

빅데이터 기법을 활용한 직업훈련 요구분석

성보경¹, 유연우^{2*}

¹한성대학교 스마트융합컨설팅학과 박사과정

²한성대학교 스마트융합컨설팅학과 교수

Analysis of Vocational Training Needs Using Big Data Technique

Bo-Kyoung Sung¹, Yen-Yoo You^{2*}

¹Doctoral Student, Dept. Of Smart Convergence Consulting, Hansung University

²Professor, Division Of Smart Management Engineering, Hansung University

요 약 본 연구는 고용노동부가 운영하는 직업훈련 통합전산망인 'HRD-NET(<http://hrd.go.kr>)'을 통해 구직자가 필요로 하는 직업훈련 정보 등이 원활하게 제공되고 있는지를 확인하기 위해 질문게시판을 빅데이터 기법에 가장 최적화된 'R' 프로그램을 이용해서 추출하였다. 따라서, 이를 통해 직업훈련제도의 유효성, 적절성, 시각화, 빈도 분석, 연관분석 등을 실시하였으며, 연구결과는 다음과 같다. 첫째, 직업훈련 카드발급 및 동영상 시청, 공인인증서 문제, 등록오류 이 발견되었으며, 둘째, 내일배움카드에 대한 노동관서에서의 관리 및 처리절차가 복잡하고 까다로워 제도개선이 필요한 것으로 나타났다. 또한, 교육훈련의 수강에 있어 훈련지종 및 과정, 훈련기관에 따라서 차등화 된 훈련비 시스템과 환급구조가 예외요인으로 작용하는 것으로 분석되었다. 본 논문 기초로 하여 향후 고용노동부의 훈련시스템 뿐만 아니라 정부부처의 다양한 훈련 전산망 시스템 에 대한 전반적인 빅데이터 분석을 통한 개선점 등을 연구하고자 한다.

주제어 : 융합, 직업훈련, 고용노동부직업훈련전산망, 구직자, 실업자교육, 텍스트마이닝

Abstract In this study, HRD-NET (<http://hrd.go.kr>), a vocational and training integrated computer network operated by the Ministry of Employment and Labor, is used to confirm whether job training information required by job seekers is being provided smoothly The question bulletin board was extracted using 'R' program which is optimized for big data technique. Therefore, the effectiveness, appropriateness, visualization, frequency analysis and association analysis of the vocational training system were conducted through this, The results of the study are as follows. First, the issue of vocational training card, video viewing, certificate issue, registration error, Second, management and processing procedures of learning cards for tomorrow 's learning cards are complicated and difficult. In addition, it was analyzed that the training cost system and the refund structure differentiated according to the training occupation, the process, and the training institution in the course of the training. Based on this paper, we will study not only the training system of the Ministry of Employment and Labor but also the improvement of the various training computer system of the government department through the analysis of big data.

Key Words : Convergence, Vocational Training, HRD-NET, Job seeker, Unemployed Education, Text Mining

*This research was financially supported by Hansung University.

*Corresponding Author : Yen-Yoo You(threey0818@hansung.ac.kr)

Received March 13, 2017

Revised April 26, 2018

Accepted May 20, 2018

Published May 28, 2018

1. 서론

1.1 배경 및 목적

산업의 발달과 경제적인 성장으로 인해 기술교육에 대한 인식이 재인식 되고 있으며(김정숙, 2011)[1], 경제 침체여파로 취업난이 계속되면서 대학을 졸업하고도 거꾸로 직업교육을 받기 위해 직업학교에 입학하는 구직자가 매년 증가하였고(이인희, 2003)[2], 고학력자들의 실업으로 인해 기술교육에 대한 필요성을 인식하게 되면서 기술교육을 위한 정부의 교육비 지원시스템이 활성화 되었다. 2008년부터 시범 도입된 ‘내일배움카드제’는 구직자 및 취약근로자를 대상으로 1인당 일정금액을 구직에 관련된 교육에 사용할 수 있도록 노동부에서 다양한 훈련과정이 세분화 하여 운영하고 있다. (김주영, 2010)[3]. 최근 온라인 커뮤니티를 통하여 네티즌들의 다양한 의견이나 경험, 지식 등을 표현한 비정형화된 텍스트 데이터 형태의 고객 리뷰들이 방대하게 존재하고 있고 증가하고 있는 추세이며, (김근형, 2011)[4]. 네티즌 또는 고객들의 품평에 대해 신속하게 관리하는 기업이나 조직들은 대량의 온라인 상품평이나 고객 리뷰에서 상품이 나 서비스의 속성들에 대한 고객들의 주관적 의견을 상품의 개선 사항을 도출하기도 하고, 고객의 인식 변화에 발 빠르게 대처하고 있는 기업은 트위터와 인터넷에 올라온 기업 관련 댓글을 실시간으로 분석하여 자사 이미지를 파악하고 대응전략을 세우고 있다(정용찬, 2011)[5]. 따라서 고용노동부가 운영하는 직업능력개발종합정보망 ‘HRD-Net’은 중앙정보고용원에서 운영하는 인적자원개발 종합정보망으로서 직업능력개발에 대한 수요자의 다양한 요구를 반영하는 종합정보망의 기능을 수행하고 있으며,(김석진, 2013)[6]. 구직자 입장에서 필요로 하는 직업훈련 정보, 홍보 및 안내사항, 참여방법, 훈련기관 및 훈련의 종류, 취업처 정보 등 실질적으로 구직자에게 필요로 하는 정보가 내재되어 있다. 그러나 직업훈련 수강자들은 끊임없이 운영상의 문제점 등을 제기하고 있으며, 이를 정확하게 파악하여 어떠한 문제점이 주요이슈인지 확인하여 향후 직업훈련 전산망 운영 및 직업훈련의 방향성을 수요자의 니즈를 적극적으로 반영할 필요성이 있다고 하겠다.

1.2 연구의 범위

본 연구에서는 고용노동부의 직업훈련 전산망인

‘HRD-NET’의 질문게시판 크롤링 하여 빅 데이터 분석을 시행하였다. 이는 HRD-NET 전산망의 유효성 및 적절성을 판단하고 전산망이 개편된 2012년부터 2017년까지의 6년간의 질문 키워드를 시계열 분석을 통해 수요자가 궁금해 하고 필요로 하는 부분이 어떤 것인지 파악하기 위함이다. 궁극적으로 본 연구는 직업훈련전산망을 이용하는 훈련수요자의 요구사항을 분석하였다. 특히, 빅 데이터를 이용한 연구방법은 기존 연구방법들이 지녔던 표본의 한계성을 극복하고 더욱 정확성이 높은 결과를 제시할 가능성이 높다. (김경애 외, 2017)[7]. 고용노동부가 추진해 온 직업훈련서비스에 대한 사람들의 VOC(Voice of customer) 질문 빅데이터를 추출하여 직업능력개발 정책 서비스에 대한 이슈 및 요구수요 등을 파악하여 시행의 문제점 과 각각의 키워드의 연관분석을 통한 연계된 시스템 및 동시에 요구되는 시스템 등을 파악하고자 한다.

1.3 연구의 방법

세계경제포럼은 가장 주목할 기술로 빅데이터(big data)를 지목하였다(전채남·서일원, 2013)[8]. 빅데이터는 대용량의 데이터로 저장, 수집, 발굴, 분석하는 일련의 과정을 거쳐 현상파악, 미래예측, 패턴 분석, 마케팅 등 다양한 분야에서 도입기 단계를 활용하고 있다(유인호, 2012)[9]. 스마트시대의 도래로 인한 통신수단의 획기적인 발달로 비정형 데이터가 전체데이터의 80%를 상회하는 것이 현실이며, (송민구·김선배, 2013)[10]. 또한, 빅데이터는 실시간 생성되는 대규모의 데이터의 텍스트, 사진, 동영상 등의 비정형 데이터를 포함한 높은 다양성의 특징을 지닌 정보 자산이다(Douglas, 2012)[11]. 특히 빅데이터의 텍스트 마이닝 기법은 통계적인 접근을 통하여 개념 간의 연결성과 영향 관계를 파악하고 이를 시각화함으로써 데이터가 보여주는 의미를 도출하는 데에 유용하다고 하였다(김선아 외, 2016)[12]. 따라서 비정형화된 텍스트 데이터인 교육 훈련생의 질문게시판 키워드 추출을 통해서 그 의미분석을 통해서 문제점, 개선점, 시사점 등을 도출하고자 한다.

본 연구를 위한 분석 자료의 수집은 빅데이터 분석 솔루션인 R시스템을 통해 진행되었다. 구체적으로 분석 자료를 수집하기 위한 키워드는 HRD-NET 질문게시판 자료로써 2012년부터 2017년 11월 현재까지의 23,112개의 전처리 작업을 거친 최종 키워드를 도출하였다. 전체 기

간에 걸쳐 가장 이슈가 되는 키워드를 분석하여 시각화하여 제시하였다. 최종적으로는 전체 키워드 들 간의 연관분석을 통해서 선행키워드와 이에 따른 부속 워드를 분류함으로써 향후 시스템 개선 등에 타당성과 신뢰성을 줄 수 있는 분석을 시행하였다.

2. 용어의 정의

2.1 텍스트 마이닝(Text Mining)

데이터 마이닝은 방대한 양의 자료 속에서 의미 있는 패턴과 규칙을 찾아내기 위해 자동적이거나 반자동적인 방식으로 자료를 탐색하고 분석하는 것을 말하며, 이 중 텍스트 형태의 데이터를 분석 가능한 형태로 만들기 위하여 형태소 분석 및 자연어 처리 기술 등이 활용되고 있는데, 이를 텍스트 마이닝 기술이라고 한다(전채남, 2016)[13].

비정형 텍스트 데이터 분석은 텍스트 마이닝 기술을 이용한 내용 분석(Context Mining) 과 네트워크 분석 기술을 이용한 구조 분석(Structure Mining)으로 구분할 수 있는데, 내용 분석은 전처리 과정에서 형태소 분석, 불용어 처리를 통해 정제된 데이터를 갖고 분류·군집·주제식별·개념체계 자동구성 등을 수행하며, 구조 분석의 경우 언어 네트워크 분석(Semantic Social Network Analysis)을 대표적이라 할 수 있는데 이는 데이터에 나타난 단어의 공유된 의미를 토대로 체계적 구조를 분석하는 데 주안점을 두고 있다. 즉 핵심 단어 사이의 의미론적 연관을 통해 텍스트 데이터로부터 구조화된 형태의 정보를 추출함으로써 단어의 맥락적 의미까지 고려하여 전체 데이터에 대한 구조화된 자료를 시각적으로 나타낼 수 있다(박치성 외, 2013)[14].

2.2 의미분석(Semantic analysis)

‘의미분석 (semantic analysis)은 자연 언어 이해 기법의 하나로, 문장의 의미에 근거해서 그 문장을 해석하는 방법이며, 문장이 어떻게 구성되었는지를 나타내 주는 규칙들로 구성된 일종의 형식시스템 이다.’ 라고 정의되어 있다(위키백과, 2016)[15]. 또한, ‘의미분석은 단편적으로는 문장을 구성하는 단어들의 의미를 구분하고, 통합적으로는 문장 구성 성분들 사이의 의미적 관계(agent-predicate-object)를 논리적으로 밝혀내어 문장의 전체적

의미를 파악하는 것이다(황영숙, 2013)[16].

2.3 R프로그램

소셜 미디어 등 비정형 데이터의 증가로 인해 분석기법들 중에서 텍스트 마이닝, 오피니언 마이닝, 소셜네트워크 분석, 군집분석 등은 기존의 SPSS같은 정형데이터 분석 툴로는 분석할 수 없는 문제점을 개선한 프로그램이다. 본 논문의 경우에도 시각화 및 연관분석은 R프로그램이 가장 손쉽게 최적화를 시켜준다. R은 데이터 분석을 위한 통계 분석도구로서 다양한 분야의 패키지들을 다운로드 하여 사용할 수 있어 빅데이터 활용 도구로서 많이 사용되고 있다고 하였으며, 오픈 소스 프로젝트 R은 통계 계산 및 시각화를 위한 언어 및 개발환경을 제공하며, R 언어와 개발 환경을 통해 기본적인 통계 기법부터 모델링, 최신 데이터 마이닝 기법까지 구현·개선이 가능한 것이 차별점 이다. (김효종 외, 2017)[17].

3. 자료수집 및 가공

3.1 자료수집

인터넷에서 제공되는 수많은 웹문서를 일정한 간격마다 자동으로 탐색하고 수집하는 기술을 웹크롤러 라고 한다. (이지희 외, 2013)[18]. 따라서 본 자료는 웹크롤러를 이용하여 고용노동부의 직업훈련 통합 전산망인 HRD-NET (<http://hrd.go.kr>) 의 2012년도부터 2017년 11월 현재까지의 고객센터- 질문게시판 데이터를 크롤링 하여 사용하였다.

3.2 자료가공

HRD-NET을 통해 크롤링 된 자료는 확장자 가 txt인 ASCII형식의 파일로 저장되어 R Studio에서 인터프리터 방식으로 읽어서 실행하였다. 데이터는 자연어 상태에서 불용어 처리 등 전처리 과정을 수행하여 필터링을 수행하였으며, 이때 gsub 함수를 활용하여 숫자, 특수문자, 웹사이트에 메뉴 등을 제거하여 최종적으로 필요한 단어만 추출하였다. 이후에 KoNLP, wordcloud, arules, igrph, arulesViz 패키지 등을 활용하여 키워드 분석과 연관분석을 실시하였다. 시각화된 자료는 연도별 핵심 키워드들로 확인 가능하도록 했으며, 최종적으로 전시기에 걸쳐 가장 연관성이 높은 키워드를 분석하였다.

4. 분석결과

4.1 상위키워드 분석

자료는 고용노동부의 직업훈련 통합 전산망인 HRD-NET(<http://hrd.go.kr>)의 고객센터 질문게시판을 이용하여 크롤링 한 분석대상 키워드의 개수는 아래와 같다. 텍스트 마이닝 기법 중에 하나인 워드클라우드(Word Cloud)는 데이터에 대한 시각화 기법으로 사용되고 있으며,(조남희 외, 2017)[19]. 아래는 연도별 상위 빈도 키워드를 wordcloud 패키지를 이용한 분석 결과이다.

Table 1. Final analysis data

year	Word count
2012	386
2013	2,616
2014	2,638
2015	3,993
2016	4,983
2017	8,516



Fig. 1. Top keywords visualization

HRD-NET의 모든 질문내용을 가지고 종합 분석한 결과 신청(735), 오류(610), 동영상(609), 훈련(552), 카드(435), 인증서(427)등 교육훈련에 등록 및 동영상 시청 등 절차적인 부분의 애로사항이 드러나는 것으로 나타났으며, 정보보안 등 공인인증서 처리, 웹 상의 가입 및 로그인 등에 대한 시스템적 질의내용이 많은 것으로 나타나 교육훈련 수강 및 등록 처리 등의 일련의 과정에 훈련수요자들의 절차적 복잡성을 느끼는 것으로 추정된다.

4.2 연관분석(Associative Analysis)

연관분석이란 자료에 존재하는 항목간의 연관규칙(association rule)을 발견하는 분석으로, 일반적으로 상

품을 구매하거나 서비스를 받는 등의 거래나 사건들의 연관성에 대한 규칙을 이용하여, 마케팅에서 손님의 장바구니에 들어있는 품목간의 관계를 알아본다는 의미에서 장바구니분석 (market basket analysis)이라고도 한다.

연관분석에 있어 사용되는 분석지표로 지지도(SUPPORT), 신뢰도(CONFIDENCE), 향상도(LIFT)가 있으며 지지도는 예를 들어 전체 사건(Event)에서 특정 이벤트 A와 B가 동시에 일어나는 비중으로 해당 규칙이 얼마나 의미가 있는 규칙인지를 나타낸다. 따라서 지지도 = $P(A \cap B)$ 즉, A와 B가 동시에 일어난 횟수 / 전체 사건 횟수로 표시 할 수 있다. 신뢰도는 A를 포함하는 사건 중 A와 B가 동시에 발생하는 비중으로, A라는 사건이 발생했을 때 B가 발생할 확률이 얼마나 높은지를 말해준다. 따라서 신뢰도 = $P(A \cap B) / P(A)$ 즉, A와 B가 동시에 일어난 횟수 / A가 일어난 횟수로 표시된다. 향상도는 A와 B가 동시에 거래된 비중을 A와 B가 서로 독립된 사건일 때 동시에 거래된 비중으로 나눈 값이다. 즉, A와 B가 우연에 의해서 같이 거래된 확률보다 A와 B 사이의 관계가 얼마나 더 끈끈한지를 보는 지표이다. 따라서 향상도 = $P(A \cap B) / P(A) * P(B) = P(B|A) / P(B)$ 즉, A와 B가 동시에 일어난 횟수 / A, B가 독립된 사건일 때 A, B가 동시에 일어날 확률로 표시가능하며, 사건 A와 B 사이에 아무런 관계가 상호 관계가 없다면 향상도는 1이 된다. 향상도가 1보다 높아질 수록 이 규칙은 우연히 일어나지 않았다는 강한 표시가 될 수 있다.

Table 2. Major Keyword support, confidence, lift

lhs	rhs	support	confidence	lift
seeing and hearing	video	0.061	0.962	9.291
enrollment	authorized	0.031	0.440	6.809
authorized	error	0.013	0.205	1.977
card	error	0.016	0.220	2.126
training	video	0.010	0.115	1.11
proposal	error	0.029	0.232	2.241
charge, own expense	refunds	0.011	0.833	34.010
authorized, certification	login	0.011	0.183	4.204
Employee, card	error	0.011	0.268	2.582
own expense, refunds	charge	0.011	0.783	47.448
certification	real name	0.012	0.961	41.529
attended	management	0.017	0.694	30.009
served	Tomorrow learning card	0.011	0.411	7.260
card	issued	0.011	0.158	5.649
secession	request	0.012	0.420	16.045
video	confirm	0.011	0.106	3.184
training	agency	0.029	0.313	9.445

[8] C. N. Jun & I. W. Su. (2013). A Study on the Application of Technology Marketing for Big Data Analysis. □ Marketing Bulletin□ 21(2), 181-203.

[9] I. H. Yu. (2012). Web Korea Annual, 2012. Seoul : Impress Media

[10] M. G. Song & S. B. Kim(2013). A Study of improving reliability on prediction model by analyzing method Big data, Journal of Digital Convergence, 11(6), 103-112

[11] Douglas. L.(2012). The importance of 'big data', Connecticut: Gartner

[12] S. A. Kim, J. H. Park, H. J. Lee & Y. J. Jung. (2016). Research on multicultural art education using text mining techniques. Multicultural Education Research, 9(2), 203-227.

[13] C. N. Jun. (2016). View the world with language materials. Seoul : National Korean Language Institute

[14] C. S. Park & C. W. Jung (2013).Text network analysis: Policy through social recognition network analysis Stakeholder-shared meanings. Government research . 29(2), 73-108.

[15] Wikipedia (2016). Semantic analysis. https://ko.wikipedia.org/wiki/%EC%9D%98%EB%AF%B8_%EB%B6%84%EC%84%9D

[16] Y. S. Hwang. (2013). Introducing Natural Language Processing and Opinion Mining Platform. <https://readme.skplanet.com/?p=3749>

[17] H. J. Kim, J. H. Lee & S. S. Sin(2017). Multi-threaded Web Crawling Design using Queues, Journal of Convergence for Information Technology, 7(2), 43-51

[18] J. H. Lee, J. S. Lee & J. W. Son (2016). R programming based unstructured construction data analysis. Journal of the Architectural Institute of Korea. 32(5), 37-44.

[19] N. H. Jo, E. Y. Nam(2017). Analysis of Domestic Research on Depression and Stress : Focused on the Treatment and Subjects. Journal of Convergence for Information Technology. 7(6), 53-59.

성 보 경(Sung, Bo, Kyoung) [정회원]



- 2002년 2월 : 방송통신대학교 법학과(법학사)
- 2004년 8월 : 세종대학교 정보통신학과(공학석사)
- 2017년 3월 ~ 현재 : 한성대학교 스마트융합 컨설팅 학과 박사과정

3학기 재학 중 · 관심분야 : 경영컨설팅, 4차산업혁명

▪ E-Mail : career_hrd@naver.com

유 연 우(You, Yen Yoo) [정회원]



- 1996년 2월 : 숭실대학교 정보과 학대학원 산업경영(석사)
- 2007년 2월 : 한성대학교 일반대학원 행정학과(행정학 박사)
- 1981년 7월 ~ 2002년 1월 : 해외 건설협회(기획, 전산, 해외금융, 전략/IT컨설팅)

- 2002년 2월 ~ 2008년 4월 : 중소기업 기술정보진흥원 (컨설팅, 경영혁신, CSR, IT, 서비스R&D, 기술혁신)
- 2008년 9월 ~ 2018년 3월 한성대학교 지식서비스&컨설팅학과교수
- 2018년 3월 ~ 현재 : 한성대학교 스마트경영공학부 컨설팅트랙 교수
- 2010년 1월 ~ 현재 : 서울산업통산진흥원 BS산업육성위원회 위원
- 2011년 1월 ~ 현재 : 소상공인시장진흥공단 평가 운영 위원
- 2016년 7월 ~ 현재 : (재)장애인기업종합지원센터 평가위원
- 2011년11월 ~ 현재 : 제주관광공사 성과평가 위원
- 2015년 1월 ~ 현재 : 중소기업기술정보진흥원 평가위원
- 2016년 1월 ~ 현재 : 한국산업기술평가관리원 평가위원
- 2018년 6월 ~ 현재 : 정보통신산업진흥원(NIPA) 평가위원
- 관심분야 : Consulting(Stratgy, PM, 성과평가, MOT), CSR, Technology Innovation, Management Innovation, Service R&D, Franchise, 1인창조기업, 지식재산, 장애인기업지원