발 간 등 록 번 호
11-1240245-000057-14

2023년 연구보고서

2025 인구총조사 산업·직업 분류항목 자동완성 기능 도입에 관한 연구

2024. 4.

http://kostat.go.kr/sri

ISSN 2288-1166(Print)
ISSN 2733-4120(Online)

통계청
통계개발원

연구보고서 2023-20

2025 인구총조사 산업·직업 분류항목 자동완성 기능 도입에 관한 연구

우찬균

발간사

기업 경영, 개인의 일상에 이르기까지 합리적 의사결정의 근간인 통계에 대한 중요성이 점점 커지고 활용범위도 넓어지고 있으며, 특히 국가통계는 정책결정에 필수적으로 활용되면서 그 중요성이 더욱 증대되고 있습니다.

이러한 시대의 변화에 따라 통계청은 빅데이터의 활용, 조사자료와 행정자료 간의 연계 등과 같은 통계생산방식의 혁신을 통해서 응답자 부담은 최소화하면서 동시에 보다 정확하고 사용자 친화적인 통계를 만들고자 끊임없이 노력하고 있습니다.

통계개발원은 국가통계의 중추를 담당하는 통계청의 싱크탱크로써 전략적인 연구를 수행하고 있는 국내의 유일한 「국가통계 전문연구기관」 입니다. 2006년에 설립된 이래 기존의 조사통계를 보다 효율적으로 작성하기 위한 각종 기법과 관련된 통계방법론적 연구는 물론 데이터에 기반한 국가정책이 수립될 수 있도록 경제·사회현상에 대한 심층 분석 연구를 강화하고 있습니다.

또한 저출산·고령사회 현상 등으로 인해 대내외적으로 관심이 높아지고 있는 인구집단 및 인구동향에 관한 분석연구 및 인구동태 관련 방법론 연구를 밀도 깊게 수행하고 있습니다. 이러한 연구의 구체적인 결과를 중심으로 통계개발원은 「2023년도 연구보고서」를 발간하게 되었습니다.

이번「2023년도 연구보고서」에는 AI 통계분류 결과분석 및 실무활용성 제고방안 연구 등 데이터과학 연구, 2025년도 인구주택총조사 등 조사표 개선 연구, 경제·사회·환경 변화를 반영한 인구통계, 격자통계를 활용한 도시화 현상 분석 등 인구통계 연구, 인구감소지역과 생활밀접업종 관계 분석 등 경제통계 연구, 비확률표본을 위한 통계적 추론 등 국가통계 방법론 연구, 위성영상을 활용한 국토그린지표 개발 기초연구 등 SDG 지표 관련 연구 등을 수록하고 있습니다.

본 연구보고서는 통계개발원이 전년에 국가통계 개선·개발을 위해 수행한 연구과제로서 국가통계 생산자의 통계개발 및 개선에 유용한 자료로 활용되고 될 수 있기를 기대합니다. 앞으로도 통계개발원이 “국가통계 전문연구기관”으로서 대내외적으로 선도적인 역할을 할 수 있도록 독자 여러분의 지속적인 관심을 부탁드립니다.

통계개발원은 본 연구보고서가 데이터 이용자의 통계 활용에 도움이 되고, 통계 작성자의 통계 개발 및 개선에 유용한 자료로 활용될 수 있기를 기대합니다. 앞으로도 국가통계의 통계연구에 대한 독자 여러분의 지속적인 관심을 부탁드립니다. 아울러 실용적이고 품질 높은 연구 결과를 도출하기 위해 최선을 다한 연구진에게 따스한 감사를 전합니다.

2024년 4월

통계개발원장

목 차

요 약

통계청에서 실시되는 인구총조사는 5년마다 전체 국민의 20%를 전수조사한다. 이때 많은 조사 항목들 중 유일하게 주관식 형태로 텍스트를 입력받는 문항이 바로 산업과 직업과 관련된 문항이다.

이렇게 조사된 산업과 직업 관련 문항은 조사가 완료되면 표준분류 숫자코드 형태로 변환을 해야 집계가 가능하기 때문에 조사가 끝나면 가장 먼저 표준분류 코드로 변환하는 작업을 시작한다.

인구총조사는 대규모 조사이기 때문에 조사 대상이 많고 응답자가 응답한 텍스트의 품질에 따라 코드로 변환하는 작업의 규모가 달라지게 된다. 뿐만 아니라 코로나 펜데믹 이후 비대면조사 선호현상으로 인해 조사원의 면접으로 진행되는 면접조사보다 직접 자기기입식 조사가 늘어나고 있다. 그런데 자기기입식 조사는 응답자가 텍스트로 응답한 내용을 자료처리 기간 전에는 알 수가 없다는 단점이 있다. 따라서 이러한 자기기입식 조사의 단점을 보완하고자 텍스트조사의 응답 품질을 높이는 방안을 이 연구에서 고민해 보았다.

AI통계분류시스템에서는 표준분류를 제공하는 것뿐만 아니라 조사항목에 맞는 키워드를 추천해 주는 자동완성 기능을 제공하고 있다. 그런데 이 자동완성 기능을 잘 이용하면 응답자의 응답 품질을 높일 수 있다. 그리고 미국 센서스국의 사례처럼 조사내용을 구체적으로 받을 수 있게 조사표를 구성하는 것도 응답 품질을 높이는 하나의 방법이지만, 이 방법은 응답자의 응답포기율이 높아지는 단점이 존재한다.

시험조사 결과 각 조사 항목에 맞는 키워드 선정과 키워드들을 어떤 우선순위에 따라 상위로 노출시킬 것인가가 가장 중요한 문제로 확인되었다.

데이터 분석 결과 대부분 단답형 단어 중심의 응답이기 때문에 분류의 특징을 가장 잘 나타내는 키워드 선정 및 추천이 중요하고, 직업문항의 경우 장기적으로는 직업이름을 묻는 방향으로 조사표의 변경이 필요해 보인다.

주요 용어 : 인구총조사, 산업분류, 직업분류, 자동완성, AI통계분류시스템

제 1 장

서 론

통계청은 통계법 제22조 1항에 의거하여 통계청장이 표준분류를 제정하게 되어 있다. 그리고 통계법 제22조 2항에 의거하여 통계작성기관의 장이 통계를 작성할 때 꼭 통계청에서 고시한 표준분류를 사용하게 되어있다. 따라서 표준분류는 통계청이 실시하는 대부분의 조사에서 사용이 되고 있다. 대표적으로 가장 많이 쓰이는 표준분류는 산업분류와 직업분류이다.

통계법이 제정된 시기의 조사방식은 대부분 종이조사방식이었다. 그래서 대부분 대면조사방식으로 조사가 진행되었고 응답자의 응답을 조사표 종이에 적는 형태였다. 표준분류는 사회가 복잡해지고 다양해지면서 분류가 세분화되어 가고 많아지기 시작했다. 산업분류의 경우는 이제 일반 전문가가 아닌 사람이 자신의 표준분류를 직접 선택하기에는 어려운 환경이 되었다. 그리하여 통계조사에서 표준분류를 응답자나 조사원이 선택을 하는 것이 아니라 표준분류 처리에 필요한 자료를 얻기 위한 항목이 조사내용에 포함된 현재의 조사방식이 되었다. 산업분류의 경우 가구 조사 시 사업체 이름, 사업체가 하는 일을 응답자에게 텍스트로 받아서 적는다. 직업분류의 경우는 회사에서의 직위, 부서명, 내가 한 일에 대한 텍스트를 받아서 적는다. 이러한 조사표의 응답내용을 가지고 분류전문가가 각 표준분류에 맞게 분류하게 되는데, 이러한 작업을 코딩이라고 한다. 분류코딩은 분류가 점점 복잡해지고 응답자나 조사원이 분류와 관련된 지식을 단시간에 습득하기 어려워서 조사 당시에는 조사표에 응답 항목들을 작성하고 이후에 분류 코드를 부여하는 방식으로 지금까지 진행되어 왔다.

이러한 코딩작업에는 고려해야 할 여러 가지 문제점이 발생하는데, 첫 번째로 가장 큰 문제는 시간이 오래 걸린다는 것이다. 특히 인구총조사와 같은 대규모 조사에서는 더욱더 많은 시간과 노력이 필요하다. 사회조사의 경우 조사 시 분류의 단위가 세분류 4자리인 경우가 많은데 산업분류와 직업분류는 490개가 넘는다. 이 분류의 기준을 모두 기억해서 조사 자료를 보고 바로 분류로 코딩을 할 수 있는 사람은 거의 없을 것이다. 응답자가 응답한 내용도 봐야 하고 응답내용에 문제가 있다면 보완도 해야 한다. 분류에 필요한 정보가 부족하다면 다른 조사항목의 성별, 나이, 지역, 학력 등을 고려해서 최종적으로 산업, 직업분류를 코딩하기 때문에 코딩작업은 시간

이 매우 오래 걸린다.

두 번째로는 코딩작업에는 인력과 돈이 많이 든다. 조사가 종료되면 바로 처리가 가능한 형태의 조사 항목들은 자료를 집계하거나 내검을 통해 처리 결과를 수정하기도 한다. 하지만 텍스트 조사인 산업, 직업분류 항목은 바로 집계를 할 수가 없다. 더군다나 자료처리 기간이 짧은 경우라면, 많은 사람이 투입이 되어서 분류코딩을 해야 한다.

마지막으로는 여러 사람이 코딩작업을 하다 보니 코딩 담당자들 간 의견이 달라 문제가 발생할 수 있다. 산업분류의 경우 동일한 텍스트 내용으로 조사되었다 하더라도 다른 요소들 때문에 분류가 달라지는 경우가 있다. 응답자의 학력, 지역, 기업의 매출액 등 다양한 자료를 참고하기 때문이다. 이러한 다양한 요소들을 확인하고 최종적으로 담당자의 판단에 의해서 분류코드가 부여가 되기 때문에 동일한 조사내용에서도 다른 의견이 나올 수 있다. 그래서 이러한 부분들을 보완하기 위해서 관리자 검토, 전산내검 도입 등 통계청에서는 다양한 노력을 하고 있다.

제 2 장

해외 사례 연구

제1절 미국사회조사 산업직업분류 코딩 프로세스

통계청에서 사용하고 있는 산업, 직업분류는 국제표준을 따르고 있다. 그래서 다른 나라의 분류체계와 비슷한 구조를 가지고 있기 때문에 다른 국가 통계기관들의 분류코딩 경험은 우리에게 아주 좋은 사례가 될 수 있다. 미국 센서스국의 산업분류(NAICS), 직업분류(SOC)도 한국 통계청의 분류체계와 동일한 계층형 체계를 가지고 있고, 조사의 역사가 우리보다 오래되었기 때문에 참고하기에 아주 좋은 사례이다.

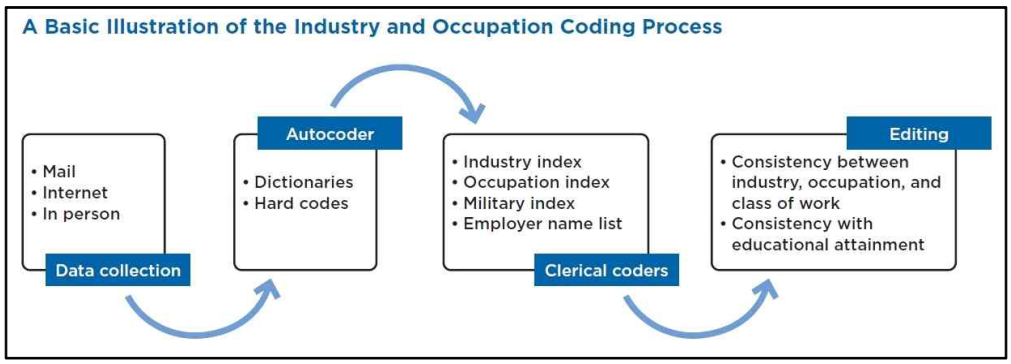
미국사회조사는 연간 350만 명을 조사하는 규모가 큰 조사이다. 사회조사이기 때문에 산업과 직업에 관련된 문항이 있고 이러한 조사 결과를 처리하기 위한 처리절차가 아주 잘 갖추어져 있다.

산업직업분류 코딩절차는 크게 4단계로 나뉘어 있다. 첫 번째로 데이터를 수집하는 단계이다. 데이터는 이메일이나 인터넷 면접원의 대면조사로 이루어진다.

두 번째로는 자동코딩 단계이다. Logistic Regression model을 통해서 1차적으로 자동분류를 하게 된다. 그리고 Dictionaries를 통해서 과거 조사에서 나왔던 키워드들을 색인하고 최종적으로 Quality Score를 부여하게 된다.

다음으로는 Quality Score가 낮은 조사 자료를 모아서 NPC(National Processing Center)로 보낸다. NPC에서는 품질점수가 낮은 조사 자료를 모아서 부가적인 자료들을 참고한 후 최종적으로 Clerical coders가 분류코드를 부여한다. 여기에서 참고하는 자료는 산업직업분류 색인어, 군 관련 색인어, 고용주 이름 리스트(Employer Name List)이며, 이를 사용해서 코드를 부여한다.

그리고 최종적 단계에서는 산업분류와 직업분류와의 관계, 교육 정도 등의 부가자료들도 활용하게 된다.



<그림 2-1> 미국사회조사 산업과 직업분류 코딩 프로세스

제2절 유사도 지수를 이용한 미국사회조사 직업분류

다음으로 미국 센서스국에서 발간한 워킹페이퍼를 통해서 직업을 분류하는 새로운 방법을 발견했다. “Using a Similarity Index to Understand the Measurement and Meaning of Occupations”라는 제목의 워킹페이퍼이고, 발간일이 2023년 9월로 아주 시의성이 있는 자료라고 볼 수 있다.

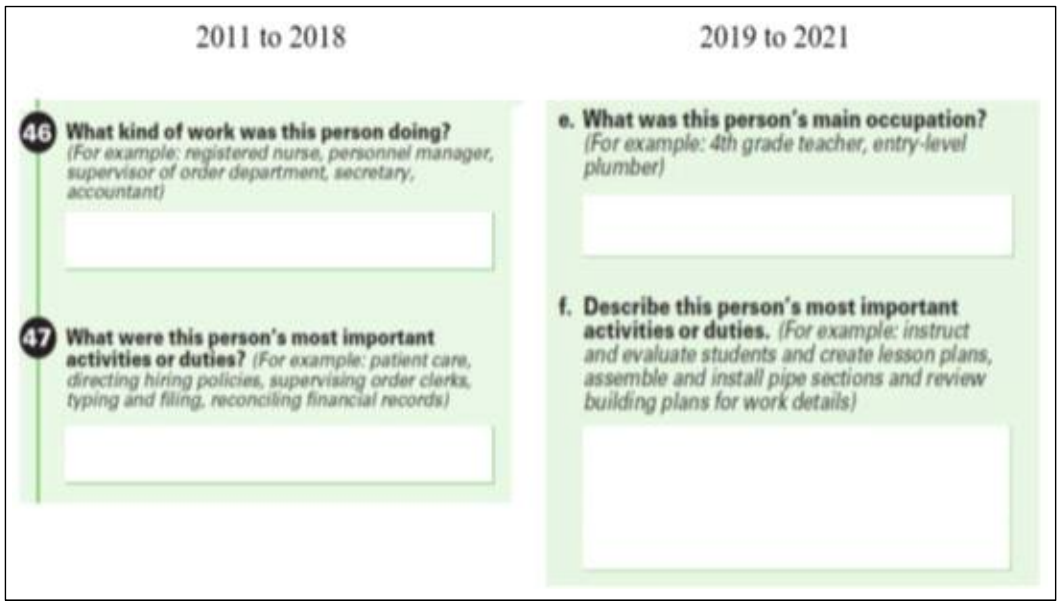
여기에는 조사 자료를 토대로 어떻게 직업을 분류해 내는지에 대한 실험과 설명이 자세하게 나와 있다. 직업의 경우 미국 사회조사에서는 직업 이름과 본인의 직업에 대한 설명을 자세한 응답 자료로 받아서 그것을 토대로 직업을 분류한다고 되어 있다.

실험에서 사용한 데이터는 미국사회조사 350만 가구의 데이터이다. 기간은 2011년부터 2021년까지의 자료를 사용했고 2013년과 2020년 자료는 제외가 되었다. 그리고 스페인어로 조사된 자료도 제외시켰다.

실험에서 가장 주목해야 할 부분은 바로 직업문항 질문의 변화이다. 2011년부터 2018년까지는 미국사회조사에서 ‘이 사람이 어떤 일을 하고 있었나요?’와 ‘이 사람의 가장 중요한 활동이나 임무는 무엇이였습니까?’라고 질문을 하였다. 이러한 질문들에 응답자는 첫 번째 질문에는 간호사, 인사관리자, 감독자, 회계사 등으로 응답하게 되어 있고, 두 번째 질문에는 환자치료, 채용정책 감독, 재정조정 등의 단답형으로 응답하게 되어 있다.

하지만 미국사회조사는 2019년 조사문항에 변화를 주었다. 지금까지는 어떤 일을 하는지 물어보았다면 2019년부터는 직업을 구체적으로 묻는 방식으로 조사 문항이 변경되었다. 그래서 ‘이 사람의 가장 주된 직업은 무엇이였나요?’와 ‘이 사람의 가장 중요한 활동이나 임무를 설명해 주세요.’로 직업 관련된 질문이 변경되었다. 이렇게

질문이 변경되면 첫 번째 질문에 응답자는 4학년 교사, 초급배관공 등 구체적인 직업을 응답하게 되고, 두 번째 질문에는 '학생들을 지도 및 평가하고 수업 계획을 작성합니다.' 또는 '파이프 섹션을 조립 설치하고 작업 세부 사항에 대한 계획을 검토합니다.'라고 답변을 하게 되었다. 질문의 변경에 있어 가장 큰 변화 중 하나는 바로 두 번째 문항에 길게 답변할 수 있도록 한 것인데, 과거에는 종이조사표에 1줄로 응답했던 내용을 3줄로 더 길게 응답할 수 있게 조사표를 변경했고, 온라인으로 응답할 때에는 60~100자를 작성할 수 있게 변화를 주었다.



<그림 2-2> 미국사회조사 직업문항의 변화

일반적인 직업분류의 방식과 다르게 워킹페이퍼에서 주목했던 부분은 바로 유사도이다. 유사도를 어떻게 측정을 해서 직업을 이해하고 분류했는지는 아래와 같은 절차들을 통해 확인할 수 있다.

1. 데이터 전처리

첫 번째로 해야 할 일은 데이터 전처리다. 오픈형으로 직업과 관련된 조사를 하기 때문에 응답자나 조사원이 여러 가지 실수를 할 가능성이 있다. 그리고 이러한 실수는 응답자가 의도하던 것이 아니기 때문에 그대로 놔둘 경우 원하지 않는 의미로 해석될 수도 있고, 잘못 해석될 수도 있다. 따라서 여러 가지 데이터 처리기법을 통해 텍스트에 있는 오류를 처리한 후 기계나 사람이 잘 이해할 수 있도록 변환해서 직업

유사도를 측정할 수 있게 만든다.

문법이 완벽하게 작성된 글을 만들려는 목적이 아닌 데이터 처리가 주된 목적이기 때문에 여러 다양한 기법이 존재한다. 대문자로 되어 있는 문자를 모두 소문자로 변경한다. 그리고 의미가 없는 단어는 제거한다. ‘director of operations’를 예로 들면, 여기에서 ‘of’는 없어도 의미를 파악하는데 지장이 없기 때문에 ‘director operations’로 변경한다. 그리고 스펠링이 잘못 작성되어 있는 단어나 문장도 올바른 스펠링으로 변경한다. ‘direct fund-raising’에서처럼 글을 쓸 때 사용되는 구두점 ‘-’을 ‘direct fund raising’과 같이 제거한다. 또, 컴퓨터로 글을 쓸 경우에 2칸 이상의 공백이 입력된 경우, 필요 없는 공백들은 모두 제거한다. 마지막으로는 Lemmatization을 한다. ‘medical assistant’를 ‘medical assist’로 변경해서 데이터가 잘 처리될 수 있게 전처리를 수행한다.

Table 1. Text Pre-processing Steps

Step	Before	After
Lowercase	REGISTERED NURSE	registered nurse
Remove stop words	director of operations	director operations
Correct misspellings	maintaince	maintenance
Remove punctuation	director fund-raising	director fundraising
Collapse multiple spaces	fire chief	fire chief
Lemmatize	medical assistant	medical assist

<그림 2-3> 텍스트 전처리 과정

2. 유사도 측정방법

페이퍼에서 사용한 유사도 측정방식은 코사인유사도(cosine similarity) 측정방식이다. 두 항목 간의 유사성 수준을 계산하여 텍스트 분석에 사용되는 측정값이다. 유사도 측정 대상인 직업이름과 직업설명 간의 유사도를 측정했다. 미국사회조사의 직업분류를 살펴보면 대부분 직업분류명이 직업이름으로 되어 있다. 그래서 직업분류에 있는 직업이름과 미국사회조사에서 응답자가 작성한 직업설명 간의 텍스트 유사도를 측정해서 유사성을 판단했다.

문장의 길이나 순서에는 상관 없이 단지 작성되어 있는 단어만 가지고 유사성을 판단했다. 측정값의 범위는 0~1 사이의 값이며, 0은 유사도가 가낭 낮고 1은 유사도가 가장 높은 값이다. 텍스트 간의 유사도 측정은 컴퓨터 프로그래밍상에서 처리하며, 언어 처리과정은 단어를 벡터로, 즉 숫자로 만들어 계산하여 이루어진다. 두 벡터의 내적값을 계산해서 단어 간의 유사도를 측정한 것이다. <그림 2-4>의 식에서 xi, xj는 조사에서 응답자가 응답한 텍스트 벡터이다.

$$sim(x_i, x_j) = cos(\theta) = \frac{x_i \cdot x_j}{\|x_i\| \, \|x_j\|}$$

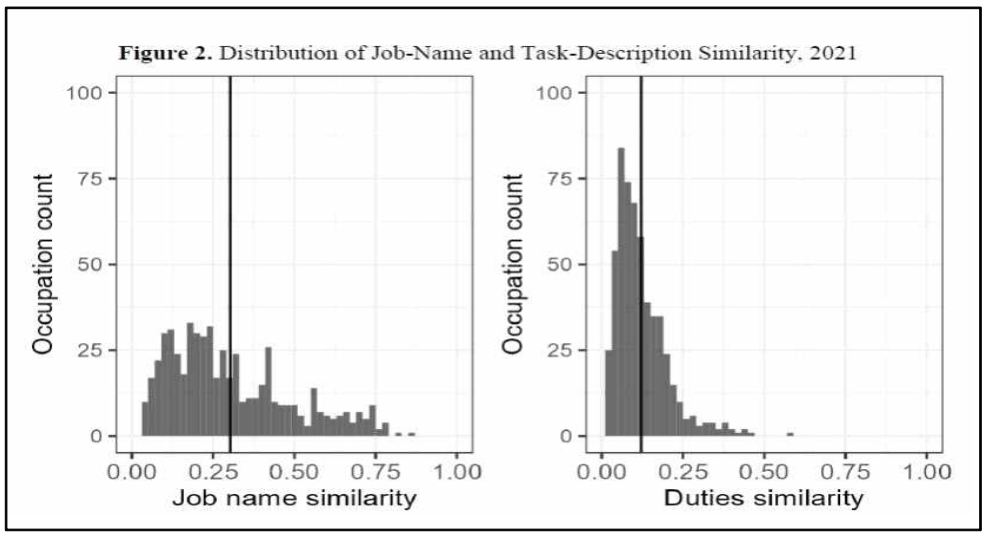
$$C = \frac{\sum_{i=1}^{P} \sum_{j=1, j \neq i}^{P} sim(x_i, x_j)}{|P|(|P|-1)}$$

<그림 2-4> 코사인 유사도 측정 산식

3. 직업이름과 직업설명 유사도 측정결과

2021년 미국사회조사 결과를 이용해 그래프를 만들어 보았다. 2021년은 조사표 문항이 변경된 해로, 변경된 직업문항 '이 사람의 주된 직업은 무엇이였습니까?'와 직업설명 문항 '이 사람의 가장 중요한 활동이나 임무를 설명해 주세요'에 대한 응답 결과를 분석한 것이다. 직업이름과 직업에 대한 설명 자료를 가지고 유사성을 비교했을 때 직업이름 유사도가 직업설명 유사도보다 더 유사성이 높은 것으로 나타났다. 유사도 측정이라는 것이 단어와 단어 간의 벡터값을 비교하기 때문에 아무래도 단어가 많은 직업설명 문항의 유사도가 직업문항에 비해 낮은 것이 정상이다.

실험 자료는 모두 전처리를 마무리한 데이터를 사용했으며, 소득이 0이거나 직업이 없다고 응답한 개인은 제외했다.



<그림 2-5> 직업이름과 직업설명 유사도 측정 결과

4. 유사도가 가장 높은 직업과 낮은 직업의 예

유사도를 측정한 결과 뚜렷한 특징들을 발견할 수 있었다. 직업이름은 승무원, 물리치료사, 실무간호사, 호흡기치료사와 같이 직업의 의미가 직관적이고 명확한 경우 유사도가 매우 높은 것으로 결과가 나왔다. 반면에 기타섬유, 의류, 가구근로자, 기타 자재 운반작업자와 같이 특징이 명확하지 않거나 기타로 분류가 된 경우는 유사도가 매우 낮은 것으로 결과가 나왔다.

직업설명은 페인트공, 우편배달원, 기계운영자가 가장 유사도가 높은 것으로 결과가 나왔다. 유사도 지수만 비교한다면 직업이름에 비해서 전체적으로 유사도는 낮지만, 위의 직업설명은 유사도가 높은 것으로 결과가 나왔다. 그리고 비즈니스 운영전문가, 기타 잦 운반근로자, 기타 모두 등 직업이름과 비슷하게 기타로 분류되는 직업설명은 유사도가 가장 낮은 것으로 결과가 나왔다.

Table 2. Examples of Most and Least Internally Similar Occupations, 2021

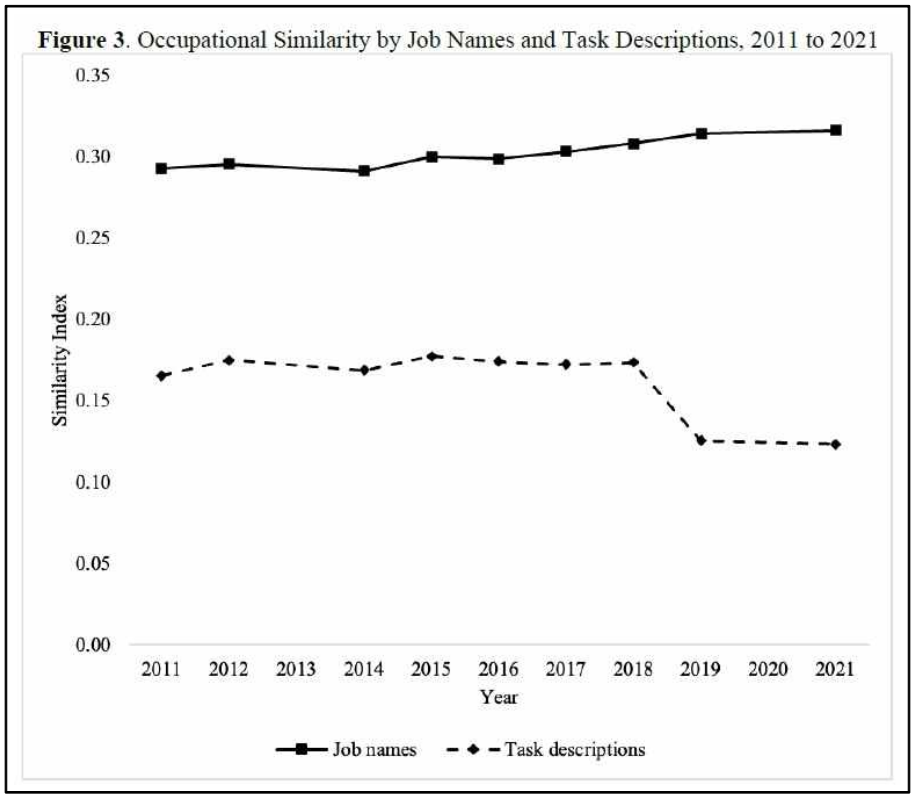
Job-Name Similarity		Task-Description Similarity	
Most similar			
Flight attendants	0.864	Postal service mail carriers*	0.573
Physical therapists*	0.814	Painters and paperhangers	0.469
Nurse practitioners	0.786	Postal service mail sorters, processors, and processing machine operators	0.448
Sales engineers	0.786	Dishwashers	0.441
Respiratory therapists	0.778	Meter readers, utilities	0.427
Least similar			
Other textile, apparel, and furnishings workers	0.031	Business operations specialists, all other*	0.013
Computer occupations, all other	0.038	Other material moving workers	0.015
Textile machine setters, operators, and tenders	0.041	Office and administrative support workers, all other	0.016
Other material moving workers	0.041	Agricultural and food science technicians	0.016
Other entertainment attendants and related workers*	0.045	Social science research assistants	0.020

<그림 2-6> 유사도가 가장 높은 직업과 낮은 직업의 예

5. 직업이름 및 직업설명 연도별 유사도 결과 비교(2011-2021)

직업이름과 직업설명에 대한 유사도의 시계열을 비교해 보았다. 직업이름 유사도는 매년 비슷한 추세를 보이는 반면 직업설명 유사도는 2019년에 가파르게 하락하는 현상을 볼 수 있는데, 2019년은 조사표가 변경되었던 해이다. 그러나 직업이름은 변

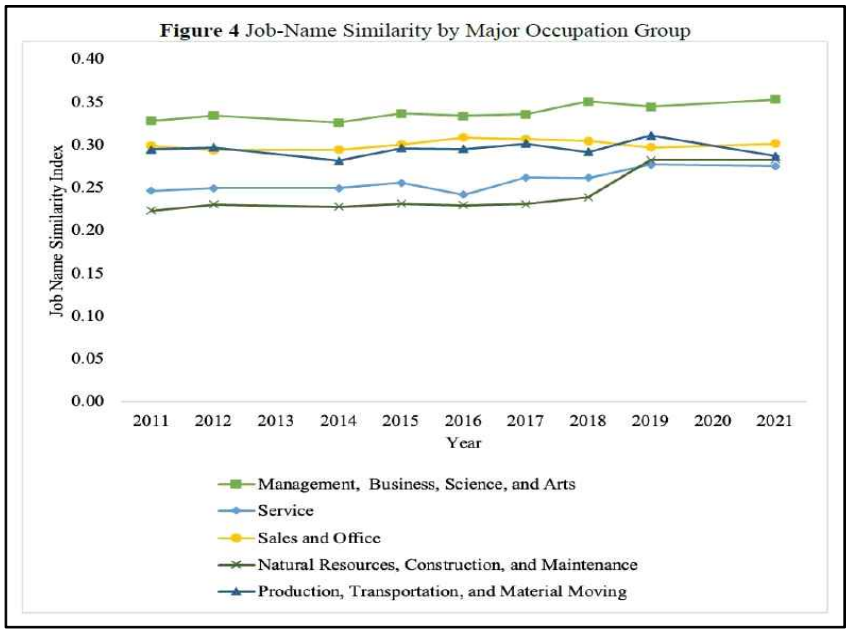
함없이 꾸준하게 약간의 상승세를 보인 반면 직업설명 유사도는 많이 하락했다. 하락의 원인 중 하나는 2019년부터 직업설명 문항의 길이를 늘렸기 때문으로 해석된다. 과거보다 더 많은 양의 단어들 사이의 유사도를 측정했기 때문에 다른 해에 비해서 더 많이 유사도가 하락한 것으로 보인다.



<그림 2-7> 직업이름과 직업설명 유사도 측정 결과 (2011~2021)

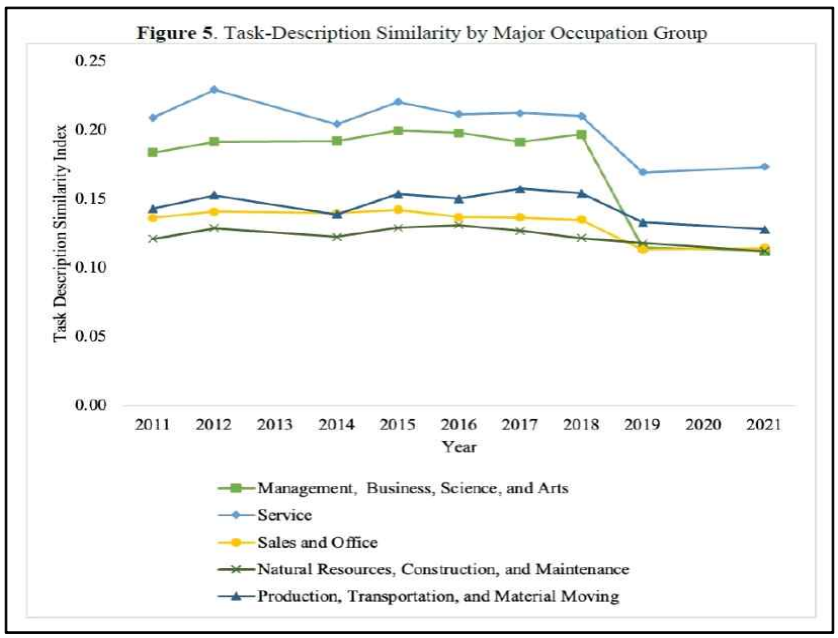
6. 주요 직업군별 직업이름, 직업설명 유사도

다음의 그래프는 5개 주요 직업이름에 대한 유사도를 보여주고 있다. 5개 그룹 중 서비스 직업을 뺀 4개의 그룹은 직업이름 유사도가 약간 증가했다. 2018년에는 “천연자원, 건설 및 유지관리” 직업이름 유사도가 가장 많이 증가했고, 2019년에는 “수학자”, “건축가”, “전기 기술자”, “임상 및 상담 심리학자”, “침술사”의 유사도가 가장 많이 증가했다. 이렇게 많이 증가한 원인인 질문 문구의 변경이 이러한 직업들의 유사도를 증가시켰다고 볼 수 있다.



<그림 2-8> 주요 직업군의 직업이름 유사도(2011~2021)

아래의 그래프는 5개 주요 직업설명에 대한 유사도를 보여주고 있다. 직업설명에 대한 유사도는 5개 그룹 모두에서 감소를 보인다. “경영, 비즈니스, 과학, 예술” 분야에서 가장 급격한 감소세를 보였고, 그다음으로는 “서비스” 관련 직업설명이 감소세를 이어갔다.



<그림 2-9> 주요 직업군의 직업설명 유사도(2011~2021)

제 3 장

AI통계분류시스템

제1절 자동완성이란?

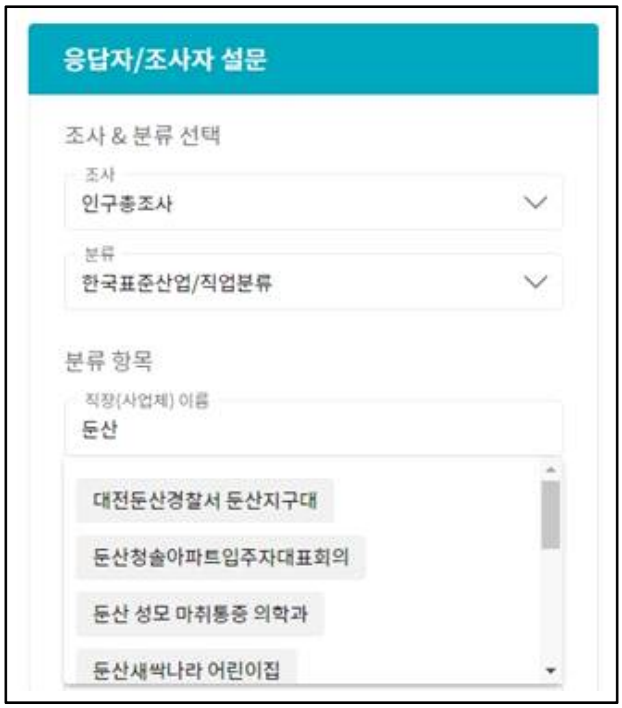
자동완성 기능은 인터넷이나 모바일 기기에서 이용자에게 입력 편의를 제공하는 기능이다. 대표적으로 많이 사용하는 인터넷 포털이나 검색엔진에 정보 입력 시 많이 경험해 보았을 것이다. 모든 단어를 입력하지 않아도 우리가 원하는 단어나 문장을 입력할 수 있고 영어의 스펠링이 생각이 나지 않을 때 추천되는 단어를 보면서 어렵지 않게 영어 단어도 입력할 수 있게 도움을 준다.

자동완성 기능을 이용하는 이유는 크게 두 가지로 나뉜다. 첫 번째로 입력의 편의성이다. 사람들은 본능적으로 귀찮은 걸 싫어한다. 우리가 원하는 정보를 얻기 위해서 현재 사용하는 디바이스의 입력 방식은 대부분 키보드 자판을 통한 입력이다. 인공지능 기술의 발달로 음성인식이나 이미지 검색 방식으로 원하는 정보를 검색하기도 하지만, 아직까지는 텍스트를 입력해서 원하는 정보를 얻는 경우가 가장 많다. 관련 정보를 디바이스상에 정확히 입력을 해야 원하는 정보를 얻을 수 있기 때문에 사람들은 컴퓨터나 테블릿의 키보드 자판을 통해서 정보를 입력하게 된다. 그런데 업무상 반복적으로 많은 정보를 찾아야 한다면 동일한 단어나 문자의 반복적인 입력이 비효율적이라는 것을 느낄 수 있을 것이다. 사람들의 이러한 불편함을 해소해 주기 위해서 자동완성 기술이 보편화된 것이다.

두 번째로는 입력 오류를 줄여주기 때문이다. 사람들은 키보드로 입력할 때 키보드 자판을 누르기 전 머릿속으로 무엇을 찾아야겠다고 생각하게 된다. 그런데 찾기로 한 단어가 생각이 잘 안 나거나 스펠링이 생각이 안 날 때가 있다. 이럴 때 자동완성 기능이 있는 포털에서 검색을 한다면 스펠링을 모르거나 단어가 생각이 안 나더라도 앞의 한두 자만 입력해 봐도 원하는 단어들의 예시가 나오기 때문에 그 예시를 보면서 진짜로 원하는 키워드를 쉽게 입력할 수 있다. 이러한 이유로 자동완성 관련 기술들을 검색회사에서 아주 널리 사용하게 된 것이다.

이러한 입력의 편의성과 편리함을 주는 기능을 인구총조사 전자조사표에 적용을 해 보자는 것이 본 연구의 목적이다. 본 연구는 이미 AI통계분류시스템에서 제공하

고 있는 자동완성 기술과 관련된 것이 아니다. 본 연구의 주된 목적은 서론에서 이야기한 것처럼 응답자 또는 조사원이 오픈형으로 조사문항을 입력받는 산업직업분류 관련 조사문항 입력 시 편의를 제공하고 오류를 줄이는 것에 있다. 그러기 위해서 AI통계분류시스템에서 어떠한 방식으로 자동완성 기능이 제공되는지를 확인해 본다.



<그림 3-1> AI 통계분류시스템 자동완성 기능

제2절 엘라스틱서치(Elasticsearch)

1. 엘라스틱서치

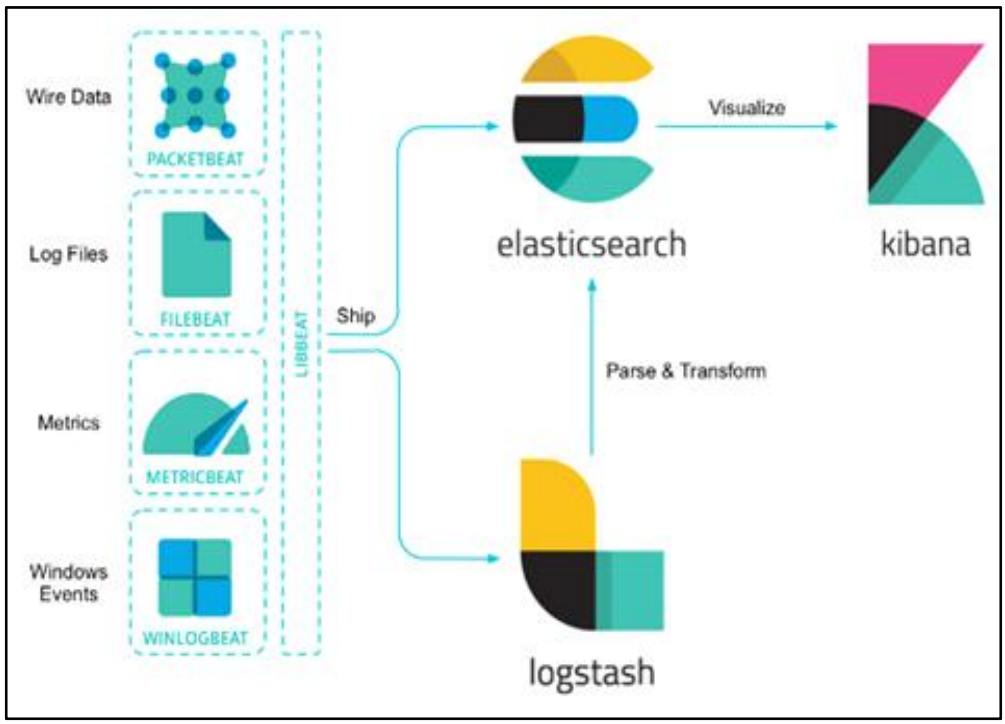
AI통계분류시스템은 자동완성 기능을 위해 엘라스틱서치 검색엔진을 사용했다. 엘라스틱서치는 아파치 루씬 기반의 무료 검색 및 분석엔진이다. 텍스트 분석에 특화된 검색엔진이라 주로 웹 검색에 사용된다. 엘라스틱서치에서는 다양한 기능을 제공하지만 본 연구에서 주목한 것은 어플리케이션 검색 기능이다.

검색엔진을 사용하지 않는다면 보통은 RDBMS를 사용하게 된다. RDBMS에서는

주로 like 문을 이용해서 검색을 한다. like로 검색할 경우 검색 대상의 데이터가 많다면 속도가 느리고 동의어나 유의어에 대한 검색은 되지 않는다.

엘라스틱서치의 특징은 크게 4가지로 구분이 된다. 첫 번째로 가장 중요한 기능은 전문검색(Full Text Search)이다. 자동완성 기능을 사용하는 사용자가 입력기에 한 자 한 자 타이핑할 때마다 거기에 맞는 추천 단어가 화면에 제시되어야 한다. 그런데 아주 빠르게 자동완성이 제시되지 않는다면, 사용자는 추천 단어가 나오기 전에 이미 텍스트 입력을 마칠 수도 있는 것이다. 두 번째 기능은 스키마리스이다. 스키마가 없어도 스스로 데이터를 분석해서 필드를 생성하고 저장하는 기능이 있다. 세 번째는 RESTful API 기능의 제공이다. REST API를 통해서 다른 시스템 간에 엘라스틱서치 기능을 제공할 수 있다. 마지막으로는 역색인 구조의 제공이다. 데이터가 들어오면 미리 색인을 해 놓기 때문에 검색어의 입력 시 빠른 검색이 가능하다.

엘라스틱서치의 단점도 존재한다. RDBMS와는 다르게 실시간(Real Time) 처리는 불가능하다. 데이터가 변경이 되었을 때 반영되기까지 1초 정도 차이가 나서 Near Real Time을 제공한다고 한다. 그리고 트랜젝션이나 롤백 기능을 제공하지 않는다. 그래서 데이터 관리에 더 관심을 가져야 하는 단점이 있다.



<그림 3-2> 엘라스틱서치 구조도

2. 역색인 구조

엘라스틱서치 역색인 구조는 검색을 위한 아주 중요한 부분이다. 이 구조는 데이터를 검색하기 위해서 아주 효율적으로 구성이 되어 있어 빠르고 정확한 검색이 가능하게 된다. 역색인 구조는 텍스트를 검색하기 위해서 사용이 되며 문서에 있는 단어를 효율적으로 찾기 위한 구조이다. 여기에서 '역'이란 용어는 단어에 해당하는 색인키(Key)를 의미하는 것이다. 역색인 구조를 사용하면 특정 단어가 포함된 문서를 아주 빠르게 찾을 수 있으며 이를 통해서 검색 속도를 빠르게 하고 정확한 검색을 제공해 주게 된다. 다음은 역색인 구조의 구성요소이다.

<표 3-1> 역색인 구조의 구성요소

기능	설명
어휘 사전 (Vocabulary)	엘라스틱서치는 각 단어를 어휘에 저장한다. 이 사전은 문서 내 모든 고유한 단어들의 목록을 가지고 있다.
포스팅 리스트 (Posting List)	각 단어는 해당 단어를 포함하는 문서의 목록을 가지고 있다. 이를 포스팅이라고 하는데 포스팅 리스트는 문서 ID나 위치 정보와 같은 추가 정보를 포함할 수 있다.
텀 (Term)	텍스트에서 추출한 단어를 텀이라고 한다. 텀은 어휘 사전에 등록이 되며 해당 텀이 어떤 문서에서 어떤 위치에 나타나는지에 대한 정보가 포스팅 리스트에 저장이 된다.
문서ID와 텀 위치정보	포스팅 리스트는 특정 텀이 나타나는 문서의 ID를 기록하고 필요한 경우 위치정보도 기록할 수 있다. 이는 특정 단어를 검색할 때 해당 단어를 포함하는 문서를 빠르게 찾을 수 있도록 도와준다.

3. 애널라이저(analyzer)

엘라스틱서치에서 애널라이저는 텍스트데이터를 색인할 수 있는 핵심요소 중 하나이다. 애널라이저는 텍스트를 분석 후 어휘 사전을 구축하고 역색인 구조를 생성한다. 이렇게 색인을 만들어 놓은 데이터는 이후에 데이터를 검색할 때 빠르고 효율적인 검색을 가능하게 해 준다. 애널라이저는 세 가지로 구성되어 있는데 다음의 표와 같은 특징이 있다.

<표 3-2> 에널라이저의 세 가지 구성요소

구성요소	내용
문자필터 (Character Filters)	문자필터는 텍스트를 색인화하기 전에 문자열을 정제하거나 변환하는 데 사용한다. 특수문자를 제거하거나 문자열을 다른 문자열로 변경할 수 있다.
토크나이저 (Tokenizer)	토크나아저는 문자열을 토큰(token)으로 분할하는 역할을 한다. 토큰은 어휘 사전 단어를 사용하고 검색의 기본단위가 된다. 토큰을 나눌 수 있는 단위는 공백이나 구두점 기준으로 한다.
토큰필터 (Token Filters)	토큰필터는 생성된 토큰을 변경하거나 추가 작업을 진행할 때 사용된다. 형태소 분석, 불용어 제거(stop words), 대소문자 변환 등의 작업이 토큰 필터를 통해서 이루어진다.

제3절 AI통계분류시스템 엘라스틱서치

1. AI통계분류시스템 애널라이저

AI통계분류시스템에서는 전자조사표에서 자동완성 기능을 사용하기 위해서 애널라이저를 커스터마이징해서 만들었다. 기본적으로 자동완성 기능을 제공해 주는 애널라이저 외에 초성검색 기능 제공을 위한 애널라이저, 키워드 앞부분에 좀 더 가중치를 부여해 주기 위한 애널라이저도 만들었다. AI통계분류시스템에서 사용되고 있는 애널라이저는 아래의 표와 같다.

<표 3-3> AI통계분류시스템 에널라이저의 기능

애널라이저	기능
ac_search_analyzer	자동완성 검색
ac_index_analyzer	자동완성 인덱싱
chosung_search_analyzer	초성 검색
chosung_index_analyzer	초성 인덱싱
ac_index_sub_analyzer	키워드 앞부분 일치에 대한 가중치 부여
chosung_front_analyzer	초성 앞부분 일치에 대한 가중치 부여
title_nori_analyzer	형태소 분석

애널라이저 내에 토큰을 처리하는 토큰필터도 만들어서 추가적인 데이터를 처리하고 있다. 사용되고 있는 필터는 한글 자모를 분리해 주는 한글 자모 분리 필터(JamoFilter), 초성을 분리해 주는 초성 분리 필터(ChosungFilter), 한글을 영어로 변경해 주는 필터(HanToEngFilter), 영어를 한글로 변경해 주는 필터 (EngToHanFilter)가 있다.

2. AI통계분류시스템 데이터 저장구조

AI통계분류시스템은 오라클 데이터베이스(RDB)를 쓰고 있다. 자동완성 사전에 쓰이는 원데이터로 조사시스템의 자료를 가져오기 때문에 오라클 데이터베이스에 원자료를 저장하고 오라클에 있는 자료를 불러와서 엘라스틱서치의 색인작업을 통해서 저장하게 된다. 데이터 저장 구조는 자동완성 사전의 사용 목적에 맞게 설계가 되어 있고 그 구조에 따라 색인을 하게 되어 있다.

```
{
  "actvtnYn": "Y", // 활성화 여부
  "ctgrySn": 11, // 카테고리 (시소러스사전 카테고리정보)
  "atcptType": "20", // 10: 랜딩(홈), 20: 단건&API
  "count": 3, // 검색수
  "id": "OG_20_((주))삼성전자", // 자체 부여 id
  "entityTag": "OG", // TAG (ex) OG: Organization
  "word": "((주))삼성전자" // 워드정보
}
```

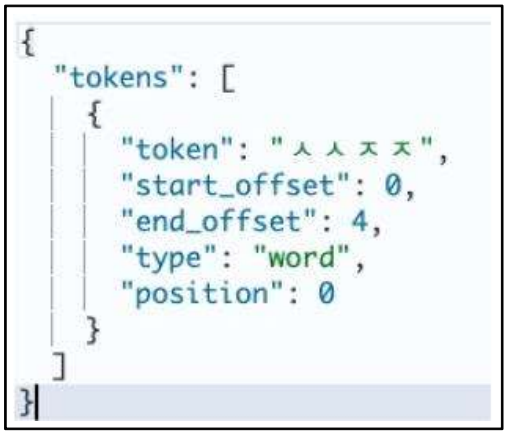
<그림 3-3> 자동완성 기능 데이터 저장구조

3. AI통계분류시스템 인덱싱 저장 구조

인덱싱은 역색인 구조를 만들어서 검색을 빠르게 하는 데 목적이 있다. 그래서 자동완성 사전에 단어를 추가하면 색인을 만들어 놓는데, AI통계분류시스템은 두 가지로 인덱싱을 한다. 첫 번째로는 초성검색 기능을 지원하기 위해서 초성인덱싱을 하게 된다. 예를 들어 “삼성전자” 라는 키워드를 입력할 때 “ㅅㅅㅈㅈ”만 입력을 해도 “삼성전자” 단어를 추천해 줄 수 있게 미리 색인을 해 놓는 것이다.

```
GET local_atcpt_word_230517/_analyze
{
  "analyzer": "chosung_index_analyzer",
  "text": "삼성전자"
}
```

<그림 3-4> 초성검색 애널라이저

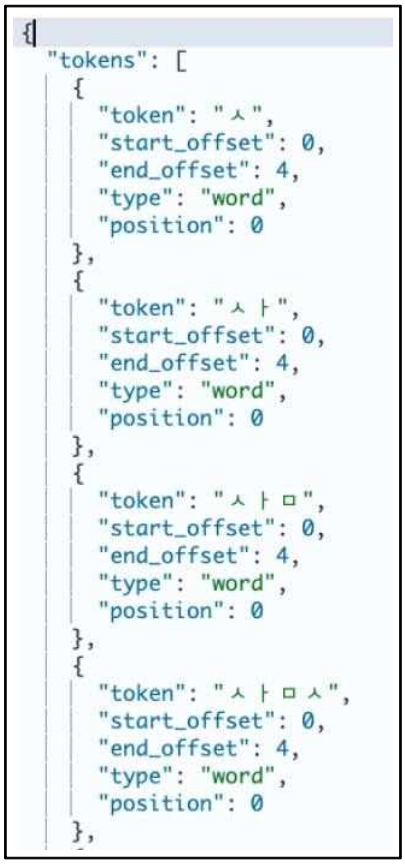


<그림 3-5> 초성검색 인덱싱

다음으로는 자동완성 인덱싱을 만들어 놓는다. 동일하게 “삼성전자”로 예를 들어 보겠다. “삼성전자”를 이용자가 입력할 때에는 ㅅ, 사, 삼, 삼ㅅ, 삼서, 삼성... 이러한 식으로 한글을 순차적으로 입력한다. 자동완성 기능에서는 많은 글자를 타이핑하지 않아도 입력하려는 글자의 입력을 편하게 하고 타이핑 절약를 위해서 사전에 미리 그림과 같이 인덱싱을 해 놓아서 사용자의 입력을 도와줄 수 있다.

```
GET local_atcpt_word_230517/_analyze
{
  "analyzer": "ac_index_analyzer",
  "text": "삼성전자"
}
```

<그림 3-6> 자동완성 애널라이저



<그림 3-7> 자동완성 인덱싱

제4절 AI통계분류시스템 자동완성사전 추가

2025 인구총조사 2차 시험조사는 2025 인구총조사의 성공적인 추진을 위해서 시스템, 업무 프로세스, 조사표 등 총조사 전반적인 과정을 종합적으로 검토하는 시험조사이다. 시스템이 새로운 차세대 시스템으로 변경되면서 AI통계분류시스템 API 서비스를 연계 개발해서 기능적으로 자동완성 기능을 사용할 수 있는 환경이다.

하지만 AI통계분류시스템 개발 당시의 자동완성 사전은 인구총조사에 맞는 데이터가 부족한 상태이다. 자동완성 사전은 각 조사에 맞는 키워드를 등록해서 해당 키워드가 조사표상에서 추천될 수 있게 설계되어야 한다. 따라서 2차 조사에서 데이터 사전의 활용성 검토를 위해 연구 기간에 자동완성 사전을 추가하였다.

자동완성 사전에 등록하는 자료로는 기본적으로 정제 과정을 거친 과거의 조사 자료를 사용하는데 조사 자료에는 전처리 후에도 예상치 못한 단어들이 존재할 수 있기에, 비교적 정제가 되어 있을 것으로 예상되는 전국사업체조사 자료를 활용했다.

전국사업체조사는 인구총조사와는 달리 사업체조사로 조사문항이 다르다. 그래서 모든 문항의 자동완성 사전을 추가할 수가 없었고 그 중 분류에 있어서 가장 중요한 항목인 “사업체 이름” 항목의 자동완성 사전만 추가하기로 했다.

데이터는 2021년,2022년 전국사업체조사 자료 중 “사업체 이름” 항목을 사용하였다. 전체 자료 중에 중복을 제거한 후 데이터를 추출했는데 예상했던 것보다 특수기호가 많아서 전처리 과정에 많은 시간이 소요가 되었다.

전처리 방법은 주로 특수기호를 처리하는 방법으로 진행했다. 첫 번째로는 추천되는 단어에 오타가 있으면 안 되기 때문에 한글 오타가 있는 자료들을 제거했다. 다음으로는 전각 문자를 반각문자로 변환했다. 전각문자는 키보드를 이용해서 작성하기 어렵기 때문에 추천되는 단어에서 필요가 없다. 다음으로는 괄호가 한쌍을 이루지 않거나 괄호 안에 데이터가 없는 경우 제거했다. 그다음으로는 “%, ·(가운뎃점), ‘(인용부호)” 등 다양한 기호들도 자동완성 추천단어에는 적절하지 않아 삭제 처리했다. 마지막으로는 자, 모가 분리된 한글문자를 추출했다. 자, 모가 분리된 문자의 경우 오타로 인식하는 경우가 많아 모두 정상적인 문자로 수정해서 데이터를 업데이트해 주었다.

이렇게 전처리가 완료된 자료는 AI통계분류시스템 시소러스 지식관리에 등록을 하고 엘라스틱서치 검색엔진에 색인을 거쳐 API를 통해서 자동완성 기능에서 사용할 수 있도록 등록을 마쳤다.

제 4 장

결론 및 시사점

자동완성 기능은 응답자에게 타이핑을 절약해 주고 입력문자의 정확도를 높여주는 아주 유용한 기능으로 AI통계분류시스템에서는 자동완성 사전을 관리하면서 이 기능을 필요로 하는 시스템에 API 형태로 서비스를 제공해 주고 있다. 연구 결과 시스템으로 기능을 제공해 주고 있지만 좀 더 보완해야 할 부분이 확인되었다.

수많은 단어가 자동완성 사전에 등록되어 있지만, 해당 조사에 맞는 키워드 가중치가 없어서 응답자가 원하는 단어가 상위에 노출되지 않고 입력한 비슷한 단어가 상위에 노출되었다. 이러한 유용한 기능을 조사에 활용하기 전에 이러한 부분들은 정비가 되어야 한다.

게다가 응답자마다 원하는 키워드 우선순위가 다를 수 있으니 입력 당시 다른 부가정보들도 참고해서 키워드를 추천해 준다면 좀 더 완성된 자동완성 서비스를 제공할 수 있을 것이다.

다음으로는 우려가 되는 사항이다. 자동완성 기능은 분명 응답을 도와주는 좋은 기능이지만 잘못 사용하게 된다면 조사의 품질을 떨어뜨릴 수 있다. 응답자가 무응답을 피하기 위한 수단으로 아무 키워드나 또는 상위 노출된 키워드로 응답할 가능성이 있다. 따라서 이러한 우려를 보완하기 위해서 자기기입식 전자조사의 경우 조사 입력시간 등 파라데이터를 수집하고 조사 신뢰도 자료로 활용해서 자동완성 기능 오남용 사례를 파악하고 시정해야 할 것이다.

다음으로는 데이터분석과 해외사례를 분석한 결과 인구총조사 산업, 직업 질문 문항이 좀 더 구체적일 필요성이 제기된다. 특히 직업의 경우 부서나 직위 정보가 없는 응답자가 많다. 부서나 직위 정보가 없을 경우 응답자가 하는 일에 대한 작성만으로는 그 응답자의 직업분류를 코딩하기는 쉽지 않아 보인다.

그리고 표준분류에는 많은 분류가 있지만 대부분의 응답자가 해당되는 분류는 상위 10위 안에 있는 분류가 대부분인 것으로 파악이 되었다. 게다가 상위 분류는 특징이 명확하고 관련 키워드가 많지 않아서 이러한 분류와 관련된 키워드를 모아서 조사를 한다면 추후에 분류코딩 내검 시 많은 시간을 절약할 수 있을 것이다.

참고문헌

권철민. (2020). **파이썬 머신러닝 완벽가이드.**

웨스 맥키니. (2019). **파이썬 라이브러리를 활용한 데이터 분석**.

유경준. (2016). “제10차개정 한국표준산업분류”.

전창욱, 최태균, 조중현,.신성진. (2020). **텐서플로2와 머신러닝으로 시작하는 자연어 처리.**

통계정책국 통계기준과. (2017). “제10차개정 한국표준산업분류”.

통계청. “제7차개정 한국표준직업분류”.

Julia B. Beckhusen. (March 2020). “Recent Change in the Census Industry and Occupation Classification Systems”, U.S. Census bureau.

Ananda Martin-Caughey. (September 19,2023). “Using a Similarity Index to Understand the Measurement and Meaning of Occupations”, U.S. Census bureau.

OpenAI. (2021). ChatGPT https://chat.openai.com/

Abstract

A study on the introduction of automatic completion function of the industry and occupation classifications to the 2025 Population Census

Chankyun Woo

Statistics Korea carries out the Population Census every 5 years, which covers 20% of the total population. In the census, many survey items are investigated, and survey items filled out only in a subjective form are those related to industry and occupation.

These survey items can be aggregated only after being converted to numeric codes of the standard classification after the census is completed. Therefore, during the first business process, these survey items should be converted into standard classification codes.

Because the Population Census, which is a large-scale survey, covers a huge number of survey targets, the scale of conversion into codes varies, depending on the quality of responses. Besides, due to the preference for non-face-to-face interviewing after the COVID-19, the self-interviewing method by respondents is more widely applied than the interviewing method by enumerators. The disadvantage of the self-interviewing method is that the respondents' text responses cannot be known before the data processing. Therefore, this study aims to improve the quality of text responses to compensate for the shortcomings of the self-interviewing method.

The AI statistical classification system not only provides the standard classification, but also provides an automatic completion function that recommends keywords matching the survey items. Through a smart use of this automatic completion function, the quality of responses can be improved. As in the case of the United States Census Bureau, composing the questionnaire so that responses can be filled out in detail can be an option, but it has the disadvantage of increasing the non-response rate.

The result of the pilot survey showed that the selection of keywords appropriate for each survey item was the most important and which keywords were exposed at the top position was the most important issue.

The result of data analysis shows that most of the responses are word-centered short answer type. Therefore, it is important to select and recommend keywords that best represent the characteristics of the classification. In the case of occupation-related questions, it is necessary to change the questionnaire to ask for occupation names in the long term.

Key words: Population Census, industrial classification, occupational classification, automatic completion, AI statistical classification system

연구진

○ 우찬균 (통계청 통계개발원 통계방법연구실 주무관)

* 연구진의 소속 및 직급은 연구과제 완료 시 기준임을 알려드립니다.

연구보고서 2023-20

인구총조사 산업·직업 분류항목 자동완성 기능 도입에 관한 연구

인 쇄	2024년 4월
발 행	2024년 4월
발 행 인	박상영
발 행 처	통계청 통계개발원 35220 대전광역시 서구 한밭대로 713 TEL.(042)366-7100 FAX.(042)366-7123
홈페이지	http://sri.kostat.go.kr
ISSN(Online)	2733-4120

통계청
통계개발원