벤처창업연구 제12권 제5호 (통권53호)
Asia-Pacific Journal of Business Venturing and Entrepreneurship Vol.12 No.5 pp.1-12

텍스트마이닝을 이용한 창업 관련 연구 동향 분석

한성수 (호서대벤처대학원 박사과정)*
양동우 (호서대벤처대학원 교수)**

국 문 요 약

본 연구는 현재까지의 국내에서의 창업 관련 연구 동향을 알아보기 위한 것이다. 이를 위해서 RISS에 등록되어 있는 창업 관련 석·박사 논문들 중 한글 초록이 제공되는 논문들을 2009년까지의 창업연구 전기(前期), 2010년부터의 창업연구 후기(後期)로 나누어 최빈 사용 빈도 분석, TF-IDF 분석 및 토픽 분석 등 3가지 분석 방법을 사용하여 메타 분석을 수행하였다.

본 연구의 결과로 보면 3가지 분석에서 공통적으로 드러나는 것은 창업 교육 및 정부 정책이나 지원은 계속되는 연구주제이며, 및 소상공인 창업 등도 계속되는 연구 주제이나 창업 연구 후기에 더욱 강조됨을 알 수 있었다. 또 창업 연구 후기에서는 실증 분석이 강화됨을 알 수 있었다. TF-IDF 분석에 의해 드러나는 것은 창업 연구 전기에는 퇴역군인 관련 연구도 많이 진행되었으며, 후기에는 문화콘텐츠 및 노령화 사회를 맞아 노년층 관련 연구가 많이 진행됨을 알 수 있었다. 토픽 분석을 통해 추가 확인된 사항은 전 기간에 브랜드 관련 연구가 진행되었으며, 창업 연구 전기에 벤처 관련 연구, 창업가 특성 및 창업 동기, 창업 전략 등 창업 준비 관련 연구가 많이 진행되었으며, 여성 창업도 연구되었음을 알 수 있었다. 후기에는 창업 성과를 중시하고 산학 협력, 창업 투자, 사회적 기업 등 창업 관련 연구가 다양화 되었다는 것을 알 수 있었다.

본 연구는 창업 관련 메타 분석에 텍스트마이닝, 토픽 분석 등 최근 이슈가 되고 있는 방법을 적용해 보았다는 데에 그 의의가 있다 하겠다. 향후 창업 관련하여 더욱 세분화된 다양한 주제로 연구가 필요할 것이다.

핵심주제어: 창업, 창업연구, 메타분석, 텍스트마이닝, TF-IDF, 토픽 분석

Ⅰ. 서론

취직의 어려움과 다양한 형태의 정부의 창업지원 프로그램으로 20, 30대의 청년창업이 늘어나고 있으며, 베이비부머 세대의 본격적인 퇴직으로 은퇴자들의 소상공인 창업도 늘어나고 있는 추세이다. 이에 따라 창업 교육 및 창업 관련 연구도 확대되고 있다. 수십 년간의 연구 성과로 이제 창업 관련 연구 논문들도 많이 쌓여 있어, 이 논문들의 다양한 연구동향의 흐름을 파악해 보는 것도 의미 있는 일일 것이다.

기존 논문들에 대한 메타분석은 연구경향을 파악하기 위해 매우 중요하며, 과거 많은 메타 연구들이 행해져 왔다. 메타분석은 과거 또는 최신의 연구 동향을 파악한다던가, 실증적 연구에서 공통 사항을 뽑아내 정리하는 등의 기존의 타겟팅 된 연구들에서는 찾아 낼 수 없는 의미를 찾아내는 중요한 방법론이다. 그런데 사회가 발전하고 복잡해지면서 갈수록 발표되는 논문의 수는 기하급수적으로 증가하고 있고 기존에 사용하였던 메타 분석 방법론으로 이를 분석하는 것은 점점 더 어려워지고 있다. 이에 따라 새로운 방법론이 필요해지고 있다. 과거에도 타 학문에서의 방법론을 적용하여 학문의 발전을 이뤄나가는 경우가 종종 있어왔다. 사회과학, 특히 경영학 중에서도 마케팅 분야는 통계학의 회귀분석 방법론을 적용하여 많은 발전을 하게 되었다. 학문들 간의 융·복합 현상은 최근에 더욱 심화되고 있는데, 정보시스템학 분야와의 융합을 통해 기존 논문에 텍스트마이닝을 적용하므로써 의미 있는 정보를 추출하려는 많은 시도들이 존재하고 있다.

본 연구는 학술연구정보서비스(RISS) 논문 데이타베이스에 있는 창업관련 석·박사 논문 중 텍스트 파일 형태로 한글 초록을 획득할 수 있는 662건의 논문들에 대하여 논문의 초록을 텍스트마이닝을 통하여 분석하여 각 기간별 의미 있는 키워드들을 뽑아내어 창업 관련 연구의 특징 및 주요 연구 주제 등에 대해서 알아보고자 한다.

Ⅱ. 이론적 배경

2.1 창업 연구

* 제1저자, 호서대벤처대학원 박사과정, sshan1@naver.com
** 교신저자, 호서대벤처대학원 교수, dwyang@office.hoseo.ac.kr
· 투고일: 2017-09-04 · 수정일: 2017-10-12 · 게재확정일: 2017-10-25

창업은 사전적 의미로는 사업을 새로 시작하는 것이다. 창업은 일반 기업의 경영과 맥을 같이 하고 있고, 경영의 여러 요소들을 포함하고 있지만, `창업은 변화가 많고 불확실성이 훨씬 높은 특성을 지닌다. 또한 창업자에게는 더 높은 수준의 기업가정신이 요구된다. 창업기업 경영에는 급격한 상황 변화에 따른 빠른 의사결정이 필요하며, 의사 결정의 미숙과 상황의 불확실성에 따라 자원도 불합리하게 분배되는 경우가 발생하게 된다. 이런 창업의 특성으로 창업학은 경영학의 분화된 학문이지만, 창업 관련 연구는 기존의 경영학 방법론 외에 다양한 방법론을 필요로 한다.

창업 관련 연구의 시작은 미국에서 창업학 관련 강좌가 개설된 1960년대 말부터라고 볼 수 있으나 국내 창업학 연구는 1986년 창업학이라는 용어를 사용한 후부터 본격적으로 시작되었다고 할 수 있다. 그 이후 창업에 대한 연구가 계속 확대되면서 2003년에는 창업학회가 설립되어 학회 차원의 지원이 이루어지게 되었다. 비록 학문으로서 정착된 기간은 길지 않지만 창업학은 경영학뿐만 아니라 교육학, 심리학, 사회학, 경제학, 공학 등의 다른 학문으로부터 이론적 체계나 실천적 방법론 등을 흡수하면서 계속 발전해 오고 있다(박재환·박명수, 2011). Low & MacMillan(1988)은 기업가 정신(Entrepreneurship) 연구와 관련하여 목적, 이론적 관점, 초점(focus), 분석 수준, 시간 골격(Time Frame), 방법론 등 6가지 연구 분야에 대한 통일된 정의를 제안했으며, Dean et al.(2007)은 기업가정신(Entrepreneurship) 연구에 사용된 연구 기술 관련 데이터 분석 방법을 추적하여, 1976년부터 1985년까지의 기간에는 기술적 통계(記述的 統計), 비모수 통계 및 상관 분석이 주류를 이루고, 1986년-1995년의 기간에는 비모수 통계, 기술적 통계 및 다중 회귀 분석이 주류를 이루고, 1996년-2004년의 기간에는 계층적 회귀분석, 단순/다중 회귀분석이 주류를 이룬다는 것을 확인하였다. Vesper(1976)는 창업학 연구 영역을 10개 영역으로 재분류하였으며, 한주희·고언정(2007)은 창업개론, 환경분석 및 창업준비, 기업가정신, 창업경영 및 전략, 분야별 창업, 사례연구 및 세미나 등 6개의 대분류, 32개의 중분류를 포함하는 연구대상의 창업학 교과과정 분류표를 만들었다.

창업 관련 연구 및 실무는 벤처창업, 소상공인 창업 등의 창업 형태에 대한 연구, 창업자에 대한 특성 연구와 기업의 특성 연구 등이 있고, 사업기회 포착과 평가 및 창업 실무 등이 있다. 이 외에도 창업 관련 연구에는 창업정책, 창업교육 및 창업투자 등이 있다.

2.2 메타 분석

메타 분석은 과거부터 현재까지의 연구경향을 파악하기 위한 연구방법으로서, 기존 논문들에 대한 메타분석은 연구경향을 파악하기 위한 측면에서 매우 중요하다.

메타 분석은 과거 많은 연구들이 행해져 왔는데, 인문, 사회과학, 자연과학 및 의학 등에서도 메타 분석은 활발하게 연구되고 있으며(이건혁, 2016; 이용희 외, 2015), 이현우·손영곤(2016)은 위기관리 커뮤니케이션의 실질적 효과가 어떻게 되는가를 확인하기 위하여 43편의 논문을 가지고 메타분석을 수행하여 위기 커뮤니케이션 효과에 대한 일반적인 결론을 도출하였다. 송민선 외(2016)은 한국학 연구 논문에서 기존의 논문의 형식 구조 검색시스템에 비하여 텍스트 구조 기반 메타데이터 시스템이 학술적 정보 요구에 맞는다는 것을 시스템구축을 통하여 확인하였다.

창업 관련해서도 메타 분석들이 이루어지고 있는데, 이용희 외(2015)는 국내 벤처기업들에 대하여 창업 성공에 관한 연구 동향을 분석하였다. Barringer et al.(2005)는 비즈니스 유형 및 현재 활동, 경영진, 문화/가치/인센티브, 혁신적인 접근법 그리고 미래 계획의 5가지 분야에 대해 메타분석을 수행하여, 창업하는 기업가(Entrepreneur)의 개인적 특성이 회사의 발전에 영향을 미친다고 하였고, 이에 가장 중요한 변수는 관련 업계 경험, 대학 교육 및 기업가 정신이라고 하였다. 또한 과거 논문들에서의 일반적인 경향을 찾아내기 위한 메타 분석 연구도 이루어지고 있는데, 류준호(2014)는 1998년부터 2013년까지의 382편의 논문을 대상으로 메타분석 연구를 하여 IT산업 분야, 외식산업분야, 콘텐츠산업분야에 대한 창업 관심이 높이 나타나고, 창업지원분야 중 창업자, 창업교육, 외부환경 등이 주로 연구되고 있으며, 주요 창업연구 대상으로는 초기에는 벤처기업 중심 연구이었다가 최근에는 소상공인 창업 연구가 활발히 진행되고 있으며, 창업자 특성 측면에서는 청년창업이 주된 연구 대상이며 최근에는 시니어 창업에 대한 관심이 높아진다고 하였다. 김희수(2016)는 951편의 학위 논문을 대상으로 연구를 진행하여 1997년까지의 연구들은 창업기업, 창업금융, 창업정책, 창업인, 창업보육, 창업교육 등의 창업 생태계 기반에 관한 연구들이 주를 이뤘으며, 1998년부터 2008년까지의 연구들은 창업보육차원의 연구가 활발히 진행되었고, 2009년부터 2015년까지는 창업인 차원의 연구들이 대폭 증가하였다는 시사점을 도출하였다.

기존의 메타 분석 연구들은 논의의 성격, 연구주제, 방법론, 저자 등의 형식구조에서 답을 찾는 것이 일반적이며(이건혁, 2016), 독립변인들이 종속변인에 영향을 미치는 효과를 검증하는 등의 변인들 간의 관계를 검증하거나, 구조적 모형을 점검하는 등의 경우도 있었으나, 비교적 적은수의 개별 연구들을 대상으로 메타분석을 실시하였다는 한계가 있었고, 샘플의 한계로 또 다른 조절변인들이 존재하는지 여부를 확인하는 것은 어려웠다(이현우·손영곤, 2016). 정보시스템의 발달로 최근 들어 발생하는 데이터양은 기하급수적으로 증가하고 있으며 이를 위해 빅데이터 처리에 사람들이 관심을 가지고 있다. 빅데이터는 데이터가 규모(Volume)가 크고, 데이터의 형태가 다양(Variety)하며, 분석되는 처리 속도(Velocity)가 빠른 경우를 일컫는다(McAfee & Brynjolfsson, 2012).

사회 과학에서 매년 나오는 논문의 양은 직접 코딩하여 분석하기에는 어려운 정도로 기하급수적으로 증가하기 때문에, 이것을 기존의 방법론대로 일일이 수작업으로 분류하여 분석

한다는 것은 점점 어려운 작업이 되어 가고 있다. 즉 논문들에 대한 메타 분석도 빅데이터 분석이 되어야 하는 경우가 많아지게 된다. 사회과학/경영학 논문의 빅데이터 분석을 위해서는 데이터마이닝의 한 부류인 텍스트마이닝을 이용하여 수행이 필요하다. 텍스트마이닝 기법은 여러 종류가 있는데, 기본적인 빈도분석과 TF-IDF 기법을 이용한 분석, SNA 분석, 네트워크 텍스트 분석, 군집 분석 등의 분석 방법이 많이 사용되었다. 최근에는 토픽 분석으로도 많이 연구되고 있는데, 특히 LDA가 많이 사용되고 있다.

메타 분석은 이러한 창업연구의 특성을 파악하기 위한 좋은 방법론이다. 국내외에서 창업 및 기업가/기업가정신 관련 메타 분석이 많이 이루어지고 있지만, 빅데이터 분석 및 텍스트마이닝 방법에 의한 메타 분석은 국내에서는 아직 이루어지지 않았고, 해외에서도 사례가 거의 없는 실정이다.

2.3 TF-IDF 분석

TF-IDF(Term Frequency - Inverse Document Frequency)는 정보 검색 및 기계학습 등에서 많이 사용하는 단어 중요도의 가중치로 여러 문서로 이루어진 문서군이 주어질 때 특정 단어가 한 문서 내에서 얼마나 중요한 지를 나타내는 통계적 수치이다. Salton & Buckley(1988)는 1988년 논문에서 단지 하나의 단어에 가중치를 둠으로써도 다른 정교한 텍스트 표현보다 텍스트 색인에 있어서 우수한 결과를 얻는다는 것을 보여 주었다. TF(Term Frequency, 단어빈도)는 여러 가지 형태로 표현할 수가 있는데, 가장 간단한 형태는 문서 d내의 단어 t의 총빈도를 $f(t,d)$라 할 때, 가장 간단한 형태는 다음과 같으며, 특정 단어가 문서 내에서 얼마나 자주 등장하는지를 나타내는 값으로, 이 값이 높을수록 단어의 반복출현 횟수가 높으므로 그 단어가 문서에서 중요할 가능성이 있다.

$$tf(t,d) = f(t,d) \tag{1}$$

이를 그대로 사용하는 경우 문서의 길이가 큰 경우에는 단어가 등장할 빈도수가 커질 가능성이 높으므로, 이에 대한 보완이 필요한데, 불린빈도(Boolean Frequency)는 단지 해당 단어가 문서 내에 있는지 여부만 따지며 다음과 같이 표현된다.

$$tf(t,d) = \begin{cases} t\text{가 } d\text{에 한번이라도 나타나면 } 1 \\ \text{그렇지 않으면 } 0 \end{cases} \tag{2}$$

또 다른 형태로서는 정규화된 빈도로서 증가빈도(Augmented Frequency)가 있다. 이는 문서의 길이에 따라 단어의 빈도값을 조절하는 효과를 가진다. 특정 단어 t와 문서 d에 대한 증가빈도 수식은 다음과 같다.

$$tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{\max\{f(w,d) : w \in d\}} \tag{3}$$

그러나 창업연구에 있어서 '창업'이나 '연구'같은 단어는 거의 모든 문서에 존재하지만 정보로서의 가치는 떨어지게 된다. 이를 보완하기 위해 IDF의 개념이 등장한다. DF(Document Frequency, 문서 빈도)는 특정 단어 자체가 문서군 내에서 자주 사용되는 경우 그 단어가 흔하게 등장하는 일반적인 용어라는 의미이며, IDF(Inverse Document Frequency, 역문서 빈도)는 DF의 역수로 이 값이 높으면 그 단어가 일부 문서들에만 나타난다는 의미이며 그 단어는 핵심어가 될 가능성이 있다. 어떤 단어가 단 하나의 문서에만 존재하는 경우에 IDF는 가장 높은 값을 가지게 된다. 실제로는 전체 문서수를 해당 단어를 가지는 문서 수로 나눈 값에 로그를 취하게 된다.

$$idf(t) = \log \frac{D}{|\{d \in D : t \in d\}|} \tag{4}$$

TF-IDF는 앞의 TF 값과 IDF 값의 곱으로 주어지는데, 여기에서는 식(2)의 증가빈도와 식(4)의 IDF 값의 곱으로 하였다.

$$\begin{aligned} tfidf(t) &= tf(t,d) \times idf(t) \\ &= [0.5 + \frac{0.5 \times f(t,d)}{\max f(w,d) : w \in d}] \log \frac{D}{|\{d \in D : t \in d\}|} \end{aligned} \tag{5}$$

특정 문서 내에서 단어 빈도가 높을수록, 그리고 전체 문서들 중 그 단어를 포함하는 문서가 적을수록 TF-IDF 값이 높아진다(김주영, 2017). 즉 TF-IDF 값이 높을수록 중요한 단어라는 의미가 되므로 여러 문서 집합으로부터 특정 단어가 특정 문서에서 얼마나 중요한가를 판단하는 기준이 된다(김현정· 외, 2015). 이 TF-IDF는 목적에 따라서 여러 가지 변형이 있을 수 있는데, 이성직·김한직(2009)은 TF-IDF에 대해 6가지 변형식을 사용하여 인터넷 포털 사이트에 있는 대용량 문서 집합에서 분야별 주제를 제시할 수 있는 키워드 추출 기법을 제안하였다(이성직·김한준, 2009). Ramos(2003)는 쿼리가 주어졌을 때 해당되는 문서들을 찾는 데에 TF-IDF가 이용가능하다는 것을 연구하였다.

TF-IDF의 결과는 표로서 나타낼 수가 있는데, 이는 식별성이 떨어지므로 이를 다시 워드 클라우드 등을 이용하여 표시할 수 있다. 워드 클라우드는 주어진 상위 빈도 단어들에 대하여 각 단어들의 빈도수에 비례하는 글자 크기로 시각적으로 표현하므로, 각 키워드의 상대적 출현 빈도를 직관적으로 나타낼 수 있다.

2.4 토픽(주제) 분석

토픽 분석 방법은 각 텍스트 문서는 여러 개의 토픽(주제)으로 구성되어 있다고 보고, 논문에 숨겨져 있는 토픽을 찾아내는 분석 방법이다(Blei et al., 2003; Blei, 2012). 문서군에 출현하는 단어들의 연관관계를 이용하여 군집화 하는 방법이라고 할 수 있다. 각 단어들은 단어자루(Bag of Words)를 가정하고

있어 문법이나 단어의 순서 등은 무시하게 되며 군집화된 토픽들에 포함되는 단어들은 중복이 허용된다. 이 감춰진 토픽들이 몇 개가 있는지를 확인하고, 이 토픽들을 표현하는 단어들이 무엇인지를 알아내는 위해 주로 통계적 분석 방법을 사용한다. 현재 주로 사용되는 토픽 모델링 분석 방법은 LSA(Latent Semantic Analysis), pLSA(probabilistic Latent Semantic Analysis), LDA(Latent Dirichlet Allocation) 와 CTM(correlated topic model) 등이 있다(Blei & Lafferty, 2006; Sidorova et al, 2009; Lee et al, 2010).

LSA, LDA는 토픽 및 단어들 간의 상관관계가 없다고 가정을 하지만, CTM은 토픽들 간의 상관관계를 가정하기 때문에 실제 문서의 모형화에 더 유용하다. 그러나 CTM은 계산시간이 오래 걸리고, 토픽 분포의 변동 모수의 분포에 대한 업데이트에서 기울기 최적화에 의한 적합이 이루어져야만 하는 문제점이 있다(Blei & Lafferty, 2006; 김규하·박철용, 2015). LDA는 의미적으로 일관성 있는 토픽들을 생산한다는 장점을 가지고 있기 때문에 텍스트 분석에서 많이 사용된다(Mimno & McCallum, 2008: 강범일 외, 2013에서 재인용).

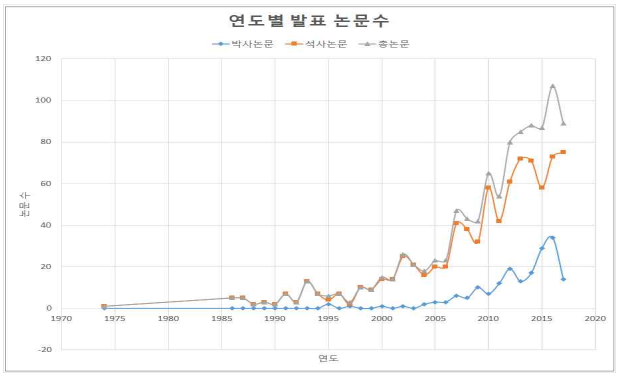
본 논문에서는 현재 토픽 모델링 기법 중에서 사회과학계에서 가장 많이 사용되고 있는 LDA(Latent Dirichlet Allocation; 잠재 디리클레 할당)를 사용하여 분석한다. 이는 Blei 등이 제안한 확률 그래프 모델로서 한 토픽에 대한 단어들의 포함 확률을 디리클레 분포를 이용하여 모델링하는 것이다(Blei et al, 2003). 이는 대량의 문서들을 통계학적으로 처리해 검색 등에 활용할 수 있는 분석 방법이다(Blei, 2012). LDA는 단어의 순서와 무관한 단어자루 가정 하에서, 각 토픽(주제)은 확률분포에 의해 생성된다고 가정한다. 즉 한 토픽 안에서 동시에 나타나는 단어들은 관찰값이 높게 나타나고, 이를 근거로 잠재적 토픽을 추론하게 된다. 각 문서는 이 토픽들의 확률분포로 구성된다고 가정한다.

박자현·송민(2013)은 문헌정보학의 주요 학술지들에 실린 논문 초록을 이용하여 LDA로 토픽 모델링 분석하였으며, 토픽 모델링에서 도출된 연구주제와 문헌정보학의 주제분류표와 비교·분석하였다. 강병일 등(2013)도 LDA 기반의 토픽 모델링을 사용하여 신문 기사를 대상으로 오피니언 마이닝을 수행하여 대선 이슈에 대한 각 언론매체들의 입장을 분석하였다. Chandra, et al.(2016)은 2,334 명의 사회적 기업가(Social Entrepreneur)들에 대한 토픽 모델링 연구를 통해서 사회적 기업가 정신에 대해서 연구를 하였다. 이 토픽들의 확률적분포를 직접적으로 계산하는 것은 불가능하기 때문에 본 논문에서는 붕괴된 깁스 샘플링(Collapsed Gibbs Sampling) 방법을 사용하여 LDA를 계산한다. 붕괴된 깁스 샘플링 방법을 사용하기 위해서는 사전에 토픽의 수를 결정해야 하며 토픽의 수 결정은 분석 결과에 영향을 주기 때문에 중요하다. 본 연구에서는 토픽의 수의 결정에 있어서 주어진 확률모델이 얼마나 잘 측정되는지를 나타내는 혼잡도(Perplexity)를 사용하여 판별하는데(Brown et al, 1992), 본 논문에서는 혼잡도 및 교차검증값을 계산하기 위해서 은닉 토픽의 수에 따른 로그우도(Log-Lokelihood)의 조화평균값을 계산하는 방법을 사용한다(Grün & Hornik, 2011; Ponweiser, 2012). 즉 토픽의 수를 2개부터 순차적으로 증가시켜 로그우도의 조화평균 값이 가장 크게 나오는 모형의 토픽의 수를 최적 토픽의 수로 결정하며, 최적 모형이라고 판단한다(김지은, 2017).

III. 연구 방법

3.1 분석 대상

학술연구정보서비스(RISS)에 등록되어 있고 제목에 창업이 들어가 있는 창업 관련 석·박사 논문은 1974년부터 2017년 2월까지 1,010편이다. <그림1>에 1,010편에 대한 연도별 창업 관련 학위 논문 수를 나타내었다. 1974년 창업 관련 첫 석사 논문이 발표되었으며, 박사논문은 1995년에서야 첫 논문이 발표 되었다. 발표 논문의 수는 2000년 대 들어 증가세가 커지는 것을 알 수 있으며, 2007년과 2010년에 증가세가 큰 폭으로 증가하는 것을 알 수 있다.



<그림 1> 연도별 창업 관련 발표 논문 수

2007년의 창업관련 논문의 수가 급증한 것은 2004년 중앙대, 2005년 호서대, 한밭대, 진주산업대, 예원예술대에 창업대학원이 설립되면서 창업학 연구가 본격화되었기 때문으로 풀이된다(목영두, 2011). 2010년의 증가세는 2008년 말의 글로벌 금융위기 후에 2009년도의 구조조정 및 부실기업 퇴출 등의 정책(누초경, 2010)과 일자리 창출 정책 등 정부의 정책 등에 기인한 바가 크다고 하겠다. 본 연구에서는 컴퓨터 프로그램에 의한 분석이 가능하도록 창업 관련 학위 논문 중 RISS에서 한글 초록을 획득할 수 있는 662편을 대상으로 하였다. <그림 1>에서 보면 2007년 및 2010년에 논문 수가 급증한 것을 알 수 있다. 2006년까지는 분석 대상의 논문이 너무 적어 분석의 의미가 없으므로 2009년, 2010년을 기준으로 하여 전후의 논문 흐름을 살펴보고자 한다. 즉, 1994년부터 2009년까지의 120 편과 2010년부터 2017년2월까지의 542 편으로 나누어서 한글 초록을 추출하여분석 대상으로 하여 비교 분석하

였다. 2006년으로 나누지 않은 이유는 2005년까지의 논문의 수가 작아 의미 있는 결과를 나타내지 못할 가능성이 높기 때문이다.

<표 1> 분석 대상 학위 논문 수

기간별	분석대상 논문 수		
	박사	석사	합계
창업연구 전기(1994~2009)	14	106	120
창업연구 후기(2010~2017.2)	113	429	542
소계	127	535	662

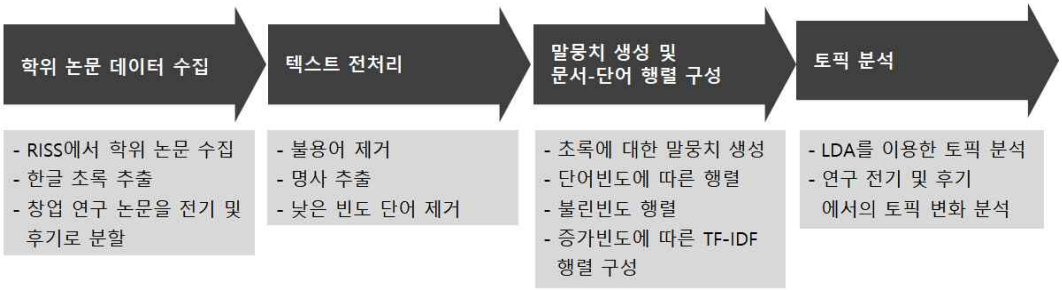
<표1>에서 보면 창업연구 후기는 창업연구 전기보다 기간이 짧음에도 불구하고 창업 관련 논문이 훨씬 많다는 것을 알 수 있다. 특히 박사 논문의 수가 급격히 증가하고 있는 것을 알 수 있다.

3.2. 연구방법

본 연구는 창업 관련 학위논문을 텍스트마이닝으로 메타 분석하는 것으로 다음과 같은 과정을 거친다.

3.2.1. 학위 논문 데이터 수집

RISS에서 석·박사 학위논문의 서지사항을 엑셀의 xlsx의 형태로 출력하였다. 이 출력은 서지 정보가 순차적으로 기록되어 있는 형태이므로, 이를 행으로는 각 학위논문에 대응시키고, 열에는 연도, 초록 등의 필요한 정보를 정리하였다. 이중에서 한글 초록이 존재하는 662건을 추출하였으며, 2009년까지의 논문과 2010년 이후의 논문으로 나누어 분류하였다. 또한 한자어와 한글어의 혼용은 텍스트마이닝을 어렵게 하므로 한자어를 전부 한글화하여 저장하였다.



<그림 2> 창업 논문 텍스트마이닝 분석 방법

3.2.2. 텍스트 전처리

논문의 초록은 일반적인 문장이라 컴퓨터로 처리하기에는 적절치 않은 비구조적 데이터이며, 따라서 이를 구조적 형태로 변경하는 것이 필요하다. 이를 위해 R프로그램을 사용하여 특수문자, 구두점 등을 제거하였으며, 주요단어(Keywrod)는 주로 명사 형태를 가지므로 R프로그램의 'KoNLP' 패키지를 이용하여 명사만을 추출하였다. 한 글자 단어의 경우에는 '편', '수' 등의 의존명사가 많고 독립성이 없어서 독자적인 의미를 가지지 못한 경우가 많으며, 정보로서의 가치가 없는 경우가 많기 때문에 두 글자 이상 단어로 한정하였다. 또한 '창업', '연구' 등은 창업 관련 논문에는 빈번히 들어가는 단어들이므로 이런 단어들은 불용어로서 제거하였다.

이렇게 추출된 단어집합을 가지고 각 연도별 학위논문들의 빈출 단어들을 정리하였고, 이것은 식(1)의 2글자 이상의 명사 단어들에 대한 기본적인 단어빈도(Term Frequency)의 문서-단어 행렬(Document-Term Matrix) 형태이다. 이 행렬이 말뭉치(Corpus)[1] 생성을 위한 입력이 되며, 또한 후에 토픽 분석의 입력으로 사용하였다.

3.2.3. 말뭉치 생성 및 문서-단어 행렬 구성

텍스트마이닝을 하기 위해서 텍스트 전처리된 내용들을 먼저 말뭉치 구조로 변경해야 한다. 이를 위해서 R 프로그램의 텍스트마이닝 패키지인 'tm'을 이용하여 말뭉치로 변경하였다.

이 말뭉치를 바탕으로 식(2)의 TF 불린빈도의 문서-단어 행렬을 만들었다. 이 불린빈도는 초록의 크기에 따라 단어빈도 값이 크게 나오는 것은 막을 수 있지만 일반적으로 자주 사용되는 단어인 경우에는 값이 커지는 것을 막을 수 없는 단점이 있다. 이를 보완하기 위해서 식(5)의 TF-IDF 문서-단어 행렬을 만들었다. 이 두 행렬은 장단점이 있으므로 이 두 가지 결과를 비교하면서 분석하여 의미 있는 결과를 얻어낼 수 있다. 이 행렬들은 표로서 나타낼 수가 있는데, 본 연구에서는 다시 이를 워드 클라우드 형태로 나타내어 시각적으로 쉽게 중요 단어를 판단할 수 있게 하였다. 각 행렬들은 창업연구 전기와 후기로 나누어 도출하고 서로 비교하여, 공통적인 부분과 각 기간별 주로 연구되는 부분을 확인하였다.

3.2.4. 토픽 분석

본 연구에서는 토픽 분석 중에 가장 많이 쓰이고 있는 LDA를 사용하여 분석하였다. LDA를 사용하기 위해서는 토픽(주제)의 수가 사전에 결정되어야 하는데, 본 논문에서는 토픽의 수를 결정하기 위하여 Grün & Homik(2011)이 제안한 로그우드(Loglokelihood)의 조화평균값을 계산하는 방법을 사용하였

1) 말뭉치(corpus)는 언어 연구를 위해 텍스트를 컴퓨터가 읽을 수 있는 형태로 모아 놓은 언어 자료

다. 즉 토픽의 수를 2에서 100까지 순차적으로 변화시키면서 로그우드 조화평균의 크기를 보고 가장 큰 값에 해당하는 값을 토픽의 수로 결정하였다.

이렇게 해서 나온 토픽들의 개수는 많기 때문에 이를 전부 분석하는 것은 바람직하지 않다. 따라서 상위 10개의 토픽에 대해서만 검토를 하였다. 추출된 각 토픽에서 포함된 단어들을 확인할 수 있는데, 각 토픽들에 대한 토픽명은 연구자가 각 토픽에 포함된 단어들을 보고 정해야 한다. 이 부분은 각 연구자의 경험과 직관이 필요한 부분이다. 이 토픽 분석도 창업연구 전기 및 후기로 나누어 비교 분석하였다.

Ⅳ. 연구 결과

4.1. 단어 빈도분석

662편의 논문들의 초록에서 명사들을 추출한 후, 2글자 이상의 단어들에 대한 단어 빈도분석을 수행하였다. <표2>는 본 연구에서 사용한 각 연도별 창업관련 학위 논문 수 및 각 연도 논문들에서 사용된 최빈 10개 단어들에 대한 내용이다. 연도별 최빈 단어 분석을 2002년부터 수행한 이유는 2001년 이전에는 논문의 수가 적어 의미 있는 결과가 나오지 않기 때문이다. 전체적으로 보면 해마다 논문의 수가 증가하는 것을 알 수 있고, 논문의 수가 증가하면서 포함된 주요 단어(명사)의 수도 증가하는 것을 알 수 있다.

<표 2> 연도별 초록의 상위 빈출 단어 구성

	발행연도	논문 수	주요 단어 수	상위 10개 빈출 단어 수
창업 연구 전기	2002	12	244	지원 교육 경제 창업자 경영 성공 개발 서비스 특성 창업자금
	2003	6	148	기술 사업 성장 지원 창업자 정부 경제 준비 센터 기술력
	2004	4	61	교육 사업 지원 창업자 프로그램 성장 활성화 정보 직업 산업
	2005	6	116	인식 경제 변화 성공 개발 벤처 사업 산업 창업자 지원
	2006	10	253	지원 기업 정부 경제 환경 사업 기술 성장 효과 실시
	2007	23	414	사업 창업자 경제 기술 성공 지원 성장 검증 실시 선행연구
	2008	18	337	지원 개발 사업 교육 성과 경제 실시 프로그램 정부 창업자
	2009	30	482	사업 지원 기술 환경 성과 실증 성장 정부 창업자 검증
창업 연구 후기	2010	54	718	경영 사업 창업자 교육 실시 지원 성과 검증 기술 실증
	2011	41	657	실증 검증 경제 교육 기술 성과 시사점 경영 정부 창업가
	2012	60	777	교육 검증 성과 효과 사업 실증 창업자 지원 시사점 가설
	2013	72	843	교육 실시 지원 창업자 검증 경제 프로그램 활성화 정부 시사점
	2014	79	952	교육 창업자 지원 실시 사회 사업 경제 검증 정부 실증
	2015	72	867	검증 교육 지원 창업자 기술 유의 성과 실시 가설 프로그램
	2016	83	976	교육 지원 창업자 사업 검증 정부 실증 프로그램 경제 성과
	2017	81	980	교육 기업 지원 검증 성과 경제 정부 창업자 활성화 기술

<표 2>연도별 빈출 단어들이다. 여기서는 기업, 특성, 요인, 목적, 대상, 영향, 효과 등의 의미 없는 단어들을 불용어로서 빼고, 상위 빈도의 단어들을 나타내었다. 2009년을 전후하여 비교해 보면 2009년까지의 전기에는 경제에 대한 내용이 상대적으로 많고, 2010년 이후의 후기에는 실증 및 검증 등의 내용이 많아 실증 분석이 많이 이루어짐을 알 수 있다. 교육도 계속 나타나 창업 교육은 계속되는 연구과제이며, 지원도 전 연도에 대해 골고루 나와 창업은 정부 지원과 깊은 관련이 있는 것을 알 수 있다.

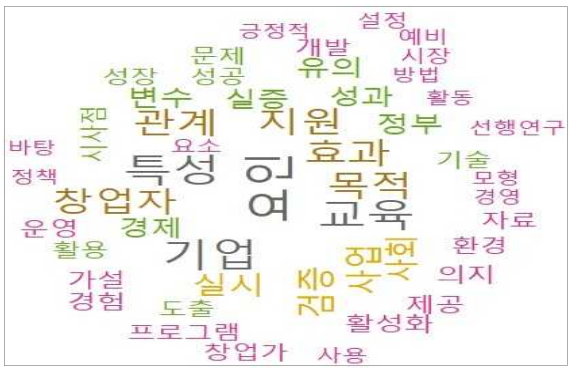
<그림 3>은 창업연구 전기(1995~2009)의 최빈출 단어 50개를 이용한 워드클라우드이다. 여기서는 각 문서에 들어간 횟수를 하나로 간주하는 불린빈도 방법을 사용하였는데, 이는 문서의 크기에 대한 효과를 제거하기 위함이다.

<그림 3>에서 보면 정부, 지원 등이 보이고 프로그램, 성과 등이 보이며, 산업, 경제, 성장 등의 용어가 나타난 것으로 볼 때, 창업연구 전기 기간에는 정부의 기업 지원 정책 및 이를 통한 산업 발전, 경제 성장 등에 연구의 주안점이 주어져 있는 것을 알 수 있다.

<그림 3> 창업연구 전기의 주요 키워드에 대한 워드 클라우드

<그림 4>는 창업연구 후기(2010~2017.2)까지의 최빈출 단어 50개를 이용한 워드클라우드이다. <그림 4>에서 보면 기업, 교육, 창업자 등이 상대적으로 많이 나오고 있으며, 지원, 정부, 성과 등도 빈도수가 높다는 것을 알 수 있다. 상대적으로 보면 전반기 보다 교육이 크게 나타나고, 요인, 검증, 변수, 실증 등이 크게 나타난 것으로 미루어 창업연구 전기 보다 실증 연구가 많이 진행 되었다는 것을 알 수 있다.

전·후반기에 걸쳐 전반적으로 보면 정부, 지원, 성과 등이 공통적으로 나와 창업과 정부 등의 지원과의 관계에 대한 연구가 계속 이루어짐을 알 수 있었다.



<그림 4> 창업연구 후기의 주요 키워드에 대한 워드 클라우드

4.2. TF-IDF 분석

<그림 5>는 창업연구 전기(1994~2009)까지의 TF-IDF 가중치를 두었을 때의 값이 큰 단어 50개를 이용한 워드 클라우드이다. 이 종류의 워드 클라우드에서는 IDF 효과로 특정 단어가 논문 1개에만 있을 때에 그 값이 클 가능성이 높다.

<그림 5>에서 보면 호두과자, 맥주, 요리, 인터넷마케팅, 쇼핑몰 등 특정 논문에만 존재하는 단어가 높은 값을 가지는 것을 알 수 있다. 이 단어들을 유추해보면 소상공인 창업 관련 아이템에 대한 연구 진행이 많다는 것을 알 수 있다. 제대, 전역 등 퇴직군인들에 의한 창업 연구도 진행되었다는 것을 알 수 있다.

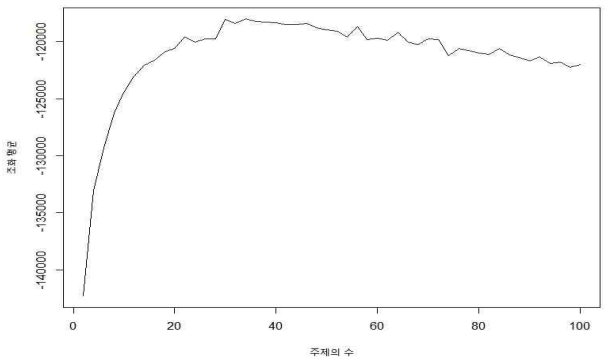
<그림 5> 창업연구 전기의 TF-IDF 가중치 워드 클라우드

<그림 6>은 창업연구 후기(2010~2017.2)까지의 TF-IDF 가중치를 두었을 때의 값이 큰 단어 50개를 이용한 워드 클라우드이다. <그림 6>에서 보면 애니메이션, 공예 등이 높은 값을 나타내 문화콘텐츠에 대한 관심이 나타난 것을 알 수 있으며, 교회 등이 높은 값을 차지하고 있는 것을 알 수 있으며, 한정식, 탄산수 등 외식업 관련 연구도 높은 값을 보이고 있다. 이외에도 현장체험의 중요성이 나오기 시작하였으며, 정보서비스, 온톨로지, 교회 등이 등장하였고, 노령화 시대에 맞춰 노년층에 대한 관심이 있다는 것을 알 수 있다. 또 사내기업가가 상위 키워드여서 사내기업가에 대한 연구도 진행되었다는 것을 알 수 있다.

<그림 6> 창업연구 후기의 TF-IDF 가중치 워드 클라우드

4.3 토픽 분석

LDA는 문서들에 들어 있는 토픽을 확률통계적으로 분석할 수 있는 좋은 도구이지만 사전에 토픽(주제)의 개수를 지정해야 한다. <그림 7>은 창업연구 전기의 초록들에서 토픽의 수를 2에서 100까지 변화시키면서 로그우도(Log-Likelihood)의 조화평균의 값을 계산하였다. 조화평균이 최대가 되는 점은 토픽(주제)의 수가 34개일 때이며, 이 때 값으로 LDA를 계산하여 토픽들을 추출하였다.



<그림 7> 창업연구 전기에서의 토픽(주제)의 수에 따른 조화평균의 값

<표 3> 창업연구 전기에서의 주요 토픽 및 관련 단어 들

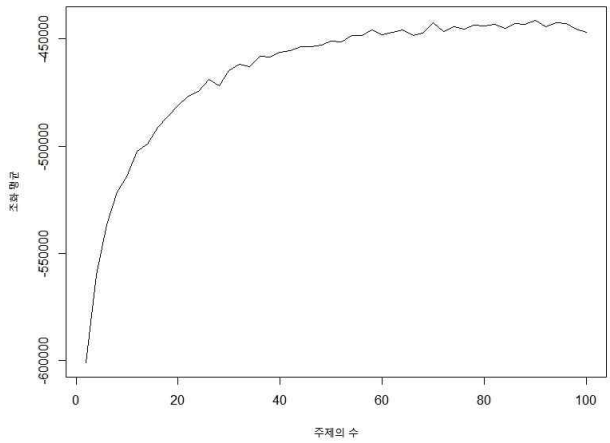
토픽번호	토픽명	토픽에 따른 단어들
T1	창업 전략	시장 전략 네트워크 산업 진입 제품 경쟁 선택 협력 포지셔닝
T2	창업 동기	기회 소규모 산업 사례 내용 기준 경험 변화 김치 품질
T3	창업가 특성	유형 특성 성향 위험 혁신 업무 음식 활용 변화 행동
T4	이미지/브랜드	이미지 브랜드 구매 맥주 제품 소비자 하우스 기업이미지 의도 자동차
T5	기술 혁신	기업 지향성 기술 시장 기술혁신 효과 개발 조직 활동
T6	창업 조사	분야 국내 문헌 기술 방법 이용 관시 외국 조사 우리나라
T7	창업 교육	교육 학생 고등학교 직업 청소년 학교 참여 프로그램 목적 진로
T8	기술 평가	기술 평가 가치 사업 확보 활성화 모형 국내 반영 금융
T9	창업 환경	프로그램 효과 대학원 요인 비교 교육 협력 발견 유의미 독립
T10	여성 창업	여성 분야 기업가 활동 사회 특성 정보기술 경제 동기 가능성

<표 3>은 이 경우의 상위 10개의 토픽에 따른 단어들이다. 토픽의 명칭은 단어들을 보고 유추한 것이다. 창업 연구 전기에는 창업 교육, 창업가 특성등과 창업을 준비하고 시작하기 위한 창업 동기, 창업전략, 창업 조사, 창업 환경 등이 주요 토픽이며, 기술 혁신, 기술 평가 등 벤처/기술 관련 연구도 진행되고 있다는 것을 알 수 있었다. 그 외에 여성 창업 및 디자인 관련해서도 활발히 연구되고 있었다.

<그림 8>은 창업연구 후기의 초록들에서 토픽의 수를 2에서 100까지 변화시키면서 로그우도 조화평균의 값을 계산한 결과이다. 조화평균이 최대가 되는 점은 토픽이 90개일 때이며, 이 이후는 점차로 감소함을 알 수 있다. 이 90개를 토픽의 수로 하여 LDA를 계산하여 토픽들을 추출하였다. 창업연구 전기의 34에 비하여 후기에는 토픽의 수가 90개로 증가하였는데, 이는 전기의 대상 논문 120건에 비해 후기는 대상 논문이 542개로 증가한 때문으로 여겨진다.

<표 4>는 이 경우의 상위 10개의 토픽에 따른 단어들이다. 각 토픽의 명칭은 역시 단어들을 보고 유추한 것이다. 후기에는 전기부터 계속되어진 창업 교육과 브랜드/마케팅이 연구 되고 있었고, 전기보다는 창업 성과를 중시하고 실증 연구가 강조되었으며, 소상공인 창업이나 외식업 창업이 강조되었고, 그 외 창업 정책, 산학 협력, 창업 투자, 사회적 기업 등 창업 관련 연구가 전기에 비해 다양화 되었다는 것을 알 수 있었다.

전, 후기를 비교해 보면 브랜드 및 창업 교육은 공통적으로 계속되는 연구 주제라는 것을 알 수 있었으며, 창업 관련 정부의 정책은 계속되는 연구 주제이나 후기에 들어서 훨씬 연구가 많이 되는 주제라는 것을 알 수 있었다. 전기의 창업 연구에는 창업전략이나, 창업 동기, 창업조사 창업 환경 등 창업 준비 관련 연구나 창업자 특성 등에 주안점이 있었다. 후기 들어 두각을 나타내는 연구 주제는 창업 성과를 나타내고, 실증 연구가 더욱 강조되고 있는 것을 알 수 있었으며, 창업 투자, 산학 협력 등 창업 환경 관련 주제가 많이 연구가 되고 있으며, 문화콘텐츠, 소상공인, 외식 등 창업 분야의 다양화에 따른 다양한 연구 및 사회적 기업에 대한 연구도 많이 연구되는 주제라는 것을 알 수 있었다.



<그림 8> 창업연구 후기에서의 토픽(주제)의 수에 따른 조화평균의 값

<표 4> 창업연구 후기에서의 주요 토픽 및 관련 단어 들

토픽번호	토픽명	토픽에 따른 단어들
T1	창업 정책	문화콘텐츠 가치 분야 정책 공예 조사 관점 방안 산업 활성화
T2	창업 성과	성과 무적 재무 시사점 검증 설문 마케팅 시스템 실증 모형
T3	브랜드/마케팅	브랜드 마케팅 시장 변화 스타트업 패션 중심 기회 부채 발견
T4	실증 연구	가설 변수 검증 모형 설정 유의 실증 관계 바탕 채택
T5	창업 교육	대학원 진학 지식 상관 유의 선택 온라인 결정 전문 학문
T6	산학 협력	대학 증가 이전 강좌 변화 산학 협력 개설 토대 설문조사
T7	창업 금융	투자 조합 산업 수익률 벤처캐피탈 시장 제도 벤처 영화 조합원
T8	소상공인 창업	소상공인 운영 인식 소셜커머스 의미 제공 가치 상황 사업체 독립
T9	사회적 기업	사회 사회적기업 증가 직접 문제 경제 바탕 실증 지속적 연결
T10	외식업 창업	레스토랑 선택 내부 고려 점포 고객 모형 관리 음식 외식

4.4 분석 종합 및 선행문헌과의 차이

단어 빈도 분석 및 TF-IDF에 의한 분석의 경우, 전 기간에 정부, 지원, 성과 등은 계속 높은 관심을 갖는 주제이었으며, 인터넷마케팅, 쇼핑몰, 공예 등 소상공인 창업 관련 연구도 계속 되는 주제인데 후반기에는 더욱 강조됨을 알 수 있었다. 창업 연구 전기에는 산업 및 경제 성장 등도 주요한 주제였으며, 제대군인들의 창업도 주된 관심사 이었다는 것을 알 수 있었다. 후기에는 전기에 비해 실증연구가 많이 진행되었으며, 문화콘텐츠 및 노년층 관련 연구도 진행되었고, 사내기업가에 대한 연구도 주요 연구 주제라는 것을 알 수 있었다.

토픽 분석의 결과로 보면, 브랜드 및 창업 교육은 공통적으로 계속되는 연구 주제라는 것을 알 수 있었으며, 창업 관련 정부의 정책도 계속되는 연구주제로 후기에 연구가 더 많이 진행되고 있는 것을 알 수 있었다. 후기 들어서는 창업 성과 및 실증 연구가 더욱 강조되고 있는 것을 알 수 있었으며, 창업 투자, 산학 협력 등 창업 환경, 창업 보육 관련 연구도 많이 진행되고 있으며, 창업 관련 연구로 사회적 기업 및 산학 협력, 창업 금융 등을 포함 다양화 되었다는 것을 알 수 있었다.

류준호(2014)의 결과와 비교해 보면 IT산업분야, 외식산업분야, 콘텐츠산업분야의 창업 관심이 높게 나타난 것으로 연구된 것은 본 연구의 TF-IDF 분석 및 토픽 분석의 결과와 일치함을 알 수 있었다. 또 창업 연구 대상으로 초기에 벤처기업 중심이었다가 최근 소상공인 창업 중심 연구가 주로 연구되고 있다는 것은 토픽 분석에서 전기에서 기술 혁신, 기술 평가 등이 많이 연구되는 점으로 벤처 관련성이 있다고 할 것이고, 후기에서는 소상공인 창업, 외식업 창업이 활발히 연구되고 있다는 점에서 맥락이 같다고 할 것이다. 최근 시니어 창업에 관심이 높다는 것은 본 연구에서도 TF-IDF 분석에서 노년층 관련 단어가 중심어로서 일맥 상통한다고 할 것이나, 청년 창업이 주된 연구 대상이라는 것은 본 연구에 명확히 나타나지는 않았다. 김희수(2016)와의 차이점은 1997년까지의 연구에서 창업기업, 창업금융, 창업정책, 창업인, 창업보육, 창업교육 등의 창업 생태계 기반에 관한 연구들이 주를 이뤘다고 주장하였는데, 본 연구에서는 창업 금융 관련한 부분은 크게 드러나지 않았으며, 2009년부터 2015년까지는 창업인 차원의 연구들이 대폭 증가하였다는 부분도 본 연구에서는 크게 드러나지 않았다. 앞의 선행 연구들에서 나타나지 않은 부분은 토픽 분석을 통해 브랜드가 관심 있게 계속 연구되고 있었으며, 창업 연구 후기에는 창업 연구 분야가 다양화 되었다는 것을 알 수 있었다.

Barringer et al.(2005)은 메타분석을 통하여, 창업하는 기업가(Entrepreneur)의 개인적 특성이 회사의 발전에 영향을 미친다고 하였고, 중요한 것이 관련 업계 경험, 대학 교육 및 기업가 정신이라고 하였다. 대학 교육 및 기업가 정신은 본 연구의 결과인 창업 연구에서 중요하게 다루어지는 내용인 창업 교육, 창업가 특성과 관련이 있다고 하겠으나, 관련 업계 경험은 분명히 나타나지 않은 점은 차이점이라 하겠다.

V. 결론

본 연구에서 창업 관련 학위논문들에 대하여, 한글 초록 추출이 가능한 1994년부터 2017년 2월까지의 논문 662편에 대해 텍스트마이닝 방법에 의한 메타 분석을 수행하였다. 연구의 흐름을 살펴보고자 2009년과 2010년을 기준으로 2009년까지의 창업 연구 전기와 2010년 이후의 창업연구 후기로 나누어 여러 가지 형태로 연구 동향을 살펴보았다. 적용된 방법은 빈도분석, TF-IDF분석, 토픽 분석이며 각 분석에 따라 다른 결과가 나왔다.

본 연구의 결과로 보면 3가지 분석에서 공통적으로 드러나는 것은 창업 교육 및 정부 정책이나 지원은 계속되는 연구 주제이며, 및 소상공인 창업 등도 계속되는 연구 주제이나 창업 연구 후기에 더욱 강조됨을 알 수 있었다. 또 창업 연구 후기에서는 실증 분석이 강화됨을 알 수 있었다. TF-IDF 분석에 의해 드러나는 것은 창업 연구 전기에는 퇴역군인 관련 연구도 많이 진행되었으며, 후기에는 문화콘텐츠 및 노령화 사회를 맞아 노년층 관련 연구가 많이 진행 됨을 알 수 있었다. 토픽 분석을 통해 추가 확인된 사항은 전 기간에 브랜드 관련 연구가 진행되었으며, 창업 연구 전기에 벤처 관련 연구, 창업가 특성 및 창업 동기, 창업 전략 등 창업 준비 관련 연구가 많이 진행되었으며, 여성창업도 연구되었음을 알 수 있었다. 후기에는 창업 성과를 중시하고 산학 협력, 창업 투자, 사회적 기업 등 창업 관련 연구가 다양화 되었다는 것을 알 수 있었다.

본 연구를 통해서 창업연구의 동향에 대해 기존의 메타 분석에서 확인되지 않았던 사실들을 확인하였으나, 빅데이타 분석으로서의 텍스트마이닝 기법을 다양하게 적용하지는 못한 관계로 한계가 있었으며, 추가 연구할 부분도 나타났다. 첫째, 창업이라는 테마가 너무 넓고 광범위하여 결과 도출이 명확하지 않은 면이 있었다. 주제를 창업 교육이나 정부의 창업 지원, 소상공인 창업 등 소주제로 하였으면 더 좋은 결과가 나왔지 않을까 생각된다. 둘째, 텍스트마이닝에는 본 연구에서 사용하지 않았던 여러 기법들이 있는 바, 본 연구에서 사용하지 않은 연관규칙이나 네트워크 분석 등 다른 방법을 병용하여 사용하여 분석하였다면 다른 측면에서 더 나은 결과가 도출될 가능성이 있을 것이다. 셋째, 창업 관련하여 학위논문만이 아니라 학술지 논문을 포함하여 분석할 필요성이 있었다. 학술지에는 학생들의 논문뿐만 아니라 그 분야의 전문가들의 논문도 포함되어 있어 결과의 차이가 있었을 가능성이 있다.

본 연구에서는 의미를 가지는 복합명사들을 개별 단어들로 분리하여 처리하였는데, 이를 통해서 의미를 찾기가 어려운 점이 있었다. 향후 의미 가지는 복합명사들을 하나의 의미 가지는 주제어로서 처리하면 더 좋은 결과가 나타날 것이다.

본 연구는 처음으로 창업 관련하여 텍스트마이닝을 적용하였다는 점에서 의의가 있다. 그러나 향후 네트워크 텍스트 분석, 연관 규칙 분석, 군집 분석 등 더 다양한 분석방법과 여성 창업, 소상공인 창업 등 창업 관련하여 더욱 세분화된 다양한 주제로 연구가 필요할 것이다. 또한 석·박사 학위논문을 분리하여 비교 분석하면 더 의미 있는 결과가 나올 수 있을 것이며, 본 연구의 학위논문과 관련 학술지 논문에 분석한 결과를 비교하는 것도 의미가 있을 것이다.

REFERENCE

강범일·송민·조화순(2013). 토픽모델링을 이용한 신문 자료의 오피니언 마이닝에 대한 연구, *한국문헌정보학회지*, 47(4), 315-334.

김규하·박철용(2015). 토픽 모형 및 사회연결망 분석을 이용한 한국데이터정보과학회지 영문초록 분석, *한국데이터정보과학회지*, 26(1), 151-159.

김주영(2017). *개인 맞춤형 상품 리뷰 정보 시각화 방법론 연구*, 석사학위논문, 숭실대학교

김지은(2017). *빅데이터를 활용한 대학구조개혁 평가의 키워드 및 토픽 분석*, 박사학위논문, 서울대학교

김현정·조남옥·신경식(2015). 항공산업 미래유망분야 선정을 위한 텍스트 마이닝 기반의 트렌드 분석, *지능정보연구*, 21(1), 65-82.

김희수(2016). *창업관련 학위논문 추세에 관한 탐색적 연구*, 석사학위논문, 호서대학교

누초경(2010). *2008년 글로벌 금융위기에 대한 한국경제 대처방안*, 석사학위논문, 대구카톨릭대학교

류준호(2014). 창업 관련 연구에 대한 메타분석, *벤처창업연구*, 9(1), 51-67.

목영두(2011). *대학의 창업교육 체계화를 위한 창업학 교육과정 개발 모형 연구*, 박사학위논문, 중앙대학교

박자현·송민(2013). 토픽모델링을 활용한 국내 문헌정보학 연구동향 분석, *정보관리학회지*, 30(1), 7-32.

박재환·박명수(2011). 창업학의 전략적 선택과 연구 과제, *경영교육연구*, 26(4), 315-339.

송민선·고영만·이승준(2016). 한국학 연구 논문의 텍스트 구조 기반 메타데이터 검색 시스템 개발 연구, *정보관리학회지*, 33(3), 155-176.

이건혁(2016). 언론학연구DML 메타분석 연구, *언론학연구*, 20(1), 29-51.

이성직·김한준(2009). TF-IDF 의 변형을 이용한 전자뉴스에서의 키워드 추출 기법, *한국전자거래학회지*, 14(4), 59-73.

이용희·홍광표·박수홍(2015). 국내 벤처기업의 창업 성공에 관한 연구동향 분석, *벤처창업연구*, 10(6), 15-26.

이현우·손영곤(2016). 국내 위기관리 커뮤니케이션 연구에 대한 메타분석, *홍보학연구*, 20(3), 139-172.

한주희·고언정(2007). 창업학 대학원의 교과과정에 대한 한국과 미국의 비교연구, *직업교육연구*, 26(2), 105-131.

Barringer, B., Jones, F., & Neubaum, D.(2005). A quantitative content analysis of the characteristics of rapid-growth firms and their founders., *Journal of Business Venturing*, 20(5), 663-687.

Blei, D. M.(2012). Probabilistic Topic Models, *Communications of the ACM*, 55, 77-84.

Blei, D. M., Ng, A. & Jordan, M.(2003). Latent Dirichlet Allocation, *Journal of Machine Learning and Research*, 3, 993-1022.

Blei, D. M., & Lafferty, J. D.(2006). Correlated topic models, *Advances in neural information processing systems,* 18, 147-154.

Brown, P. F., Pietra, V. J. D., Mercer, R. L., Pietra, S. A. D. & Lai, J. C.(1992). An estimate of an upper bound for the entropy of English, *Computational Linguistics,* 18(1), 31-40.

Chandra, Y., Jiang, L. C. & Wang, C. J.(2016). Mining social entrepreneurship strategies using topic modeling, *PloS one*, 11(3), e0151342.

Dean, M., Shook, C., & Payne., G.(2007). The Past, Present, and Future of Entrepreneurship research: Data Analytic Trends and Training, *Entrepreneurship Theory and Practice*, 25, pp.101-113.

Grün, B., & Hornik, K.(2011). topicmodels: an R package for fitting topic models, *Journal of Statistical Software,* 40(13), 1-30.

Hahn, J. H., & Ko. U. J.(2007). Comparative Study of Entrepreneurship Programs in the Graduate Schools of Korea and the US, *The Journal of Vocational Education Research,* 26(2), 105-131

Kang, B. L., Song, M., & Jho, W. S.(2013). A Study on Opinion Mining of Newspaper Texts based on Topic Modeling, *Journal of the Korean Library and Information Science Society,* 47(4), 315-334.

Kim, G. H., & Park, C. Y.(2015). Analysis of English abstracts in Journal of the Korean Data & Information Science Society using topic models and social network analysis, *Journal of the Korean Data & Information Science Society,* 26(1). 151-159.

Kim. H. S.(2016). *An Exploratory Study on the Startup Research*, Master dissertation, Graduate School of Hoseo University.

Kim, H. J., Jo, N. Y., & Shin, K. S.(2015). Text Mining-Based Emerging Trend Analysis for the Aviation Industry, *Journal of Intelligence and Information Systems*, 21(1), 65-82.

Kim, J. E.(2017). *Keyword and Topic analysis on the College and University Structural Reform Evaluation Using Big Data*, Doctoral dissertation, The Graduate School of Seoul National University

Kim, J. Y.(2017). *Information Visualization Methods for Personalized Product Reviews*, Master dissertation, Graduate School of Soongsil University.

Lee, G. H.(2016). The Meta-analysis of "Journalism and Communication", *Journalism and Communication*, 20(1), 29-51.

Lee, H. O., & Sohn, Y. K.(2016). A Meta-Analytic Review of Crisis Communication Research in Korea, *Journal of Public Relations,* 20(3), 139-172.

Lee, S., Song, J., & Kim, Y.(2010). An Empirical Comparison of Four Text Mining Methods, *Journal of Computer Information Systems*, 51, 1-10.

Lee, S. G., & Kim, H. J.(2009). Keyword Extraction from

News Corpus using Modified TF-IDF, *The Journal of Society for e-Business Studies*, 14(4), 59-73.

Lee, Y. H., Hong, K. P., & Park, S. H.(2015). Analysis of Research Trends in the Successful Establishment of Venture Companies: with Priority Given to Domestic Articles Between 1998 and 2014, *Journal of Business Venturing and Entrepreneurship*, 10(6), 15-26.

Lou, X. K.(2010). *Cause and Alternatives for 2008 USA's Financial Crisis*, Master dissertation, The Graduate School of Catholic University of Daegu.

Low, M. B., & MacMillan, I. C.(1988). Entrepreneurship: Past research and future challenges, *Journal of management*, 14(2), 139-161.

McAfee, A., & Brynjolfsson, E.(2012). Big data: the management revolution, *Harvard business review*, 90(10), 60-68.

Mimno, D., & McCallum, A,(2008). Topic models conditioned on arbitrary features with Dirichlet-multinomial regression, *The 24th Conference on Uncertainty in Artificial Intelligence,* 411-418.

Mok, Y. D.(2011). *A Study on the Entrepreneurship Curriculum Development Model designed to Systemize Entrepreneurship Education in Undergraduate School*, Doctoral dissertation, The Graduate School of Chung-ang University.

Park, J. W., & Park, M.(2011). The Strategic Choice and Research Themes of Venturing Entrepreneurship, *Journal of Management Education*, 26(4), 315-339.

Park, J. H., & Song, M.(2013). A Study on the Research Trends in Library & Information Science in Korea using Topic Modeling, *Journal of the Korean Society for Information Management*, 30(1), 7-32.

Ponweiser, M.(2012). *Latent Dirichlet Allocation in R*, Doctoral dissertation, Vienna University of Economics and Business.

Ramos, J.(2003). Using tf-idf to determine word relevance in document queries, *Proceedings of the first instructional conference on machine learning*, 242, 133-142.

Salton, G., & Buckley, C.(1988). Term-weighting approaches in automatic text retrieval, *Information processing & management*, 24(5), 513-523.

Sidorova, A., Evangelopoulos, N., & Valacich, J.(2008). Uncovering the Intellectual Core of the Information Systems Discipline, *MIS Quarterly*, 32(3), 467-482.

Song, M. S., Ko, Y. M., & Lee, S. J.(2016). A Study on Developing a Metadata Search System Based on the Text Structure of Korean Studies Research Articles, *Journal of the Korean Society for Information Management*, 33(3), 155-176.

Vesper, K. H.(1997). Sub-Fields of Entrepreneurship Research. *Academy of Management Proceedings*, Academy of Management, 1, 440-444.

Yoo, J. H.(2014). Meta-analysis about the study related with foundation: As the center from 1998 to 2013 treatises, *Journal of Business Venturing and Entrepreneurship*, 9(1), 51-67.

Analysis of Research Trends Related to Start-Up Using Text Mining

Han, Sung-Soo*
Yang, Dong-Woo**

Abstract

The purpose of this study is to investigate the trends of the start-up research in Korea. To accomplish this, meta-analysis was carried out using text mining methodology by dividing the entrepreneur-related master's and doctoral theses registered in RISS into the first term of entrepreneurship research by 2009 and the second term of entrepreneurship research from 2010.

As a result of this study, it can be seen from the three different analysis that the entrepreneurship education and government policy and support are the subject of continuous research topics in the whole period and that the researches on small business start-ups have been studied continuously and conducted more in the second half. In addition, empirical analysis is strengthened in the latter stage of entrepreneurial research. The TF-IDF analysis reveals that many researches on veterans have been carried out in the field of entrepreneurship research, and in the latter period, it was found that many studies related to the elderly were conducted with cultural contents and aging society. In addition, research on brand-related research has been carried out throughout the entire period, and research on venture-related research, characteristics of entrepreneurs, entrepreneurship motivation and start-up strategy have been conducted a lot and female entrepreneurship was also studied. In the latter period, we have emphasized entrepreneurial achievements and found that research on start-ups such as industry-academia cooperation, start-up investment, and social enterprise diversified.

This study is meaningful to apply the method which is becoming a recent issue such as text mining and topic analysis to the meta-analysis related to start-up. Future research will need to be undertaken on a variety of more detailed topics related to entrepreneurship.

Keywords: start-up, entrepreneurship, meta-analysis, text mining, TF-IDF, topic modeling

* First Author, Graduate School of Venture, Hoseo University, sshan1@naver.com
** Corresponding Author, Professor, Graduate School of Venture, Hoseo University, dwyang@office.hoseo.ac.kr