정보검색 기법을 이용한 산업/직업 코드 자동 분류 시스템*

임 희 석†

요 약

본 논문은 통계청에서 실시하는 인구 주택 총조사와 사업체 기초통계조사 시 실시되는 수작업에 의한 표준 산업/직업 코드 분류 시 발생하는 막대한 비용과 시간, 일관성의 결여 등을 해소하기 위한 표준 산업/직업 코드 자동 분류 시스템을 제안한다. 제안한 시스템은 정보 검색 기법과 문서 분류 기법을 이용하여 자연어로 기술된 레코드를 입력받아 입력 레코드에 해당하는 분류 코드를 생성한다. 수작업으로 올바른 코드가 할당되어 있는 산업 분류 레코드 46,762개와 직업 분류 코드 36,286개를 이용하여 10-fold cross-validation evaluation을 수행한 결과, 제안한 시스템은 완전 자동 모드에서 2수준의 산업 분류에 대해서 87.08%, 5수준에 대해서는 66.08%의 생성률을 보였으며 반자동 모드에서는 각각 99.10%와 92.88%의 성능을 보였다. 직업 분류 코드에 대한 성능은 산업 분류 코드에 대한 성능보다는 약간 저하된 성능을 보였다. 제안한 시스템은 아직 수작업을 완전히 대체할 수 있는 완전 자동 분류기로서는 많은 개선의 여지를 가지고 있지만 수작업을 최소화할 수 있는 반자동 도구나 수작업의 정확도를 검증할 수 있는 보조 도구로써 충분히 활용될 수 있을 것으로 기대된다.

키워드 : 산업 분류 코드, 직업 분류 코드, 자동 코드 분류

An automated Classification System of Standard Industry and Occupation Codes by Using Information Retrieval Techniques

Heui Seok Lim

Abstract

This paper proposes an automated coding system of Korean standard industry/occupation for census which reduces a lot of cost and labor for manual coding. The proposed system converts natural language responses on survey questionnaires into corresponding numeric codes using information retrieval techniques and document classification algorithm. The system was experimented with 46,762 industry records and occupation 36,286 records using 10-fold cross -validation evaluation method. As experimental results, the system show 87.08% and 66.08% production rates when classifying industry records into level 2 and level 5 codes respectively. The system shows slightly lower performances on occupation code classification. We expect that the system is enough to be used as a semi-automate coding system which can minimize manual coding task or as a verification tool for manual coding results though it has much room to be improved as an automated coding system.

Keyword : automated coding system, standard industry code, standard occupation code, informaiton retrieval

1. 서론

통계청에서 실시하는 인구 및 주택 조사는 매 5년(0년, 5년)마다 실시되고 있다. 인구주택 총조사의 방법으로는 전체 가구를 조사하는 전수 조사와 전체의 10%만을 발췌하여 조사하는 표본 조사가 있다. 사업체 기초통계조사는 연단위로 이루어져 통계청에서 주관하는 사업 중 연속성이 있는 조사

† 종신회원: 한신대학교 소프트웨어학과 교수
논문접수: 2004년 7월 2일, 심사완료: 2004년 7월 10일
* 본 논문은 2004년도 한신대학교 학술연구비 지원에 의하여 연구되었음

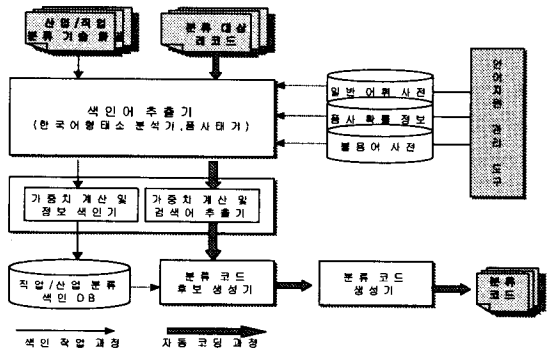
이다.

인구 주택 총조사나 사업체 기초통계 조사 방법은 조사원들이 조사 대상을 방문하여 그들에게 직접 문의하는 방법으로 이루어진다. 조사 항목은 크게 거주 지역, 출생 지역, 성씨, 나이, 표준산업분류 코드, 표준 직업분류 코드 등 정형화된 데이터와 각 개인이 근무하고 있는 사업체 명, 사업체의 주된 사업 내용, 자신의 직책, 그리고 직무 등을 나타내는 비정형화된 데이터로 구분된다. 정형화된 데이터 중 표준산업분류코드와 표준 직업분류 코드는 국가의 경제, 산업, 예산 등의 국가 기본 정책을 수립하는데 있어서 기반이 되는 중요한 지식이다. 표준산업분류 코드와 표준 직업분류 코드는 조사원이 가구 조사에서 얻은 사업체명, 사업체의 주된 사업 내용, 직책, 그리고 직무에 대한 자연어로 기술한 설명에 근거하여 작성된다. 현재까지 표준산업분류코드와 직업분류코드 분류 작업은 한국표준산업분류 책자와 표준직업분류 책자를 참조하여 수작업으로 코딩을 하고 있으며, 이로 인하여 아래와 같은 문제점이 발생하고 있다.

- **수작업을 수행하기 위한 작업자 교육 및 활용에 많은 비용이 소요**
- **막대한 수작업 량과 고비용 발생** : 표준산업분류 코드와 표준직업분류 코드를 분류할 처리 레코드는 약 300만개에 해당된다. 따라서 300만개의 레코드를 일일이 사람이 코드를 부여할 경우 막대한 수작업이 필요하며, 이에 따른 비용이 매우 크다.
- **매번 반복 작업에 따른 인력 및 비용 소모** : 일년에 한번 또는 5년에 한번씩 이루어지는 조사에 따른 반복된 인력 동원 및 수작업으로 인하여 비용 소모가 매우 크다.
- **구축된 자료의 부정확성** : 수작업으로 코드를 부여할 경우 작업자의 심리 상태, 작업 환경 등에 따라서 실수로 코드를 잘못 부여할 경우가 종종 발생하게 된다. 또한 수작업자들이 대부분 일정 기간 교육을 받아 분류 작업에 투입된 비전문가들이기 때문에 코딩 결과의 정확도가 떨어진다.
- **구축된 자료의 일관성 유지의 어려움** : 자료가 매우 방대하기 때문에 여러 사람이 작업을 나누어 수행하게 되므로 같은 산업이나 직업에 대해서도 각자가 정확한 분류를 찾지 못한다면 부여된 코드가 서로 다를 수 있다. 수작업자의 산업과 직업에 대한 선입관과 관점에 따라 코딩 결과가 다양성을 띠게 된다.

위와 같은 수작업에 의한 표준 코드 분류 작업의 문제점을 극복하기 위한 방법은 가구 조사에서 얻은 자연어의 응답을 표준 분류 코드로 분류할 수 있는 자동 코드 분류 시스템을 개발하여 활용하는 것이라 할 수 있다. 미국과 캐나다와 같은 외국에서는 1980년대부터 이미 자동 코드 분류 시스템에 대한 연구를 시작하여 현재까지 꾸준히 수행하고 있으며, 높은 정확도를 보이는 시스템이 표준 산업/직업 코드 자동 분류를 위하여 활용되고 있다[1,2,3]. 이에 반하여 국내의 통계 조사를 위한 자동 분류 시스템에 대한 연구는 매우 미흡한 실정이다. 한국어로 쓰인 가구 조사 내용을 표준 산업/직업 코드로 분류하는 시스템은 한국어의 특성상 외국에서 개발된 시스템을 직접 활용하기에는 무리가 따르며, 자체 개발의 필요성이 매우 요구 된다. 본 논문은 조사원들로부터 획득된 각 개인들의 '근무 사업체명', '사업체의 주된 업무', '직책', 그리고 '직무'에 대한 내용을 입력받아 한국표준산업 분류 코드와 한국표준직업 분류 코드를 자동 생성하는 산업/직업 코드 자동 분류 시스템을 제안하고, 실험을 통한 결과를 제시함으로써 한국 통계 조사를 위한 자동 분류 시스템의 가능성에 대하여 논하고자 한다.

2. 시스템 개요



(그림 1) 산업/직업 코드 자동 분류 시스템 구성도

본 논문이 제안하는 시스템은 조사원들이 수작업으로 산업/직업 코드를 분류할 때 참조하는 한국표준산업분류 책자와 표준직업분류 등 색인 대상 데이터에서 색인어를 추출하여 색인 DB로 구성한다. 코드를 분류할 때는 실제 조사원들이 조사해온 자연어 응답 데이터를 입력받아 검색을 수행하여 가장 유사도가 높은 산업/직업 코드를 후보로 생성하고, 후보 코드 중 올바른 분류 코드를 생성한다. (그림 1)은 본 논문이 제안하는 산업/직업 코드 자동 분류 시스템을 도식화한 것이며, 색인 시스템, 정보 검색 기법에 의하여 분류 코드 후보를 생성하는 분류 코드 후보 생성기, 그리고 후보 코드들의 정보를 이용하여 최종 분류 코드를 생성하는 코드 생성기로 구성되어 있다. 색인 시스템과 분류 코드 후보 생성기는 색인어 및 검색어를 추출하기 위하여 한국어 형태소 분석기, 품사 태거, 그리고 이와 관련된 언어 자원들로 구성된다. 본 시스템에서 사용되는 언어 자원으로는 한국어 형태소 분석 및 명사 추출을 위하여 일반 어휘 사전이 사용되며 명사 추출기에서 추출된 명사 중 색인어로써 가치가 없는 명사를 제거하기 위하여 사용되는 불용어 사전, 그리고 품사 태깅을 위한 확률 정보가 사용된다.

본 논문의 구성은 다음과 같다. 3장에서는 색인어를 추출하고 색인어의 가중치를 계산하는 색인어 시스템에 대하여 설명한다. 4장은 자연어로 기술된 조사원들의 응답에서 검색어를 추출하고 이와 유사한 코드를 검색하는 분류 코드 후보 생성기에 대하여 설명하고, 5장에서는 분류 후보 코드 중 올바른 분류 코드를 생성하는 분류 코드 생성기에 대하여 설명한다. 6장에서는 제안 시스템의 실험 결과 및 분석 결과를 설명하고 7장에서 결론을 맺고자 한다.

3. 색인시스템

색인시스템은 색인 DB 구축에 사용되는 색인 대상 데이터와 색인어 추출 및 가중치 계산기로 구성되며 각 모듈에 대하여 자세히 설명한다.

3.1. 색인 대상 데이터

색인 대상 데이터는 수작업으로 코드를 분류할 때 참조하는 한국 표준산업 분류 책자와 한국 표준직업 분류 책자의 텍스트 데이터, 표준 코드를 대표할 수 있는 색인어 목록, 그리고 과거 용어 집합으로 구성된다. 표준 산업 분류 코드와 표준 직업 분류 코드는 1수준에서부터 5수준까지 계층적으로 분류되어 있으며 각 수준의 분류 코드의 수는 <표 1>과 같다.

<표 1> 한국표준산업(직업) 분류 코드 분류 체계

코드 \ 수준	1	2	3	4	5
산업분류	20	63	194	442	1,121
직업분류	11	46	162	447	1,404

한국 표준 산업/직업 분류 책자 내의 각 분류 코드에 대한 설명은 표준 코드, 코드에 해당하는 산업/직업명, 코드의 산업/직업에 대한 설명, 예시, 제외 등의 데이터를 포함하고 있으며, <표 2>는 한국 표준 산업분류 코드 중 01121의 내용을 보인 것이다.

<표 2>의 01 줄의 '01121'은 표준코드 번호를 의미하며 '채소작물 재배업'은 해당 코드에 부여된 산업명을 의미한다. 02~03까지 해당 코드에 대한 설명이 있으며, 05 <예시>는 코드에 해당되는 예를 설명한 부분이다. 10 <제외> 부분은 해당 코드와 산업 내용이나 직업 내용이 유사하지만 제외되어야 하는 코드를 설명하는 부분이다. 색인 대상 데이터 중 표준 코드를 대표할 수 있는 색인어 목록은 특정 코드의 색인어로 사용될 가능성이 높은 명사 및 명사절을 "코드 색인어" 형식으로 기술한 레코드의 집합으로 통계청에서 수작업 코딩을 수행하는 전문가에 의해서 구축되었다. 현재 산업 분류를 위하여 23,431 레코드와 직업분류를 위한 17,184개의 레코드가 사용된다. <표 3>과 <표 4>는 각각 산업 분류를 위한 색인어 목록과 직업 분류를 위한 색인어 목록의 예를 나타내고 있다. 색인어 목록은 전문가에 의해서 작성된 것으로 코드 자동 분류 시 색인어 목록에 나타난 단어들은 코드 분류를 위하여 결정적인 실마리를 제공할 수 있다.

<표 2> 한국 표준직업분류 책자의 일부

01:	**01121**	**채소작물 재배업**
02:		노지에서 각종 채소작물을 재배하는 산업활동을 말한다. 노지에서 깻잎 등과 같이 채
03:		소로 사용하기 위하여 각종 작물을 재배하는 경우에도 여기에 분류된다.
04:		
05:		<예 시>
06:		· 풋마늘 및 풋고추 재배 · 파 및 양파 재배
07:		· 아스파라거스 재배 · 참외, 수박 및 멜론 재배
08:		· 잎 및 열매 채소 재배
09:		
10:		<제 외>
11:		· 채소작물의 종자 및 묘목 생산(01123)
12:		· 커피, 코코아, 차, 겨자, 붉은 고추 등의 음료용 또는 향신용 작물 재배(01132)
13:		· 딸기 노지 재배(01131)
14:		· 콩나물, 버섯 및 기타 채소작물의 시설재배(0115)

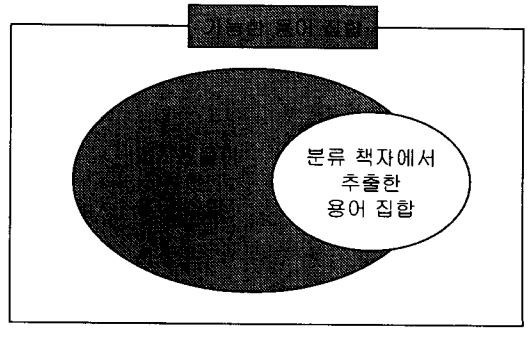
<표 3> 산업 분류를 위한 색인어 목록의 예

:	
01110	감자 재배
01110	고구마 재배
01110	곡물작물 재배
:	
01121	시금치 재배
01121	아스파라거스 재배
01121	양배추 재배
:	
12112	다이나스어스 채취
12112	마전토등 채취
12112	뮬라이트 채취
12112	샤모트 채취
12112	홍주석 채취
:	

<표 4> 직업 분류를 위한 색인어 목록의 예

:	
01110	감자 재배
01110	고구마 재배
01110	곡물작물 재배
:	
01121	시금치 재배
01121	아스파라거스 재배
01121	양배추 재배
:	
12112	다이나스어스 채취
12112	마전토등 채취
12112	뮬라이트 채취
12112	샤모트 채취
12112	홍주석 채취
:	

색인 대상 데이터 중 과거 용어 집합은 수작업으로 산업/직업 분류 코드가 정확하게 분류된 레코드에서 추출된 명사나 명사구로서 사람들이 각 코드를 기술할 때 실제 사용된 용어들을 모아둔 집합을 의미하며 과거 용어 집합의 활용 목적은 다음과 같다. 제안하는 시스템은 색인 대상 데이터에서 색인어를 추출하여 색인 DB로 구성하고, 실제 조사원들이 조사해온 자연어 응답의 데이터를 입력받아 검색을 수행하여 가장 유사도가 높은 코드를 후보로 생성하고, 후보 코드 중 올바른 분류 코드를 생성한다. 그러나 표준 산업/직업 분류 코드의 책자에서 코드를 기술하기 위하여 사용된 용어들은 해당 코드와 관련된 많은 용어들을 포함하고 있지 못하다. 따라서 조사원들이 조사한 자연어 응답에서 추출한 용어와 분류 책자에서 추출하여 구성한 색인 DB간의 용어들이 불일치 현상이 높게 나타난다.



(그림 2) 용어 집합간의 불일치 현상

(그림 2)에서 보인 것처럼 조사원들은 매우 다양한 용어를 사용하나 분류 책자에서 사용하는 용

어들의 집합은 한정되어 있다. 따라서 표준 산업/직업분류 책자의 용어들만으로 색인 DB를 구성할 경우 조사원들이 사용한 용어들로 색인 DB를 탐색하여 코드를 자동 생성하려는 경우 탐색되지 않는 용어들이 많이 발생하며 이로 인하여 코드 분류의 정확도가 낮아지게 된다. 이러한 문제점을 해결하기 위한 한 가지 방법은 전거어 사전을 사용하는 것이나 현재로서는 사용할 수 있는 전거어 사전이 존재하지 않는다. 따라서 본 논문은 과거 용어 집합을 색인 대상 데이터에 포함시켜 색인어로 사용함으로써 용어 집합간의 용어 불일치 문제를 완화시키고자 한다.

과거 용어 집합은 과거에 코드가 정확하게 부여된 레코드에서 코드, 사업체명, 사업체의 주요 업무, 직무, 하고 있는 일의 종류 등 4가지의 필드에서 추출하여 구성한다. 이중 사업체명, 사업체의 주요 업무 필드는 산업 분류를 위한 용어로, 주요 업무, 직무, 하고 있는 일 필드는 직업 분류를 위한 용어 집합으로 사용된다. 현재 산업 분류를 위하여 46762개의 레코드, 직업 분류를 위하여 36,286개의 레코드가 사용된다. <표 5>와 <표 6>은 각각의 레코드의 예를 나타내고 있다.

<표 5> 산업 분류를 위한 과거 용어 집합의 예

011	가사 농사 우리논 자두 벼농사
:	
01110	우리 논 벼농사
:	
55211	순천식당 숯불갈비 음식점 쉘덤 승미 식당 한식
:	

<표 6> 직업 분류를 위한 과거 용어 집합의 예

0110	시의회의장 시의정활동 시의원 관내의정활동
:	
5120	종사원 군복 판매 주유 배달
5120	연락 종사원 식품 판매
:	
8424	운전 계란수송 공장자재운반 기구운반
:	

3.2. 색인어 추출 및 가중치 계산

색인어와 검색어 추출을 위해서는 음절 단위 단어 인식 모델을 이용한 [4]에서 제안한 명사 추출기를 사용하고 색인어의 가중치는 (식 1)과 같이 정보검색 분야에서 전통적으로 사용되는 TF/IDF 기법을 이용하여 계산한다[5, 6].

$$w_{ij} = \frac{Tf_{ij}}{\max_l TF_{lj}} \times \log \frac{N}{n_i} \qquad \text{(식 1)}$$

(식 1)에서 w_{ij}는 색인어 i의 코드 j에서의 중요도이고 TF_{ij}는 코드 색인어 i 가 코드 j에서의 출현 빈도이고, N은 전체 색인된 코드 개수며, n_i는 전체 코드 집합에서 색인어 i를 포함하고 있는 코드의 개수를 의미한다. 산업/직업 코드는 색인어 추출과 가중치 계산 작업에 의하여 t 차원의 코드 벡터, $\vec{c_j} = (w_{1j}, w_{2j}, \ldots, w_{tj})$로 표현된다. 코드 벡터는 산업/직업 분류 코드의 논리적 상(logical view)이며 t값은 색인 대상 전체에서 출현한 색인어의 총 수를 의미한다.

4. 분류 코드 후보 생성

분류 코드 후보 생성 모듈은 조사원이 자연어로 기술한 산업체와 개인의 정보를 입력받아 입력과 관련된 정보를 포함하고 있는 산업/직업 분류 코드 후보 집합을 생성하는 역할을 수행한다. 산업 분류를 위하여 입력되는 산업체와 관련된 정보는 사업체명과 사업의 주된 내용이며 직업 분류를 위하여 사용되는 입력은 직책과 하고 있는 일을 자연어로 기술한 내용이다.

분류 코드 후보 생성기는 입력 레코드와 관련 있는 후보 코드를 생성하기 위하여 벡터 공간 검색 모델을 사용한다. 즉 정보 색인 시 사용한 색인어 추출기와 가중치 계산기를 이용하여 입력 레코드의 논리적인 상인 입력 벡터를 생성하고 코사인 유사도를 이용하여 입력 벡터와 유사한 분류 코드 벡터 후보 집합을 산출한다. 입력 레코드 $\vec{q} = (w_{1q}, w_{2q}, \ldots, w_{tq})$와 분류 코드 벡터

$\vec{c_j} = (w_{1j}, w_{2j}, \ldots, w_{tj})$와의 코사인 유사도 계산식은 (식 2)와 같다.

$$\mathrm{sim}(\vec{c_j}, \vec{q}) = \frac{\vec{c_j} \cdot \vec{q}}{|\vec{c_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \quad \text{(식 2)}$$

(식 2)에서 분자는 검색 시에 계산할 수 있고, 분모에서 $\sqrt{\sum_{i=1}^{t} w_{ij}^2}$는 가중치 부여기에서 코드마다 미리 계산해 놓고, $\sqrt{\sum_{i=1}^{t} w_{iq}^2}$는 모든 코드에 동일한 값이므로 랭킹에 영향을 주지 않기 때문에 실제 계산에서는 사용하지 않는다.

분류 코드 후보 생성기는 유사도에 의하여 입력 레코드와 관련있는 분류 코드를 정렬하고 정렬된 분류 후보 코드들은 kNN(k nearest neighbors) 학습[9]을 이용한 분류기에서 입력 질의와 유사한 k개의 후보를 생성하듯 입력 레코드와 유사한 k개의 후보 코드 집합, *candidate set*을 생성한다.

5. 분류 코드 생성

분류 코드 생성기는 분류 코드 후보 집합에서 올바른 코드를 생성하는 역할을 수행하여 완전 자동 분류 또는 반자동 분류 방식으로 동작하며, 본 논문은 (식 3)과 (식 4)와 같은 DVF(discrete value function)와 SF(similarity-based function)를 정의하고 이를 분류 코드 생성 함수로 사용하고자 한다.

$$\hat{f}\,(candidate\,set\,,\,l)_p = argmax_{c \in C_l} \sum_{i=1}^{k} \zeta\,(c,\, candidate_i) \quad \text{(식 3)}$$
$$where\, \zeta\,(a,\,b) = 1 \;\; \text{if } a_l = b_l,\, 0\; otherwise$$

$$\hat{f}\,(candidate\,set\,,\,l)_p = argmax_{c \in C_l} \sum_{i=1}^{k} \zeta\,(c,\, candidate_i) \quad \text{(식 4)}$$
$$where\, \zeta\,(a,\,b) = sim\,(\vec{q}\,,\, \overrightarrow{candidate_i})$$
$$\text{if } a_l = b_l,\, 0\; otherwise$$

(식 3)과 (식 4)에서 *candidate set*은 분류 코드 후보 생성기의 출력 결과를 의미하고 l은 분류 수준(1~5)을 의미한다. a_l과 b_l은 코드열 a와 b의 처음부터 l개까지의 숫자로 이루어진 코드를 의미한다. 예를 들어 12345_3은 12345 코드열의 처음부터 3개까지의 숫자로 이루어진 123을 의미하는 것이다. p값은 분류 코드 생성기의 동작 모드를 지정하는 값으로 p가 1인 경우에는 완전 자동 분류 방식으로 동작하며, 2이상인 경우에는 상위 p개의 분류 코드를 생성하고 이 중 올바른 것을 사람이 수작업으로 선정할 수 있도록 지원하는 반자동 방식을 의미한다.

(식 3)을 사용하는 DVF 방법은 후보 코드 집합에서 동일한 c_l을 포함하는 코드 수가 많은 c_l 코드를 올바른 코드로 생성한다. 이 방법은 구현이 간단하다는 장점이 있지만 입력 레코드와 검색된 레코드간의 유사도를 코드 분류에 사용하지 못하는 문제점이 있을 수 있다. 이러한 문제점을 극복하기 위한 방법이 (식 4)에 정의한 SF 방법이며, 이 방법은 동일한 c_l로 이루어진 코드 각각과 입력 벡터 $\vec{q}$와의 유사도의 합이 큰 c_l을 올바른 코드로 생성하는 것이다. 예를 들어, 분류 후보 집합이 <표 7>과 같고, $p=3$, $l=4$라고 가정하였을 때 (식 3)과 (식 4)를 이용한 경우의 분류 코드 생성 결과를 나타내면 <표 8>과 같다.

<표 7>에서 '코드수에 의한 순위'는 (식 3)에 의하여 계산된 순위를 의미하며 '유사도 합에 의한 순위'는 (식 4)에 의하여 계산된 순위를 위미한다. 즉 (식 3)에 의해서는 1111, 2222, 3456의 3가지 코드가 최종 결과이며 (식 4)에 의해서는 1111, 3456, 1114의 결과 값이 산출된다.

<표 7> 분류 후보 코드 집합의 예

코드	유사도	코드	유사도
11110	0.5	2222	0.1
11115	0.1	22222	0.3
11119	0.6	12347	0.1
33337	0.2	4444	0.4
11118	0.2	11117	0.5
1111	0.2	1114	0.4
12345	0.5	11141	0.5
3333	0.3	44444	0.4
34567	0.3	34560	0.6
22225	0.3	34569	0.4

<표 8> 후보 순위의 예

코드	코드수	DVF에 의한 순위
1111	6	1
2222	3	2
3456	3	3
1234	2	4
3333	2	5
4444	2	6
1114	2	7
코드	유사도 합	SF에 의한 순위
1111	2.1	1
2222	0.7	5
3456	1.3	2
1234	0.6	6
3333	0.6	7
4444	0.8	4
1114	0.9	3

6. 실험 결과

제안된 시스템의 성능 평가를 위해서는 정확한 산업분류코드와 직업분류코드가 할당된 자료가 필요하다. 본 논문은 용어의 불일치 현상을 완화하기 위하여 사용하였던 기존의 코드가 부여된 산업 분류 코드와 직업 분류 코드의 일부분을 실험 자료로 사용하였다. 즉 코드가 이미 할당되어 있는 산업 분류 코드의 46,762개의 산업 분류 레코드 중 90%는 용어 불일치 현상을 완화하기 위하여 색인 데이터에 포함시켰고 나머지 10%를 실험 자료로 사용하였으며, 직업 분류 코드의 실험을 위해서도 36,286개의 직업 분류 레코드에 동일한 방법을 적용하였다. 또한 특정 실험 자료에 의존적인 결과를 지양하기 위하여 10-fold cross-validation 방법으로 성능을 평가하였다. 10-fold cross-validation은 전체 자료를 10개의 부분 집합으로 구분하여 9개의 부분 집합(90%)을 색인 자료로 사용하고 나머지 1개의 부분 집합(10%)을 실험 자료로 사용하여 성능을 평가하고, 다른 9개의 부분 집합에 대해서도 동일하게 적용하여 평가하여 총 10회의 실험 자료에 대한 성능을 계산하는 방법이다.

성능의 평가를 위해서는 생성률을 정의하여 사용하였으며, 생성률 PR_p의 정의는 (식 5)와 같다. (식 5)에서 p 는 시스템이 올바른 코드 후보로 산출하게 되는 후보 집합 크기를 나타나며 (식 3)과 (식 4)에서의 값과 동일한 의미를 갖는다. 즉 $p = 1$은 올바른 코드로 1개의 코드를 출력하는 완전 자동 분류 시의 생성률을 뜻하며, $p \geq 2$의 경우 다수개의 올바른 후보 코드를 출력하여 전문가가 수작업으로 올바른 코드를 선택할 수 있도록 지원하기 위한 반자동 분류 모드의 생성률을 의미한다. 생성률 PR_p의 의미는 전체 입력 레코드의 개수와 상위 p 개의 결과 내에 올바른 코드를 포함하고 있는 레코드의 수의 백분율을 뜻한다.

$$PR_p = \frac{\#\ of\ correctly\ assigned\ cases}{\#\ of\ input\ cases} \times 100 \qquad \text{(식 5)}$$

<표 9>는 몇 가지의 p값에 따른 산업 분류 코드의 자동 분류 결과를 나타낸 것이다. '색인어 확장 후'의 결과는 용어 불일치 문제 해결을 위한 과거 용어 집합 사용의 효용성을 알아보기 위하여 과

거 용어 집합을 색인 대상에 포함시켰을 경우의 성능을 나타낸 것이다.

< 표 9> 산업 분류 코드 자동 분류 성능

	p	2수준	3수준	4수준	5수준
색인어 확장 전	1	70.03	66.16	60.23	57.08
	2	81.04	80.91	75.02	67.23
	3	83.52	81.10	77.02	72.98
	10	84.55	83.02	80.70	78.38
색인어 확장 후	1	87.08	82.46	73.12	66.08
	2	95.16	90.91	85.14	77.01
	3	97.22	94.16	90.03	82.65
	10	99.10	98.22	95.80	92.88

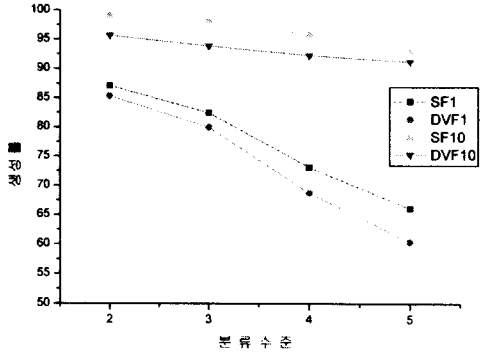
<표 9>의 결과에서 보듯이 과거 용어 집합의 사용은 산업 분류 책자 내의 용어와 실제 사람들이 사용하는 용어간의 불일치 문제를 완화하는데 매우 효과적임을 알 수 있다. 색인어 확장 후 생성률은 완전 자동 분류 모드에서는 2수준에서 87.08%, 5수준에서는 66.08%를 보였다. 완전 자동 분류 모드로 5수준에서의 성능은 그리 높은 편이 아니지만 $p \geq 2$ 이상인 반자동 모드의 경우 생성률이 급격히 증가하여 $p = 10$의 경우 2수준에서 99.10%, 5수준에서 92.88%의 높은 성능을 보였다.

<표 10>은 직업 분류 코드의 자동 분류 실험 결과를 나타낸 것이다. 직업 분류에서도 과거 용어 집합의 활용은 생성률 향상에 많은 기여를 한 것으로 나타났다. 또한 완전 자동 분류 모드의 경우 2수준에서 75.08%, 5수준에서 64.16%로 비교적 낮은 생성률을 보였지만 $p \geq 2$의 경우 생성률이 급격히 증가하여 $p = 10$인 경우 2수준에서 96.20%, 5수준에서 92.88%의 높은 생성률을 보였다.

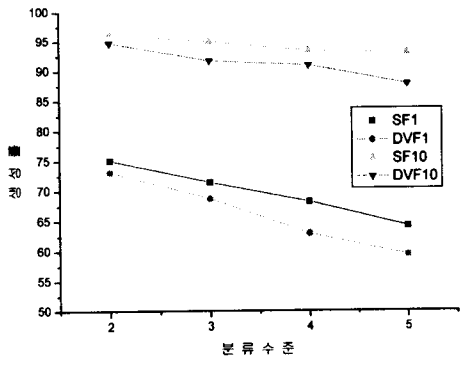
<표 10> 직업 분류 코드 자동 분류 성능

	p	2수준	3수준	4수준	5수준
색인어 확장 전	1	66.03	60.25	55.83	47.30
	2	71.04	65.72	58.45	57.23
	3	73.52	70.84	62.71	61.72
	10	74.55	73.99	78.80	73.27
색인어 확장 후	1	75.08	71.36	68.12	64.08
	2	84.67	78.56	85.14	77.01
	3	89.14	83.65	90.03	82.65
	10	96.20	94.86	93.29	92.88

(그림 3)과 (그림 4)는 코드 생성 함수인 DVF와 SF의 성능을 비교하기 위하여 직업 분류 코드와 산업 분류 코드 분류에 두 방법을 적용한 결과를 나타내고 있다. 그림에서 SFn은 $p = n$일 때 SF 방법에 의한 결과 값을 나타내며, DVFn은 $p = n$일 때 DVF 방법에 의한 결과 값을 나타내고 있다. 실험 결과, p의 값과는 상관없이 모든 경우에 있어서 입력 레코드와 분류 코드와의 유사도를 코드 분류에 이용하는 SF 방법이 우수한 성능을 보임을 알 수 있었다.



(그림 3) 분류 함수의 성능비교:산업분류코드 실험



(그림 4) 분류 함수의 성능비교:직업분류코드 실험

올바른 코드를 생성하지 못한 경우를 분석한 결과 입력 레코드에 존재하는 띄어쓰기 오류와 한 음절 명사의 색인어 추출 실패가 가장 주요한 원인임을 알 수 있었다. 입력 레코드에 존재하는 띄어쓰기 오류는 조사원들의 조사 내용을 입력 전담 인력들이 입력할 때 발생하는 오류로 실험 자료에 많은 띄어쓰기 오류가 존재하였다. 띄어쓰기 오류로 인한 성능 저하 문제는 자동 분류 작업 수행 이전에 띄어쓰기 자동 교정기 등을 이용한 전처리 작업으로 해결 가능할 것으로 생각된다. 두 번째 주요한 원인이었던 한 음절 명사의 색인어 추출 실패는 색인어 추출기의 명사 추출기에서 기인한 것으로 기본적으로 명사 추출기는 2음절 이상의 명사만을 추출하는 것을 가정하고 있기 때문이다. 명사 추출기가 2음절 이상의 명사만을 추출하는 이유는 한국어의 경우 거의 모든 1음절어가 명사에 해당되므로 추출할 색인어의 수가 과도하게 많아질 뿐만 아니라 검색 정확률을 저하시키는 결과를 초래할 수 있기 때문이다. 1음절 명사에 의한 자동 분류 실패의 문제를 극복하기 위해서는 1음절 명사 추출 방법 또는 다른 대안에 대한 추가적인 연구가 필요할 것이다.

7. 결론

본 논문은 인구주택 총조사와 사업체 기초통계 조사 시 수작업으로 분류하던 표준 산업/직업 코드를 자동으로 분류할 수 있는 정보검색 기법을 이용한 산업/직업 코드 자동 분류 시스템을 제안하였다. 제안한 시스템은 수작업 코딩 작업 시 작업자가 참조하는 정보를 색인하여 색인 DB로 구축하는 색인 시스템, 개인의 직업 정보나 사업체의 사업 정보 등을 자연어로 기술한 내용을 입력 받아 입력된 내용과 관련 있는 다수개의 분류 코드를 검색하는 분류 코드 후보 생성기, 그리고 분류 코드 후보 생성기의 입력을 받아 올바른 분류 코드를 생성하는 분류 코드 생성기로 구성된다. 색인기는 명사 추출기를 이용하여 색인어를 추출하고 정보검색에서 가중치 계산을 위하여 전통적으로 사용하는 TF/IDF 기법을 이용하여 가중치를 계산하여 역화일 형태로 색인 DB를 구성한다. 분류 코드 후보 생성기는 코사인 유사도를 이용하여 입력 코드 벡터와 유사한 코드 벡터 집합을 결과로 생성한다. 분류 코드 생성기는 분류 코드 후보 집합에서 입력 레코드에 해당하는 분류 코드를 생성하는 역할을 하며 완전 자동 모드와 반자동 모드로 동작이 이루어지며, 본 논문은 분류 코드 생성을 위한 DVF와 SF를 정의하여 사용하였다.

제안한 방법에 대한 실험은 수작업으로 산업/직업 분류 코드가 할당되어 있는 46,762개의 산업 분류 레코드와 36,286개의 직업 분류 레코드를 이용하여 10-fold cross-validation evaluation 방법으로 수행되었으며, 생성률 값으로 성능을 평가하였다. 실험 결과, 입력 레코드와 분류 코드간의 유사도를 코드 분류에 이용하는 SF 방법이 DVF 방법보다 우수한 성능을 보였으며, 산업 분류 책자와 입력 레코드 간의 용어 불일치 현상을 완화하기 위하여 과거 용어 집합을 사용한 방법이 매우 효과적임을 알 수 있었다. 현재 코드 자동 분류기는 산업 분류 코드에 대해서 완전 자동 모드로 동작할 때 2수준의 코드 생성 시 87.08%, 5수준에서는 66.08%의 성능을 보이며, $p = 10$일 때 2수준에서 99.10%, 5수준에서 92.88%의 생성률을 보이고 있다. 직업 분류 코드에 대해서는 완전 자동 모드에서 2수준의 코드 생성 시 75.08%, 5 수준에서 64.08%의 성능을 보이며, $p = 10$일 때 2수준에서 96.20%, 5수준에서는 92.88%의 생성률을 보이고 있다. 올바른 코드를 생성하지 못한 이유를 분

석한 결과, 입력 레코드에 존재하는 띄어쓰기 오류와 1음절 명사의 색인어 추출의 실패가 주요한 원인임을 알 수 있었는데, 이러한 문제점 해결을 위한 방법으로 띄어쓰기 교정기를 이용한 전처리기 사용과 1 음절 명사의 색인어 추출에 대한 추후 연구가 수행되어야 할 것으로 생각된다.

제안한 시스템은 아직 사람의 수작업을 대체할 수 있는 완전 자동 분류기로 사용하기에는 성능의 개선이 요구되지만, ㄷ값을 5 또는 10으로 설정하여 분류기의 결과 중 올바른 코드를 수작업으로 분류할 수 있도록 하여 수작업의 양을 최소화시킬 수 있는 반자동 도구로 활용하거나 수작업에 의한 코드 분류의 검증을 위하여 충분히 사용될 수 있을 것으로 기대된다.

참고문헌

[1] Apeel, M. V. and Hellerman, E., Census Bureau Experiments with Automated Industry and Occupation Coding, Proceedings of the American Statistical Association, 32-40, 1983.

[2] Rowe, E. and Wong, C., An Introduction to the ACRT Coding System, Bureau of the Census Statistical Research Report Series No. RR94/02 (1994)

[3] Gilman, D. W. and Appel, M. V., Automated Coding Research At the Census Bureau. U.S. Census Bureau, http://www.census.gov/srd/papers/pdf/rr94-4.pdf

[4] Do-Gil Lee, Hae-Chang Rim, and Heui-Seok Lim, "A Syllable Based Word Recognition Model for Korean Noun Extraction", Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), pp.471-478, 2003.

[5] Baeza-Yates and Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, 1999.

[6] Salton, G. and McGill, M.J., Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

[7] Chen, B., Creecy, R. H., and Appel, M. On Error Control of Automated Industry and Occupation Coding, Journal of Official Statistics, Vol. 9, No. 4, 729-745, 1993.

[8] Creecy, R. H., Masand, B. M., Smith, S.J., and Walts, D. L., Trading MIPS and Memory for Knowledge Engineering, Communications of the ACM, Vol. 35, No.8, 48-64, 1992.

[9] Tom M. Mitchell, Machine Learning, Mc Graw Hill, 1997.

임 희 석

1992. 2 : 고려대학교 컴퓨터학과(학사)
1994. 2 : 고려대학교 컴퓨터학과(석사)
1997. 9 : 고려대학교 컴퓨터학과(박사)
1997. 9~1999. 2 : 삼성종합기술원 전문연구원
1999. 3~2004. 2 : 천안대학교 정보통신학부 교수
2004. 3~ 현재 : 한신대학교 소프트웨어학과 교수
관심분야 : 자연어처리, 정보검색, 인공지능, 인지신경과학
E-Mail : limhs@hs.ac.kr