토픽 모델링을 활용한 꿀 연구 동향 분석***

김은영
농촌진흥청 농산업경영과
e-mail: key2022@korea.kr

Honey research trend analysis Using Topic Modeling

Eun-Young, Kim
Rural Development Administration

요 약

양봉산물의 생태적 가치와 영향력이 증대되고 있지만, 사양꿀로 인한 국내산 천연꿀의 신뢰도 하락, 관세율 인하로 인한 수입꿀 가격경쟁력으로 국내 양봉시장이 큰 타격을 입을 것으로 우려된다. 이에 본 연구에서는 꿀 R&D의 재고와 향후 미래연구에 대한 지표를 목적으로 비지도학습 알고리즘인 토픽모델링을 사용하여 국내 꿀 연구 동향과 방향성을 살펴보았다. 분석 결과 4개의 토픽으로 정의할 수 있으며, '사양꿀 성분', '화학적 특성', '의료적 기능', '품질 평가'로 생산단계에 대한 연구가 대다수로 유통, 소비 연구는 미미한 실정이다. 향후 수입꿀에 대응하기 위해서는 소비자 수요를 반영한 신제품 개발과 경쟁력 강화를 통한 신뢰도 개선 등의 지속적인 연구개발의 확대가 필요하다.

1. 서론

꿀벌이 작물 생산에 미치는 영향력과 생태적 이해가 증대되면서 그 가치는 확대되고 있다. 1차 생산물인 꿀을 넘어 양봉산물을 통한 고부가가치 창출 뿐만 아니라 환경보호 및 생태계 보전 등의 다양한 공익적 가치를 창출한다. 그러나 최근 천연꿀 수입량이 늘어나면서 국내산을 대체할 수 있을 것이라는 우려가 제기되고 있다. 우리나라 기후 특성상 사계절이 뚜렷하여 겨울철과 장마철와 같은 재밀기가 아닌 시기에 벌의 생존을 위해 설탕을 먹인 사양꿀 시장이 성장했는데, 주요 수출국인 미국과 캐나다, 베트남 등에는 사계절 내내 꽃이 피는 등 밀원이 풍부하여 사양꿀의 유형이 없다. 사양꿀과 천연꿀은 육안으로 구분하기 어려워 등급을 나누기 위한 연구들이 활성화되면서 기원 식물을 확인할 수 있는 탄소동위원소비가 식품의 기준 및 규격으로 도입되었다. 탄수화물을 생성하는 식물의 광합성 경로는 밀원 자원으로 이용하는 C3 식물군과 옥수수, 사탕수수 등의 C4 식물군으로 구분된다. C3 식물군의 탄소동위원소비는 -23‰이하, C4 식물 중에서 사탕수수로 만든 설탕의 탄소동위원소비는 -11% 수준으로 측정된다. 이에 따라 식품의약품안전처에서 고시한 식품공전에 따르면, 탄소동위원소비가 -22.5‰이하이면 천연꿀 -22.5‰를 초과하면 사양꿀로 규정한다. 그러나 C3 식물군인 사탕무 설탕으로 생산된 사양꿀의 탄소동위원소비가 -22~-30‰로 측정되어 천연꿀과 판별하기 어려운 문제가 존재한다(김소민 외, 2018). 이로 인해 사양꿀을 천연꿀로 둔갑하여 저렴한 가격에 판매하는 사례가 적발되면서 국내산 꿀에 대한 신뢰가 하락하였다. 더불어 수입산 천연꿀이 국내산 천연꿀보다 가격경쟁력에서 우위에 있어 국내시장의 위험 요인으로 작용할 수 있다. 2015년 정식 발효된 한-베트남 FTA로 천연꿀 관세율은 매년 16.2% 낮춰 수입단가가 kg당 2달러 가량으로 낮아지고 있으며, 2029년 관세 철폐가 예정되어 있어 국내 양봉시장에 타격을 줄 것으로 예상된다. 실제로 관세청 수출입 실적에서 국내 천연꿀의 수출량은 2019년 대비 2022년 73% 감소하였고 수입량은 95% 증가하였다. 양봉산업이 전체적인 위기상황에 있는 만큼 양봉산업의 활성화와 수출시장 확대, 지속적인 농가소득을 기대하기 위해서는 꿀에 대한 다양한 R&D 동향 분석으로 향후 연구 방향성을 수립하고 생산품목을 확대해 나가야 할 필요가 있다.

이에 본 연구는 국내 꿀 연구 동향을 살펴봄으로써 현재까지 진행된 연구의 진척도와 연구 방향성을 살펴보고자 한다. 국내 학술 데이터를 수집하여 비지도 학습 알고리즘인 토픽 모델링을 통해 연구를 유형화하고 중점 연구를 파악함으로써 미래 연구를 수립하는데 참고 자료로 활용되기를 기대한다.

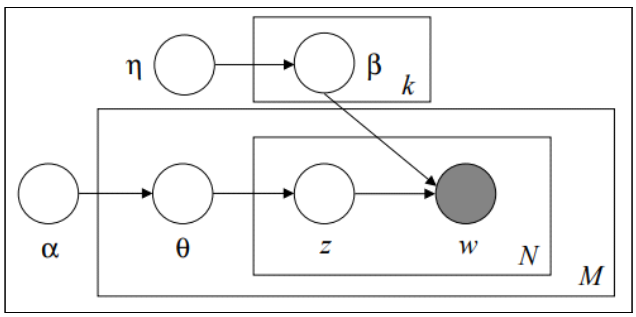
* 본 성과물은 농촌진흥청 연구사업(과제번호: RS-2020-Rd009133)의 지원에 의해 이루어진 것임
** 본 연구는 2023년도 농촌진흥청 전문연구원 과정 지원사업에 의해 이루어진 것임

2. 분석 자료

데이터는 학술연구정보서비스인 RISS(Research Information Sharing Service) 웹페이지에서 제공하는 연구 데이터이다. RISS는 전국 대학이 생산하고 보유하며 구독하는 학술자원을 공동으로 이용할 수 있도록 개방된 대국민 서비스이다. RISS에서 보유하고 있는 연구 데이터는 학위논문(2,291,160건), 음성논문(1,038,190건), 국내학술논문(6,304,375건), 해외학술논문(62,608,654건), 학술지(184,584건), 단행본(12,523,772건), 연구보고서 등(167,166건)이다. robots.txt를 통해 데이터 크롤링에 대한 허용 범위를 살펴보면, myriss와 mylibrary를 제외한 모든 페이지에 대한 접근을 허용한다. 이에 따라 본 연구에서는 '꿀'에 대한 국내학술논문과 학위논문을 중심으로 데이터를 크롤링하여 수집하였다. 데이터 수집에 사용된 프로그램은 Python 3.8 버전이다. 모듈은 BeautifulSoup과 Selenium을 사용하였다. 키워드는 "꿀"이며, 국내학술논문과 학위논문으로 분류된 연구에서 연구 연도와 연구 제목 데이터를 수집하였다. 수집된 데이터는 1965년부터 2023년까지 발간된 학술연구로 총 852건이었으며, 중복 데이터를 제거하고 남은 316건의 데이터를 분석 자료로 사용하였다.

3. 분석 방법

본 연구는 텍스트 마이닝 기반의 토픽 모델링(Topic Modeling)으로 분석을 시행한다. 토픽모델링은 문서 집합의 숨겨진 주제를 찾아내는 비지도 학습 알고리즘 중 하나이다. 분석에 사용될 모델은 LDA(Latent Dirichlet Allocation)이다. LDA는 연속 확률분포의 하나인 디리클레 다항 분포를 따르며, 문서 단어 행렬(Document-Term Matrix)의 차원을 축소해가면서 축소된 차원에서 근접 단어들을 토픽으로 묶는 기법이다. 단어의 교환성을 가정하며 예를 들어 '사과는 빨간색이다'와 '빨간색은 사과이다' 간에 차이가 없다고 간주하는 것으로 단어의 순서와 상관없이 출현 빈도에만 초점을 둔 단어의 집합이다.



[그림 1] 평활화된 LDA의 그래픽 모델
(David M. Blei. et al, 2003)

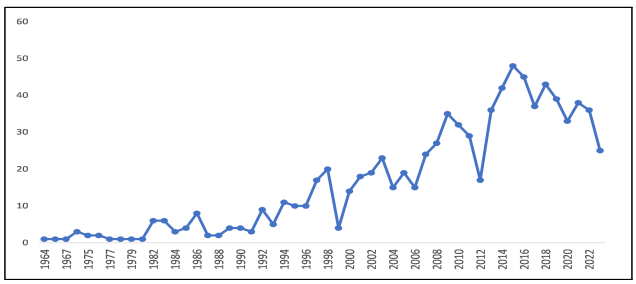
전체 토픽의 수가 k개라고 가정하는 경우, k개의 토픽에서 하나의 토픽을 고르는 것이 다항분포에 해당한다. 이때 사전 확률은 디리클레 분포로 두어 문서별 토픽 분포(θ)와 토픽별 단어 분포(β)는 디리클레 분포를 따른다고 가정한다. 문서별 토픽 분포(θ)의 하이퍼파라미터(α)와 토픽별 단어 분포(β)의 하이퍼파라미터(η)는 문서의 토픽 분포 밀도를 조절하는 디리클레 하이퍼파라미터로, 확률이 0이 되는 것을 방지하기 위해 0이 아닌 값을 가진다. 문서별 토픽 분포(θ)에 의해 단어에 대한 토픽 할당(z)이 이루어지고 토픽들에 해당하는 단어 가중치(β)를 통해 관측 가능한 단어(w)가 추정된다. 이때 결합분포는 다음의 식과 같다.

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{i=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

관측 가능한 단어(w)를 추정하면서 단어에 따라 적합한 토픽(z)을 정한다. 모든 토픽(z) 중에서 가장 가능도가 높은 토픽(z)을 찾아 문서 내 각 단어를 어떤 토픽(z)에 배정할지 최종적으로 추론한다(전은정, 2023). 본 연구에서는 Python의 konlpy 모듈을 사용한 형태소 분석과 gensim 모듈을 사용하여 LDA 분석을 시행하였다.

4. 분석 결과

먼저 꿀 연구에 대한 시계열 변화를 살펴보면, 국내산 벌꿀의 경쟁력 확보를 위해 벌꿀등급제가 시행된 2014년부터 급증하기 시작한 것을 알 수 있다.



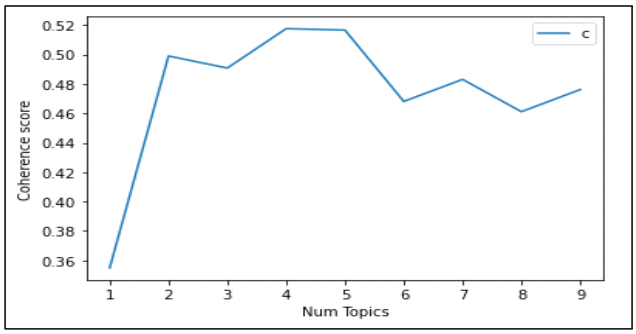
[그림 1] 꿀 연구에 대한 시계열 변화

단어 출현 빈도를 살펴보면, 꿀과 벌꿀 등을 제외하고 '연구'라는 키워드가 가장 많은 빈도를 차지했다. 다음으로 '특성'과 '품질', '분석', '이용', '생산' 등의 순으로 나타났다. 대부분의 연구가 생산단계에서 꿀의 화학 구조적 특성과 기능 성분에 대한 연구가 주를 이루는 것으로 파악되었다.

[표 2] 문서 단어 빈도

순위	단어	빈도	순위	단어	빈도
1	꿀	260	11	영향	21
2	연구	66	12	아카시아	21
3	특성	49	13	국내산	20
4	벌꿀	49	14	방법	19
5	품질	38	15	화학	18
6	분석	32	16	효과	18
7	꿀벌	32	17	토종	16
8	이용	31	18	국산	13
9	생산	27	19	변화	13
10	성분	21	20	관리	13

토픽모델링은 토픽의 수를 지정해야 한다. 최적 토픽의 수는 주제 일관성을 나태내는 지표인 coherence score를 통해 검증할 수 있으며, 토픽의 수가 4개일 때 0.5176으로 가장 높았다.



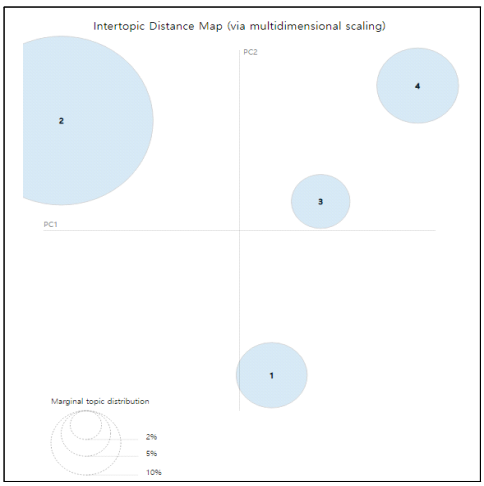
[그림 2] 최적 토픽 수

토픽모델링에 사용할 단어 사전은 사전에서 단어가 차지하는 비율과 빈도수의 기준을 정할 수 있다. 이 연구에서는 전체의 90% 이상 차지하는 단어를 제외하였다. topic1은 '사양꿀 성분', topic2는 '화학적 특성', topic3은 '의료적 기능', topic4는 '품질 평가'으로 토픽을 라벨링하였다.

[표 2] LDA 분석 결과

구분	단어	주제
Topic1	0.099*"중심" + 0.080*"사양" + 0.056*"국내산" + 0.048*"지역" + 0.047*"성분" + 0.045*"분리" + 0.041*"폐놀" + 0.040*"연구"	사양꿀 성분
Topic2	0.191*"꿀" + 0.057*"연구" + 0.048*"특성" + 0.044*"벌꿀" + 0.039*"분석" + 0.037*"이용" + 0.036*"효과" + 0.027*"품질" + 0.024*"화학"'	화학적 특성
Topic3	0.111*"환자" + 0.069*"급" + 0.064*"탐구" + 0.063*"등급" + 0.042*"땅" + 0.032*"유전자" + 0.025*"후" + 0.023*"의미" + 0.020*"추출"'	의료적 기능
Topic4	0.083*"평가" + 0.072*"피부" + 0.059*"변화" + 0.049*"판별" + 0.042*"메밀" + 0.038*"함유" + 0.038*"꿀벌" + 0.036*"토종"	품질 평가

전체 4개의 토픽 중 topic2가 가장 지배적이었으며, 다음으로 4>1>3 순이었다. 꿀 연구는 생산단계 연구였다.



[그림 3] 다차원 척도법

5. 결론 및 시사점

본 연구는 꿀의 학술적 연구 동향을 고찰하고자 토픽 모델링 분석을 시도하였다. 토픽 모델링을 시행하기 위해 연구에서 사용한 모델은 LDA이다. 분석 결과 4개의 토픽이 도출되었으며, topic1은 '사양꿀 성분', topic2는 '화학적 특성', topic3은 '의료적 기능', topic4는 '품질 평가'으로 주제를 정의하였다. 그 중 topic2가 전체 토픽 중에 지배적이었는데, 벌꿀의 품질과 화학 구조적 특성에 대한 연구였다. 또한 생산단계에서 벌꿀에 대한 연구가 주로 이루어졌으며, 유통이나 소비에 대한 연구는 이루어지지 않고 있다. 1차 생산물을 활용한 양봉산물 제품 다양화와 소비자 수요를 고려한 신제품 개발 및 브랜딩을 통한 경쟁력 강화 등으로 수입산에 대처할 수 있는 연구가 지속되어야 할 것이다. 이에 농촌진흥청에서는 지역 특화 상품으로 충청북도 괴산군에서 만든 국내산 천연벌꿀을 신제품으로 개발하였다. 밀원 종류별로 아카시아꿀, 피나무꿀, 때죽나무꿀, 야생화꿀 4가지 종류의 꿀로 소비자 선택의 폭을 확장하였으며 소량화·간편화 소비트렌드를 고려한 스틱형의 제품으로 활용도를 제고하였다.

한편, 본 연구에서 사용한 토픽 모델링은 콘텐츠, 교육, 전자공학 등 여러 분야에서 시도되고 있는 분석 방법이지만, 농식품 관련 생산물을 중심으로 분석한 연구는 거의 없다. 따라서 새로운 분야에 적용하여 분석을 시도했다는 것에 의의를 두며 연구의 결과가 관련 연구자들에게 신제품 및 기존 제품 보완 지표로 활용, 정책 입안자의 정책 수립 및 제도 보완, 농가 컨설팅 및 농업인 교육 자료로 참고할 수 있는 자료로 활용되기를 기대한다.

참고문헌

[1] 김소민, 김병희, 김문정, 김정민, Truong A Tai and 윤병수. Journal of Apiculture, "사탕무(Beta vulgaris) 설탕으로 제조된 사양꿀에서 사탕무 고유 유전자의 검출". 제 33권 3호, pp. 213-219, 2018년.

[2] Blei, David M., Andrew Y. Ng and Michael I. Jordan. the Journal of machine Learning research, "Latent dirichlet allocation". vol 3, pp. 993-1022. 2003.

[3] 전은정, 이화여자대학교 박사학위논문, " LDA 토픽모델링의 적정 표본크기 분석 연구 : 고교학점제 뉴스기사를 중심으로", 2023.