운 시장규모 추정 방법을 제시하고 있다. 이와 관련된 대부분의 선행연구는 사전에 정해져 있는 분류체계를 기준으로 집계된 통계데이터에 근거하여 시장규모를 추정하는 방식을 제안하였다. 반면, 본 연구에서는 텍스트마이닝을 통해 사전에 정해지지 않은 품목 혹은 제품군에 대한 시장규모를 추정하는 방법을 제안함으로써, 확장성과 활용성을 크게 개선하였다는 점이 가장 큰 차별점이다.

2.2 Word2Vec을 활용한 연구

Word2Vec 모델은 자연어를 벡터화하여 벡터 공간에 맵핑하는 단어 임베딩(Word Embedding) 알고리즘의 하나로, 2013년 Google사에서 제안한 인공신경망 기반의 언어 학습 모델(NNLM; Neural Net Language Model)이다(Mikolov et al., 2013). '유사한 의미를 가지는 단어는 유사한 문맥에서 등장한다.'는 언어학의 분산 가설(Distributional hypothesis)에 기반하여 데이터 학습을 통해 의미가 비슷한 단어들을 서로 가까운 벡터공간에 위치하도록 할당한다(Harris, 1954; Le and Mikolov, 2014). Word2Vec의 학습 방식은 CBOW (Continuous Bag Of Words)와 Skip-gram의 두 가지가 존재한다. CBOW는 주변 단어들을 입력받아 이로부터 중심(목표) 단어를 예측하는 방식이고, Skip-gram은 입력받은 단어를 중심으로 주위에 같이 등장할 확률이 높은 단어들을 예측하는 방식이다. 일반적으로 Skip-gram 학습 방식이 정확성 등의 측면에서 더 우수한 성능을 보이는 것으로 알려져 있다(Mikolov et al., 2013).

Word2Vec 알고리즘은 벡터 공간을 활용한 분산 표현을 통해 단어의 의미와 단어 간 관계를 효율적으로 추정하는 방법으로서 자연어 처리 분야에서 비약적인 정밀도 향상을 가능하게 하며 다양한 분야의 연구에서 활용되어 왔다(Kim and Lee, 2015). 대표적으로 의미있는 정보나 시사점을 보다 효율적으로 검색하고 추출하기 위해 활용되거나(Heu, 2018; Ngo et al., 2016; Park et al., 2017), 특정 주제에 따라 특허, 신문기사, SNS, 상품 리뷰 등 다양한 종류의 문서를 분류하는 데 활용되었다(Kim and Koo, 2017; Kim and Park, 2019; Lilleberg et al., 2015; Stein et al., 2019; Yang et al., 2019). 또한 사용자의 특성과 구매 실적, 기호 등에 따라 구매가 예상되는 상품 및 서비스를 추천해주는 상품 추천 분야(Grbovic et al., 2015; Kang, 2019; Vasile et al., 2016), 긍/부정 태도와 사용자 의견 등을 분석하는 감성 분석 및 오피니언 마이닝(Heo and Ohn, 2017; Lee et al., 2017; Liu, 2017; Park and Lee, 2018; Xue et al., 2014) 등에도 활용되었다.

다른 한 방향으로 Word2Vec 모델은 영어 단어의 임베딩에 최적화되어있기 때문에 한국어 데이터를 적용할 때 부딪히게 되는 여러 가지 문제점들을 해결하고, 한국어에 적합하도록 모델을 개선하고 성능을 향상시키기 위한 연구들도 활발하게 진행되고 있는 추세이다. 교착어에 속하는 한국어의 특성을 반영하여 파라미터를 튜닝 및 최적화하고, 정확한 성능 평가를 위해 한국어로 구성된 평가 데이터셋을 구축하는 등 효율적이고 효과적인 한국어 기반의 모델을 구현하기 위한 다양한 연구들이 이루어지고 있다(Choi et al., 2016; Kang and Yang, 2019; Kang and Yang, 2018; Yang et al., 2015).

본 연구는 Word2Vec 알고리즘을 시장규모 추정에 최초로 적용하고 그 활용성을 타진해 보았다는 점에 의의가 있다.

3. 연구방법

3.1 실험 데이터

본 연구에 활용한 데이터의 종류는 크게 3가지로 다음과 같으며, 각 데이터를 활용한 전체 연구 프로세스에 대해서는 3.2절에서 설명한다.

국내 기업들이 생산하는 제품 및 이들의 매출액 정보에 대한 원시데이터로서 통계청의 2015

경제총조사 중 제조업(C) 부문의 마이크로데이터를 활용하였다(RDC19021001). 조사항목 중 사업체의 주사업 종류가 존재하는 종사자수 9인 이하 사업체에 대한 데이터(345,103건)를 추출하여 기본 데이터셋으로 활용하였다. 연구에 사용된 항목(필드)은 사업체고유번호, 매출액, 사업의 종류(주사업)이며, 사업의 종류(주사업)의 세부 항목에는 최종생산제품명칭, 매출액 비중(%), 산업분류부호(한국표준산업분류(KSIC; Korean Standard Industrial Classification) 세세분류코드)가 포함되어 있다. Table 1은 데이터셋에 대한 기초통계량 분석 결과를 나타낸다. 사업의 종류는 주사업과 부사업으로 구분되는데, 개별 기업의 사업 중 주사업이 차지하는 비중의 평균은 99.1%, 주사업 비중이 100%인 기업은 전체의 97%를 차지하였다. 즉, 해당 기업들의 사업 중 주사업의 비중이 매우 높은 것으로 나타나 본 연구에서는 주사업만을 대상으로 분석을 실시하였다. 분석 대상 기업들의 주사업 매출액 총합은 153조 원, 주사업 매출액 평균은 444백만 원인 것으로 파악되었다. 총 345,103개 사업체 데이터 중 결측치를 제외한 최종생산제품명칭 텍스트 데이터는 337,581건으로 나타났다.

〈Table 1〉 The summary of raw data (Economic Census Korea 2015 (9 or less workers)) (unit: number, million KRW)

KSIC section	C (Manufacturing)
No. of establishments	345,103
Total sales	154,721,910
Major business portion (%)*	99.1
Total sales of major business	153,141,818
Mean sales of major business	444
Max. of major business sales	247,170
Min. of major business sales	0
No. of text (product name) data	337,581

* The ratio of establishments that major business portion is 100%: 97.0%

본 연구에서 제안하는 방법을 이용하면, 어떠한 용어의 분석 대상에 대해서도 시장규모를 산출할 수 있다. 그러나 결과를 보다 명확하게 설명하고 제품군 도출 시 기준으로 활용하기 위해 통계청에서 제공하는 KSIC 해설서의 색인어를 기준 제품명으로 활용하였다. 통계청에서는 KSIC 각 산업분류코드별로 분류명과 분류내용(정의, 예시 및 제외)을 설명하는 해설서를 제공하고 있으며, 세세분류 단계에서는 색인어를 함께 제공한다. 색인어의 경우 해당 산업에 포함되는 비교적 상세한 수준의 품목 유형에 대한 정보를 제공하기 때문에 이들을 제품 종류를 구분하는 기준으로 활용할 수 있다. 본 연구에서는 KSIC 9차 기준 제조업 분야 461개 세세분류에 대한 11,654개 색인어를 제품군을 도출하는 기준 제품명으로 활용하였다(Table 2).

또한 본 방법론의 성능을 검증하기 위해서는 연구 모형을 통해 추정된 시장규모를 실제 집계된 수치와 비교해 볼 필요가 있다. 이를 위해 통계청의 광업·제조업조사 데이터와의 비교를 수행하였다. 광업·제조업조사 품목편에서는 약 2,100여개 품목별로 사업체수, 생산액, 출하액, 완제품 연말재고액 등의 정보를 매년 조사하여 발표하고 있다. 일반적으로 10인 이상 사업체(약 75,000개)에 대해서만 조사하여 발표하기 때문에 본 연구에서 활용한 종사자수 9인 이하 사업체에 대해서는 공표된 데이터가 없지만, 2015년의 경우 경제총조사로 통합 실시되었기 때문에 9인 이하 사업체에 대해서도 품목별로 집계된 데이터가 존재한다.

〈Table 2〉 Part of index words by KSIC sub-class

KSIC Sub-class	Name	Index words*
C17110	펄프 제조업	가성소다펄프, 고지재생펄프, 기계가공펄프, 기계목재펄프, 기계펄프, 대나무펄프, 목재펄프, 미표백펄프, 반표백펄프, 섬유질셀룰로오스재료펄프, 섬유질셀룰로오스펄프, 섬유펄프, 소다펄프, 쇄목펄프(기계펄프), 아황산펄프, 재생펄프, 표백펄프, 화학목재펄프, 화학펄프, 황산펄프
C28410	전구 및 램프 제조업	LED(발광다이오드)램프, UV(ultravioletray, 자외선)램프, 가스충전이중(dual, 가열및방전)램프, 가스램프(가스방전등), 가스주입램프, 검사기기용램프, 경보등(경보장치용), 고전압용수은등, 광고용램프, 글로(glow)방전램프, 나트륨증기램프, 나트륨등(전구및조명장치), 나트륨램프, 네온가스방전램프(조명용), 네온램프, 네온사인용등(네온등), 네온전구, 디스플레이후면광원용램프, 램프(전기용), 메탈할라이드램프, 메탈할라이드(metalhalide)램프, 방전등, 방전램프, 백열등, 백열램프, 백열전구(램프), 보안등, 빔램프(beamlamp), 사진용섬광전구,...
C33402	영상게임기 제조업	DDR(dancedancerevolution)(오락기), 게임용구(영상수상기형), 게임용품(비디오수상기형), 게임장용게임기(비디오수상기형), 고정소프트웨어내장전자게임기, 문방구용영상게임기, 비디오게임기, 비디오게임용구(텔레비전수상기연결형), 비디오게임용품, 비디오게임장비, 아케이드게임기, 영상게임용구, 영상게임기, 오락실용게임기, 전자게임기, 전자오락게임장비, 전자오락실용게임기, 텔레비전연결게임기, 휴대용오락게임기

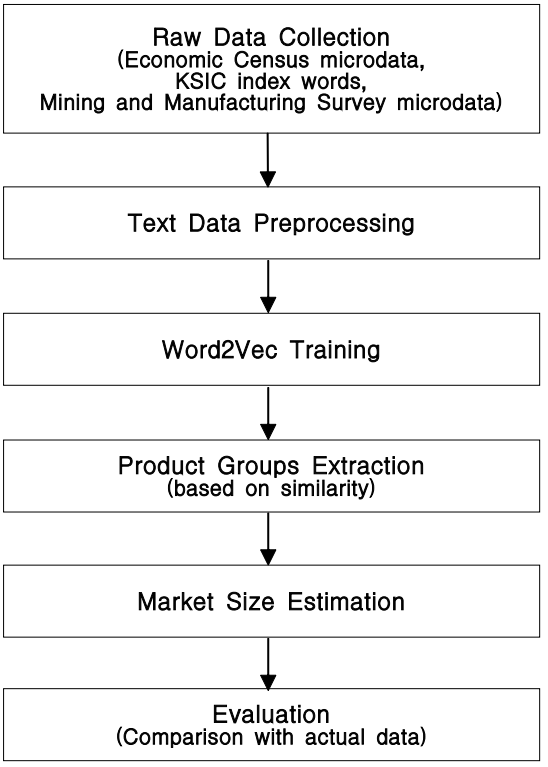
* Index words were refined by text data preprocessing criteria applied in this study.

따라서 통계청 마이크로데이터로부터 9인 이하 사업체의 2015년 품목별 출하액 데이터를 확보(RDC19021001)하여 각 품목에 대한 실제 시장규모 수치와 본 방법론을 통해 추정된 시장규모와의 비교 평가를 수행하였다.

3.2 연구 모형

본 연구 방법론의 전체적인 모형을 Figure 1에 나타내었다. 먼저 실험에 사용할 제품정보 관련 데이터들을 수집한 뒤 이들을 정제, 가공 및 연계하는 데이터 전처리를 수행한다. 다음으로 전처리한 데이터를 Word2Vec 알고리즘을 적용하여 학습시킴으로써 벡터 공간에 임베딩하고, 벡터 공간에서의 제품명간 거리(유사도)를 기준으로 특정 제품명과 유사한 제품명들을 추출하여 제품군을 도출한다. 이어서 추출된 각각의 제품명에 연결되어 있는 매출액 정보를 호출 및 연산하여 해당 제품군의 매출액 규모를 추정한다. 마지막으로 특정 품목을 기준으로 실제 집계된 수치와 비교함으로써 본 방법론의 성능을 평가한다. 전체 연구 모델은 오픈소스 기반의 언어 프로그램인 R version 3.5.3을 이용하여 구현하였다.

각 단계를 구체적으로 살펴보면, 데이터 수집 단계에서는 앞에서 설명한 바와 같이 통계청 경제총조사 마이크로데이터, KSIC 해설서 색인어, 통계청 광업·제조업조사 품목편 마이크로데이터를 수집하여 활용하였다. 경제총조사 마이크로데이터의 경우 추후 매출액 추정 단계에서 다시 호출하여 제품군별 매출규모 연산에 사용되었다. 데이터 전처리 단계에서는 각 데이터셋으로부터 제품명에 대한 텍스트 데이터를 모두 추출하여 정제 작업을 수행하고, Word2Vec 학습을



〈Figure 1〉 Research model

위해 KSIC 세세분류에 따라 구분하여 하나의 문서로 만들었다. 즉, 동일한 KSIC 세세분류에 해당하는 제품명끼리 하나의 문서에 포함되도록 학습용 데이터셋을 구성하였다.

Word2Vec 학습 단계에서는 구성된 데이터셋을 Skip-gram 방식을 사용하여 학습시켰다. 벡터 차원의 경우 일반적으로 한국어 임베딩에 적합하다고 알려져 있는 300 차원을 적용하였으며(Choi et al., 2016), 학습 윈도우의 크기(주변의 몇 개 단어를 학습할 것인지)는 최적화를 수행하였다. 제품군 도출 단계에서는 KSIC 세세분류별 색인어를 기준으로 가까운 벡터 공간에 위치하는 제품명칭들을 추출하는 알고리즘을 구현하였다. 벡터 공간 내 거리는 코사인 거리(Cosine distance), 유클리드 거리(Euclidean distance), 맨하탄 거리(Manhattan distance), 마할라노비스 거리(Mahalanobis distance) 등을 통해 구할 수 있는데, 본 연구에서는 다음과 같이 코사인 거리(코사인 유사도)를 기준으로 근접해있는 제품명들을 추출하였다. 또한 기준이 되는 코사인 유사도 수치를 조절함으로써 추출되는 제품군의 범위를 조절하였다.

$$similarity = \cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

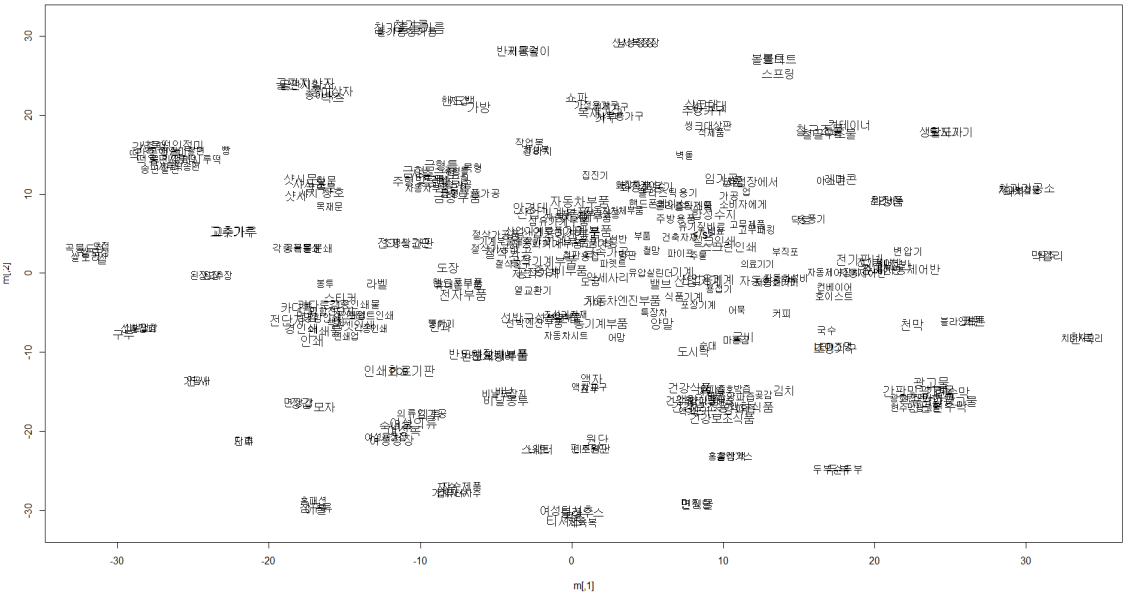
매출액 추정 단계에서는 앞의 제품군 도출 단계에서 결정된 제품군의 범주에 따라 시장규모(매출규모)를 자동으로 산출하였다. 즉, KSIC 색인어를 기준으로 추출된 각 제품명칭에 대해 앞에서 구축한 기본 데이터셋으로부터 개별 기업의 제품별 매출액 데이터를 호출하여 합산하였다. 따라서 제조업 분야 11,654개 색인어를 기준으로 상세한 제품 수준의 시장규모를 추정할 수 있었다. 그러나 색인어를 기준으로 한 시장규모는 현재 공표된 수치가 없으므로 추정된 시장규모가 얼마나 정확한 것인지 파악하기 어렵다. 따라서 마지막 평가 단계에서는 기 조사 및 집계된 실제 데이터인 2015년 광업·제조업조사의 2,077개 품목별 매출액 수치와 본 모형을 통해 동일한 품목명을 기준으로 산출된 시장규모가 얼마나 유사한 값을 가지는 지를 피어슨 상관계수 등을 통해 비교하였다.

4. 연구결과

4.1 실험 결과

텍스트 데이터를 Word2Vec 학습에 사용하기에 앞서 전처리 작업을 수행하였다. 문장부호(쉼표(,), 마침표(.), 가운뎃점(·) 등) 및 공백을 제거하였고(전체 데이터의 20.3%에 포함), 일반명사이면서 생산제품명칭에 다수 포함되어있는 “제조”, “제작”, “생산” 등의 단어를 제거하였다(전체 데이터의 29.5%에 포함). 전처리 수행 후 Word2Vec 학습을 위해 텍스트 데이터의 형태를 재구조화 하였다. 즉, 행렬 형태의 데이터로부터 KSIC 세세분류에 따라 동일한 KSIC 분류별로 제품명들이 하나의 문서에 포함되도록 형태를 변경하였다. 총 35만여 개 제품명 텍스트 데이터(경제총조사 마이크로데이터(337,581개), KSIC 색인어(11,654개), 광업·제조업조사 품목편의 품목명(2,077개))를 제조업 세세분류에 따라 분류하여 총 461개의 문서로 구조화하였다.

다음으로 R의 WordVectors 라이브러리를 이용하여 Word2Vec 학습을 수행하였다. 정확한 시장규모를 산출하기 위해 모든 제품명이 학습되어야 하는 본 방법론의 특성상, 단어의 최소출현빈도(min_count)는 1로 설정하였다. 학습이 끝난 후에는 t-SNE (t-distributed Stochastic Neighbor Embedding) 알고리즘을 적용하여 벡터 공간에 맵핑된 제품명을 2차원으로 축소시켜 학습된 결과를 시각적으로 확인하였다(Figure 2). t-SNE는 비선형 차원 축소 기법 중의 하나로, 고차원 데이터를 2차원 또는 3차원의 저차원으로 축소하여 가시화 하는데 유용하게 사용된다(Maaten and Hinton, 2008). Figure 2에 나타난 바와 같이 대체적으로 유사한 제품명칭끼리 가까운 공간에 위치하는 것을 직관적으로 확인할 수 있다. 완전



〈Figure 2〉 Visualization of Word2Vec training result using t-SNE

히 동일한 단어가 중복적으로 표시된 경우는 실제 개별 기업들의 원시데이터에 같은 용어로 기재된 경우가 다수 존재하였기 때문이다.

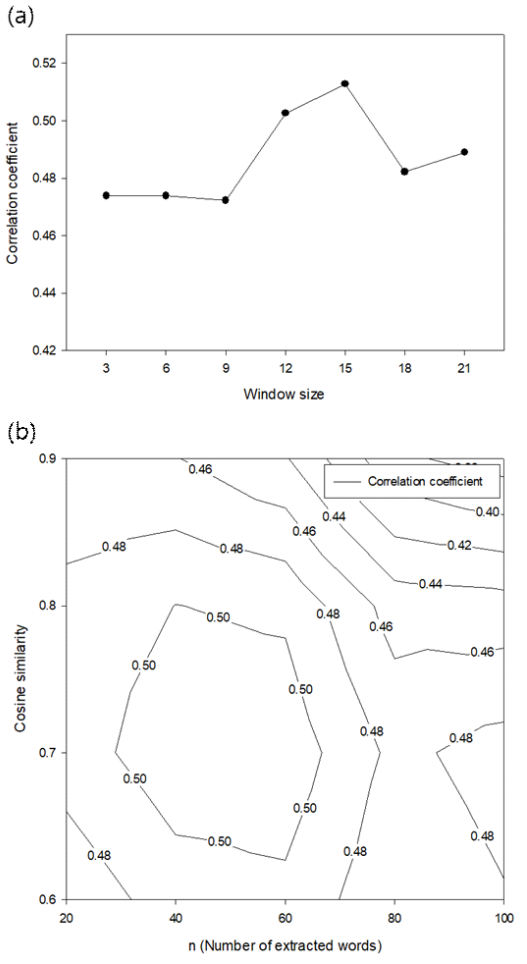
벡터화된 제품명을 토대로 KSIC 색인어를 기준으로 각각의 색인어와 유사한 제품명들을 추출하여 하나의 제품군으로 그룹화 하였다. 코사인 거리를 기준으로 거리가 가까운 순으로 제품명들을 추출하는 방식을 적용하였다. Table 3은 KSIC 색인어 기준 추출된 제품명들의 일부를 나타낸다. Cosine similarity는 1-코사인거리로 계산된 값을 의미한다. 추출 기준으로 적용할 유사도 수치를 설정하면 이에 따라 기준을 만족하는 제품명들만 추출되게 된다.

그러나 KSIC 색인어를 기준으로 유사한 제품명들을 추출하고 매출규모를 산출할 경우 모든 색인어에 대해 알려진 시장규모 수치가 존재하지 않기 때문에 추정된 값이 얼마나 정확한지 확인하기 어렵다. 따라서 본 방법론의 성능을 검증하기 위해 광업·제조업조사 품목편 데이터와 본 방법론을 통해 도출된 결과값을 비교하였다.

비교에 앞서 실험 파라미터의 최적화를 수행하였다. 먼저 학습 윈도우의 크기를 조절하면서 윈도우 크기에 따라 얼마나 성능 차이가 있는지를 살펴보았다. 성능은 품목편 데이터와 추정치 간의 Pearson 상관계수를 산출하여 비교하였다. Figure 3(a)에서 볼 수 있듯이 학습 윈도우의 크기가 15인 경우에 가장 좋은 성능을 보이는 것으로 나타났다. 다음으로 코사인 유사도와 추출 단어의 개수 변화에 따른 차이를 살펴보았다. 제품명 추출을 위해 WordVectors의 nearest_to 함수를 사용하였는데, 코사인 유사도를 기준으로 특정 단어와 유사한 단어들을 추출해 준다. 따라서 nearest_to 함수를 이용하여 단어들을 추출하고 결과를 비교해보았다. Figure. 3(b)에 나타난 바

〈Table 3〉 Part of extracted product names based on KSIC index words

KSIC Index	The extracted product names	Cosine Similarity
그림액자표구제품	그림액자표구제품	1.00000
그림액자표구제품	족자	0.97558
그림액자표구제품	표구사	0.97350
그림액자표구제품	표구처리	0.97186
그림액자표구제품	병풍	0.96755
그림액자표구제품	동양화서양화	0.96408
그림액자표구제품	액자주문	0.96396
그림액자표구제품	액자표구등	0.96334
그림액자표구제품	액자틀표구	0.96321
그림액자표구제품	병풍및표구	0.96279
그림액자표구제품	액자표구소매	0.96250
그림액자표구제품	POP액자아크릴상자스카시등	0.96225
그림액자표구제품	문구류전사인쇄	0.96144
그림액자표구제품	액자표구현황판	0.96057
그림액자표구제품	서예표구	0.96040
그림액자표구제품	전사	0.96034
그림액자표구제품	케이스마케팅용품(학용품)	0.96014
그림액자표구제품	플라스틱표면인쇄	0.95989
그림액자표구제품	서예및그림표구	0.95988
그림액자표구제품	족자무역업	0.95970
커튼	커튼	1.00000
커튼	커텐	0.96158
커튼	커튼블라인드	0.94331
커튼	커텐블라인드	0.94096
커튼	커텐브라인드	0.93779
커튼	커텐주문	0.92920
커튼	브라인드커텐	0.91817
커튼	커튼및유사제품	0.91452
커튼	커텐홈패션	0.91217
커튼	실내용커텐	0.91128
커튼	각종커텐	0.90771
커튼	가정용커텐	0.90747
커튼	커텐등	0.90715
커튼	블라인드	0.90585
커튼	커튼홈패션	0.90412
커튼	롤스크린커튼	0.90401
커튼	롤스크린커텐	0.90307
커튼	커텐제공	0.90298
커튼	커튼및블라인드	0.90246
커튼	로만쉐이드블라인드	0.90200



〈Figure 3〉 Optimization of parameters. (a) window size, (b) number of extracted words and cosine similarity.

와 같이 추출 단어개수 60개 및 코사인 유사도 0.7의 조합에서 가장 좋은 성능을 나타냈다.

최적화된 조건에서 품목편 데이터와의 성능 비교 결과를 살펴보면, 피어슨 상관계수(r)는 0.513, 예측값/실측값 비율(Ratio)의 중앙값은 1.24로 나타났다(Table 4). 이는 품목편 데이터와 본 연구의 예측값이 어느 정도 일치성을 보인다는 것을 의미한다. 그러나 Ratio가 1보다 높거나 낮은, 즉 과대 추정되거나 과소 추정된 품목들도 다수 발견되었다. 특히 Ratio가 10 이상인 경우가 전체의 9.15%를 차지하여 과대 추정된 경우가 다소 많은 것으로 나타났다.

〈Table 4〉 Comparison with actual market size (Ratio=predicted value/actual value)

Indicator	Values
Correlation coefficient (r)	0.513
Median of Ratio	1.24
Percentage of Ratio>10	9.15%

Table 5는 본 연구모형을 활용하여 시장규모를 추정한 결과의 예시를 나타낸다. 최적화된 조건에서 “산업용저울”을 키워드로 제품명을 추출한 결과, “산업용저울”, “전자저울”, “일반저울” 등 코사인유사도 순으로 연관성이 높은 생산제품명들이 추출되었다(Table 5(a)). 또한 이들의 매출액을 연산한 결과로서 시장규모 추정액은 49,743백만 원으로 산출되었다. 실제 출하규모(30,952백만 원)와 비교해보면 다소 과대 추정된 수치임을 알 수 있다. 마찬가지로 “이불”을 키워드로 추출한 결과, 시장규모 추정액이 264,531백만 원으로 실제 출하규모인 202,863백만 원보다 과대 추정되었다(Table 5(b)). Table 5(c)와 5(d)는 실제 출하액보다 과소 추정된 결과를 보여주고 있다. “승강기류부품” 또는 “산업용액체펌프”를 기준제품명으로 추출한 결과, 이들의 시장규모 추정액은 각각 실제 출하규모의 0.45배, 0.57배로 과소 추정된 결과를 나타냈다.

〈Table 5〉 Examples of market size estimation using the proposed model

(a)

Extraction Keyword	The extracted product names	Cosine similarity	Estimated Market size
산업용 저울	산업용저울	1.00000	49,743 (million KRW)
	전자저울	0.96432	
	일반저울	0.92642	
	중량측정기	0.91971	
	산업기계저울	0.91551	
	산업용저울계량기부품	0.90939	
	아르곤용접	0.90851	
	키중계측정기	0.90523	
	자동차계근용저울	0.90441	
	도량형기(일반저울)	0.90434	
	산업용계량식저울	0.90200	
	팩킹가스켓보운제	0.90162	
	일반저울,피스톤	0.90118	
	저울유포	0.89863	
	감광저울을제외한가정상업 산업용의각종저울	0.89720	
	전자저울및계측기	0.89649	
	게근대	0.89251	
	저울	0.89181	
	저울(일반기계식)	0.89091	
	가정상업용저울	0.89071	
	스푼저울	0.88971	
	저울계량기	0.88959	
	가정용저울	0.88921	
	전자저울정밀기기	0.88861	
	계량기저울(차량용계근대)	0.88819	
	…	…	

(b)

Extraction Keyword	The extracted product names	Cosine similarity	Estimated Market size
이불	이불	1.00000	264,531 (million KRW)
	침구	0.93485	
	이불베개	0.93284	
	침구류	0.93003	
	이불베게	0.91613	
	이불침구류	0.91443	
	누비이불	0.91269	
	각종침구류	0.90387	
	이불요	0.90169	
	이불방석	0.89661	
	이불베게등	0.89553	
	홈패션	0.89071	
	방석	0.88891	
	베개이불	0.88506	
	베개	0.88472	
	베개방석	0.87993	
	담요	0.87819	
	솜이불	0.87487	
	이불배게	0.87445	
	침구및수예품	0.87321	
	쿠션	0.86850	
	베게이불	0.86846	
	침구류및관련제품	0.86770	
	베개피	0.86700	
	각종이불	0.86687	
	방석홈패션	0.86575	
	…	…	

(c)

Extraction Keyword	The extracted product names	Cosine similarity	Estimated Market size
승강기류 부품	승강기류부품	1.00000	46,159 (million KRW)
	승강기상승과속방지장치	0.98259	
	엘리베이터필터	0.97935	
	승강기부품(엘리베이터문벽등)	0.97651	
	벨트케이블(엘리베이터부품)	0.97623	
	고압차단기엘리베이터하우징	0.97540	
	승강기부분품	0.97373	
	기계식주차기설비	0.97317	
	하역기계연속이동식	0.97264	
	엘리베이터부품조립	0.97249	
	하역운반설비연속이동식	0.97246	
	승강기및유사장치	0.97214	
	승강기본체조립(엘리베이터)	0.97192	
	케이블작동식견인기	0.97191	
	컨베이어수평이동장치	0.97169	
	체인형컨베이어	0.97159	
	자동화기기리프트	0.97124	
	조립금속제품오이스트출입 리프트	0.97084	
	승강기본체	0.97070	
	엘러베이트추	0.97046	
	엘리베이터제어판	0.97010	
	…	…	

(d)

Extraction Keyword	The extracted product names	Cosine similarity	Estimated Market size
산업용 액체펌프	산업용액체펌프	1.00000	91,121 (million KRW)
	공업용액체펌프	0.92403	
	펌프모타플랜트기계기계부품	0.91684	
	수처리기계액체펌프업	0.91172	
	산업용특수펌프	0.90993	
	양수기발전기	0.90884	
	에바라펌프펌프모터	0.90623	
	윤활펌프	0.90619	
	양수기펌프	0.90547	
	모터펌프	0.90035	
	냉난방히트펌프	0.89845	
	요소수주입기	0.89643	
	내산펌프	0.89617	
	액체펌프업	0.89551	
	펌프(액체용)	0.89506	
	공업용펌프	0.89425	
	액체용펌프가공	0.89242	
	산업용수중펌프	0.88944	
	펌프선풍기스토브	0.88941	
	자동그리스주유펌프장치	0.88892	
	온도자동콘트롤수중펌프	0.88762	
	펌프	0.88704	
	수중펌프	0.88678	
	액체모타펌프도매	0.88620	
	유체이용펌프(정량펌프)	0.88576	
	…	…	

4.2. 토의

중소기업을 비롯한 정보수요자들은 자신이 관심을 가지고 있는 가급적 세부적인 품목에 대한 시장정보를 필요로 한다. 국내외 시장 진입 시 가장 일차적으로 동일 또는 유사한 제품 범주 내에서 다른 기업의 제품들과 경쟁하기 때문이다.

본 연구에서 제안한 시장규모 추정모형을 이용하면, 분류체계 등에 의해 사전에 정의되지 않은 세부적인 수준의 품목이나 제품군에 대해서도 시장규모를 추정할 수 있다. 예를 들어, 본 연구에서는 11,654개 KSIC 색인어와 2,100여개 품목 각각에 대해 시장규모를 추정하여 활용할 수 있음을 보여주었고, 그 결과가 통계적으로 집계된 데이터와 어느 정도 일치함을 확인하였다. 그러나 본 연구에서 활용한 데이터는 2015년도 기준 데이터로서, 추정된 시장규모 정보가 현재 시점과는 차이가 존재한다. 따라서 실제 정보수요자 제공을 위한 제품군별 시장규모를 추정할 경우에는 가장 최신 시점의 데이터를 수집하여 모형에 적용함으로써 산출 정보의 시의성을 확보할 수 있다.

본 연구 모형의 성능을 평가한 결과 정확성 향상이 보다 필요한 것으로 나타났는데, 이는 다음과 같은 몇 가지 한계점에서 기인한 것일 수 있으며, 이들을 개선 및 보완함으로써 성능을 더욱 향상시킬 수 있을 것으로 기대된다.

첫째, Word2Vec 학습용 데이터셋 구성 시 단어의 배열(순서)에 주의를 기울일 필요가 있다. 실험에 사용된 데이터는 제품명칭으로만 이루어져 있기 때문에 일반적으로 문맥을 갖는 문장으로 구성된 문서와는 차이가 있다. 현재 모형에서는 데이터의 수집 및 활용 가능한 항목(필드)의 제한으로 인해 적절한 순서 부여에 어려움이 있었으나, 추가적으로 제품명 배열에 순서를 부여할 수 있는 기준을 적용한다면 정확도/신뢰도를 포함한 모델의 성능을 더욱 향상시킬 수 있을 것으로 예상된다. 또한 기존의 연구에서 활용된 바와 같이 Word2Vec 알고리즘과 더불어 단어간 형태적 유사도를 계산하는 음절 단위의 자카드(Jaccard) 유사도를 추가적으로 반영(Lee and Kim, 2018)함으로써 유사한 제품명 추출의 정확성을 더욱 높일 수 있을 것으로 기대된다.

둘째, 본 모형 평가 결과의 정확성을 향상시키지 못한 또 다른 이유로 출하액과 매출액의 차이에서 발생하는 오류를 꼽을 수 있다. 성능 평가 시 광업·제조업조사 품목별 출하액과 본 모형을 통해 추정한 매출액을 비교하였다. 그러나 출하액은 전체 매출액에서 구입상품의 매출액을 제외하고 공장 내부간 거래금액을 가산/감산한 수치이기 때문에 일정 수준의 차이가 존재한다. 따라서 이러한 비교 데이터의 차이로 인해 일부 결과 차이가 발생했을 수 있다.

셋째, 광업·제조업조사의 품목별 출하액과의 비교 결과 과대 추정된 경우가 많았는데, 이는 품목편의 경우 품목의 집계 범위가 엄격하고 보수적으로 산출되는 경향이 있는 반면, 본 연구모형으로부터 추정된 매출액의 경우 일부 포함 관계에 있는 제품까지 포괄하는 경우가 있기 때문인 것으로 보인다. 예를 들어 Table 5(b)의 “이불”을 키워드로 추출한 경우에서 볼 수 있듯이, “홈패션”, “침구및수예품”, “방석홈패션” 등 이불제품을 포함하는 일부 확대된 범위의 제품들이 추출되었고, 따라서 실제 출하액 규모보다 과대 추정된 결과를 나타냈다. 그러나 본 방법론의 경우 추출된 제품명을 확인한 후 사용자가 생각하는 기준에 따라 추출 기준을 다시 변경함으로써 포함되는 제품의 범위를 조절할 수 있으며,

시장규모 추정 시 유사 또는 포함관계에 있는 제품들을 함께 고려해야 하는 경우도 제품의 유형과 시장정보 사용 목적에 따라 다수 존재하므로 이는 본 방법론의 장점으로 작용할 수 있다. 예를 들어 '소파'를 생산하는 기업의 경우, '소파'에 한정된 시장규모뿐만 아니라 '소파'와 유사한 제품으로서 대체재(비슷한 유용성으로 인해 한 재화의 수요가 늘어나면 그에 따라 수요가 줄어드는 재화) 성격이 강한 '가정용 의자'를 포함한 시장규모를 파악해야 하는 경우가 있다. 또한 '소파'의 보완재(한 재화의 수요가 늘어나면 동시에 수요가 늘어나는 재화) 성격인 '테이블', 'TV 장식장' 등을 포함하는 거실용 가구 전체 시장 대한 정보도 다각적인 시장 파악과 효과적인 마케팅 전략 수립을 위해 필요한 경우가 존재한다. 따라서 본 모형의 제품군 추출 단계에서 추출된 제품명들을 확인한 후 이용자의 주관에 따라 제품군을 확대 또는 축소할 필요가 있을 경우 추출기준이 되는 유사도 임계값을 증가 또는 감소시킴으로써 추출되는 제품군의 범위를 조절할 수 있다.

한 편 과소 추정된 결과의 경우, 본 모형의 최적화된 조건에서 기준제품명과 동일한 제품들을 모두 포괄하여 정확히 추출하지 못하였기 때문에 결과적으로 과소 추정된 시장규모가 산출된 것으로 보인다. Table 5(c)의 "승강기류부품" 추출 결과에서 볼 수 있듯이, 추출된 제품명들은 다양한 종류의 승강기부품들로서 관련 없는 제품이 잘못 추출된 경우는 없었지만, 더 추출됐어야 하는 다른 승강기부품 제품들을 충분히 추출하지 못한 것을 확인하였다.

따라서 본 모형의 성능 향상을 위한 보완 연구가 필수적으로 뒤따라야 할 것이며, 다음과 같은 방법을 통해 성능 향상을 기대할 수 있다. 먼저 Word2Vec 알고리즘은 학습데이터의 양이 많을수록 추론의 정확도가 높아지는 특징이 있기 때문에 한국어 어휘에 대해 사전에 학습된 대용량 데이터 또는 백과사전, 위키피디아, 뉴스 기사, 산업·시장분석 보고서 등의 데이터셋을 활용하여 학습데이터 양을 확대하는 방안을 고려해 볼 수 있다. 또한 제품군 유형화 시 다른 방식의 기계학습 알고리즘을 적용해 볼 수 있다. 벡터 공간 내 제품명 데이터의 분포 밀도가 다양하므로 밀도 기반 클러스터링(DBSCAN; Density-Based Spatial Clustering of Applications with Noise) 기법을 적용하거나, k-평균 클러스터링(k-means clustering) 또는 계층적 클러스터링(Hierarchical clustering) 알고리즘을 적용해 볼 수 있다.

이러한 과정을 통해 본 연구에서 제시한 기본 모형을 지속적으로 보완 및 개선해나감으로써 모델을 더욱 고도화하고 발전시킬 수 있을 것으로 기대된다.

5. 결론

본 연구에서는 딥러닝 기반의 시멘틱 단어 임베딩 모델인 Word2Vec을 활용하여 자연어로 존재하는 개별 기업의 제품 데이터로부터 bottom-up 방식으로 상세한 제품 수준의 시장규모 추정이 가능하고 제품군의 범주를 자유롭게 조절할 수 있는 기계학습 기반의 시장규모 추정 방법론을 제안하였다.

방법론의 구현을 위해 먼저 제조업 분야 국내 기업별 생산제품 매출 데이터를 대상으로 제품명에 대한 Word2Vec 학습을 통해 제품명칭들을 벡터 공간에 임베딩 하였다. 그 후 특정 제품명(KSIC 색인어)에 대해 코사인 유사도를 기준으

로 유사한 제품명들을 추출하고, 이들의 개별 매출액을 연산하여 해당 제품군의 시장규모를 추정하는 모형을 구축하였다. 마지막으로 실제 조사·집계된 데이터와의 비교를 통해 모형의 성능을 검증하였다.

본 연구의 의의를 살펴보면, 먼저 본 연구는 학문적인 관점에서 최근 들어 활발해지고 있는 한국어 Word2Vec 모델을 구현함에 있어 새로운 사례연구로서 경험을 축적하고, 특히 문장이 아닌 한국어 단어만으로 구성된 문서를 임베딩하는데 적합한 학습 파라미터를 실험을 통해 최적화함으로써 앞으로 이어질 한국어 기반 모델 연구의 효율화와 가속화에 기여한다는 점에서 의의가 있다.

또한 본 연구는 특히 실무적 활용의 관점에서 다음과 같은 장점을 가진다. 단어 임베딩과 같은 기계학습 모델링 기법은 그동안 다양한 분야에서 활용되어 왔으나, 대용량의 제품정보에 대한 기계학습을 통해 이들을 유사한 제품끼리 구분하여 제품군별 상세한 시장규모를 산출하는 데 적용된 바는 없었다. 시장규모 추정 방법에 인공지능 알고리즘을 적용하여 시장규모 추정 프로세스를 지능화, 자동화함은 물론, 획일적인 기준이 아닌 사용 목적에 따라 제품군의 범주를 조절할 수 있도록 시장규모 추정 방식을 유연화하였다는데 본 연구의 차별성과 의의가 있다.

또한 상세한 제품 수준의 시장정보에 대해 다양한 분야에서 지속적인 수요가 있어왔으나, 그동안 제공되던 정보는 일부 품목에 국한되거나 산업 분야별로 파편적으로 존재하여 정보 획득에 어려움이 있었다. 본 방법론을 적용함으로써 세부적인 제품군별 시장규모 및 매출액 추정이 필요한 다양한 공공기관(기술사업화 지원기관, 산업분야별 협회 등) 또는 민간 부문(기술가치평가 수행기관, 연구개발서비스 기업 등)에서 폭넓게 활용되어 미충족 수요를 해결할 수 있을 것으로 기대된다.

아직까지 정확도 및 신뢰성 등을 향상시켜야 하는 한계점이 존재하지만 한국어에 특화된 텍스트 마이닝 및 기계학습에 대한 연구개발이 계속적으로 이어지고 있는 만큼, 개선된 알고리즘의 적용 및 알고리즘 간 결합을 통해 본 모형의 성능이 더욱 향상되고 발전할 수 있음은 자명한 사실이다. 후속 연구를 통해 문제점들을 보완해 나갈 계획이며, 제조업 외에 전 산업으로 적용분야를 확대하는 한편, 연도별 데이터의 연계를 통해 시장의 성장률 및 미래 시장수요를 예측하고, 제품 단위의 경쟁정보를 포함시켜 도출되는 시장정보의 범위를 확대하고자 한다.

참고문헌(References)

An, J., S.-H. Lee, E.-H. An and H.-W. Kim, "Fintech Trends and Mobile Payment Service Anlaysis in Korea: Application of Text Mining Techniques", *Informatization Policy*, Vol.23, No.3(2016), 26~42.

Balachandra, R. and J. H. Friar, "Factors for Success in R&D Projects and New Product Innovation: A Contextual Framework", *IEEE Transactions on Engineering Management*, Vol.44, No.3(1997), 276~287.

Chakraborty, G. and M. Krishna, "Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining", SAS global forum, (2014), 1288~2014.

Choi, B.-O. and B. S. Kim, "A Forecasting on the Market Size of Korean Solar Salt", *Journal of*

Korea Academia-Industrial cooperation Society, Vol.14, No.10(2013), 4812~4818.

Choi, S., J. Seol and S.-g. Lee, "On Word Embedding Models and Parameters Optimized for Korean", *Proceedings of the 28th Annual Conference on Human and Cognitive Language Technology*, (2016), 252~256.

Chun, S.-Y. and S.-G. Kim, "Estimation for Market Size of the Purchasing Costs for the Textbooks and Reference Books", *The Journal of Economics and Finance of Education*, Vol.19, (2010), 95~124.

Grbovic, M., V. Radosavljevic, N. Djuric, N. Bhamidipati, J. Savla, V. Bhagwan and D. Sharp, "E-Commerce in Your Inbox: Product recommendations at scale", *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2015), 1809~1818.

Harris, Z. S., "Distributional Structure", *Word*, Vol.10, No.2-3(1954), 146~162.

Heo, C. and S.-Y. Ohn, "A Novel Method for Constructing Sentiment Dictionaries Using Word2vec and Label Propagation", *Journal of Korean Institute of Next Generation Computing*, Vol.13, No.2(2017), 93~101.

Heu, J.-U., "Korean Language Clustering Using Word2vec", *The Journal of The Institute of Internet, Broadcasting and Communication*, Vol.18, No.5(2018), 25~30.

Jang, E.-S., Y.-H. Baek and S.-W. Lee, "The Survey on the State and Scale of Constitutional Medical Service Market in Korea", *Journal of Sasang Constitutional Medicine*, Vol.25, No.1(2013), 43~50.

Jun, S.-P., T.-E. Sung and S. Choi, "A Data-Based Sales Forecasting Support System for New Businesses", *Journal of Intelligence and Information Systems*, Vol.23, No.1(2017), 1~22.

Jung, Y. L., H. S. Yoo, J. H. Kim, H. G. Kim and E. S. Kim, "Estimating Apparatus for Market Size, and Control Method Thereof", Patent Application Number 10-2019-0112446, Korean Intellectual Property Office, Republic of Korea, 2019.

Kang, B.-S., "A Study on the Accuracy Improvement of Movie Recommender System Using Word2vec and Ensemble Convolutional Neural Networks", *Journal of digital convergence*, Vol.17, No.1(2019), 123~130.

Kang, H. and J. Yang, "Optimization of Word2vec Models for Korean Word Embeddings", *Journal of Digital Contents Society*, Vol.20, No.4(2019), 825~833.

Kang, H. and J. Yang, "The Analogy Test Set Suitable to Evaluate Word Embedding Models for Korean", *Journal of Digital Contents Society*, Vol.19, No.10(2018), 1999~2008.

Kang, J. and J. Cho, "The Demographic Structure, Firm Age and Economic Performance: A Local Level Analysis", *Economic Analysis*, Vol.24, No.4(2018), 101~128.

Kim, D. and M.-W. Koo, "Categorization of Korean News Articles Based on Convolutional Neural Network Using Doc2vec and Word2vec", *Journal of KIISE*, Vol.44, No.7 (2017), 742~747.

Kim, D. J., D. I. Park and J. S. Park, "Study on the Change of Marketing Strategy through Data Mining Technique", *Korea Business Review*, Vol.22, No.2(2018), 177~194.

Kim, K. and C. Park, "Automatic Ipc Classification of Patent Documents Using Word2vec and Two Layers Bidirectional Long Short Term

Memory Network", *The Journal of Korean Institute of Next Generation Computing*, Vol.15, No.2(2019), 50~60.

Kim, S. and S. Lee, "Automatic Extraction of Alternative Word Candidates Using the Word2vec Model", *Journal of Korean Institute of Information Scientists and Engineers*, Vol.23, No.12(2015), 769~771.

Le, Q. and T. Mikolov, "Distributed Representations of Sentences and Documents", *International conference on machine learning*, (2014), 1188~1196.

Lee, D.-Y., J.-C. Jo and H.-S. Lim, "User Sentiment Analysis on Amazon Fashion Product Review Using Word Embedding", *Journal of the Korea Convergence Society*, Vol.8, No.4(2017), 1~8.

Lee, K., K.-Y. Kim and Z. Lee, "Factors Affecting Purchase Intention of Smart Mobility : Integrating Text Mining and Mental Accounting Theory", *Korea Journal of Business Administration*, Vol.31, No.11(2018), 2147~2168.

Lee, M. and H.-J. Kim, "Construction of Event Networks from Large News Data Using Text Mining Techniques", *Journal of Intelligence and Information Systems*, Vol.24, No.1(2018), 183~203.

Lee, Y.-J., J.-H. Seo and J.-T. Choi, "Fashion Trend Marketing Prediction Analysis Based on Opinion Mining Applying Sns Text Contents", *Journal of Korean Institute of Information Technology*, Vol.12, No.12(2014), 163~170.

Lee, Y., Y.-J. Lee and H. Kang, "A Study on Estimating the Size of the Fashion Market through Sample Survey", *Journal of The Korean Data Analysis Society*, Vol.14, No.3(2012), 1281~1290.

Lilleberg, J., Y. Zhu and Y. Zhang, "Support Vector Machines and Word2vec for Text Classification with Semantic Features", *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing*, (2015), 136~140.

Lim, J. and H. Oh, "A Study on the New Product Forecasting Methodology", *Journal of the Korean Institute of Industrial Engineers*, Vol.18, No.2(1992), 51~63.

Liu, H., "Sentiment Analysis of Citations Using Word2vec", arXiv preprint arXiv:1704.00177, (2017).

Maaten, L. v. d. and G. Hinton, "Visualizing Data Using t-SNE", *Journal of machine learning research*, Vol.9, No.Nov(2008), 2579~2605.

Mikolov, T., K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space", arXiv preprint arXiv:1301.3781, (2013).

Nam, Y., “Analysis on the Determinants of Exit of Self-Employed Businesses in Korea”, *BOK working paper*, Vol.5, (2017), 1~37.

Ngo, D. L., N. Yamamoto, V. A. Tran, N. G. Nguyen, D. Phan, F. R. Lumbanraja, M. Kubo and K. Satou, "Application of Word Embedding to Drug Repositioning", *Journal of Biomedical Science and Engineering*, Vol.9, No.01(2016), 7~16.

Park, H. and J. Ahn, "Demand Forecasting for G2B E-Commerce Using Public Data : A Case Study of Public Procurement Service", *The Journal of Korean Institute of Information Technology*, Vol.12, No.10(2014), 113~121.

Park, S. S. and K. C. Lee, "Effective Korean Sentiment Classification Method Using Word2vec and Ensemble Classifier", *Journal of Digital Contents Society*, Vol.19, No.1 (2018), 133~140.

Park, Y.-J., Y.-B. Kim, S.-Y. Jeong, Y. J. Kim and S.-W. Son, "Network Analysis in Korean Presidential Speeches by Using Word2vec", *New Physics: Sae Mulli*, Vol.67, No.5(2017), 569~574.

Shin, M. C., *Basic Statistics for Business and Economics*, Changmin, Seoul, 2010.

Son, N. S., Y. Lee and H. Chun, "Growth and Failure of Manufacturing Plants in Korea: Single-Unit Versus Multi-Unit Plants", *Journal of Market Economy*, Vol.47, No.1(2018), 1~27.

Statistics Korea, “Report on the Census on Establishments”, 2015.

Statistics Korea, “Report on the Economic Census - Whole Country”, 2017.

Stein, R. A., P. A. Jaques and J. F. Valiati, "An Analysis of Hierarchical Text Classification Using Word Embeddings", *Information Sciences*, Vol.471, (2019), 216~232.

Vasile, F., E. Smirnova and A. Conneau, "Meta-Prod2vec - Product Embeddings Using Side-Information for Recommendation", *Proceedings of the 10th ACM Conference on Recommender Systems,* (2016), 225~232.

Weiss, S. M., N. Indurkhya and T. Zhang, *Fundamentals of Predictive Text Mining*, Springer, 2015.

Xue, B., C. Fu and Z. Shaobin, "A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec", *2014 IEEE International Congress on Big Data*, (2014), 358~363.

Yang, H., Y.-I. Lee, H.-j. Lee, S. W. Cho and M.-W. Koo, "A Study on Word Vector Models for Representing Korean Semantic Information", *Phonetics and Speech Sciences*, Vol.7, No.4(2015), 41~47.

Yang, Y.-J., B.-H. Lee, J.-S. Kim and K. Y. Lee, "Development of an Automatic Classification System for Game Reviews Based on Word Embedding and Vector Similarity", *The Journal of Society for e-Business Studies*, Vol.24, No.2(2019), 1~14.

Yoo, H. S., J. H. Seo, S.-P. Jun and J. Seo, "A Study on an Estimation Method of Domestic Market Size by Using the Standard Statistical Classifications", *Journal of Korea Technology Innovation Society*, Vol.18, No.3(2015), 387~415.

Yoon, Y. S., J.-C. Park, and S. S. Cho, "A Study on the Market Size Distribution of Artificial Information Industry", *Journal of Industrial Economics and Business*, Vol.29, No.6(2016), 2179~2198.

Abstract

A Study on Market Size Estimation Method by Product Group Using Word2Vec Algorithm

Ye Lim Jung* · Ji Hui Kim** · Hyoung Sun Yoo***

With the rapid development of artificial intelligence technology, various techniques have been developed to extract meaningful information from unstructured text data which constitutes a large portion of big data. Over the past decades, text mining technologies have been utilized in various industries for practical applications. In the field of business intelligence, it has been employed to discover new market and/or technology opportunities and support rational decision making of business participants.

The market information such as market size, market growth rate, and market share is essential for setting companies' business strategies. There has been a continuous demand in various fields for specific product level-market information. However, the information has been generally provided at industry level or broad categories based on classification standards, making it difficult to obtain specific and proper information.

In this regard, we propose a new methodology that can estimate the market sizes of product groups at more detailed levels than that of previously offered. We applied Word2Vec algorithm, a neural network based semantic word embedding model, to enable automatic market size estimation from individual companies' product information in a bottom-up manner. The overall process is as follows: First, the data related to product information is collected, refined, and restructured into suitable form for applying Word2Vec model. Next, the preprocessed data is embedded into vector space by Word2Vec and then the product groups are derived by extracting similar products names based on cosine similarity calculation. Finally, the sales data on the extracted products is summated to estimate the market size of the product groups. As an experimental data, text data of product names from Statistics Korea's microdata (345,103 cases) were mapped in multidimensional vector space by Word2Vec training. We performed parameters

* Div. of Information Analysis, Korea Institute of Science and Technology Information
** Div. of Information Analysis, Korea Institute of Science and Technology Information
*** Corresponding Author: Hyoung Sun Yoo
Div. of Information Analysis, Korea Institute of Science and Technology Information
66 Hoegi-ro, Dongdaemun-gu, Seoul 02456, Korea
Tel: +82-2-3299-6173, Fax: +82-2-3299-6139, E-mail: hsyoo@kisti.re.kr

optimization for training and then applied vector dimension of 300 and window size of 15 as optimized parameters for further experiments. We employed index words of Korean Standard Industry Classification (KSIC) as a product name dataset to more efficiently cluster product groups. The product names which are similar to KSIC indexes were extracted based on cosine similarity. The market size of extracted products as one product category was calculated from individual companies' sales data. The market sizes of 11,654 specific product lines were automatically estimated by the proposed model. For the performance verification, the results were compared with actual market size of some items. The Pearson's correlation coefficient was 0.513.

Our approach has several advantages differing from the previous studies. First, text mining and machine learning techniques were applied for the first time on market size estimation, overcoming the limitations of traditional sampling based- or multiple assumption required-methods. In addition, the level of market category can be easily and efficiently adjusted according to the purpose of information use by changing cosine similarity threshold. Furthermore, it has a high potential of practical applications since it can resolve unmet needs for detailed market size information in public and private sectors. Specifically, it can be utilized in technology evaluation and technology commercialization support program conducted by governmental institutions, as well as business strategies consulting and market analysis report publishing by private firms.

The limitation of our study is that the presented model needs to be improved in terms of accuracy and reliability. The semantic-based word embedding module can be advanced by giving a proper order in the preprocessed dataset or by combining another algorithm such as Jaccard similarity with Word2Vec. Also, the methods of product group clustering can be changed to other types of unsupervised machine learning algorithm. Our group is currently working on subsequent studies and we expect that it can further improve the performance of the conceptually proposed basic model in this study.

Key Words : Word2Vec, machine learning, text mining, market size estimation, market analysis

Received : December 9, 2019 Revised : February 12, 2019 Accepted : February 28, 2019
Publication Type : Regular Paper(Fast-track) Corresponding Author : Hyoung Sun Yoo

저 자 소 개

정 예 림

한국과학기술원에서 생명화학공학으로 공학석사 및 공학박사 학위를 취득하고 현재 한국과학기술정보연구원 데이터분석본부 기술사업화센터에서 선임연구원으로 재직 중이다. 관심 연구분야는 산업시장분석 방법론, 과학기술혁신, 기계학습, 바이오 기술사업화, 보건의료 데이터분석 등이다.

김 지 희

한양대학교에서 재료공학으로 공학석사 및 공학박사 학위를 취득하고 현재 한국과학기술정보연구원 데이터분석본부 기술사업화센터에서 선임연구원으로 재직 중이다. 관심 연구분야는 산업시장분석 방법론, 중소기업 R&D 기획, 비즈니스콘텐츠 분석 등이다.

유 형 선

한국과학기술원에서 공학 박사학위를 취득하고 현재 한국과학기술정보연구원 책임연구원으로 재직 중이며, 과학기술연합대학원대학교 과학기술정책학과 부교수를 겸임 중이다. 관심 연구분야는 기계학습 기반의 산업시장 자동분석, 과학기술정책, 시계열 수요예측, 복잡계 네트워크 등이다.