*The Korean Journal of Applied Statistics* (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.7.1207
**27**(7), 1207–1217

# Text Mining for Korean: Characteristics and Application to 2011 Korean Economic Census Data

Juna Goo[a,b] · Kyunga Kim[a,1]

[a]Biostatistics and Clinical Epidemiology Center, Samsung Medical Center
[b]Department of Statistics, Sookmyung Women's University

(Received October 14, 2014; Revised November 12, 2014; Accepted November 21, 2014)

**Abstract**

2011 Korean Economic Census is the first economic census in Korea, which contains text data on menus served by Korean-food restaurants as well as structured data on characteristics of restaurants including area, opening year and total sales. In this paper, we applied text mining to the text data and investigated statistical and technical issues and characteristics of Korean text mining. Pork belly roast was the most popular menu across provinces and/or restaurant types in year 2010, and the number of restaurants per 10000 people was especially high in Kangwon-do and Daejeon metropolitan city. Beef tartare and fried pork cutlet are popular menus in start-up restaurants while whole chicken soup and maeuntang (spicy fish stew) are in long-lived restaurants. These results can be used as a guideline for menu development to restaurant owners, and for government policy-making process that lead small restaurants to choose proper menus for successful business.

Keywords: Text mining, dictionary construction, big data, Korean economic census.

## 1. 서론

오늘날 현대사회는 빅데이터 시대를 맞이하여 엄청난 용량(volume)과 빠른 생성 속도(velocity), 다양한 형태와 구조(variety)의 특징을 가지고 있는 데이터로부터 의미 있는 가치(value)를 발견하는 것이 중요시되고 있다. 이러한 빅데이터 시대에 정보 활용의 대표적인 방법론으로는 데이터 마이닝을 꼽을 수 있다. 과거 데이터 마이닝은 대부분 정형화된 수치 데이터를 위주로 발달하였으나 최근에는 수치 데이터뿐만 아니라 텍스트, 영상 등과 같이 형태와 구조가 복잡한 비정형 혹은 비구조화 데이터를 대상으로 한 연구들이 많이 보고되고 있는 추세이다(Vijayarani와 Vinupriya, 2013; Bartere와 Deshmukh, 2012; Ko 등, 2011; Choi와 Kim, 2009). 비구조화 데이터 중에서 텍스트 데이터를 기반으로 한 데이터 마이닝의 방법론을 텍스트 마이닝(text mining)이라 하며 이는 Feldman과 Dagan(1995)에 의해 처음 언급된 이후 다양한 분야에서 발전되어 왔다.

과거 텍스트 마이닝은 주로 영어권 문서를 중심으로 연구되어 왔다. 영어 기반 텍스트 마이닝 사례를 몇 가지 소개하자면, Lin (2003)은 저널의 개별 기사에 수록된 제목과 내용 사이의 구(phrase)를 매치

[1]Corresponding author: Professor, Biostatistics and Clinical Epidemiology Center, Samsung Medical Center, Irwon-ro 81, Gangnam-gu, Seoul 135-710, Korea. E-mail: kyunga.j.kim@gmail.com

시키는 텍스트 마이닝 알고리즘을 제시하였고 Mittermayer과 Knolmayer (2006) 는 뉴스 기사에 대한 단기 시장 반응을 예측하는 데 있어 개발된 텍스트 마이닝 기반의 시제품들(prototypes)을 비교 및 평가하였다. 그밖에 van Driel 등 (2006)는 5000개 이상의 인간의 표현형(human phenotype)을 분류하는데 텍스트 마이닝을 사용하였으며 또한 표현형간의 유사성이 생물학적 기능의 측면에서 관련된 유전자간의 상호작용을 반영한다는 것을 밝혀냈다. 근래에는 영어뿐만 아니라 한국어 기반의 텍스트 마이닝 연구도 그 힘을 싣고 있다. Kam과 Song (2012)은 텍스트 마이닝을 활용하여 경향신문, 한겨레, 동아일보 세 개의 신문기사의 내용 및 논조 차이점을 단순빈도 분석과 군집분석(clustering), 분류분석(classification)의 결과를 통해 비교하였고, Yang과 Ko (2011)은 텍스트 마이닝의 한 분야인 비교 마이닝(comparison mining)을 이용하여 ‘ 보다’와 같은 조사가 포함된 한국어 비교문장에서 비교 요소를 자동적으로 추출하는 방법을 제시하였다. Ahn (2011)은 한국어 웹 문서를 바탕으로 오피니언 마이닝(opinion mining)을 위한 체계적인 사전을 구축하고 사용자의 오피니언 문장을 분류하였는데 여기서 오피니언 마이닝은 전산 언어학(computational linguistics)이 접목된 분야로 텍스트 마이닝의 일종이라 볼 수 있다.

본 연구에서는 2011년에 최초로 전국의 사업체를 대상으로 실시한 경제총조사의 데이터 중 한식 음식점업 사업체를 대상으로 설문조사의 주관식 응답 형태로 조사된 텍스트 자료에 한국어 텍스트 마이닝을 적용하여 한식 음식점업 사업체가 취급하는 대표 메뉴의 현황과 특성을 고찰하였다. 이 때 함께 조사된 수치 자료(예. 매출액, 창설년월) 및 범주 자료(예. 행정구역, 체인점 가입 여부)와 대표 메뉴명의 텍스트 자료를 연계하여 한식 음식점업 사업체의 현황에 대한 좀 더 다각적인 분석을 시도하였다. 또한 경제총조사라는 빅데이터에 한국어 텍스트 마이닝을 적용할 때 발생하는 문제점들을 단계별 과정에 따라 관찰 및 정리하고 이에 대한 해결 방법을 제안함으로써 향후 한국어 텍스트 마이닝 관련 연구에 가이드라인을 제시하고자 하였다. 우선 2장에서는 전반적인 텍스트 마이닝의 단계별 과정을 간략하게 요약하였으며, 3장에서는 본 논문의 연구대상인 2011 경제총조사의 특징을, 4장에서는 한국어 텍스트 마이닝을 적용한 실제 자료 분석 결과를 각각 기술하였다. 끝으로 5장에서는 결론과 관련된 후속 연구를 제시하였다.

## 2. 텍스트 마이닝의 단계별 과정

텍스트 마이닝은 단계별 과정에 따라 크게 자료 처리과정(data processing)과 자료 분석(data analysis)으로 나눌 수 있다. 자료 처리과정이란 정보 검색(information retrieval), 정보 추출(information extraction), 자연어 처리(natural language processing) 등을 기반으로 텍스트 데이터를 가공하는 단계이다. 자료 분석은 데이터 마이닝, 기계학습(machine learning), 통계학(statistics) 등을 활용하여 텍스트로부터 의미 있는 정보를 추출하는 단계를 말하는데(Hotho 등, 2005), 자료 처리과정상에서도 자료 분석상의 기법들을 적용하여 자료 처리과정의 성능을 높이는 연구도 활발히 진행 중이다(Jeong 등, 2013; Choi등, 2009). 구조화 데이터를 사용하는 데이터 마이닝에 비해, 텍스트 마이닝과 같이 비구조화 데이터를 다루는 경우는 자료 처리과정의 중요도가 상대적으로 높다. 특히 텍스트 마이닝은 한국어나 영어 등 문맥에 의존적인 자연어를 대상으로 하기 때문에 자료 처리과정에서의 자연어 처리는 필수적이며 자료 분석 결과의 질에 큰 영향을 준다.

자료 처리과정에서 전처리(preprocessing)가 완료된 텍스트를 저장하는 것을 사전(dictionary) 구축이라 한다. 사전에는 세종 말뭉치(Lee 등, 2010), 한나눔 한국어 형태소 분석기(SWRC, 1999) 등과 같이 자료에 관계없이 개발되고 제공되는 범용 사전과 특정 자료에 맞게 사용자가 직접 구축한 사용자 사전이 있다. 사전을 구축하는 과정에서 텍스트 데이터베이스의 크기를 효과적으로 줄여줄 수 있는데, 대표적인 방법으로 불용어(stopwords) 제거가 있다(Hotho 등, 2005). 불용어란 내포하고 있는 정보의 양이 극히 적은 단어를 말하며, 사전에서 불용어를 제거함으로써 큰 정보의 손실 없이 데이터베이스의 크

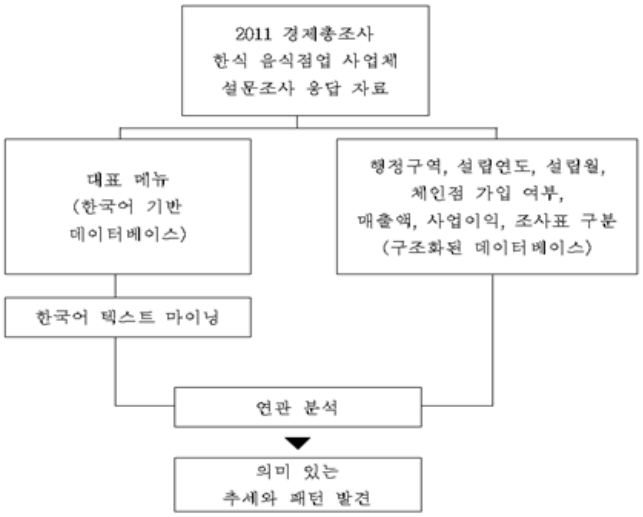


**Figure 3.1.** Flow chart of data analysis

기를 축소시킬 수 있다. 불용어는 특정 분야에서만 불용어가 되는 전문 불용어와 분야에 관계없이 항상 불용어가 되는 일반 불용어가 있다(Lee 등, 1997). 텍스트 마이닝에서 범용 사전을 사용할 경우 일반 불용어에 해당하는 단어가 사전에 등재되지 않았거나 등재는 되어있으나 전문 불용어로 지정해야 하는 경우가 존재하기 때문에 불용어 선정에 유의해야 한다. 다시 말해 범용 사전은 생성 및 소멸 주기가 빠른 자연어의 특징상 특정 자료에 한해서는 부적합할 수 있다. 반면 특정 자료에 맞게 구축된 사용자 사전의 경우 구축 단계에서 수작업 과정이 많으며 편의(bias)가 발생할 수 있다는 단점이 있으나 특정 자료에 적합하도록 설계되었기 때문에 범용사전을 사용하는 것보다 분석 목적을 달성하는 데 효율적이다. 사전 구축이 완료되면 이를 바탕으로 자료 분석을 실시하여 텍스트에 대한 빈도 분석 등을 통해 의미 있는 추세와 패턴을 탐색한다.

## 3. 실제 자료의 특징 및 분석 개요

2011 경제총조사는 우리나라 최초로 전국의 사업체를 대상으로 실시한 대규모 경제 분야 전수조사로 기준년도는 2010년이며 기존의 서비스업총조사와 산업총조사를 통합한 것이다. 본 연구에서는 통계청의 산업 분류 기준에 따라 서비스업에서 숙박 및 음식점업의 하위 범주인 음식 및 주점업, 그 중에서도 총 280,717개의 한식 음식점업 사업체에 국한하여 분석하였다. 자료의 특기할 만한 사항은 매출액, 사업이익 등 구조화된 데이터와 한국어 텍스트로 이루어진 비구조화된 데이터를 포함하고 있다는 점이다. 따라서 일차적으로 한국어 텍스트 마이닝을 단계별로 진행하였으며 이차적으로는 가공된 텍스트 데이터와 구조화된 데이터베이스를 병합하여 연관 분석을 실시하였다. 이상의 분석 절차는 Figure 3.1에 도식화되어 있다.

### 3.1. 텍스트 자료

2011 경제총조사의 숙박 및 음식점업에 대한 조사표에서 한국어 텍스트로 기록하는 주관식 설문 문항은 '사업의 종류'로 식재료를 나타내는 '무엇을 가지고'와 조리법을 나타내는 '어떤 방법으로' 그리고 대표

메뉴를 나타내는 '생산·제공 하였는가' 등 3개의 항목이다. 본 연구에서는 '무엇을 가지고'와 '어떤 방법으로'를 분석 대상에서 제외하였는데, 그 이유는 이들 항목에 대한 응답이 의미가 없거나 식재료와 조리법과 관계없는 경우가 대다수여서 분석의 대상으로 적절하지 않다고 판단되었기 때문이다. 따라서 한식 음식점업 사업체의 대표 메뉴가 기록된 '생산·제공 하였는가' 항목만을 텍스트 기반 데이터베이스로 선정하여 분석하였다. '생산·제공 하였는가' 항목에 대한 응답들의 특성을 살펴보면 대부분 응답이 '회', '삼겹살' 등과 같이 대표 메뉴로 입력되어 있다. 특히 복수의 대표 메뉴가 열거되는 응답이 많았으며, 반점(,)이나 온점(.) 혹은 공백 등 다양한 구분자가 사용되었다. 또한 메뉴와 전혀 관련이 없거나(예. '서비스제공'), 메뉴가 구체적으로 명시되지 않은(예. '식사', '음식') 응답도 있었으며, 입력 상의 오타나 잘못된 띄어쓰기 문제도 빈번하게 발견되었다. 동일한 메뉴를 지역마다 다르게 부르는 방언(예. '칼치', '갈치')이나 수식어 등에 의해 조합된 메뉴로 인해 혼란이 있기도 하였다. 이는 2011 경제총조사 자료를 수집 및 기록하는 과정에서 조사원들에게 대표 메뉴에 대한 표준화된 지침서를 제공하지 않아서 생긴 오류라고 생각된다.

### 3.2. 구조화 자료

각 사업체가 취급하는 대표 메뉴에 대한 텍스트 자료와 함께 분석할 구조화 데이터는 '행정구역', '설립연도', '설립 월', '체인점 가입 여부', '매출액', '사업이익', '조사표구분'이다. '행정구역'은 시도별 행정구역 분류번호이며 '설립연도', '설립 월'은 해당 사업체의 창설연월이다. '체인점 가입 여부'는 한식 음식점업 사업체의 프랜차이즈 가입 여부를 나타낸다. '매출액(단위: 백만원)'은 2010년 영업활동을 통한 총 수입액을 말하며 '사업이익(단위: 백만원)'은 매출액에서 사업(영업)비용을 제외한 금액을 말한다. '조사표 구분'은 각 사업체의 종사자수에 따라 조사표 (1)과 (6)으로 구분하여 각각 4인 이하 사업체와 5인 이상 사업체를 대상으로 한다. 여기서 '설립연도', '설립 월', '조사표구분'을 이용하여 2개의 파생 변수를 생성하였는데, '영업 기간' 변수는 영업 기간에 따라 사업체를 신생, 일반, 장수 사업체로 구분하는 범주형 변수이며 '사업체 규모' 변수는 사업체의 종사자 수에 따라 대형 및 영세 사업체로 구분하는 범주형 변수이다. 이들 파생변수에 대한 가변수(dummy variable)의 생성 과정은 다음과 같으며, 그 결과는 Table 3.1에 정리되어 있다.

첫째, '영업 기간' 변수를 만들기 위해 2011 경제총조사 본 조사 종료시점인 2011년 6월 24일을 기준으로 하여 각 사업체의 '설립 연도'와 '설립 월'을 이용하여 총 영업년수를 계산하고 이를 통해 사업체들을 세 개의 범주, '신생 사업체'(총 영업년수 2년 미만, 즉 설립연월이 2009년 7월 이후), '장수 사업체'(총 영업년수 10년 이상, 즉 설립연월이 2001년 6월 이전), '일반 사업체'(총 영업년수가 2년 초과 및 10년 미만)로 범주화하였다. 이 때 '신생 사업체'와 '장수 사업체' 범주에 각각 해당하면 1, 그렇지 않으면 0이 되는 2개의 가변수를 도입하였다. 둘째, '사업체 규모'는 기존의 '조사표 구분'을 이용하여 만든 파생변수로 값이 1 이면 종사자수가 5인 이상인 '대형 사업체', 0이면 4인 이하인 '영세 사업체'로 분류하는 가변수로 표현하였다. 따라서 본 연구에서는 '행정구역', '체인점 가입 여부', '매출액', '사업이익'의 기존 변수와 '영업 기간', '사업체 규모'의 파생 변수를 이용하여 대표 메뉴와 취급하는 사업체간의 연관성을 알아봄으로써 각 대표 메뉴를 취급하는 사업체의 특징을 파악하고자 하였다.

## 4. 한국어 텍스트 마이닝을 이용한 실제 자료 분석

### 4.1. 자료 처리과정

본 연구에서는 자립형태소인 명사로 된 대표 메뉴를 추출하는 자연어 처리를 중심으로 다음 6가지 자료처리과정 규칙을 제안한다.

**Table 3.1.** Two derived variables with dummy coding.

| 변수명 | 범주 | 코딩 |
|---|---|---|
| 영업 기간 | 신생 사업체 | 1 0 |
| | 일반 사업체 | 0 0 |
| | 장수 사업체 | 0 1 |
| 사업체 규모 | 대형 사업체 | 1 |
| | 영세 사업체 | 0 |

(처리 1) *문장부호(punctuation)를 공백(즉, 한 칸 띄어쓰기)으로 대체한다.* 텍스트 마이닝에서 빈도 분석은 띄어쓰기 단위로 이루어지기 때문에 반점(,) 이나 온점(.) 등의 구두점 대신 공백을 사용한다. 예를 들어 '닭발.떡볶이.국수'는 '닭발 떡볶이 국수'로 정정한다.

(처리 2) *불용어를 선정하고 제거한다.* 본 연구에서는 일반 불용어로 '을', '를', '랑', '와' 등의 조사와 '및', '함께', '또는' 등의 부사를 선정하였고, 전문 불용어로는 정보는 없으나 빈도가 높아 실제 분석에 방해가 되는 불필요한 명사(예. '제공', '종류', '판매', '서비스' ) 와 메뉴명의 조합에 추가되어 빈도 분석에 혼란을 야기하는 명사(예. '양념', '한방', '진흙') 등 두 가지 경우를 선정하였다.

(처리 3) *오타 및 맞춤법 오류를 정정한다.* 국립국어원의 「표준국어대사전」에 의거하여 입력 과정상의 오타와 맞춤법의 오류를 표준어로 정정한다. 본 자료에서 빈번하게 발견된 오타 및 오류는 '찌게, 째개, 지개, 지게(찌개)', '뽁음, 복음(볶음)', '낚지(낙지)', '덥밥(덮밥)' 등이 있다.

(처리 4) *띄어쓰기와 붙여쓰기를 정정한다.* 예를 들어 '비빔밥김치찌개된장찌개'는 '비빔밥 김치찌개 된장찌개'처럼 올바른 띄어쓰기로 정정하였다. 경우에 따라서는 붙여쓰기를 하는 것이 바람직할 수 있는데, 그 예로 '고기 구이'는 빈도 분석 과정에서 의미없는 단어인 '고기'와 '구이'로 각각 집계되므로 붙여쓰기를 하여 '고기구이'로 정정하였다. 본 자료에서는 '구이', '주물럭', '탕', '찜', '볶음' 등 메뉴명이 조리법으로 끝나는 경우에 잘못된 띄어쓰기 문제가 유독 많은 것을 확인하였고 이를 붙여쓰기 하는 것으로 정정하였다.

(처리 5) *동일한 메뉴에 대한 다양한 명명(命名)을 표준화한다.* 한식 메뉴명은 지역별 방언이나 외래명 등에 따라 다르게 명명되는 경우가 많기 때문에 국립국어원의 「표준국어대사전」과「두산백과」등 한식 메뉴명이 게재된 문헌 및 사이트를 참고하여 메뉴명을 표준화하였다. 그 예로는 '설농탕(설렁탕)', '국시(국수)' 등이 있다.

(처리 6) *생략된 부분을 복원한다.* 조사원이 응답을 연달아 입력하는 과정에서 단어의 일부가 생략되어 정보의 손실이 있는 경우라 판단되면 앞의 단어를 참고로 생략된 부분을 복원한다. 예를 들어 '오리구이,탕'은 '오리구이', '오리탕'으로 정정하였다.

이상의 6가지 자연어 처리를 총 280,717개의 한식 음식점업 사업체에 대한 자료에 대해 적용한 결과, 총 261,045개의 한식 음식점업 사업체에 대한 자료가 최종 분석 대상으로 남게 되었다. 다시 말해 전체 사업체의 7.008%에 해당하는 19,672개의 사업체의 '생산·제공 하였는가' 항목에 대한 응답이 대표 메뉴명에 적합하지 않은 응답으로 이루어져 자료 처리과정에서 삭제되었다. 여기서 자료 처리과정이 완료된 텍스트는 범용 사전이 아닌 본 자료에 맞게 구축된 사용자 사전의 형태로 저장되었다. 범용 사전을 이용하지 않은 이유는 한식 음식점업 사업체에서 취급하는 메뉴명이 모두 등재되어 있는 표준화된 범용 사전이 현재로서는 없기 때문이다.

**Table 4.1.** Top 5 menus offered by Korean restaurants during year 2010 across provinces.

| 지역 | 1위 | 2위 | 3위 | 4위 | 5 위 |
|---|---|---|---|---|---|
| 전국 | 삼겹살구이 (8.68%) | 생선회 (8.67%) | 백반 (8.41%) | 매운탕 (4.44%) | 된장찌개 (4.41%) |
| 서울특별시 | 백반 (9.28%) | 삼겹살구이 (7.63%) | 생선회 (6.10%) | 김치찌개 (4.11%) | 순댓국 (4.09%) |
| 부산광역시 | 생선회 (13.69%) | 된장찌개 (7.01%) | 삼겹살구이 (6.69%) | 매운탕 (5.98%) | 아귀찜 (5.35%) |
| 대구광역시 | 생선회 (9.53%) | 된장찌개 (7.88%) | 삼겹살구이 (6.63%) | 칼국수 (6.42%) | 돼지국밥 (3.99%) |
| 인천광역시 | 백반 (12.41%) | 생선회 (9.94%) | 삼겹살구이 (8.35%) | 매운탕 (4.58%) | 순댓국 (4.56%) |
| 광주광역시 | 백반 (12.65%) | 삼겹살구이 (12.15%) | 생선회 (7.87%) | 해장국 (3.65%) | 매운탕 (3.51%) |
| 대전광역시 | 백반 (14.68%) | 삼겹살구이 (12.62%) | 칼국수 (7.55%) | 생선회 (4.71%) | 해장국 (3.37%) |
| 울산광역시 | 생선회 (10.65%) | 삼겹살구이 (9.19%) | 매운탕 (5.33%) | 된장찌개 (5.29%) | 돼지갈비 (4.26%) |
| 경기도 | 백반 (9.49%) | 삼겹살구이 (9.39%) | 생선회 (5.70%) | 김치찌개 (5.38%) | 된장찌개 (4.31%) |
| 강원도 | 생선회 (10.30%) | 백반 (8.80%) | 삼겹살구이 (8.22%) | 매운탕 (5.61%) | 칼국수 (4.46%) |
| 충청북도 | 삼겹살구이 (11.53%) | 백반 (10.58%) | 칼국수 (5.17%) | 생선회 (4.47%) | 해장국 (4.44%) |
| 충청남도 | 백반 (10.76%) | 삼겹살구이 (10.12%) | 생선회 (7.92%) | 매운탕 (6.28%) | 김치찌개 (4.69%) |
| 전라북도 | 백반 (22.27%) | 삼겹살구이 (9.48%) | 생선회 (7.38%) | 매운탕 (5.59%) | 순댓국 (3.26%) |
| 전라남도 | 백반 (20.38%) | 생선회 (12.81%) | 삼겹살구이 (8.14%) | 매운탕 (7.71%) | 닭백숙 (3.27%) |
| 경상북도 | 생선회 (11.74%) | 삼겹살구이 (8.08%) | 된장찌개 (7.92%) | 칼국수 (4.26%) | 매운탕 (3.82%) |
| 경상남도 | 생선회 (13.85%) | 삼겹살구이 (8.78%) | 매운탕 (6.96%) | 된장찌개 (5.86%) | 아귀찜 (3.92%) |
| 제주특별자치도 | 생선회 (11.25%) | 매운탕 (6.23%) | 해장국 (5.80%) | 돼지갈비 (4.57%) | 김치찌개 (3.70%) |

### 4.2. 자료 분석

자료 분석에서는 사용자 사전에 있는 각 대표 메뉴에 대해 해당 메뉴를 취급하는 한식 음식점업 사업체들의 수 (빈도)와 함께 매출액 및 사업이익의 평균, 중위수, 표준편차 등의 기술통계량을 구하였다. 이 때 빈도와 기술통계량들은 전체 사업체를 대상으로 산출하였고, 행정구역(시도), 영업 기간(신생, 일반, 장수), 사업체 규모(대형, 영세), 체인점 가입 여부의 범주별로도 산출하였다.

**4.2.1. 시도별 비교** 본 절에서는 전체 한식 음식점 사업체들을 대상으로 메뉴별로 취급하는 사업체의 빈도를 측정하여 메뉴의 순위를 집계하여 빈도가 높은 인기 대표 메뉴를 조사하였다. 취급하는 사

업체의 수가 많은 상위 5개 인기 메뉴로는 삼겹살구이(8.68%), 생선회(8.67%), 백반(8.41%), 매운탕(4.44%), 된장찌개(4.41%)가 있었다. 또한 각 시도별로도 상위 5개의 인기 대표 메뉴를 살펴보았다(Table 4.1). 상당수 지역에서 생선회나 백반이 1위를 차지하였으나 전국적으로는 삼겹살구이를 취급하는 사업체의 수가 가장 많았다. 이는 사업체수가 많은 지역들에서 생선회나 백반보다 삼겹살구이가 1위~3위 등 최상위권 내에 더 많이 들었기 때문으로 생각된다. 생선회의 경우 주로 해안가와 인접한 부산광역시, 울산광역시 등에서 취급하는 사업체가 많은 것으로 나타났는데, 내륙인 대구광역시에서도 다른 메뉴보다 취급하는 사업체가 많은 점이 특이하였다. 대구광역시에서 생선회를 취급하는 사업체의 특징을 알아보기 위해 해당 사업체의 영업 기간과 사업체 규모 등을 살펴본 결과, 신생 사업체와 장수 사업체의 비중이 각각 24.56%와 16.25%를 차지하여 장수 사업체보다 신생 및 일반 사업체에서 생선회를 더 많이 취급하였다. 또한 생선회를 취급하는 사업체 중 영세 사업체의 비율은 90.77%로 전체 메뉴에 대한 영세 사업체의 비율(88.57%)보다 높았다. 따라서 대구광역시에서 생선회를 취급하는 사업체는 대부분 총 영업 년수가 10 년 미만인 영세 사업체로 구성되어 있으며 장수 사업체의 비중이 낮음을 알 수 있다.

다음으로 전국에서 가장 많이 취급되는 메뉴인 삼겹살구이에 대해 취급 사업체 수의 시도별 분포 및 특성을 살펴보았다(Figure 4.1). 삼겹살구이를 취급하는 사업체가 가장 많은 지역은 경기도, 서울특별시, 경상남도 등의 순으로 인구 밀도가 높을수록 사업체수가 증가하는 경향이 있다(Figure 4.1 의 (a)). 그러나 인구로 보정하여 인구 만 명당 취급 사업체수를 살펴본 결과 충청북도, 강원도, 대전광역시 등의 순으로 취급 사업체 수가 많았다(Figure 4.1의 (b)). 인구 대비 취급 사업체수가 많은 지역은 동종 사업체가 과밀된 지역으로 볼 수 있으므로, 이러한 정보는 동종 사업체의 창업을 억제하는 등의 과밀화를 해소 방안을 구상할 때 활용될 수 있을 것이다. 이때 인구가 상주인구만 포함할 뿐 유동인구는 포함하지 않는다는 점에 유의해야 한다. 예를 들어 강원도의 경우는 인구밀도가 낮고 관광지가 많이 분포해있다는 점을 고려할 때, 삼겹살 구이를 취급하는 사업체들이 상주인구보다는 관광객 등 유동인구를 주요 대상으로 영업하고 있는 것으로 유추된다.

삼겹살구이를 취급하는 사업체들의 특성을 알아보기 위하여 시도별 매출액과 사업이익 등을 비교하였다. 매출액과 사업이익은 유사한 경향을 나타내었으며, 매출액의 중위수와 표준편차를 Figure 4.1의 (c)와 (d)에 도시하였다. 서울특별시와 경기도를 포함한 수도권 일대와 광주광역시, 대전광역시, 대구광역시 등의 광역시 지역이 상대적으로 높은 매출을 보였고, 강원도와 경상북도가 낮은 매출을 보였다. 특히 취급 사업체 수가 많은 서울특별시와 경기도를 포함한 수도권과 인구 대비 취급 사업체 수가 많은 강원도는 취급 사업체의 특성이 대비되었다. 즉, 수도권의 경우 매출액과 사업이익의 중위수가 크지만 표준편차도 큰 것으로 보아 삼겹살구이를 취급하는 사업체간의 매출 및 사업이익 구조의 양극화가 큰 것으로 보이며, 강원도의 경우 매출액과 사업이익의 중위수와 표준편차가 모두 작아 영세 사업체들이 주류를 이루는 것으로 보인다.

**4.2.2. 사업체 유형별 인기 메뉴와 메뉴별 사업체 유형의 비중** 본 절에서는 사업체들을 특징에 따라 신생, 장수, 대형, 영세, 체인점가입 사업체로 유형을 분류하고, 특정 메뉴를 취급하는 사업체 중 해당 유형의 사업체 비중을 조사하여 비중이 높은 상위 10개 메뉴를 유형별로 알아보았다(Table 4.2). 총 261,045개의 한식 음식점 사업체를 '영업 기간' 별로 신생 사업체와 장수 사업체로 분류하면 각각의 비중이 전체 사업체의 25.96% 와 19.98%를 차지하였다. 신생 사업체의 비중이 높은 메뉴는 육회, 돈가스, 갈비찜, 족발 등이며, 특히 육회(49.5%)와 돈가스(44.6%)를 취급하는 사업체들 중 신생 사업체의 비중이 매우 높았다. 참고로 전체 사업체를 대상으로 구한 가장 인기 있는 메뉴인 삼겹살구이의 경우 신생 사업체의 비중이 27.01%이었다. 장수 사업체의 비중이 높은 메뉴는 닭백숙, 매운탕, 보신탕 등을 비

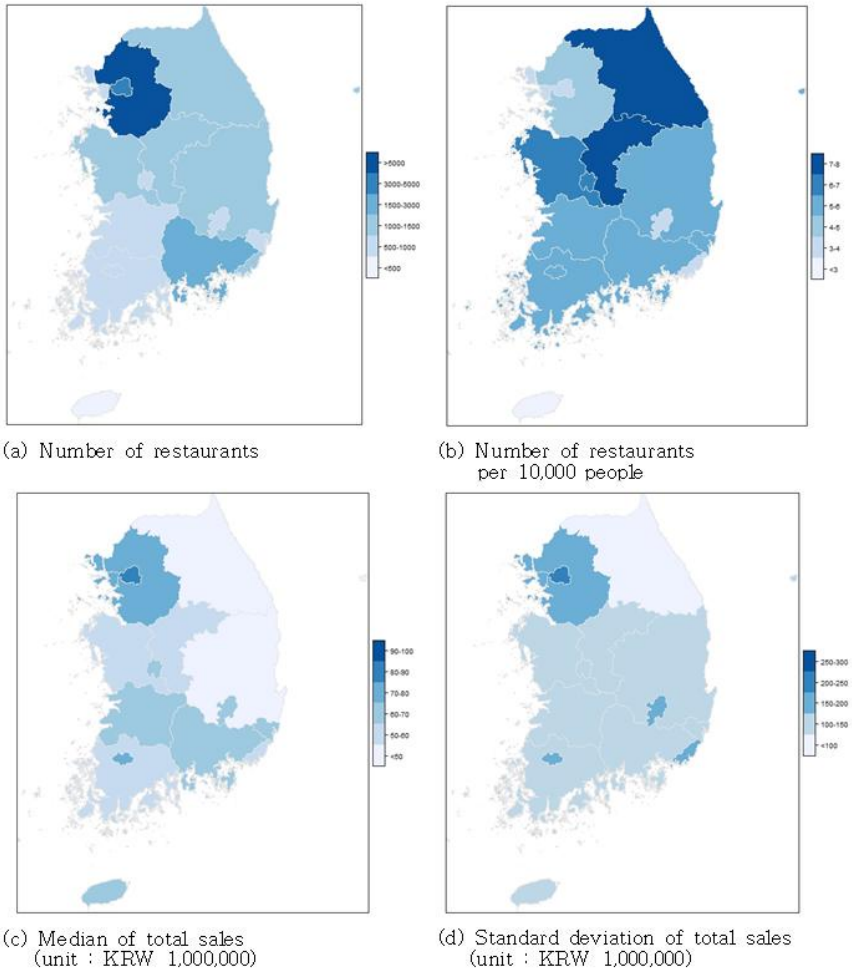


**Figure 4.1.** Distribution and characteristics of Korean restaurants offering Samgyupsal Gui on the menu during year 2010

롯한 탕, 국, 백숙 류가 주를 이루었다. 특히 닭을 재료로 한 메뉴들은 조리 방법에 따라 사업체의 유형별 비중이 달라지는 특징을 볼 수 있었는데, 닭백숙과 닭볶음탕은 장수 사업체의 비중이 높았고 닭갈비와 닭찜은 신생 사업체의 비중이 높았다. 다음으로 '사업체 규모'별로 유형을 나누면 전체 사업체의 대부분인 88.57%가 영세 사업체였으며, 대형 사업체의 비중은 11.43%에 불과하였다. 영세 사업체의 비중이 95% 이상으로 특히 높은 메뉴는 곱창, 닭볶음탕 등이 있었다. 뷔페, 샤브샤브, 소갈비 등의 메뉴는 취급 사업체 중 대형 사업체의 비중이 각각 50.3%, 41.8%, 30.9%로 매우 높았는데, 이들은 취급하는 사업체들의 매출과 사업이익도 높은 메뉴들이다. '체인점 가입 여부'에 따라 유형을 나누면 체인점 가입 사업체는 전체 사업체에서 6.15%로 그 비중이 미미하였다. 이에 비하면, 죽과 도시락의 경우 체인점 가입 사업체 비중이 50%를 넘어 매우 높은 수치이다. 특히, 죽을 취급하는 사업체들의 특징은 종사자수가 4인 이하의 영세 사업체로 체인점의 형태가 많음을 알 수 있는데, 이는 최근 소규모 체인점으로 창업하는 경우가 많기 때문인 것으로 보인다. 죽을 취급하는 사업체들의 사업이익을 살펴보면 다른 메뉴에 비

**Table 4.2.** Top 10 menus offered by Korean restaurants during year 2010 across restaurant types.

| 사업체 유형 | 신생 | 장수 | 대형 | 영세 | 체인점 가입 |
|---|---|---|---|---|---|
| 1위 | 육회 (49.51%) | 닭백숙 (36.82%) | 뷔페 (50.34%) | 곱창 (96.61%) | 죽 (56.22%) |
| 2위 | 돈가스 (44.57%) | 매운탕 (31.29%) | 샤브샤브 (41.83%) | 닭볶음탕 (96.37%) | 도시락 (50.55%) |
| 3위 | 갈비찜 (35.69%) | 보신탕 (30.57%) | 소갈비 (30.91%) | 보신탕 (96.18%) | 샤브샤브 (27.14%) |
| 4위 | 족발 (34.48%) | 막국수 (29.96%) | 한정식 (28.8%) | 소고기국밥 (95.95%) | 보쌈 (25.63%) |
| 5위 | 뷔페 (34.44%) | 오리탕 (29.87%) | 갈비탕 (24.18%) | 김치찌개 (95.55%) | 족발 (16.06%) |
| 6위 | 곱창 (33.26%) | 닭볶음탕 (29.08%) | 냉면 (23.19%) | 죽 (95.47%) | 부대찌개 (15.95%) |
| 7위 | 조개구이 (33.17%) | 오리백숙 (28.6%) | 오리고기구이 (19.61%) | 동태찌개 (95.3%) | 뼈다귀감자탕 (15.04%) |
| 8위 | 닭갈비 (32.64%) | 비빔밥 (28.09%) | 보쌈 (18.8%) | 돼지불고기 (95.26%) | 닭찜 (14.65%) |
| 9위 | 생선구이 (32.48%) | 육개장 (27.95%) | 돼지갈비 (18.03%) | 선짓국 (95.1%) | 육회 (14.4%) |
| 10위 | 닭찜 (31.74%) | 백반 (26.61%) | 설렁탕 (17.26%) | 된장찌개 (95.04%) | 낙지볶음 (12.85%) |

해 낮은 편이었다. 이에 비해 샤브샤브를 취급하는 사업체들은 체인점 가입 비중이 상대적으로 높은 대형 사업체라는 특성이 있으며, 이는 반조리 상태로 제공되어 조리 과정이 손님 앞에서 지속적으로 이루어지는 메뉴의 특성 상 종사자수가 큰 대형 사업체의 형태가 적합하기 때문인 것으로 보인다. 샤브샤브를 취급하는 사업체들의 사업이익은 상대적으로 높은 편이었다.

## 5. 결론

본 연구에서는 설문조사의 주관식 응답 형태로 조사된 텍스트 자료를 포함하는 2011 경제총조사에 텍스트 마이닝 기법을 적용하여 텍스트 자료와 구조화된 자료를 연계함으로써 한식 음식점업 사업체의 대표 메뉴의 현황 및 특성을 다각도로 탐색하였다. 또한 빅데이터인 경제총조사에 한국어 텍스트 마이닝을 적용함으로써 텍스트 기반 데이터베이스의 규모가 매우 큰 경우 발생하는 문제점들을 단계별 과정에 따라 관찰 및 정리하였다. 특히 설문조사의 주관식 응답 형태로 조사된 텍스트 자료라는 점에서 입력 과정상의 오타, 맞춤법의 오류, 잘못된 띄어쓰기 및 붙여쓰기 등의 문제가 빈번하게 발견되었고, 문법상의 오류는 없으나 동일한 개체에 대한 다양한 표기로 인해 빈도 분석의 정확도가 떨어지는 문제를 발견하였다. 본 연구에서는 이러한 문제점을 해결하기 위하여 자연어 처리를 중심으로 한 6가지의 자료 처리과정 규칙을 제안하였다. 제안한 규칙은 향후 설문조사의 주관식 응답을 기록하는 과정에 대한 표준화된 지침을 만드는데 기여할 것으로 기대한다. 자료 분석에서는 우리나라 전체 한식 음식점 사업체들이 취급하는 대표 메뉴에 대한 빈도 분석을 토대로 한식 음식점 사업체들의 현황을 시도별, 영업 기간별, 사업체 규모별, 체인점 가입 여부에 따라 다각적으로 탐색하였고, 매출액 및 사업이익에 대한 자료와 연계하여 현 시점에서 인기 메뉴의 현황과 메뉴별의 기대 매출액, 사업이익, 지역별 특징 등에 대한 정보를 도출하였다. 자료 분석의 결과로 얻어진 정보는 취급 메뉴의 변경을 고려하고 있는 현 업주 및 한식 음

식점 창업을 계획하는 예비 업주들에게 메뉴 선택의 가이드라인을 제시하는 데 기여할 것이라 예상한다. 또한 관련 정부 부처가 특정 메뉴를 취급하는 한식 음식점업체들의 과잉 경쟁 및 영세화 현황을 파악하고 이를 기반으로 영세 사업체들의 적절한 메뉴 변경 유도를 통한 폐업 방지 및 성공하는 창업을 위한 메뉴 선정의 유도 등의 정책을 마련하는데 도움이 될 것으로 기대된다. 본 연구에서는 텍스트 마이닝의 결과와 구조화된 데이터베이스의 연관성을 알아보기 위하여 중위수 및 표준편차 등의 기술 통계량만을 활용하였으나 후속 연구에서는 좀 더 다양한 데이터 마이닝 방법론이 활용되기를 기대해본다.

## References


Ahn, A. (2011). A Study of a Lexicon and Syntactic Patterns for an Automatic Classification of Korean Opinion Sentences, *Master Thesis*, Hankuk University of Foreign Studies.

Bartere, M. M. and Deshmukh, P. R. (2012). Cluster Oriented Image Retrieval System, *IJCA Proceedings on Emerging Trends in Computer Science and Information Technology (ETCSIT2012) etcsit1001*, ETCSIT**(3)**, 25–27.

Choi, S., Jeong, C., Choi, Y. and Myaeng, S. (2009). Relation extraction based on extended composite kernel using flat lexical features, *Journal of KIISE: Software and Applications*, **36(8)**, 642–652.

Choi, W. and Kim, D. (2009). A Study of Measuring Text Distances using the Hierarchical Clustering Method in Application to Pansori Narratives, *Seoul National University the Journal of Humanites*, **62**, 203–229.

Feldman, R. and Dagan, I. (1995). Forecasting item production with ARIMA model, *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, **KDD-95**, 112–117.

Ko, G., Jung, W., Shin, Y., Park, S. and Jang, D. (2011). A Study on Development of Patent Information Retrieval Using Textmining, *Journal of the Korea Academia-Industrial Cooperation Society*, **12(8)**, 3677–3688.

Hotho, A., A. Nurnberger, G. Paabß. (2005). A Brief Survey of Text Mining, *Ldv Forum*, **20(1)**, 19–62.

Jeong, D., Kim, J., Kim, G., Heo, J., On, B. and Kang, M. (2013). A Proposal of a Keyword Extraction System for Detecting Social Issues, *Journal of intelligence and information systems*, **19(3)**, 1–23.

Kam, M. and Song, M. (2012). A Study on Differences of Contents and Tones of Arguments among Newspapers Using Text Mining Analysis, *Journal of Intelligence and Information Systems*, **18(3)**, 53–77.

Lee, D., Yeon, J., Hwang, I. and Lee, S. (2010). KKMA : A Tool for Utilizing Sejong Corpus based on Relational Database, *Journal of KIISE: Computing Practices and Letters*, **16(11)**, 1046–1050.

Lee, H., Lee, J. and Lee, S. (1997). Noun Phrase Indexing using Clausal Segmentation, *Journal of KIISE: Software and Applications*, **24(3)**, 302–311.

Lin, X. (2003). Text-Mining Based Journal Splitting, *Proceedings of International Conference on Document Analysis and Recognition*, 1075–1079.

Mittermayer, M. and Knolmayer, G. (2006). Text Mining Systems for Market Response to News: A Survey, *Working paper in Institut fur Wirtschaftsinformatik der Universitat Bern*, **184**, 1–17.

SWRC (Semantic Web Research Center) (1999). HanNanum: Korean Morphological Analyzer, *Software*, Available from: http://semanticweb.kaist.ac.kr

van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G. and Leunissen, J. A. (2006). A Text-Mining Analysis of the Human Phenome, *European Journal of Human Genetics*, **14(50)**, 535–542.

Vijayarani, S. and Vinupriya, M. M. (2013). An Efficient Edge Detection Algorithm for Facial Images in Image Mining, *International Journal of Engineering Sciences & Research Technology*, **2(10)**, 2880–2884.

Yang, S. and Ko, Y. (2011). Extracting Comparative Elements for Korean Comparison Mining, *Journal of Korean Institute of Information Scientists and Engineers (KIISE): Software and Applications*, **38(12)**, 689–696.

# 한국어 텍스트 마이닝의 특성과 2011 한국 경제총조사 자료에의 응용

구주나[a,b] · 김경아[a,1]

[a]삼성서울병원 의생명정보센터, [b]숙명여자대학교 통계학과

(2014년 10월 14일 접수, 2014년 11월 12일 수정, 2014년 11월 21일 채택)

**요 약**

한국 전체 사업체 대한 최초의 전수조사인 2011 경제총조사 중 한식 음식점업 사업체 자료는 취급 메뉴에 대한 텍스트 자료와 영업 지역, 창립연월, 매출액 등 사업체의 특성을 나타내는 구조화 자료로 구성되어 있는 빅데이터이다. 본 연구에서는 취급 메뉴 자료에 텍스트 마이닝을 실시하는 과정에서 발생하는 통계 및 기술적 문제점들을 살펴보고, 이를 통해 한국어 텍스트 마이닝의 특징을 고찰하였다. 또한 텍스트 마이닝의 결과를 사업체 특성 자료와 결합하여 한식 메뉴와 이를 취급하는 사업체 특성 간의 연관성을 탐색하였다. 2010년 기준 가장 많은 사업체가 취급하는 인기 메뉴는 삼겹살구이로 특히 강원도와 대전광역시에 인구 대비 취급 사업체가 많았다. 신생 사업체의 인기 메뉴는 육회와 돈가스였고, 닭백숙과 매운탕 등이 장수 사업체가 많이 취급하는 메뉴였다. 이러한 결과들은 한식 음식점 창업 시 메뉴 선정 가이드라인으로 활용될 수 있으며 관련 정부 부처가 영세 사업체들의 메뉴 변경 유도를 통한 폐업 방지 등의 정책을 마련하는데 도움이 될 것이다.

주요용어: 텍스트 마이닝, 사전 구축, 빅데이터, 한국 경제총조사.

[1]교신저자: (135-710) 서울 강남구 일원로 81, 삼성서울병원 의생명정보센터, 연구조교수.
E-mail: kyunga.j.kim@gmail.com