J. of Practical Agriculture & Fisheries Research ISSN 2234-1129
2020 Dec. : 22(2): 99-114

한국농수산대학 신입생 자기소개서의 텍스트 마이닝과 연관규칙 분석 (2)

Text Mining and Association Rules Analysis to a Self-Introduction Letter of Freshman at Korea National College of Agricultural and Fisheries (2)

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 from the self introduction letter of freshman at Korea National College of Agriculture and Fisheries(KNCAF) in 2020. The analysis items of the 3rd question were and the 4th question were the motivation for applying to college, the academic plan and the career plan.

The text mining to the 3rd question showed that the frequency of 'friends' was overwhelmingly high, followed by keywords such as 'thought', 'time', 'opinion', 'activity', and 'club'. In the 4th question, keyword frequency such as 'thought', 'agriculture', 'KNCAF', 'farm', 'father' was high.

The result of association rules analysis for each question showed that the relationship with the highest support level, which means the frequency and importance of the rule, was the {friend} <=> {thought}, {thought} <=> {KNCAF}. The confidence level of a correlation between keywords was the highest in the rules of {teacher}=>{friend}, {agriculture, KNCAF}=>{thought}. Also the lift level that indicates the closeness of two words was the highest in the rules of {friend} <=> {teacher}, {knowledge} <=> {professional}.

These keywords are found to play a very important roles in analyzing betweenness centrality and analyzing degree centrality between keywords. 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.

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 학생들이 제출하는 자기소개서(이하 자소서)는 수험생이 자유롭게 기술한 학업 생활과 과정을 구체적으로 확인할 수 있는 만큼 학생부종합전형(학종)의 핵심 평가 자료로 쓰인다. 학생부종합전형의 서류인 자소서는 수험생이 준비할 수 있고, 준비 시간과 노력 정도에 따라 평가가 달라질 수 있다. 따라서 수험생들이 지원하려는 대학·학과와 관련해 지원동기와 성장 가능성 등 자신의 강점을 얼마나 잘 표현하느냐가 관건이라 할 수 있다. 자소서는 학생 스스로 작성하는 서류란 점에서 의미가 깊으며 자신의 강점을 드러내고, 약점을 보완할 수 있다는 데 효용이 있다.

한국농수산대학(이하 한농대) 자소서는 문항 1~문항 4로 구성되는데, 문항 1~문항 3은 전국 대학 모두 한국대학교육협의회가 지정한 공통 문항이며, 문항 4는 한농대가 학교 특성에 맞게 정한 질문이다. 선행연구[1]에서 분석한 문항은 교과학습 발달 상황과 비교과 활동에 관련된 두 개의 질문이었다. 본 연구에서 다루는 문항 3은 학교생활 중 배려·나눔·협력·갈등 관리 등의 인성과 리더십, 그리고 문항 4는 지원동기와 학업계획을 중심으로 향후 진로계획(영농·영어계획)에 관련한 질문이다.

본 연구는 선행연구의 후속으로서 2020년 한농대 신입생 자소서의 특성 파악 및 평가 자료를 획득하기 위하여 문항 3과 문항 4를 대상으로 하였다. 분석 방법은 지난 연구와 동일하게 비정형 데이터 처리 방법인 텍스트 마이닝에 의한 토픽 분석과 워드 클라우드에 의한 시각화, 그리고 연관분석을 통한 단어와 단어 사이의 연관성 분석과 연관어 네트워크에 의한 시각화를 통한 규칙과 패턴을 추출이었다.

Ⅱ. 연구내용

1. 분석 도구 및 기법

분석에는 선행연구와 동일하게 R 프로그램 작업을 편리하게 지원하기 위한 통합 개발환경 프로그램인 RStudio(버전 3.6)을 활용하였다. RStudio는 재현성 (Reproducibility)을 위한 R 스크립트 및 Project 관리, 시각화에 특히 강점이 있는 프로그램이다. 따라서 명확하고 접근하기 쉬운 프로그래밍 도구로서 통계적 추론, 데이터 분석, 머신러닝 알고리즘 등에 활용된다. 현재 R 프로그래밍 언어는 교육용이나 학계에서 사용될 뿐 아니라 구글, 우버, 에어비엔비, 페이스북 등 많은 대기업들이 다양한 빅 데이터 분석 및 예측 분석 등을 포함한 고급 분석 기술들의 연구 및 개발에 많이 활용하고 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 연관분석 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.

1) 주진수 외 5인. (2020). 한국농수산대학 신입생 자기소개서의 텍스트 마이닝과 연관규칙 분석 (1). 현장농수산연구지 Vol. 22(1), No.1: 113-130.

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

전체	도시인재 전형	농수산인재 전형	일반 전형
550	82	110	358

2. 텍스트 마이닝

텍스트 마이닝은 구조화되지 않은 비정형 텍스트 데이터에서 패턴이나 관계를 추출하고 그 안에서 의미 있는 정보나 가치를 발굴하여 해석하거나 의사결정을 지원하는 일련의 과정 또는 기술을 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 이해가 필요하다.

RStudio로 텍스트 마이닝을 하려면 먼저 패키지들을 설치해야 하는데 한글 자연어 분석 패키지인 KoNLP(Korean Natural Language Processing)이 필요하고, rJava 패키지와 memoise 패키지 등을 설치해줘야 한다. 또한 단어들 검사에는 useNIADic() 명령어로 NIA 한글 사전을 사용하였으며, 사전에 없는 단어는 본 연구에서 5,835개를 사전에 추가하였다. 추가한 단어는 선행연구보다 900여개 증가하였다. 추출한 키워드는 한눈에 텍스트 맥락을 이해할 수 있도록 word cloud 패키지를 사용하여 시각화하였다.

3. 연관분석

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

연관분석에서는 신입생의 문항별 자소서 텍스트를 개인별로 한 줄 문장으로 편집하여 각 문항을 550(학생수)개의 문장으로 수정하여 프로그램에 입력하였다. 분석 그룹은 2개의 문항(문항 3, 문항 4), 4개의 전형(도시인재 전형, 농수산인재 전형, 일반 전형, 전체)으로 구분하였다.

연관분석에서는 학생별로 하나의 문장으로 편집해야 하는데, 방법 및 이유는 선행연구에서 설명하고 있다. 연관규칙 분석은 어떤 두 아이템 집합이 빈번히 발생하는가를 알려주는 일련의 규칙들을 생성하는 알고리즘으로서 arules Package의 apriori() 함수를 사용하며, 시각화를 위하여 igraph와 arulesViz Package를 사용한다.

연관규칙의 측도는 크게 support(지지도), confidence(신뢰도) 그리고 모델의 성능 평가를 위한 lift(향상도)를 고려하였다. 지지도는 전체 글 중에서 어떠한 A 와 B를 둘 다 가지고 있는 글의 확률로서 추출한 규칙이 데이터에서 발생하는 빈도를 측정한 것으로 지지도가 크면 클수록 중요성이 높다는 것을 의미한다. 신뢰도는 예측 능력이나 정확도의 측정치로서, 신뢰도가 크면 클수록 두 단어 사이의 연관성이 높은 것을 의미한다. 향상도는 A→B의 연관규칙에서 랜덤으로 B가 사용되는 경우에 비해 A와의 관계가 고려되어 사용되는 경우의 비율이다. 즉 A, B가 우연히 나타날 확률보다 A, B 사이 관계가 밀접한지를 보는 지표로서 1보다 크면 우연히 발생하지 않았다는 의미이다. 뒤에서 반드시 사용되는 단어이다. 지지도와 향상도는 대칭 척도로서 좌우 키워드에 상관없이 값이 동일하며, 신뢰도는 비대칭 척도로서 좌우 키워드에 따라 값이 변한다.

이 함수에 의해 단순한 테이블로 정리되는 결과는 네트워크 그래프로 시각화하여 각 단어 간의 관계 파악과 그 핵심 단어를 한눈에 알 수 있게 하였다. 연관 키워드의 네트워크 그래프는 Node(노드), edge(선)로 구성된다. 본 연구에서 사용한

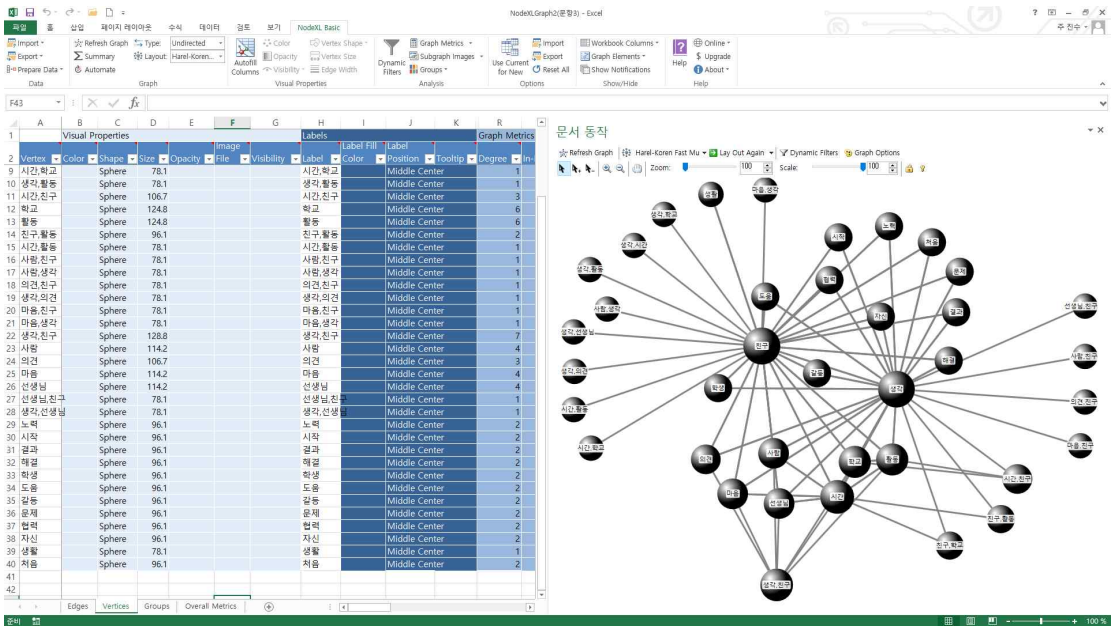

Fig. 1. Screen Capture of NodeXL

척도는 degree() 함수를 이용하여 노드에 연결된 edge의 수에 따라 중요도를 파악하는 '연결 중심성'(Fig. 1)과 전체 네트워크에서 해당 노드와 다른 노드들 사이 최단 경로를 얼마나 많이 가졌는지 측정하는 방법, 즉 노드 간의 매개체 역할을 하는 노드를 알려주는 '관계 중심성' 척도이다.

Ⅲ. 결과 및 고찰

1. 텍스트 마이닝 및 시각화

가. 문항 3

수험생들이 배려·나눔·협력·갈등관리 등을 실천한 사례와 그 과정에서 배운 점 등 인성과 리더십을 기술한 문항 3에 대한 키워드 분석 결과를 Table 2와 Table 3에 나타내었다. 문항 3에서 추출한 명사는 53,493 단어로 전처리 작업을 통하여 키워드를 정리하였다.

Table 2는 키워드의 빈도 분석 결과로서, 평균 빈도는 각 전형별로 나타난 각각의 키워드 빈도를 모집 전형별 인원으로 나누어 표준화한 값을 의미한다. 모든 전형에서 '친구'는 1순위 키워드로 나타났으며, 농수산인재 전형(5.73회)에서 가장 높으며, 전체 학생의 사용 빈도는 1인당 평균 5.47회로 나타났다. 또한 '생각'은 모든 전형에서 2순위로 나타났으며, 전체 평균 빈도는 2회로 나타났다.

이외에도 10위까지의 키워드에는 '시간', '의견', '활동', '학교', '선생님', '동아리' 등의 키워드가 추출되었다. 이는 학교생활 중 동아리 활동을 비롯하여 다양한 집단에 속하여 활동한 사례와 그 과정에서 배우고 느낀 배려·나눔·협력·갈등 관리 등을 서술해야 하는 문항의 성격이 나타난 결과로 볼 수 있다.

전체 학생의 10위까지 키워드(붉은색 표기)를 기준으로 각 전형별 순위를 보면 큰 차이는 없으나, 키워드 '동아리'는 도시인재 전형(27위)과 농수산인재 전형(14위)에서 차이가 나타났다. 표에는 나타내지 않았으나 도시인재 전형에서 '동아리'의 평균 빈도는 0.1회로서 다른 입학 전형에 비해 동아리 활동에 대한 서술이 적었다.

Table 2. Keywords by the admission selection in question 3 (단위 : 회/인)

순위	전 체		도시인재 전형		농수산인재 전형		일반 전형	
	키워드	평균 빈도	키워드	평균 빈도	키워드	평균 빈도	키워드	평균 빈도
1	친구	5.47	친구	4.72	친구	5.73	친구	5.56
2	생각	2.00	생각	2.22	생각	1.94	생각	1.97
3	시간	1.19	사람	1.26	학교	1.35	시간	1.21
4	의견	1.12	시간	1.20	학생	1.25	의견	1.13
5	활동	1.07	활동	1.13	의견	1.25	활동	1.07
6	사람	1.04	의견	0.88	시간	1.13	사람	1.03
7	학교	1.00	학교	0.87	활동	1.05	학교	0.92
8	선생님	0.88	갈등	0.82	선생님	1.03	선생님	0.85
9	학생	0.87	선생님	0.78	사람	0.92	학생	0.83
10	동아리	0.69	마음	0.77	갈등	0.75	동아리	0.76
11	갈등	0.69	협력	0.73	참여	0.70	마음	0.67
12	마음	0.68	학급	0.68	청소	0.68	문제	0.67
13	문제	0.66	상황	0.63	문제	0.67	해결	0.66
14	해결	0.65	봉사	0.60	동아리	0.66	갈등	0.64
15	도움	0.61	해결	0.60	마음	0.66	도움	0.64
16	학급	0.61	문제	0.57	해결	0.65	노력	0.61
17	노력	0.57	도움	0.54	아이	0.63	학급	0.61
18	협력	0.56	아이	0.52	준비	0.62	생활	0.55
19	생활	0.53	결과	0.49	배려	0.61	결과	0.54
20	결과	0.52	서로	0.49	도움	0.57	협력	0.52

Table 3. Word clouds by the admission selection in question 3

전체	도시인재 전형
	
농수산인재 전형	일반 전형
	

Table 3은 각 전형별로 추출한 상위 50위까지의 키워드를 시각화한 워드 클라우드이다. 그래프의 키워드 크기는 Table 2에 나타난 바와 같이 '친구', '생각', '시간', '의견', '활동', '사람' 등의 빈도순으로 작아지는 것을 알 수 있다. 특히 다른 단어에 비하여 사용 빈도가 탁월하게 높은 '친구'는 비교할 수 없을 만큼 크게 표현되는 것을 알 수 있다.

나. 문항 4

대학마다 질문 내용이 다른 문항 4는 지원동기와 학업계획을 중심으로 향후 진로계획(영농·영어계획) 등의 서술을 요구하고 있다. 문항 4에 대한 키워드 분석 결과를 Table 4와 Table 5에 나타내었다. 문항 4에서 추출한 명사는 84,129개로 문항 3보다 많은 단어가 추출되었다.

Table 4는 키워드의 빈도 분석 결과로서 모든 전형에서 '생각', '농업', '한농대', '농장', '지식' 등의 키워드가 순서는 약간 다르지만 상위 5위 안에 나타난 것을 알 수 있다. 전형별로는 도시인재 전형에서 추출된 키워드 빈도가 다른 전형에 비하여 대부분 높게 나타난 것을 알 수 있다. 즉 다른 전형 학생들에 비하여 해당 키워드를 많이 사용하는 것으로 나타났다.

빈도수 상위 10위 안에 나타난 키워드 가운데 '생각'은 선행연구 문항 1과 2에서와 마찬가지로 본 연구의 문항 3과 4에서도 신입생들이 많이 사용한 단어로서 1, 2위의 높은 빈도수를 나타냈다. 그러나 '아버지', '한농대', '전문', '농장' 등의 단어는 문항 4에서만 상위에 나타나는 특징을 나타냈다, 이러한 결과는 대학 지원동기, 학업계획 및 향후 영농·영어계획을 기술한 한농대의 특성을 잘 나타낸 것이라고 할 수 있다. Table 4에는 나타내지 않았으나 농수산인재 전형에서 '아버지'의 빈도는 0.81(28위)로 다른 전형에 비하여 낮은 빈도를 나타냈다.

Table 4. Keywords by the admission selection in question 4 (단위 : 회/인)

순위	전 체		도시인재 전형		농수산인재 전형		일반 전형	
	키워드	평균 빈도	키워드	평균 빈도	키워드	평균 빈도	키워드	평균 빈도
1	생각	2.94	생각	3.26	농업	2.78	생각	2.93
2	농업	2.63	농업	3.02	생각	2.72	농업	2.49
3	한농대	1.71	농장	2.12	지식	1.84	한농대	1.68
4	농장	1.54	한농대	1.83	한농대	1.70	농장	1.41
5	지식	1.45	지식	1.50	농장	1.51	지식	1.31
6	공부	1.24	사람	1.32	공부	1.45	사람	1.22
7	사람	1.18	관심	1.28	전문	1.28	공부	1.20
8	관심	1.15	아버지	1.28	재배	1.25	관심	1.15
9	전문	1.14	전문	1.21	관심	1.09	아버지	1.11
10	아버지	1.07	공부	1.17	생산	1.09	전문	1.08
11	기술	1.07	기술	1.15	운영	1.04	기술	1.07
12	재배	1.01	작물	1.12	기술	1.01	경험	0.97
13	경험	0.98	경험	1.10	지원	0.98	부모님	0.94
14	지원	0.95	다양	1.09	판매	0.96	지원	0.93
15	생산	0.91	재배	1.05	사람	0.95	재배	0.92
16	부모님	0.88	지원	1.01	경험	0.94	농사	0.90
17	농사	0.86	관련	1.00	관련	0.94	생산	0.86
18	다양	0.85	경영	0.98	실습	0.93	학교	0.81
19	작물	0.84	농사	0.95	다양	0.90	졸업	0.80
20	졸업	0.84	졸업	0.94	작물	0.90	다양	0.78

Table 5는 문항 3과 같이 각 전형별 상위 빈도 50위까지의 키워드를 시각화한 워드 클라우드이다. 키워드는 Table 4에 나타난 바와 같이 '생각', '농업', '한국농수산대학', '농장', '지식'. '관심', '농장' 등의 빈도순으로 작아지는 것을 알 수 있다. 특히 붉은색 타원으로 표시한 '아버지' 크기를 비교해 보면, 빈도가 낮게 나타난 농수산인재 전형(28위)에서 작게 시각화된 것을 알 수 있다.

Table 5. Word clouds by the admission selection in question 4

전체	도시인재 전형
	
농수산인재 전형	일반 전형
	

2. 연관분석 및 시각화

가. 문항 3

Table 6은 문항 3의 분석 결과이다. 분석은 불필요한 연산을 줄이고 좋은 규칙을 찾기 위하여 기준값을 support(지지도)=0.3, confidence(신뢰도)=0.35로 설정하였다. itemMatrix 사이즈는 550×5,921의 메트릭스로 분석되었다.

Table 2에서 빈도 1순위 키워드인 '친구'를 중심으로 살펴보면, {친구} <=> {생각} 규칙의 지지도가 0.645로 가장 높게 나타났다. 이 규칙은 문항 3에서 가장 중요성이 높은 단어이며, 전체 글 중에서 '친구'와 '생각'을 둘 다 가지고 있는 글의 비율이 약 64.5%라는 것을 의미한다. 이 규칙의 신뢰도는 0.787로서 두 단어 사이의 연관성은 약 78.7%의 비교적 높은 것으로 나타났다. 이 때 '친구'가 '생각'을 사용하도록 유발하는 신뢰도와 '생각'이 '친구' 사용을 유발하는 비대칭 척도인 신뢰도(0.808) 값이 같지 않다는 점에 주의할 필요가 있다. 이것은 자소서 3번 문항에서 학생들은 '친구'를 사용할 때 '생각'을 함께 사용하는 비율이 78.7%인 반면, '생각'을 사용할 때 '친구'를 함께 사용하는 비율은 80.8% 연관된다는 것을 의미한다. 그러나 두 단어의 밀접성을 나타내는 향상도(0.986)는 1보다 작아 한 단어가 나타나면 반드시 뒤에서 나머지 단어가 나타나는 것은 아닌 관계로 나타났다. {친구} <=> {선생님} 규칙의 지지도와 신뢰도는 {친구} <=> {생각}과 차이가 있지만

Table 6. Association rules in question 3

번호	규칙 {lhs} => {rhs}	지지도 (support)	신뢰도 (confidence)	향상도 (lift)	빈도 (frequency)
1	{도움} => {친구}	0.301818	0.846939	1.032852	166
2	{친구} => {도움}	0.301818	0.368071	1.032852	166
3	{노력} => {친구}	0.312727	0.855721	1.043563	172
4	{친구} => {노력}	0.312727	0.381375	1.043563	172
5	{선생님} => {생각}	0.301818	0.775701	0.971835	166
6	{생각} => {선생님}	0.301818	0.378132	0.971835	166
7	{선생님} => {친구}	0.343636	0.883178	1.077046	189
8	{친구} => {선생님}	0.343636	0.419069	1.077046	189
9	{마음} => {생각}	0.34	0.827434	1.036648	187
10	{생각} => {마음}	0.34	0.425968	1.036648	187
11	{마음} => {친구}	0.349091	0.849558	1.036046	192
12	{친구} => {마음}	0.349091	0.425721	1.036046	192
13	{의견} => {생각}	0.329091	0.826484	1.035458	181
14	{생각} => {의견}	0.329091	0.412301	1.035458	181
15	{의견} => {친구}	0.316364	0.794521	0.968927	174
16	{친구} => {의견}	0.316364	0.385809	0.968927	174
17	{사람} => {생각}	0.410909	0.840149	1.052578	226
18	{생각} => {사람}	0.410909	0.514806	1.052578	226
19	{사람} => {친구}	0.396364	0.810409	0.988304	218
20	{친구} => {사람}	0.396364	0.48337	0.988304	218
21	{활동} => {생각}	0.374545	0.792308	0.992641	206
22	{생각} => {활동}	0.374545	0.469248	0.992641	206
23	{활동} => {친구}	0.38	0.803846	0.9803	209
24	{친구} => {활동}	0.38	0.463415	0.9803	209
25	{학교} => {시간}	0.301818	0.584507	0.995291	166
26	{시간} => {학교}	0.301818	0.513932	0.995291	166
27	{학교} => {생각}	0.405455	0.785211	0.98375	223
28	{생각} => {학교}	0.405455	0.507973	0.98375	223
29	{학교} => {친구}	0.44	0.852113	1.039162	242
30	{친구} => {학교}	0.44	0.536585	1.039162	242
31	{시간} => {생각}	0.463636	0.789474	0.98909	255
32	{생각} => {시간}	0.463636	0.580866	0.98909	255
33	{시간} => {친구}	0.501818	0.854489	1.04206	276
34	{친구} => {시간}	0.501818	0.611973	1.04206	276
35	{생각} => {친구}	0.645455	0.808656	0.986166	355
36	{친구} => {생각}	0.645455	0.78714	0.986166	355
37	{사람,생각} => {친구}	0.327273	0.79646	0.971293	180
38	{사람,친구} => {생각}	0.327273	0.825688	1.034461	180
39	{생각,친구} => {사람}	0.327273	0.507042	1.036703	180
40	{생각,학교} => {친구}	0.345455	0.852018	1.039046	190
41	{친구,학교} => {생각}	0.345455	0.785124	0.983641	190
42	{생각,친구} => {학교}	0.345455	0.535211	1.036501	190
43	{생각,시간} => {친구}	0.390909	0.843137	1.028216	215
44	{시간,친구} => {생각}	0.390909	0.778986	0.97595	215
45	{생각,친구} => {시간}	0.390909	0.605634	1.031265	215

향상도는 1보다 크게 나타나 두 단어 사용은 우연히 발생하지 않는 밀접한 관계의 규칙으로 나타났다.

Table 6의 연관규칙 분석 결과에서 규칙의 발생 빈도이자 중요성을 나타내는 지지도가 가장 높게 나타난 관계는 {친구} <=> {생각}이며, 두 단어 사이에 연관성을 나타내는 신뢰도가 가장 높게 나타난 관계는 {선생님} => {친구}이다. 또한 두 단어 간 밀접성을 나타내는 향상도가 가장 높게 나타난 관계는 {친구} <=> {선생님} 관계에서 나타났다.

Fig. 2는 Table 6에서 {LHS} => {RHS} 키워드 간의 연관규칙을 지지도와 향상도의 관계를 igraph와 arulesViz Package를 이용한 시각화 결과이다. 지지도가 높을수록 원이 크게 나타나며, 향상도가 높을수록 원 내부의 색상이 더욱더 붉은 색으로 나타난다. 앞에서 설명한 지지도가 가장 높은 {친구} <=> {생각} 규칙의 원이 크게 나타났으며, 향상도가 높은 {친구} <=> {선생님} 규칙에서 붉은색이 진하게 나타났다. 또한 우측 수직축의 {RHS} 키워드 {친구}와 {생각} 축에 {LHS} 키워드들이 많이 연관되어 있어 이들 키워드가 중심 매개체 역할의 키워드라는 것을 알 수 있다.

Fig. 3은 키워드 간의 매개체 역할을 하는 키워드를 알려주는 '관계 중심성' 분석 결과로서 '친구', '생각', '학교', '시간' 및 '사람' 등이 다른

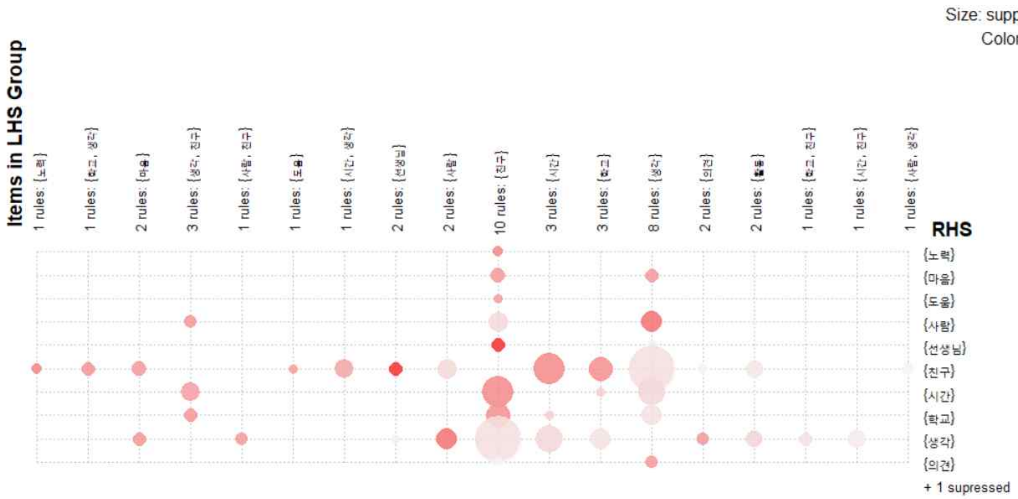

Fig. 2. Grouped graph for association rules to question 3

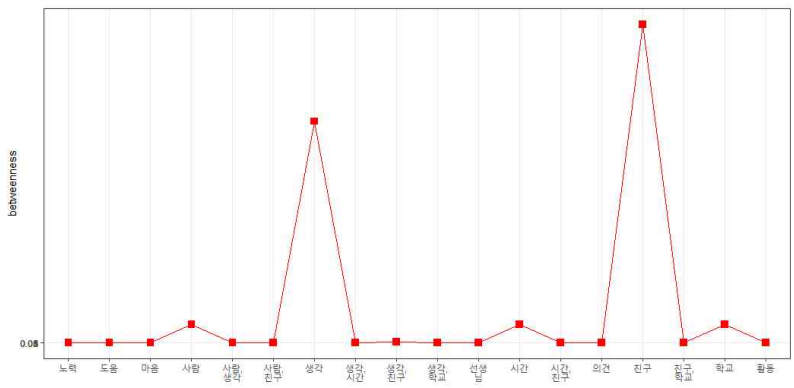

Fig. 3. Association graph for betweenness centrality to question 3

키워드들 사이에 최단 경로를 많이 가지고 있어 관계 중심성이 높은 것을 알 수 있다.

Fig. 4는 노드와 edge로 구성되는 네트워크 그래프로서 키워드를 연결하는 edge 수에 따라 노드의 라벨 크기로 표현되는 '연결 중심성' 분석 결과로서, 시각화 효과를 위하여 기준값(sup. = 0.25, conf. = 0.35)을 Table 6과 다르게 설정하여 추출한 117개 규칙의 연관도이다. '친구', '생각', '학교' 및 '시간'의 라벨이 크게 나타나 이들 키워드는 다른 키워드 사이에서 연결 중심성이 높은 것을 알 수 있다.

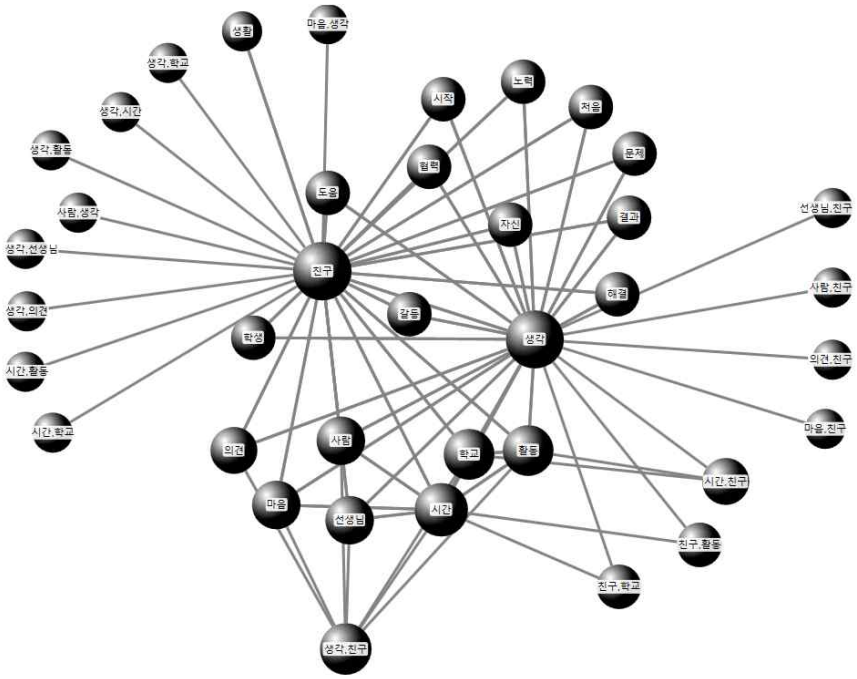

Fig. 4. Association graph for degree centrality tor question 3

Fig. 5. Word cloud for question 3 in association rules analysis

Fig. 5는 연관분석으로 추출한 빈도수 상위 50위까지 키워드를 시각화한 워드 클라우드이다. 앞에서 나타낸 Table 3(전체)의 텍스트 마이닝에 의한 워드 클라우드와 비슷한 듯 보이나 Table 7(상위 20위)에 나타낸 바와 같이 키워드의 순위에 차이가 있음을 알 수 있다. 이는 한 사람이 하나의 단어를 반복하여 사용하는 경우 그 단어를 하나로 처리하는 연관분석 기법에 따른 결과이다.

Table 7은 텍스트 마이닝 결과와 연관분석 결과의 상위 20위까지를 비교한 표이다. 텍스트 마이닝에 의한 ②평균 빈도는 키워드 ①빈도를 전체 학생수로 나눈 값이며, ④실제 평균 빈도는 해당 키워드를 사용한 학생수로 나눈 값을 의미한다. 표를 보면 키워드 '친구'는 텍스트 마이닝 결과에서는 1인당 평균 5.4회가 사용되었으나 실제로 이 단어를 사용한 학생을 고려한 연관분석에 의한 실제 평균 빈도는 6.67회로 약간 높아지는 것을 알 수 있다. 즉 키워드 '친구'는 대부분의 학생들이 다른 어느 단어들보다 많이 사용하는 키워드라는 것을 알 수 있다.

또한 표에는 모두 나타내지 않았으나 키워드 '동아리'의 텍스트 마이닝에 의한 ②평균 빈도는 상위 10위에 나타났으나 연관분석에서는 실제로 '동아리'를 사용한 학생은 99명으로 45위에 나타났다. 그러나 ④실제 평균 빈도를 구해보면 학생 한 사람당 '동아리' 사용 빈도는 평균 3.86회로 '친구' 다음으로 높게 나타났으며, 이는 소수 학생만이 반복적으로 사용한 키워드라는 것을 알 수 있다.

Table 7. The frequency of key wordsby text mining and association rules analysis to question 3

순위	텍스트 마이닝			순위	연관분석		실제 평균 빈도**
	키워드	빈도①*	평균 빈도②**		키워드	사용자③***	④ = ①/③
1	친구	3,007	5.47	1	친구	451	6.67
2	생각	1,100	2.00	2	생각	439	2.51
3	시간	656	1.19	3	시간	323	2.03
4	의견	614	1.12	4	학교	284	1.94
5	활동	591	1.07	5	사람	269	2.13
6	사람	574	1.04	6	활동	260	2.27
7	학교	550	1.00	7	마음	226	1.66
8	선생님	482	0.88	8	의견	219	2.80
9	학생	476	0.87	9	선생님	214	2.25
10	동아리	381	0.69	10	노력	201	1.57
11	갈등	379	0.69	11	시작	199	1.43
12	마음	376	0.68	12	도움	196	1.71
13	문제	361	0.66	13	해결	196	1.82
14	해결	356	0.65	14	갈등	193	1.96
15	도움	335	0.61	15	결과	192	1.50
16	학급	334	0.61	16	협력	192	1.60
17	노력	315	0.57	17	학생	185	2.57
18	협력	307	0.56	18	문제	184	1.96
19	생활	294	0.53	19	자신	178	1.51
20	결과	288	0.52	20	처음	176	1.26

* : (회), ** : (회/명), *** : (명)

나. 문항 4

Table 8은 문항 4의 연관규칙 분석 결과로서 기준값은 sup. = 0.41, conf. = 0.65으로 설정하였다. itemMatrix는 550×8,211의 메트릭스로 분석

Table 8. Association rules in the question 4

번호	규칙 {lhs} => {rhs}	지지도 (support)	신뢰도 (confidence)	향상도 (lift)	빈도 (frequency)
1	{사람} => {생각}	0.4709091	0.8809524	1.0052361	259
2	{기술} => {한국농수산대학}	0.4109091	0.8042705	1.0239555	226
3	{기술} => {생각}	0.4545455	0.8896797	1.0151947	250
4	{경험} => {한국농수산대학}	0.4309091	0.8116438	1.0333428	237
5	{경험} => {생각}	0.4763636	0.8972603	1.0238447	262
6	{공부} => {한국농수산대학}	0.4763636	0.8213166	1.0456577	262
7	{공부} => {생각}	0.5127273	0.8840125	1.008728	282
8	{졸업} => {지원}	0.4109091	0.7040498	1.0938628	226
9	{졸업} => {지식}	0.4254545	0.728972	1.0806862	234
10	{졸업} => {한국농수산대학}	0.4709091	0.8068536	1.0272441	259
11	{졸업} => {생각}	0.52	0.8909657	1.0166621	286
12	{농업} => {한국농수산대학}	0.4490909	0.771875	0.9827112	247
13	{농업} => {생각}	0.5236364	0.9	1.026971	288
14	{전문} => {지식}	0.4509091	0.7630769	1.1312461	248
15	{지식} => {전문}	0.4509091	0.6684636	1.1312461	248
16	{전문} => {한국농수산대학}	0.4781818	0.8092308	1.0302707	263
17	{전문} => {생각}	0.5290909	0.8953846	1.0217044	291
18	{관심} => {한국농수산대학}	0.4872727	0.7928994	1.0094784	268
19	{관심} => {생각}	0.5472727	0.8905325	1.0161678	301
20	{지원} => {지식}	0.4436364	0.6892655	1.0218222	244
21	{지식} => {지원}	0.4436364	0.6576819	1.0218222	244
22	{지원} => {한국농수산대학}	0.5181818	0.8050847	1.0249922	285
23	{한국농수산대학} => {지원}	0.5181818	0.6597222	1.0249922	285
24	{지원} => {생각}	0.5709091	0.8870056	1.0121434	314
25	{생각} => {지원}	0.5709091	0.6514523	1.0121434	314
26	{지식} => {한국농수산대학}	0.5472727	0.8113208	1.0329315	301
27	{한국농수산대학} => {지식}	0.5472727	0.6967593	1.0329315	301
28	{지식} => {생각}	0.5927273	0.8787062	1.002673	326
29	{생각} => {지식}	0.5927273	0.6763485	1.002673	326
30	{한국농수산대학} => {생각}	0.6909091	0.8796296	1.0037268	380
31	{생각} => {한국농수산대학}	0.6909091	0.7883817	1.0037268	380
32	{공부,한국농수산대학} => {생각}	0.42	0.8816794	1.0060657	231
33	{공부,생각} => {한국농수산대학}	0.42	0.8191489	1.042898	231
34	{졸업,한국농수산대학} => {생각}	0.4254545	0.9034749	1.0309361	234
35	{생각,졸업} => {한국농수산대학}	0.4254545	0.8181818	1.0416667	234
36	{농업,한국농수산대학} => {생각}	0.4109091	0.9149798	1.044064	226
37	{농업,생각} => {한국농수산대학}	0.4109091	0.7847222	0.9990676	226
38	{전문,한국농수산대학} => {생각}	0.4254545	0.8897338	1.0152565	234
39	{생각,전문} => {한국농수산대학}	0.4254545	0.8041237	1.0237686	234
40	{관심,한국농수산대학} => {생각}	0.4327273	0.8880597	1.0133461	238
41	{관심,생각} => {한국농수산대학}	0.4327273	0.7906977	1.0066753	238
42	{지원,한국농수산대학} => {생각}	0.4581818	0.8842105	1.0089539	252
43	{생각,지원} => {한국농수산대학}	0.4581818	0.8025478	1.0217622	252
44	{생각,한국농수산대학} => {지원}	0.4581818	0.6631579	1.0303301	252
45	{지식,한국농수산대학} => {생각}	0.4872727	0.8903654	1.0159772	268
46	{생각,지식} => {한국농수산대학}	0.4872727	0.8220859	1.0466371	268
47	{생각,한국농수산대학} => {지식}	0.4872727	0.7052632	1.0455384	268

되었다. Table 4에서 평균 빈도가 가장 높은 키워드 '생각'을 중심으로 살펴보면, {생각} => {한국농수산대학} 규칙 지지도는 0.69로 가장 높게 나타났다. 이 규칙은 문항 4에서 가장 중요성이 높은 단어이며, 전체 글 중에서 '생각'과 '한국농수산대학'을 둘 다 동시에 사용하는 글의 비율이 약 69%라는 것을 의미한다. 이 규칙의 신뢰도는 0.788로서 두 단어 사이의 연관성은 78.8%로 비교적 높게 나타났으며, 두 단어의 밀접성을 나타내는 향상도는 1보다 큰 양의 관계로서 두 단어가 우연히 함께 사용되지 않고 서로 밀접한 관계가 있는 것으로 나타났다.

Table 8에서 규칙의 발생 빈도이자 중요성을 나타내는 지지도가 가장 높은 관계는 {생각} => {한국농수산대학}이며, 두 단어 사이에 연관성을 나타내는 신뢰도가 가장 높게 나타난 관계는 {농업, 한국농수산대학} => {생각}이다. 또한 두 단어 간 관계의 밀접성을 보는 향상도가 가장 높게 나타난 관계는 {지식} => {전문} 규칙으로 나타났다.

Fig. 6은 Table 8에 나타낸 연관규칙을 지지도와 향상도의 관계로 시각화 한 그래프이다. 앞에서 설명한 지지도가 가장 높은 {생각} <=> {농수산대학} 규칙의 원이 크게 나타났으며, 향상도가 높은 {전문} <=> {지식} 규칙에서 붉은색이 진하게 나타났다. 우측 수직축의 {RHS}의 {생각}과 {한국농수산대학} 축에 {LHS} 키워드들이 많이 연관되어 있어 이들 키워드가 중심 매개체 역할의 키워드라는 것을 알 수 있다.

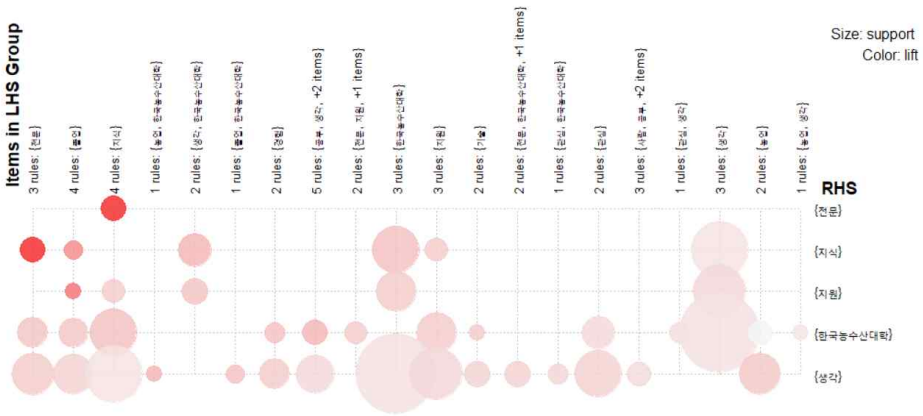

Fig. 6. Grouped graph for association rules to the question 4

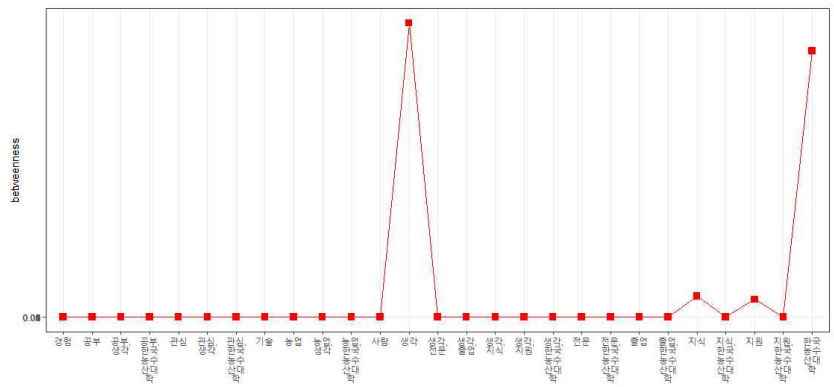

Fig. 7. Association graph for betweenness centrality to the question 4f

Fig. 7은 키워드 간의 매개체 역할을 하는 키워드를 알려주는 '관계 중심성' 분석 결과로서 '생각', '한국농수산대학', '지식' 및 '지원' 등이 다른 키워드들 사이에 최단 경로를 많이 가지고 있어 관계 중심성이 높은 것을 알 수 있다.

Fig. 8은 연결 중심성의 시각화를 보다 효과적으로 나타내기 위하여 함수 apriori()의 기준값을 sup.=0.35, conf.=0.5로 설정하여 추출한 연관도이다. 이 결과를 보면 Table 9의 연관분석 결과의 상위 키워드로 나타난 '생각', '한국농수산대학', '지식', '지원', '관심' 및 '졸업' 등의 노드 라벨이 명확하게 크게 나타나 이들 키워드가 다른 키워드 사이에서 연결 중심성이 높은 것을 알 수 있다.

Fig. 9는 연관분석으로 추출한 키워드의 빈도수 상위 50위까지를 시각화한 워드 클라우드로서 '생각', '한국농수산대학', '지식', '지원', '관심' 등

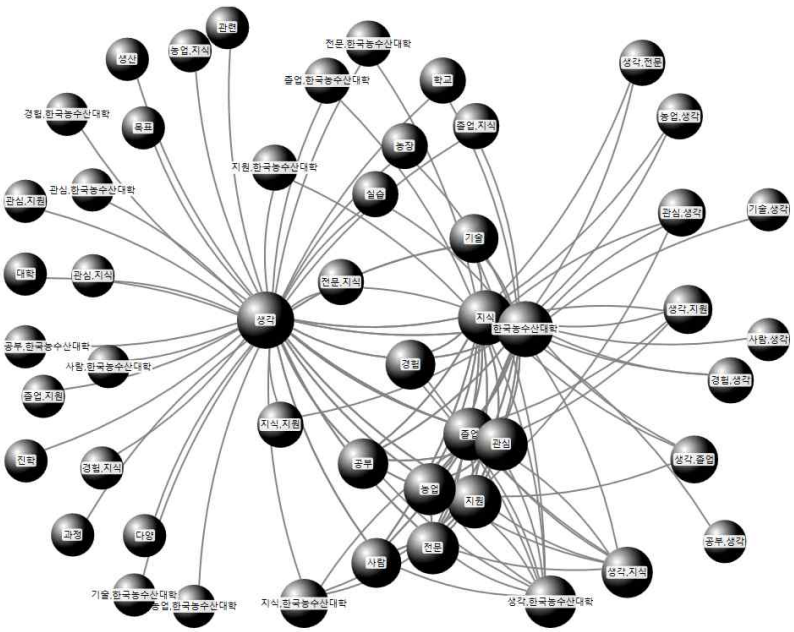

Fig. 8. Association graph for degree centrality to the question 4

Fig. 9. Word cloud for the question 4 in association rules analysis

상위 키워드가 눈에 띄게 나타난 것을 알 수 있다. Fig. 5에서 설명한 바와 같이 반복 단어를 제외하는 연관분석 기법의 특성 때문에 텍스트 마이닝에 의한 키워드 빈도 순위와는 차이(Table 9 참고)가 나타났다.

Table 9는 문항 4의 텍스트 마이닝 결과와 연관분석 결과를 상위 20위까지 비교하여 나타낸 표이다. ① 빈도, ② 평균 빈도, ③ 사용자 및 ④ 실제 평균 빈도는 Table 7에서 설명한 바와 같다. 연관분석 결과를 보면 '생각', '한국농수산대학', '지식', '지원', '관심' 등의 순으로 ③ 사용자가 많지만 ④ 실제 평균 빈도는 '농업', '생각', '농장', '지식' 순으로 순위가 변하는 것을 알 수 있다. 표에는 모두 나타내지 않았으나 텍스트 마이닝에 의한 20위까지의 키워드 가운데 '아버지', '부모님', '농사' 및 '작물'은 연관분석에 의한 ③ 사용자 순위는 35~39위에 나타났으나, ④ 실제 평균 빈도는 각각 3.19(4위), 2.45(8위), 2.52(7위) 및 2.58(5위)로 높게 나타났다. 이러한 현상은 적은 수의 학생들이 이 단어들을 반복적으로 많이 사용했기 때문이며, 문항 4에 대한 특성이 잘 반영된 결과라 할 수 있다.

Table 9. The frequency of key words by text mining and association rules analysis to the question 4

순위	텍스트 마이닝			순위	연관분석		실제 평균 빈도**
	키워드	빈도①*	평균 빈도②**		키워드	사용자③***	④ = ①/③
1	생각	1,616	2.94	1	생각	482	3.35
2	농업	1,447	2.63	2	한국농수산대학	432	2.17
3	한국농수산대학	939	1.71	3	지식	371	2.53
4	농장	845	1.54	4	지원	354	1.48
5	지식	795	1.45	5	관심	338	1.88
6	공부	684	1.24	6	전문	325	1.93
7	사람	647	1.18	7	졸업	321	1.44
8	관심	635	1.15	8	농업	320	4.52
9	전문	627	1.14	9	공부	319	2.14
10	아버지	591	1.07	10	사람	294	2.20
11	기술	589	1.07	11	경험	292	1.85
12	재배	554	1.01	12	기술	281	2.10
13	경험	539	0.98	13	농장	253	3.34
14	지원	525	0.95	14	학교	251	1.73
15	생산	498	0.91	15	다양	243	1.92
16	부모님	486	0.88	16	생산	237	2.10
17	농사	472	0.86	17	과정	234	1.82
18	다양	467	0.85	18	관련	234	1.84
19	작물	464	0.84	19	실습	231	1.65
20	졸업	461	0.84	20	목표	229	1.57

* : (회), ** : (회/명), *** : (명)

Ⅳ. 적요

본 연구는 2020년 한농대 입학생의 자소서에 서술된 학생들의 다양한 교내외 활동, 대학 지원 동기, 학업계획 및 향후 영농·영어계획 등의 텍스트 데이터를 대상으로 텍스트 마이닝에 의한 토픽 분석과 연관성 분석을 하였다.

텍스트 마이닝 결과에서 문항 3의 동아리 활동을 비롯한 다양한 활동 사례와 그 과정에서 배우고 느낀 점에 대한 키워드는 '친구' 빈도가 압도

적으로 높았으며, ‘생각’, ‘시간’, ‘의견’, ‘활동’, ‘사람’, ‘학교’, ‘선생님’, ‘학생’, ‘동아리’ 등의 키워드 순으로 많이 사용되었다. 문항 4의 대학 지원동기 및 졸업 후 진로계획에 대한 서술 데이터에는 ‘생각’, ‘농업’, ‘한농대’, ‘농장’, ‘지식’, ‘공부’, ‘사람’, ‘관심’, ‘전문’, ‘아버지’ 등의 키워드 빈도가 높게 나타났으며, 이 가운데 ‘아버지’, ‘한농대’, ‘전문’, ‘농장’ 등의 키워드는 다른 질문에 비하여 상위에 나타나는 특징을 보였다.

연관규칙 분석 결과에서 키워드 간 규칙의 발생 빈도이자 중요성을 나타내는 지지도는 문항 3에서 {친구} <=> {생각}, 문항 4에서 {생각} <=> {한국농수산대학} 규칙에서 가장 높게 나타났다. 두 단어 사이 연관성을 나타내는 신뢰도는 문항 3에서 {선생님} => {친구}, 문항 4에서 {농업, 한국농수산대학} => {생각}에서 높게 나타났다. 두 단어 간 밀접성을 나타내는 향상도는 문항 3에서 {친구} <=> {선생님}, 문항 4에서 {지식} <=> {전문}에서 높게 나타났다. 즉 두 단어는 우연히 함께 사용되지 않고 한 단어가 나타나면 뒤에 반드시 나머지 단어가 사용되었다는 것을 의미한다. 또한 키워드 간의 매개체 역할의 분석, 즉 키워드들 사이에 최단 경로를 파악하는 관계 중심성 분석과 연결 edge 수를 평가하는 연결 중심성 분석에서 문항 3은 ‘친구’, ‘생각’, ‘학교’, ‘시간’ 및 ‘사람’, 문항 4는 ‘생각’, ‘한국농수산대학’, ‘지식’ 및 ‘지원’ 등의 키워드의 중심성이 매우 높은 결과를 나타냈다.

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. (2020). 한국농수산대학 신입생 자기소개서의 텍스트 마이닝과 연관규칙 분석 (1). 현장농수산연구지 Vol. 22(1), No.1: 113-130.
9. https://is-this-it.tistory.com/39
10. https://magician-of-c.tistory.com/23
11. https://needjarvis.tistory.com/59
12. https://tour-analyst.tistory.com/3
13. https://r-pyomega.tistory.com/18

논문접수일 : 2020년 10월 30일
논문수정일 : 2020년 11월 25일
게재확정일 : 2020년 12월 3일