J. of Practical Agriculture & Fisheries Research
2020 June. : 22(1): 113-129
ISSN 2234-1129

한국농수산대학 신입생 자기소개서의 텍스트 마이닝과 연관규칙 분석 (1)

Text Mining and Association Rules Analysis to a Self-Introduction Letter of Freshman at Korea National College of Agricultural and Fisheries (1)

주진수
J. S. Joo
국립한국농수산대학
농어업·농어촌연구소[1]
nongsusan@af.ac.kr

이소영
S. Y. Lee
국립한국농수산대학
농수산비지니스학과[2]
lsy2000@korea.kr

김종숙
J. S. Kim
국립한국농수산대학
농수산비지니스학과[2]
jskimsy@korea.kr

신용광
Y. K. Shin
국립한국농수산대학
농수산비지니스학과[2]
ykshin22@korea.kr

박노복*
N. B. Park*
국립한국농수산대학
화훼학과[3]
noubogpark@naver.com

Abstract

In this study we examined the topic analysis and correlation analysis by text mining to extract meaningful information or rules from the self introduction letter of freshman at Korea National College of Agriculture and Fisheries in 2020. The analysis items are described in items related to 'academic' and 'in-school activities' during high school. In the text mining results, the keywords of 'academic' items were 'study', 'thought', 'effort', 'problem', 'friend', and the key words of 'in-school activities' were 'activity', 'thought', 'friend', 'club', 'school' in order.

As a result of the correlation analysis, the key words of 'thinking', 'studying', 'effort', and 'time' played a central role in the 'academic' item. And the key words of 'in-school activities' were 'thought', 'activity', 'school', 'time', and 'friend'.

The results of frequency analysis and association analysis were visualized with word cloud and correlation graphs to make it easier to understand all the results. In the next study, TF-IDF(Term Frequency-Inverse Document Frequency) analysis using 'frequency of keywords' and 'reverse of document frequency' will be performed as a method of extracting key words from a large amount of documents.

Key words : Association rules analysis, Betweenness centrality, Degree centrality, Word cloud

*교신저자
1 Korea National College of Agriculture and Fisheries, 1515, Kongwipatjwi-ro, Deokjin-gu, Jeollabuk-do, 54874, Korea
2 Department of Agriculture and Fisheries Business, Korea National College of Agriculture and Fisheries
3 Department of Floriculture, Korea National College of Agriculture and Fisheries

Ⅰ. 서론

대학을 지원하는 학생들은 입학원서와 함께 자기소개서(이하 자소서)를 작성하여 제출해야 한다. 자소서는 지원자 본인이 작성해야 하고 사실에 근거하여 정직하게 지원자 자신의 능력이나 특성, 경험을 기술하여야 한다. 자소서는 대입을 준비하는 수험생이 자신을 소개하기 위해 작성하여 제출하는 서식으로 제한된 공간에서 자신의 장점과 자신의 성격 등을 잘 알려야 한다.

한국농수산대학(이하 한농대) 자소서는 기본적으로 한국대학교육협의회가 지정한 세 문항을 공통으로 하고 있다. 자소서 1번은 학업에 기울인 노력과 학습 경험을 통해, 배우고 느낀 점, 2번은 본인이 의미를 두고 노력했던 교내 활동을 통해 배우고 느낀 점, 3번은 배려·나눔·협력·갈등관리 등을 실천한 사례와 그 과정에서 배운 점, 그리고 4번은 지원동기와 학업계획을 중심으로 향후 진로계획(영농, 영어계획)을 기술하도록 하고 있다.

본 연구에서는 비정형의 텍스트 자료인 2020년 한농대 신입생 자소서의 특성을 파악하여 대학 입시 평가에 활용하기 위하여 1번과 2번 문항을 대상으로 두 가지 분석 방법을 활용하였으며 나머지 문항은 다음 연구에서 검토하도록 하였다.

분석 방법은 비정형 데이터 처리 방법인 텍스트 마이닝에 의한 토픽 분석과 워드 클라우드에 의한 시각화, 그리고 연관 분석을 통한 단어와 단어 사이의 연관성 분석과 연관어 네트워크에 의한 시각화를 통하여 규칙과 패턴을 추출하였다.

Ⅱ. 연구내용

1. 분석 도구 및 기법

본 연구에서 활용한 분석 프로그램은 R 프로그래밍 언어에 기반한 RStudio(버전 3.6)다. R 프로그래밍 언어는 오픈소스 프로젝트로 통계 계산 및 시각화를 위한 언어 및 개발 환경을 제공한다. 또한 이를 통해 기본적인 통계 기법부터 최신 데이터 마이닝 기법까지 구현이 가능하며 통계적 컴퓨팅 언어로 다양한 통계 분석에 용이하다. 현재 R 프로그래밍 언어는 다양한 빅 데이터 분석 및 예측 분석 등을 포함한 고급 분석 기술들의 연구 및 개발에 많이 활용되고 있다.

본 연구에서는 빅 데이터 분석 기법의 하나인 텍스트 기반의 데이터로부터 정보 검색, 추출, 체계화, 분석을 포함하는 Text-processing 기술 및 처리 과정인 텍스트 마이닝(Text Mining) 기법을 활용하여 자소서의 문항별 주요 단어의 추출, 단어의 연관 분석 및 시각화를 하였다.

분석 자료는 2020년 한농대 신입생 550명의 자소서이며, 컴퓨터에 입력하기 위하여 자료를 담당 부서로부터 엑셀 파일 형식으로 받은 후 띄어쓰기, 오자 수정 등 몇 가지 문법적 처리를 한 후 UTF-8 형식의 텍스트 파일로 변환하였다.

2. 텍스트 마이닝

텍스트 마이닝은 다양한 형식의 데이터를 이용한 데이터 마이닝으로 구조화되지 않은 비정형 텍스트를 사용하여 패턴이나 관계를 추출하고 그 안에서 의미 있는 정보나 가치를 발굴하여 해석하거나 의사결정을 지원하는 일련의 과정을 통칭한다. 텍스트 마이닝은 문서 요약, 문서 분류, 문서 군집, 특성 추출로 크게 4가지의 기능이 있으며 텍스트 분석을 위해서는 해당 언어, 문화 및 관습에 대한 깊은 이해도가 필요하다.

텍스트 마이닝에 활용한 RStudio에서 한글 자연어 처리를 위해서는 문서 안에 있는 단어들의 품사를 정확하게 알기 위한 패키지 KoNLP를 설치해야 한다. 또한 단어들 검사에는 useNIADic() 명령어로 NIA 한글 사전을 사용하였으며, 사전에

Table 1. Number of freshmen by the admission process (단위 : 명)

전체	도시인재 전형	농수산인재 전형	일반 전형
550	82	110	358

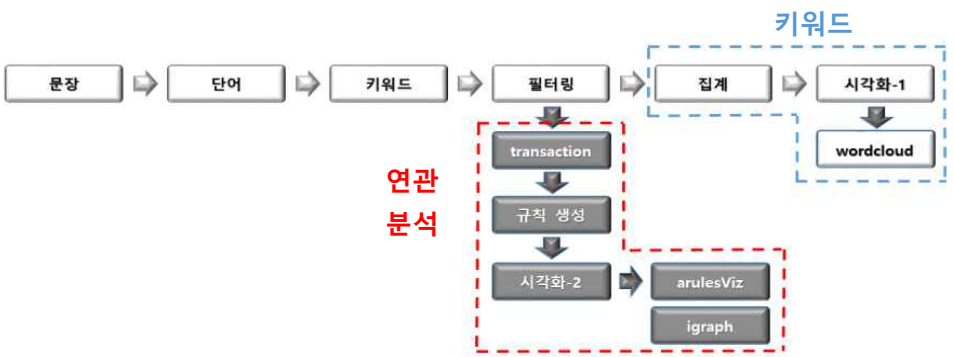

Fig. 1. Conceptual maps for keyword analysis and associative rules analysis

없는 단어는 개인적으로 4,993개를 사전에 추가하였다.

자소서에서 추출한 주요 단어는 word cloud 패키지를 사용하여 시각화하였다(Fig. 1). word cloud에 의한 도표화는 한눈에 텍스트의 맥락을 이해할 수 있게 하는 장점이 있다.

3. 연관 분석

마케팅과 웹 마이닝에서 많이 사용되는 연관 분석은 장바구니 분석으로 잘 알려진 분석으로서 대용량 데이터베이스에서 변수들 사이에 흥미로운 관계를 탐색하기 위해 고안된 자율학습법의 하나이다.

연관 분석에서는 먼저 각각의 신입생이 작성한 글이 여러 개의 문장으로 구성된 경우 하나의 문장으로 편집하여 각 문항을 전체 550개의 문장으로 수정하여 프로그램에 입력하였다. 분석 그룹은 2개의 문항(문항 1, 문항 2), 4개의 전형(도시 인재 전형, 농수산 인재 전형, 일반 전형, 전체)으로 구분하였다.

각각의 자소서를 하나의 문장으로 편집하는 이유는 연관 분석은 보통 하나의 글마다 나오는 단어를 분석한 후 어떠한 단어들이 연계되어 자주 나오는가를 분석하기 때문에 키워드 분석에 사용한 데이터를 사용하게 되면 한 줄마다 단어의 연관을 분석하여 결과로 아무것도 나오지 않을 가능성이 있기 때문이다. 그러나 앞에서 설명한 텍스트 마이닝에서는 이러한 절차가 필요하지 않다.

연관 분석을 위해서는 arules Package의 apriori() 함수를 사용했다. 또한 연관 분석의 최종적인 목표는 단어 간의 연관성 및 키워드를 그래프로 보여주는 시각화이기 때문에 그래프를 그리기 위해 igraph Package와 arulesViz Package를 설치하였다. apriori() 함수로 얻어진 연관 규칙의 흥미도 측정은 크게 support(지지도)와

Table 2. The evaluation scales of association rules analysis

척 도	수 식	설 명
support (지지도)	$s(X\rightarrow Y)=\frac{n(X\cup Y)}{N}$	전체 자소서 중에서 단어 *X, Y*가 동시에 포함되는 자소서의 비율
confidence (신뢰도)	$c(X\rightarrow Y)=\frac{n(X\cup Y)}{n(X)}$	단어 *X*를 포함하는 자소서 중에서 단어 *X, Y*를 모두 포함하는 자소서일 확률
Lift (향상도)	$Lift(X\rightarrow Y)=\frac{c(X\rightarrow Y)}{s(Y)}$	단어 *Y*를 사용한 자소서 대비 단어 *X*를 사용한 자소서에 대한 확률

$s(Y)=n(Y)/N$ (N : 전체 transaction 개수)

confidence(신뢰도) 그리고 모델의 성능 평가를 위한 lift(향상도)를 고려하였다(Table 2).

지지도는 좋은 규칙(빈도가 높거나 구성비가 높은)을 찾거나 불필요한 연산을 줄일 때 기준으로 사용하며 전체 단어 사용 경향을 파악할 수 있다. 신뢰도는 그 정도가 높을수록 유용한 규칙일 가능성이 크며 일관성의 정도를 파악할 수 있으며, 향상도는 상관관계를 파악할 수 있다.

이 함수에 의해 단순한 테이블로 정리되는 결과는 한눈에 규칙을 파악하기 어려운 단점이 있음으로 네트워크 그래프로 시각화하여 각 단어 간의 관계 파악과 그 핵심 단어를 한눈에 알 수 있게 하였다. 연관어 네트워크의 시각화에는 plot.igraph() 함수 및 degree() 함수를 이용하였다(Fig. 1).

연관 키워드의 네트워크 지도는 Node(노드), edge(선)로 구성되는데, 각각의 키워드를 노드로 정의하고 네트워크 그래프로 나타난 노드의 중요도를 파악하였다. 네트워크 지도는 ① degree() 함수를 이용하여 노드에 연결된 edge의 수에 따라 중요도를 파악하는 '연결 중심성(degree centrality)'과 ② 전체 네트워크에서 해당 노드와 다른 노드들 사이 최단 경로를 얼마나 많이 가졌는지 측정하는 방법, 즉 노드 간의 매개체 역할을 하는 노드를 알려주는 '관계 중심성(betweenness centrality)'으로 표현하였다.

Ⅲ. 결과 및 고찰

1. 텍스트 마이닝 및 시각화

가. 문항 1

신입생들이 고교재학 기간 중 학업 노력과 학습 경험을 기술한 문항 1에 대한 키워드 분석 결과를 Table 3과 Table 4에 나타내었다. 문항 1에서 추출한 명사는 약 61,000개였으며 전처리 작업을 통하여 키워드를 정리하였다.

먼저 Table 3은 키워드의 빈도 분석 결과로서 평균 빈도는 각 전형별로 나타난 각각의 키워드 빈도를 모집 전형별 인원으로 나누어 표준화한 값을 의미한다. 모든 전형에서 1순위로 나타난 '공부'는 도시 인재 전형에서 가장 높게 나타났으며, 전체적으로 신입생들은 평균 2.7회 정도 사용한 것으로 나타났다. 이러한 결과는 자소서 문항 1에서 요구하는 '학업'과 '학습'의 경험· 활동을 서술하라는 내용과 연관성이 높은 것으로 여겨진다. 전체 신입생의 10위까지의 키워드(붉은색 표기)를 기준으로 하여 전형별로 분포를 비교해 보면 도시 인재 전형과 농수산 인재 전형에서 키워드 '농업', '선생님', '수업', '노력' 등의 순위와 빈도가 차이를 보이는 것을 알 수 있다. 특히 7위의 '농업'은 도시 인재 전형에서 평균 빈도 0.7회이나 농수산 인재 전형에서 1.4회로 나타나 모

Table 3. Key words of the question 1 in self-introduction letter by the admissions process

(단위 : 회/인)

순위	전체		도시 인재 전형		농수산 인재 전형		일반 전형	
	키워드	평균 빈도	키워드	평균 빈도	키워드	평균 빈도	키워드	평균 빈도
1	공부	2.7	공부	2.8	공부	2.5	공부	2.7
2	생각	1.9	생각	2.0	생각	2.0	생각	1.9
3	노력	1.1	문제	1.2	농업	1.4	노력	1.3
4	문제	1.1	친구	1.1	시간	1.1	문제	1.1
5	시간	1.1	시간	1.0	문제	1.0	시간	1.1
6	친구	1.0	관심	1.0	학교	1.0	친구	1.1
7	농업	0.9	노력	1.0	친구	0.9	농업	0.8
8	이해	0.8	이해	1.0	내용	0.9	수업	0.8
9	선생님	0.8	수학	0.9	이해	0.9	선생님	0.8
10	수업	0.8	과목	0.9	방법	0.9	이해	0.8
11	관심	0.7	방법	0.8	선생님	0.9	성적	0.8
12	내용	0.7	내용	0.8	지식	0.8	관심	0.7
13	방법	0.7	과정	0.7	노력	0.8	결과	0.7
14	결과	0.7	농업	0.7	흥미	0.8	과정	0.7
15	학교	0.7	결과	0.7	작물	0.7	고등학교	0.6
16	과정	0.7	선생님	0.6	결과	0.7	내용	0.6
17	성적	0.7	흥미	0.6	수업	0.7	학교	0.6
18	과목	0.6	수업	0.6	다양	0.6	활동	0.6
19	고등학교	0.6	경험	0.5	과정	0.6	방법	0.6
20	활동	0.6	관련	0.5	실습	0.6	시작	0.6

Table 4. Word clouds of the question 1 in self-introduction letter by the admissions process

전체	도시 인재 전형
	
농수산 인재 전형	일반 전형
	

집 전형별 특징을 반영한 결과로 판단된다.

Table 4는 각 전형별로 추출한 키워드 가운데 상위 빈도 50위까지의 키워드를 가시화한 워드 클라우드이다. 워드 클라우드를 이용한 시각화는 Table 3에 나타난 바와 같이 빈도가 높은 '공부', '생각', '노력', '문제', '시간' 등의 키워드가 크게 보이고 있어 더욱더 쉽게 문맥을 알아볼 수 있다.

나. 문항 2

문항 2는 신입생들이 고교 재학 기간 중 본인이 의미를 두고 노력했던 교내 활동에서 배우고 느낀 점을 중심으로 3개 이내를 기술하도록 하고 있다. 문항 2에 대한 키워드 분석 결과를 Table 5와 Table 6에 나타내었다. 문항 2에서 추출한 명사는 약 88,000개로 문항 1보다 많은 단어가 추출되었으며 전처리 작업을 통하여 키워드를 정리하였다.

Table 5는 키워드의 빈도 분석 결과로서 '활동'은 모든 전형에서 1순위로 나타났으며, 전체 신입생으로 계산해 보면 모든 학생은 평균 3.0회 정도 사용한 것으로 나타났다. 가장 높은 빈도를 나타낸 키워드 '활동'은 자소서 문항에서 요구하는 '교내 활동'에 대한 서술어로 기술된 단어로 여겨진다.

전체 신입생의 10위까지의 키워드(붉은색 표기)를 기준으로 하여 전형별로 분포를 비교해 보면 도시 인재 전형과 농수산 인재 전형에서 키워드 '농업', '관심', '노력' 등의 순위와 빈도에 차이를 있는 것을 알 수 있다. 특히 도시 인재 전형에서 '활동', '친구', '생각', '동아리' 등의 키워드 평균 빈도수는 농수산 인재나 일반 전형 신입생보다 매우 높게 나타났으며, 이는 '친구'들과 '동아리 활동' 등 다양한 활동을 한 내용을 기술한 결과로 판단된다.

Table 5. Key words of the question 2 in self-introduction letter by the admissions process

(단위 : 회/인)

순위	전체		도시 인재 전형		농수산 인재 전형		일반 전형	
	키워드	평균 빈도	키워드	평균 빈도	키워드	평균 빈도	키워드	평균 빈도
1	활동	3.0	활동	3.9	활동	2.8	활동	2.9
2	생각	2.8	친구	3.3	생각	2.6	생각	2.8
3	친구	2.1	생각	3.2	동아리	1.7	친구	2.0
4	동아리	1.6	동아리	2.2	학교	1.6	학교	1.6
5	학교	1.5	시간	1.5	친구	1.6	동아리	1.4
6	시간	1.4	관심	1.3	시간	1.3	시간	1.4
7	사람	1.1	사람	1.3	경험	1.1	사람	1.1
8	관심	0.9	학교	1.1	사람	1.0	농업	1.0
9	농업	0.9	선생님	1.0	학생	1.0	노력	1.0
10	노력	0.9	경험	0.8	노력	0.9	관심	0.9
11	선생님	0.9	과정	0.8	작물	0.9	선생님	0.9
12	경험	0.9	농업	0.8	선생님	0.9	경험	0.8
13	과정	0.8	식물	0.8	과정	0.9	과정	0.8
14	학생	0.8	학생	0.8	농업	0.8	문제	0.8
15	문제	0.8	의견	0.8	재배	0.8	학생	0.7
16	다양	0.7	노력	0.8	관심	0.8	참여	0.7
17	방법	0.7	부원	0.8	방법	0.8	다양	0.7
18	참여	0.7	다양	0.7	문제	0.8	방법	0.7
19	식물	0.6	문제	0.7	발표	0.8	식물	0.7
20	결과	0.6	조사	0.7	결과	0.7	결과	0.6

Table 6은 문항 1과 같이 각 전형별로 추출한 키워드 가운데 상위 빈도 50위까지의 키워드를 가시화한 워드 클라우드이다. Table 5에 나타난 바와 같이 빈도수가 높은 '활동', '생각', '친구', '동아리', '시간', '학교' 등의 키워드가 눈에 띄게 나타나고 있다.

Table 6. Word clouds of the question 2 in self-introduction letter by the admissions process

전체	도시 인재 전형
	
농수산 인재 전형	일반 전형
	

문항 1, 2의 텍스트 마이닝 결과(Table 3, Table 5)에서 '생각', '노력', '시간', '친구', '농업' 등의 키워드가 상위 10위 안에 공통으로 나타났다. 이들 5개 키워드와 함께 문항 1에서는 키워드 '공부', '문제', '이해', '선생님', '수업' 등과 같은 학업에 기울인 노력과 학습 경험을 기술하는 단어들이 나타났으며, 문항 2에서는 '활동', '동아리', '학교', '사람', '관심' 등 의미를 두고 노력한 교내활동을 기술하는 단어들이 상위 빈도 10위 안에 나타나는 결과를 보였다.

2. 연관 분석 및 시각화

가. 문항 1

Table 7은 apriori() 함수를 사용하여 얻어진 문항 1의 연관 규칙 분석 결과이다. 분석은 불필요한 연산을 줄이고 좋은 규칙을 찾기 위하여 기준값으로 support(지지도)는 0.3, confidence(신뢰도)는 0.5로 설정하여 46개의 규칙을 추출하였다. itemMatrix 사이즈는 550×7,774의 메트릭스로 분석되었다.

Table 3에서 평균 빈도가 가장 높게 나타난 키워드 '공부'와 연관한 규칙을 예로 살펴보았다. 먼저 {성적} => {공부} 규칙에서 지지도는 0.309로

Table 7. Association rules for the question 1 in self-introduction letter

번호	규칙 {lhs} => {rhs}	지지도 (support)	신뢰도 (confidence)	향상도 (lift)	빈도 (frequency)
1	{} => {노력}	0.563636	0.563636	1	310
2	{} => {시간}	0.534545	0.534545	1	294
3	{} => {공부}	0.710909	0.710909	1	391
4	{} => {생각}	0.798182	0.798182	1	439
5	{성적} => {공부}	0.309091	0.885417	1.245471	170
6	{과목} => {공부}	0.3	0.854922	1.202576	165
7	{시작} => {공부}	0.307273	0.786047	1.105692	169
8	{시작} => {생각}	0.32	0.818605	1.025587	176
9	{경험} => {생각}	0.3	0.778302	0.975093	165
10	{이해} => {공부}	0.314545	0.797235	1.12143	173
11	{이해} => {생각}	0.301818	0.764977	0.958399	166
12	{고등학교} => {공부}	0.34	0.806034	1.133808	187
13	{고등학교} => {생각}	0.347273	0.823276	1.031439	191
14	{방법} => {생각}	0.316364	0.78733	0.986405	174
15	{과정} => {생각}	0.323636	0.809091	1.013667	178
16	{학교} => {공부}	0.309091	0.73913	1.039698	170
17	{학교} => {생각}	0.345455	0.826087	1.034961	190
18	{결과} => {공부}	0.321818	0.753191	1.059477	177
19	{결과} => {생각}	0.336364	0.787234	0.986284	185
20	{문제} => {공부}	0.314545	0.748918	1.053465	173
21	{문제} => {생각}	0.347273	0.82684	1.035904	191
22	{선생님} => {노력}	0.3	0.684647	1.214697	165
23	{노력} => {선생님}	0.3	0.532258	1.214697	165
24	{선생님} => {공부}	0.336364	0.767635	1.079793	185
25	{선생님} => {생각}	0.347273	0.792531	0.992921	191
26	{친구} => {공부}	0.345455	0.766129	1.077675	190
27	{친구} => {생각}	0.372727	0.826613	1.03562	205
28	{관심} => {생각}	0.341818	0.789916	0.989644	188
29	{노력} => {시간}	0.309091	0.548387	1.025894	170
30	{시간} => {노력}	0.309091	0.578231	1.025894	170
31	{노력} => {공부}	0.425455	0.754839	1.061794	234
32	{공부} => {노력}	0.425455	0.598465	1.061794	234
33	{노력} => {생각}	0.44	0.780645	0.978029	242
34	{생각} => {노력}	0.44	0.551253	0.978029	242
35	{시간} => {공부}	0.383636	0.717687	1.009534	211
36	{공부} => {시간}	0.383636	0.539642	1.009534	211
37	{시간} => {생각}	0.430909	0.806122	1.009948	237
38	{생각} => {시간}	0.430909	0.539863	1.009948	237
39	{공부} => {생각}	0.570909	0.803069	1.006123	314
40	{생각} => {공부}	0.570909	0.715262	1.006123	314
41	{공부, 노력} => {생각}	0.338182	0.794872	0.995853	186
42	{노력, 생각} => {공부}	0.338182	0.768595	1.081144	186
43	{공부, 생각} => {노력}	0.338182	0.592357	1.050955	186
44	{공부, 시간} => {생각}	0.312727	0.815166	1.021278	172
45	{생각, 시간} => {공부}	0.312727	0.725738	1.02086	172
46	{공부, 생각} => {시간}	0.312727	0.547771	1.024741	172

'성적' 단어를 사용할 때 '공부' 단어를 동시에 사용하는 비율은 높지 않으나 신뢰도가 0.854로 매우 높게 나타나 '성적' 단어를 사용하는 경우에 '공부' 단어를 사용하는 비율이 높음을 나타냈다. 또한, 향상도는 1보다 큰 단어 사이에는 연관성이 매우 높음을 의미하는데 {성적} => {공부} 규칙의 향상도가 1.245로 최댓값을 나타내 '성적' 단어를 사용하는 경우는 그렇지 않은 경우에 비하여 '공부' 단어를 함께 사용하는 비율이 높음을 나타냈다. {과목} => {공부} 규칙도 이와 비슷한 경향을 나타냈다. 또한 {공부} => {생각}의 규칙은 지지도(0.57)와 신뢰도(0.8)가 높게 나타났으며, 향상도 역시 1보다 크게 나타나 학생들이 이들 두 단어를 동시에 사용하는 비율이 높으며 단어 간 연관성도 높은 결과를 나타냈다.

향상도가 1로 나타난 1번~4번까지의 규칙은 서로 연관성이 없는 독립적 관계를 의미하며, 1보다 작은 {생각}과 맺어진 8개의 규칙은 음의 상관관계로서 연관성이 없음을 의미한다.

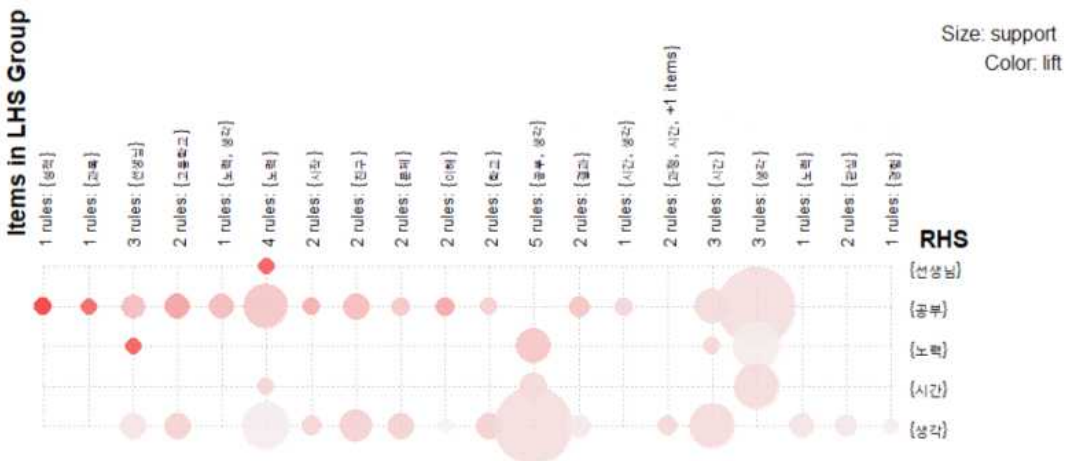

Fig. 2. Grouped matrix for the question 1 by association rules

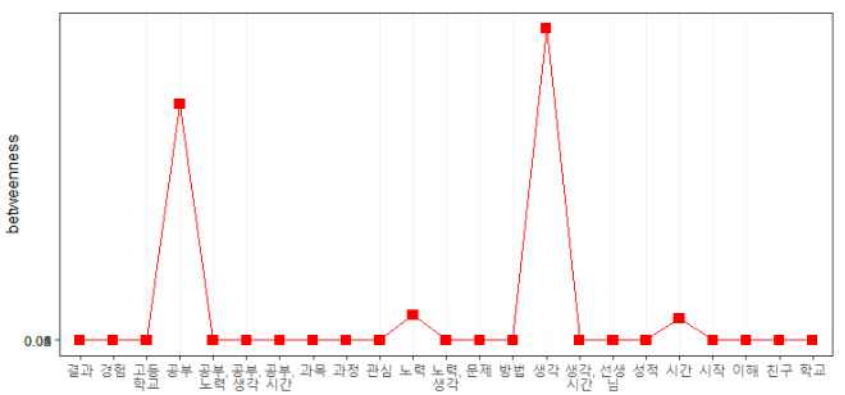

Fig. 3. Association graph for the question 1 by betweenness centrality

연관 분석에서 키워드 간의 연관성 및 키워드의 시각화에는 igraph Package와 arulesViz Package를 이용하였다. Fig. 2는 Table 7에서 독립관계인 1번~4번을 제외하고 {LHS} => {RHS} 키워드 간의 연관 규칙을 지지도와 향상도의 관계로 나타낸 결과이다. 지지도가 높으면 원형이 크게 나타나며, 향상도가 크면 색상이 붉은색으로 진하게 나타난다. 우측의 수직축에 나타낸 {RHS} 키워드를 중심으로 보면 {공부}와 {생각} 키워드 축에 {LHS} 키워드들이 많이 연관되어 있음을 쉽게 알 수 있다.

Fig. 3은 키워드 간의 매개체 역할을 하는 키워드를 알려주는 '관계 중심성' 분석 결과로서 '생각', '공부', '노력' 및 '시간'이 다른 키워드들 사이에 최단 경로를 많이 가지고 있어 관계 중심성이 높은 것을 알 수 있다.

Fig. 4는 노드와 edge(선)로 구성되는 네트워크 그래프로서 edge 수에 따라 노드의 라벨(Label) 크기로 표현되는 '연결 중심성' 분석 결과로서 기준값(support=0.25, confidence=0.35)을 Table 7과 다르게 설정하여 추출한 137개 규칙의 연관도이다. '생각', '공부', '노력' 및 '시간'의 라벨이 크게 나타나 키워드 간의 연결 중심에 있음을 나타내고 있다.

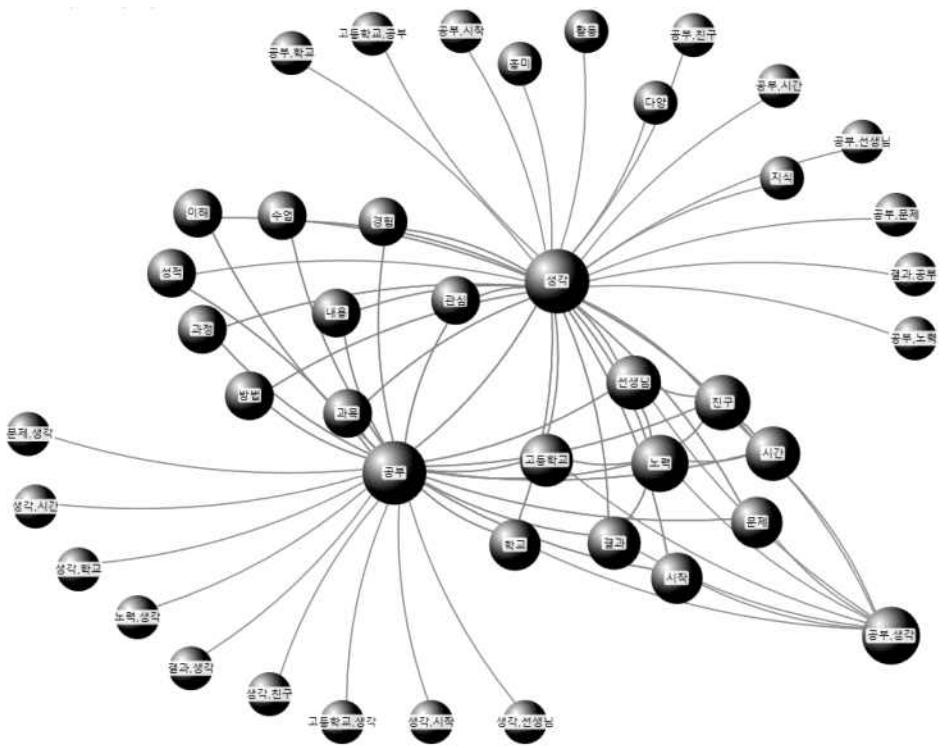

Fig. 4. Association graph for the question 1 by degree centrality

Fig. 5는 연관 분석으로 추출한 키워드의 빈도수(Table 8) 상위 50위까지를 시각화한 워드 클라우드이다. 앞에서 나타낸 Table 4(전체)의 텍스트 마이닝에 의한 워드 클라우드와 비슷한 듯 보

이나 Table 8에 비교하여 나타낸 바와 같이 키워드의 순위에 차이가 있음을 알 수 있다. 이는 한 사람이 하나의 단어를 반복하여 사용하는 경우 그 단어를 하나로 처리하는 연관 분석 기법에 따른 결과이다. 예를 들어 키워드 '공부'는 텍스트 마이닝 분석에서는 1인당 평균 2.7회를 사용했다는 의미이며, 연관 분석의 빈도를 고려하면 실제 평균 빈도는 3.79로 더욱 높아지는 것을 알 수 있다. 이러한 결과는 일부 학생은 '공부' 단어를 많이 사용하지 않았다는 것이며, 실제로 이 단어를 사용한 학생만을 고려하게 되므로 평균 빈도가 높아지는 것이다.

Fig. 5. Word cloud for the question 1 by association rules analysis

Table 8. The frequency of key words for the question 1 by text mining and association rules analysis

순위	텍스트 마이닝			순위	연관 분석		실제 평균 빈도
	키워드	빈도①*	평균 빈도*		키워드	사용자②*	= ①/②*
1	공부	1,481	2.7	1	생각	439	2.40
2	생각	1,055	1.9	2	공부	391	3.79
3	노력	616	1.1	3	노력	310	1.99
4	문제	612	1.1	4	시간	294	2.03
5	시간	598	1.1	5	친구	248	2.32
6	친구	575	1.0	6	선생님	241	1.75
7	농업	507	0.9	7	관심	238	1.71
8	이해	453	0.8	8	결과	235	1.64
9	선생님	422	0.8	9	고등학교	232	1.48
10	수업	413	0.8	10	문제	231	2.65
11	관심	407	0.7	11	학교	230	1.66
12	내용	393	0.7	12	방법	221	1.76
13	방법	388	0.7	13	과정	220	1.68
14	결과	385	0.7	14	이해	217	2.09
15	학교	381	0.7	15	시작	215	1.54
16	과정	369	0.7	16	경험	212	1.49
17	성적	365	0.7	17	수업	206	2.00
18	과목	356	0.6	18	내용	199	1.97
19	고등학교	343	0.6	19	과목	193	1.84
20	활동	336	0.6	20	성적	192	1.90

** 단위 : 빈도(회), 평균 빈도(회/명), 사용자(명), 실제 평균 빈도(회/명)

Table 9. Association rules for the question 2 in self-introduction letter

번호	규칙 {lhs} => {rhs}	지지도 (support)	신뢰도 (confidence)	향상도 (lift)	빈도 (frequency)
1	{} => {친구}	0.649091	0.649091	1	357
2	{} => {학교}	0.652727	0.652727	1	359
3	{} => {시간}	0.649091	0.649091	1	357
4	{} => {활동}	0.807273	0.807273	1	444
5	{} => {생각}	0.854545	0.854545	1	470
6	{선생님} => {생각}	0.412727	0.897233	1.049954	227
7	{동아리} => {활동}	0.427273	0.883459	1.094374	235
8	{동아리} => {생각}	0.416364	0.860902	1.007439	229
9	{사람} => {활동}	0.44	0.86121	1.066814	242
10	{사람} => {생각}	0.450909	0.882562	1.032786	248
11	{경험} => {활동}	0.414545	0.829091	1.027027	228
12	{경험} => {생각}	0.438182	0.876364	1.025532	241
13	{관심} => {활동}	0.418182	0.839416	1.039817	230
14	{관심} => {생각}	0.447273	0.89781	1.050629	246
15	{노력} => {활동}	0.416364	0.809187	1.002372	229
16	{노력} => {생각}	0.449091	0.872792	1.021352	247
17	{친구} => {학교}	0.447273	0.689076	1.055687	246
18	{학교} => {친구}	0.447273	0.685237	1.055687	246
19	{친구} => {시간}	0.447273	0.689076	1.061601	246
20	{시간} => {친구}	0.447273	0.689076	1.061601	246
21	{친구} => {활동}	0.543636	0.837535	1.037487	299
22	{활동} => {친구}	0.543636	0.673423	1.037487	299
23	{친구} => {생각}	0.567273	0.87395	1.022707	312
24	{생각} => {친구}	0.567273	0.66383	1.022707	312
25	{학교} => {시간}	0.434545	0.665738	1.025647	239
26	{시간} => {학교}	0.434545	0.669468	1.025647	239
27	{학교} => {활동}	0.530909	0.81337	1.007554	292
28	{활동} => {학교}	0.530909	0.657658	1.007554	292
29	{학교} => {생각}	0.565455	0.866295	1.01375	311
30	{생각} => {학교}	0.565455	0.661702	1.01375	311
31	{시간} => {활동}	0.532727	0.820728	1.016668	293
32	{활동} => {시간}	0.532727	0.65991	1.016668	293
33	{시간} => {생각}	0.565455	0.871148	1.019429	311
34	{생각} => {시간}	0.565455	0.661702	1.019429	311
35	{활동} => {생각}	0.694545	0.86036	1.006805	382
36	{생각} => {활동}	0.694545	0.812766	1.006805	382
37	{친구,활동} => {생각}	0.481818	0.886288	1.037145	265
38	{생각,친구} => {활동}	0.481818	0.849359	1.052134	265
39	{생각,활동} => {친구}	0.481818	0.693717	1.068752	265
40	{학교,활동} => {생각}	0.465455	0.876712	1.02594	256
41	{생각,학교} => {활동}	0.465455	0.823151	1.019669	256
42	{생각,활동} => {학교}	0.465455	0.670157	1.026703	256
43	{시간,활동} => {생각}	0.465455	0.87372	1.022438	256
44	{생각,시간} => {활동}	0.465455	0.823151	1.019669	256
45	{생각,활동} => {시간}	0.465455	0.670157	1.032455	256

나. 문항 2

Table 9는 apriori() 함수에 의한 문항 2의 연관 규칙 결과이다. 분석 기준값은 support(지지도) = 0.4, confidence(신뢰도) = 0.6으로 설정하였으며, 45개의 규칙을 만들었다. itemMatrix 사이즈는 550×10,594의 메트릭스로 분석되었다.

Table 9의 향상도는 모두 1보다 크게 나타나서 {LHS} => {RHS} 키워드 사이에는 양의 관계가 성립되어 좌측의 {LHS} 단어를 사용하는 경우가 그렇지 않은 경우에 비해 {RHS} 단어를 함께 사용하는 비율이 높은 것으로 나타났다.

또한, Table 5에서 평균 빈도가 가장 높게 나타난 키워드 '활동'과 연관한 규칙을 예로 보면, {활동} => {생각} 규칙의 지지도는 0.695로 비교적 높게 나타나 이들 두 단어를 동시에 사용하는 비율이 높게 나타났으며, 두 단어의 신뢰도는 0.86으로 '활동' 단어를 사용하는 경우에 '생각' 단어를 사용하는 비율이 매우 높아 유용한 규칙으로 판단된다. {동아리} => {활동}의 규칙은 지지도 0.427로 '동아리'를 기술할 때 '활동'을 기술하는 경우가 흔하다고 보기는 어려우나, 신뢰도가 0.883으로 '동아리'를 기술하는 경우에 '활동'을 기술하는 비율이 매우 높아 두 단어의 사용에는 일관성이 있는 결과로 나타났다. 또한 향상도는 1.094로 다른 규칙보다 높은 결과를 보이며, 학생들은 양의 상관관계인 두 단어를 동시에 기술하는 확률이 높은 것으로 나타났다.

키워드 간의 연관 규칙 및 키워드 연관성의 시각화 결과는 다음과 같다. 먼저 Fig. 6은 Table 9에서 독립관계인 1번~5번을 제외하고 {LHS} => {RHS} 키워드 간의 연관 규칙을 지지도와 향상도의 관계로 나타낸 결과이다. 지지도가 높으면 원형이 크게 나타나며, 향상도가 크면 색상이 붉은색으로 진하게 나타난다. 우측의 수직축에 나타낸 {RHS} 키워드를 중심으로 보면 {활동}과 {생각} 키워드 축에 {LHS} 키워드들이 많이 연관되어 있음을 쉽게 알 수 있다.

Fig. 9는 연결 중심성의 시각화를 보다 효과적으로 나타내기 위하여 기준값(support=0.3, confidence=0.5)을 Table 9와 다르게 설정하여 추출한 182개 규칙의 연관도이다. 이 결과를 보면 Table 10의 상위 5위까지에 나타난 키워드 '생각', '활동', '친구', '학교' 및 '시간' 등의 노드 라벨이 명확하게 크게 나타나 이들 키워드의 연결 중심성이 높은 것을 알 수 있다.

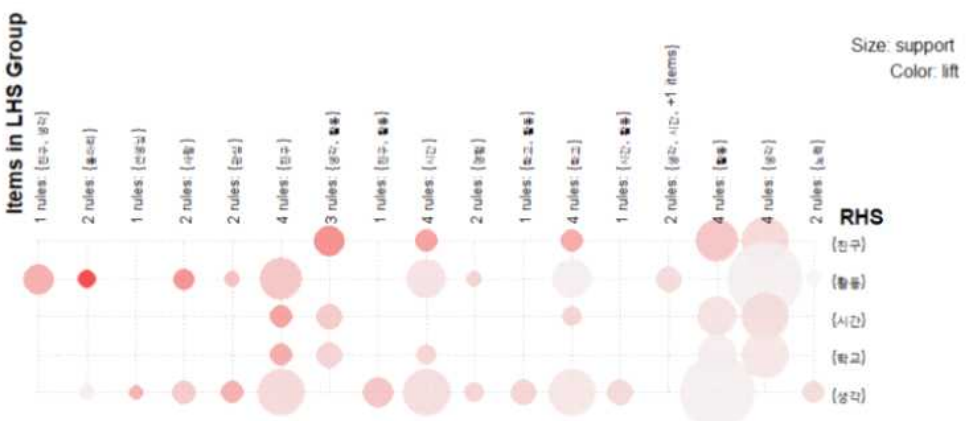

Fig. 6. Grouped matrix for the question 2 by association rules

Fig. 7은 키워드 간의 매개체 역할을 하는 키워드를 알려주는 '관계 중심성' 분석 결과로서 '생각', '활동'은 다른 키워드들 사이에 최단 경로를 많이 가지고 있어 관계 중심성이 가장 높게 나타났으며, '친구', '학교' 및 '시간' 키워드도 관계 중심성이 있는 것으로 나타났다.

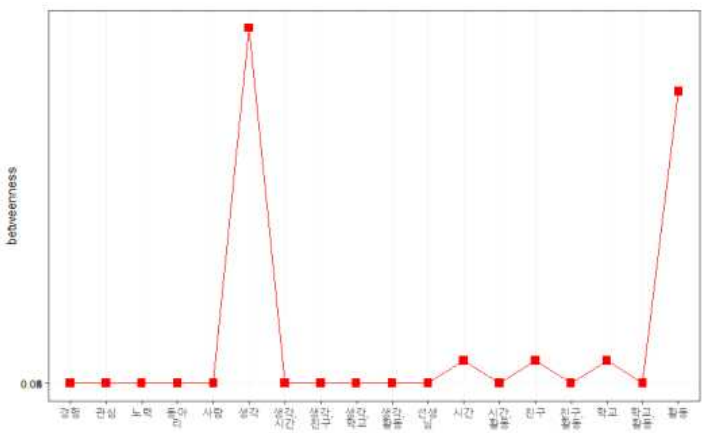

Fig. 7. Association graph for the question 2 by betweenness centrality

Fig. 8은 연관 분석으로 추출한 키워드의 빈도수(Table 8) 상위 50위까지를 시각화한 워드 클라우드로서 '생각', '활동', '학교', '시간', '친구' 등의 상위 키워드가 눈에 띄게 나타난 것을 알 수 있다. 앞에서 설명한 바와 같이 반복 단어를 제외하는 연관 분석 기법의 특성 때문에 텍스트 마이닝에 의한 키워드 빈도 순위와는 약간의 차이가 나타났다.

Fig. 8. Word cloud for the question 2 by association rules analysis

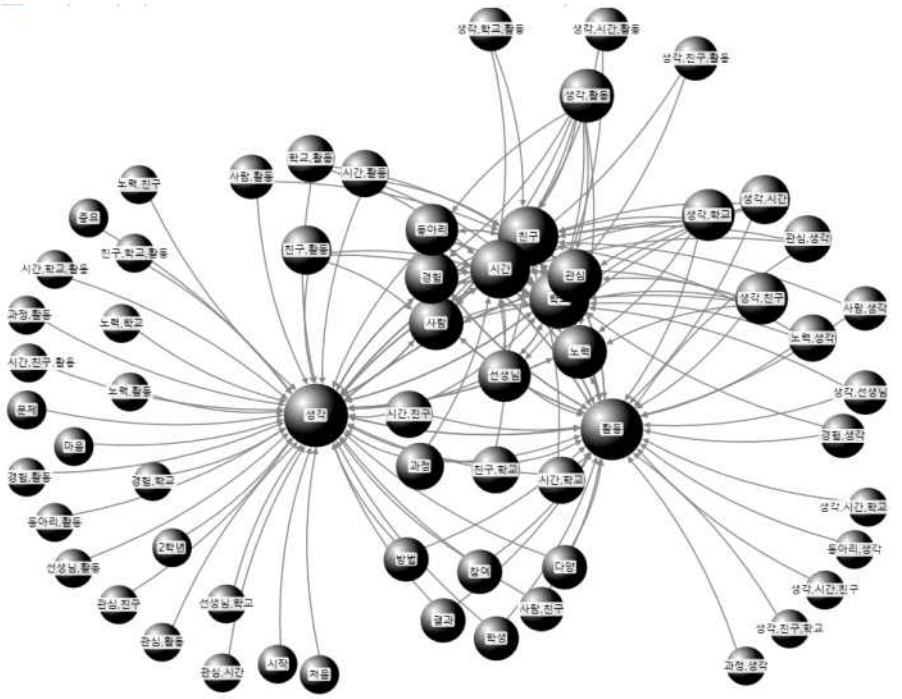

Fig. 9. Association graph for the question 2 by degree centrality

Table 10. The frequency of key words for the question 2 by text mining and association rules analysis

순위	텍스트 마이닝			순위	연관 분석		실제 평균 빈도
	키워드	빈도①*	평균 빈도*		키워드	사용자②*	= ①/②*
1	활동	1,671	3.0	1	생각	470	3.3
2	생각	1,563	2.8	2	활동	444	3.8
3	친구	1,140	2.1	3	학교	359	2.3
4	동아리	861	1.6	4	시간	357	2.1
5	학교	826	1.5	5	친구	357	3.2
6	시간	761	1.4	6	노력	283	1.8
7	사람	598	1.1	7	사람	281	2.1
8	관심	517	0.9	8	경험	275	1.8
9	농업	508	0.9	9	관심	274	1.9
10	노력	507	0.9	10	동아리	266	3.2
11	선생님	500	0.9	11	과정	253	1.8
12	경험	485	0.9	12	선생님	253	2.0
13	과정	451	0.8	13	결과	219	1.6
14	학생	438	0.8	14	다양	217	1.8
15	문제	418	0.8	15	참여	217	1.7
16	다양	384	0.7	16	방법	206	1.8
17	방법	370	0.7	17	시작	205	1.6
18	참여	370	0.7	18	학생	202	2.2
19	식물	354	0.6	19	처음	197	1.3
20	결과	346	0.6	20	중요	192	1.4

* 단위 : 빈도(회), 평균 빈도(회/명), 사용자(명), 실제 평균 빈도(회/명)

Table 10에서 키워드 '활동' 빈도수는 텍스트 마이닝 분석에서는 전체 평균 3.0회 사용하였으나 연관 분석의 빈도를 고려하면 실제 평균 빈도는 3.8로 높아지는 것을 알 수 있다. 이는 연관 분석은 실제로 해당 단어를 사용한 학생만을 고려하는 분석이므로 실제 평균 빈도가 3.8로 높아지는 결과를 보이게 된다. 한편 키워드 '동아리'는 텍스트 마이닝에서 빈도수는 4위이나 연관 분석에서 사용자는 10위로 나타나 이 단어를 사용한 학생(266명)들의 반복 사용이 많았음을 의미한다. 키워드 '동아리'는 Table 9에서 '활동'과 가장 높은 향상도를 보였으나, 다른 키워드와의 연결 중심성은 예상과 다르게 약하게 나타났다.

Ⅳ. 적요

본 연구는 2020년 한농대 입학생의 비정형 텍스트인 자소서에서 의미 있는 정보 혹은 규칙을 추출하기 위하여 고교 재학 중 '학업 및 학습 경험'과 '교내 활동'을 기술한 두 개 문항에 대하여 텍스트 마이닝에 의한 토픽 분석과 연관성 분석을 하였다.

모집 전형을 구분하지 않은 텍스트 마이닝 분석 결과에서 '학업 및 학습 경험' 항목과 관련된 주요 키워드는 '공부', '생각', '노력', '문제', '친구' 등의 순으로 많이 나타났으며, '교내 활동' 항목과 관련된 주요 키워드는 '활동', '생각', '친구', '동아리', '학교' 등의 순으로 빈도가 높게 나타났다. 그러나 도시 인재 전형과 농수산 인재 전형 신입생들의 키워드 빈도 순위는 두 항목 모두 전형 특성에 따른 약간의 차이를 나타냈다. 빈도 분석에 결과는 빈도수 상위 50위까지의 키워드를 워드 클라우드로 시각화하여 키워드를 알기 쉽게 표현하였다.

연관 분석은 apriori() 함수를 사용하였으며 적정한 계산을 위하여 support(지지도)와 confidence(신뢰도)의 기준값을 항목별로 설정하였다. 먼저 '학업' 항목에 대한 연관 규칙은 46개를 추출하였으며, 그 가운데 {공부} => {생각}, {성적} => {공부} 및 {과목} => {공부} 등의 규칙에서 높은 연관성을 볼 수 있었다. 이 규칙을 바탕으로 매개체 역할의 키워드를 평가하는 관계 중심성 평가와 노드에 연결된 edge의 수에 따라 중요도를 파악하는 연결 중심성 평가에서는 '생각', '공부', '노력', '시간' 등의 키워드가 중심적인 역할을 하는 정보를 획득하였다. 다음으로 '교내 활동' 항목에서는 45개의 연관 규칙을 생성하여 {활동} => {생각}, {동아리} => {활동} 등의 규칙에서 높은 연관성을 볼 수 있었으며, 관계 중심성 평가와 연결 중심성 평가에서는 '생각', '활동', '학교', '시간', '친구' 등의 키워드가 중심 키워드라는 결과를 얻었다.

다음 연구에서는 자소서의 나머지 두 개의 문항 '배려·나눔·협력·갈등관리' 항목과 한농대 '지원동기와 향후 진로계획' 항목을 분석한다. 분석에는 '키워드의 빈도'에 '문서 빈도의 역수'를 곱하여 주로 다량의 문서에서 핵심어를 추출하는 TF-IDF(Term Frequency-Inverse Document Frequency) 분석을 추가한다.

Ⅴ. 참고문헌

1. 김경태, 안정국, 김동현. (2018). 빅 데이터 활용서 (Ⅰ). 시대인.
2. 김영우. (2017). 쉽게 배우는 R 데이터 분석, 이지스퍼블리싱.
3. 나종화. (2017). R 데이터마이닝, 자유아카데미.
4. 남길임, 조은영. (2017). 한국어 텍스트 감성 분석, 커뮤니케이션북스.
5. 조민호. (2019). 데이터 분석 전문가를 위한 R 데이터 분석. 정보문화사.

6. 주진수 외 3인. (2018). 한국농수산대학 졸업생 영농정착 성공 사례집의 Text Mining. 현장농수산연구지 Vol. 20, No.2: 57-72.
7. 주진수 외 5인. (2019). 비정형 데이터 마이닝을 활용한 한국농수산대학 재학생의 학교생활 감성 분석(1). 현장농수산연구지 Vol. 21(1), No.1: 99-114.
8. 주진수 외 5인. (2019). 한국농수산대학 재학생의 학교생활 감성 분석 및 영농의지에 관한 연구. 현장농수산연구지 Vol. 21(2), No.1: 103-114.
9. https://ark1st.tistory.com/25
10. https://data-traveler.tistory.com/34
11. https://is-this-it.tistory.com/39
12. https://magician-of-c.tistory.com/23
13. https://needjarvis.tistory.com/59
14. https://tour-analyst.tistory.com/3
15. https://r-pyomega.tistory.com/18

논문접수일 : 2020년 5월 18일
논문수정일 : 2020년 6월 8일
게재확정일 : 2020년 6월 12일