학습 관리 시스템 챗봇 데이터의 군집화 전략: 실루엣 스코어를 중심으로

이나영○ 이동헌

고려대학교 수학과

leenysky@gmail.com, holy@korea.ac.kr

Clustering Strategies for Learning Management System Chatbot Data: Based on Silhouette Scores

Nayoung Lee Donghun Lee

Department of Mathematics, Korea University

요 약

온라인 학습 시스템 내에서 발생하는 챗봇 대화 로그는 양이 방대하고 복잡하기 때문에 효율적으로 관리하고 분석하는 과정이 중요하다. 본 연구는 다양한 클러스터 수와 계층적 군집화 레벨을 조정하며 실루엣 스코어를 기반으로 군집화 전략의 효율성을 평가하였다. 또한, 대화 로그의 특성을 고려한 데이터 전처리 방법을 제시하고, 계층적 접근 방식이 클러스터링 품질에 미치는 영향을 분석해 방대한 양의 데이터를 분석하는 과정에서 효과적인 방법을 보여주었다. 단일 레벨 군집화가 다중 레벨 군집화보다 일반적으로 더 높은 실루엣 점수를 보였으며, 이는 단일 레벨에서의 군집 내 데이터 포인트들의 밀집도가 더 높다는 것을 의미한다. 이러한 결과는 교육 기술과 챗봇 시스템의 설계에 중요한 개선점을 시사한다.

1. 서 론

본 연구에서는 학습 관리 시스템(Learning Management System, LMS)에서 수집된 데이터의 챗봇 대화 로그를 분석한다. LMS 는 교육과 훈련 과정의 효율적인 전달과 더 쉽게 관리를 가능하게 하는 핵심 기술로, 전 세계 교육 산업에 혁명을 일으켰다. 챗봇은 사용자의 질문에 대응하는 소프트웨어 프로그램으로, 텍스트나 음성 기반의 상호작용을 지원한다.

챗봇의 사용은 이러한 시스템에서 더 체계적이고 사용자 친화적인 환경을 제공하여, 적절한 시간에 필요한 정보를 효과적으로 탐색할 수 있게 한다. 인공지능 기술의 적용은 이 과정에서 핵심적 역할을 하며, 교육 기술의 발전을 통해 교육의 질을 향상시키고 학습 경험을 개인화하는 데 기여한다[1,2].

최근 챗봇 연구 동향에 따르면, 이러한 챗봇의 효과적인 활용을 더욱 발전시키고자, 인공지능을 활용한 데이터 기반 분석을 통해 사용자에게 더 적합한 챗봇 대화 추천 시스템을 개발하는 것이 최근 활발히 연구되고 있다[3,4].

이러한 연구 동향에 발맞추어, 본 연구는 LMS 시스템 내에서의 챗봇 데이터 분석을 통해 최적의 군집화 전략을 탐색하고자 한다.

데이터를 기반으로 운영되는 여러 챗봇 데이터는 챗봇 시스템 개편을 위하여 다양하게 활용되고 있다. 예를 들어 텍스트 마이닝 기법을 활용하여 대화 내용에서 자주 등장하는 키워드와 이와 관련된 용어를 추출하여 주요 관심사 항목을 식별해 이를 기반으로 사용자를 비슷한 관심사 또는 질문 유형에 따라 여러 그룹으로 클러스터링하였다[5]. 이와 유사하게 사용자 행동 데이터를 수집해 클러스터링을 진행하여 사용자 세분화와 맞춤형 챗봇 서비스 제공에 기여한 연구 동향도 있다[6]. 이렇듯 챗봇 대화 로그에 클러스터링 기법을 적용하는 것은 사용자의 요구와 상호작용 패턴을 깊이 이해할 수 있는 효과적인 방법으로, 챗봇 서비스의 개선과 사용자 만족도 향상에 크게 기여한다[7].

실루엣 스코어가 높다는 것은 클러스터 내의 데이터가 서로 밀접하게 모여 있고, 다른 클러스터와는 잘 구분되어 있다는 것을 나타낸다. 따라서 본 연구에서는 높은 군집도를 달성하고자 한다. 이러한 맥락에서, 다양한 클러스터 수와 계층적 군집화 레벨의 조정을 통해 실루엣 스코어가 어떻게 변화하는지를 분석한다. 이를 통해 대화 로그의 효과적으로 분류하는 방법을 찾아 사용자의 요구 및 상호작용 패턴의 깊은 이해를 가능하게 하는 군집화 전략을 도출하고자 한다.

2. 본 론

2.1. 분석 데이터

본 연구에서는 블랙보드 학습 시스템 내에서 사용자와 챗봇 간의 대화 로그를 분석하기 위해, 우선 세 가지 다른 기간 동안 대학교에서 수집된 대화 로그들을 통합하였다. 순서대로 2022 년 1 월 1 일부터 2022 년 6 월 30 일, 2022 년 7 월 1 일부터 2022 년 12 월 31 일, 2023 년 1 월 1 일부터 2023 년

6 월 30 일까지의 기간 동안 사용자 질문, 챗봇 응답, 사용자 반응 등을 포함하고 있다.

2.2. 전처리 과정

데이터 전처리 과정에서는 자연어 처리(NLP) 기술을 활용하여 텍스트 로그에서 불필요한 요소를 제거하기 위해 데이터의 텍스트 컬럼에서 정규 표현식(re)을 사용하여 특수 문자와 숫자를 제거하였다. 이후 무의미하게 반복되는 키워드가 포함된 문장과 빈 문자열을 포함하는 행을 필터링하여 제외시켰다. 챗봇 데이터 중 영어 텍스트에 대해서는 NLTK 라이브러리[8]를 사용하여 텍스트 로그를 소문자로 변환, 토큰화, 그리고 영어 불용어를 제거하는 과정을 거쳤다. 한국어로 된 데이터에 대해서는 한국어 형태소 분석기(Okt)[9]를 이용해 텍스트를 토큰화하고, 이 토큰들을 기반으로 TF-IDF 벡터화 방법을 적용하여 각 대화에서 자주 등장하지만 다른 대화에서는 드물게 나타나는 단어들에 높은 가중치를 부여함으로써 대화의 주요 특징을 수치화하였다.

2.3. 클러스터링

마지막 분석 단계에서는 MiniBatchKMeans[10]을 활용하여 각 대화 로그의 TF-IDF 벡터에 기반하여 계산된 유사도를 주요 키워드 사용빈도와 문맥상 유사성이 높은 대화 로그들을 그룹화했다. MiniBatchKMeans 알고리즘은 클러스터링 시 전체 데이터셋을 소규모 배치로 분할하여 처리함으로써 대규모 데이터셋에 대한 계산 비용을 절감하고 처리 속도를 향상시키는 장점이 있어 본 논문의 대규모 데이터셋을 분석하는 데에 매우 적합했다.

계층적 군집화를 통해 대화 로그를 여러 수준에 걸쳐 분류할 수 있다. 즉, 클러스터링 수준은 데이터를 얼마나 상세하게 나누어 분석할 것인지 결정하는 것으로, 수준이 높아질수록 데이터는 더 많은 소그룹으로 세분화된다.

2.4. 실루엣 점수

클러스터링 이후, 각 클러스터의 일관성과 분리도를 평가하기 위해 실루엣 스코어(Silhouette Score)[11]를 계산했다. 실루엣 스코어는 -1 부터 1 까지의 값을 가지며, 값이 높을수록 클러스터 내의 데이터가 잘 모여 있고, 다른 클러스터와는 잘 구분되어 있음을 의미한다.

실루엣 스코어 계산은 각 데이터 포인트에 대해 해당 포인트가 속한 클러스터 내의 다른 포인트들과의 평균 거리(응집도)와 가장 가까운 클러스터의 포인트들과의 평균 거리(분리도) 사이의 차이를 기반으로 한다.

실루엣 스코어를 기반으로 본 연구에서는 대화 로그 데이터가 어떤 상황에서 효과적으로 분류되었는지 비교하였다. 따라서 본 연구에서는 실루엣 스코어가 높을수록, 클러스터링이 효과적으로 된 것으로 본다.

2.5. 탐색적 데이터 분석

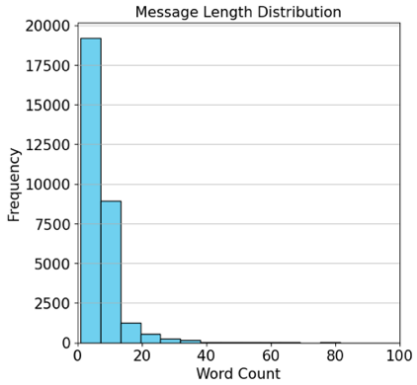

그림 1 챗봇 대화 길이 분포

챗봇 대화 로그의 각 메시지별 단어 수를 계산하고, 그 분포를 분석하여 대화 로그의 길이가 어떻게 분포하는지를 시각화하여, 사용자의 질문 복잡성과 상호작용의 깊이를 유추할 수 있었다. 대부분의 메시지 길이가 평균적으로 짧은 것으로 나타났기 때문에, 이는 사용자의 의도가 간결하게 표현되었음을 의미한다. 따라서, 각 대화 로그를 독립적인 데이터 단위로 취급하고, 이를 바탕으로 클러스터링 분석을 수행하기로 결정했다.

2.6. 실루엣 점수 분석

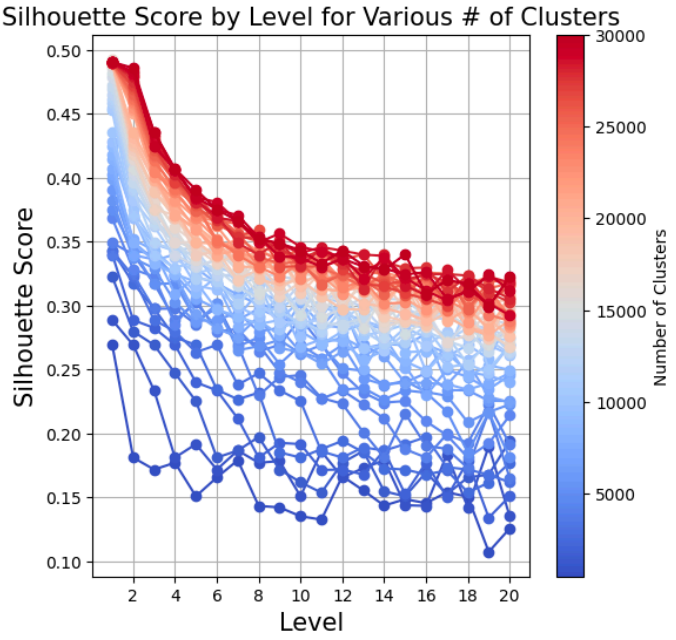

그림 2 다양한 클러스터 수에 대한 수준별 실루엣 점수의 변화

그림 2 는 500 부터 30,000 까지 500 간격으로 한 다양한 클러스터 수에 대하여 클러스터링의 수준을 1 부터 20 까지 조정하여 실루엣 스코어가 어떻게 변화하였는지를 나타낸다. 대체적으로 동일 클러스터 수에 대하여 수준이 깊어질수록 더

낮은 성능의 클러스터링 성능을 나타낸 것을 확인할 수 있다. 또한, 동일 수준에서는 더 많은 클러스터를 사용하면 실루엣 스코어가 전반적으로 향상됨을 확인하였다.

분석 결과, 레벨 1(단층 클러스터링)과 레벨 20(다층 클러스터링) 사이의 실루엣 스코어 차이의 평균이 -0.1824 로 나타났다. 또한, 60 개의 군집 각각의 20 개 실루엣 스코어에 대해 수행된 선형 회귀 분석에서 얻어진 모든 계수가 음수가 나왔다. 이는 계층적 군집화가 진행될수록 실루엣 스코어, 즉 클러스터 내 응집도가 저하되는 경향을 보여준다.

3. 결론 및 향후 연구

본 연구의 분석 결과는 계층적 군집화를 사용할 때 클러스터링의 깊이와 클러스터 수를 결정하는 데 있어 중요한 고려 사항을 제공한다. 특히, 실루엣 스코어가 클러스터링 품질의 지표로 사용될 때, 군집화의 깊이에 따른 응집도의 변화를 고려하는 것이 중요하다. 이는 교육 기술 분야에서 챗봇 시스템의 설계와 개선에 있어서, 데이터의 군집화 전략을 결정하는 데 있어 실질적인 가이드라인을 제공한다.

초기 단계에서는 가장 유사한 클러스터들이 먼저 합쳐지기 때문에, 초기에 형성된 클러스터들은 비교적 높은 응집력을 유지할 수 있다. 그러나 군집화의 깊이가 깊어짐에 따라, 상대적으로 유사성이 떨어지는 클러스터나 데이터 포인트들이 합쳐지면서, 클러스터 내의 데이터 포인트들 사이의 거리가 증가한다. 이는 응집도의 저하로 이어지며, 실루엣 스코어의 감소를 초래한다.

본 연구에서 사용된 방법론을 확장하여, 대화 로그 데이터 분류에 바텀업(bottom-up) 클러스터링과 탑다운(top-down) 클러스터링 방식을 비교 분석하는 연구로 이어질 수 있다. 바텀업 클러스터링은 더 세분화된 클러스터부터 시작하여 점차 통합하는 방식이며, 탑다운 클러스터링은 큰 클러스터에서 시작하여 더 작은 클러스터로 세분화하는 방식이다. 본 연구에서 제안된 효과적 단층 클러스터링 기법을 바탕으로, 이들을 점진적으로 통합하여 챗봇 대화 로그 분석에 활용할 수 있다.

또한, 본 연구의 결과를 기반으로 클러스터링 결과의 적용 가능성을 탐색할 수 있다. 예를 들어, 클러스터링을 통해 식별된 사용자 그룹별 맞춤형 대화 시나리오를 개발하고, 이를 챗봇에 적용하여 사용자 개인별 경험을 개선해 사용자의 만족도를 높이는 연구를 진행할 수 있다.

참 고 문 헌

[1] Shukla, Vinod Kumar, and Amit Verma. "Enhancing LMS experience through AIML base and retrieval base chatbot using R language," in 2019 International Conference on Automation, Computational and Technology Management (ICACTM), pp. 561-567, 2019.

[2] Murad, Dina Fitria, Adhi Gustian Iskandar, Erick Fernando, Tica Shinta Octavia, and Deryan Everestha Maured. "Towards smart LMS to improve learning outcomes students using LenoBot with natural language processing," in 2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE), pp. 1-6, 2019.

[3] Rooein, Donya. "Data-driven EDU chatbots," in Companion proceedings of the 2019 World Wide Web Conference, pp. 46-49, 2019.

[4] Akhtar, Mubashra, Julia Neidhardt, and Hannes Werthner. "The potential of chatbots: analysis of chatbot conversations," in 2019 IEEE 21st Conference on Business Informatics (CBI), vol. 1, pp. 397-404, 2019.

[5] Akhtar, Mubashra, Julia Neidhardt, and Hannes Werthner. "The potential of chatbots: analysis of chatbot conversations," in 2019 IEEE 21st Conference on Business Informatics (CBI), vol. 1, pp. 397-404, 2019.

[6] Hwang, S., Kim, B., & Lee, K., "A data-driven design framework for customer service chatbot," Design, User Experience, and Usability, Springer, 2019.

[7] Xu, Luxun, Vagelis Hristidis, and Nhat XT Le. "Clustering-based summarization of transactional chatbot logs," in 2019 IEEE International Conference on Humanized Computing and Communication (HCC), pp. 60-67, 2019.

[8] Hardeniya, Nitin, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, and Iti Mathur. *Natural Language Processing: Python and NLTK*. Packt Publishing Ltd, 2016.

[9] Moon, Sangwhan, Won Ik Cho, Hye Joo Han, Naoaki Okazaki, and Nam Soo Kim. "OpenKorPOS: Democratizing Korean Tokenization with Voting-Based Open Corpus Annotation," in Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 4975-4983, 2022.

[10] Hicks, S. C., Liu, R., Ni, Y., Purdom, E., & Risso, D., "mbkmeans: Fast clustering for single cell data using mini-batch k-means," Public Library of Science, 2021.

[11] Shahapure, Ketan Rajshekhar, and Charles Nicholas. "Cluster quality analysis using silhouette score," in 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pp. 747-748, 2020.