효율적인 추천 시스템을 위한 학습 컨텐츠의 상대적 특성 업데이트 방법 연구

운봉영, 아가르왈 판카즈
㈜태그하이브

wbyoung@tag-hive.com, pankaj@tag-hive.com

A Method to Update the Relative Characteristics of Learning Contents For Efficient Recommendation System

Woon BongYoung, Agarwal Pankaj
TagHive, Inc.

요 약

본 논문은 추천 시스템의 핵심 요소인 컨텐츠 관리 방법에 대하여 연구한다. 학습 컨텐츠의 경우 난이도, 스킬, 챕터 등의 특성을 부여 받는데 사용자 군집의 실력에 따라 상대적인 특성이 발생하게 된다. 이를 해결하기 위하여 사용자 군집의 문제 풀이 데이터에 기반하여 사전에 부여된 난이도를 업데이트하고, 로지스틱 회귀모델을 활용하여 문제 정답 확률 예측을 위한 학습의 결과로 생성된 스킬 분포 값을 통해 스킬 적합도를 판단하는 방법을 제안한다.

Ⅰ. 서론

전 세계적으로 코로나 시대가 도래하면서 다양한 사회의 모습이 변화하고 있다. 교육 시장의 경우 대면 학습이 어려워지면서 온라인 교육 시장이 급격히 성장하였고 이와 더불어 에듀테크(Edu-Tech)에 관심도 높아지고 있다. 이전의 온라인 교육(이하 비대면 학습)의 경우 소비자가 사전에 만들어진 컨텐츠를 온라인으로 시간의 자유롭게 소비할 수 있다는 장점이 주요했지만 최근에는 각 사용자의 실력 및 특성 등을 정확하게 분석하고 이를 기반으로 적절한 컨텐츠 및 학습 방향을 추천하며 학습의 효율성을 극대화하는 다양한 개인화 추천 시스템이 개발되고 있다[1].

개인화 추천 시스템의 성능을 높이기 위해서는 다양한 유저에 대한 정확한 특성 분석과 양질의 많은 컨텐츠가 요구된다. 학습 컨텐츠의 경우 난이도, 스킬 등의 특성들이 정확하게 부여되어 있어야 하지만 컨텐츠 개발자의 주관 또는 컨텐츠 사용 분야, 시기에 따라 상대적인 차이가 발생한다. 예를 들어 현재 가장 높은 난이도를 부여 받은 컨텐츠는 사용자 군집의 기본 실력 수준에 따라 상대적으로 낮은 레벨로 분류 될 수 있고, 시간이 지나면서 전체적인 교육 수준의 향상으로 인하여 낮은 레벨로 분류 될 수 있다. 또한, 새로운 교육 정책에 따라 챕터 구성이 달라지고 스킬 이름이 달라지는 경우 기존 컨텐츠에 부여된 특성을 수정해야만 한다. 이처럼 추천 시스템을 높은 성능 확보를 위해서는 환경에 따라 발생하는 상대적인 특성을 반영하는 것이 중요하다.

이에 따라 본 논문에서는 사용자 데이터 통계를 활용하여 사용자 군집 실력에 따라 난이도를 업데이트 하고 스킬의 적합도를 판단하여 컨텐츠의 상대적 특성을 효율적으로 반영할 수 있는 방법을 제안한다.

Ⅱ. 본론

본 논문에서는 데이터에 기반하여 사용자 군집의 실력의 변화에 따라 컨텐츠 난이도를 업데이트하고 정답 확률 예측 모델을 활용하여 컨텐츠의 특성 적합도를 판단하는 방법을 제안한다.

1) 난이도 업데이트 시스템

난이도 업데이트는 Fig 1 의 순서로 이루어진다. 먼저 현재 사용자 군집의 문제 풀이 데이터를 수집하고 각 문제의 정답자, 오답자의 평균 실력을 계산한다. 사용자의 실력은 문제에 대한 지식 수준 정도를 구분하고 문제의 정답 유무에 따라 레벨을 부여한다[2]. 이후 새로운 난이도는 식 1 과 같이 생성한다. 계산식에서 해당 문제의 정답 확률과 오답자의 평균 실력을 곱하는 이유는 정답률이 높은 경우 현재 사용자 군집에게는 해당 문제가 쉬운 문제로 받아들여지기 때문에 난이도를 낮춰야 하고, 반대로 정답률이 낮은 경우 현재 사용자 군집에게 해당 문제가 어렵게 받아들여져 난이도를 높이는 방향으로 업데이트 되기 때문이다.

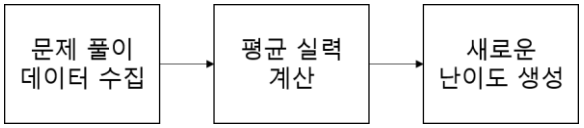

Fig 1. 난이도 업데이트 시스템 블록도

$$\Delta = (\beta * \mu) + ((1-\beta) * \nu) \quad \cdots\cdots\cdots\cdots\cdots\cdots \text{식 1}$$

Δ = 새로운 난이도
β = 해당 문제의 정답률
μ = 오답자 평균실력
ν = 정답자 평균 실력

최종 생성된 새로운 난이도는 현재 난이도와 λ 매개변수를 통해 평균을 계산한다. 단일 횟수로 난이도를 업데이트 하는 경우 현재 사용자 군집의 실력 분포에 의존도가 높아지는 문제가 발생하기 때문에 문제 풀이 데이터가 충분히 쌓였을 경우 난이도를 복수 횟수로 재 업데이트하여 정규분포를 고려한다.

2) 스킬 적합도 판단 시스템

본 논문에서의 스킬은 어떠한 과목 챕터의 하위 개념이며 문제를 풀기 위해 필요한 능력을 일컫는다. 스킬의 적합도를 판단은 Fig 2 와 같은 과정으로 진행한다. 현재 사용자 군집의 문제 풀이 데이터를 수집하고 정답 확률 예측 모델을 학습한다. 학습 결과로 얻은 데이터 기반의 스킬 분포 값과 현재 컨텐츠에 부여된 스킬을 비교하여 스킬의 적합도를 분석한다.

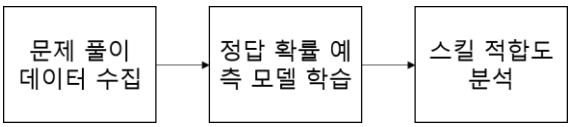

Fig 2. 스킬 적합도 판단 시스템 블록도

정답 확률 예측 모델의 경우 특정 사용자가 아직 풀지 않은 어떠한 문제에 대하여 정답을 맞출 수 있는 확률을 계산하기 위하여 사용되었으며 대표적으로 아래와 같은 방법들이 있다.

① Collaborative filtering: 어떠한 사용자가 새로운 문제를 접할 때 기존 사용자 군집에서 현재 사용자와 비슷한 특성을 지닌 사용자의 사전 정보를 활용하여 예측
② Logistic regression: 사용자 군집의 정보를 학습하여 어떠한 문제를 맞출 확률에 대한 예측

본 연구에서 활용하는 정답 확률 예측 단계는 학습 결과로 얻은 스킬 분포 값을 활용하는 것이 주된 목적이다.

따라서 현재 사용자 군집의 문제 풀이 데이터를 기반으로 컨텐츠의 특성을 분석하기 위한 방법이 요구하고 사전 유사 군집의 정보 없이 독립적인 학습을 진행하기 때문에 로지스틱 회귀 모델을 이용하여 사용자의 문제 정답 확률을 예측하였으며, 현재 사용자 군집(U)의 스킬별 실력(S) 데이터(U-S matrix)와 각 문제(Q)의 스킬 분포(S') 초기 랜덤값(Q-S' matrix)을 입력 데이터로 사용하고 문제 풀이 데이터 Y-matrix 를 학습 목표 데이터로 활용한다. 식 2 에 따라 생성한 정답 확률 예측 값을 Y-matrix 와 비교하여 학습하게 되고 스킬 적합도를 판단하기 위해서 Q-S' matrix 를 학습 매개변수로 사용한다. U-S matrix 에는 각 사용자의 스킬의 현재 실력 값이 들어가고, Q-S' matrix 의 경우 각 문제에 대한 랜덤 초기 값을 갖는다. 또한 Y-matrix 의 경우 사용자가 해당문제를 맞춘 경우 1, 틀린경우 0 으로 값을 지정한다.

본 연구는 정답 확률 예측이 아닌 스킬 분포 값을 계산하는 것이 목적이기 때문에 Fig 3 과 같이 학습을 수행하고 나면 매개변수인 Q-S' matrix 는 각 문제에 대한 스킬 분포 값을 얻게 되고 기존의 문제에 부여된 스킬의 개수에 따라(본 연구팀 컨텐츠의 경우 보통 1~2 개) 각 문제의 가장 높은 peak 값 1 개 혹은 2 개 값을 갖는 스킬을 추출한다.

U-S matrix * Q-S' matrix = 정답 확률 예측……식 2

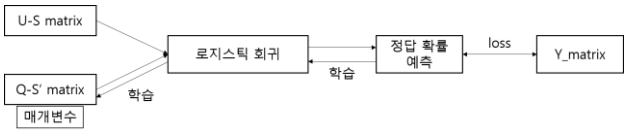

Fig 3. 정답 확률 예측 시스템 블록도

본 과정은 사용자 군집의 실력 값과 학습 결과로 생성된 문제의 스킬 분포 값을 통해 사용자의 문제 정답 확률을 정확하게 예측 한다면 매개변수인 Q-S' matrix 는 해당 문제에 대해 적절한 스킬 분포 값이 부여되었다고 판단할 수 있다. 따라서 학습된 Q-S' matrix 에서 peak 값을 갖는 스킬과 실제 각 문제에 부여된 스킬을 비교함으로써 스킬 적합도를 판단하여 스킬 특성을 업데이트 한다.

Ⅲ. 실험 결과

본 챕터에서는 앞에서 제안한 1) 난이도 업데이트 시스템, 2) 스킬 적합도 판단 시스템의 실험 및 결과에 대하여 설명한다. "Class saathi"는 인도의 6-10 학년의 수학, 과학 문제를 제공하며 초반 유저의 실력을 분석한 뒤 Fig4 와 같이 실력에 맞는 문제를 추천하여 단기간 실력 향상에 도움을 주는 앱으로 본 실험은 "Class Saathi" 앱의 데이터에 기반한다.

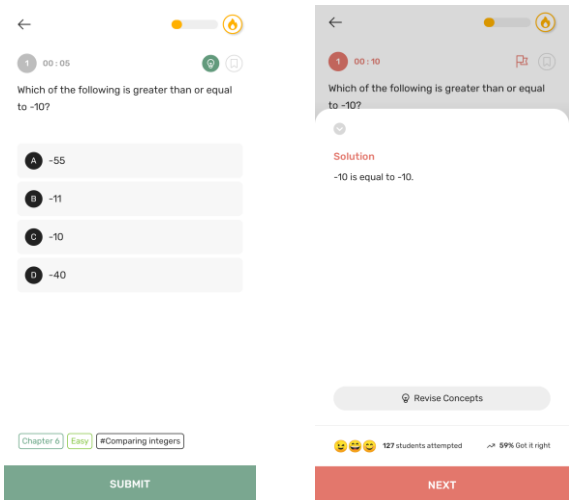

Fig 4. Class Saathi 퀴즈 화면 예시

1) 난이도 업데이트 실험

Fig4 의 왼쪽 그림에서와 같이 각 문제는 각기 다른 난이도를 가지고 있고, 앱의 화면에서는 Easy, Medium, Hard 로 표기되는 것과 달리 DB 에는 0~1 의 값으로 상세하게 분류 되어 있다. 각 문제를 접하는 사용자 그룹군의 실력에 적합한 난이도를 재부여 하기 위해서 최소 100 번 이상 풀어진 문제에 대하여 난이도 업데이트를 실시한다.

문제 ID (5 개 예시)	기존 난이도	새로운 난이도
06AeF110e0sBT8SOwgXp	0.66	0.527856

7THy1ZZsWAXJmcsnvk6B	0.33	0.516037
02XoWLZtk26RCt5SEtV1	0.33	0.612946
hg32mmDvbQkbgWR5EHtL	0.66	0.48694
01K10LcUw6TVU4trmKXb	0.2	0.319829

Table 1. 문제 난이도 업데이트 결과

식 1 에 따라 문제의 새로운 난이도를 생성한 결과는 Table 1 과 같다. 이를 통해 기존 문제의 난이도는 새로운 난이도로 대체하여 부여되며 유저의 문제 풀이 기록이 충분히 쌓인 뒤 난이도가 업데이트 된 문제의 데이터 통계를 분석한다. 분석 기준은 각 문제의 정답율이 약 70%가 되는 것이 목표이다.

문제 ID (5 개 예시)	정답율 (%)	유저에 의해 풀어진 횟수
06AeF110e0sBT8SOwgXp	74.8	127
7THy1ZZsWAXJmcsnvk6B	78.23	124
02XoWLZtk26RCt5SEtV1	76.42	106
hg32mmDvbQkbgWR5EHtL	64.71	102
01K10LcUw6TVU4trmKXb	64.58	96

Table 2. 난이도 업데이트 문제의 정답율

각 문제는 난이도가 업데이트 된 이후의 유저에 의해 풀어진 횟수가 많은 순으로 나열되어 있으며, 정답율은 최저 64%에서 최고 74%로 목표 정답율인 70%에 근접하였다. 각 문제들은 분기 혹은 풀어진 횟수에 따라 1 년 기준으로 약 4-6 회 업데이트 되어 사용자 그룹군의 실력의 정규 분포에 맞는 난이도를 부여 받게 된다.

2) 스킬 적합도 판단 실험

Fig 4 와 같이 각 문제에는 스킬이 1 또는 2 개 부여되어 있다. 업데이트 되어 사용자 그룹군의 실력의 정규 분포에 맞는 난이도를 부여 받게 된다. 해당 스킬들은 문제 생성자에 의해 사전에 부여 되며, 본 실험에서는 유저의 문제 풀이 데이터를 활용하여 이에 대한 적합도를 판단한다.

Logistic model 을 활용하여 Q-S' matrix 를 학습하며 이후 결과는 Fig5 와 같다. 왼쪽은 실제 스킬이 부여된 인덱스 번호이고 오른쪽은 학습 뒤 얻어진 Q-S'의 결과이다. 이처럼 각 문제에 대하여 얻어진 Q-S'의 결과 값을 통하여 실제 문제의 스킬 인덱스와 비교하며 스킬 적합도를 판단 할 수 있다. 하지만 Fig6 과 같이 스킬 적합도의 성능의 경우 해당 문제가 유저에게 풀어진 데이터 셋의 크기에 따라 성능의 차이가 발생할 수 있다.

Fig6 은 해당 문제의 Q-S' matrix 의 학습이 완료된 전체 값을 보여준다. 일반적으로 학습된 Q-S' matrix 의 max 값을 취해 비교하고 문제에 스킬이 복수로 부여되어 있는 경우 max 2 개 값까지 추출 한 뒤 비교한다.

Fig 5. 문제(v1iNuUfvpd8NJqnhqUni)에 부여된 스킬 인덱스와 Q-S' matrix 값

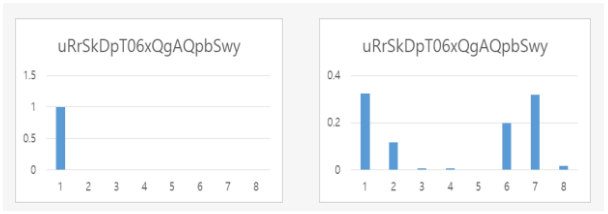

Fig 6. 문제(uRrSkDpT06xQgAQpbSwy)에 부여된 스킬 인덱스와 Q-S' matrix 값

Ⅳ. 결론

본 논문에서는 추천시스템의 높은 성능을 확보하는데 기반이 될 수 있도록 컨텐츠의 상대적 특성을 업데이트 하는 방법에 대하여 소개하였다. 본 연구는 현재 사용자 군집의 데이터에 기반하여 새로운 난이도를 부여하고 컨텐츠의 스킬 적합도를 판단함으로써 컨텐츠가 갖는 사용자 군집의 상대적 특성 차이 의존도를 낮추고 활용도를 높였다. 각 사용자의 군집의 데이터를 충분히 수집하면 제안 방법의 효과를 극대화할 수 있으며 난이도 업데이트 시스템, 스킬 적합도 판단 시스템 모두 단일 횟수의 적용 보다는 복수의 횟수를 반복함으로써 사용자 군집의 특성이 정규 분포 형태로 반영될 수 있도록 하는 것이 중요하다. 추후 연구에서는 단일 횟수의 업데이트 만으로 상대적 특성을 업데이트 하고 대량의 사용자 군집 데이터를 실시간으로 활용할 수 있는 방법에 대하여 논의한다.

ACKNOWLEDGEMENT

본 연구는 2020 년도 중소벤처기업부의 기술개발사업 지원에 의한 연구임. [S2814142]

참 고 문 헌

[1] 이석준, and 이희춘. "협업 필터링 추천에서 대응평균 알고리즘의 예측 성능에 관한 연구." *Information Systems Review* 9.1 (2007): 85-103.

[2] Jia, Bing, Yongjian Yang, and Jun Zhang. "Study on Learner Modeling in Adaptive Learning System." *JCP* 7, no. 10 (2012): 2585-2592.