정보검색 기법을 이용한 산업/직업 코드 분류 도구

○
임희석*, 박두순**
*천안대학교 정보통신학부
**순천향대학교 정보기술공학부

An automatic Industrial/Occupational Code Classification Tool Using Information Retrieval Technique

○
Heui-Seok Lim*, Doo-Soon Park**
*Dept. of Information and Communications, CheonAn University
*Dept. of Information Technique Engineering, SoonChunYang University

요약

본 논문은 통계청에서 실시하는 인구주책 총조사로부터 획득된 각 개인의 직업 및 직종을 기술하고 있는 자연어를 입력받아 입력된 자연어가 의미하는 한국 표준 산업/직업 분류 코드의 후보들을 생성하는 산업/직업 코드 분류 도구를 제안한다. 코드 분류는 분류할 코드를 문서 범주로 간주하면 문서 분류와 동일한 문제로 생각할 수 있다. 하지만 본 산업/직업 코드 분류 문제는 입력되는 자연어의 길이가 한 두 문장 정도로 매우 짧아 문서 분류에 사용될 자질들이 개수가 적어 기존의 문서 분류 기법을 적용하기 어렵다. 이에 본 논문은 표준 코드를 기술하고 있는 내용을 미리 색인하고 입력된 자연어로부터 질의어를 생성하여 벡터공간모델로 질의어를 검색후 질의어와 일치율이 가장 높은 코드들을 분류될 후보 코드로 제시하는 정보검색 기법을 이용한 산업/직업 코드 분류 도구를 개발하였다.

1. 서론

통계청에서는 매 5년(0년, 5년)마다 인구주택 총조사를 실시하고 있다. 인구주택 총조사를 통하여 얻어지는 표준산업분류 코드와 표준직업분류 코드는 국가의 기본 정책을 수립하는데 있어서 매우 중요한 지식으로 사용된다. 현재까지 표준산업분류 코드와 표준직업분류 코드의 분류 작업은 가구 조사에서 조사원이 얻은 자연어로 쓰여진 개인이 근무하고 있는 사업체명, 사업체의 주된 사업 내용, 직책, 그리고 직무의 내용과 한국 표준산업/직업 분류 책을 이용하여 코드 분류 전문가에 의해서 수작업으로 진행되어 왔다. 이런 수작업은 아래와 같은 문제점을 갖고 있다.

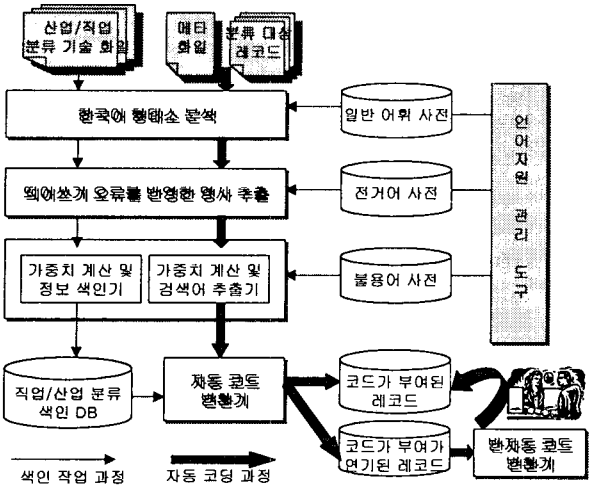
- 수작업을 수행하기 위한 작업자 교육 및 활용에 많은 비용이 소요
- 매년 반복 작업에 따른 인력/비용 소모
- 막대한 수작업 량과 고비용 발생
- 오랜 시간 소요

산업/직업 코드를 문서 분류시 할당될 문서의 범주로 간주하면 산업/직업 코드 분류 작업은 문서 분류 작업으로 생각할 수 있다. 기존의 문서 분류 기법을 사용하기 위해서는 문서를 표현할 수 있는 자질들(feature)로 문서를 벡터화할 필요가 있다. 하지만 산업/직업 코드 분류 작업에서는 입력으로 사용되는 자연어의 길이가 매우 짧아 입력된 자연어를 벡터화하기 위하여 사용될 수 있는 자질의 개수가 매우 적어 기존의 문서 기법을 사용하기에 어려움이 따른다. 이에 본 논문은 분류될 코드를 기술하고 있는 내용을 미리 색인하여 색인 DB를 구축하고, 입력된 자연어로부터 검색어를 추출하여 벡터공간모델로 검색된 결과를 분류될 후보 코드로 사용하는 정보검색 기법을 이용한 산업/직업 코드 분류 도구를 제안한다.

본 연구는 정보통신부의 ITRC 사업에 의해 수행된 것임

2. 시스템 개요

[그림 1]은 본 시스템의 전체적인 시스템 구성도를 나



[그림 1] 시스템 개요

타내고 있다. 색인 시스템과 검색 시스템은 색인어 및 검색어를 추출하기 위한 한국어 형태소 분석기, 명사 추출기[1] 및 관련 언어 자원들로 구성된다.

색인 시스템은 산업/직업 분류 코드를 설명한 분류표[5]를 색인어의 가중치 값을 계산하여 역화일로 색인한다. 검색 시스템은 코드 분류를 위하여 사용되며 조사원들로부터 조사된 '사업체명', '사업체의 주된 사업내용', '부서 및 직책', '하고 있는 일의 종류'를 기술한 자연어를 입력받아 구성한 질의어에 의하여 검색된 결과를 일치율에 따라 정렬된 코드를 생성한다.

본 시스템은 검색 시스템에 의하여 정렬된 코드 중 일치율이 최상인 코드 하나만을 생성할 경우 완전 자동 코드 분류 시스템으로 사용될 수 있으며, 상위 몇 개의 코드들을 출력하여 그 결과중에서 올바른 코드를 찾아 할당함으로써 수작업량을 감소시키며 작업 속도를 증가시킬 수 있는 반자동 코드 분류 도구로 활용될 수 있다.

3. 색인 및 검색

3.1. 색인 데이터

코드를 색인하기 위하여 사용되는 색인 데이터는 통계청에서 제공한 한국 표준산업 분류 책자[]와 한국 표준직업 분류 책자의 텍스트 데이터, 표준 코드를 대표할 수 있는 색인어 목록으로 구성된다. 한국 표준 산업/직업 분류 책자의 텍스트 데이터는 표준 코드, 코드에 해당하는 산업/직업명, 코드의 산업/직업에 대한 설명, 예시, 제외 등의 데이터를 포함하고 있다. 색인어 목록은 특정 코드의 색인어로 사용될 가능성이 높은 명사 및 명사절을 "코드 색인어" 형식으로 기술한 레코드의 집합으로 현재 산업 분류를 위하여 23,431 레코드와 직업분류를 위한 17,184개의 레코드가 사용된다.

분류 책자 데이터와 색인어 목록을 이용하여 색인어 DB를 구성하고 코드 분류 작업을 수행한 결과 예상하던 것보다 매우 저조한 성능을 보였다. 그 이유를 분석한 결과 저조한 성능의 가장 큰 이유는 용어의 불일치에 의한 것이었다. 표준 산업/직업 분류 코드의 책자에서 코드를 기술하기 위하여 사용된 용어들은 해당 코드와 관련된 많은 용어들을 포함하고 있지 못하다. 따라서 조사원들이 기입한 데이터에서 추출한 용어와 책자에서 추출하여 구성한 색인 DB안의 용어들이 불일치하는 경우가 많이 발생한다. 본 논문은 이와 같은 용어 불일치 문제를 완화하기 위하여 '과거 용어 집합'을 사용한다. '과거 용어 집합'은 과거 연도의 조사 결과로 코드가 정확하게 부여된 레코드에서 코드, 사업체명, 사업체의 주요 업무, 직무, 하고 있는 일의 종류 등 4가지로 구성된다.

3.2. 색인기

색인기는 형태소 분석 및 명사 추출기에 의해서 얻어진 색인어들을 이용하여 역화일을 구성하고 색인어들의 가중치를 부여하는 역할을 수행한다[3,4]. 색인어는 검색이 용이하도록 역파일구조로 색인하며 색인 DB는 1) 산업/직업 분류 코드의 일련 번호, 2) 코드의 설명, 3) Posting File의 시작위치, 4) 코드의 빈도 등의 정보를 포함한다. 산업/직업 분류 코드의 일련 번호와 코드의 설명은 산업/직업 분류 책자내에 기술되어 있는 코드의 일련 번호와 설명을 의미하며 Posting File의 시작 위치와 코드의 빈도는 아래

에 설명한다.

3.2.1. Posting File의 시작 위치

Posting File의 시작 위치는 색인어가 나타난 코드와 각 코드에서의 빈도를 저장하고 있는 파일이 posting file 이다. 같은 색인어라도 코드마다 그 코드에서의 중요도가 달라지기 때문에, 색인어가 나타난 코드마다 따로 빈도를 저장하고 있어야 한다.

색인단계에서는 색인어가 몇 개의 코드에서 나타날지 알 수 없고, 색인어마다 따로 posting file을 만드는 것도 불가능하기 때문에, 이 posting file이 하나의 파일에 linked list구조로 만들어진다. 이때 만들어지는 posting file의 구조는 다음 그림과 같다.

CodeID	Freq	Next

CodeID는 코드마다 부여되는 고유번호이고, Freq는 CodeID코드에서 색인어가 나타난 빈도이고, Next는 다음에 색인어가 나타난 정보를 저장하고 있는 위치이다. 이 Next정보는 속도를 위해서 사실은 앞에 나타난 정보의 위치를 가리키게 된다. 예를 들어, 다음 그림과 같은 Posting file은 다음과 같은 정보를 가지고 있다.

위치	0	1	2	3	4	5	6	7	8
정보	1	3	-1	1	2	-1	2	5	0

색인어 A가 Posting file의 시작위치로 6을 가지고 있다면, 색인어 A는 2번 코드에 5번 1번 코드에 3번 나타났다는 것을 알 수 있다.

3.2.2. 코드 빈도

코드 빈도란 색인어가 나타난 코드의 개수를 말한다. 색인어의 중요도를 계산하기 위해서 필요하고, 개선된 Posting file의 빠른 탐색을 위해서도 필요하다. 코드 빈도가 높은 색인어는 코드의 특성을 설명하는데 좋지 않으므로 중요도가 떨어지게 된다.

3.3. 가중치 부여기

가중치 부여기는 색인어가 해당 코드를 식별하는데 얼마나 중요하게 사용되는지를 나타내는 중요도를 계산하여 색인어에 할당하는 역할을 수행한다. 중요도를 계산하는 척도로는 색인어의 코드 내 빈도와 코드간 빈도를 사용하며 색인어의 중요도를 계산하는 식은 다음과 같다[3,4].

$$w_{ij} = f_{ij} \times \log \frac{N}{n_i}$$

여기서, w_{ij}는 색인어 i가 코드 j에서의 중요도이고 f_{ij}는 코드 j에서 색인어 i가 나타난 빈도이고, N은 전체 색인된 코드의 개수이며, n_i는 색인어 i가 나타난 코드 빈도이다.

3.4. 자동 코드 변환

자동 코드 변환기는 색인 시스템에서 색인해 놓은 정보를 이용하여 코드를 검색하여 일치율에 의해서 내림차순으로 정렬된 결과를 생성한다. 본 시스템은 벡터공간모델(vector space model)을 사용하여 사용자가 입력한 질의에 적합한 코드를 검색하도록 한다. 벡터공간모델은 사용자 질의와 코드를 벡터로 표현하고, 이 두 벡터 사이의 유사도를 이용하여 코드를 검색해낸다. 사용자 질의 $\vec{q} = (w_{1q}, w_{2q}, \ldots, w_{tq})$와 코드$j$ 벡터 $\vec{d_j} = (w_{1j}, w_{2j}, \ldots, w_{tj})$와의 유사도는 아래의 수식으로 계산한다.

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

위 수식에서 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 계산하지 않는다.

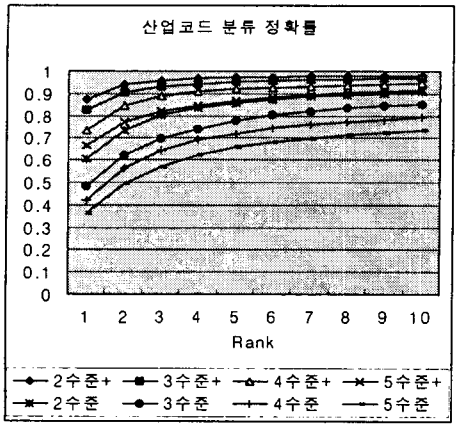
질의를 분석하고, 모델에 따라서 검색을 하게 되면 코드들이 질의에 적합한 정도가 가중치로 표현이 된다. 사용자에게 가장 적합한 코드를 먼저 보여주기 위해서 이 가중치에 따라서 코드를 정렬한다. 코드의 정렬에는 quick sorting algorithm을 사용하였다.

4. 실험 및 결과

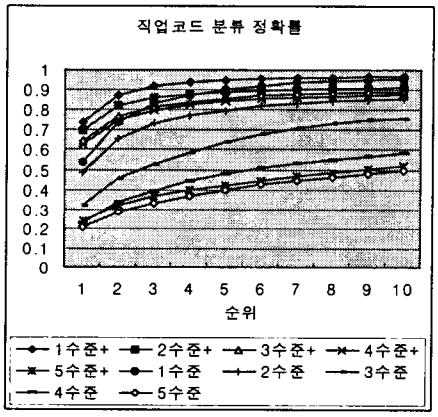
본 논문은 제안한 시스템의 평가를 위하여 산업 분류 코드가 이미 부여된 46,762개의 레코드와 직업 분류 코드가 이미 부여된 36,287개를 실험 집합으로 사용하였다. '과거 용어 집합'을 구축하기 위하여 실험 집합의 90%를 학습 데이터로 사용하여 색인 DB에 색인하고 나머지 10%를 테스트 집합으로 사용하여 실험하였다. 실험 방법은 교차평가(cross-evaluation)을 위하여 테스트 집합을 10개를 구축하여 총 10회의 실험을 수행하였다. 평가를 위하여 사용한 기준은 코드 생성율과 정확률을 사용하였으며 다음과 같이 계산된다

$$\text{생성율} = \frac{\text{코드가 할당된 레코드 수}}{\text{입력 레코드 수}}$$

$$\text{정확률} = \frac{\text{정확한 코드를 포함하는 레코드 수}}{\text{코드가 할당된 레코드 수}}$$



[그림 2] 산업코드 분류 결과



[그림 3] 직업코드 분류 결과

산업/직업 코드는 0, 01, 011, 0111, 01111과 같이 1수준에서 5수준까지 계층적으로 구성이 되어 있다. 따라서 코드를 분류할 때 어느 수준까지 결정할 것인가가 코드 분류기이 파라미터로 사용될 수 있다. [그림 2]와 [그림 3]은 각각 산업 코드 분류와 직업 코드 분류 결과의 정확률을 보이고 있다. 여기에서 'x수준+'는 '과거 용어 집합'을 사용하여서 x수준으로 분류한 결과를 나타내며 'x수준'은 과거 용어 집합을 사용하지 않은 결과를 나타낸다. 또한 모든 정확율은 테스트 집합 10를 이용한 실험 결과의 평균값을 나타낸다. 'Rank'는 검색 결과 상위 몇번째 순위의 결과 값까지를 정확률 계산에 포함한 것인가를 나타낸다[1].

위의 결과를 보면 산업코드와 직업 코드 분류에 있어서 과거 용어 집합을 사용하였는가와 상관없이 Rank값이 올라갈수록 정확률이 증가함을 알 수 있다. 도한 산업 코드 및 직업 코드를 분류하기 위하여 '과거 용어 집합'을 사용한 경우가 사용하지 않은 경우에 비하여 정확도가 월등히 높아짐을 알 수 있다. [그림 2]의 산업코드 분류 결과에서는 과거 용어 집합을 이용하여 5수준까지 분류한 정확률이 과거 용어 집합을 사용하지 않은 경우의 2수준까지의 분류 정확률보다도 더 높았음을 볼 수 있다. 이러한 결과로 산업/직업 코드 분류에 있어서 용어 불일치 현상을 완화하기 위한 과거의 조사 데이터를 사용한 방법이 매우 효과적이었음을 알 수 있다.

5. 결론

본 논문은 정보 검색 기법을 이용하여 자연어 입력에 해당하는 산업/직업 분류 코드 후보를 생성하는 산업/직업 코드 분류 도구를 제안하였다. 제안된 시스템은 질의어의 검색 결과를 일치율이 가장 높은 코드를 생성하게 함으로써 완전 자동 산업/직업 코드 분류기로 사용할 수 있다. 하지만 아직까지 완전 자동 도구로 사용되기에는 정확율이 낮으므로 검색된 상위 일정개수의 코드를 수작업자들에게 제시하여 올바른 코드를 빠른 시간에 정확하게 찾을 수 있도록 지원하는 도구로 사용하는 것이 효과적일 것이다. 완전 자동화 도구로 개발하기 위해서는 코드 생성율이 낮아지더라도 정확률을 높일 수 있는 방법을 고안되어야 할 것이다. 즉 코드 생성율이 70%이더라도 정확율이 99%에 가까운 성능을 보인다면 30%의 양만을 수작업을 수행하여 정확율 99%에 가까운 코드 분류 결과를 생성할 수 있다. 현재 자동 코드 분류기로 사용할 수 있도록 정확율 향상을 연구하고 있다.

참고문헌

[1] 이도길, *명사 출현 환경을 고려한 빠른 색인어 추출 시스템*, 고려대학교 컴퓨터학과 석사학위논문, 2000.

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

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

[4] Baeza-Yates, Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.

[5] 통계청, 한국 표준 산업/직업 분류 코드, 2000.

1) 실험 결과 생성율은 산업/직업 코드의 여부, 과거 용어 집합을 사용한 여부, Rank에 상관없이 0.974~1.000의 높은 결과를 보였으며 생성율의 결과는 지면 관계상 생략하였다.