통계정보 분류의 자동코딩 성능 실험 연구*

An Experimental Study on the Automatic Coding System for Statistical Information Classification in Korea

남영준(Young-Jun Nam)**, 안동언(Dong-Ein Ahn)***

목 차

초 록

인구센서스와 같은 국가 통계정보는 국가의 미래 투자계획과 정책수립을 위한 중요한 기초데이터이다. 그러나 데이터의 코딩과정이 모두 수작업으로 이루어지기 때문에 결과의 일관성 결여와 시간과 인력이 너무 많이 소요된다는 것 등이 문제점으로 지적되고 있다. 따라서 본 연구에서는 한국 산업표준 분류표에 근거한 자동코딩시스템을 개발하여 코딩과정을 수작업으로 처리할 때 발생하는 문제점을 해결하였다. 시스템의 지식베이스로는 학습이론을 사용하여 저자가 새로이 개발한 복수의 전거어 사전들을 활용하였다. 실험한 결과, 생성률은 99.5%를, 정확률은 83.3%라는 결과를 얻었다 따라서 이 시스템은 실제 통계데이터의 자동코딩과정에 사용될 수 있으며, 국가 통계정보의 효율적 분석에 매우 유용하게 사용될 수 있을 것이다.

ABSTRACT

National statistical data such as Korean Census is fundamental data for national administration. In this paper, we present an automatic coding system utilizing morphological analyser and knowledge dictionaries. Knowledge bases are constructed based on an authority dictionaries which were developed by authors utilizing a newly learning theory. Test data indicates 99.5% of productivity and 83.3% of accuracy. The presented methods can be effectively applied to analyze statistical information.

키워드: 통계정보, 분류, 자동코딩 프로그램, 전거어 사전, 한국표준산업분류표

* 이 연구는 1999년도 전주대학교 학술연구지원금에 의해 연구되었음
** 전주대학교 사회과학부 부교수
*** 전북대학교 전자정보공학부 조교수
■ 논문 접수일 : 2000년 10월 28일

1 머리말

선진국에서는 국가의 정책수립과 미래국가경영 방향 등을 결정하기 위해 계량화된 기초통계데이터를 필요로 한다. 이를 위해 각 국은 정부 차원에서 정기적으로 인구센서스를 실시하여 기초통계데이터를 수집하고 있다. 우리 나라도 국가 경영을 위한 기초통계데이터를 수집하기 위해 통계청이 주관이 되어 5년마다 인구주택 총조사를 전국적으로 실시하고 있다. 데이터의 수집은 전문조사요원들이 각 가정을 방문하여 면담방식으로 이루어진다. 수집된 데이터 가운데 산업 및 직업관련 항목은 각각 한국 표준산업분류표와 한국 표준 직업분류표에 근거하여 분류번호를 수작업으로 부여하고 있다. 그러나 조사항목수의 증가와 인구수의 증가로 인하여 코딩에 소요되는 절대시간이 늘어나게 되었다. 이는 통계데이터수집에 상대적으로 많은 경비와 인력이 소요된다는 점과 통계데이터의 수집과 가공에 절대시간이 부족함에 따라 적기에 국가경영에 필요한 데이터를 제공하지 못하는 점(time-lag)이 문제로 지적되고 있다. 또한, 다양한 코딩전문요원의 지적 배경으로 인해 코딩결과의 일관성이 결여될 수 있어 기초 통계데이터로써 분석결과에 의문이 제기될 수 있다는 점도 지적되고 있다. 특히, 조사항목 가운데 산업과 직업조사 관련 항목은 비정형화된 데이터로써 이를 분석하고 분류하는데 많은 시간과 경비가 소요되어 이 항목들에 대한 자동화가 요구되고 있다. 따라서 본 연구에서는 두 항목 가운데 우리 나라 인구주택 총조사에서 조사된 데이터 가운데 산업 관련항목을 국가경영에 필요한 데이터로 가공하는데 유용하게 사용될 수 있는 자동코딩 시스템과 알고리듬을 개발하고자 한다.

자동 코딩시스템은 키워드를 추출하기 위한 프로그램과 효율적인 코딩결과를 보장하는 지식베이스로 구성된다. 본 시스템에서 사용된 프로그램으로는 문장분석용 형태소 해석기와 색인기 등이다 지식베이스로는 한국 표준 산업분류표에 열거된 명사(구)를 기반하여 개발된 산업 일반 전거어 사전과 전문 조사요원이 조사한 데이터에 나타난 용어에 기반하여 개발된 산업 공동 전거어 사전을 사용한다. 특히, 산업 공동 전거어 사전은 전문용어보다는 실제 생활에서 사용되고 있는 일반용어들과 표현 방식을 중심으로 수집하여 개발하고자 한다. 이 사전에는 전문용어보다는 국민들이 현장에서 사용하고 있는 일반명사와 파생용언들이 상대적으로 많이 등록된다. 지식베이스의 확장을 위해 본 연구에서는 학습이론을 도입하여 데이터 생성의 오류결과를 용어풀에 저장하고 일정 기준에 도달할 때에 이를 공동전거어 사전에 등재하는 방법을 사용한다. 실험 대상은 2000년 4월에 통계청이 실시한 산업조사 데이터 가운데 통계청에서 수작업으로 코딩한 6,428건을 사용한다. 실험결과는 전문코딩 요원들이 수작업으로 코딩한 결과와 본 연구에서 개발한 시스템으로 코딩된 결과를 비교하여 생성률과 정확률로 시스템의 효율성을 평가한다.

2 한국표준산업분류의 자동코딩

우리 나라에서는 5년을 주기로 하여 인구주택 총조사를 전국적으로 실시한다. 이를 통해 국가는 우리 나라의 모든 인구와 주택의 총수는 물론 개별 특성까지 파악하여 각종 경제 사

회 발전계획의 수립 및 평가와 각종 학술연구, 민간부문의 경영계획수립에 필요한 기본 자료를 입수한다. 조사 범위는 인구와 가구, 주택에 관한 항목을 전수와 표본으로 나누어 총 50개 항목을 조사한다. 2000년 인구주택 총조사에는 인구에 관해 29개 (전수 8개, 표본 21개) 항목, 가구에 관해 16개(전수 7개, 표본 9개) 항목, 그리고 주택에 관해 5개(전수 5개) 항목 등 총 50개 항목을 조사한다. 인구주택 총조사는 조사원들이 조사대상 가구나 사업체를 방문하여 필요한 자료들을 수집하는 면담방식으로 이루어진다. 조사항목은 크게 정형화된 항목과 비정형화된 항목으로 구분될 수 있다. 예를 들면, 성별과 생년월일 등과 같은 항목은 정형화 항목이기 때문에 구술(국민)과 기술(조사자)에 있어 거의 오류가 발생하지 않는다. 그러나 산업과 직업 항목은 구술자의 주관이나 기술자의 주관에 따라 매우 비정형화된 데이터로 표현이 되기 때문에 데이터의 오류발생 확률이 상대적으로 높은 편이다. 한편, 국가 전략상 필요한 데이터는 대부분 비정형화된 데이터가 많이 활용되며, 국가 거시지표 등을 수립하는데 필요한 데이터로 가공하기 위해서 분류 코드를 전문적으로 부여(코딩)하는 지역별 전문코딩요원들이 한국표준산업분류와 한국표준직업분류에 근거하여 해당항목들을 수작업으로 코딩하고 있다. 이 코딩은 전문조사요원들이 수집한 직업/산업 필드에 기록된 내용에 근거하여 한국 산업/표준분류표에 열거된 해당 분류번호를 부여하는 행위이다. 코딩요원들은 통계청 지방 사무소에 상주하는 직원들이며 코딩을 위한 자체 교육을 받고 2개월에 걸쳐 코딩작업을 수행한다. 국가통계를 위해 산업 및 직업을 코딩하는 이유는 국내 근로자들이 수행하는 생산활동을 경제적 성질의 유사성에 따라 체계적으로 분류함으로써 산업구조 파악은 물론 생산성, 경쟁력 등 경제분석을 위한 기본정보를 입수할 수 있기 때문이다(통계청, 2000). 또한 분류에 대한 일관성 있는 데이터와 국가간 비교할 수 있는 데이터를 확보하여 일반행정 및 산업정책관련 법령에서 그 법령의 적용대상과 산업영역을 한정하는 기준으로 활용할 있는 기본 정보를 입수할 수 있기 때문이다. 이를 위해 통계청에서는 인구주택 총조사 과정에서 다음과 같이 산업과 직업관련 항목을 조사하고 있다.[1)]

- 산업 : 직장 사업체 이름/주된 사업 내용
- 직업 : 부서 및 직책/하고 있는 일의 종류
  예) oo 화학 영등포 공장 / 가정용 플라스틱 탱크(용기) 성형 제조기술 개발부 과장 / 플라스틱 제품을 위한 원료 시험 분석

2.1 한국표준산업분류표의 구조

대한민국 통계청에서 활용하고 있는 한국표준산업분류표(이하 산업 분류표)는 5단계 계층구조를 갖고 있다. 각 계층별 항목은 대분류(20), 중분류(63), 소분류(194), 세분류(442), 세세분류(1,121)로 구성되어 있다. 괄호 안의 숫자는 분류항목의 숫자이며, 산업 분류표에 열거된 분류항 가운데 대분류 전체와 중분류의 첫 번째에 열거된 항목을 살펴보면 다음과 같다.

A 농업, 수렵업 및 임업
  01 농업 외 1개항

1) 통계청, 2000년 인구주택총조사표, 직업 및 산업관련항목의 예시

B 어업
  05 어업
C 광업
  10 석탄, 원유 및 우라늄 광업 외 2개항
D 제조업
  15 음·식료품 제조업 외 22개항
E 전기, 가스 및 수도사업
  40 전기, 가스 및 증기업 외 1개항
F 건설업
  45 종합 건설업 외 1개항
G 도매 및 소매업
  51 도매 및 상품 중개업 외 1개항
H 숙박 및 음식점업
  55 숙박 및 음식점업
I 운수업
  60 육상운송 및 파이프라인 운송업 외 3개항
J 통신업
  64 통신업
K 금융 및 보험업
  65 금융업 외 2개항
L 부동산 및 임대업
  70 부동산업 외 1개항
M 사업서비스업
  72 정보처리 및 기타 컴퓨터 운영 관련업 외 3개항
N 공공행정, 국방 및 사회보장행정
  76 공공행정, 국방 및 사회보장행정
O 교육 서비스업
  80 교육 서비스업
P 보건 및 사회복지사업
  85 보건업 외 1개항
Q 오락, 문화 및 운동관련 서비스업
  87 영화, 방송 및 공연산업 외 1개항
R 기타 공공, 수리 및 개인 서비스업
  90 하수처리, 폐기물처리 및 청소관련 서비스업 외 3개항
S 가사 서비스업
  95 가사 서비스업
T 국제 및 외국기관
  99 국제 및 외국기관

2.2 한국표준산업분류표의 특징

통계청의 산업분류표는 총 1,398개의 분류항을 갖고 있으며, 이 분류항은 기존의 도서관에서 사용하는 분류표의 항목과는 달리 모분류(상위항목)와 자분류(하위항목)에 기술된 표목들이 동일한 구절 혹은 단어로 구성이 되어 분류항이 반복되는 특성을 갖고 있다. 예를 들면, 농업항목 가운데 "작물재배 및 축산 복합농업"은 소분류, 세분류, 세세분류의 분류표목이 다음과 같이 각 계층별로 반복되어 나타난다.

013 작물재배 및 축산 복합농업
0130 작물재배 및 축산 복합농업
01300 작물재배 및 축산 복합농업

이와 같이 분류항이 반복되어 출현하기 때문에 전문조사요원 혹은 분류담당자가 조사된 데이터를 구별하기 위해서는 해당 분류항목의 해설을 참조해야 한다. 그러나 한국 산업분류표에는 일부 분류항에 대해서만 해설이 기술되어 있기 때문에 반복되어 출현한 분류항에 대해서는 조사원의 주관에 따라 상위개념번호와 하위개념번호의 부여가 결정된다. 다음은 하위 분류항목에 대한 해설만이 주어진 예이다. 즉, 다음과 같이 상위개념번호(013과 0130)에 대한 해설은 전혀 없으며 최하위 개념어(01300)에만 해설이 되어있다.

013 작물재배 및 축산 복합농업
0130 작물재배 및 축산 복합농업
01300 작물재배 및 축산 복합농업
작물 재배활동과 축산활동을 복합적으로 운영하는 산업활동으로서 이중 한편의 전문화율이 66% 미만으로 운영되는 산업활동을 말한다
〈예 시〉
· 작물과 동물의 복합생산

이와 같은 분류항의 중복과 해설문의 생략은 궁극적으로 산업분류번호의 부여가 완전히 조사자의 주관적인 판단에 의해 이루어질 수 있음을 의미한다. 이는 분류의 목적 가운데 일관

성에 배치되는 것으로서 궁극적으로는 국가 통계조사의 부정확성을 초래할 수 있다. 이러한 주관적인 수작업 코딩과정에서 나타나는 문제점을 제시하면 다음과 같다(충남대학교 소프트웨어연구소, 1999: 6-8).

① 구축되는 자료의 정확성이 떨어진다.

계층분류표를 이용한 코딩작업에서는 분류전문가의 수준에 따라 분류의 질이 결정된다(Ruiz, Miguel E., Srinivasan, Padmini, 1999: 281) 따라서 코딩 작업자들이 이 전체를 이해하고 파악하는 것은 어렵기 때문에, 수작업으로 코딩할 때에 전문코딩 요원의 기억력에 의존한 부정확한 코딩이 이루어질 수 있다.

② 구축되는 자료의 일관성 유지가 어렵다.

자료가 방대하기 때문에 여러 사람이 작업을 하게 되므로 같은 산업내용에 대해서 각자가 정확한 분류를 찾지 못한다면 동일한 사실을 놓고 조사원들간에 서로 상이한 코드가 부여될 수 있다. 즉 코딩 전문요원의 산업과 직업에 대한 선입관과 관점에 따라 코딩 결과가 군집화되지 않을 수가 있다. 궁극적으로 일관성을 유지하지 못하여 구축된 데이터베이스가 일관성이 결여된다.

③ 데이터 가공의 지연(time-lag)이 발생한다.

수작업은 자동 방식에 비해 코딩과정에 막대한 시간과 인력이 요구된다. 이는 국가 정책을 수립하는데 있어서 최단기간에 조사 분석결과가 제시되어야 함에도 불구하고 코딩이 수작업으로 이루어지기 때문에 상대적으로 많은 시간이 소요된다. 이는 적기에 국가 통계데이터가 제시되지 못하는 결과가 초래되어 인구주택 총조사자체가 무의미하게 될 수 있다.

따라서 이러한 문제점들을 해결하기 위해 산업 분류의 코딩 자동화 시스템 개발이 필요하며, 이를 통해 다음과 같은 효과를 거둘 수 있다.

· 통계 자료의 정확성 향상 : 수작업 코드 부여시 발생할 수 있는 부정확성을 개선하는 효과가 있다.

· 통계 자료의 일관성 보장 : 객관적이며 기계적인 판단 규칙에 따라 동일한 표현은 항상 동일한 코딩 결과를 보장한다.

· 통계 자료의 신뢰성 확보 : 통계 자료의 정확성과 일관성을 통하여 국가 통계 자료의 신뢰성을 확보할 수 있다.

· 시간과 인력 절감을 통한 비용 절감

· 과거 자료의 소급 변환 가능

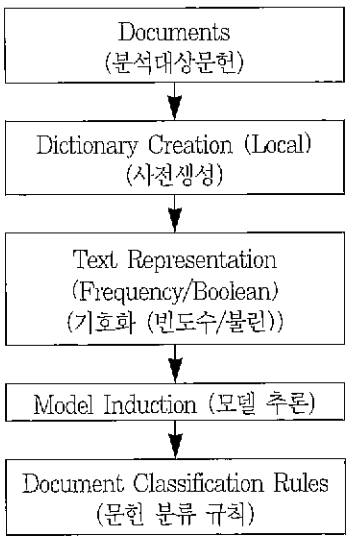
궁극적으로 신뢰성 있는 국가 통계 데이터를 확보함으로써 국가를 경영하는데 있어서 근거를 가지고 국가정책을 수립할 수 있다. 또한, 계획성 있는 투자를 통하여 산업발전의 고도화에 이바지할 수 있다. 대외적으로는 대외 신인도 향상에도 기여할 수 있다.

2.3 기존 자동코딩의 검토

자동코딩과 자동분류는 입력된 데이터를 기준이 되는 분류표의 구조에 따라 정확하게 배정하는 것을 목적으로 한다. 이 두 방법의 가장 큰 차이점은 자동코딩은 산업분류표에 열거된 분류번호만을 사용해야 하나, 자동분류의 경우는 분류표를 근거로 보조기호표를 이용하여 본표에 열거되지 않은 분류번호를 분류자의 주관에 따라 나름대로 조합이 가능하여 본표에 없는 분류번호가 처리대상자료에 부여될 수 있다. 또한 자동분류의 경우는 통계적 기준에 중점을 두고 있는 반면 자동코딩은 어의적 분석에 중점을 두고 있다. 따라서 자동분류는 서명을 포함한 초록, 본문과 같이 분석대상문헌의 내용을

중심으로 분석이 이루어지기 때문에 통계적 알고리듬을 주로 활용하고 있으며, 자동코딩의 경우는 정형된 필드에 정해진 길이의 제한된 정보만으로 분석이 이루어지기 때문에 어의적 알고리듬이 활용된다. 자동분류에 대한 선행연구는 문헌자료를 대상으로 이루어진 것이 대부분이다. Diane Vizine-Goetz 등은(Vizine-Goetz, Diane and Jean Godby 1996, 127-135) DDC의 본표를 활용하여 인터넷에 존재한 문서들을 DDC 항목에만 있는 분류항으로 분류하는데 한계를 보이고 있는 점에 착안하여 LCSH(LC Subject Headings)을 보조 분류기호로 활용한 자동분류를 실험하였다. 분류표를 활용한 자동 분류 실험은 본표에 열거된 분류번호에 한정하여 실질적으로 수작업으로 구분표를 이용한 논리조합을 통한 분류작업과는 큰 차이를 보이고 있다. 즉, 기존의 분류표를 활용한 자동분류 연구 가운데 본표상에 열거된 분류항을 대상으로 새로이 입수되는 자료가 어디에 배정되는지를 확인하는 연구가 주로 이루어지기 때문에 자동코딩의 알고리듬과 유사해짐을 알 수 있다. 한편, 대부분의 자동분류는 일반적으로 학습(learning) 알고리듬을 이용하여 〈그림 1〉과 같은 순서로 분류작업을 수행한다(Apte, Chidanand and Damerau, Fred, Weiss, Sholorm M. 1994. 24). 분류번호 부여(기호화 : Text representation)를 위해서는 주로 통계정보를 활용하며, 이를 위해서는 적정한 규모의 용어가 추출될 수 있는 분석대상을 필요로 한다.

이에 비해 자동코딩은 조사 항목의 일부 필드에 기재된 정보에 의존하는 것이 일반적이기 때문에 충분한 통계정보를 입수하지 못하게 된다. 이러한 제한된 정보만으로 보다 정확한 코딩결과를 얻기 위해 미국은 통계국(U.S. Census Bureau)주관으로 체계적인 통계정보의 분류를 목적으로 자동코딩화 사업(The Automated Industry and Occupation Coding System; AIOCS)을 추진하고 실제 2000년 센서스에 적용될 수 있는 체계를 수립하기 위한 실험을 실시하였다(Gillman, D. W. and Appel, M V. 1994). 미국 통계국에서 직업 및 산업분류에 사용한 자동코딩 알고리즘은 오류에 기반한 학습과 지식베이스(직업/산업분류표 항목 데이터베이스와 전거어 사전)의 활용으로 오류율을 개선하는 방식이다. 즉, 생성이 이루어지지 않은 데이터의 경우에 그 패턴을 조사하여 차후 분석에서는 생성의 오류가 이루어지도록 않도록 설계하였다. 미국은 1990년 인구센서스데이터를 대상으로 한 분석결과에서 산업분류코딩은 57%, 직업분류코딩에서는 37%의 생성률을 보였으며, 생성된 데이터 가운데 잘못 분류된 것은 산업분류코딩에서는



〈그림 1〉 자동 분류 순서도

6.2%가, 직업분류코딩에서는 11.8%의 결과를 보이고 있다.

프랑스 통계청(INSEE, French National Institute of Statistics)은 1982년부터 QUID 소프트웨어를 이용하여 직업분류코딩에 자동화를 시험하였다. 이 시험에서 사업의 효용성과 정확률을 높이기 위해 여섯 개의 모듈로 이루어진 SICORE라는 프로그램을 사용하였으며, 모듈 중 '참조파일'과 '코드화된 파일의 기록파일'을 지식베이스로 사용하는 새로운 방법을 사용하였다(Schuhl, Pierrette, 1996).

가. 참조파일

나. 자동코딩을 위해 문장분석을 비롯한 언어처리용 규칙

다. 변수를 수용할 수 있는 논리규칙(logical rule)

라. 변환규칙: 논리규칙을 변수로 이해할 수 있는 변환규칙

마. 코드화된 파일의 기록파일

바. 학습알고리즘의 파라미터

이 시스템은 stemming 알고리듬, 학습이론, 전거어 사전의 활용 및 문장 분석 규칙도 사용하였다. 프랑스는 자동코딩실험에서 약 66%의 생성률을, 90% 이상의 정확률(4자리수)을 제시하였다. 한편, 2자리수의 코딩은 95% 이상의 효율을 보이고 있어 4자리수에 비해 높은 코딩결과를 보여 복잡한 분류는 정확률이 저하되고 있었다.

일본은 자동코딩에 대한 연구를 1차로 국가차원에서 실시하였으며, 2차로 위탁연구방식으로 외부기관과의 공동 연구를 수행하였다. 공동연구는 일본 통계정보연구개발센터에서 산업분류에 대한 자동코딩이 가능한지를 연구하고, 그의 효용성에 대해서도 연구하였다.

동 센터와 일본전기(NEC)의 공동연구에서는 구조화된 분류표를 기반으로 표제항과 일치하는 정확한 산업분류코드를 부여하기 위해서 단어일치와 상품명(고유명사) 등을 활용한 여러 분석기법 등을 사용하였다(米澤, 1998). 각 알고리듬은 조사된 데이터의 수준에 따라 서로 다르게 적용하였으며, 그 가운데 '상품명을 표기하지 않은 전문일치방식'으로 산업분류코드 실험대상이 되고 있는 조사표 가운데 1/3이 99%의 정확률을 나타내어 가장 높은 정확률을 보였다. 100%가 되지 않은 이유는 조사원의 직접 기입에 따른 입력 오류에 기인한 것으로 판단하고 있다. 이러한 입력오류가 나타나는 것은 다음과 같은 원인에 기인한다고 분석하였다.

가. 조사표의 기입내용과 사업소의 실체가 상이하며 현장의 조사담당자가 수기한 표현대로 기록하였기 때문이다.

나. 기입내용을 분석할 수 있는 분석대상정보의 절대량이 작았기 때문이다.

다. 겸업을 할 경우 자동부여가 불가능하기 때문이다.

연구결과, 생성된 통계데이터의 자동코딩은 대체적으로 모든 방식이 센터에서 의도한 추천기준치보다 자동코딩에 따른 효율이 높았으나, 이를 산업분류에 적용하기에는 미흡한 점이 많다고 판단하여 실용화를 연기하였다. 왜냐하면, 실험이 생성된 1/3의 데이터를 기준으로 하였기 때문에 실제 생성률을 고려할 경우에 자동코딩효율은 33% 미만이기 때문이다. 그러나 이 연구에서는 여러 실험조건 가운데에서 보조키워드(전거어 및 유사어 사전)를 사용하였을 경우에 가장 높은 생성률과 정확률을 얻을 수 있음을 주장하고 있었다.

3 자동코딩 프로그램의 설계

인구주택조사표에서 산업 및 직업 조사 항목은 피조사대상인 주민들이 구술한 것을 조사원들이 이를 하나의 문장이나 혹은 어절로 기술하도록 구성되어 있다. 산업 및 직업에 관한 필드명은 〈표 1〉과 같은 구조를 갖고 있다.

실제로 조사과정에서 기술된 데이터의 예를 들면 아래와 같다.

- 사업체명의 예 : "우리밭에서 농업"
- 사업내용의 예 : "정수기 앤드 가정용품 등/가지고다니며 판매"
- 부서 및 직책 : "가정을 방문하면서 피아노 교습"
- 하고 있는 일 : "모심어 쌀나고 씨앗심어 채소재배"

이와 같이 전문조사요원들이 직업과 산업조사항목을 분명하게 구분하여 기술하지 못하고 구술자(피조사자)가 구술한 내용을 그대로 받아 적은 데이터가 많기 때문에 각 필드별 정보를 산업과 직업필드를 분명하게 구분지을 수 없다. 이런 이유 때문에 산업에 관련된 많은 정보들이 직업관련 필드에 기술되어 있다. 따라서 본 연구에서는 직업관련 필드에 기재된 정보도 산업관련 필드에 기술된 정보로 분류가 불가능할 경우에 한하여 이를 활용하고자 한다.

산업 및 직업 관련 필드에 기술된 내용의 특징은 명사(구)와 명사절 혹은 문장의 형태로 되어 있기 때문에 키워드 추출 알고리듬이 명사(구)의 추출뿐만 아니라 용언의 분석도 함께 고려되어야 한다. 이를 위해 기본적으로 맞춤법검사기, 형태소분석기, 주요어(키워드)추출기 등과 같은 자연어처리프로그램이 필요하다. 또한 구어체와 문어체로 기술된 문장 내에서 주요어(키워드)를 추출하기 위해서는 문장을 분석하는 위의 3가지의 모듈 외에 코드변환기를 사용한다. 이를 정리하면 다음과 같다.

모듈	기능
철자검사기(A)	산업/직업 관련 필드들을 읽어들여 맞춤법 오류, 띄어쓰기 오류, 비표준어 오류 등을 제거하는 기능.
철자검사기(A)	(A)의 출력에 대하여 형태소를 분석하고, 색인기를 거치면서 의미 있는 단어들을 추출하는 기능
코드변환기(B)	(B)의 출력을 바탕으로 지식베이스를 검색하여 적합한 코드부여

이 가운데 색인기는 형태소분석이 완료된 데이터에서 산업명으로 사용할 수 있는 용어를 추출하는 역할을 수행한다. 추출된 용어들은 정보검색의 입장에서는 대부분 명사들이나, 통계정보에서는 형태소 분석의 대상이 명사나 복합명사로만 되어 있는 것만은 아니다. 즉, 서술형

〈표 1〉 인구 주택 총조사표 구조 (직업/산업 부분)

구 분	산업관련필드		직업관련필드	
필드명	사업체명	사업내용	부서 및 직책	하고 있는 일
필드타입	문자	문자	문자	문자

이나 명사형을 이루고 있는 경우가 많이 있다. 예를 들면 "농사짓는 일", "각종 빵 소매", "각종 빵 만듦". "가족일을 도움" 등으로 기술되어 있기 때문에 이를 근거로 키워드를 생성할 필요가 있다. 또한, "짓는", "만듦" 등의 단어는 명사는 아니지만 의미를 가지는 동사의 명사형이기 때문에 이러한 용언형 단어들도 "생산", "제조" 등의 용어로 변환되기 위하여 반드시 추출되어야 한다. 즉, 용언 중에서 키워드어로 사용될 수 있는 용어도 추출대상이 되었다.

3.1 전거어 사전 구축

인구주택 총조사는 불특정다수인과 대상물을 조사하는 국가적 사업이다. 따라서 조사내용은 다양한 계층의 주민과 다양한 산업에 따라 동일한 내용에 대해 여러 이형이 나타날 수 있다. 이러한 이형에 대한 체계적이고 망라적인 지식베이스가 없는 자동코딩은 불가능하다. 전거어 사전은 자동코딩과정의 효율을 결정짓는 주요한 지식베이스이다. 특히, 본 개발에서는 전거어 사전이 유사어와 파생어[2]에 관한 정보도 체계적이고 구조적으로 갖고 있어야 한다. 예를 들면, 서비스에 대한 표기가 전문 조사요원에 따라 다음과 같이 다양하게 표현될 수 있기 때문이다.

① 맞춤법통일안: 서비스
② 유사어: 봉사, service, 접대
③ 이형전거어: 써비스, 서어비스, 싸비스, 써어비스, 서빙, 써빙, 써어빙

시스템의 효율성은 유사어와 이형동의어들이 적정하게 군집화되어 있는 전거어 사전을 구축하는 것이 관건이다. 따라서 본 연구에서는 크게 두 개의 전거어 사전을 개발하였다.

3.1.1 산업 일반 전거어 사전

한국 표준산업분류표는 항목간 계층구조를 갖고 있으며 세세부 항목에는 해당 항목에 관련된 해설이 기재되어 있다. 이 분류표는 수작업을 위한 분류 도구로서 분류항과 해설이 문장형태로 이루어져 있기 때문에 기계처리에는 적합치 않은 구조를 갖고 있다. 따라서 자동코딩에 사용할 수 있도록 각 분류번호에 대한 용어 테이블 데이터베이스를 구축하였다. 철도운송업에 대한 분류항과 해설, 이를 분석한 테이블을 살펴보면 아래와 같다.

산업분류표의 원문 :
· 60100 철도 운송업 : 철도차량을 이용하여 여객 및 화물을 운송하는 산업활동(도시철도 제외)을 말한다.

〈예 시〉
· 도시간 철도운송
· 철도 여객운송, 도시간
· 철도 화물운송, 도시간

· 테이블 :

분류번호	키워드
60100	철도운송업, 철도차량운송업, 철도여객운송, 철도 화물운송, 도시간 철도운송, 관광열차

· 일반 전거어 사전 : 도시간 철도운송
=철도운송, 철도여객운송
=도시간 철도 여객운송
철도 화물운송
=도시간 철도 화물운송

3.1.2 산업 공동 전거어 사전

산업분류표로 생성된 키워드를 기준으로 각 용어별에 대한 전거어를 구축할 경우에 단순이형 전거어 사전의 크기가 엄청나게 커지기 때

2) '만들다'라는 으뜸꼴 외에 '만듦', '만드는' 등과 같이 파생어를 의미한다.

문에 이의 관리가 현실적으로 불가능하며, 시스템의 효율도 떨어진다. 예를 들면, '제작'이라는 단어의 경우에 이형동의어들로 '만들기, 만듬, 해주기, 만드는 일, 만드는 곳 등'을 들 수 있다. 이러한 용어에 대해 공동전거어를 구축하지 않고 이 단어가 나열된 모든 용어에 대해 전거어를 만들 경우에 산업 분류표에 해당하는 표준항목들의 수는 기하급수적으로 증가한다. 특히, 복합명사의 경우는 그 이형형태가 너무 많아지기 때문에 이를 처리할 수 있는 새로운 형태의 전거어 사전이 필요하다. 한편, 산업조사 때 피조사들이 구술한 용어들은 전문용어보다는 조사요원들이 이해하는 수준의 일반적인 수준의 용어들이 대부분이다. 이에 비해 인구주택 총조사의 '사업내용' 필드에서 기술 형태가 전문용어나 명사로 이루어진 명사(구)와 단순한 절의 형태를 갖고 있는 경우도 많기 때문에 상대적으로 이형동의어들이 많이 나타나고 있다. 예를 들면, 한식당에 대한 전거어로서 '한식만들기, 한식만듬, 한식해주기, 한식만드는일, 한식만드는곳'과 같은 전거어가 필요하며, 또한 일식당에 대해서는 '일식만들기, 일식만듬, 일식해주기, 일식만드는일, 일식만드는곳'과 같은 전거어가 필요하다. 이외에도 중식, 양식 등 식당과 관련된 용어에 대해서는 모두 동일한 수의 전거어가 생성되어야 한다. 따라서 접미사로 사용될 수 있는 명사 혹은 용언들을 기존 전거어 사전과는 별도로 구축하여 접미사나 접두사 혹은 중간어로 활용될 수 있는 공동전거어 사전을 활용함으로써 기존의 전거어 사전의 규모를 크게 줄일 수 있다.

기존의 전문용어에 대한 전거어 사전이 이형명사형태만을 수집한 지식베이스라면, 공동전거어 사전은 이형명사형 이외에 포괄적인 의미를 갖는 명사나 용언, 파생어에 대한 정보를 수용한 새로운 형태의 지식베이스이다. 예를 들면, 철도에 대한 공동 전거어 파일을 구축하고, '철도수리'라는 명사에 대해 공동전거어 사전을 참조하여 다음과 같은 파생전거어 지식베이스가 생성된다.

· 입력데이터 : 철도수리

· 파생된 전거어 사전 : 열차 수리, 기차 수리, 기관차 수리

즉, 철도라는 용어에 "철도=열차=기차=기관차"라는 공동 전거어 사전이 구축되어 있기 때문에 일반어 전거어 사전에 철도수리에 대한 전거어가 없더라도 자동적으로 3개의 전거어 사전이 생성된다. 다음은 '판다'라는 용언에 대해 생성된 산업 공동 전거어 사전의 예이다.

· 판다=장수, 판매, 판매하다, 장사, 팔다, 파는 일, 판매하는 일, 판매하다, 판매하는 곳

3.2 전체 프로그램

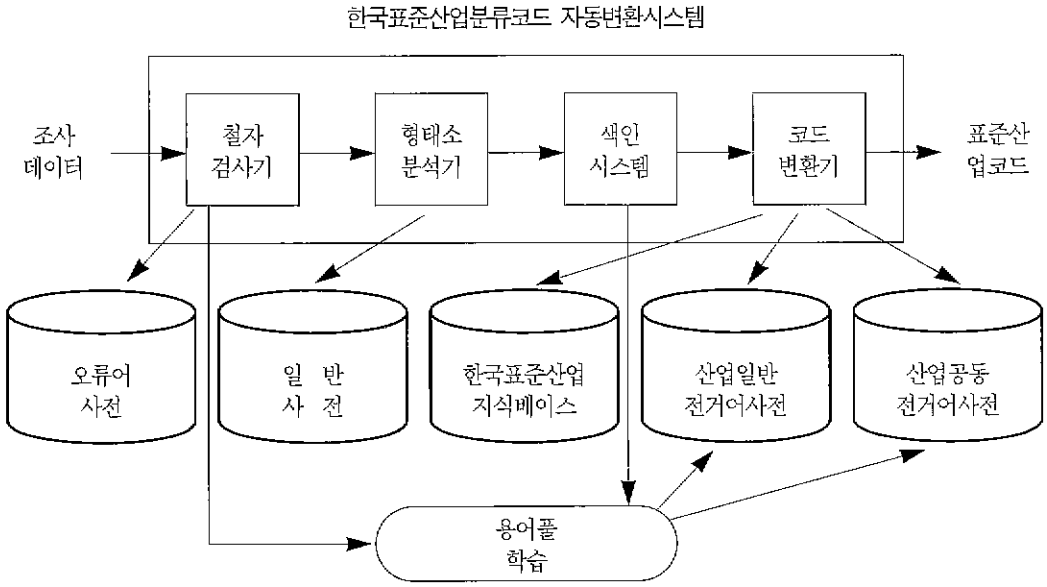
본 연구에서는 국민들의 산업활동을 조사하여 이를 한국 표준 산업분류표에 근거한 분류번호를 자동코딩하는 것을 목적으로 한다. 조사된 데이터는 전문요원에 따라 매우 다양한 형태로 기술되기 때문에 이를 산업분류표에 내포된 의미나 혹은 단어로 변환하여야 한다. 즉, 언어로써 표현한 비정형화된 기호를 산업분류표에 열거된 정형화된 기호로 변환하는 작업이다. 이를 위한 첫 번째 작업은 기술된 데이터를 정련화하는 작업으로써 철자 검사기를 이용한 단계이다. 철자 검사의 결과가 완전할수록 코딩의 효율이 높아질 것이기 때문이다. 철자 검사기가 역할을 수행하기 위해서는 오류어 사전이 필요하며, 이 과정에서는 맞춤법 오류와 띄어쓰기 오류 등과 같은 기술

(descriptive)의 문제를 해결한다.

다음 과정은 기술된 문장을 대상으로 형태소 분석기를 이용하여 산업관련 필드와 직업관련 필드에서 키워드를 추출하는 과정이다(안동언, 1999, 173-178). 한국어는 교착어이기 때문에 영문과 같이 단순한 stemming 프로그램으로 형태소를 분석할 수 없다. 또한, 입력된 산업관련 정보와 직업관련 데이터들은 명사(구)나 혹은 복합명사와 같은 전형적인 키워드의 형태 이외에도 서술어가 포함된 절과 문장의 형태를 갖고 있다. 이 과정에서 일반사전(불용어 사전 포함)이 사용된다. 다음 단계에서 형태소 분석 결과에 근거하여 색인시스템을 이용하여 한국표준산업분류표의 항목과 비교를 위한 키워드를 추출 생성한다. 생성된 키워드는 산업분류표에 근거한 분류테이블에 있는 용어와 일치하는지를 검색한다. 일치하는 것이 있으면 해당 분류 코드를 출력한다. 일치하는 것이 없을 경우에는 일반어 전거어 사전과 공동전거어 사전을 이용하며, 이 과정에서는 방언이나 외래어 표기와 같은 구술에 따른 이형데이터를 처리한다. 이 단계에서 산업분류표와 매칭이 되지 않는 용어들은 용어풀로 저장된다. 구축된 미매칭 용어들은 수작업으로 두 개의 전거어 사전에 신규로 등록한다. 코드 변환기는 산업분류표에 근거한 코딩 테이블과 분석된 용어를 비교하는 작업을 수행한다. 이상과 같은 순서를 그림으로 표현하면 〈그림 2〉와 같다.

4 자동코딩 실험

이상에서 제안한 알고리듬과 프로그램을 입증하기 위해 2000년 4월에 통계청에서 실시한 산업관련조사 자료로 실험하였다. 실험대상은 통계청에서 본 실험을 위해 제공해준 약 9,000



〈그림 2〉 자동 코딩 시스템 흐름도

〈표 2〉 인구 주택 총조사 데이터의 예

산업분류	사업체명	평균	부서 및 직책	하고 있는 일
4631	수원종합전기	전기설비공사업	무급가족종사자	전화받고 사무실관리
4631	한창전업사	전기설치 및 보수공사	자기가게운영	전기설치 및 보수공사
4631	영수전업사	전기설치 판매	부장	영업 설치 전기수리
4631	전기공업	전기자재	기술부사원	전기공사

건의 데이터이다. 이 데이터는 전체 280여만 건의 표본데이터에서 기관자체 분석용으로 통계청이 중복삭제, 중요도, 업무량 등을 감안하여 추출 보관하고 있는 핵심자료 2만여 건 중 일부이다. 본 연구에서는 협조받은 약 9,000건 가운데 통계청의 코딩요원들이 수작업으로 분류번호가 이미 부여된 6,428건만을 최종 실험 데이터로 활용하였다. 실험 데이터에는 '사업체명' 필드에만 기록된 것과 '사업체명' 필드와 '사업내용' 필드까지만 기록된 불완전한 데이터들도 포함되어 있다.

실험데이터는 조사원들이 입력한 형태를 수정 없이 그대로 실험을 수행하였다. 단 맞춤법 검사기로 오류가 검출된 용어 가운데 명백한 입력오류[3]라고 판단된 것에 한하여 수작업으로 수정하였다. 입력오류에 대한 수작업 처리건수는 전체 데이터를 기준으로 2% 미만이었다.

실험데이터의 기본 구조는 〈표 2〉와 같이 마이크로소프트사의 엑셀파일로 입력이 되어 있으며, 산업분류필드는 조사 데이터를 전문코딩요원들이 산업분류표에 의거하여 수작업으로 분류번호를 부여한 것이다. 본 실험에서는 전문코딩요원들이 코딩한 분류번호를 정확한 분류번호로 가정하였다.

4.1 용어열 단순 비교 방법

이 방법은 산업 분류표에 열거된 분류항을 포함한 해설부분과 조사된 산업관련 데이터에서 명사 혹은 명사구를 선정하여 이를 단순 비교하여 실험한 것이다. 단, 조사데이터는 각 필드별로 데이터의 양이 다르기 때문에 이에 대해 가중치를 부여하였다. 즉, 조사 데이터의 각 필드별로 명사(구)를 선정하여 필드별 가중치를 동일하게 부여하고, 4개의 필드에서 계산된 용어값을 합산하여 해당 용어들을 산업분류표의 항목(명사, 명사구)들과 비교하였다. 추출된 각 용어들이 서로 다른 영역에 배정이 될 경우에, 추출한 용어 가운데 가장 높은 값을 갖고 있는 용어가 속해있는 분류번호를 채택하였다. 예를 들면, 다음과 같다.

사업체명	두리방(주)
사업내용	미싱작업으로의류/만듬
부서 및 직책	미싱사
하고 있는 일	의류/미싱작업해서만듬
원데이터	181
자동코딩	181

3) 외래어(캣샤:현금출납원)를 발음대로 표기한 것은 수정하지 않았다
수정 예) 불빔어로→불법어로

사업체명에서 추출한 '두리방'은 해당 필드 내에서 한 단어만 출현하여 1의 값을 갖지만 등록된 용어가 아니기 때문에 0으로 처리하였다. '미싱작업'과 '의류만듬'은 '부서 및 직책' 필드에서 2개의 단어가 출현하여 단어당 각 1/2의 값을 갖는다. '미싱사'는 하나의 용어가 하나의 필드에서 나왔기 때문에 1의 값을 갖는다. '의류'와 '미싱작업'은 하나의 필드에서 두 단어가 추출되어 각 단어가 1/2의 값을 갖는다. 따라서 이를 가중치순으로 정리하면 아래와 같다.

· 주요단어순 : 미싱작업(3/2), 미싱사(1), 의류만듬(1/2), 의류(1/2)

'원데이터' 필드에 기재된 분류번호는 통계청에서 수작업으로 분류번호를 부여한 결과이고, '자동코딩' 필드에 기재된 숫자는 위 방법으로 코딩한 결과이다. 미싱작업과 미싱사도 산업분류표에 등록하지 않는 단어이기 때문에 이에 대한 값은 부여하지 않는다. 따라서 위의 주요단어 가운데 등록된 것은 '의류만듬'과 '의류'만이다. 이 두 단어가 동시에 출현한 테이블은 산업 분류표에서 '181' 부분이기 때문에 이 분류번호가 부여되었다. 이상과 같은 방법으로 전체 6,428개의 실험데이터를 분석한 결과는 다음과 같았다.

분류번호를 부여하지 못한 건수가 3,857건이었다. 이는 조사된 데이터 가운데 분류항목과 일치하지 않은 단어가 전혀 없을 경우에는 코딩을 하지 않도록 조치하였기 때문이다. 즉, 이 방법의 생성률[4]은 39.9%였다. 생성된 데이터 가운데에서 수동코딩과 자동코딩의 결과가 완전하게 일치한 것은 136건이었다. 수동코딩과 4자리수 이상이 매칭된 데이터는 12건이었으며, 3자리수가 매칭된 것은 299건이었고, 2자리수가 매칭된 것은 117건이었다. 또한 코딩은 되었으나 전혀 다른 번호가 부여된 것은 1,970건이었다.

이는 타국의 자동코딩효율과 비교하여 생성률과 정확률이 매우 저조한 결과이다. Bushnell (Bushnell and Diane, 1995, 25-35)은 '지식베이스를 이용하기보다는 용어간 매칭 프로그램이 훨씬 효율적인 코딩시스템이 될 수 있다'고 영국의 연구결과를 발표하였으나 본 연구의 실험에서는 지식베이스가 없는 자동코딩은 실용화와 거리가 있는 것으로 나타났다. 이러한 결과가 나타난 것은 형태소 해석기와 같은 프로그램 문제가 아니라 지식베이스의 부족에 따른 것으로 판단된다. 이러한 문제 때문에 일본의 경우도 상품명을 전거어로 사용함으로써 자동코딩의 효율을 높이고 있었다(충남대학교 소프트웨어연구센터, 1999, 34-37).

프랑스도 일반적인 매칭방법으로는 효과적인 자동코딩 효율을 얻지 못함으로써 참조파일의 구조화와 학습과정 알고리듬(전거어 및 유사어 사전의 강화)을 도입해 생성된 데이터의 정확률을 90% 이상을 유지할 수 있었다(Schuhl and Pierrette 1996).

4.2 지식베이스 활용 방법

산업체 조사는 조사요원들의 성격과 학력, 주변환경의 차이 때문에 동일한 결과의 기술에도 많은 차이를 보이고 있다. 예를 들면, 산업

4) 생성률(Production Rate)은 수작업으로 조사된 데이터를 자동코딩프로그램에 처리하였을 경우에 정확과 오류에 관계없이 새로운 코드를 생성한 비율을 뜻한다. 이는 각 국의 자동코딩 프로그램의 효율을 측정할 때 사용된다.

분류표에는 의복제조로 되어있으나 일부 조사 데이터에서는 '의류만듬. 미싱으로 옷만듬, 옷만듬, 의상만듬 등'으로 다양하게 표현하기 때문에 같은 분류번호가 부여되는 내용임에도 불구하고 해당 필드에 기술하는 방식은 조사요원들간에 많은 차이를 보이고 있다. 따라서 정확한 코딩을 위해서는 분류항목에 대응되는 일상생활에 사용하는 용어로 구성된 지식베이스를 필요로 한다.

이에 따라 본 연구에서는 학문적 성격보다는 일상적 성격의 지식베이스를 구축하였다. 지식베이스로 사용한 것은 2개의 코딩용 전거어사전이며 일반사전에는 불용어 사전도 포함되어 있다. 일반사전은 불용어를 제외하고 기존에 책자형 사전의 품사정보만이 부여된 형태로서 가장 일반적인 구조를 갖고 있다.

① 산업 일반 전거어 사전

산업 전거어 사전은 기존의 전거어 사전의 항목과 거의 유사한 것으로 산업관련 용어중심의 표제항을 수집하여 구축하였다. 대표적인 형태는 다음과 같다.

· **주류**=소주=비어=맥주=양주[5]

② 산업 공동 전거어 사전

기존의 전거어 사전이 명사중심의 전거어 사전이라면 본 연구에서 제시한 산업 공동 전거어 사전은 용언중심의 전거어 사전이다. 명사로서 일반 전거어 사전을 구축한 예를 들면 다음과 같다.

· **서비스**=써비스, 서어비스, 싸비스, 써어비스, 써빙, 서빙, 써어빙

용언으로 구축한 일반어 전거어 사전을 구축한 예를 들면 다음과 같다.

· **제작하다**=제작, 만들다, 만들기, 만듬, 해주기, 만드는 일, 만드는 곳

③ 불용어 사전

본 연구에서 사용한 불용어 사전은 형태소 해석과정에 분석대상명사를 생성한 후에 지식베이스로 사용한다. 이때 불용어 사전에 등록된 용어들은 해당 용어를 분석대상용어에서 완전히 제외한다. 예를 들면, 밀양한국은행지점이라는 사업체명에 기재된 데이터가 형태소 해석과정에서 불용어 사전을 검색하여 '지점'이 불용어로 등재되어 있을 경우에 분류표에 비교할 용어는 '밀양한국은행'이 사용된다.

④ 학습과정

본 연구에서는 분류번호 부여의 생성과 누락에 관계없이 각 필드에 기재된 데이터 가운데 사전에 등록되지 않은 용어나 혹은 분석하지 못한 구절에 대해서는 일련의 학습과정을 거친다. 전자는 주로 조사원들이 사용하는 지역방언(예: 뻥티기 만듬)이나 일본어식 표현(예: 오뎅)과 같은 외래어, 영어와 같은 외국어 등이다. 이러한 용어들은 생성률과 정확률을 높이기 위해 학습이 필요하며 모두 용어풀에 수집된다. 용어풀에 수집된 용어들은 계산된 빈도정보에 따라 사전이나 전거어 사전에 등재된다. 후자는 주로 형태소해석의 실패로 이루어졌을 경우에 수작업을 통하여 해당정보를 다음 형태소 해석시에 활용할 수 있도록 하였다.

⑤ 필드별 가중치 부여

용어열 단순 비교 방법으로 실험한 결과, 각 필드별로 동일한 가중치를 부여함으로써 정확률이 상대적으로 떨어졌다고 판단하여, 본 실험에는 필드별 우선 순위를 부여하여 분류작업을 수행하도록 하였다. 즉, 산업분류이기 때문에

5) 굵은 글자형태는 산업분류표에 열거된 용어이다. 그밖의 것은 조사데이터에서 입수하였다

산업관련 필드인 '사업체명' 필드와 '사업내용' 필드에 기재된 정보만으로 분류번호를 우선 부여한다. 분석된 용어가 2개 이상이고 각 용어가 속해있는 체계가 대분류(2자리수: 05)가 서로 상이할 경우[6]를 제외하고는 '사업체명' 필드에 들어 있는 정보는 상위개념에 해당하는 것으로 간주하고, '사업내용' 필드에 들어있는 정보는 하위개념으로 간주한다. 따라서 우선 '사업내용'에 들어 있는 정보로 분류번호를 검색한다. 그러나 '사업내용'에 들어있는 정보로 분류번호 부여 작업이 실패했을 경우에는 '사업체명'에 들어있는 정보만으로 분류번호를 부여한다.

즉, '사업체명' 필드에 기재된 단어를 가장 중시한다. 따라서 이 필드에서 추출한 용어가 산업분류표에 열거된 항목과 매칭이 되면 해당 분류번호를 우선적으로 부여한다.

한편, 산업에 관련된 두개의 필드에서 아무런 데이터를 얻지 못하면, 직업에 관련된 세 번째 '부서 및 직책' 필드를 참조한다. 여기서도 원하는 데이터를 못 얻을 경우에 네 번째 '하는 일' 필드를 참조한다. 예를 들면, 다음과 같다.

· 세 번째 데이터로 확인하는 경우 : 남의가정집/가정일,돌봐줌/파출부/공란

이 경우에 '남의가정집'이나 '가정일,돌봐줌'으로 적합한 분류번호를 찾지 못하기 때문에 세 번째 필드에 있는 '파출부'로 해당 분류번호를 추출한다.

· 네 번째 데이터로 확인하는 경우 : 르네상스빌딩/여러직종,사무실/관리직/주차,정산관리

'빌딩'이란 단어와 '사무실', '관리직'이란 단어로 적정한 대응 분류번호를 찾지 못할 경우에 '주차'로 해당 분류번호를 추출할 수 있다.

4.3 지식베이스 적용후의 성능평가

이러한 지식베이스를 사용하고도 일부 데이터에 대해서는 분류번호가 누락되고, 오분류가 이루어졌다. 그 원인을 분석하면 아래와 같다.

① 분류번호 부여 누락

학습이 이루어지고도 자동코딩이 이루어지지 않은 경우가 있었다. 코딩이 대부분의 실패한 경우에는 각 필드에 들어있는 정보가 극히 적은 경우이다. 예를 들면, 다음과 같이 '사업체명' 필드에만 '누리A'라는 정보가 있으며 다른 필드에는 전혀 정보가 없는 경우이다. 또한 '누리A'라는 정보차제도 축약된 표현으로 이루어진 것이다. 여기서 A는 아파트의 축약된 형태이지만 이를 전거어 사전에 등록할 경우에 다른 정보의 오류가 발생하여 오히려 자동분류에 더욱 큰 혼동을 야기시킬 수 있기 때문에 'A'라는 기호는 전거어로 채택할 수 없었다.

분류번호	70211
사업체명	누리A
사업내용	공란
부서 및 직책	공란
하고있는일	공란

② 분류번호의 부여실패

학습이 이루어진 후에도 코딩이 잘못 이루어진 경우가 있었다. 이는 대부분 해당 정보의 동음이의어나 혹은 외래어 표기로 이루어진 경우에 발생하였다. 예를 들면, 산업에 관련된 필드

6) 예를 들면, 다음과 같은 정보가 나타났을 경우이다. "구로구청(분류번호:76113)/모여/목회활동(91912)" 이러한 것은 수동으로 처리한다.

에서 적절한 용어를 추출하였음에도 불구하고 외래어와 동음이의어로 코딩이 잘못 이루어진다. 예를 들면 다음과 같다.

분류번호	70211
자동코딩결과	55212
사업체명	신세계FOOD
사업내용	중식제공

위와 같이 분류가 잘못 이루어진 것은 '사업내용'에서 중식을 중국음식으로 전거어 사전에 등재되어 있기 때문에 점심(lunch)의 의미로 해석되지 못하였다. 또한 '사업체명' 필드에서 '신세계 식당'으로 기술되었을 경우에 최소한 근사분류번호를 부여할 수 있었으나 이도 영어(food)로 기술되어 '사업내용'에 있는 정보만으로 분류되기 때문에 자동코딩 결과는 55212(중국음식점)로 잘못 분류되는 결과를 얻게 되었다.

③ 분류번호의 근사치부여

분류번호의 근사치 부여는 수작업 처리결과의 분류번호와 거의 유사하게 부여된 결과들을 의미한다. 이는 코딩의 실패일 수도 있으나, 수작업 처리결과자체도 대략적으로 분류번호를 부여한 경우가 있으며, 입력의 오류가 상당수 있는 것을 감안할 경우에 근사치로 분류번호가 부여된 것을 완전한 분류의 실패로 볼 수 없다고 판단한다. 다음은 분류번호를 수작업으로 분류하면서 나타난 오류의 예이다.

· 입력오류: 아래의 예에서 분류번호(26211)는 벽돌 제조업체에 관련된 것이지만, 사업체명은 전혀 무관한 내용이기 때문에 이 자체가 맞는 것인지 틀린 것인지 실제로는 확인이 불가능한 경우이다.

분류번호	26211
사업체명	꼬마동네옷가게
사업내용	내화벽돌제조생산업체

· 분류오류 : 아래의 예에서 일식집이 상호를 초밥여행으로 짓고, 음식점(분류번호: 55212)을 하고 있음에도 분류전문요원이 이를 여행사(6331)로 간주하여 잘못 분류한 예이다.

분류번호	6331
사업체명	초밥여행
사업내용	일식

· 근사치 분류 : 아래의 예에서 '사업체명' 필드에 있는 식당이란 정보만으로 본 시스템에서는 근사치 분류번호를 제공하나 산업분류체계에서 집단 급식에 대한 별도의 분류번호를 갖고 있기 때문에 그에 대한 세부적인 분류번호는 부여되지 못하고 근사치 분류번호만 부여된 것이다.

분류번호	55215
코딩번호	552
사업체명	기숙사내 식당
사업내용	사원들에게 식사제공
부서 및 직책	식당아주머니
하고있는일	사원들에게 식사제공

따라서 수작업 분류에서의 오류는 전혀 엉뚱한 번호가 부여된 것에 비하여 본 연구에서 근사치 분류는 최소한 2자리수 이상의 분류번호가 부여됨에 따라 전혀 상관없는 번호가 부여

되지 않았음을 보이고 있다.

④ 실험결과

이러한 오류에도 불구하고 본 실험의 결과는 생성률의 경우는 선행연구에 비해 비교적 긍정적인 결과를 얻을 수 있었다. 다음 〈표 3〉은 본 실험에서 학습 후에 얻어진 코딩 결과이다.

〈표 3〉 실험결과(생성률과 정확률)

	조사건수	비 율
생성률	6396/6428	99.5%
정확률	5331/6396	83 3%
근사치율	662/1065	62.2%
누락률	32/6428	0.5%

〈표 3〉에서 정확률은 학습이 완료된 후에 조사된 것으로 생성이 이루어진 6,396건을 기준으로 5,331건이 정확하게 일치한 경우이다(83.3%). 근사치율은 정확하게 분류되지 않은 1,065건 가운데 수작업으로 부여된 분류번호와 비교하여 앞의 2자리수 이상이 동일한 분류번호가 부여된 662건을 계산한 비율(62.2%)이다. 누락률은 분류번호를 부여하지 못한 경우이다.

이 결과들을 각국 통계자료의 자동코딩을 실험한 연구와 비교하면 〈표 4〉와 같다. 이 가운데 일본(米澤 1998)은 생성률에 대한 데이터가 없이 정확률 측정에서 상품명을 첨부한 코딩실험에서 정확한 생성이 이루어진 데이터 가운데 1/3만을 선정하여 조사하면 99% 이상의 정확률을 보인다고 기술되어 있으나 생성률에 대한 데이터가 없어 정확한 비교대상이 될 수 없다고 판단된다.

이상과 같이 본 연구에서 나타난 생성률이 다른 연구결과에 비해 상대적으로 높은 것은 크게 두 가지 원인에 기인하기 때문이다. 첫 번째는 학습 과정을 여러 단계에 걸쳐 전거어 사전의 수준이 높아진 것에 기인한다. 왜냐하면 조사원들이 기술한 조사데이터를 컴퓨터로 생성하지 못하는 것은 대부분 전거어 사전이 포괄적이지 못하기 때문이었다. 반면에 정확률이 떨어지는 것은 산업 분류표에 기술된 용어들이 너무 하부 항목(5자리수)까지 세부적으로 전개되어 있어 조사된 데이터에 특정한 용어가 기술된 경우 외에는 이에 대한 정확한 매칭이 어려웠기 때문이다. 따라서 근사치 분류건수까지를 정확률에 포함하였을 경우에는 생성률은 93.7%로 오히려 외국의 경우보다 높은 자동코딩결과를 보이고 있었다.

5 맺는 말

인구센서스 데이터를 자동코딩하기 위해서는 일반적으로 자연언어처리기법에 기반한 단어열 매칭을 이용한 코딩방법과 전거어 사전과 같은 지식베이스를 이용한 코딩방법이 널리 사용되

〈표 4〉 각국 자동코딩 효율 비교표

	미국	프랑스	일본	본 연구	비고
생성률	63	66	-	99.5	
정확률	90	90	99%	83.3	93.7

고 있다. 본 연구에서는 단어열 매칭방법으로 자동코딩의 효율성을 측정하여 결과의 불안정성을 확인하고, 이 방법의 단점을 보완하여 지식베이스를 이용한 자동코딩 알고리듬을 제시하였다.

실험 데이터로는 통계청에서 제공받은 2000년 4월에 산업조사 데이터를 활용하였으며, 산업분류표에 근거한 매칭테이블 구축과 조사데이터에서 주요어를 추출하기 위해 자연어처리 프로그램들을 사용하였다. 또한, 테이블과 키워드간의 매칭효율을 높이기 위해 지식베이스로 산업 일반 전거어 사전과 산업 공동 전거어 사전을 사용하였다. 두 개의 전거어 사전을 확장 관리하기 위해 학습이론을 사용하였다. 학습의 원리는 학습을 통해 이형데이터를 확보하고, 해당 용어들을 용어풀에 보관하고 용어의 출현빈도가 적정한 기준에 도달하면 이를 전거어 사전에 등재하는 과정의 반복이었다.

이러한 알고리듬에 기반하여 자동 코딩 시스템으로 실험데이터를 분석한 결과, 생성률은 99.5%를 얻을 수 있었으며, 정확률은 83.3%를 얻을 수 있었다. 이는 외국의 자동코딩방식에 비해 생성률이 최저 33.5% 이상 높은 수치이며, 정확률은 6.7%가 낮은 수치이다. 그러나 정확률에 있어서는 근사치까지 분류가 이루어진 것을 정확하게 분류가 된 것으로 간주하여 합산할 경우에 93.7%로서 궁극적으로는 정확률도 3.7%정도가 오히려 높아졌음을 확인할 수 있었다. 이렇게 상대적으로 월등한 생성률과 정확률을 보이고 있는 것은 학습과정에서 두 개의 전거어 사전을 확대 구축하여 전문조사요원들이 사용하고 있는 조사패턴을 거의 입수하였기 때문으로 판단된다. 또한 생성률의 경우에 산업관련 필드들에서 적합한 코딩정보를 입수하지 못하였을 경우에 직업관련 필드에 기술된 정보까지도 확장하여 활용하였기 때문에 상대적으로 높은 수치를 얻었기 때문이다. 향후 연구는 이 시스템으로 2000년 11월에 조사될 인구주택 총조사에서 입수된 전체 데이터를 대상으로 분석이 이루어질 필요가 있다.

그 결과 본 연구에서 제시한 코딩효율이 확보될 경우에 본 연구에서 제시한 알고리듬과 시스템은 국가의 미래계획 수립에 가장 중요한 국가 통계데이터를 수집할 수 있는 중요한 이론과 시스템으로 활용될 수 있을 것이다.

참 고 문 헌

남영준. 1995. 『색인어 형태분석에 의한 한국어 자동색인기법연구』. 중앙대학교 대학원 박사학위논문.

米澤. 1998. "平成9年度の産業分類自動格付システムの研究及び改良報告," 日本 統計廳 情報企劃室.

신동욱, 안동언. 1996, 『MIDAS 기반 정보검색 시스템 개발』. 최종보고서, 한국전자통신연구소

안동언. 1999, "좌우접속정보를 이용한 명사추출기," 『1999년도 제11회 한글 및 한국어 정보처리 학술대회 및 제1회 형태소 분석기 및 품사태거 평가 워크숍 논문집』. 한국정보과학회. 한국인지과학회.: 173-178.

안동언. 1997. 『한국어 형태소분석기 개발에 관

한 연구. 삼성전자.

안동언 1998. 확장 품사 사전 규칙과 보급 패키지. 국어정보처리기술 개발 제3차년도 최종보고서. 한국과학기술원.

충남대학교 소프트웨어연구센터 1999. 산업/직업 분류 코딩 자동화 시스템 개발을 위한 기초 연구. 통계청.

통계청. 2000. 한국표준산업분류. 통계청. http://www.nso.go.kr/stat/indclass/k-industry.htm

Apte, Chidanand and Damerau(1994), Fred, Weiss, Sholorm M. "Towards Language Independent Automated Learning of Text Categorization Models." In *Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Bushnell, Diane (1995). "Computer Assisted Occupation Coding." *IBUC 1995 3rd Annual International Blaise Users Conference Helsinki*: 25-35. (http://www.census.gov/prod/2/gen/96arc/ixbschuh.pdf)

D. W. Gillman and M. V. Appel(1994). "Automated Coding Research at The Census Bureau." U.S. Census Bureau (http://www.census.gov/srd/www/abstract/rr94-4.htm)

Nordbotten, S(1993), "Statistical Meta-Knowledge and -Data." *Statistical Journal of UN/ECE*, 10(2): 101-111.

Ruiz, Miguel E and Srinivasan, Padmini(1999), "Hierarchical Neural Networks for Text Categorization." In *proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in information Retrieval*: 281-282.

Schuhl, Pierrette (1996), "SICORE, The INSEE Automatic Coding System." US Census Bureau. *1996 Annual Research Conference Proceedings.*

Vizine-Goetz, Diane and Jean Godby(1996). "Library Classification Schemes and Access to Electronic Collections: Enhancement of the Dewey Decimal Classification with Supplemental Vocabulary." In *Advances in classification research Volume 7: proceedings of the 7th ASIS SIG/CR classification research workshop*: 127-135.