Journal of the Korea Convergence Society
Vol. 10. No. 11, pp. 159-164, 2019

ISSN 2233-4890
https://doi.org/10.15207/JKCS.2019.10.11.159

반려동물 사료 추천시스템을 위한 유사성 측정 알고리즘에 대한 연구

김삼택
우송대학교 IT융합학부 교수

A Study of Similarity Measure Algorithms for Recomendation System about the PET Food

Sam-Taek Kim
Professor, School of Information Technology Convergence, Woosong University

요 약 ICT 기술 발전으로 강아지와 고양이등 반려동물 돌보기와 건강에 대한 관심도가 높아지고 있다. 본 논문에서는 반려동물 산업의 다양한 분야에 활용될 수 있도록 반려동물 사료의 성분 데이터를 기반으로 군집분석을 수행하고 적합한 서비스에 대해 고찰한다. 군집분석을 위해 시중에서 유통되고 있는 300여 개의 강아지 및 고양이 펫푸드를 대상으로 성분별 상관관계를 분석하여 유사성을 측정하며, Hierarchical, K-Means, Partitioning around medoids(PAM), Density-based, Mean-Shift 등의 다양한 클러스터링 기법을 활용하여 군집화 하여 분석한다. 또한 반려동물의 개인화 추천시스템도 제안한다. 본 논문의 연구 결과는 반려동물을 대상으로 한 사료 추천시스템 등의 맞춤형 개인화 서비스에 활용할 수 있다.

주제어 : 빅데이터, 군집분석, 반려동물, 애완동물사료, 추천시스템, 유사성측정

Abstract Recent developments in ICT technology have increased interest in the care and health of pets such as dogs and cats. In this paper, cluster analysis was performed based on the component data of pet food to be used in various fields of the pet industry. For cluster analysis, the similarity was analyzed by analyzing the correlation between components of 300 dogs and cats in the market. In this paper, clustering techniques such as Hierarchical, K-Means, Partitioning around medoids (PAM), Density-based, Mean-Shift are clustered and analyzed. We also propose a personalized recommendation system for pets. The results of this paper can be used for personalized services such as feed recommendation system for pets.

Key Words : Big data, Cluster Analysis, Pet, PET food, Recommendation system, Similarity measure

1. 서론

최근 4차 산업혁명의 핵심기술인 빅데이터, 사물인터넷, 인공지능 등의 기술을 활용하여 인간의 삶을 바꾸려는 노력과 더불어 반려동물에게도 적용하려는 시도가 늘고 있다. 특히, 빅데이터 기술을 활용하여 반려동물에게 맞춤형 추천서비스 등의 헬스케어 서비스에 대한 연구가 늘어나고 있다. 반려동물에 대한 맞춤형 헬스케어 서비스를 구현하기 위해서는 우선적으로 반려동물이 매일 섭취하는 사료에 대한 성분을 통계적 모델을 기반으

*This research is based support of 2019 Woosong University Academic Research Funding.
*Corresponding Author : Sam-Taek Kim(stkim@wsu.ac.kr)
Received October 16, 2019 Revised November 8, 019
Accepted November 20, 2019 Published November 28, 2019

로 분석할 필요성이 있다.

따라서 본 논문에서는 260여개의 반려동물 사료 데이터의 특성을 분석하여 콘텐츠 기반 추천시스템에 활용될 수 있는 효율적인 유사성 측정 방법과 군집방법에 대해 제안한다[1-3].

논문의 구성은 다음과 같다. 제2장에서는 추천시스템, 군집분석에 대한 최근 연구를 살펴보고, 제3장에서는 반려동물 개인화 추천시스템의 구조를 설계하고, 제 4장에서는 실 사료 데이터를 기반으로 다양한 유사성 및 군집화 알고리즘을 적용하고 분석한다. 제5장에서는 실험 결과를 기반으로 실제 추천시스템에 적용할 수 있는 분야에 대해 고찰한다. 마지막으로 제6장에서는 유사성측정 방법에 대한 결론 및 앞으로 연구에 대해 논한다.

2. 관련 연구

2.1 추천시스템(Recommendation System)

추천시스템은 사용자 구매와 선호도를 예측하는 머신러닝 기법 중 하나로써 기존 연구되어온 추천시스템은 크게 협업 필터링, 콘텐츠 기반 추천시스템 지식기반 추천시스템으로 나누어질 수 있다.

2.1.1 협업 필터링(Collaborative Filtering) 추천시스템

협업하여 필터링하는 추천시스템은 사용자 간의 선호도를 고려하여 많은 선택 사항들로부터 아이템을 자동적으로 필터링하는 방법을 사용한다. 크게 사용자와 아이템 기반과 협업 필터링으로 구분하는데 사용자 기반협업 필터링(User-based CF) 은 사용자 간에 본인이 원하는 유사도를 분석해 추천하는 방식으로 아이템에 대한 분석이 불필요하다는 장점이 있지만 사용자가 늘어날수록 계산량이 급격하게 증가하거나 신규 사용자에 대한 추천 정확도가 떨어진다는 단점이 있다. 아이템을 기반으로 한 협업 필터링은 아이템의 유사도를 측정하여 특정 아이템을 구매한 사용자에게 그와 유사한 아이템을 추천해 주는 방식이다. 신규 이용자가 아이템에 대한 개인적인 평가를 가지고 있지 않은 경우 신규 사용자에게 정확한 추천이 가능한 장점이 있는 반면 초기 적은 데이터양일 때 추천 정확도가 떨어지는 단점이 있다[4,5].

2.1.2 콘텐츠를 기반으로 한 추천 시스템

콘텐츠 기반 추천시스템은 아이템들과 사용자들 간의 유사도를 측정해 사용자에게 적합한 아이템을 추천한다. 따라서 정확한 추천을 위해 주변 사용자들의 선호도 정보보다는 아이템의 속성 및 특징과 사용자의 과거 선호도를 고려하여 추천한다. 본 논문에서의 성분분석을 통한 유사 사료의 분류는 콘텐츠 기반 추천시스템에 활용되어질 수 있다[3-5].

2.1.3 지식 기반 추천시스템

지식 기반 추천시스템은 아이템의 특징과 명시적인 질문을 통해 획득한 추천 범위와 사용자 선호도에 대한 정보를 고려해 추천하는 시스템이다. 이 시스템은 사용자들의 구매 이력이 적은 경우에 효율적이다. 모델의 정확도는 추천된 아이템이 얼마나 사용자에게 유용한가를 기반으로 평가되어진다.[1-4].

2.2 차원축소(Dimensionality Reduction)

추천시스템을 구축하는 과정에서 겪는 문제 중 하나는 다양한 특징이 존재함으로 인한 고차원 데이터의 처리 문제이다. 고차원 데이터의 차원을 축소하기 위한 일반적인 방법으로 주성분 분석 PCA, SVD, NMF 등이 있다. 주성분 분석은 선형적으로 연관성이 없는 값의 집합, 즉, 주성분에 상관관계가 높은 변수들의 정사영을 사용한다. 이는 높은 상관관계가 있는 특징을 고려하여 데이터에서 변동성이 많은 것을 의미하며, 결론적으로 첫 번째의 주성분과 직접 교차하는 상관관계가 가장 낮은 특징을 사용하는 것으로 각 요소는 높은 분산성질을 갖게 된다[6-8].

2.3 군집분석(Cluster Analysis)

2.3.1 유사도 측정 알고리즘

데이터의 군집화를 위해서는 두 객체 사이 거리 또는 차이를 수치화하는 유사도 측정이 반드시 필요하다. 유사도를 측정하는 알고리즘으로는 유클리디안(Euclidean) 기반. 코사인(Cosine) 기반, 쟈카드(Jaccard) 계수 기반, 피어슨 상관계수, 맨하튼 거리 등이 있다.

코사인 유사도 측정 알고리즘은 측정하는 함수의 계산식에 다음과 같이 두개의 벡터 곱을 가지고 있다[4].

$$S_{\cos}(x,y) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=0}^{n}(x_i)^2} \times \sqrt{\sum_{i=0}^{n}(y_i)^2}}$$

위 식에서 , x, y는 주어진 두 벡터를 나타내고 각 벡

터는 n개의 차원을 가지고 있으며, 이들 두 벡터의 코사인 유사도는 $S_{\cos}(x, y)$으로 나타내어진다.

코사인 유사도 알고리즘은 추천시스템에서의 사용자 특성, 성향으로 구성된 이미지 검색, 사회연계서비스 데이터에서 이용자 클러스터링 등 도메인 항목에서 주데이터가 벡터로 모델링되어 데이터 간 유사도 측정의 특징 및 경향 혹은 방향성이 중요시되는 분야에서 폭넓게 이용되고 있다[6-7].

2.3.2 K-means 군집분석

K-means 군집분석은 비계층적 군집분석 방법을 사용하는 알고리즘으로써 계산 부분이 적기 때문에 다수의 데이터를 빠르게 처리할 수 있는 장점이 있다. K-means 군집 분석의 알고리즘 과정은 K개의 군집 중심점을 분석자가 설정하여 랜덤하게 선정한다. 다음에 관측한 데이터를 군집 중심에 가장 가까이 할당한 후 군집 중심을 새롭게 계산한다. 끝으로 기존에 설정된 중심과 새로이 계측한 군집 중심이 같아질 때까지 계속하여 반복한다. 다음으로 가장 근접한 군집 중심에 관측 데이터를 할당한 후에 군집 중심을 새로이 계산한다. 마지막으로 새로 계산한 군집 중심과 기존의 중심이 같아질 때까지 반복하여 수행 한다[8-10].

2.3.3 계층적 군집분석(Hierarchical Clustering)

계층적 군집분석은 데이터들을 특정 알고리즘에 연결하여 계층적인 방법으로 클러스터를 구성해 나가는 방식이다. K-means 클러스터링 방법과는 달리 장점으로는 최초에 클러스터의 개수를 가정할 필요가 없다.

군집 사이의 거리를 정의하는 방법으로는 단순, 완전, 평균연결, centroid, ward의 방법이 있다[11,12].

3. 반려동물 개인화 추천 시스템

반려동물용 사물인터넷기기, SNS, 쇼핑몰에서 추출되는 방대한 데이터를 바탕으로 수학적 모델 기반의 유사성을 분석하여 개인 맞춤형 서비스를 수행하는 플랫폼을 개발하였다. 반려동물 개인화서비스 통합 플랫폼의 구성은 다음 Fig. 1과 같이 데이터 수집 및 분석, 개인화서비스, 데이터 공유서비스, 데이터 시각화 서비스 시스템으로 구성된다.

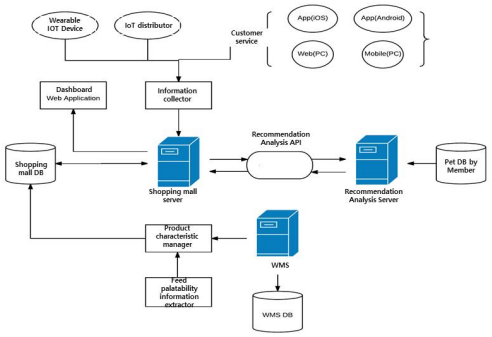

Fig. 1. Companion diagram of pet personalization service integration platform

첫째, 고객 반려동물 정보 수집기 개발로 플랫폼(PC웹, 모바일웹, Andorid, iOS)내에서 회원 가입 시, 종 품종 나이 무게 알러지 등 반려동물 관련 정보를 입력하게 한다. 둘째, 고객 행동 데이터 수집기 개발로 서비스 플랫폼 내에서 반려동물 보호자를 대상으로 어떤 행동을 하는지 분석을 위하여, 상품 클릭정보, 카테고리 클릭정보, 검색정보 등을 수집한다. 플랫폼 내에서 발생하는 고객 행동데이터는 주로 고객의 관심에 관한 것이다. 셋째, 고객 성향 추천 분석기 개발로 고객은 반려동물 쇼핑몰 플랫폼에서 상품을 고를 때, 주목적은 “반려동물을 위한 상품을 찾는 것”인 것을 설문성 이벤트 통계결과로 확인하였고, 추천분석을 위한 알고리즘은, 현재 널리 쓰이고 있는 CF(Collaborative Filtering)을 채택하였으며, 현재 데이터 수집 및 사용자 피드백 상황을 고려할 때 명시적 피드백수집이 원활하지 않으므로, 암시적 피드백을 이용하였다[13].

4. 제안 유사성 측정 방법

4.1 사료 데이터 분석

국내외 유통되고 있는 반려동물 사료의 등록성분표를 분석해 보면 크게 조지방, 조섬유, 조단백질, 조회분, 인, 수분, 칼슘등 7대 주요영양소와 비타민, 오메가-3 및 오메가-6, DHA, EPA 등의 기타 영양소로 표기되어 있다. 7대 주요영양소의 경우, 국내의 법률 「농림축산식품부 고시, 사료 등의 기준 및 규격 제9조 제1항」 에 의해 식품의약품안전처에 사료에 함유된 성분량을 정확히 등록해야 하므로 성분표에는 7대 주요영양소가 정확히 기재되어 있다. 하지만 기타 영양소의 경우 정확한 첨부량의 등록의무가 없으므로 첨부 여부만 간단히 표기되어 있다.

다음 Table 1은 사료 데이터를 분석한 예이다.

Table 1. Feed data analysis

ID	Feed name	Crude -protein	Crude -fat	Crude -fiber	crude -ash	Phosphorus	calcium	moisture	vitamin C
1	feed A	0.23	0.14	0.06	0.08	0	0	0.1	inclusion
2	feed B	0.23	0.14	0.06	0.08	0	0	0.1	inclusion
3	feed C	0.21	0.14	0.025	0.065	0.0076	0.014	0	Without
4	feed D	0.21	0.14	0.025	0.065	0.0076	0.014	0	Without
5	feed E	0.25	0.14	0.035	0.095	0.0098	0.0164	0	Without
6	feed F	0.27	0.15	0.05	0.09	0.006	0.01	0.12	Without
7	feed G	0.28	0.18	0.05	0.1	0.012	0.01	0	Without
8	feed H	0.27	0.15	0.05	0.1	0.006	0.009	0.12	Without
9	feed I	0.28	0.18	0.05	0.1	0.006	0.01	0.12	Without
10	feed J	0.28	0.18	0.05	0.1	0.006	0.009	0.12	Without

4.2 유사성 측정 방법

본 논문에서는 3.1절에서 분석한 사료 데이터의 특성을 기반으로 이에 적합한 유사성 측정 방법에 대해 제안한다.

$$\text{유사성} = \frac{(\cos(\theta) + Jaccard\,Coefficient)}{The\ number\ of\ Feature\ Class}$$

$$= \frac{\left(\frac{\sum_{i=1}^{n} \omega_i (A_i \times B_i)}{\sqrt{\sum_{i=0}^{n} (A_i)^2} \times \sqrt{\sum_{i=0}^{n} (B_i)^2}} + \frac{r}{p+q+r} \right)}{The\ number\ of\ Feature\ Class}$$

여기서, 비교 대상 사료 Ai 와 Bi에 해당되며, i는 사료 A, B의 특징 번호이다. r은 Ai가 true이고 Bi가 true일 경우의 값이며, p는 Ai가 false, Bi가 true일 경우 값이다. q는 Ai가 true, Bi가 false일 경우의 값에 해당된다. 또한 wi는 가중치 행렬의 값이 된다[14].

5. 실험 결과

5.1 데이터 준비

본 논문에서 사용한 데이터는 국내외에서 유통되고 있는 260여개의 강아지 및 고양이 사료 제품 중 100개의 데이터를 랜덤하게 샘플로 추출하여 실험하였다. 특성(feature)은 6대 주요영양소와 수분 및 기타 영양소 정보를 활용하였다. 다음 Fig. 2와 같이 6대 주요영양소는 조단백, 조지방, 조섬유, 조회분, 인, 칼슘과 수분의 비율을 활용하였으며 기타 영양소로는 비타민A, 비타민C, 비타민D3, 비타민E, 오메가-3, 오메가-6, DHA, EPA, 유산균 등 21개의 영양소 포함 여부를 사용하였다.

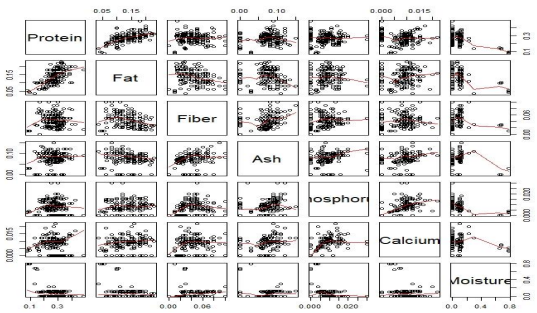

Fig. 2. Six Nutrients and Water Correlation Analysis

5.2 정규화

평균적으로 데이터 값을 0으로 놓는 정규화는 초기의 데이터 값의 분포를 정규분포로 가정 한다. 이때의 기본 생각은 평균값은 0으로 하고, 평균값에서 멀어질수록 값을 증가 시키자는 것이다. 분산으로 나누는 의미는, 값의 분포가 차이나지 않는 상황에서 1 값의 차이와, 값의 분포가 아주 큰 경우 1 차이나는 것은 분명히 다른 경우이므로 이를 분산으로 나누게 하여 원래 분포가넓게 퍼지는 효과를 줄이자는 것이다. 다음 Fig. 3는 범위를 벗어나는 데이터를 정제하는 것으로 오류값 임계치를 구하여 오류를 제거하기 위한 그림이다[15].

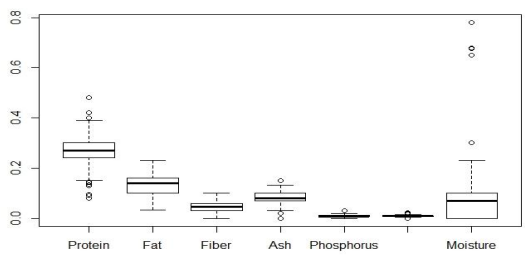

Fig. 3. Eliminate Errors by Obtaining Error Value Thresholds

5.3 군집분석 결과

5.3.1 계층적 클러스터링

다음 그림과 같이 군집분석을 위해 사료의 특징 간 유사도를 측정하였다. 유사도 측정 알고리즘은 Euclidean distance, Jaccard distance, Cosine distance, Manhattan Distance를 이용하였다[14,15]. 계층적 군집분석을 시행한 결과로 도출되는 Dendrogram을 그리면 다음 Fig. 4와 같다.

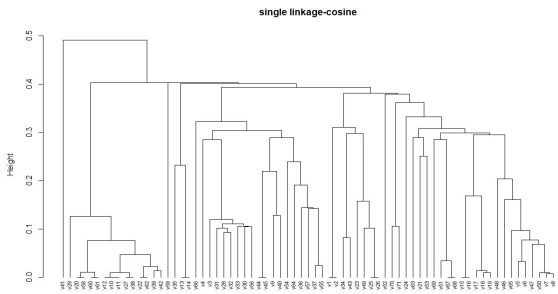

Fig. 4. Similarity Measurement Algorithm(Cosine distance)

5.3.2 제안한 유사성 측정 방법 적용 결과

국내 유통 중인 260여 개의 반려동물 사료를 대상으로 6대 중요 영양소와 수분을 분석하여 군집화 하였다. 먼저 데이터의 전처리 과정으로 특성 간 데이터의 범위를 일치시키기 위해 데이터를 정규화 하였고, 범위를 벗어난 데이터를 정제하였다. 다음으로 군집분석을 위해 사료의 특징 간 유사도를 측정하였다. 분석한 사료 데이터의 특성을 기반으로 이에 적합한 유사성 측정 방법을 4.2에서 제안했다.

6. 결론

최근 반려동물을 대상으로 한 맞춤형 헬스케어시스템 및 추천시스템에 관한 관심이 높아지고 있다. 따라서 본 논문에서는 이러한 시스템에 활용할 수 있도록 국내 유통 중인 260여 개의 반려동물 사료를 대상으로 7대 중요 영양소를 분석하여 군집화 하였다. 먼저 데이터의 전처리 과정으로 특성 간 데이터의 범위를 일치시키기 위해 데이터를 정규화 하였고, 범위를 벗어난 데이터를 정제하였다. 다음으로 군집분석을 위해 사료의 특징 간 유사도를 측정하였다. 유사도 측정 알고리즘은 Euclidean distance, Jaccard distance, Cosine distance, Manhattan Distance를 이용하였다. 군집분석은 계층적 군집분석과 K-means 군집분석을 이용하여 분석하였다. 위에서 분석한 사료 데이터의 특성을 기반으로 이에 적합한 유사성 측정 방법을 제안했다. 특히 향 후 연구할 부분으로는 추천시스템의 군집 정확도를 높이기 위해서 기존 주요 영양소 정보와 더불어 원재료 데이터를 추가적으로 분석할 필요성이 있다. 또한, 사용자의 구매정보와 반려동물의 사료 하루 섭취량을 병합하여 분석한다면 반려동물의 헬스케어스 시스템, 추천시스템 등에 고도화된 서비스가 가능할 것이다.

REFERENCES

[1] B. Sarwar, G. Karypis, J. Konstan & J. Riedl. (2000). Analysis of Recommendation Algorithms for ECommerce. *Proc. of ACM EC '00 conference*, 158-167.

[2] G. Adomavicius & A. Tuzhilin. (2005). Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. on Knowledge and Data Engineering, 17(6),* 734-749.

[3] J. S. Kim. (2016). Subway Congestion Prediction and Recommendation System using Big Data Analysis. *Journal of digital Convergence, 14(11),* 289-295. DOI : 10.14400/JDC.2016.14.11.289

[4] Dae-Sung Seo. (2019). A Study on the Autonomous Decision Right of Emotional AI based on Analysis of 4th Wave Technology Availability in the Hyper-Linkage, *Journal of Convergence for Information Technology, 9(8),* 9-19. DOI : 10.22156/CS4SMB.2019.9.8.009

[5] J. Horey, E. Begoli, R. Gunasekaran, S. Lim & J. Nutaro. (2012). Big Data Platforms as a Service: Challenges and Approach, *USENIX Workshop on Hot Topics in Cloud Computing (HotCloud).*

[6] B. Cabral, R. D. Beltro & M. G. Manzato. (2014). Combining Multiple Metadata Types in Movies Recommendation Using Ensemble Algorithms, *In Proceedings of the 20th Brazilian Symposium on Multimedia and the Web* (pp. 231-238).

[7] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen & X. Sun. (2010). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems 50*, 559-569.

[8] The R. C. Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing. The R Foundation [Online]. https://www.R-project.org/

[9] J. T. Oh & S. Y. Lee. (2019). A Music Recommendation System based on Context-awareness using Association Rules. *Journal of digital Convergence, 17(9),* 375-381. DOI : 10.14400/JDC.2019.17.9.375

[10] J. L. Herlocker, J. A. Konstan, L. G. Terveen & J. Riedl. (2004). Evaluating Collaborative Filtering Recommender Systems, *ACM Transactions on Information Systems, 22(1),* 5~53.

[11] I. H. Witten, E. Frank & M. A. Hall. (2011). *Data Mining: Practical Machine Learning Tools and Techniques,* Amsterdam : Elsevier.

[12] M. P. Callao & I. Ruisánchez. (2018) An overview of multivariate qualitative methods for food fraud detection. *Food Control, 86,* 83-293.

[13] S. H. Namn & K. S. Noh. (2015). A Study on the Effective Approaches to Big Data Planning, *Journal of*

digital Convergence, 13(1), 227-235,

[14] K. S. Noh. (2015). Convergence Analysis of Recognition and Influence on Bigdata in the e-Learning Field, *Journal of digital Convergence, 13(10),* 51-58.

DOI : 10.14400/JDC.2015.13.10.51

[15] H. J. Jung. (2015). The Analysis of Data on the basis of Software Test Data. *Journal of digital Convergence, 13(10),* 1-7.

김 삼 택(Sam-Taek Kim) 정회원

· 2005년 2월 : 중앙대학교 중앙대학원 컴퓨터공학과 (공학 박사)
· 1990년 5월 ~ 1995년 2월 : LG연구소 전임연구원
· 1995년 3월 ~ 현재 : 우송대학교 IT융합학부 교수
·관심분야 : 무선/유선 네트워킹, VoIP, 모바일, IoT, Big Data, USN
· E-Mail : stkim@wsu.ac.kr