데이터 수집과 정제의 중요성 – 멘토 매칭 및 AI 튜터링 프로젝트

최원교 [1](주저자), 최은준 [2], 양혜원 [3], 김세령 [4]

[1] 동덕여자대학교 데이터사이언스전공 학부생
[2] 동덕여자대학교 데이터사이언스전공 학부생
[3] 동덕여자대학교 데이터사이언스전공 학부생
[4] 동덕여자대학교 컴퓨터학과 학부생

20221645@dongduk.ac.kr, 20221646@dongduk.ac.kr, 20221631@dongduk.ac.kr, 20210755@dongduk.ac.kr

DCR: Importance of Data Collection and Refinement for Mentor Matching and AI Tutoring Project

Won-Kyo Choi[1](lead author), Eun-Jun Choi[2], Hye-Won Yang[3], Se-Ryeong Kim[4]
[1]Dept. of Data Science, Dong-Duk Women's University
[2]Dept. of Data Science, Dong-Duk Women's University
[3]Dept. of Data Science, Dong-Duk Women's University
[4]Dept. of Computer, Dong-Duk Women's University

요 약

본 논문은 'AI 튜터와 함께하는 매칭 플랫폼' 개발 과정에서의 데이터 수집 및 정제에 대해 다룬다. 이 플랫폼은 사용자에게 개인화된 멘토 매칭 및 강의 추천 서비스를 제공하며, 이를 위해 웹 크롤링을 통해 데이터를 수집하고, 그 데이터를 정제하는 과정을 거쳤다. 특히, 요리 레시피 데이터를 기반으로 한 취미 레벨 테스트 기능이 포함되어 있으며, 정제된 데이터를 통해 딥러닝 기반의 추천 알고리즘과 AI 튜터링 시스템을 구축했다. 본 연구는 이러한 시스템이 사용자 맞춤형 학습 경험을 제공하는 데 어떻게 기여하는지 논의한다.

1. 서론

AI 기술의 발전과 함께 개인화 추천 시스템의 중요성은 커지고 있으며, 맞춤형 서비스는 사용자 경험을 향상시키고 서비스 제공자의 성과를 극대화한다. 특히 멘토링 및 학습 분야에서는 맞춤형 추천 알고리즘이 학습 동기와 성과를 높이는 중요한 도구로 자리 잡고 있다. 본 연구에서는 MZ 세대의 다양한 학습 및 취미 요구를 충족시키기 위해 AI 기반 멘토 매칭 및 튜터링 서비스를 제공하는 플랫폼을 개발하였다.

본 논문은 개인화 추천 시스템 구현 과정에서 데이터를 수집하고 정제하는 방법론을 다루며, 웹 크롤링을 통해 데이터를 취합하고 불완전한 데이터를 정제하는 과정을 설명한다. 이를 마탕으로 딥러닝 기반 추천 시스템과 AI 튜터링 시스템을 구축하여 사용자 맞춤형 학습 경험을 제공하는 방법을 논의한다.

2. 본론

2.1 데이터 수집

본 프로젝트에서 데이터 수집은 웹 크롤링 기술을 활용하여 수행되었다. 웹 크롤링은 인터넷 상의 다양한 출처로부터 정보를 자동으로 수집하는 기법으로, 멘토 매칭, 강의 추천, 그리고 레벨 테스트 기능을 위한 기초 데이터를 확보하는 데 사용되었다.

2.1.1 AI 튜터 및 레벨테스트

웹 크롤링을 통해 요리, 음악, 미술 등 다양한 취미 활동에 대한 데이터를 수집하였다. 예를 들어 요리 레시피의 경우, 각 레시피의 단계, 재료, 도구 정보를 크롤링하여 이를 정형화된 형식으로 변환하였다.

2.2 데이터 정제

웹 크롤링을 통해 수집된 데이터는 다양한 문제를 안고 있기 때문에, 데이터 정제 과정을 통해 사용 가능한 형태로 변환하는 작업이 필요하다. 본 연구에서는 다음과 같은 형태로 변환하는 작업이 필요하다.

2.2.1 형식 정리 및 오류 검증

모든 수집된 데이터는 사전 정의된 형식에 맞춰 정리되었다. 예를 들어 레벨테스트-요리 부문에서는, 레시피가 단계별로 정렬되고, 재료와 도구 또한 명확하게 구분되었다. 레시피의 단계나 재료가 누락된 경우, 또는 비정상적으로 많은 단계나 재료가 있는 경우에는 오류를 체크하고 수정하였다.

2.2.2 중복 데이터 제거

웹 크롤링 과정에서 동일한 데이터가 여러 번 수집될 가능성이 있다. 이러한 중복 데이터를 식별하고 통합하여 단일 항목으로 정리했으며, 이는 데이터 분석과 추천 시스템의 효율성을 높이는 데 기여하였다.

2.2.3 난이도 평가 알고리즘 적용

수집 후 정제된 요리 레시피는 각 레시피의 단계 수, 재료 수, 도구 수를 기반으로 난이도를 평가하는 알고리즘이 적용되었다. 이 알고리즘은 레시피의 단계를 기반으로 점수를 부여하며, 단계 개수에 따른 난이도 분류(하, 중, 상)를 자동으로 수행한다.

2.3 하이브리드 추천 알고리즘

본 프로젝트에서 사용된 하이브리드 추천 시스템은 콘텐츠 기반 필터링과 협업 필터링의 장점을 결합하여 개인화된 추천을 제공하였다.

2.3.1 사용자-강의 매트릭스 생성

추천 시스템을 위해, 사용자-강의 행렬을 구성하였다. 이는 사용자가 특정 강의와 상호작용한 데이터를 기반으로 구축된 행렬로, 각 행은 사용자, 각 열은 강의를 나타내며, 사용자의 상호작용 빈도(예: 클릭 수)를 포함한다. 해당 행렬은 pivot_table 을 사용하여 생성되었으며, 상의를 클릭하지 않은 경우에는 0 으로 채워졌다.

2.3.2 유사도 계산

사용자 간의 유사도를 계산하기 위해 코사인 유사도(Cosine Similarity)를 사용하였다. 코사인 유사도는 두 벡터 사이의 각도를 측정하는 방식으로, 두 벡터의 방향이 비슷할수록 유사도가 높다. 사용자-강의 매트릭스를 기반으로, 사용자가 시청한 강의의 패턴을 벡터화한 후, 각 사용자 간 유사도를 측정하였다.

$$\text{Cosine Similarity}(A,B) = \frac{A \cdot B}{|A||B|}$$

여기서 A 와 B 는 두 사용자의 강의 시청 벡터이다. 계산된 유사도 행렬은 각 사용자 간 유사도를 정량화한 데이터프레임 형태로 변환되었다.

2.3.3 유사 사용자 추출

유사 사용자 추출 과정에서는 특정 사용자의 유사도를 기준으로 상위 n 명의 유사한 사용자를 찾는다. 이를 통해, 유사한 사용자가 시청한 강의 정보를 바탕으로 새로운 강의를 추천할 수 있다. 사용자가 시청하지 않은 강의 중 유사한 사용자들이 선호한 강의를 추천하며, 만약 해당 사용자의 시청 데이터가 존재하지 않을 경우 가장 많이 시청된 강의를 추천하게 된다.

2.3.4 추천 강의 제공

유사한 사용자가 시청한 강의 중, 사용자가 아직 시청하지 않은 강의를 추천하는 방식으로 알고리즘이 동작한다. 강의 추천 시, 유사 사용자들이 시청한 강의 목록을 분석하여 우선 순위를 매긴 후 추천 목록을 생성한다.

2.3.5 클러스터링 기반 추천

또한, 본 연구에서는 클러스터링 기법을 활용하여 사용자의 학습 패턴을 분석하고, 유사한 그룹을 형성한 후 추천을 제공하였다. K-Means 클러스터링을 통해 유사한 취향을 가진 사용자들을 군집화하였으며, 각 클러스터 내에서 가장 인기 있는 강의를 추천하는 방식으로 사용자에게 맞춤형 콘텐츠를 제공하였다.

3. 결론

본 프로젝트에서 데이터 수집 및 정제 과정은 ‘AI 튜터와 함께하는 매칭 플랫폼’ 프로젝트의 성공적인 구현에 필수적인 요소였다. 데이터 정제를 통해 시스템의 성능을 높이고, 추천 정확도를 극대화할 수 있었다. 특히 하이브리드 추천 시스템과 AI 튜터링을 결합하여 개인화된 멘토링과 학습 경로 설정을 가능하게 하였으며, 실시간 피드백을 통해 사용자 맞춤형 학습 경험을 제공하는 데 성공하였다.

"본 논문은 과학기술정보통신부 대학디지털교육역량강화사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물 입니다"

참고문헌

[1] Burke, R., "Hybrid Recommender Systems: Survey and Experiments," *User Modeling and User-Adapted Interaction*, Vol. 12, No. 4, pp. 331–370, 2002.

[2] Koren, Y., Bell, R., & Volinsky, C., "Matrix Factorization Techniques for Recommender Systems," *Computer*, Vol. 42, No. 8, pp. 30-37, 2009.

[3] Abadi, M., et al., "TensorFlow: A System for Large-Scale Machine Learning," *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 265-283, 2016.

[4] Van der Aalst, W., *Process Mining: Data Science in Action*, Springer, 2016.

[5] Aggarwal, C. C., *Recommender Systems: The Textbook*, Springer, 2016.

[6] Gama, J., *Knowledge Discovery from Data Streams*, CRC Press, 2010.