Journal of Knowledge Information Technology and Systems

ISSN 1975-7700

http://www.kkits.or.kr

A Study on Industry Information Analysis Methodology Based on Text Mining: PEST and Polarity Analysis Using Sentence Classification

Yoon-Sung Kim[1], Ho-Chang Lee[2], Seok Kee Lee[3], Do-Gil Lee[*4], Han-Gook Kim[5], You-Eil Kim[6]

[1,2]*Department of Computer Science & Engineering, Korea University*
[3]*College of Computer Engineering, Hansung University*
[4]*Research Institute of Korean Studies, Korea University*
[5]*UST S&T Information Science / Department of Industry Information Analysis, KISTI*
[6]*Department of Industry Information Analysis, KISTI*

ABSTRACT

Today's companies are in an environment where they have to survive in ever-increasing competition in the industry, by constantly identifying changes and trends in their industries and by periodically reflecting them in their policies and product development. For this purpose, one of the tasks that should be carried out is the analysis of industrial information. Most companies acquire industry analytical information at the cost of a large amount of time, manpower or, with the help of external professional analysts. However, since this conventional method is a somewhat heuristic and qualitative approach. The quality of these analysis results are different each time. A huge amount of industry related information is produced online in real time and when the information is reflected in the analysis as much as possible, it is required to introduce a new analytical method. In this paper, we propose a text mining methodology that extracts information from large amount of source data and automatically classifies it into each category of industry analysis framework. By constructing a sentence classifier using feature selection technique based on machine learning method, information that can be classified by indicators of universally used industry analysis framework is collected in sentence form. We performed PEST and polarity analysis by using our system and evaluated the classification accuracy of the proposed system through experiments.

KEYWORDS: Industrial analysis, PEST, SWOT, Polarity analysis, Text mining, Machine learning

ARTICLE INFO: Received 4 January 2017, Revised 10 February 2017, Accepted 10 February 2017.

[*]Corresponding author is with the Research Institute of Korean Studies, Korea University, Seoul, KOREA.

E-mail address: motdg@korea.ac.kr

1. 서 론

급변하는 기업 내·외적 환경 변화를 파악, 분석하고 이에 대응함으로써 자사의 기업 경쟁력을 계속 유지해 나가야 하는 것은 분야를 불문하고 현업에 종사하고 있는 모든 기업들이 필수적으로 수행해야 하는 과제 중의 하나이다. 현재 자신이 속해 있거나 향후 진출 분야로서 관심을 가지고 있는 산업 분야에 대한 정보 탐색 및 분석을 통해 획득한 정보는 최신 기술동향에 대한 대응, 신제품의 개발, 경쟁기업에 대한 대응전략의 수립 등에 활용됨으로써 궁극적으로 자사의 경쟁력 향상에 기여하기 위한 기본적인 정보로 활용된다. 대부분의 기업들은 이러한 정보의 획득을 위해 시장 분석 전문기관이 작성한 분석 정보를 구입하거나 혹은 자사 인력을 활용하여 직접 산업분석 정보를 획득하는 방법 등을 주로 활용하고 있다. 공통적으로 획득한 산업분석 정보는 전통적인 산업분석 프레임워크 (예: PEST[1], STEEP[1], SWOT[2], 5-Force Model[3] 등)의 형태로 표현된다. 하지만, 이러한 정보 획득의 두 가지 방식 모두 다음과 같은 측면에서 한계점을 지적할 수 있다.

첫째, 우선 비용적인 측면에서의 비효율성을 들 수 있다. 산업분석을 전문으로 하는 시장분석 전문기업으로부터 정보를 구매하는 경우 이들 기관이 보유한 다양한 이론 및 고유한 분석 방법론 및 사례를 바탕으로 한 비교적 고급의 정보를 획득하는 것이 가능하지만 아쉽게도 국내외 연구기관들이 제공하는 이러한 산업분석 서비스는 대부분 고가의 맞춤형 서비스로서 비용 측면에서 부담이 매우 크다. 내부인력이 직접 관련 정보를 수집, 가공하여 산업분석 결과를 파악하는 방법도 대기업을 제외한 대다수 국내 기업의 실정상 자사가 속한 산업의 분석을 전담할 전문 인력을 상시로 고용하고 이들 인력에 대한 비용을 부담하는 것은 현실적으로 매우 어렵다.

둘째, 다양한 형태와 원천을 가지는 데이터로부터 정성적이고 직관적인 분석을 통해 결과를 도출해야 하는 산업정보 분석 분야의 특성을 감안할 때, 내부 인력에 의한 정보 분석 과정은 표준화 내지 안정화 되어 있지 못하고 이에 따라 일정 품질 이상의 분석 결과 도출을 담보하기가 어렵다. 또한 높은 품질의 분석결과를 도출하기 위해서는 당연히 검토해야 하는 데이터의 양이 늘어나야 하고 이에 따라 분석 및 정리 작업에 소요되는 시간도 기하급수적으로 증가하는 문제가 발생한다.

셋째, 기업을 둘러싼 산업 환경의 변화 속도는 날이 갈수록 빨라지고 있으며 이러한 변화의 내용을 담고 있는 정보는 다양한 형태로 가공되어 하루에도 엄청난 양의 빅데이터로 계속 양산되고 있다. 위에서 언급한 두 가지 형태, 즉 기업들이 현재까지 주로 활용하고 있는 산업분석 정보 획득의 방법으로는 감당하기 힘든 수준의 데이터의 양(volume)이라고 할 수 있는데, 문제는 이들 데이터를 얼마나 더 신속하고 정확하게 처리하여 분석하느냐가 기업 경쟁력의 차별화에 큰 영향을 발휘한다.

이러한 현실적 문제점들을 감안할 때, 지금과는 다른 새로운 접근 방식의 산업정보 분석 방법론의 도입이 어느 때보다 필요한 시점이라고 할 수 있다. 문제를 해결하기 위해 첫 번째로 접근할 수 있는 대안은 분석과정에서 현재 양산되고 있는 빅데이터를 가능한 최대로 찾아내고 분석할 뿐만 아니라 이러한 일련의 과정을 보다 신속하게 진행할 수 있는 자동화 기법의 개발일 것이다. 이에 관련하여 빅데이터를 기반으로 한 방법론의 연구가 실험적으로 검토되거나 일부 기획된 사례는 이전 일부 연구에서도 진행된 바 있다[4-5]. 하지만 이들이 제시한 시스템은 분석 시스템에 대한 설명과 예시를 제시하였으나 정확한 성능 등의 객관적 평가가

없으며, 군집화한 키워드를 다시 사람이 분석하여 범주를 결정하는 등 사람의 개입이 필요한 반자동화 시스템이라는 한계점 등을 가지고 있다.

최근 빅 데이터가 널리 각광 받고 있고 이를 기반으로 유용한 정보를 추출할 수 있을 것이라는 당위론적인 논의는 많지만 데이터마이닝 기반의 실제적인 방법론의 설계 및 시스템 구축을 통해 그 결과를 비교하고 실제 산업 정보 분석에의 활용을 적극 모색해야 할 시점임을 감안하면 현재까지 산업정보를 내포한 문장들에 대한 자동 분류 시스템에 대한 관련 연구는 매우 부족한 형편이다.

이러한 이유로 본 연구는 대용량의 원천 데이터로부터 산업분석에 포함될 수 있는 정보를 파악하고 이를 카테고리 별로 자동으로 분류해 주는 텍스트마이닝 방법론을 제안하고자 한다. 구체적으로는 기계학습 방법 기반의 자질 선택 기법을 이용한 문장 분류기를 구축하여 산업분석 프레임워크 내에 들어갈 만한 정보로 판단되는 문장들을 분류하게 하였다. 수많은 종류의 산업분석 프레임이 존재하는 현실을 감안, 본 연구에서는 이들 프레임워크 종류 중에서 널리 사용되는 PEST에 대하여 네 가지 카테고리 (Political, Economic, Social, Technical)에 해당하는 문장을 데이터 원천으로부터 분류하여 정리하였다. 또한, 산업 분석 시 긍정적인 요소와 부정적인 요소를 가려내는 것도 의사결정에 큰 도움을 줄 수 있다. 이에 따라, 본 연구에서는 PEST뿐만 아니라, 긍부정 요소에 따른 카테고리도 같은 방식으로 분류하였다. <그림 1>과 같이, 문서 내 문장들이 산업분석 시스템에 입력되면, 분류 카테고리에 따라 문장들을 분류하여 출력한다. 이와 같이 분류된 문장들은 향후 기업의 전략 수립 및 의사결정에 활용될 수 있다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 기존 산업 분석 분야에서 연구된 자동화된 기법 및 적용 가능한 텍스트마이닝 기법 관련 연구

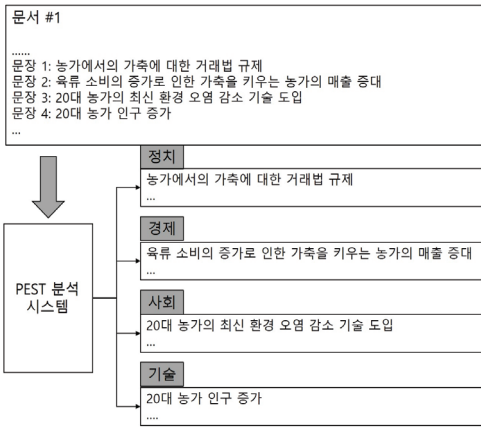

그림 1. PEST 분석의 시스템 입력 및 출력
Figure 1. Inputs and outputs of PEST analysis

들에 대해 소개한다. 3장에서는 데이터 및 모델 구축 과정을 포함, 본 연구에서 제안한 산업정보 분석을 위한 문장 분류 방법론에 대해 설명한다. 4장에서는 제안된 방법론의 성능 평가 결과에 대해 기술하며, 이에 따른 결과 분석을 제시한다. 마지막으로 5장에서는 결론 및 이 연구의 한계점에 대해 논의하고자 한다.

2. 관련 연구

2.1 기존 산업 분석 시스템

최근 데이터의 양이 늘어남에 따라 기존에 기업 분석 방법론을 자동화하고자 하는 시도가 있어 왔다. [4]는 비즈니스의 경쟁 환경을 모니터하고 분석하는 것을 목표로 하여 5세력 모형(Five Force Analysis; 이하 FFA)과 SWOT에 대하여 여러 텍스트마이닝 기술을 적용한 의사결정 모델을 제안하였다. [6]은 신문기사 등에서 SWOT 분석을 위한 키워드 추출을 하는 시스템 개발하기 위해 군집화 방식을 이용하였다. 하지만, 이들 방법론은 분석 시스템에 대한 설명과 예시를 언급하기는 하지만,

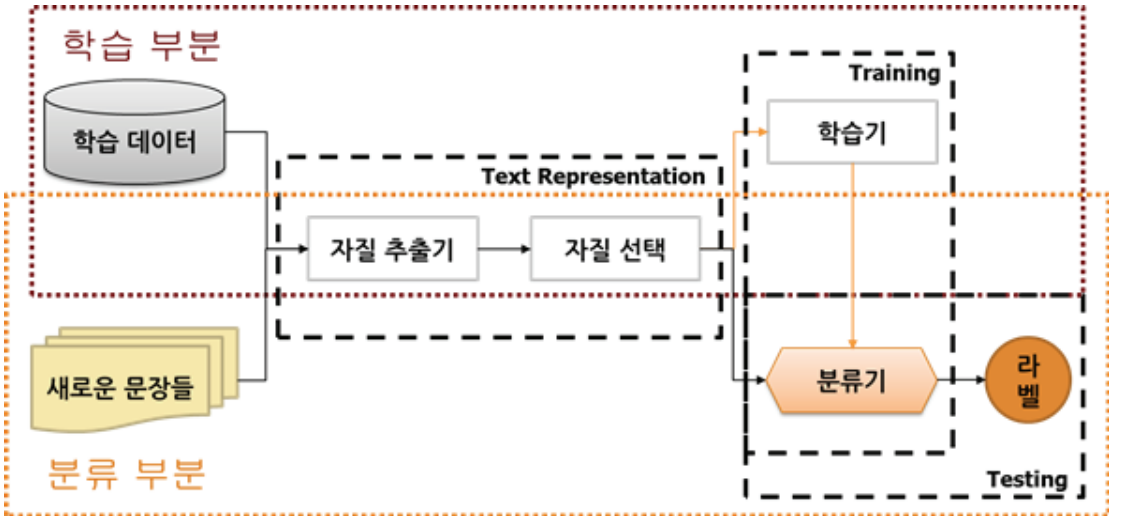

그림 2. 기계학습 기반의 산업 분석 시스템 구조도
Figure 2. System Architecture of Industry Analysis System based on Machine Learning

정확한 성능 등의 객관적 평가는 이루어지지 않았다. 또한, 군집화한 키워드를 결국 사람이 다시 분석하여 결정하여야 하는 등의 사람의 개입이 필요한 반자동화 시스템이라는 한계점을 가지고 있다.

FFA, SWOT 등의 기업 분석 이외에도 데이터마이닝 기법을 활용하여 경영자의 의사결정을 돕는 여러 연구가 시도되었다. [5]는 어떤 데이터 내 문장들이 시간, 비용 등의 유용한 정보를 가지고 있는지를 분류하는 시스템을 제안하였다. 이 과정에서 데이터마이닝에서 사용되는 Apriori Association Rule Mining 알고리즘을 사용하여 어휘 단위 자질을 추출하였으며, 이들을 기계학습 기법에 적용하였다. 또한, [7]은 연관 규칙 마이닝(association rule mining)과 change mining 기법을 이용하여 다른 두 시점 내에서 변화한 정보를 제공하는 이벤트 변화 탐지(event change detection) 시스템을 개발하였다.

2.2 텍스트마이닝과 기계학습 기법

텍스트마이닝을 위한 접근법은 크게 규칙 기반과 통계 기반으로 나눌 수 있다. 규칙 기반 접근법은 규칙을 한 번 정하면 그 규칙은 언제든 적용이 가능하므로 정확률이 높지만, 데이터가 바뀔 때마다 새로운 규칙을 지속적으로 생성해야 하는 어려움이 있다. 반면에, 통계 기반 접근법은 초기 학습 데이터가 부족하면 통계적 데이터의 신뢰도가 낮아 정확률이 낮으나, 데이터의 양이 많아질수록 그에 대하여 충분한 학습을 하므로 성능이 향상된다. 최근 정보의 양이 많아짐에 따라 데이터마이닝 분야에서는 통계 기반의 방법이 널리 사용되고 있으며, 특히 기계학습 기법을 활용하여 데이터마이닝에서 다루는 여러 문제를 해결하기 위해 시도하였다[8-10].

또한, 기계학습 기법을 이용한 분류 문제(classification problem) 해결에 있어서, 사용되는 자질의 수가 너무 많을 때에는 일부 자질은 분류에 도움이 되지 않아 분류 성능을 저해하거나, 학습 시간이 오래 걸리는 요인이 된다. 이를 해결하

기 위하여 기계학습 분야에서는 자질 선택(feature selection) 기법을 사용함으로써 분류에 도움이 되지 않는 자질을 제거하여 학습 속도의 향상과 분류기의 성능 향상을 시도한다. [11]

3. 산업 분석을 위한 문장 분류 시스템

3.1. 시스템 구조도

본 연구에서는 텍스트 데이터를 기계학습 기반으로 분류하는 산업 분석 시스템을 제안한다. <그림 1>에서와 같이, 본 시스템은 주어진 텍스트 문서를 문장 단위로 입력받아 각 문장에 산업분석 카테고리를 부착한다. 산업 분석을 위한 시스템의 구조도는 <그림 2>와 같다.

<그림 2>에서 보면, 본 시스템은 학습 부분과 분류 부분으로 나누어지며, 이는 PEST와 긍부정 분류에서 동일하게 적용된다.

학습 부분에서는 주어진 학습 데이터로부터 문장을 분류하는데 사용되는 자질을 추출하고, 이를 활용하여 산업 카테고리 분류기를 학습한다. 이를 위하여 본 시스템에서는 자질 추출기를 이용하여 학습 데이터 내에 존재하는 어휘 관련 자질을 추출한다. 그 후, 자질 선택기에서 분류에 도움이 되는 자질을 선별하여 추출한 자질들 중 선별된 자질들만 학습에 이용한다. 그 후, 학습기는 선택된 자질들을 활용하여 산업 분류 카테고리에 대한 분류기를 생성한다. 이 때, 어휘 자질을 이용하므로 매우 많은 수의 자질이 생성된다. 본 연구에서는 기계학습 분야에서 연구된 다양한 분류 알고리즘 중 다수의 어휘 자질 집합을 학습하기에 용이한 최대엔트로피(maximum entropy) 모델을 사용하였다.

분류 부분에서는 산업 분석을 수행할 새로운 문장에 대하여 학습 부분에서 학습한 분류기를 통해 산업 분석을 위한 카테고리로 분류한다. 입력된 분류 대상 문장에 대해, 학습 부분에서와 같이 자질 추출기를 이용하여 자질을 추출하고, 자질 선택기를 통하여 분류에 유용한 자질을 선별한다. 분류기는 입력 문장으로부터 선별된 자질들로 구성된 자질벡터를 입력으로 받아 카테고리 분류 결과를 출력한다. 본 연구에서는 산업 분석 카테고리로 PEST와 긍부정 분류를 수행하였다.

3.2. 문장 분류를 위한 자질 및 자질 선택

3.2.1. 사용된 자질의 종류

산업 분석을 하고자 하는 문장 텍스트를 이용하여 분류하기 위해서는 어휘 정보를 활용하여야 한다. 또한, 하나의 어휘만으로 알 수 있는 경우 외에도 연속된 두 개의 어휘, 또는 세 개 이상의 어휘열이 더 확실한 단서를 제공할 수 있다. 실제로 데이터마이닝 분야에서는 문장 분류를 수행하는 여러 문제에서 어휘 자질을 사용하였다.[8], [12], [13]

하지만, 한국어는 교착어로서 하나 이상의 형태소가 결합되어 어절을 형성하고, 어절의 어형의 변화가 많은 언어이다. 따라서 한국어에 대해 어휘 자질을 사용하기 위해서는 어절을 형태소로 분리해야 하는데, 이러한 과정을 형태소 분석이라고 한다. <표 1>은 <그림 1>에서의 문장 1에 대한 형태소 분석 결과를 보여준다.

표 1. 형태소 분석의 예
Table 1. Example of morphological analysis

어절	형태소 분석 결과
농가에서의	농가/NNG+에서/JKB+의/JKG
가축에	가축/NNG+에/JKB
대한	대하/VV+ㄴ/ETM
거래법	거래법/NNG
규제	규제/NNG

표 2. 산업분석 시스템의 어휘 자질에 대한 설명 및 예시
Table 2. Explanation and Example of Word Feature in Industry Analysis System

n-그램	자질	설명	예시
유니그램	형태소	하나의 형태소	농가, 가축, 대하, 거래법, 규제
	품사	형태소에 부착되는 품사 한 개	NNG, NNG, VV, NNG, NNG
	형태소/품사	하나의 형태소와 그에 부착된 품사	농가/NNG, 가축/NNG, 대하/VV, 거래법/NNG, 규제/NNG
바이그램	형태소	연속된 두 개의 형태소	농가_가축,가축_대하, 대하_거래법, 거래법_규제
	품사	연속된 두 형태소에 부착되는 태그	NNG_NNG, NNG_VV, VV_NNG, NNG_NNG
	형태소/품사	연속된 두 개의 형태소에 각각 부착된 품사	농가/NNG_가축/NNG, 가축/NNG_대하/VV, 대하/VV_거래법/NNG, 거래법/NNG_규제/NNG
트라이그램	형태소	연속된 세 개의 형태소	농가_가축_대하, 가축_대하_거래법, 대하_거래법_규제
	품사	연속된 세 형태소에 부착되는 태그	NNG_NNG_VV, NNG_VV_NNG, VV_NNG_NNG
	형태소/품사	연속된 세 개의 형태소에 각각 부착된 품사	농가/NNG_가축/NNG_대하/VV, 가축/NNG_대하/VV_거래법/NNG, 대하/VV_거래법/NNG_규제/NNG

형태소는 그 역할에 따라 내용어(content word)와 기능어(function word)로 나눌 수 있다. 그 중, 기능어들은 그 자체로 의미를 가지지 않으며, 특히 그 중 일부는 출현 빈도가 높아 분류에 도움이 되지 않는다. 이에 따라 본 연구에서는 문장에 속한 형태소들 중 문장 분류를 위한 자질로 내용어 및 조사(~의, ~에, ~으로, ~을 등), 받침(~ㄴ 등), 문장 기호(., !, ? 등) 등을 제외한 기능어를 사용하였으며, 제외된 기능어는 총 70개이다.

텍스트 문장 내에서 일부 단어들은 다른 단어들과 함께 사용되어 구(phrase)를 이룬다. 개별 단어로 사용되는 것보다 <그림 1>의 문장 3에서의 "환경 오염"과 같이 구를 이루어 함께 사용될 때 보다 명확한 의미를 파악할 수 있다. 실제로 텍스트마이닝 분야에서는 단어의 연쇄를 자질로써 활용하는데 [13], 자연어처리 분야에서는 이러한 단어의 연쇄를 "n-그램"이라 하며, 단어열의 길이가 각각 1,2,3인 유니그램, 바이그램, 트라이그램이 가장 널리 사용되며, 본 연구에서도 이들을 자질로 사용하였다.

텍스트마이닝 분야에서는 어휘 자질을 사용할 때 형태소 이외에도 품사를 자질로 사용하기도 하며, 여러 종류의 품사를 가질 수 있는 형태소를 구별하기 위해 형태소에 품사태그를 부착하여 '형태소/품사' 형식으로 활용하기도 한다. 이에 따라 본 연구에서는 각 n-그램 별로 형태소 이외에도 품사와 품사태그가 부착된 형태소를 자질로 활용하여 총 9개의 자질을 추출하여 학습에 사용하였으며, 이에 대한 설명과 예는 <표 2>에 있다. 또한, 본 연구에서는 이들 개별 자질을 조합하여 분류기를 학습함으로써 어떤 자질 조합을 사용할 때 가장 높은 분류 성능을 보이는지도 확인하고자 하였다.

3.2.2. 자질 선택

산업 분석을 수행할 때에 문장 내의 어휘 정보 중에서 모든 어휘 정보가 산업 분석에 도움이 되는 것은 아니다. 실제로, 문장 내 산업 요소를 분석할 때, 모든 문장에서 쓰일 수 있는 어휘들은 문장 내 분류에 도움이 되지 않을 뿐만 아니라 오히려 성능 하락의 요인이 될 수 있다.

<그림 3>은 <그림 1>의 4개의 예문의 형태소 중

표 3. PEST 분석 데이터
Table 3. Training data for PEST analysis

카테고리	문장 수	예시 문장
정책	1,619	정부는 반도체, 디스플레이 산업에 이은 차세대 신성장동력으로 태양광 산업 육성에 대한 의지가 확고
경제	1,642	2013년 3Q 상위 10개 업체의 생산용량이 17.06 GW로중국 기업 비중이 81.7%
사회	1,510	대기업 및 해외 우수기업의 투자 및 인수합병 확대에 따라국내 태양광 발전 시장 확대
기술	1,677	또한 태양광 모듈 양면에서 발전할 수 있는 Bifacial 모듈을 개발 완료
전체	6,448	

일부 단어를 bag-of-word 방식으로 표현한 것이다.

문장 1: 농가에서의 가축에 대한 거래법 규제
문장 2: 육류 소비의 증가로 인한 가축을 키우는 농가의 매출 증대
문장 3: 20대 농가의 최신 환경 오염 감소 기술 도입
문장 4: 20대 농가 인구 증가

문장	PEST 카테고리	형태소 자질								
		가축	환경	오염	거래법	농가	매출	최신	인구	...
문장 1	정치	1	0	0	1	1	0	0	0	...
문장 2	경제	1	0	0	0	1	1	0	0	...
문장 3	기술	1	1	1	0	1	0	1	0	...
문장 4	사회	0	0	0	0	1	0	0	1	...

그림 3. 예시 문장에 대한 어휘 분포 및 PEST 분석
Figure 3. Word Distribution and PEST analysis about example sentences

<그림 3>에서의 '농가'의 경우 네 가지 카테고리에 모두 분포하고 있으므로 해당 문장에 대한 특정 카테고리를 판별하는 데에 그다지 도움이 되지 않는다. 따라서 이러한 자질은 어휘의 분포를 확인하여 자동으로 제거하는 자질 선택 방법을 통해 제외할 수 있다.

본 연구에서는 ANOVA[14]와 카이제곱[15]을 자질 선택을 위한 척도로 각각 적용하였다. 모든 자질들을 각 척도에 의해 순위를 매기고, 이들 중 특정 비율 또는 임계값에 의한 상위의 자질들만을 선별하여 학습에 포함시키는 방식에 따라 자질 선택을 수행하였다.

4. 실험 및 평가

본 연구에서는 산업 분석 시스템의 효과성을 입증하기 위하여 PEST와 긍부정 분류의 성능을 평가하였다. 이를 위하여 본 절에서는 각 분류 시스템의 환경 및 분류 결과의 성능을 기술한다.

4.1. 실험 환경

본 시스템의 개발 환경은 다음과 같다.

- 운영체제: 유닉스 기반의 우분투 16.04
- 개발 언어: python 3.5
- 사용 Toolkit: numpy(Version 1.11.2 이상), scikit-learn(Version 0.18 이상)

또한, 학습 및 실험을 위한 데이터로는 KISTI (www.kisti.re.kr)가 보유하고 있는 PEST 데이터 (6,448개의 문장)와 긍부정 데이터 (9,398개의 문장)를 활용하였다. 카테고리 별 문장 수 및 카테고리 별 예시는 <표 3>과 <표 4>에 제시되어 있다.

표 4. 긍부정 분석 데이터
Table 4. Training data for polarity analysis

카테고리	문장 수	예시
긍정	4,689	고품질의 태양광 생산기술 확보
부정	4,709	공급과잉에 따른 시장가격 폭락
전체	9.398	

본 연구에서 자질 추출을 위해 [16]의 형태소 분석기를 사용하였다. 그리고 실험에 대한 평가 척도로는 정확도(accuracy)를 사용하였으며 다음 수식과 같이 계산하였다.

표 5. PEST 분석에 대한 자질 선택에 따른 개별 자질 간 성능 비교
Table 5. Performance of individual features in PEST analysis

n-그램	자질	자질 선택 하지 않음	ANOVA 자질 선택		Chi-square 자질 선택	
		정확도	정확도	최고 성능 상위 %	정확도	최고 성능 상위 %
유니그램	형태소	0.7353	0.7509	35	0.7493	35
	품사	0.3617	0.3676	60	0.3695	45
	형태소/품사	0.7380	0.7485	35	0.7470	35
바이그램	형태소	0.7217	**0.7656**	25	**0.7670**	25
	품사	0.3833	0.3946	45	0.3916	45
	형태소/품사	**0.7411**	0.7598	55	0.7600	35
트라이그램	형태소	0.6482	0.7099	20	0.7078	20
	품사	0.4081	0.4175	30	0.4157	40
	형태소/품사	0.7254	0.7552	30	0.7502	25

$$\text{정확도} = \frac{(\text{분류 결과 중 정답으로 분류한 문장의 수})}{(\text{분류하고자 하는 모든 문장의 수})}$$

또한, 평가 시 데이터의 편향 정도를 줄이고 학습 시 데이터를 최대한 활용하기 위하여 10차 교차검증(10-fold cross validation) 방식으로 성능을 평가하였다.

4.2. 실험 결과

4.2.1. PEST 분석

본 절에서는 PEST 분석 개별 자질의 성능 및 자질 선택 기법에 따른 성능 변화를 실험하였다. 또한, 이들 자질에 대한 조합을 통하여 PEST 분석의 성능도 측정하였다.

<표 5>는 각 자질별로 자질 선택 유무 및 자질 선택 기법에 따른 성능을 보여 준다. 이 때, 자질 선택 기법으로는 ANOVA, 카이제곱을 사용하였고, 자질 선택 비율을 5% 단위로 변화시켜 가며 실험하였다.

<표 5>의 결과를 통해 형태소 단위의 자질 중에서는 유니그램 형태소 자질이 가장 높은 정확도를

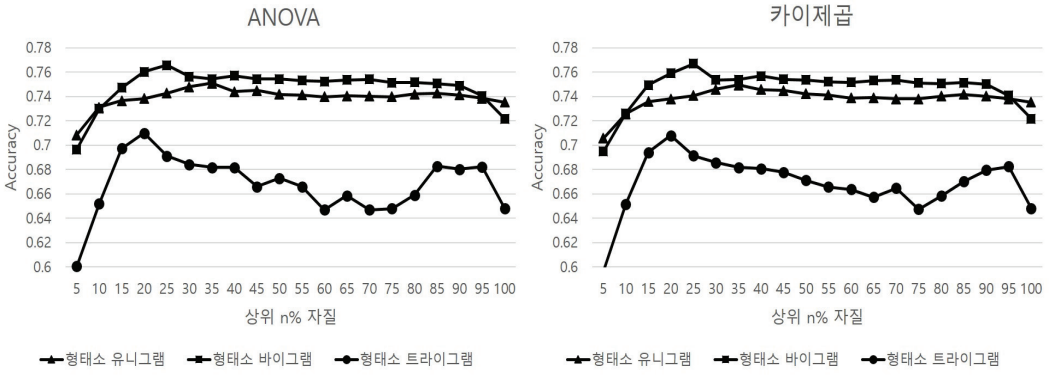

그림 4. PEST 분석에서의 자질 선택 기법 및 선택 비율에 따른 성능 변화
Figure 4. Performances of PEST Analysis selection method and selection rate

표 6. PEST 분석에서의 자질 조합 성능 비교
Table 6. Performance of combining features in PEST analysis

n-그램	자질	자질 선택 하지 않음	ANOVA	Chi-square
유니그램	형태소	0.7353	0.7509	0.7493
	형태소/품사	0.7380	0.7473	0.7470
바이그램	형태소	0.7217	0.7656	0.7670
	형태소/품사	0.7411	0.7598	0.7600
유니그램 + 바이그램	형태소	0.7642	0.7854	0.7826
	형태소/품사	0.7445	0.7648	0.7670
트라이그램	형태소	0.6482	0.7099	0.7078
	형태소/품사	0.7254	0.7552	0.7502
유니그램 +바이그램 +트라이그램	형태소	**0.7710**	**0.7940**	**0.7944**
	형태소/품사	0.7367	0.7643	0.7623

보이는 것을 알 수 있다. 이는 적절한 길이의 단어열이 학습에 도움이 됨을 보여 준다. 반면에, 품사 자질에서는 트라이그램 품사 자질이 성능이 가장 높다. 유니그램 품사 자질은 품사 개별의 성질만을 보여 주나, 트라이그램 품사 자질은 품사열을 통해 문장의 구조를 볼 수 있기 때문으로 보인다. 그리고 품사가 부착된 형태소 자질에서는 유니그램 자질은 품사가 부착되지 않는 형태소 자질보다 성능이 낮아지나, 바이그램 자질과 트라이그램 자질은 성능이 향상된다.

자질 선택을 했을 경우, 모든 자질에서 성능이 향상됨을 <표 5>를 통하여 알 수 있다. 이는 관찰 가능한 자질들 중 일부는 성능 향상에 도움이 되지 않는다는 것을 보여 준다. 또한, 두 가지 자질 선택 방법(ANOVA, 카이제곱) 모두 비슷한 성능을 보임을 알 수 있다.

<그림 4>는 형태소 자질의 세 가지 n-그램인 유니그램, 바이그램, 트라이그램에 대한 자질 선택 비율에 따른 정확도의 변화를 보여 준다. <그림 4>를 통해 두 자질 선택 기법 모두 비슷한 성능을 보인다. 그리고 지나치게 자질을 적게 선택하거나 많이 선택하는 경우에는 성능 하락의 폭이 크다.

그리고 본 절에서는 제안된 자질들의 조합 방법에 따른 성능 향상의 정도를 실험하였다. 자질을 조합할 때, 본 연구에서는 단어열의 길이 및 그들의 조합을 기준으로 실험하였으며, 이에 따라 총

표 7. 긍부정 분석에서의 개별 자질 간 성능 비교
Table 7. Performance of individual features in polarity analysis

n-그램	자질	자질 선택 하지 않음	ANOVA 자질 선택		Chi-square 자질 선택	
		정확도	정확도	최고 성능 상위 %	정확도	최고 성능 상위 %
유니그램	형태소	**0.8896**	0.9032	85	0.9041	80
	품사	0.5972	0.6003	55	0.6019	65
	형태소/품사	0.8968	0.9024	90	0.9028	85
바이그램	형태소	0.8731	0.9028	30	0.9030	30
	품사	0.6151	0.6194	35	0.6184	35
	형태소/품사	0.8978	**0.9059**	35	**0.9055**	35
트라이그램	형태소	0.8235	0.8749	20	0.8746	20
	품사	0.6259	0.6372	80	0.6372	80
	형태소/품사	0.8845	0.8993	70	0.8972	55

10가지 조합에 대하여 수행하였다. 이에 대한 결과는 <표 6>에 나타나 있다.

<표 6>의 결과를 통해, 자질을 조합하였을 때 성능이 향상됨을 알 수 있다. 특히, 유니그램, 바이그램, 트라이그램의 형태소 자질을 모두 사용하였을 때가 가장 성능이 좋은데, 이는 각 형태소 자질들이 상호 보완적 역할을 하였기 때문으로 보인다. 또한, 자질 선택을 하였을 때 월등히 성능이 좋으며, 이때에도 유니그램, 바이그램, 트라이그램의 형태소 자질을 모두 사용했을 때가 성능이 가장 좋았다.

4.2.2. 긍부정 분석

본 절에서는 긍부정 분석에 대하여 개별 자질의 성능 및 자질 선택 기법에 따른 성능 변화를 실험하였다. 그리고 본 연구에서 제안한 자질들의 조합을 통하여 긍부정 분석의 성능 향상 정도를 측정하였다.

본 연구에서는 앞에서 소개한 자질 선택의 효과를 비교하고 각각의 자질별 특성을 분석하기 위해 개별 자질의 분석 정확도에 대한 실험을 수행하였다. <표 7>은 개별 자질에 대하여 자질 선택의 유무 및 자질 선택 기법에 따라 변화하는 성능을 실험하였으며, 자질 선택 기법 및 비율은 PEST 분석 실험 때와 같다.

<표 7>의 결과를 통해, 형태소 단위의 자질 중에서는 유니그램 형태소 자질이 가장 성능이 좋다는 것을 확인할 수 있다. 이는 PEST 분석 결과와 같이, 적절한 길이의 단어열을 학습하는 것이 중요함을 의미한다. 반면에, 각 형태소에 대한 품사 자질 중에서는 트라이그램 품사 자질이 성능이 가장 좋다. 이는 유니그램 품사 자질과 달리 트라이그램 품사 자질이 문장 구조의 일부를 반영하기 때문인 것으로 보인다. 그리고 품사가 부착된 형태소 자질 중에서는 바이그램 자질의 성능이 제일 높다. 이를 통하여 품사가 부착되어 의미가 세분화된 바이그램은 성능 향상이 유니그램의 경우보다 좋을 수 있음을 알 수 있다.

자질 선택을 수행했을 경우에 ANOVA와 카이제곱 두 경우 모두 성능 향상이 있음을 <표 7>을 통해 확인할 수 있다. 이는 학습 데이터 내에서 관찰되는 자질들 중에는 학습에 도움이 되지 않는 자질의 비중이 상당함을 보여 준다. 또한, 본 연구에

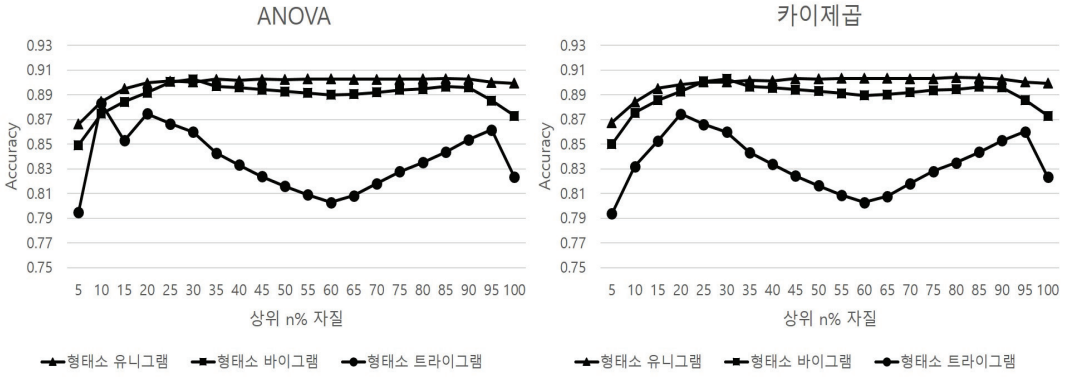

그림 5. 긍부정 분석에서의 자질 선택 기법과 선택 비율에 따른 성능 변화
Figure 5. Performances of Polarity Analysis selection method and selection rate

표 8. 긍부정 분석에서의 자질 조합 성능 비교
Figure 8. Performance of combining features in polarity analysis

n-그램	자질	자질 선택 하지 않음	ANOVA	Chi-square
유니그램	형태소	0.8996	0.9032	0.9041
	형태소/품사	0.8967	0.9024	0.9028
바이그램	형태소	0.8731	0.9028	0.9030
	형태소/품사	0.8978	0.9059	0.9055
유니그램 + 바이그램	형태소	0.9152	0.9255	0.9280
	형태소/품사	0.9091	0.9136	0.9128
트라이그램	형태소	0.8235	0.8749	0.8746
	형태소/품사	0.8845	0.8993	0.8972
유니그램 + 바이그램 + 트라이그램	형태소	0.9133	**0.9281**	**0.9301**
	형태소/품사	0.9032	0.9135	0.9185

서 수행한 두 가지의 자질 선택 기법 모두 긍부정 분석에 활용될 때 비슷한 성능을 보임을 알 수 있다.

<그림 5>는 형태소 자질의 세 가지 n-gram인 유니그램, 바이그램, 트라이그램의 자질 선택 비율에 따른 정확도를 자질 선택 기법 별로 보여 준다. <그림 5>에서 보듯이, 두 자질 선택 기법은 큰 차이를 보이지 않으며, 두 기법 모두 자질 선택을 할 경우에 더 좋은 성능을 보여 준다. 또한, 너무 적은 자질을 선택하거나, 너무 많은 자질을 선택하는 경우에는 성능이 하락하는 경향을 보아, 적절한 숫자의 자질을 선택하는 것이 분석의 정확도를 높이는 데 중요한 요인이 됨을 알 수 있다.

그리고 본 절에서는 소개한 자질들에 대하여 긍부정 분류에 있어 어떤 자질 조합이 효과적인지에 대하여 실험하였다. 자질을 조합하는 방법은 앞 절의 PEST 조합 방법과 같으며, 총 10가지 조합을 실험하였다. 이에 대한 결과는 <표 8>과 같다. <표 8>의 결과를 통해, 자질을 조합하여 학습할 경우에 분석 성능이 향상되었음을 알 수 있다. 자질 선택 유무 모두 유니그램, 바이그램, 트라이그램 형태소 자질을 사용할 때 성능이 가장 좋은데, 이는 각 형태소 자질들이 상호 보완적 역할을 수행하기 때문으로 보인다.

5. 결 론

본 연구는 텍스트마이닝 분야에서 문장 단위의 정보를 분석하고 목적에 부합하는 내용을 추출하는 기술로서 보편적으로 사용되고 있는 기계학습 기반의 분석 기술을 활용하여, 다양한 원천 데이터로부터 PEST와 SWOT와 같은 산업분석 프레임워크에 포함될 수 있는 내용을 검색하고 선택된 문장들을 추출, 카테고리 별로 분류하여 사용자에게 제시하는 텍스트마이닝 방법론을 제안하였다. 제안된 방법론은 온, 오프라인에 존재하는 방대한 양의 텍스트 데이터를 입력 데이터로 활용할 수 있다.

이에 대해 본 연구에서는 시스템의 성능 평가를 위해 KISTI (www.kisti.re.kr)가 보유하고 있는 PEST 데이터 (6,448개의 문장으로 구성)와 긍부정 데이터 (9,398개의 문장으로 구성)를 활용하여 평가를 수행하였다. 그 결과, PEST 및 긍부정 분류에서 어휘적 특성을 자질로 사용하고 이들 자질들을 조합함으로써 효과적으로 분류를 수행할 수 있음을 보였다. 또한, 자질 선택을 함에 따라 성능이 더욱 향상되는 것을 확인하였다.

기존의 산업정보 분석 과정은 대개 자료의 탐색에서 최종적인 산업정보 분석 보고서의 작성 과정까지의 전체 과정을 분석가 스스로의 노하우와 역

량에 의존하여 진행해야 하는 경우가 대부분이다. 본 연구에서 제안한 문장 분류 기반의 텍스트마이닝 시스템은 방대한 양의 데이터로부터 산업정보 분석에 활용되는 정보가 담긴 문장을 신속하고 정확하게 분류하여 사용자에게 제시할 수 있음을 실험을 통해 증명하였다. 본 연구에서 제안한 시스템을 도입할 경우, 방대한 양의 데이터를 손쉽게 분석함으로써 보다 풍부한 관련 정보를 추출하는 것이 가능해져서 결국 일정 품질 수준 이상의 산업정보 분석이 가능해 질 것으로 기대할 수 있다. 또한 분석 과정에 소요되는 많은 시간과 비용을 절약할 수 있기에 기업자원이 풍부하지 못한 중소기업 등에서도 외부 전문기관 등에 의존하지 않고 산업분석 업무를 진행, 도출된 결과를 활용할 수 있어 산업 전반에의 파급효과도 함께 기대할 수 있겠다.

다만, 본 연구에서 제안한 방법론이 아직은 충분한 검증과 보완을 거친 결과라기보다는 실험적 차원에서 구현되었다는 점, 실험 평가를 위해 일부 산업분석 프레임워크 (PEST, 긍부정)에 대해서만 적용가능하게 구현된 점 등은 현재 진행된 연구의 한계로 지적할 수 있다. 이를 해결하기 위해 첫째, 구현된 시스템에 대해 보다 세밀한 단계별로 성능을 점검, 보완하여 향후 완성도를 높이고 둘째, 특정 산업분석 프레임워크가 아닌 일반적으로 활용되는 모든 분석에 범용적으로 적용될 수 있도록 제안된 방법론의 확장성을 향상 시키는 것과 같은 후속 작업이 향후 후속 연구과제로서 이어질 수 있을 것이다.

References

[1] J. Morgan, *Participatory pest analysis*, PLA notes, pp. 84-86, 1997.

[2] T. Hill, and R. Westbrook, *SWOT analysis: it's time for a product recall*, Long range planning, Vol. 30, No. 1, pp.46-52, 1996.

[3] M. E. Porter, *The five competitive forces that shape strategy*, Harvard Business Review, 2008.

[4] Y. Dai, *MinEDec: a Decision-support model that combines text-mining technologies with two competitive intelligence analysis methods*, International Journal of Computer Information Systems and Industrial Management Applications, Vol. 3, pp. 165-173, 2011.

[5] N. Ur-Rahman, and J. A. Harding, *Textual data mining for industrial knowledge management and text classification: A business oriented approach*, Expert Systems with Applications, Vol. 39, No. 5, pp. 4729-4739, 2012.

[6] M. Samejima, Y. Shimizu, M. Akiyoshi, and N. Komoda, *SWOT analysis support tool for verification of business strategy*, 2006 IEEE International Conference on Computational Cybernetics, pp. 1-4. 2006.

[7] D. Liu, M. Shih, C. Liau, and C. Lai, *Mining the change of event trends for decision support in environment scanning*, Expert Systems with Applications, Vol 36, No. 2, pp. 972-984, 2009.

[8] B. Pang, and L. Lee, *Thumbs up?: Sentiment classification using machine learning techniques*, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, pp. 79-86, 2002.

[9] D. Buscaldi, P. Rosso, F. Pla, E. Segarra, and E. S. Arnal,, *Verb sense disambiguation using support vector machines: Impact of wordNet-extracted features*, CICLing, pp. 192-195, 2006.

[10] O. Amayri, and N. Bouguila, *A study of spam filtering using support vector machines*, Artificial Intelligence Review, Vol. 34, No. 1, pp. 73-108, 2010.

[11] B-J. Yi, D-G Lee, and H-C. Rim, *The effects of feature optimization on high-dimensional essay data*, Mathematical Problems in Engineering, 2015.

[12] L. Barbosa, and J. Feng, *Robust sentiment detection on twitter from biased and noisy data*, Proceedings of the 23rd International Conference on Computational Linguistics, pp. 36-44, 2010.

[13] T. Wilson, J. Wiebe, and P. Hoffmann, *Recognizing contextual polarity in Phrase-level sentiment analysis*, Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 347-354, 2005.

[14] S. Stigler, *The history of statistics: the measurement of uncertainly before 1900*, Cambrige, Mass:Belknap Press of Harvard University Press, 1990.

[15] F. Yates, *Contingency table involving small numbers and the chi-square test*, Supplement of the Journal of Royal Statistical Society, Vol. 1, No. 2, pp. 217-235, 1934.

[16] D-G. Lee, and H-C. Rim, *Probabilistic modeling of Korean morphology*, in IEEE Transactions on Audio, Speech, and Language Processing, Vol. 17, No. 5, pp. 945-955, 2009.

텍스트마이닝을 활용한 산업분석 방법론에 관한 연구: 문장 분류를 이용한 PEST와 긍부정 분석

김윤성[1], 이호창[2], 이석기[3], 이도길[4], 김한국[5], 김유일[6]

[1,2] *고려대학교 컴퓨터학과*

[3] *한성대학교 컴퓨터공학부*

[4] *고려대학교 민족문화연구원*

[5] *과학기술연합대학원대학교 과학기술정보과학과/ 한국과학기술정보연구원 산업정보분석실*

[6] *한국과학기술정보연구원 산업정보분석실*

요 약

오늘날의 기업들은 날로 치열해 지는 산업 내 경쟁 속에서의 생존을 위해 끊임없이 자기가 속한 산업의 변화와 동향을 파악하고 이를 자사의 정책이나 제품 개발에 주기적으로 반영하면서 생존해야 하는 환경하에 있다. 이를 위해 주기적으로 수행해야 할 업무 중 하나가 산업정보의 분석이다. 대다수의 기업들은 많은 시간, 인력을 투입하거나 혹은 적지 않은 비용을 들여 외부 전문 분석기관의 도움을 받는 형태로 산업분석 정보를 획득하고 있다. 하지만 이러한 기존의 방식이 다소 휴리스틱하고 정성적인 접근법임으로 인해 분석 결과의 품질이 매번 다르다는 점, 엄청난 양의 산업관련정보가 실시간으로 온라인에서 생산되고 있고 이들 정보를 최대한 분석에 반영할 경우 보다 높은 품질의 분석결과를 기대할 수 있음을 감안할 때 기존과는 다른 새로운 방식의 분석 기법 도입이 요구된다. 이에 본 연구에서는 대용량의 원천 데이터로부터 산업분석에 포함될 수 있는 정보를 추출하고 이를 산업분석 프레임워크의 각 카테고리별로 자동 분류해주는 텍스트마이닝 방법론을 제안한다. 기계학습 기반의 문장 분류기를 구축하여 보편적으로 활용되는 산업분석 프레임워크의 지표별로 분류될 수 있는 정보를 문장 형태로 수집하게 하였다. 제안한 시스템을 이용하여 PEST와 긍부정 분석을 수행하였으며, 실험을 통해 제안된 시스템의 분류 정확도를 평가하였다.

감사의 글

이 논문은 KISTI '중소중견기업을 위한 산업시장 분석 정보 활용체제 기반 강화 (K-16-L04-C02-S02)'과제의 지원과 2007년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2007-361-AL0013)

Yoon-Sung Kim received the B.D. in computer science from Korea University, Seoul, in 2013. He is an Ph. D. Candidate student in Korea University. His research interests include natural language processing, information retrieval, and machine learning.

E-mail address: kys0205@korea.ac.kr

Ho-Chang Lee received the B.D. in computer software engineering from Kumoh National Institute of Technology, Gumi, in 2013. He is an Ph.D. Candidate student in Korea University. His research interests include natural language processing, machine translation, and information retrieval.

E-mail address: pery@korea.ac.kr

Seok Kee Lee received the bachelor's degree in the Department of Computer Science from the Korea University in 2002. He received the M.S. degree and the Ph.D. degree in the Department of Management Engineering from KAIST in 2002 and 2009, respectively. From 2005 to 2010, he was a Professor at Dongyang Mirae University. He has been a professor in the College of Computer Engineering at Hansung University since 2012. His current research interests include data mining, recommender systems.

E-mail address: seelee@hansung.ac.kr

Do-Gil Lee received the M.S. and Ph.D. degrees in computer science from Korea University, Seoul, in 2001 and 2005, respectively. He is an HK Professor at the Research Institute of Korean Studies, Korea University. He was a research engineer at the NHN Corporation. His research interests include natural language processing, information retrieval, and machine learning.

E-mail address: motdg@korea.ac.kr

Han-Gook Kim received the M.S. and Ph.D. degrees in Industrial Engineering from Tokyo Institute of Technology in 2003 and 2007, respectively. He has been a principal researcher in the Department of industry information analysis at KISTI since 2009. His current research interests include industry information system analysis, text mining, and big data analysis.

E-mail address: hgkim712@kisti.re.kr

You-Eil Kim received the M.S. and Ph.D. degrees in biochemical engineering from Seoul National University, Seoul, in 1997 and 2002, respectively. He is an principal researcher in charge of the department of Industry and Market Analysis at the Korea Institute of Science and Technology Information. His research interests include qualitative and quantitative analysis on industries and markets related on biotechnology.

E-mail address: yekim@kisti.re.kr