生産性論集 第29卷 第1號(2015年 3月)
Productivity Review, Vol.29, No.1, March 2015

논문 접수일 : 2015. 03. 04
수정 접수일 : 2015. 03. 12
게재 확정일 : 2015. 03. 15

텍스트마이닝 방법론을 활용한 기업 부도 예측 연구

The Prediction of Corporate Bankruptcy Using Text-mining Methodology

최정원* · 한호선** · 이미영*** · 안준모****
Jung-won Choi · Ho-sun Han · Mi-young Lee · Jun-mo Ahn

<목 차>

* 주저자, 딜로이트안진, 건국대학교 경영학과 박사과정, jungwochoi@deloitte.com.
** 공동저자, 유세스파트너스, 건국대학교 경영정보학과 박사과정, hshan@ucesspartners.com
*** 교신저자, 건국대학교 경영정보학과 교수, yura@konkuk.ac.kr
**** 공동저자, 건국대학교 경영정보학과 교수, joonan@konkuk.ac.kr

Ⅰ. 서 론

기업 부도가 급격히 증가하여 한국경제 전반이 큰 위기를 맞이하였던 과거의 국가경제 위기 사태 이후, 기업의 부도 위험에 관한 관심이 높아짐에 따라 기업 부도 위험을 평가하고 측정하는 방법은 계속 발전하였다. 하지만 기존의 수많은 연구들이 이미 진행되었음에도 불구하고 기업 운영 환경이 글로벌화 되고 그 밖에 여러 경영환경의 변화가 발생함에 따라, 기업 부도 위험을 보다 정확히 예측하고 평가하는 정교한 모형의 필요성은 나날이 증가하고 있다.

기업의 부도를 예측하기 위하여 전통적으로 가장 많이 기반이 되는 자료는 재무정보이다. 재무정보를 활용한 기업 부도 예측 방법론은 기업의 현재 재무현황 및 상태를 반영하여 기업 부도의 가능성을 평가하는 모형으로서 가장 정확하고 객관적인 방법이라 할 수 있다. 하지만 재무정보가 각 기업의 결산 시점 이후에 정기적으로만 작성되므로 기업 경영현황의 급격한 변화를 즉각적으로 반영하는 데에는 한계가 있다.

이러한 단점을 보완하기 위하여 무디스(Moody's)사의 KMV 모형으로 대표되는 '주가' 정보를 이용한 부도예측 모형이 제시 되었다. 기업의 주가는 투자자들에 의하여 실시간으로 평가된 결과로 시장가격이 형성되므로, 기업의 현황 및 재무상태 등의 정보를 가장 빠르게 반영하여 주는 정보 중 하나이다. 따라서 주가를 활용하여 기업 부도예측을 수행한다면 기업 부도 징후를 다른 방법론에 비하여 선제적으로 파악할 수 있다.

하지만 주가 정보는 금융시장과 해당 산업의 상황에 따라 기업의 가치 외에 다른 영향으로 인하여 변동될 가능성이 있다. 또한 상장기업 외에는 주가 정보를 활용할 수 없는 한계점을 가지고 있다.

본 연구에서는 주가와 함께 기업에 대한 가장 선제적인 정보 원천이며 보다 다양한 기업에 대한 정보를 획득할 수 있는 뉴스 콘텐츠를 활용한 기업부도예측의 가능성에 관하여 논하고자 한다. 뉴스 콘텐츠는 다양한 매체에서 가장 빠른 정보를 제공하며, 보도의 대상 또한 제약이 없어 매우 방대한 정보 원천으로서 활용가치 할 수 있다. 더욱이 뉴스 정보가

즉각적으로 주가에 반영되는 상장기업의 경우 뉴스 정보 활용도가 주가보다 덜할 수 있겠으나, 비상장기업의 경우 뉴스 정보는 주가를 대체할 수 있는 적시성 있는 정보로서 충분한 활용 가치가 있을 것이다.

뉴스 콘텐츠를 활용한 기업 부도 예측을 위해서는 최근 빅 데이터(big-data)분석에서 많이 활용되고 있는 텍스트마이닝(text-mining) 방법론을 활용하여야 한다. 텍스트마이닝 방법론은 최근 발전된 정보처리 기술과 인프라를 활용하여 뉴스, 인터넷 등의 텍스트 문서로부터 정보를 획득, 키워드의 패턴을 분석하고 이를 토대로 예측을 수행하는 방법론으로서 최근 그 활용 영역을 확장해 나가고 있다. 텍스트마이닝은 데이터-마이닝과 유사한 개념이지만, 기존의 데이터-마이닝이 관계형 데이터베이스나 XML과 같은 구조화된 데이터들만을 처리할 수 있는 반면 텍스트문서, e-메일, HTML 파일과 같은 비정형 또는 반정형화 된 데이터를 일정한 형식과 조건을 만족하는 자료로 가공하여 분석하는 방법론을 텍스트마이닝으로 별도 구분하고 있다.

본 연구에서는 주요 포털 사이트에서 검색이 가능한 뉴스 콘텐츠를 확보하여 텍스트마이닝 방법론으로 부도와 연관되는 주요 키워드를 도출하고 분석한다. 이를 위하여 부도가 발생하는 기업과 정상 기업 간의 주요 뉴스 키워드의 빈도(frequency)분석과 연관성(association) 분석을 수행하여 부도 기업 뉴스의 특성을 분석하고, 의사결정나무 (decision Tree)를 산출하여 뉴스 텍스트를 활용한 부도 예측 모형을 산출하였다. 이러한 과정을 통하여 부도가 발생하는 기업의 뉴스에서 주로 나타나는 키워드의 특성과 이를 활용한 사전적인 부도예측의 가능성에 관하여 논하고자 한다.

1. 선행연구

국내에서 본 연구와 유사하게 뉴스 콘텐츠 분석을 통한 부도예측을 수행한 사례는 찾기 어렵다. 해외 연구로서는 Lu et al.(2013)가 기업 부도예측과정에서 텍스트마이닝을 활용하여 기업 신용위험 조기경보 모델을 도출하는 연구를 수행하였다. 이 연구는 ‘Distress Intensity of Default-Corpus(DIDC)’ 라는 텍스트마이닝으로 산출한 변수를 로지스틱 회귀

분석 변수로 활용하여 기업부도예측 모형을 도출함으로서 뉴스 정보가 기업부도 예측에 유용한 정보임을 증명하였다. 또한 Olson et al.(2012)는 기업부도 예측을 위한 데이터-마이닝 방법론의 유용성을 연구하면서 텍스트마이닝 기법의 활용 가능성을 논하였다.

비슷한 연구로 Martinez et al.(2012)은 텍스트마이닝을 활용하여 경제 기사에 나타는 투자자들의 심리 분석(sentiment analysis)을 수행하였다. 뉴스에 나타는 단어가 부정적 단어인지 긍정적 단어인지를 판별하여, 뉴스 Text 분석 시 나타나는 긍정과 부정의 신호(signal)대로 상당수의 기업 주가가 변동하는 것을 실증 분석하였다. Mittermayer and Knolmayer(2006) 또한 텍스트마이닝을 활용한 주가 예측 방법의 효과를 검증하였으며, 효과적인 감성 사전의 구축이 예측력에 매우 중요함을 설명하였다.

텍스트마이닝 방법론에 대하여는 복수의 선행연구자들이 정의를 내리고 있다. 앞서 언급 한 바와 같이 최윤정·박승수(2002)는 구조화되지 않은 대규모의 텍스트 집단으로부터 새로운 지식을 발견하는 과정을 의미하는 것이라고 정의하였다. 또한 김근형·오성렬(2009)은 텍스트마이닝을 다양한 정보원천으로부터 자동적으로 정보를 추출함으로써 이전에 알려지지 않았던 새로운 정보를 발견하는 정보기술이라고 정의하였다.

배상진·박철균(2003)은 텍스트마이닝을 4 단계로 나누었는데 문서수집, 문서 전처리, 텍스트 분석, 그리고 결과 해석 및 정제 단계이다. 전처리 과정은 다시 텍스트마이닝에 필요없는 단어 또는 기호를 정제하는 정제 과정과 문장의 정확한 의미 파악을 위해서 각 단어의 어간을 파악하고 동의어를 할당하는 정규화 과정으로 나누었다. 정규화 과정은 또 다시 한글 처리를 위해서 문장에서 최소의 의미단위를 추출해 내는 형태소 분석(morphological analysis) 단계와 통사구조를 파악하는 구문 구조 분석(syntactic analysis) 단계, 의미 구조를 추출하는 의미 분석(semantic analysis) 단계, 그리고 문장들 사이의 관계를 분석하는 문맥 분석(discourse analysis) 단계로 나누었다. 텍스트 분석 과정은 텍스트 군집화(text clustering), 텍스트 분류(text classification), 그리고 텍스트 요약(text summarization) 으로 나누어 설명하였다. 텍스트 군집화는 텍스트의 집단을 내용의 유사도에 따라 여러 개의 소집단으로 분할하는 과정으로서 데이터에 대한 기반 지식 없이 분석 초기에 행하여 결과를 분석할 수 있다는 장점이 있으며 중복 혹은 유사한 문서를 제거하고, 다른 문서의 주제와 다른 주제를 가진 문서를 구별하고, 대량의 문서집합의 개요를

획득하는 데 적용할 수 있다고 한다. 텍스트 분류란 텍스트의 내용에 따라 미리 정의해놓은 범주를 부여하는 과정인데, 군집과 달리 분류를 수행하기 위해서는 각 항목을 위한 학습데이터를 사용자가 선정하여 훈련시키는 과정이 필요하다고 정의하였다. 텍스트 요약은 문서의 전체 내용을 반영할 수 있는 일부 내용을 추출하는 과정으로 표면수준접근(surface level approach), 개체수준접근(entity level approach), 그리고 화법수준접근(discourse level approach)의 3가지 기법이 사용되는데 일반적으로 3가지 중 2가지 이상의 기법을 조합해서 이용한다.

텍스트마이닝 과정에 대하여 김근형·오성렬(2009)도 전처리 과정과 텍스트 분석 과정으로 나누어 설명하였는데, 먼저 전처리 과정은 일반적인 텍스트 데이터들을 컴퓨터가 처리하기 쉽도록 변화하는 작업으로써, 특정단어와 관련된 문서들을 신속하게 검색할 수 있도록 인덱스 파일을 만드는 것이라고 설명하고 있다. 그리고 인덱스를 만드는 방법으로 FB(Frequency-Based), IDF(Inverse Document Frequency), LSI(Latent Semantic Indexing) 등의 대표적인 방법을 열거하였다. FB는 문서 안에서 빈번히 나타나는 단어들을 그 문서를 대표하는 중요한 단어로 파악하고 가중치를 높게 주는 개념이며, IDF는 특정문서에서 중요한 단어가 무엇인지 뿐만 아니라 다른 문서와 구분을 해주는 단어가 무엇인지에 대한 정보를 포함하기 위한 계산을 한다는 것이다. LSI는 문서들이 공유(co-occurrence)하는 단어들을 파악하여 동일한 주제나 개념으로 인식함으로써 검색단어와 정확하게 일치하지 않더라도 개념이나 주제에 의하여 문서검색이 가능할 수 있게 한다는 것이다. 다음으로 텍스트 분석 과정은 전처리 과정을 거친 데이터들을 대상으로 정보 추출(information extraction), 범주화(categorization), 문서요약(summarization) 과 같은 다양한 분석을 실시하는 것으로 설명하고 있다. 정보 추출은 특정 문서 안에서 유용한 정보 즉, 사람이름, 장소이름, 전화번호, 날짜, 화폐단위 등 문서내의 개체들(entities) 및 이들 사이의 연관성을 식별하여 검색하는 기술이라고 설명한다. 범주화는 수집된 문서들 중에서 유사한 내용의 문서들을 그룹화해서 분류하는 기술로써, 비구조적으로 모여있는 문서들을 구조적으로 조직화하는 과정이라고 정의하였다.

양동현·안준모·함유근·민형진(2014)은 비정형 데이터 등과 같은 고객의 빅데이터 분석을 통해 기업 성과를 향상시킬 수 있는 다양한 접근이 필요함을 시사하였다. 또한 최근 업

계에서는 빅데이터 분석에 다양한 관심과 더불어 분석 솔루션이 등장하고 있음을 논하였으며 이를 활용하는 방안을 연구하였다.

국내 연구 중 유사영역에 텍스트마이닝을 활용한 방법으로서, 김유신·김남규·정승렬(2012)과 유은지·김유신·김남규·정승렬(2013)의 연구가 있다. 이들은 텍스트마이닝 방법론을 활용하여 기업의 주가를 예측하는 연구가 수행하였다. 주가 예측을 위하여 수집된 뉴스 콘텐츠를 활용하여 각 뉴스 내의 의미 있는 어휘 또는 말뭉치(corpus)를 추출하고, 이를 분석하여 해당 뉴스가 주가에 호재인지 악재인지 분류한 후 그 결과를 이용하여 주가에 대한 움직임을 예측하는 방법을 적용하였다.

안성원·조성배(2010)와 김민수·구평회(2013) 또한 텍스트마이닝 분석 기법을 활용하여 주가 예측 모형을 추정하였다. 특히, 김민수·구평회(2013)는 텍스트마이닝 실험 과정에서 정보획득(Information Gain)을 기반으로 한 변수 선택을 시도하였다. 즉, 주가의 등락에 영향을 미치는 변수를 선택하는 과정에서 각각의 변수마다 정보획득 수준을 계산하여 이것이 높은 순으로 정렬하여 선택하는 방식이다.

김승우·김남규(2008)는 텍스트마이닝과 오피니언 마이닝의 특징을 비교하면서 분석 목적이 의미있는 지식을 창출하는 것이면 텍스트마이닝, 텍스트 문서가 내포하고 있는 긍정/부정의 감성을 분류하는 것이면 오피니언 마이닝이라고 별도로 정의하였으며 텍스트마이닝과 차이점을 설명하였다.

기업의 부도 혹은 부실 징후 예측은 매우 많은 연구가 수행되어 왔다. 김석태(1999)는 중소기업 부도예측에 어떠한 변수가 중요한지 연구하였으며, 신동령(2006)과 도영호·김경숙·장영민(2012)은 보다 정확한 중소기업 부도예측을 위하여 생산성지표, 재무지표 등을 활용하여 기업의 부실을 사전적으로 예측하기 위한 연구를 수행하였다.

2. 분석도구

본 연구에서는 텍스트마이닝을 위하여 Open Source인 R(3.0.1 version)을 기반으로 주요 텍스트 분석 기능을 컴퍼넌트화하여 구현한 패키지(Package) 도구를 사용하였다. 해당

패키지는 텍스트 자연어 처리를 위한 형태소 분석기(NLP MoranAn2013-K)을 사용할 수 있도록 설계 되었으며, 분석 결과의 표출을 위한 여러 기능을 제공한다.

Ⅱ. 연구방법론

본 연구는 뉴스 텍스트를 기반으로 기업부도예측을 수행하기 위하여 크게 빈도 분석(frequency), 연관어 분석(concept link), 의사결정나무(decision Tree) 총 3 가지 분석 방법론을 활용 하였다.

〔표 1〕 주요 연구 방법론 요약

기능	설명
빈도분석	• 전체 문서에서 각 단어별 발생빈도를 분석 • 발생빈도가 높은 주요 단어를 표와 그림으로 표출 • 주요 이슈 키워드의 변화 추이 분석 가능
연관어 분석	• 주요 키워드의 연관 단어 분석 및 시각화 • 특정 키워드와 밀접한 단어들을 확인하여 현재의 트랜드를 확인
의사결정나무	• 의사결정규칙(Decision Rule)을 나무구조로 도표화하여 분류와 예측을 수행하는 분석 • 다른 데이터마이닝 방법(신경망, 판별분석, 회귀분석 등)에 비하여 연구자가 그 과정을 쉽게 이해하고 설명할 수 있음

1. 빈도분석 방법론

텍스트마이닝을 위한 정보추출 방법에는 다양한 목적, 조건, 환경 등으로 정보의 추출방법이 다양하며, 정보추출방법은 텍스트마이닝에서 가장 중요한 부분 중에 하나이다. 특히, 정보추출 방법에는 수많은 수학적 알고리즘과 방법들이 존재하며, 그 중 간단하면서 가장 강력한 방법으로는 TF‐IDF(TF: Term Frequency‐IDF: Inverse Document Frequency) 방식을 많이 사용하고 있다. 각 용어의 정의는 다음과 같다.

- TF(단어 빈도수, term frequency)는 특정한 단어가 어떤 범위내의 문서에서 얼마나 자주 등장하는지를 나타내는 단어빈도수 값으로, 특정단어의 빈도수가 높으면 문서 내에서 해당 단어가 중요하다고 생각할 수 있다
- DF(문서 빈도수, document frequency)는 특정한 단어가 일정한 범위의 문서들간의 자주 사용되는 지수를 나타낸다. 일반적으로 특정한 단어가 여러 문서에서 빈도수가 높은 경우, 그 단어가 흔한 단어로써 보편적인 단어라는 의미를 나타낸다.
- IDF(역 문서 빈도수, inverse document frequency)는 DF의 역수를 뜻한다. 특히, 문서간의 역수를 취함으로써 DF가 커질수록 IDF는 감소하고 IDF를 작을수록 IDF를 증가한다.
- TF-IDF (중요도 지수)는 TF와 IDF를 곱한 값이다. 이를 통해 문서의 중요도를 생성할 수 있다.

$$[TF-IDF\text{지수}] = TF \times \frac{1}{DF} \tag{1}$$

TF-IDF 방법에 대해 좀 더 자세히 살펴보면 텍스트마이닝에서 이용하는 키워드의 가중치를 구하는 방법으로서, 여러 문서로 이루어진 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다. 문서의 핵심어를 추출하거나, 검색 엔진에서 검색 결과의 순위를 결정하거나, 문서들 사이의 비슷한 정도를 구하는 등의 용도로 사용할 수 있다.

예를 들면, 특정 범위 내에서 모든 단어들의 빈도수를 산출하고, 각 문서마다의 단어들의 빈도수를 구한 후 역수를 취해 곱하며 문서의 중요도를 산출할 수 있다. 즉, 특정 단어의 순위 결정은 질의어와 질의어가 포함된 웹 문서들 간의 가중치 및 유사도를 계산하여 높은 값을 가진 가중치를 정렬함으로써 순위가 결정되는 방식을 사용하고 있다. 이러한 단어의 중요도를 응용하여 특정단어에 대한 중요도를 추출하는 정보검색에서도 TF-IDF 방법은 가장 기본적인 알고리즘으로 널리 사용하고 있다. 또한, 대부분의 검색엔진은 이 알고리즘과 다른 알고리즘(벡터방식 등)을 혼합하여 사용한다.

그러나, TF-IDF 모델은 문서 내에서 정보검색, 색인과 유사도를 계산하고 검색 순위를 결정할 수 있는 기능을 가지고 있는 반면, 문서 내 단어 개수와 문서가 많아지고 문서 내에 포함된 단어의 수가 많을수록 검색의 효율성은 떨어지고 방대한 계산으로 너무 많은 시간이 소요되는 단점이 발생하고 있다. 또한, 많은 내용으로 인해 영어단어의 키워드가 정확히 매치되지 못하는 단점이 있다. 특히 어근에 접사가 결합되어 문장 내에서의 각 단어의 기능을 나타내는 한글의 특성상 형태소를 통해 단어를 추출해야 되므로 영어보다 분석에 더 많은 시간과 노력이 소요된다.

2. 연관어 분석

연관어 분석은 연관성 분석이라는 통계기법에 그 기반을 두고 있는 데, 연관성 분석이란 상품 혹은 서비스간의 관계를 살펴보고 이로부터 유용한 규칙을 찾아내고자 할 때 이용할 수 있는 기법이다(곽청이·함유근·이미영(2014)). 즉, 구입항목의 집합에서 하나의 구입상품 또는 구입상품들 집합의 존재가 또 다른 구입상품의 존재를 암시하는 규칙을 발견하는 기법이다. 연관 규칙(association Rule) 또는 장바구니 분석(market basket analysis)이라고도 한다.

연관성 분석은 그 분석결과를 교차 판매, 판촉전략 수립, 그리고 매장 배치 등의 다양한 마케팅 전략에 활용될 수 있는 잠재적 유용성에도 불구하고 수익을 증진시킨 실제 사례는 많지 않으며 그 이유는 분석의 결과로 제시되는 연관 규칙들의 수가 너무 많다는 것에서 찾을 수 있다고 선행연구에서 지적하고 있다. 따라서, 이러한 한계를 극복하기 위하여 방대한 연관규칙들 중 의미 있는 규칙들만을 식별하는 과정이 필요하고 이를 지원하기 위하여 다양한 흥미성 척도(interesting-ness measures) 가 고안되어 왔는데 이들 척도들은 기본적으로 발생 빈도(frequency)에 근거하여 도출된다. 연관 규칙을 발견하기 위한 알고리즘으로는 Apriori, DHP, EP-growth 등이 연구되었다(김성학·안병태(2008)).

연관성을 객관적으로 판단하기 위한 지표로서 데이터마이닝에서 가장 많이 사용되는 지표로서 지지도(support), 신뢰도(confidence), 향상도(lift)를 활용한다.

- 지지도(support): 생성된 연관규칙이 전체 항목에서 차지하는 비율을 말한다. 즉, 데이터베이스에 속한 전체 거래의 개수 중 그 연관규칙을 지지하는 거래의 개수 비율을 의미하며, 전체 거래 중 X와 Y를 포함하는 거래의 정도를 나타내는 식으로 표현된다.

$$Surpport(X, Y) = \frac{P(X \cap Y)}{total} \tag{2}$$

- 신뢰도(confidence): 연관규칙의 강도를 의미하며, 전제부를 만족하는 트랜잭션이 결론부까지를 만족하는 비율을 말한다. 즉, X를 포함하는 거래 중에서 Y가 포함된 거래의 정도를 의미한다.

$$Confidence(X, Y) = P(X \mid Y) = \frac{P(X \cap Y)}{P(X)} \tag{3}$$

- 향상도(lift): 두 거래품목 간의 연관성(독립성)을 측정하는 지표이다. 두 변수가 완전한 독립이라면 1의 향상도를 나타내다. 향상도가 1 보다 크면 두 거래 품목은 양의 상관관계를, 1보다 작으면 음의 상관관계를 의미한다.

$$Lift(X, Y) = \frac{P(X \cap Y)}{P(X)P(Y)} \tag{4}$$

본 연구에서는 향상도를 활용하여 부도 기업의 뉴스와 각각의 키워드 간의 연관성이 존재하는 지를 검증하였다. 또한 통계패키지의 연관어 분석 기능을 활용하여 빈도 분석과 높은 향상도를 나타내는 키워드를 기준으로 연관단어를 상관계수로 표현하여 연관성을 분석하였다.

3. 의사결정나무

의사결정나무(decision Tree)는 의사결정규칙(decision rule)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석방법이다. 분석과정이 나무구조에 의해서 표현되기 때문에 판별 분석(discriminant analysis), 회귀분석(regression analysis), 신경망(neural networks) 등과 같은 방법들에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다.

의사결정나무는 분류 또는 예측을 목적으로 하는 어떤 경우에도 사용될 수 있으나 분석의 정확도보다는 분석과정의 설명이 필요한 경우에 더 유용하게 사용된다. 의사결정나무 분석이 활용된 분야는 의료분야나 금융분야 뿐 만 아니라 스포츠 분야(곽청이·함유근·이미영(2014))까지 다양하며 이 활용될 수 있는 응용분야는 다음과 같다.

- 세분화(segmentation) : 관측개체를 비슷한 특성을 갖는 몇 개의 그룹으로 분할하여 각 그룹별 특성을 발견하고자 하는 경우
- 분류(classification) : 여러 예측변수(predicated variable)에 근거하여 목표 변수(target variable)의 범주를 몇 개의 등급으로 분류하고자 하는 경우
- 예측(prediction) : 자료로부터 규칙을 찾아내고 이를 이용하여 미래의 사건을 예측하고자 하는 경우
- 차원축소 및 변수선택(data reduction and variable screening) : 매우 많은 수의 예측변수 중에서 목표변수에 큰 영향을 미치는 변수들을 골라내고자 하는 경우
- 교호작용효과의 파악(interaction effect identification) : 여러 개의 예측변수들이 결합하여 목표변수에 작용하는 교호작용을 파악하고자 하는 경우
- 범주의 병합 또는 연속형 변수의 이산화(category merging and discretizing continuous variable) : 범주형 목표변수의 범주를 소수의 몇 개로 병합하거나, 연속형 목표변수를 몇 개의 등급으로 범주화 하고자 하는 경우

일반적으로 의사결정나무 추정은 다음과 같은 분석과정을 거친다.

- 의사결정나무의 형성: 분석의 목적과 자료구조에 따라서 적절한 분리 기준(split criterion)과 정지규칙(stopping rule)을 지정하여 의사결정나무를 얻는다.
- 가지치기: 분류오류(classification error)를 크게 할 위험(risk)이 높거나 부적절한 규칙을 가지고 있는 가지(branch)를 제거한다.
- 타당성 평가: 이익도표(gains chart)나 위험도표(risk chart) 또는 검정용 자료(test data)에 의한 교차타당성(cross validation) 등을 이용하여 의사결정나무를 평가한다.
- 해석 및 예측: 의사결정나무를 해석하고 분류 및 예측모형을 설정한다. 이상과 같은 과정에서 정지기준, 분리기준, 평가기준 등을 어떻게 지정하느냐에 따라서 서로 다른 의사결정나무가 형성된다.

Ⅲ. 실증분석

1. 부도의 정의

부도 기업 예측연구에서 유용한 결과를 얻기 위해서 기업의 부도(도산)에 대한 명확한 정의를 하는 것이 중요하다. 증권거래소의 '상장규정' 에 부도기업이 정의되고 있긴 하지만 실제 부도에 관하여 명확한 정의를 내리는 것은 분석하는 목적과 연구자에 따라 기준이 다를 수 있다. 연구 목적에 따라 부도를 부분적인 채무불이행 위험을 포함하는 신용위험으로 광의의 개념으로 확장하여 인식하는 경우도 있으므로 부도예측에 관한 연구를 위해서는 이에 대한 범위 설정을 보다 객관적이고 구체적으로 정의할 필요가 있다.

본 연구는 한국 유가증권 시장에서 상장폐지가 결정된 기업들 중, 부도발생, 화의절차개시신청, 회사정리절차개시신청 및 은행거래정지와 같은 부도에 관련된 공시가 발생한 기업들을 부도 발생 기업을 인식하고 분석을 진행하였다. 상장폐지 이벤트(Event)는 부도와

반드시 연결된다고 볼 수는 없으나, 일반적으로 금융 시장에서 특수한 상황을 제외하고 상기와 같은 공시가 발생하여 상장폐지 되는 기업은 부도 수준으로 간주할 수 있다. 또한 부도가 발생하지 않더라도 상장폐지 이벤트는 투자자와 채권자는 큰 손실을 입을 수 있는 사건이므로 상장폐지를 부도로 인식하는 것은 보다 보수적인 기준에서 부도를 평가하는 방법이라고 할 수 있다.[1)]

2. 뉴스 텍스트 Data 수집

뉴스 콘텐츠 수집을 위하여, 주요 포탈의 뉴스검색을 활용하여 2008년 이후 상장폐지된 177개 상장기업에 대한 상장폐지 시점 직전 6개월간의 뉴스 콘텐츠를 수집하여 텍스트 DB(database, 이하 DB)를 구축하였다. 또한 부도기업과 정상기업 간의 차이를 분석하기 위하여 각 부도기업에 대응하여 공통된 산업에 속한 정상기업 160개[2)]를 매칭(matching)하여 같은 방법으로 뉴스 콘텐츠를 수집하여 비교 기업 DB를 구축하였다. 정상기업의 경우 뉴스 콘텐츠가 수작업으로 진행할 수 없을 정도로 많아, 기업 당 약 6~10개 정도씩 매월 1개의 뉴스를 기준으로 주요 뉴스 콘텐츠를 수집하였다. 정상 기업 뉴스는 부도 기업의 뉴스와 비교를 위하여 부정적 뉴스를 가능한 선택하는 방향으로 데이터를 확보하였다.

〔표 2〕 수집된 뉴스 콘텐츠 기초통계량

구분	기업 수	전체 기사 수	기업당 기사 수
부도기업	177 개	951 건	5.37건
정상기업	160 개	734 건	4.59건

분석대상 데이터 품질 향상을 위하여 뉴스 콘텐츠 수집 시 대내외적인 신뢰성이 인정되는 뉴스 매체를 대상으로만 뉴스 텍스트를 수집하였다.

---

1) 이후 연구에서 부도예측 대상을 비상장기업까지 확대한다면 상장폐지 외에 별도의 부도 정의가 필요로 할 것이다.

2) 뉴스 콘텐츠 수집은 검색포탈의 뉴스 검색을 통하여 실행하였음. 17개 정상업체는 뉴스가 없어 사후적으로 제외함

〔표 3〕 뉴스 콘텐츠 수집 대상 언론 매체

구분	매체 명
종합	경향신문, 국민일보, 뉴시스, 동아일보, 로이터, 문화일보, 서울신문, 세계일보, 연합뉴스, 조선일보, 중앙일보, 채널A, 한겨레, 한국일보, JTBC, KBS, MBC, SBS, YTN
경제	뉴스토마토, 매일경제, 머니투데이, 서울경제, 아시아경제, 이데일리, 조선비즈, 파이낸셜뉴스, 한국경제, 한국경제TV, 헤럴드경제, MBN, SBSCNBC
온라인/인터넷	데일리안, 오마이뉴스, 쿠키뉴스

3. 텍스트 데이터 분석 전처리

(1) 키워드 동의어 처리

뉴스 텍스트의 키워드는 문장 속에 포함되어 활용되기 때문에, 동일한 의미의 단어라도 다양한 형태나 다른 표현으로 사용될 수 있다. 따라서 텍스트마이닝을 위해서는 이러한 동일한 의미, 표현을 뜻하는 키워드에 대한 동의어 처리가 필요하다.

〔표 4〕 키워드 동의어 처리대상

동의어	처리대상 키워드
상장폐지	상폐, 상장폐지절차, 상장폐지사유, 상장폐지결정, 상장폐지실질심사, 상장폐지효력정지, 상장폐지가처분신청, 상장폐지심사, 상장폐지우려, 상폐가능성, 상폐기준, 상폐사유, 상폐여부, 상폐우려
지분	지분율, 보유지분, 지분인수, 지분투자, 지분변동, 지분매입
횡령	횡령배임, 횡령혐의, 배임횡령, 횡령배임설, 횡령사실, 횡령사건
자금	자금난, 회사자금, 자금지원, 신규자금, 인수자금, 자금거래, 자금압박, 시설자금, 자금력, 자금줄, 긴급자금, 외부자금, 자금관리, 자금사정, 긴급운영자금, 보유자금, 자금경색,자금유치, 자금추적, 자금확보
증자	유상증자, 유증, 일반공모유상증자, 3자배정유상증자, 유상증자대금, 주주배정유상증자, 대규모증자, 유상증자계획
법원	재판부, 서울중앙지법, 서울중앙지방법원, 수원지방법원, 서울남부지방법원, 대법원, 대구지방법원, 부산지방법원, 서울지방법원, 서울고등법원, 등

동의어	처리대상 키워드
검찰	서울중앙지검, 검찰조사, 대검찰청중앙수사부, 서울중앙지방검찰청, 검찰고발, 검찰청, 서울지방검찰청, 검찰통보, 대검찰청, 검찰수사
채권단	채권단의, 주채권은행인, 채권은행
부도	부도설, 부도처리, 최종부도, 부도나다, 부도위기, 1차부도, 부도상태, 부도금액, 부도기업, 부도내용, 부도발생은행, 부도어음
손실	영업손실, 당기순손실, 손해, 순손실, 투자손실, 손실액
파산	파산신청, 파산절차, 파산관재인, 파산선고
감사	감사인, 재감사, 회계감사, 감사업무
회생	회생절차, 회생절차개시, 회생계획안, 회생계획, 기업회생, 회생채권자, 회생담보권, 회생채권, 회생절차폐지, 일반회생, 회생담보권자, 회생신청, 회생회사, 회생가치, 회생의지, 회생절차개시결정, 기업회생계획, 회생작업
BW	신주인수권부사채
대출	대출금, 빚, 부실대출, 신규대출, 부정대출, 불법대출, 사기대출, 신용대출
조달	자금조달
금융감독원	금감원, 금융당국, 금융위
잠식	자본잠식, 자본잠식률, 자본전액잠식, 전액잠식, 완전자본잠식, 자본잠식비율, 자본잠식상태, 자본잠식액, 자본잠식탈피, 자본전액잠식설
배임	배임혐의, 배임행위, 배임수재, 배임사실, 배임발생, 배임사건
자본	자본금, 자기자본, 자본총계
주주총회	주총, 임시주주총회, 임시주총

(2) 불용어 처리

텍스트마이닝의 분석 효율성을 증대하기 위하여 다음과 같은 키워드는 분석대상에서 제외하였다.

- 검색 빈도 수 10개 이하로 산출되는 키워드제외(동의어 처리대상은 분석 포함)
- 동사, 조사, 형용사 어미 형태 모두 제외
- 두 단어 이상 명사는 구성 명사 중 내용상 필수 명사 1단어로 변환하여 활용
- 기업 활동과 상관없는 의미 명사 제외 - 특정회사/제품/인물 이름

4. 부도 연관 키워드 분석

부도기업의 뉴스 Text DB를 활용하여 기본적인 빈도분석(TF, Term Frequency) 을 수행하였다. 이를 통하여 부도와 연관되는 주요 뉴스 키워드를 도출할 수 있다.

(1) 키워드 빈도분석

빈도수: 뉴스 텍스트 내 해당 키워드의 발생 빈도 수 (전처리 이후 데이터 기준)
비중: 전체 키워드 발생 빈도 중 해당 키워드의 비중

〔표 5〕 부도기업과 정상기업 간의 키워드 빈도 비중 비교

키워드	부도기업		정상기업		Gap
	빈도수	비중	빈도수	비중	
상장폐지	616	2.89%	63	0.42%	2.47%
회생	410	1.92%	38	0.25%	1.67%
공시	393	1.84%	105	0.69%	1.15%
법원	344	1.61%	67	0.44%	1.17%
자금	332	1.56%	70	0.46%	1.10%
매출	327	1.53%	658	4.33%	-2.80%
인수	315	1.48%	209	1.38%	0.10%
결정	302	1.42%	120	0.79%	0.63%
횡령	287	1.35%	40	0.26%	1.08%
지분	273	1.28%	248	1.63%	-0.35%
매각	271	1.27%	233	1.53%	-0.26%
검찰	243	1.14%	37	0.24%	0.90%
증자	242	1.14%	89	0.59%	0.55%
손실	241	1.13%	168	1.11%	0.02%
회계법인	228	1.07%	4	0.03%	1.04%
채권단	215	1.01%	54	0.36%	0.65%
발생	213	1.00%	139	0.92%	0.08%
보유	203	0.95%	176	1.16%	-0.21%
부도	196	0.92%	3	0.02%	0.90%
협의	186	0.87%	36	0.24%	0.64%

키워드	부도기업		정상기업		Gap
	빈도수	비중	빈도수	비중	
최대주주	183	0.86%	118	0.78%	0.08%
추진	180	0.84%	159	1.05%	-0.20%
체결	172	0.81%	122	0.80%	0.00%
감사	165	0.77%	16	0.11%	0.67%
문제	151	0.71%	82	0.54%	0.17%
개발	147	0.69%	286	1.88%	-1.19%
주가	146	0.68%	151	0.99%	-0.31%
법정관리	145	0.68%	38	0.25%	0.43%
확보	140	0.66%	197	1.30%	-0.64%
감사보고서	140	0.66%	9	0.06%	0.60%
시작	136	0.64%	172	1.13%	-0.50%
이상	135	0.63%	192	1.26%	-0.63%
파산	133	0.62%	5	0.03%	0.59%
워크아웃	127	0.60%	29	0.19%	0.40%
가능성	127	0.60%	95	0.63%	-0.03%
투자	124	0.58%	176	1.16%	-0.58%
잠식	123	0.58%	35	0.23%	0.35%
상장	121	0.57%	52	0.34%	0.23%
금융당국	121	0.57%	59	0.39%	0.18%
선정	120	0.56%	106	0.70%	-0.14%
추가	119	0.56%	141	0.93%	-0.37%
BW	119	0.56%	45	0.30%	0.26%
투자자	118	0.55%	9	0.06%	0.49%
발행	118	0.55%	64	0.42%	0.13%
경영권	110	0.52%	62	0.41%	0.11%
제기	104	0.49%	76	0.50%	-0.01%
주주총회	103	0.48%	80	0.53%	-0.04%
자본	100	0.47%	63	0.42%	0.05%
설립	100	0.47%	146	0.96%	-0.49%
퇴출	95	0.45%	10	0.07%	0.38%
…	…	…	…	…	…
Total	21,315	100%	15,180	100%	

빈도분석 결과 키워드 도출 빈도 비중 갭(gap)을 기준으로 양(+)의 숫자가 크게 나타나는 경우 부도와 연관성이 높은 키워드라 할 수 있고 음(-)의 숫자가 크게 나타날수록 부도와 연관성이 떨어지는 키워드라 할 수 있다.

부도기업의 뉴스에서는 정상기업의 뉴스보다 '상장폐지', '회생', '공시', '법원', '자금' 등과 같은 키워드의 빈도 비중이 높게 도출되었다. 반면에 '매출', '지분', '매각', '보유' 등과 같은 단어는 정상 기업의 뉴스에서 빈도 비중이 더 높게 나타난다. '인수', '손실', '발생' 등의 키워드는 부도기업과 정상기업 간의 발생빈도 비중이 큰 차이가 없었다.

(2) 뉴스 시점별 키워드 빈도 분석

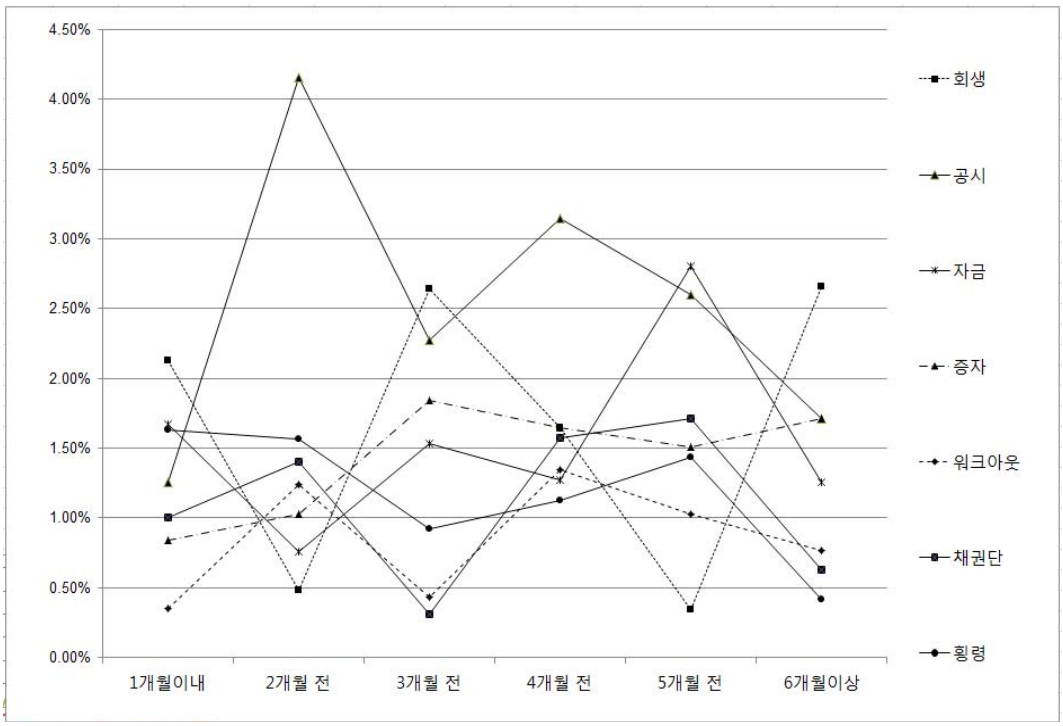
부도예측을 위해서는 부도시점보다 몇 개월 이전에 어떠한 뉴스 텍스트가 나타나는지가 중요하다. 따라 시점 별로 나누어 키워드 빈도 비중을 분석하여 보았다.

〔표 6〕 뉴스 보도 시점에 따른 키워드 빈도 비중 분석

키워드	뉴스 보도 시점 (부도발생 기준)					
	1개월이내	2개월 전	3개월 전	4개월 전	5개월 전	6개월이상
상장폐지	3.56%	5.62%	0.86%	0.42%	**1.57%**	1.05%
회생	2.13%	0.49%	**2.64%**	**1.65%**	0.34%	**2.66%**
공시	1.26%	4.16%	**2.27%**	**3.15%**	**2.60%**	**1.71%**
법원	2.10%	0.97%	**1.35%**	**1.05%**	0.21%	**1.22%**
자금	1.67%	0.76%	**1.54%**	**1.27%**	**2.81%**	**1.26%**
횡령	1.63%	1.57%	0.92%	**1.12%**	**1.44%**	0.42%
검찰	1.54%	0.81%	0.25%	**1.05%**	**1.10%**	0.35%
증자	0.84%	1.03%	**1.84%**	**1.65%**	**1.51%**	**1.71%**
회계법인	1.53%	1.13%	0.37%	0.22%	0.27%	0.42%
채권단	1.01%	1.40%	0.31%	**1.57%**	**1.71%**	0.63%
부도	1.16%	0.32%	0.25%	0.52%	0.00%	**1.43%**
혐의	1.22%	0.54%	0.25%	0.45%	0.68%	0.35%
감사	0.99%	0.92%	0.68%	0.60%	0.27%	0.24%
주가	0.76%	0.81%	0.37%	0.82%	0.00%	0.70%
법정관리	0.86%	0.38%	0.31%	0.30%	0.27%	0.77%
확보	0.61%	0.65%	0.86%	0.67%	0.75%	0.73%
파산	0.93%	0.38%	0.31%	0.00%	0.07%	0.31%
워크아웃	0.35%	1.24%	0.43%	**1.35%**	**1.03%**	0.77%
잠식	0.53%	1.46%	0.55%	0.30%	0.14%	0.63%

‘법원’, ‘검찰’, ‘혐의’, ‘회계법인’ 과 같은 키워드는 상대적으로 부도 시점이 임박해서야 빈도수가 많이 관찰되는 것을 확인할 수 있다. 또한 ‘상장폐지’와 ‘부도’, ‘파산’ 등의 경우 직접적으로 부도 발생의 결과를 보도하는 뉴스일 가능성이 많으므로 부도예측력 분석에 활용하는 것은 적합하지 않다고 판단하여 분석대상 키워드에서 제외하였다.

본 연구는 사전적인 부도예측의 정확성 향상을 목적으로 한다. 따라서 부도 이전 2~6개월 전 시점의 빈도 비중이 높은 ‘회생’, ‘공시’, ‘자금’, ‘횡령’, ‘증자’, ‘채권단’, ‘워크아웃’ 7 개 키워드를 주요 부도 키워드로 선정하고 중점적으로 이후 분석과정에 활용하였다.



〔그림 1〕 뉴스 보도 시점별 주요 키워드 비중 변화 추이

5. 부도 키워드 연관성 분석

빈도 분석결과 부도 기업 뉴스에서 발생빈도가 높은 키워드를 도출할 수 있었다. 본 장에서는 부도 기업 뉴스에서 발생빈도가 높은 키워드에 대하여 연관성(association) 분석 방법론 중 Lift 지수(향상도)를 측정하여 실제 부도와의 연관성을 계량적으로 분석하여 보았다.

$$Lift(\text{부도기업여부}, \text{키워드 포함여부}) = \frac{P(\text{부도} \cap \text{키워드포함})}{P(\text{부도})P(\text{키워드포함})} \qquad (5)$$

〔표 7〕 부도 사건과 주요 키워드 간의 연관성 분석 결과

키워드	Lift 값	키워드	Lift 값
부도	1.71	주주총회	1.14
회계법인	1.70	발행	1.14
퇴출	1.65	매각	1.12
상장폐지	1.64	제기	1.11
파산	1.64	손실	1.11
감사보고서	1.61	발생	1.10
횡령	**1.57**	최대주주	1.08
감사	1.55	가능성	1.07
검찰	1.54	체결	1.05
회생	**1.51**	경영권	1.04
혐의	1.51	금융당국	1.01
상장	1.49	추진	1.01
법원	1.49	문제	0.99
법정관리	1.45	선정	0.93
잠식	1.41	지분	0.93
공시	**1.38**	보유	0.91
채권단	**1.31**	추가	0.89
워크아웃	**1.31**	주가	0.89
결정	1.27	투자	0.78
증자	**1.27**	시작	0.78
BW	1.26	확보	0.75
투자자	1.23	설립	0.74
자본	1.20	이상	0.71
자금	**1.15**	매출	0.69
인수	1.15	개발	0.62

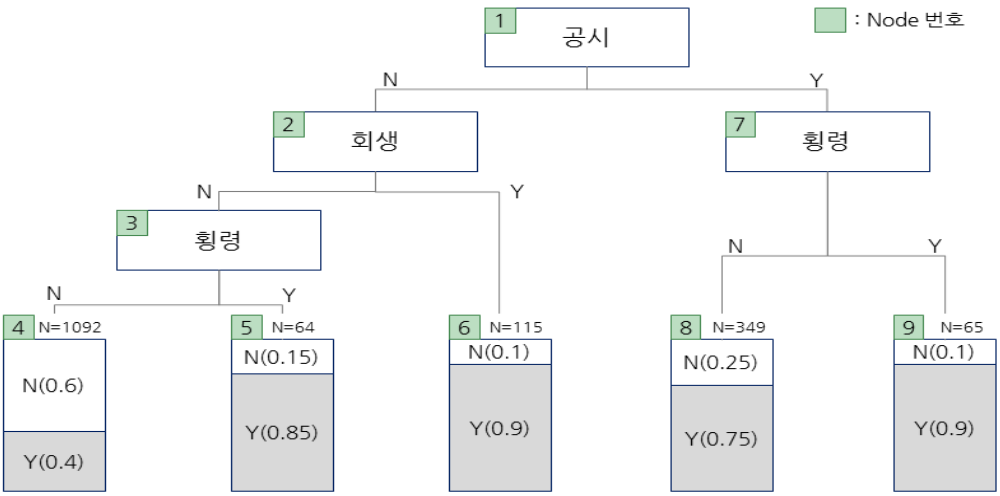
‘회생’, ‘공시’, ‘자금’, ‘횡령’, ‘증자’, ‘채권단’, ‘워크아웃’ 7 개 주요 부도 키워드는 ‘부도’, ‘퇴출’, ‘상장폐지’ 등 직접적으로 부도를 보도하는 뉴스에 비해서는 낮지만 1 이상의 향상도(Lift)가 산출됨으로서 연관성이 존재하는 것을 확인 할 수 있다.

6. 부도 키워드를 활용한 의사결정나무 도출

상기 분석 과정을 활용하여 도출된 부도 키워드를 활용하여 기업부도예측을 위한 의사결정나무(decision tree)를 도출하였다. 의사결정나무를 활용함으로서 뉴스에서 특정 키워드가 언급되었을 경우 부도 발생 가능성의 여부를 추정할 수 있다.

(1) 주요 부도 키워드 활용

우선 7개 주요 부도 키워드를 활용하여 의사결정 나무를 도출하여 보았다.



〔그림 2〕 주요 부도 키워드를 활용한 의사결정나무 도출 결과

‘공시’ 와 ‘회생’, ‘횡령’ 키워드를 기준으로 의사결정 나무가 산출 되었다.[3] 의사결정나무에 의한 분류 결과는 다음과 같다.

〔표 8〕 의사결정나무를 통한 예측력 검증 결과(1)

구분		실제		합계
		N (정상뉴스)	Y (부도뉴스)	
추정	N (정상뉴스)	621	471	1,092
	Y (부도뉴스)	113	480	593
합계		734	951	1,685

정분류율(부도&정상 판별): **65.3%** = (621+480)/1685

부도분류율(부도 판별): **50.5%** = 480/951

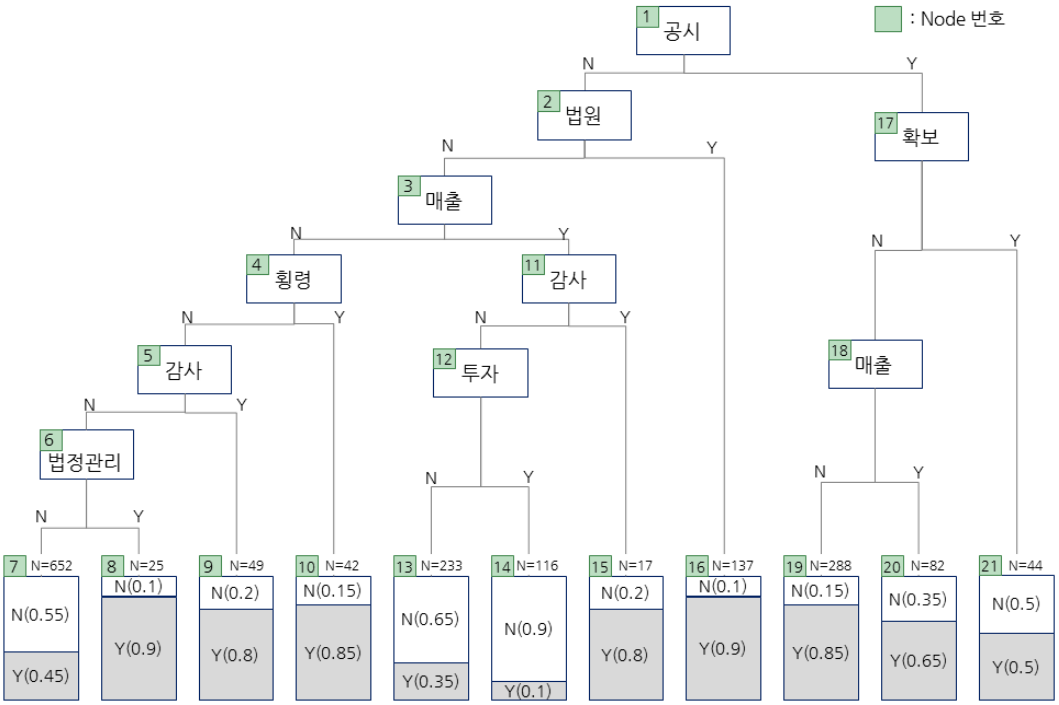
부도키워드 예측율(키워드 포함 시 부도): **80.9%** = 480/593

총 7개 주요 부도 키워드를 기반으로 추정한 의사결정나무 도출 결과, 부도 및 정상 분류를 정확하게 판별한 정분류율은 65.3%로 나타났다. 하지만 뉴스를 대상으로 부도를 예측하였을 때 실제 부도로 나타나는 뉴스로 판별한 부도 분류율은 50.5% 로 나타났다. 이러한 결과는 분류 모델로서 높은 예측력을 가진다고 볼 수 없으며, 특히 부도 예측 측면에서는 50% 내외로서 예측 모형으로서의 의미는 거의 없는 것으로 나타났다. 다만, 부도 키워드가 포함된 경우 해당 뉴스 대상 기업은 상당히 높은 비율로 실제 부도가 발생하는 것을 알 수 있다.

(2) 주요 부도 키워드 및 연관 키워드 활용

부도 키워드 및 빈도 비중 기준 상위 50개 키워드를 모두 포함하여 분석대상 키워드를 확장하여 의사결정 나무를 도출 하여 보았다. 산출결과 '공시', '법원', '매출', '횡령', '감사', '법정관리', '투자', '확보', '검찰' 키워드가 순차적인 기준으로 의사결정 나무가 산출되었다.

3) R Package 'Dtree' library 활용. 해당 함수는 판별력이 높은 키워드를 우선적으로 선택하여 의사결정나무를 산출하므로 선택되지 않은 키워드는 판별력이 낮은 키워드 임.(P-value기준)



〔그림 3〕 50개 주요 키워드를 활용한 의사결정나무 도출

의사결정 나무에 의한 예측력 검증 결과는 다음과 같다.

〔표 8〕 의사결정나무를 통한 예측력 검증 결과(2)

구분		실제		합계
		N (정상뉴스)	Y (부도뉴스)	
추정	N (정상뉴스)	625	417	1,042
	Y (부도뉴스)	109	534	643
합계		734	951	1,685

정분류율(부도&정상 판별): **68.8%** = (625+534)/1685

부도분류율(부도 판별): **56.2%** = 534/951

부도키워드 예측율(키워드 포함 시 부도): **83.1%** = 534/643

부도 키워드를 50개로 확장하여 분석한 결과 주요 키워드로 수행했던 결과보다 정분류율과 부도 분류율 모두 상승하였다. 하지만 부도 분류율 56.2% 수준은 부도 예측 모델로서 여전히 충분한 예측력을 가진다고 판단하기는 어렵다. 부도키워드 예측율 또한 83.1%로서 주요 키워드로 산출한 80.9% 수준에서 다소 상승하여 부도 키워드를 확장할 경우 판별력이 증가하는 것을 보여주고 있다.

이러한 결과가 발생한 원인은 판별하고자 하는 키워드가 뉴스 콘텐츠에 포함되지 않은 경우가 매우 많기 때문이다(전체 뉴스 & 기업 중 약 20% 이상). 즉, 키워드가 포함되지 않는 경우 '정상'으로 판별하기 때문에 부도 분류율이 낮아지는 현상이 발생하게 된다. 따라서 보다 많은 키워드를 도출하고 이를 활용하여 의사결정나무를 도출한다면 현재 보다는 훨씬 좋은 수준의 예측 결과를 도출 할 수 있을 것이라 기대할 수 있다. 현재 텍스트 DB 구축 범위를 확장하기 위해서는 다양한 Source와 품사(동사/형용사 등)의 텍스트 DB를 추가적으로 확보하여야 한다.

Ⅳ. 결론 및 제언

본 연구는 과거 부도가 발생한 기업의 뉴스 콘텐츠를 데이터로 확보하여 텍스트마이닝 기법을 통한 기업 부도예측의 가능성을 시도하여 보았다.

부도발생 기업의 뉴스에서는 빈도 분석 결과, 정상 기업의 뉴스보다 유의적으로 많이 나타나는 주요 키워드 들을 도출할 수 있었으며, 이중 '회생', '공시', '자금', '횡령', '증자', '채권단', '워크아웃' 과 같은 키워드는 부도 발생 2개월 이전에 선제적으로 많이 나타나는 키워드로 판명되었다. 이러한 키워드는 연관성 분석을 통하여 '상장폐지', '부도' 등과 같은 부도 이벤트를 나타내는 키워드와 함께 도출되는 것도 확인할 수 있었다.

또한 의사결정나무를 활용하여 상기 도출된 키워드를 활용하여 기업의 부도를 선제적으로 예측할 수 있는 가능성을 분석하여 보았다. 모형의 예측결과는 기대 수준에 비하여 낮게 산출되었는데, 가장 큰 원인은 분석대상 뉴스 콘텐츠의 부족이다. 연구과정에서 뉴스

콘텐츠 취합을 실행한 결과, 규모가 큰 기업을 제외하고 대부분의 규모가 작은 기업은 개별 기업에 대한 뉴스 콘텐츠를 충분히 확보하기가 어려웠다. 이러한 한계를 극복하기 위하여 향후 연구에서는 뉴스 콘텐츠뿐만 아니라 기업 공시 정보, 기업 투자관련 정보 공유 사이트 등 텍스트 DB를 추가적으로 확장할 수 있는 원천을 확보하여 이를 분석과정에 포함한다면 보다 우수한 모형 추정 결과를 얻을 수 있을 것이다.

예측수준 재고를 위한 또 하나의 방안은 방법론의 확장을 생각해 볼 수 있다. 본 연구에서는 예측기법으로 의사결정나무를 활용하였으나 회귀모형, 시계열모형 등 다양한 예측모형을 적용하기 위한 연구가 이루어진다면, 현재보다 높은 수준의 성과가 나타나는 예측 방법론이 될 수 있을 것이다.

본 연구의 부도예측 결과는 당장 기존의 방법론에 비하여 월등하게 좋은 결과를 산출하지 못하였다. 하지만 뉴스 텍스트를 포함 텍스트마이닝 분야는 현재도 계속 새로운 연구와 분석 기법이 활발히 연구되고 있는 분야이며, IT 기술의 성장에 따라 점점 보다 많은 텍스트 데이터를 보다 쉽게 획득할 수 있게 발전하고 있다. 또한 비상장기업의 경우 활용하지 못하고 있는 주가를 활용한 부도예측 방법론을 대체할 수 있는 적시성 있는 부도예측 방법으로서 충분한 잠재력을 확인할 수 있었다. 따라서 향후 현재 연구의 한계점을 보완한다면 기업 정보를 즉각적으로 반영할 수 있는 기업 부도예측 방법론 중 하나로서 충분히 범용적으로 활용될 수 있는 방법론으로 발전할 수 있을 것이라 기대한다.

참 고 문 헌

곽청이·함유근·이미영(2014), “인터넷마케팅에서 CRM을 통한 지불의사 상승효과에 관한 연구: 프로야구 산업을 중심으로,” Journal of Information Technology Application & Management, 제21권 1호, pp.17-34.

김근형·오성렬(2009), “온라인 고객리뷰 분석을 통한 시장세분화에 텍스트마이닝 기술을 적용하기 위한 방법론,” 「한국콘텐츠학회 논문지」, 제9권 제8호, pp.272-284.

김민수·구평회(2013), “인터넷 검색추세를 활용한 빅 데이터 기반의 주식투자전략에 대한 연구,”「한국경영과학학회지」, 제38권 제4호, pp.53-63.

김석태(1999), “중소기업 부도의 원인과 대책방안,”「한국생산성학회 생산성논집」, 13권 1호, pp. 273-295.

김성학·안병태(2008), “효율적인 클러스터링을 이용한 관심 정보 추출을 위한 웹 마이닝,”「Journal of Digital Society」, 9(2), pp.251-260.

김승우·김남규(2008), “오피니언 분류의 감성사전 활용효과에 대한 연구,”「한국지능정보시스템학회 학술대회 논문집」, 11, pp.121-128.

김유신·김남규·정승렬(2012), “뉴스와 주가 : 빅데이터 감성분석을 통한 지능형 투자의사결정모형,”「지능정보연구」, 제18권 제2호, pp.143-156.

도영호·김경숙·장영민(2012), “재무적 특성이 부도확률에 미치는 영향: 비상장 중소기업을 중심으로,”「한국생산성학회 생산성논집」, 26권 4호, pp. 127-153.

배상진·박철균(2003), “텍스트마이닝 기법의 기술정보분석 적용 가능성 연구,”「한국기술혁신학회 춘계학술대회」, pp.75-88.

신동령(2006), “기업부실예측에 있어 생산성지표의 유용성에 관한 연구,”「한국생산성학회 생산성논집」, 20권 2호, pp. 1-24.

안성원·조성배(2010), “뉴스 텍스트마이닝과 시계열 분석을 이용한 주가예측,”「한국컴퓨터종합학술대회 논문집」, 37 No.1(C), pp.364-369.

양동현·안준모·함유근·민형진(2014), “콜센터 고객정보시스템의 정보품질이 상담원 업무 성과에 미치는 영향에 관한 연구,”「한국IT서비스학회지」, 13(1), pp.87-101.

유은지·김유신·김남규·정승렬(2013), “주가지수 방향성 예측을 위한 주제지향 감성사전 구축 방안,”「지능정보연구」, 제19권 제1호, pp.95-110.

최윤정·박승수(2002), “웹 콘텐츠의 분류를 위한 텍스트마이닝과 데이터마이닝의 통합 방법 연구,”「인지과학」, 제13권 제3호, pp.33-46.

Lu, Y.C., C.H. Shen and Y.C. Wei(2013), “Revisiting early warning signals of corporate credit default using linguistic analysis,” *Pacific-Basin Finance Journal*, Vol. 24, pp.1-21.

Martinez, J. and R. Garcia, F. Sanchez(2012), "Semantic-Based Sentiment analysis in financial news," *Finance and Economics on the Semantic Web* 9, pp.38-51.

Mittermayer, M. and G. Knolmayer(2006), "Text mining system for market response to News : A Survey," *Institute of information systems Univ. of Bern : working paper No. 184*.

Olson, D.L., D. Delen and Y. Meng(2012), "Comparative analysis of data mining methods for bankruptcy prediction," *Decision Support System*, 52, pp.464-473.

The Prediction of Corporate Bankruptcy Using Text-mining Methodology

Jung-won Choi* · Ho-sun Han** · Mi-young Lee*** · Jun-mo Ahn****

Abstract

Traditional corporate bankruptcy prediction methodology basically relies on financial accounting data to objectively reflect the status of companies. However, since financial accounting data is difficult to immediately reflect changes in the status of companies, real-time financial data such as stock and bond prices are also used in order to make up for the shortcomings.

In this study, we use news text information which is a typical real-time information to study the corporate bankruptcy prediction models. In the past, news text information was difficult to use in quantitative analysis but not any more due to the recent advances of information processing technology and text-mining techniques.

For bankruptcy prediction using news information, we collect news text for six months before the bankruptcy events of companies actually occur and study the possibility of bankruptcy prediction based on the data by utilizing text-mining techniques.

Results indicate that we can not get such a high level of predictability as that of existing corporate bankruptcy prediction models, but that there exists a high potential of this approach enough to increase the predictability of bankruptcy models. Further research on bankruptcy prediction model using news text information will be promising.

Keyword : corporate bankruptcy prediction, text-mining

* Primary Author, Deloitte Anjin, Konkuk Univ. (jungwochoi@deloitte.com)
** Co-Author, Ucesspartners, Konkuk Univ. (hshan@ucesspartners.com)
*** Corresponding Author, Professor, Konkuk Univ. (yura@konkuk.ac.kr)
**** Co-Author, Professor, Konkuk Univ. (joonan@konkuk.ac.kr)