creative commons
COMMONS DEED

저작자표시-비영리-동일조건변경허락 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

동일조건변경허락. 귀하가 이 저작물을 개작, 변형 또는 가공했을 경우에는, 이 저작물과 동일한 이용허락조건하에서만 배포할 수 있습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학석사 학위논문

행렬 분해 및 잠재 변수를 활용한 협업 필터링 기반 텍스트 컨텐츠 추천

A Collaborative Filtering Approach to Content Recommendation using Matrix Factorization and Textual and Social Latent Variables

2014 년 1 월

서울대학교 대학원

산업공학과

윤주식

행렬 분해 및 잠재 변수를 활용한 협업 필터링 기반 텍스트 컨텐츠 추천

지도 교수 박 종 헌

이 논문을 공학석사 학위논문으로 제출함
2014 년 1 월

서울대학교 대학원
산업공학과
윤주식

윤주식의 공학석사 학위논문을 인준함
2014 년 1 월

위 원 장 조 성 준 (인)

부위원장 박 종 헌 (인)

위 원 이 재 욱 (인)

초 록

온라인 쇼핑 서비스나 멀티미디어 컨텐츠 서비스에서는 사용자들이 자신이 원하는 컨텐츠를 쉽게 찾을 수 있도록 추천 시스템을 활용하고 있다. 포탈 커뮤니티 사이트나 소셜 네트워크 서비스도 협업 필터링 등의 추천 방법론을 활용하여 서비스 내에서 발생하는 텍스트 컨텐츠를 추천할 경우, 사용자들이 더 다양한 컨텐츠를 쉽게 접할 수 있다.

그러나 텍스트 컨텐츠의 경우 컨텐츠에 대한 선호가 평점 또는 구매와 같은 형태로 직접적으로 나타나지 않기 때문에 정확하게 파악하기 어렵다. 또한 사용자의 선호를 정확히 파악할 수 있다고 해도, 웹 커뮤니티나 소셜 네트워크는 텍스트 컨텐츠의 생성 및 소비 속도가 매우 빠르기 때문에 동일한 컨텐츠에 대한 사용자들의 선호가 누적되지 않는다. 따라서 일반적인 협업 필터링 기반의 추천 방법론을 그대로 활용하는 것은 적절하지 않다.

본 논문에서는 이러한 텍스트 컨텐츠를 효과적으로 추천하기 위한 협업 필터링 기반 추천 방법론을 제안한다. 제안한 모델은 사용자와 컨텐츠 간의 상호 작용을 통해 사용자의 컨텐츠에 대한 선호를 점수 또는 순위로 추정하고, 이를 활용하여 잠재 변수 기반의 행렬 분해 모델 학습 및 추천을 수행한다. 해당 모델은 주제 모델링을 통해 얻은 텍스트 컨텐츠(제목 및 본문)의 주제, 컨텐츠의 작성자 정보와 사용자간의 잠재 변수를 사용하고 그 외 다양한 정보를 성향 정보로 활용하여 텍스트 컨텐츠 추천의 콜드 스타트 문제를 보완한다.

실험 결과 논문에서 제안한 점수 기반 모델과 순위 기반 모델, 그리고 컨텐츠 제목의 단어 비중을 높이는 제목 단어 빈도 증가 모델의 텍스트 컨텐츠 추천 결과가 일반적인 행렬 분해 기반 협업

필터링 추천 모델보다 정밀도 측면에서 더 좋은 결과를 보였다. 또한 모델에 사용한 다양한 요소가 모두 추천 결과를 개선하는데 기여함을 확인할 수 있었다.

주요어 : 협업 필터링, 행렬 분해, 잠재 변수, 텍스트 컨텐츠 추천, 주제 모델링

학 번 : 2012-21065

목 차

제 1 장 서 론 1

1.1. 연구 배경 1

1.2. 연구 목적 3

1.3. 연구 내용 3

제 2 장 배경 이론 및 관련 연구 5

2.1. 배경 이론 5

2.1.1. 추천 시스템 5

2.1.1.1. 컨텐츠 기반 추천 6

2.1.1.2. 협업 필터링 6

2.1.1.3. 콜드 스타트 문제 7

2.1.2. 행렬 분해 7

2.1.3. 잠재 디리클레 할당 9

2.2. 관련 기존 연구 11

제 3 장 제안 모델 14

3.1. 선호도 14

3.2. 주제 모델링 15

3.2.1. 잠재 디리클레 할당을 통한 주제 모델 학습 15

3.2.2. 주제 필터링 16

3.3. 행렬 분해 모델 18

3.3.1. 잠재 변수 18

3.3.2. 성향 20

3.3.3. 최종 모델 22

3.3.4. 기존 모델과의 차이 23

3.4. 모델 학습 24

3.4.1. 점수 기반 학습 25

3.4.2. 순위 기반 학습 27
3.5. 제목 본문 구분 모델 28
3.5.1. 제목 문서 분리 모델 29
3.5.2. 제목 단어 빈도 증가 모델 31
3.6. 모델 결과를 통한 추천 32
제 4 장 실험 및 결과 34
4.1. 실험 데이터 34
4.2. 실험 환경 35
4.3. 실험 결과 36
4.3.1. 기본 모델 37
4.3.1.1. 점수 기반 학습과 순위 기반 학습 비교 38
4.3.1.2. 잠재 변수 및 성향 비중 39
4.3.1.2. 구성 요소 별 정밀도 41
4.3.2. 제목 본문 구분 모델 43
4.3.2.1. 제목 문서 분리 모델 44
4.3.2.1. 제목 단어 빈도 증가 모델 45
4.3.3. 성능 비교 46
4.3.3.1. 정밀도 비교 46
4.3.3.2. 결과 크기 비교 48
제 5 장 결론 49
5.1. 결론 49
5.2. 향후 발전 방향 51
참고문헌 53
Abstract 58

표 목차

표 1. 잠재 디리클레 할당을 통해 얻은 주제 모델링 결과 예제 10
표 2. 사용자 및 컨텐츠 관련 성향 요소 21
표 3. 추천 모델 파라미터 35
표 4. 요소 비중에 따라 점수 기반 학습한 정밀도 결과 37
표 5. 요소 비중에 따라 순위 기반 학습한 정밀도 결과 38
표 6. 제목 문서 분리 모델의 제목 비중에 따른 정밀도 43
표 7. 제목 단어 빈도 증가 모델의 제목 비중에 따른 정밀도 44

그림 목차

그림 1. 유튜브의 연관 비디오 추천 시스템 1
그림 2. 사용자와 컨텐츠 간 관계 행렬의 분해 8
그림 3. 잠재 디리클레 할당 모델 10
그림 4. 사용자-컨텐츠 작성자 간 잠재 변수 비중에 따른 컨텐츠 작성자의 상대적 다양성 40
그림 5. 점수 기반 학습의 각 요소 별 정밀도 42
그림 6. 순위 기반 학습의 각 요소 별 정밀도 42
그림 7. 모델 간 정밀도 비교 47

제 1 장 서 론

1.1. 연구 배경

인터넷 속도의 증가와 하드웨어 용량의 증가 등의 요인에 의해 온라인 쇼핑 서비스 또는 컨텐츠 서비스에서 제공하는 컨텐츠의 양이 크게 늘어났다. 대표적인 온라인 동영상 스트리밍 서비스인 유튜브(YouTube)의 경우 홈페이지에서 제공하는 공식적인 통계 페이지[15]에 따르면 매 분마다 100시간 분량의 새로운 영상이 업로드 되고 있으며, 온라인 사진 업로드 서비스 플리커(Flickr)에는 매일 3천 5백만 개의 새로운 사진이 업로드 되고 있다[10].

컨텐츠의 양이 증가한다는 것은 사용자에게 제시되는 컨텐츠의 다양성이 증가함을 의미한다. 그러나 컨텐츠의 증가 속도가 빨라지고 컨텐츠의 수가 많아질 수록 사용자들은 자신이 원하는 컨텐츠를 찾는데 어려움을 겪는다.

그림 1. 유튜브의 연관 비디오 추천 시스템

이러한 단점을 해소하기 위한 방법으로 많은 온라인 컨텐츠 서비스에서 추천 시스템을 활용한다. 추천 시스템은 사용자들이 서비스 내에서 필요로 하는 컨텐츠를 추천해주어 접하는 컨텐츠의 다양성을 늘려주고[9], 결과적으로 컨텐츠 서비스를 효과적으로 이용하는데 도움을 준다. 실제로 유튜브의 경우 비디오 조회수의 약 30%가 그림 1과 같은 연관 비디오 추천 시스템에서 발생한 것으로 나타났으며[27], 아마존은 매출의 20%~35%가 추천 시스템에서 발생하고[19], 넷플릭스는 75%의 구매가 추천을 통해 발생하는 것으로 알려져 있다[1]. 이러한 통계를 볼 때 추천 시스템이 실제 온라인 컨텐츠 서비스에서 사용자들에게 효과적인 추천을 수행하고 있음을 확인할 수 있다.

온라인 쇼핑이나 멀티미디어 컨텐츠 서비스 외에, 포탈 커뮤니티 사이트나 소셜 네트워크 등의 웹 커뮤니티 서비스의 경우에도 서비스 내부에서 발생하는 텍스트를 컨텐츠로 본다면 역시 추천 방법론을 통해 사용자들의 컨텐츠 이용에 도움을 줄 수 있다. 사용자가 읽고 싶어하는 컨텐츠를 추천해줄 수 있다면, 이는 사용자 경험을 높이고 장기적으로는 해당 서비스에 대한 사용자의 인식을 긍정적인 방향으로 변화시킬 것이다.

그러나 이러한 텍스트 컨텐츠의 추천은 온라인 쇼핑 상품이나 멀티미디어 컨텐츠에 비해 상대적으로 쉽지 않다. 텍스트 컨텐츠는 다른 컨텐츠에 비해 일반 사용자가 생성하기 쉬워 그 양이 매우 많으며, 재배포가 용이한 텍스트 컨텐츠의 특성 상 동일한 컨텐츠에 대한 여러 사용자들의 반복적인 선호 데이터를 얻는 것이 쉽지 않다.

또한 텍스트 컨텐츠의 경우 사용자의 컨텐츠에 대한 선호를 직접적으로 파악하기 힘들다. 멀티미디어 컨텐츠 또는 쇼핑 상품의

경우 컨텐츠에 대한 평점, 구매 등의 형태로 사용자의 직접적인 선호를 확인할 수 있지만, 텍스트 컨텐츠의 경우 컨텐츠에 대한 조회, 댓글 등의 상호 작용 요소를 활용하여 사용자의 컨텐츠 선호 여부를 간접 추론하여야 한다.

따라서 이러한 데이터 부족 및 선호도 추론 문제를 고려하여 텍스트 컨텐츠에 적합한 추천 시스템을 제시할 필요가 있다.

1.2. 연구 목적

본 연구에서는 웹 서비스의 텍스트 컨텐츠 추천을 위한 방법론을 제안한다. 제안하는 방법론은 사용자와 컨텐츠 간의 상호작용 기록을 통해 사용자의 컨텐츠에 대한 선호도를 추정하여 계산하고, 컨텐츠 제목 및 본문의 내용, 서비스, 컨텐츠 작성자, 그 외 글의 선호를 추론할 수 있는 요소를 반영한 협업 필터링(collaborative filtering) 모델을 학습하여 사용자가 선호할만한 새로운 텍스트 컨텐츠를 추천을 할 수 있도록 한다.

해당 모델은 행렬 분해(matrix factorization)를 기반으로 하고 있으며, 텍스트 컨텐츠의 특성을 고려하여 잠재 디리클레 할당(latent Dirichlet allocation) 기반의 주제 모델링을 활용함으로써 일반적인 협업 필터링 추천에서 발생하는 선호 데이터의 부족 문제를 보완하고 컨텐츠 추천 방법론의 장점도 포함할 수 있도록 한다.

1.3. 연구 내용

본 연구는 사용자의 컨텐츠에 대한 알려지지 않은 선호를 예측하기 위해, 순위 기반 트위터 컨텐츠 추천 시스템[6]에서 제시한 트위터 서비스 및 컨텐츠의 특성을 반영하는 행렬 분해 방법론을 일부 변형한 모델을 제안한다. 모델에 반영하는 특성은 사용자와 컨텐츠(제목 및 본문) 주제 간의 관계, 사용자와 컨텐츠 작성자 간의 관계, 사용자의 영향력, 컨텐츠의 영향력 등이며, 각 특성은 행렬 분해 모델에서 잠재 변수, 성향(bias) 등 다양한 방법으로 표현 가능하다.

이후 데이터를 활용하여 제안한 방법론의 모델을 학습하고 추천 결과를 평가한다. 모델의 학습은 점수 기반 학습의 경우 유저들이 컨텐츠와의 상호 작용 기록을 계산하여 얻은 선호도를, 순위 기반 학습의 경우 유저에 대하여 상대적으로 선호도가 다른 컨텐츠 쌍을 학습 세트로 하여 진행한다. 학습한 모델을 통해 추천된 결과가 정확한지 확인하고, 각 모델에서 잠재 변수, 성향 등의 요소가 결과에 미치는 영향을 평가한다. 그리고 해당 모델이 기존의 협업 필터링 추천 방법론에 비해 텍스트 컨텐츠 추천에 효과적인지 확인한다.

제 2 장 배경 이론 및 관련 연구

2.1. 배경 이론

2.1.1. 추천 시스템

일반적으로 사용자에게 적합한 컨텐츠를 제공하기 위해서는 추천 시스템 또는 검색 두 가지 방법[2]을 활용할 수 있다. 검색은 사용자로부터 직접 정보를 받는 수동적인 방법인 반면에, 추천 시스템은 주어진 데이터만을 활용하여 추천할 데이터를 선정하는 더 능동적인 방식이다.

추천 시스템은 기계학습의 중요한 응용 중 한 분야로 동영상 및 음악 추천, 상품 추천 같은 마케팅 분야의 주제를 주로 다루고 있다. 추천 시스템의 가장 보편적인 형태는 사용자가 컨텐츠에 대해 평가를 한 정보가 있을 시 이를 기반으로 가장 적절한 컨텐츠를 추천하는 방식을 취한다. 일반적으로 컨텐츠의 내용을 기반으로 하는 내용 기반 추천(content-based recommendation) 시스템과 컨텐츠를 추천하고자 하는 사용자와 유사한 패턴을 지닌 사용자를 기반으로 하는 협업 필터링 시스템 두 가지[25]로 나뉜다.

많은 웹 서비스에서 추천 시스템 방법론으로 주로 협업 필터링 방법을 사용하고 있으며[8][13][17], 두 방법론을 결합하여 장점을 얻으려는 하이브리드 시스템[2]도 활용되고 있다.

2.1.1.1. 컨텐츠 기반 추천

컨텐츠 기반 추천 기법은 컨텐츠의 메타 정보를 활용한다. 사용자가 선호를 나타낸 컨텐츠의 메타 정보를 분석하여 다른 컨텐츠의 메타 정보와의 유사도를 찾고, 이를 통해 관심을 보인 컨텐츠와 가장 유사도가 높은 컨텐츠를 추천하는 방식이다. 컨텐츠간 유사도의 평가는 코사인 유사도(cosine similarity), 자카드 계수(jaccard coefficient) 등의 기준을 사용할 수 있다.

컨텐츠 기반 추천 방법론은 사용자의 기존 선호와 가까운 컨텐츠를 추천해줄 수 있다는 장점이 있으나, 추천 대상의 범위가 협업 필터링 등의 방법론에 비해 좁다는 한계가 있다. 예를 들어 사용자가 A라는 정보에 대하여 선호를 표현하였을 경우, 컨텐츠 기반 추천 방법론은 A와 유사한 컨텐츠만을 추천할 수 있다. 따라서 사용자가 A와 거의 유사하지 않은 B라는 정보를 선호 할 가능성이 있다고 하여도 이를 알아내는 것이 어렵다.

2.1.1.2. 협업 필터링

협업 필터링은 컨텐츠 기반 추천 방법론과는 달리, 사용자가 컨텐츠에 대하여 평가 또는 선호를 표현한 결과를 사용하여 추천을 수행한다. 평가 정보를 바탕으로 유사한 평가를 하는 사용자 그룹을 찾아낸 후, 그룹 내 사용자들의 평가 정보를 바탕으로 다른 사용자들에게 컨텐츠에 대한 평가 예측을 제공하는 방식이다.

협업 필터링 방법론은 사용자들간의 유사함을 통해 컨텐츠를 추천하므로, 컨텐츠 기반 추천 방법론의 단점을 보완할 수 있다. 사용자가 C라는 컨텐츠에 대한 선호를 표현하지 않았더라도,

사용자와 비슷한 선호 패턴을 보인 다른 사용자가 C 컨텐츠에 선호를 보였다면 해당 사용자 역시 C 컨텐츠를 선호 할 가능성이 있다.

그러나 협업 필터링 방법론은 새로운 컨텐츠의 경우 추천이 어렵다는 단점이 있다. 새로운 컨텐츠는 사용자들의 선호가 아직 나타나지 않았기 때문에 사용자들의 선호 유사도를 통한 추천 대상이 되기 어렵다. 이는 컨텐츠 기반 추천 방법론에 비해 협업 필터링 방법론이 지닌 약점이다.

2.1.1.3. 콜드 스타트 문제

컨텐츠 기반 추천 방법론과 협업 필터링 방법론 모두 사용자들의 선호 데이터가 적은 초기 상황에서는 추천이 어렵다는 단점[24]이 있다. 이러한 문제를 해결하기 위해서 잠재 변수 또는 확률 기반 모델을 활용하여 추천하거나, 컨텐츠 이외의 간접적인 요소를 활용하여 간접 추천[12]을 하거나, 컨텐츠를 다양한 요소로 분해거나 저 차원 요소로 매핑하고 이를 기반으로 추천[21]을 시도하는 방법론 등이 연구되고 있다.

2.1.2. 행렬 분해

행렬 분해는 협업 필터링 기반 추천 시스템에서 주로 쓰이는 방법 중 하나이다. 사용자가 컨텐츠에 표현한 선호는 사용자와 컨텐츠 간의 행렬 형태로 표현 가능한데, 이를 저 차원의 잠재 요소 차원에 매핑된 두 개의 행렬로 분해하는 것이 행렬 분해 기법이다.

행렬 분해 기반의 협업 필터링은 이웃 기반 협업 필터링 방법론에 비해 효과적[13]인 것으로 알려져 있다.

사용자와 컨텐츠 간의 관계 행렬을 분해 한 결과는 그림 2와 같은 모습을 보인다. 분해 완료된 행렬을 곱하여 단일 관계 행렬을 계산할 경우, 기존에 존재하였던 값들에 한하여 분해 하기 전 행렬의 값에 근사한 값을 가진다. 이는 사용자와 컨텐츠 간의 관계가 잠재 요소 차원을 통해 표현되었기 때문이다. 이러한 잠재 요소 차원을 통하여 사용자와 컨텐츠 간의 알려지지 않은 선호 관계를 확인하는 것이 가능하며, 추천 시스템은 이를 통하여 사용자에게 다른 컨텐츠를 추천한다. 예를 들어 어떤 사용자가 A라는 컨텐츠를 선호한다고 할 때, B라는 컨텐츠를 잠재 요소 차원에 매핑한 결과가 A 컨텐츠를 잠재 요소 차원에 매핑한 결과와 비슷하다면 해당 사용자는 B 컨텐츠도 선호할 것이라고 보는 것이다.

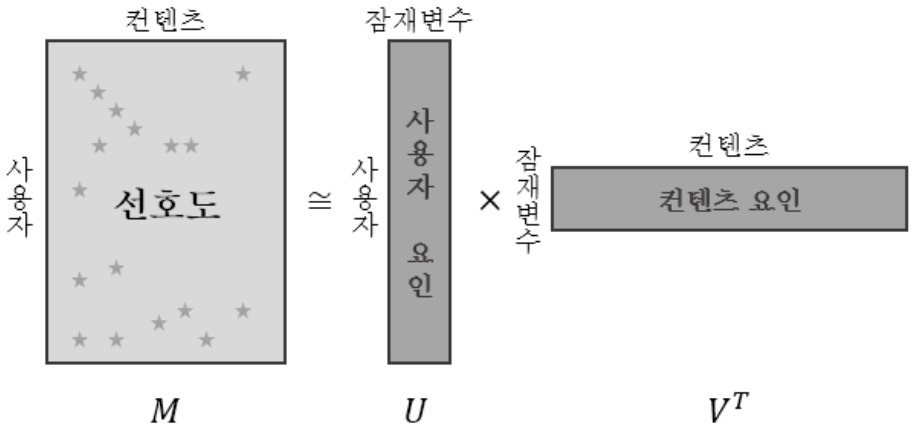

그림 2. 사용자와 컨텐츠 간 관계 행렬의 분해

따라서 잠재 변수를 활용한 행렬 분해 기법을 활용하면 특정

사용자와 컨텐츠 간의 모든 선호 관계를 계산한 전체 행렬을 유지할 필요가 없다. 사용자와 컨텐츠를 잠재 변수에 매핑한 잠재 변수 차원 행렬을 오프라인에서 학습해두면 온라인에서 두 행렬간의 단순 곱을 통해 빠른 속도로 결과를 얻을 수 있기 때문이다.

잠재 변수에 대한 사용자와 컨텐츠 행렬 학습은 기존 데이터 행렬과 잠재 변수 행렬의 내적 곱을 통해 얻은 행렬 간의 제곱 오차(squared error)를 최소화 하는 방식으로 진행된다. 주어진 사용자와 컨텐츠 간의 관계 데이터를 사용하여 분해한 행렬의 곱이 기존 행렬에 가깝도록 학습을 수행하며, 과적합을 방지하기 위하여 정규화(regularization) 등의 방법론을 사용한다. 이에 해당되는 방법론으로는 확률 기울기 하강법(stochastic gradient descent)과 교차 최소 제곱법(alternating least squares) 등이 있다.

2.1.3. 잠재 디리클레 할당

잠재 디리클레 할당은 문서를 다양한 주제의 조합에서 도출된 단어의 조합으로 보는 확률 모델[4]이다. 하나의 문서를 작성한다고 할 때, 해당 문서에는 여러가지 주제가 각각 특정한 확률로 나타날 가능성이 있다. 그 중 한가지 주제를 선택하였을 때 해당 주제에 관련된 단어도 각각 특정한 확률로 나타난다. 잠재 디리클레 할당은 이러한 방법을 반복하여 하나의 문서를 만든다는 가정을 한 모델이다. 따라서 잠재 디리클레 할당을 통해 얻은 하나의 문서는 여러 개의 주제로 구성된 다수의 단어의 집합으로 볼 수 있다. 해당 과정은 그림 3[4]과 같이 표현할 수 있다.

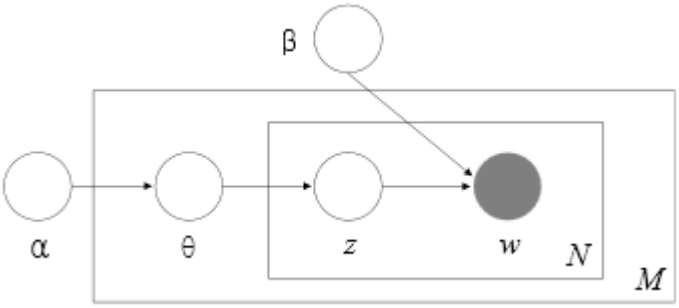

그림 3. 잠재 디리클레 할당 모델

θ는 문서의 주제 가중치 벡터, z는 선택한 주제에 속한 단어 가중치 벡터로 이를 통해 단어 w를 선택한다. α와 β는 문서 텍스트 집합에 따라 달라지는 파라미터이다.

잠재 디리클레 할당으로 생성한 주제 모델은 확률 기반이므로, 충분한 문서 텍스트를 학습 데이터로 사용하여 모델을 학습한다면 이를 이용하여 주어진 문서 텍스트의 각 단어가 어떤 주제로부터 나온 것인지를 추정하는 것이 가능하다. 이를 반복하면 주어진 문서에서 특정한 주제가 등장할 확률을 역산해낼 수 있다.

예를 들어 모델을 통해 표 1과 같은 주제 모델링 결과를 얻었다고 가정하자.

표 1. 잠재 디리클레 할당을 통해 얻은 주제 모델링 결과 예제

주제	단어
1	수면 잠 피로 장애 밤 하루 수면장애 피곤 불면증
…	…
19	자궁 자궁경부암 바이러스 미혼

20	식중독 보관 야외활동 위생 깨끗 바이러스

각 주제와 연결된 단어는 잠재 디리클레 할당 모델의 주제 z의 단어 벡터를 확률로 정렬하여 상위 확률 값을 지니는 것을 뽑은 것이다. 잠재 디리클레 할당 모델은 위와 같이 모델링한 주제를 통해 각 문서가 어떠한 주제에 어느 정도의 확률로 등장하는 지 계산해낸다. 예를 들어 문서 A에서 주제 1이 등장할 확률이 매우 높다면 해당 문서는 불면증 또는 수면 장애에 관련된 내용일 것이다. 또한 문서 B와 문서 C가 동일하게 바이러스라는 단어를 포함하고 있더라도, 문서 B에서는 주제 19가, 문서 C에서는 주제 20의 확률이 높다면 각각 인유두종 바이러스와 식중독 바이러스에 관한 주제에서 추출된 단어임을 알아낼 수 있다.

따라서 잠재 디리클레 할당을 통해 얻은 문서의 주제 확률은 문서 집합을 주제 요소 차원에 투영한 것으로 볼 수 있다[5]. 주제 요소 차원은 일반적으로 문서의 수에 비해 작으므로 문서의 공간 복잡도를 줄일 수 있으며, 주제 유사도를 통해 문서간의 관계를 표현하거나 문서를 분류하는 것도 가능하다. 이러한 방법은 컨텐츠 추천 방법론에서 활용 가능한 유사도 값을 계산하거나, 협업 필터링 추천에 컨텐츠 기반 추천을 결합하는 하이브리드 방법론을 만드는 데 사용할 수 있다.

2.2. 관련 기존 연구

많은 추천 관련 연구가 넷플릭스의 추천 서비스 개선 컨테스트인 넷플릭스 프라이즈(Netflix Prize) 데이터를 바탕으로 이루어졌다. 그 중에서는 협업 필터링을 활용하거나 확장한 추천

시스템을 많이 찾아볼 수 있다.

Salakhutdinov 등[23]은 인공신경망 모델의 일종인 제한 볼츠만 머신(RBM: Restricted Boltzmann Machines)과 기울기 하강(gradient descent) 기반의 행렬 분해 기반 추천을 통해 기존의 Bell 등의 k-근접이웃(k-Nearest-Neighborhood) 기반 협업 알고리즘[3]을 크게 개선하였다.

Zhou 등[28]은 교차 최소 제곱법 알고리즘을 통해 병렬적으로 행렬 분해를 수행하여 유사한 결과를 도출하였다. 제시한 방법론은 기울기 하강 방법론과 달리 병렬적으로 학습 가능하기 때문에 대용량의 데이터에 대하여 학습을 진행할 수 있으며, 더 많은 수의 잠재 변수에 대하여 빠른 속도로 학습할 수 있었다.

Koren 등[13]은 사용자와 컨텐츠의 성향, 시간 역학 정보 등의 간접적인 요소를 협업 필터링 학습에 적용하였다. 행렬 분해를 통한 행렬 최적화 결과 외에도 사용자와 컨텐츠간의 간접적인 요소를 모델 학습에 활용한 결과 더 좋은 추천 결과를 도출할 수 있었다.

주제 모델링을 통해 효과적인 추천을 하려는 시도도 이루어졌다. Ramage 등[22]은 마이크로블로그의 텍스트 컨텐츠를 잠재 디리클레 할당 방법론을 사용하여 주제화하고 이를 기반으로 다른 사용자 등의 컨텐츠를 추천할 수 있는 방법론을 제시하였다.

협업 필터링에 컨텐츠 기반의 추천을 결합하려는 하이브리드 방법론의 시도도 있었다. Blei 등[4]은 잠재 디리클레 할당 모델을 제시하면서 이를 활용하여 영화 평점 데이터에 대한 협업 필터링 점수 예측을 할 수 있음을 보인 바 있다.

Marlins[18]는 잠재 디리클레 할당 모델에서 문서를 주제의 집합으로 보는 것을 확장하여, 각 사용자를 추천을 수행하는

프로파일의 집합으로 보는 모델을 제시하고 이를 기반으로 협업 필터링 추천을 수행하였다.

Chen 등[6]은 컨텐츠 기반 추천 방법론과 협업 필터링 방법론의 장점을 종합하기 위하여 단어나 주제 정보를 활용하여 컨텐츠와 사용자간의 잠재 변수를 할당하는 행렬 분해 모델을 제시하였다. 트위터의 리트윗을 컨텐츠에 대한 선호 여부로 보고, 해당 지표가 존재하지 않는 컨텐츠와 쌍으로 학습 데이터를 만든 후 행렬 분해 추천 모델을 학습하는 순위 기반 학습 방법론을 사용하였다.

제 3 장 제안 모델

본 연구에서는 트위터 컨텐츠 추천 방법론[6]에서 제시하는 모델을 일부 확장한 협업 필터링 기반 행렬 분해 모델을 제안하고 여러가지 방법으로 학습을 수행한다. 해당 방법론은 주제 모델링을 통해 학습 데이터의 텍스트 컨텐츠의 주제를 얻은 후, 일반적인 행렬 분해 모델의 사용자-컨텐츠 간 잠재 변수 대신 사용자-컨텐츠 주제 간 잠재 변수, 사용자-컨텐츠 작성자 간 잠재 변수, 컨텐츠 및 커뮤니티 성향 등을 도입한 모델을 활용한다. 해당 잠재 변수 및 성향은 모델에 컨텐츠 간의 유사성, 컨텐츠 작성자에 대한 선호, 그 외 선호 여부를 추론할 수 있는 간접적인 요소들을 주어진 비중에 따라 제공하여 일반적인 행렬 분해 모델 기반 협업 필터링 추천에서 발생할 수 있는 콜드 스타트 문제를 보완한다.

3.1. 선호도

행렬 분해 모델로 컨텐츠를 추천하기 위해서는 우선 사용자의 컨텐츠에 대한 선호를 수치 값으로 정의 할 필요가 있다. 사용자 u의 컨텐츠 i에 대한 선호 r_{ui}는 댓글, 뷰 로그 등 사용자와 텍스트 컨텐츠 간의 상호 작용 요소의 발생 횟수 v를 통해 다음과 같이 표현한다.

$$r_{ui} = \sum_{j} w_j v_j \qquad (1)$$

v_j는 j번째 상호 작용 요소의 발생 횟수를 나타내는 정수값이며, w_j 는 v_j 의 비중이다. 예를 들어 j 번째 상호 작용 요소가 댓글 달기라고 하면, v_j는 사용자 u가 컨텐츠 i에 작성한 댓글의 수이다.

각 상호 작용 요소는 해당 요소 또는 커뮤니티의 특성에 따라 그 중요도가 다를 수 있으므로, 각 요소의 비중 w_j를 중요도에 따라 조절하여 적용하여야 한다. 계산을 통해 얻은 선호도 r_{ui} 는 사용자의 컨텐츠에 대한 절대적인 선호 값, 또는 여러 컨텐츠 간의 선호 정도를 비교할 수 있는 간접적인 선호 잣대로 보고 모델에 적용할 수 있다.

제시한 선호도 r_{ui}을 이용하면 모든 사용자와 컨텐츠 간의 선호 관계를 표현한 행렬을 만들 수 있다. 이 행렬은 이후 행렬 분해 모델의 분해 대상 행렬이 된다.

3.2. 주제 모델링

3.2.1. 잠재 디리클레 할당을 통한 주제 모델 학습

행렬 분해 모델에 컨텐츠 주제를 반영하기 위해서는 우선 주제 모델링을 통해 텍스트 컨텐츠를 주제의 조합으로 분석할 필요가 있다. 이는 컨텐츠 집합 내에 l 개의 주제가 존재한다고 할 때, 컨텐츠 집합 내의 모든 텍스트 컨텐츠를 $\mathbb{R}^l$ 주제 차원에 매핑하는 것으로 해석할 수 있다.

컨텐츠의 최소 단위는 단어이므로 주제 차원 대신 단어 차원을 사용한 모델을 만드는 것도 가능하다. 그러나 단어는 텍스트에 포함된 수가 매우 많아 주제에 비해 차원 효율 측면에서

비효율적이며, 동일한 표기의 단어가 상황에 따라 다양한 의미를 지니는 경우를 해결하지 못한다. 반면에 주제는 차원의 수를 모델 생성시 직접 조절할 수 있으며, 동일한 단어의 의미 해석 문제가 발생하지 않는다. 따라서 단어보다는 주제 차원 기반의 모델링을 수행하는 것이 바람직하다.

텍스트 컨텐츠를 주제로 분석하는 데에는 잠재 디리클레 할당 방법론을 사용한다. 각 컨텐츠를 잠재 디리클레 할당 모델의 문서로 보고 l개의 주제에 대하여 학습하도록 하면 해당 학습 데이터에 대하여 하나의 잠재 디리클레 할당 모델을 구할 수 있다.

학습한 모델을 활용하면 주어진 컨텐츠에 대하여 l개의 주제가 각각 나타날 확률을 구할 수 있다. 예를 들어, 5개의 주제에 대하여 학습한 잠재 디리클레 할당 모델로부터 어떠한 컨텐츠 i에서 주제가 나타날 확률을 구하면 다음과 같은 형태의 벡터를 얻을 수 있다.

$\boldsymbol{z}$	1	2	3	4	5
$\boldsymbol{P(i,z)}$	0.6	0.05	0.15	0.15	0.05

$P(i,z)$는 잠재 디리클레 할당을 통해 얻은 컨텐츠에 주제 z가 존재할 확률이다.

3.2.2. 주제 필터링

잠재 디리클레 할당은 충분한 데이터를 가지고 학습하였을 경우 주어진 컨텐츠에 어떠한 주제가 존재할 확률을 0 이상으로 출력한다. 즉, 컨텐츠 내에 모든 주제가 특정한 확률로 존재하는

것이다. 그러나 컨텐츠는 실제로 l 개의 주제를 모두 담고 있지 않으므로, 적절한 임계값(threshold)을 활용하여 컨텐츠의 주제 집합을 추려낼 필요가 있다.

임계값을 활용하여 얻은 컨텐츠 i의 주제 집합은 다음과 같다.

$$Z_i = \{z | P(i,z) > \alpha\} \quad (2)$$

Z_i는 텍스트 컨텐츠 집합 내에 존재하는 주제의 집합, α는 주제 확률의 임계값이다. (2)를 통해 얻은 집합 Z_i에 속하는 주제 z의 확률 $P(i,z)$를 모아 컨텐츠 i의 주제 확률 벡터 θ_i를 구할 수 있다. 해당 벡터는 다음과 같은 특성을 가지도록 벡터 크기를 변경한다.

$$\sum_{z \in Z_i} \theta_{iz} = 1 \quad (3)$$

θ_{iz} 는 컨텐츠 i 에 주제 z 가 존재할 보정 확률이다. 즉 최종적으로 얻은 벡터 θ_i는 임계값 이상의 확률로 나타나는 주제에 대해서만 다시 확률을 계산한 보정 벡터이다. 앞에서 제시한 예에서 $\alpha = 0.1$이라 하고 θ_i를 계산하면 다음과 같다.

$\boldsymbol{z}$	1	3	4
$\boldsymbol{\theta_{iz}}$	0.6667	0.1667	0.1667

주제 모델 학습 및 필터링을 통해 얻은 값 θ_{iz} 는 이후 행렬

분해 모델에서 사용하며, 컨텐츠 간의 내용 유사성을 모델의 잠재 변수에 반영 하는 역할을 한다.

3.3. 행렬 분해 모델

잠재 변수를 활용하는 일반적인 행렬 분해 모델[7]은 다음과 같다.

$$\hat{y}_{ui} = b + p_u^T q_i \quad (4)$$

$\hat{y}_{ui}$는 사용자 u와 컨텐츠 i간의 관계가 표현된 기존 행렬의 값 y_{ui}에 대한 실수 추정값, b는 모델, 사용자, 컨텐츠 등의 평균, 상태 등을 나타내는 성향값, $p_u^T q_i$는 k 차원의 잠재 변수 공간에 대하여 $p_u \in \mathbb{R}^k$, $q_i \in \mathbb{R}^k$로 분해한 두 잠재 변수 행렬의 곱이다. 성향은 모델 전역 성향, 사용자 u의 사용자 성향, 컨텐츠 i의 컨텐츠 성향 등이 존재하며, 필요에 따라 생략하거나 다양한 값으로 대체 가능하다. 잠재 변수도 필요에 따라 단순한 사용자-컨텐츠 간의 잠재 변수가 아닌 다양한 종류의 잠재 변수를 도입할 수 있다.

3.3.1. 잠재 변수

3.3.1.1. 사용자-컨텐츠 주제 간

트위터 컨텐츠 추천 방법론[6]에서 제안한 단어 및 주제 기반 잠재 변수를 3.2의 주제 모델링 결과를 활용하도록 변형하여 사용자 u의 컨텐츠 i의 주제에 대한 잠재 변수로 사용한다. 이를 활용한 행렬 분해 모델은 다음과 같다.

$$\hat{y}_{ui} = b + p_u^T \left(\sum_{z \in Z_i} \theta_{iz} q_z \right) \quad (5)$$

q_z는 주제 z에 대한 잠재 변수, Z_i는 3.2에서 정의한 컨텐츠 i의 주제의 집합, θ_{iz}는 컨텐츠 i에서 주제 z가 등장할 확률이다. 위와 같은 사용자-컨텐츠 주제 간 잠재 변수를 사용한 행렬 분해를 통해 사용자가 컨텐츠의 주제에 관심이 있는지를 알 수 있다. 이는 사용자가 자신이 기존에 관심 가졌던 주제를 더 선호한다는 가정에 기반한다.

해당 모델을 사용할 경우 새로운 컨텐츠를 주제의 모음으로 볼 수 있기 때문에, 다른 사용자의 해당 컨텐츠에 대한 상호 작용이 적더라도 효과적으로 추천이 가능하다. 또한 주제를 기반으로 할 경우, 잠재 디리클레 할당 모델의 주제 정보를 통해 저 차원에서 컨텐츠간 유사 관계를 표현할 수 있으므로 일반적인 잠재 변수 기반 행렬 분해 모델에 비해 최종적으로 도출되는 행렬의 크기가 더 작다[11].

3.3.1.2. 사용자-컨텐츠 작성자 간

사용자 u의 컨텐츠 i의 작성자에 대한 잠재 변수는 트위터

컨텐츠 추천 방법론[6]에서 제안한 소셜 관계 잠재 변수를 사용한다. 해당 잠재 변수를 활용한 행렬 분해 모델은 다음과 같다.

$$\hat{y}_{ui} = b + p_u^T s_{pub(i)} \quad (6)$$

$s_{pub(i)}$는 컨텐츠 i의 작성자 $pub(i)$에 대한 잠재 변수다. 위와 같은 사용자-컨텐츠 작성자 간 잠재 변수를 사용한 행렬 분해 모델을 통해 사용자가 컨텐츠의 작성자에게 관심이 있는지를 알 수 있다. 이는 사용자가 기존에 선호했던 컨텐츠의 작성자가 작성한 다른 컨텐츠를 더 선호한다는 가정에 기반한다.

해당 모델을 사용할 경우 작성자에 대한 사용자들의 기존 상호작용 결과를 활용 가능하므로, 사용자-문서 내용 간 잠재 변수와 마찬가지로 새로운 컨텐츠에 대하여 효과적으로 추천이 가능하다. 또한 컨텐츠 작성자의 수는 컨텐츠의 수보다 작으므로 모델에서 도출되는 행렬의 크기가 감소한다.

3.3.2. 성향

성향은 사용자와 컨텐츠가 속한 커뮤니티의 특성, 사용자 또는 컨텐츠의 특성에 따라 달라질 수 있는 요소로, 각 성향 값은 사용자의 컨텐츠에 대한 선호를 간접적으로 추론하는 데 사용할 수 있다[6][13]. 사용자 u의 컨텐츠 i에 대한 성향은 다음과 같이 표현 가능하다.

$$b = \sum_{k} \beta_k b_k \qquad (7)$$

b_k 는 k 번째 성향 요소이며, β_k 는 b_k 의 비중이다. 각 성향 요소는 그 중요도에 따라 비중이 다를 수 있다. 각 성향 요소는 값의 크기(또는 단위)로 인한 문제를 최소화 하기 위하여 모두 0과 1 사이의 값으로 정규화(normalization)하였다.

모델에 사용할 성향 요소는 두 가지 카테고리로 나뉘며, 표 2과 같다.

표 2. 사용자 및 컨텐츠 관련 성향 요소

카테고리	성향
사용자 관계	친구 여부
	관계 유사도
컨텐츠 정보	뷰 수
	댓글 수
	컨텐츠 길이

3.3.2.1. 사용자 관계 성향

친구 여부 성향은 사용자와 컨텐츠 작성자가 친구 관계인지를 표시하는 성향이다. 두 사람이 친구일 경우 서로의 컨텐츠를 더 선호할 것이라는 가정하에 사용한다.

관계 유사도 성향은 사용자와 컨텐츠 작성자 각각에 대하여

친구이거나 팔로우 등의 관계가 존재하는 사용자를 모두 구한 후, 이 중 공통인 사용자의 비율을 계산하는 것이다. 두 사용자의 관계가 유사할수록 비슷한 주제를 선호할 확률이 높고, 따라서 해당 컨텐츠 작성자의 글을 선호할 확률이 높아진다는 가정하에 사용한다. 자카드 계수 등의 방법을 통해 이러한 유사도를 구할 수 있다.

3.3.2.2. 컨텐츠 정보 성향

뷰 수와 댓글 수는 컨텐츠의 영향력을 표현한다. 뷰 수와 댓글의 수가 많은 컨텐츠라면 다른 컨텐츠에 비해 인기(또는 관심)이 많기 때문에 사용자가 선호할 확률이 높아진다고 가정할 수 있다.

컨텐츠 길이의 경우, 컨텐츠의 길이가 길수록 사용자가 관심을 가지는 주제가 나타날 확률이 더 높아진다고 가정한다.

3.3.3. 최종 모델

잠재 변수와 여러 성향 요소를 포함한 최종적인 행렬 분해 모델은 다음과 같다.

$$\hat{y}_{ui} = \gamma_{bias} \sum_{k} \beta_k b_k + p_u^T \left(\sum_{z \in Z_i} \theta_{iz} q_z + \gamma_{pub} s_{pub(i)} \right) \qquad (8)$$

γ_{pub}는 사용자-컨텐츠 주제 간의 잠재 변수에 대한 사용자-컨텐츠 작성자 간의 잠재 변수의 비중, γ_{bias}는 사용자-컨텐츠 주제 간의 잠재 변수에 대한 성향 값의 비중이다. 본 모델에서는 사용자-컨텐츠 내용 간의 관계가 컨텐츠 추천에서 가장 중요하므로, 이를 기준으로 사용자-컨텐츠 작성자 간의 잠재 변수와 성향 값이 가지는 상대적인 영향력을 정한다.

3.3.4. 기존 모델과의 차이

Chen 등이 트위터 컨텐츠 추천 방법론[6]에서 제시한 모델을 3.3에서 사용한 변수 정의를 통해 정의하면 다음과 같다.

$$\hat{y}_{ui} = \sum_{k} \beta_k b_k + p_u^T \left(\frac{1}{N} \sum_{z \in Z_i} q_z + \gamma_{pub} s_{pub(i)} \right) \qquad (9)$$

Z_i는 단어 또는 주제의 집합, N은 사용자-컨텐츠 단어 또는 주제간의 잠재 변수 q_z를 정규화(normalize)하는 정규화값이다.

3.3.3에서 제시한 최종 모델은 기존 모델인 (9)에 주제 필터링을 도입하였다. 이는 학습 과정에서 계산에 필요한 주제의 수를 줄여 모델의 학습 시간을 줄여주는 이점이 있다. 또한 실제 컨텐츠 내용과 관계 없는 주제가 모델에 반영되지 않으므로 모델의 추천 정확도의 향상을 기대할 수 있다.

또한 사용자-컨텐츠 주제 간의 잠재 변수에 대한 성향 값의 비중인 γ_{bias}를 추가하였다. 성향값은 잠재 변수와 달리 간접적인 정보를 수치화 한 값이므로 그 영향력을 적절하게 설정하지 않을

경우 모델의 정확도에 오히려 악영향을 미칠 위험이 있다. 따라서 성향값의 영향력을 조절할 수 있도록 하여 모델의 유연성을 높였다.

3.4. 모델 학습

사용자 u의 컨텐츠 i에 대한 추정 선호도 값 $\hat{r}_{ui}$를 정확하게 예측하기 위해서는 $\hat{r}_{ui}$을 r_{ui}에 최대한 근사해야 한다. 행렬 분해 모델을 학습하기 위해 해결해야 하는 목표 함수[6][7][13]를 일반화하면 다음과 같다.

$$\min \sum_{d \in D} \mathcal{L}(r_{ui}, \hat{r}_{ui}) \qquad (10)$$

D는 모델 학습에 사용할 학습 데이터의 집합, d는 사용자와 컨텐츠의 정보로 이루어진 학습 데이터, $\mathcal{L}$은 손실 함수이다. 모델의 학습은 주어진 학습 데이터의 사용자 u의 컨텐츠 i의 실제 선호도 값 r_{ui}와 추정 선호도 값 $\hat{r}_{ui}$의 차이를 최소화 하는 방향으로 진행한다.

해당 모델의 학습은 확률 기울기 하강(stochastic gradient descent)를 통해 진행한다. 확률 기울기 하강은 변수가 목표함수를 개선하는 기울기 방향으로 이동하도록 조정하여, 최종적으로 위에서 제시한 최소값 최적화 문제의 목표 함수의 지역 최적해를 찾는 학습 방법론이다.

학습 과정에서는 L2 정규화(L2 regularization)를 수행한다. L2 정규화는 목표 함수를 최적화 하기 위해 모델을 학습할 때

페널티를 부여하여 모델의 변수가 과적합(overfit) 되지 않도록 막아준다. 제시한 모델에서는 확률 기울기 하강 학습 방법에서 기울기 하강 정도에 대하여 기존 값 대비 페널티를 부여하여 변수의 변화 정도를 억제하는 방식으로 L2 정규화를 수행하였다.

본 논문에서는 위와 같은 방법을 통해 점수 기반 학습과 순위 기반 학습을 수행한다. 점수 기반 학습은 선호도를 직접적인 점수로 보고 모델을 학습하며, 순위 기반 학습은 각 사용자마다 컨텐츠 간의 상대적인 점수를 비교하여 모델을 학습한다. 점수 기반 학습은 3.1에서 정의한 선호도 값이 사용자의 실제 선호도를 정확하게 표현할 수록 더 좋은 모델을 얻을 수 있다. 반면 순위 기반 학습은 선호도 값의 정의가 부정확하여도, 상대적인 선호도의 순위를 매길 수 있다면 모델의 학습 결과에 큰 영향을 미치지 않는다는 장점이 있다.

3.4.1. 점수 기반 학습

선호도의 추정값 $\hat{r}_{ui}$와 행렬 분해 모델의 기존 행렬 추정값 $\hat{y}_{ui}$는 다음과 같은 관계로 표현할 수 있다.

$$\hat{r}_{ui} = f(\hat{y}_{ui}) \quad (11)$$

$f(\cdot)$는 단조 증가하는 활성 함수(activation function)이다. 본 연구에서는 다음과 같은 시그모이드(sigmoid) 활성 함수와 네거티브 로그-우도 손실 함수(negative log-likelihood loss function)[7]를 사용하여 점수 기반 학습을 수행한다.

$$f(\hat{y}) = \frac{1}{1+e^{-\hat{y}}} \quad (12)$$

$$\mathcal{L}(r,\hat{r}) = r\ln\hat{r} + (1-r)\ln(1-\hat{r}) \quad (13)$$

해당 활성 함수와 손실 함수를 기반으로 계산한 사용자 u의 컨텐츠 i에 대한 모델 목표함수 및 각 변수의 기울기 하강 값은 다음과 같다.

$$L = \min \sum_{d(u,i)\in D} (r_{ui}\ln f(\hat{y}_{ui}) + (1-r_{ui})\ln(1-f(\hat{y}_{ui}))) \quad (14)$$

$$\frac{\partial L}{\partial p_u} = \hat{e}\left(\sum_{z\in Z_i}\theta_{iz}q_z + \gamma_{pub}s_{pub(i)}\right) - \lambda_1 p_u \quad (15)$$

$$\frac{\partial L}{\partial q_z} = \hat{e}\theta_{iz}p_u - \lambda_2 q_z \quad (16)$$

$$\frac{\partial L}{\partial s_{pub(i)}} = \hat{e}\gamma_{pub}p_u - \lambda_3 s_{pub(i)} \quad (17)$$

$$\frac{\partial L}{\partial b_j} = \hat{e}\gamma_{bias}\beta_j - \lambda_4 b_j \quad (18)$$

$d(u,i)$는 학습 데이터 집합 D에 존재하는 사용자 u와 컨텐츠 i에 대한 선호도 값 r_{ui}으로 이루어진 학습 데이터, $\hat{e}$는 실제 선호 값 r_{ui}와 예측 선호 값 $\hat{r}_{ui}$의 차이, λ는 L2 정규화 페널티 비율이다. 정규화 과정에서 변수 종류에 따라 페널티 값을 다르게 조절할 수 있도록 하여 모델의 정확도를 높일 수 있도록 하였다. L의 각 변수의 기울기 하강 값은 항등(identity) 활성 함수를 쓴 경우와

크기의 차이만 있을 뿐 방향은 동일하므로 이를 그대로 사용하였다.

3.4.2. 순위 기반 학습

순위 기반 학습의 경우 사용자의 두 컨텐츠에 대한 선호도의 상대적인 비교를 통해 모델을 학습한다. 단조 증가 활성 함수에 대하여, 사용자 u의 컨텐츠 m에 대한 선호가 n에 대한 선호보다 높을 확률은 다음과 같다[7].

$$P(r_{um} > r_{un}) = f(\hat{y}_{um} - \hat{y}_{un}) \quad (19)$$

순위 기반 모델의 학습 데이터는 위와 같이 상대적인 선호 비교가 가능한 데이터 짝을 사용한다. 사용자의 선호가 컨텐츠 m에 대하여 n보다 높은 학습 데이터에 대하여, 목표 함수와 각 변수의 기울기 하강 값은 다음과 같다.

$$L = \min \sum_{d(u,m,n)\in D} \ln(1 + e^{-(\hat{y}_{um} - \hat{y}_{un})}) \quad (20)$$

$$\frac{\partial L}{\partial p_u} = \hat{e}\left(\sum_{z\in Z_m} \theta_{iz}^{m} q_z - \sum_{z\in Z_n} \theta_{iz}^{n} q_z + \gamma_{pub}\left(s_{pub(m)} - s_{pub(n)}\right)\right) - \lambda_1 p_u \quad (21)$$

$$\frac{\partial L}{\partial q_z^m} = \hat{e}\theta_{iz}^{m} p_u - \lambda_2 q_z^m \quad (22)$$

$$\frac{\partial L}{\partial q_z^n} = -\hat{e}\theta_{iz}^n p_u - \lambda_2 q_z^n \quad (23)$$

$$\frac{\partial L}{\partial s_{pub(m)}} = \hat{e}\gamma_{pub} p_u - \lambda_3 s_{pub(m)} \quad (24)$$

$$\frac{\partial L}{\partial s_{pub(n)}} = -\hat{e}\gamma_{pub} p_u - \lambda_3 s_{pub(n)} \quad (25)$$

$$\frac{\partial L}{\partial b_j} = \hat{e}\gamma_{bias}\left(\beta_j^m - \beta_j^n\right) - \lambda_4 b_j \quad (26)$$

$d(u,m,n)$ 는 학습 데이터 집합 D 에 존재하는 사용자 u 와 상대적 선호 컨텐츠 m, 상대적 비선호 컨텐츠 n으로 이루어진 학습 데이터, $\hat{e}$ 는 실제 선호도간 순위와 예측한 선호도간 순위가 다를 확률 $1 - P(r_{um} > r_{un})$ 이다. o^m 과 o^n 은 각각 컨텐츠 m 과 n 에 해당되는 변수이다.

3.5. 제목 본문 구분 모델

텍스트 컨텐츠는 많은 경우에 제목과 본문으로 분리된 형태로 구성되어 있다. 제목은 본문의 내용을 압축적으로 표현하고 있으며, 일반적으로 사용자는 제목을 보고 본문 컨텐츠를 읽을 지 여부를 결정한다.

따라서 본 논문에서는 컨텐츠의 제목에 컨텐츠의 본문과 관련된 주제 정보가 강하게 나타날 것이라는 가정을 하고, 기본 제안 모델에 제목의 영향력을 추가로 반영할 수 있는 두 가지 변형 모델을 추가로 제안한다.

3.5.1. 제목 문서 분리 모델

잠재 디리클레 할당 모델의 문서는 단어의 집합이므로, 텍스트 컨텐츠는 다음과 같은 형태의 문서로 볼 수 있다.

$$T_{doc} = T_{title} + T_{cont} \quad (27)$$

T_{title}은 제목에 포함된 단어 집합, T_{cont}는 본문에 포함된 단어 집합, T_{doc}은 제목과 본문을 포함한 전체 단어 집합을 의미한다. 단어 집합 내 단어는 중복될 수 있다.

제목 문서 분리 모델은 텍스트 컨텐츠를 하나의 단일 문서가 아닌 제목 문서와 본문 문서로 분리한다. 잠재 디리클레 할당을 통한 주제 모델링 과정은 동일하나, 제목과 본문이 각각 별개의 문서이므로 행렬 분해 모델의 잠재 변수도 따로 사용하여야 한다. 즉, 사용자-컨텐츠 주제 간의 잠재 변수가 사용자-컨텐츠 본문 주제 간의 잠재 변수와 사용자-컨텐츠 제목 주제 간의 잠재 변수로 분리하는 것이다. 이를 기반으로 한 행렬 분해 모델은 다음과 같다.

$$\begin{aligned}\hat{y}_{ui} = \gamma_{bias} \sum_{k} \beta_k b_k \\ + p_u^T \left(\sum_{z \in Z_{cont(i)}} \theta_{iz} q_z + \gamma_{title} \sum_{z' \in Z_{title(i)}} \theta_{iz'} q_{z'} \right. \\ \left. + \gamma_{pub} s_{pub(i)} \right)\end{aligned} \quad (28)$$

$Z_{title(i)}$는 컨텐츠 i의 제목 문서 T_{title}의 주제의 집합, $Z_{cont(i)}$는 컨텐츠 i의 본문 문서 T_{cont}의 주제의 집합, γ_{title}는 사용자-컨텐츠 본문 주제 간의 잠재 변수가 선호 결과에 미치는 영향에 비해 사용자-컨텐츠 제목 주제 간의 잠재 변수가 가지는 영향력의 비중을 의미한다. γ_{title}의 값을 높임으로써 모델에 미치는 제목 문서의 영향력을 높일 수 있다.

제안한 행렬 분해 모델은 3.4와 동일하게 확률 기울기 하강 방법론을 통해 학습할 수 있다. 점수 기반 학습을 하기 위한 각 변수의 기울기 하강 값은 각각 다음과 같다.

$$\frac{\partial L}{\partial p_u} = \hat{e}\left(\sum_{z \in Z_{cont(i)}} \theta_{iz} q_z + \gamma_{title} \sum_{z' \in Z_{title(i)}} \theta_{iz'} q_{z'} + \gamma_{pub} s_{pub(i)}\right) - \lambda_1 p_u \quad (29)$$

$$\frac{\partial L}{\partial q_z} = \hat{e}\theta_{iz} p_u - \lambda_2 q_z \quad (30)$$

$$\frac{\partial L}{\partial q_{z'}} = \hat{e}\gamma_{title}\theta_{iz'} p_u - \lambda_3 q_{z'} \quad (31)$$

$$\frac{\partial L}{\partial s_{pub(i)}} = \hat{e}\gamma_{pub} p_u - \lambda_4 s_{pub(i)} \quad (32)$$

$$\frac{\partial L}{\partial b_j} = \hat{e}\gamma_{bias}\beta_j - \lambda_5 b_j \quad (33)$$

순위 기반 학습을 위한 각 변수의 기울기 하강 값은 각각 다음과 같다.

$$\frac{\partial L}{\partial p_u} = \hat{e}\left(\sum_{z \in Z_{cont(m)}} \theta_{iz}^{m} q_z - \sum_{z \in Z_{cont(n)}} \theta_{iz}^{n} q_z \right.$$

$$+ \gamma_{title}\left(\sum_{z' \in Z_{title(m)}} \theta_{iz'}^{m} q_{z'} - \sum_{z' \in Z_{title(n)}} \theta_{iz'}^{n} q_{z'} \right) \quad (34)$$

$$\left. + \gamma_{pub}\left(s_{pub(m)} - s_{pub(n)}\right) \right) - \lambda_1 p_u$$

$$\frac{\partial L}{\partial q_{z'}^{m}} = \hat{e}\gamma_{title}\theta_{iz'}^{m} p_u - \lambda_2 q_{z'}^{m} \quad (35)$$

$$\frac{\partial L}{\partial q_{z'}^{n}} = -\hat{e}\gamma_{title}\theta_{iz'}^{n} p_u - \lambda_2 q_{z'}^{n} \quad (36)$$

$$\frac{\partial L}{\partial q_{z}^{m}} = \hat{e}\theta_{iz}^{m} p_u - \lambda_2 q_{z}^{m} \quad (37)$$

$$\frac{\partial L}{\partial q_{z}^{n}} = -\hat{e}\theta_{iz}^{n} p_u - \lambda_2 q_{z}^{n} \quad (38)$$

$$\frac{\partial L}{\partial s_{pub(m)}} = \hat{e}\gamma_{pub} p_u - \lambda_3 s_{pub(m)} \quad (39)$$

$$\frac{\partial L}{\partial s_{pub(n)}} = -\hat{e}\gamma_{pub} p_u - \lambda_3 s_{pub(n)} \quad (40)$$

$$\frac{\partial L}{\partial b_j} = \hat{e}\gamma_{bias}\left(\beta_j^{m} - \beta_j^{n}\right) - \lambda_4 b_j \quad (41)$$

3.5.2. 제목 단어 빈도 증가 모델

제목 단어 빈도 증가 모델은 텍스트 컨텐츠를 제목과 본문을 합친 하나의 문서로 보되, 제목에 포함된 단어 집합의 빈도를 늘려 제목의 영향력을 조절하는 방법이다. 따라서 텍스트 컨텐츠는 잠재

디리클레 할당 모델 기준으로 다음과 같은 문서로 표현 가능하다.

$$T_{doc} = \alpha_{title} T_{title} + T_{cont} \quad (42)$$

α_{title}는 본문 대비 제목 단어 집합의 비중이다. $\alpha > 1$일 경우 텍스트 컨텐츠를 표현한 문서 T_{doc}에 제목의 단어 집합이 더 빈번하게 포함되어 있으므로, 잠재 디리클레 할당 모델을 통해 문서의 주제 벡터를 추론 시 제목에 연관된 주제가 나타날 확률이 더 높을 것이라는 가정이다.

제목 단어 빈도 증가 모델은 잠재 디리클레 모델을 학습하고 각 텍스트 컨텐츠의 주제 확률 벡터를 얻는 과정에 변화를 준 것이므로, 기존 모델과 동일한 확률 기울기 하강 방법론을 통해 학습할 수 있다.

3.6. 모델 결과를 통한 추천

제시한 행렬 분해 모델의 각 잠재 변수에 임의의 초기값을 부여하고 모델을 학습하면, 이를 이용하여 사용자마다 컨텐츠 별 추정 선호도 값을 계산할 수 있다. 계산한 추정 선호도 값을 크기로 정렬하여 얻은 상위 컨텐츠가 해당 사용자에게 추천 할 컨텐츠이다.

그러나 이러한 추천 결과는 동일한 방식으로 반복적으로 모델을 학습하여도 매번 다른 결과를 도출한다. 이는 추천 과정에서 사용되는 방법론이 동일한 결과를 보장하지 않기 때문이다. 주제 모델링 과정에서 사용되는 잠재 디리클레 할당 방법론은 확률에 기반하고 있으며, 행렬 분해를 학습 하는데 사용되는 확률 기울기

하강 방법론은 최적화 문제의 전역해를 보장하지 않는다.

따라서 모델로부터 추천 결과를 안정화 하기 위하여, 행렬 분해를 통해 얻는 사용자의 컨텐츠에 대한 선호 결과를 여러 모델에서 얻은 후 이를 선형적으로 합산하는 방식을 사용한다. 이 과정에서 여러 모델에서 반복적으로 좋은 결과를 보인 추천 결과만이 살아남으므로, 단일 모델 기반인 경우에 비해 더 안정적인 결과를 기대할 수 있다.

제 4 장 실험 및 결과

4.1. 실험 데이터

헬스 포털 업체 A사의 헬스 포털 커뮤니티 서비스의 실험용 데이터를 모델 검증에 사용하였다. 해당 서비스는 건강 관련 정보를 다루고 있으며, 사용자는 이러한 건강 정보 컨텐츠를 선택하여 읽거나, 컨텐츠를 다른 SNS에 공유하거나, 댓글을 작성하는 등의 상호 작용을 할 수 있다. 실험에서는 해당 커뮤니티 사이트의 텍스트 컨텐츠 및 사용자들과 컨텐츠 간의 상호작용을 기록한 실험용 로그 데이터를 사용하였다.

데이터는 총 1,994개의 텍스트 컨텐츠로 구성되어있으며, 57명의 사용자가 268개의 컨텐츠에 대하여 상호작용한 기록을 포함하고 있다. 실험에서는 데이터의 상호작용 로그를 사용하여 선호도를 다음과 같이 정의하였다.

$$r_{ui} = 2v_{sns} + v_{com} + v_{view} \quad (43)$$

v는 사용자 u의 컨텐츠 i에 대한 공유, 댓글 작성, 뷰 등의 상호작용 횟수이다. v_{sns}는 컨텐츠를 SNS에 공유한 횟수, v_{com}은 댓글을 작성한 횟수, v_{view}는 컨텐츠를 본 횟수를 의미한다.

데이터를 기반으로 계산한 선호도 값은 사용자 및 컨텐츠와 묶어 학습 데이터로 사용하였다. 점수 기반 모델 학습에서는 사용자, 컨텐츠, 선호도 묶음을, 순위 기반 모델 학습에서는 사용자, 상호

작용을 수행한 컨텐츠와 상호 작용을 수행하지 않은 컨텐츠 쌍의 묶음을 학습 데이터로 사용하였다.

학습한 모델의 평가는 서비스 운영 측의 헬스 정보 관련 전문가가 작성한 정답셋을 사용하였다. 정답셋은 사용자의 프로필 정보 및 상호작용 정보를 바탕으로 하여 사용자들에게 추천되어야 할 내용을 선정하고, 이를 바탕으로 각 사용자 별로 1,994개의 텍스트 컨텐츠에 대하여 해당 컨텐츠가 해당 사용자에게 추천되는 것이 적절한 지 여부를 판단하는 방식으로 작성되었다.

4.2. 실험 환경

행렬 분해 및 추천을 위한 코드는 자바(Java) 언어를 사용하여 구현하였다. 주제 모델링의 사전 작업인 한글 형태소 분석은 루씬 아리랑 분석기[16]를, 잠재 디리클레 할당 모델링은 mallet 툴킷[20]을, 희소 행렬(sparse matrix) 구현체 사용 및 행렬 연산은 la4j 라이브러리[14]를 사용하였다.

실험을 수행하기 위해 모델에 필요한 공통 파라미터를 설정하였으며, 각각의 값은 표 3와 같다. 잠재 변수 수와 주제 수, 학습 반복 수 등의 경우 결과의 비교를 위해 동일한 값을 사용하였으며, 그 외 정규화 페널티 값 등의 파라미터의 경우 최적의 결과를 얻기 위해 각각 10여회의 반복적인 실험을 수행하여 획득하였다.

표 3. 추천 모델 파라미터

파라미터	값

잠재 변수 수	100
잠재 디리클레 할당 주제 수	300
주제 확률 임계값 (α)	0.005
사용자 변수 정규화 페널티 값 (λ_1)	0.0065
컨텐츠 주제 변수 정규화 페널티 값 (λ_2)	0.00065
컨텐츠 작성자 변수 정규화 페널티 값 (λ_3)	0.0065
성향 변수 정규화 값 (λ_4)	0.0065
선형 합산 모델 수	3
학습 반복 수	5000

4.3. 실험 결과

학습한 모델은 총 1,994개의 텍스트 컨텐츠에 대하여 각 사용자 별로 상위 10개의 텍스트 컨텐츠를 추천하였다. 각 사용자가 작성한 컨텐츠와 이미 해당 사용자가 상호 작용을 하였던 컨텐츠는 추천 대상에서 제외하였다.

평가 기준으로는 사용자가 추천 받은 컨텐츠가 실제로 해당 사용자에게 추천되어도 되는지 여부를 사용하였다. 총 57명의 사용자에 대하여 추천된 570개의 텍스트 컨텐츠를 정답셋 기준으로 추천 가능 여부를 판별한 후, 이를 바탕으로 다음과 같이 정밀도(precision)를 계산하였다.

$$precision = \frac{\sum_{u \in U} correct(u, R(u))}{\sum_{u \in U} n_{R(u)}} \quad (44)$$

U는 사용자 집합, $R(u)$는 사용자 u에게 추천된 컨텐츠 집합, $n_{R(u)}$ 는 추천된 컨텐츠 수, $correct(u, R(u))$ 는 정답셋 기준으로 $R(u)$의 컨텐츠 중 사용자 u에게 추천되어도 되는 컨텐츠의 개수를 출력하는 함수이다.

4.3.1. 기본 모델

주어진 학습 데이터를 사용하여 3.3에서 제안한 행렬 분해 모델을 점수 기반과 순위 기반으로 각각 학습하고 평가하였다. 각 모델은 사용자-컨텐츠 주제 잠재 변수에 대한 사용자-컨텐츠 작성자 잠재 변수의 비중 γ_{pub} 와, 성향의 비중 γ_{bias} 을 조절하여 만들었다.

여러 비중 값에 대하여 점수 기반 학습과 순위 기반 학습을 수행하고 추천 결과를 얻어 정밀도를 구한 결과는 각각 표 4, 표 5와 같다. 각 실험은 3회씩 반복하여 이루어졌으며 결과는 3회 실험에서 얻은 정밀도 결과의 평균 값이다.

표 4. 요소 비중에 따라 점수 기반 학습한 정밀도 결과

γ_{bias} \ γ_{pub}	0.01	0.1	1
0.01	0.606433	0.640936	0.682456

0.1	0.640351	**0.692982**	0.667251
1	0.520175	0.519298	0.514035

표 5. 요소 비중에 따라 순위 기반 학습한 정밀도 결과

γ_{bias} \ γ_{pub}	0.01	0.1	1
0.01	0.55614	0.588304	0.552047
0.1	0.567251	**0.626316**	0.512281
1	0.563743	0.601754	0.433333

점수 기반 학습은 비중에 따라 최대 0.69, 순위 기반 학습은 최대 0.62의 정밀도를 보인다. 정밀도는 파라미터를 조절하거나, 잠재 디리클레 할당을 활용한 주제 모델링에 사용되는 한글 형태소 분석기의 성능 조정 등을 통해 더 향상될 수 있다. 모델의 추천 정밀도 결과를 고려해볼 때, 4.1에서 사용자와 컨텐츠 간의 상호작용을 통해 정의한 선호도 값이 실제 사용자의 컨텐츠에 대한 선호도에 근사하였다고 볼 수 있다.

4.3.1.1. 점수 기반 학습과 순위 기반 학습 비교

전체적으로 점수 기반 학습이 순위 기반 학습에 비해 더 좋은 결과를 보이는 것을 알 수 있다. 성향의 비중이 1인 경우를 제외한 모든 경우에서 점수 기반 학습의 정밀도가 순위 기반 학습의

정밀도보다 높게 나타난다.

이는 순위 기반 학습의 학습 과정으로 인한 한계로 생각된다. 사용자의 컨텐츠에 대한 선호도를 단독으로 학습 데이터로 사용하는 점수 기반 학습과 달리, 순위 기반 학습은 사용자가 선호를 표현하지 않았거나 상대적으로 선호도가 낮은 컨텐츠를 함께 학습 데이터로 사용한다. 그러나 이러한 데이터는 선호 정도가 상대적으로 낮을 뿐이거나 선호하는 여부가 아직 나타나지 않았을 뿐, 실제로 사용자가 해당 컨텐츠를 선호하지 않는다고 할 수 없다. 따라서 순위 기반 학습은 위와 같이 컨텐츠의 선호 여부를 정확하게 파악하여 학습 데이터를 선정하지 않는 한 점수 기반 학습보다 좋은 결과를 보여주는 것은 어려울 것으로 생각된다.

4.3.1.2. 잠재 변수 및 성향 비중

사용자-컨텐츠 작성자 간 잠재 변수의 비중 γ_{pub} 이 0.1인 모델이 대체로 높은 정밀도를 보였다. 비중이 증가 함에 따라 정밀도가 단조 증가 하지 않는 이유는 추천 컨텐츠 집합의 작성자 다양성(diversity)을 통해 어느 정도 추정 가능하다. 작성자의 다양성은 다음과 같이 계산할 수 있다.

$$diversity = \frac{\sum_{u \in U} n_{pub(R(u))}}{n_U} \qquad (45)$$

n_U 는 사용자 집합의 사용자 수, $pub(R(u))$ 는 사용자 u 에게 추천된 컨텐츠 집합을 작성한 고유 작성자(unique publisher)의 집합, $n_{pub(R(u))}$ 는 고유 작성자 집합의 작성자 수이다. 즉, 각

사용자 별로 추천된 컨텐츠 집합 내에 평균적으로 얼마나 다양한 작성자가 존재하는 지를 표현한 값이다.

그림 4는 점수 기반 학습을 기준으로 사용자-컨텐츠 작성자 간 잠재 변수의 비중 γ_{pub}에 따라 사용자 별로 추천된 컨텐츠의 작성자 다양성을 계산한 결과이다. γ_{pub}이 0.01인 경우를 기준으로 하여, 각 γ_{pub} 값 별로 다양성의 값을 상대적으로 계산한 결과를 그래프로 표현하였다.

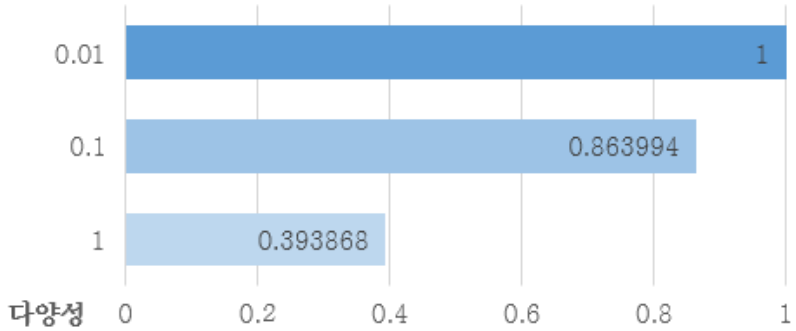

그림 4. 사용자-컨텐츠 작성자 간 잠재 변수 비중에 따른 컨텐츠 작성자의 상대적 다양성

γ_{pub} 이 커질수록 사용자에게 추천되는 컨텐츠 작성자의 다양성은 줄어드는 것을 알 수 있다. 이는 사용자-컨텐츠 작성자 간 잠재 변수는 사용자가 선호하는 컨텐츠의 작성자가 만든 다른 컨텐츠를 역시 좋아할 것이라는 가정을 포함하고 있기 때문이다. 비중이 높아질수록 모델은 기존에 선호를 표시한 컨텐츠의 작성자의 다른 컨텐츠 위주로 추천하는 것이다.

가정에 따르면 컨텐츠의 작성자는 기존의 자신이 작성한 컨텐츠와 비슷한 주제의 컨텐츠를 작성할 확률이 높고, 따라서

사용자는 해당 작성자가 작성한 새로운 컨텐츠도 역시 선호 할 확률이 높다. 그러나 작성자의 모든 컨텐츠의 주제가 사용자의 선호 주제와 같지는 않다. 따라서 사용자−컨텐츠 작성자 간 잠재 변수의 비중이 지나치게 높아질 경우, 기존에 선호한 컨텐츠의 작성자의 다른 컨텐츠만을 추천하므로 오히려 정밀도가 낮아지는 것으로 해석할 수 있다.

성향 값의 비중 γ_{bias} 도 0.1인 모델이 가장 높은 정밀도를 보였다. 이는 성향이 사용자의 요소를 간접적으로 추론할 수 있는 값이기 때문인 것으로 추정된다. 적당한 비율로 사용될 경우 선호 여부를 예측하는데 도움이 되지만, 지나치게 비중이 높을 경우 오히려 사용자−컨텐츠 주제 간의 잠재 변수의 영향력을 감소시켜 정밀한 예측을 방해하는 것이다.

결국 두 요소 모두 예측의 정밀도를 높이는 데 도움을 줄 수 있으나, 그 비중이 지나치게 높을 경우 오히려 모델의 성능을 하락시킨다고 볼 수 있다. 이를 확인하기 위해서는 각 요소가 모델 전체의 정밀도 값에 어떠한 영향을 미치는 지를 확인하여야 한다.

4.3.1.2. 구성 요소 별 정밀도

모델을 구성하고 있는 세 요소(사용자−컨텐츠 주제 간 잠재 변수, 사용자−컨텐츠 작성자 간 잠재 변수, 성향)가 모델의 정밀도에 미치는 영향력을 비교해보기 위하여, 각 요소만을 단일하게 고려하여 정밀도를 계산하였다.

γ_{pub} 과 γ_{bias} 가 0.1인 모델에 대하여 점수 기반 학습과 순위 기반 학습을 수행하였으며, 각 요소 별 정밀도 결과는 그림 5, 그림 6와 같다.

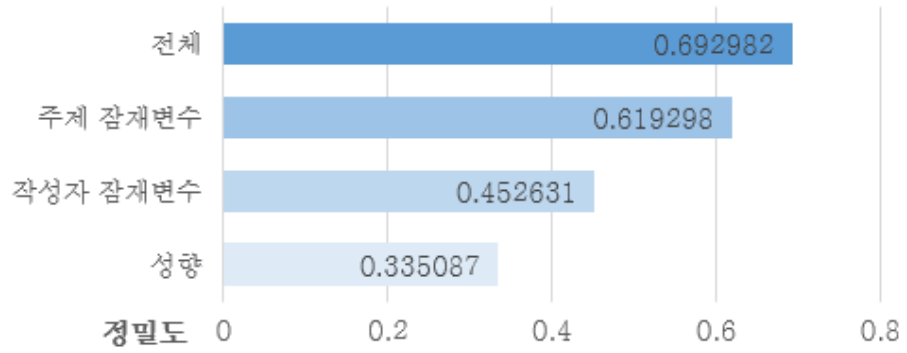

그림 5. 점수 기반 학습의 각 요소 별 정밀도

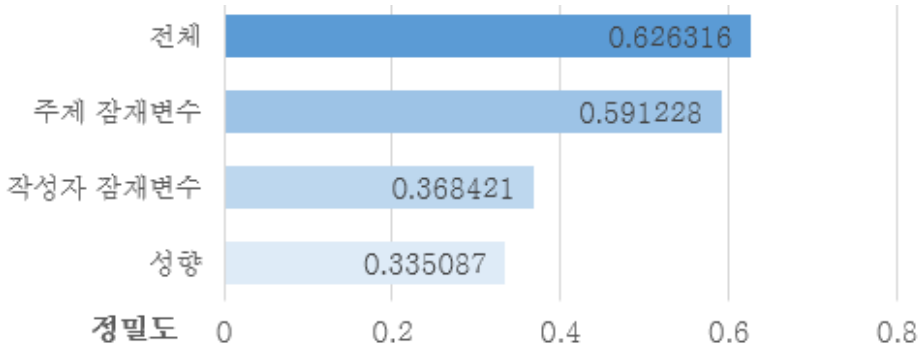

그림 6. 순위 기반 학습의 각 요소 별 정밀도

사용자-컨텐츠 주제 간 잠재 변수가 가장 정밀하고, 사용자-컨텐츠 작성자 간 잠재 변수와 성향 값은 상대적으로 정밀도가 낮다. 따라서 모델 전체의 정밀도에 가장 큰 영향력을 가지는 요소는 사용자-컨텐츠 주제 간 잠재 변수임을 알 수 있다.

한편 모든 요소를 고려한 모델이 사용자-컨텐츠 주제 간 잠재 변수만 고려한 경우에 비해 정밀도가 높다는 것을 볼 때, 사용자-작성자 간 잠재 변수와 성향 값도 정밀도 향상에 기여한다는 것을

알 수 있다. 그러나 사용자-컨텐츠 주제 간 잠재 변수에 비해 정밀도에 미치는 영향력은 낮으므로, 잠재 변수 및 성향 비중 항목에서 언급한 것처럼 적당한 비중으로 적용하여야 모델 전체의 정밀도 향상에 도움이 된다.

4.3.2. 제목 본문 구분 모델

기본 모델과 동일한 학습 데이터를 사용하여 제목 문서 분리 모델과 제목 단어 빈도 증가 모델을 평가하였다. 제목 문서 분리 모델은 사용자-컨텐츠 본문 주제 잠재 변수에 대한 사용자-컨텐츠 제목 주제 잠재 변수의 영향력 비중 γ_{title}를, 제목 단어 빈도 증가 모델은 텍스트 컨텐츠 문서의 본문 대비 제목 단어 집합의 비중 α_{title}을 조절하여 만들었다.

제목 문서 분리 모델과 제목 단어 빈도 증가 모델을 여러 비중 값에 대하여 학습하고 추천을 수행하여 정밀도를 구한 결과는 각각 표 6, 표 7와 같다. 각 실험은 3회씩 반복하여 이루어졌으며 결과는 3회 실험에서 얻은 정밀도 결과의 평균 값이다.

표 6. 제목 문서 분리 모델의 제목 비중에 따른 정밀도

γ_{title} / 학습	1	3	5
점수 기반 학습	0.645614	0.57193	0.235088
순위 기반 학습	0.592982	0.587719	0.331579

표 7. 제목 단어 빈도 증가 모델의 제목 비중에 따른 정밀도

학습 \ α_{title}	1	3	5
점수 기반 학습	0.689912	**0.750877**	**0.768421**
순위 기반 학습	0.6	0.54386	0.592982

4.3.2.1. 제목 문서 분리 모델

제목 문서 분리 모델의 경우 사용자-컨텐츠 제목 주제 잠재 변수의 비중 γ_{title}이 증가할수록 모델의 정밀도가 감소한다. 또한 γ_{title}이 1인 경우에도 4.3.1에서 구했던 기본 모델의 정밀도 0.69와 0.62에 미치지 못하는 결과를 보인다.

이러한 결과는 사용자-컨텐츠 제목 주제 잠재 변수가 텍스트 컨텐츠를 효과적으로 표현하지 못함을 의미한다. 이는 텍스트 컨텐츠의 제목을 본문과 분리하여 단독 문서로 본 것이 원인으로 생각된다. 실제로 데이터를 살펴본 결과 제목 문서의 주제와 본문 문서의 주제 간의 불일치를 발견할 수 있었는데, '빼빼 마른 나 왜 뚱뚱해 보일까?'라는 제목을 가진 컨텐츠의 경우 잠재 디리클레 할당 모델을 통해 주제를 분석한 결과 제목 문서에서는 다이어트 및 건강 관련 주제가 높은 확률로 나타났으나, 본문 문서에서는 거식증 및 정신과 관련 주제가 나타났다.

즉, 제목의 경우 단어의 수가 적기 때문에 제목에 핵심적인 단어가 포함되어 있다고 하더라도 잠재 디리클레 할당을 통해 주제가 나타날 확률을 분석하였을 땐 실제 본문의 내용과는 다른 주제들의 확률이 높게 나타날 수 있다. 따라서 사용자-컨텐츠 주제

잠재 변수를 제목과 본문으로 분리하는 제목 문서 분리 모델은 기본 모델에 비해 오히려 낮은 정밀도를 보인다고 해석할 수 있다.

4.3.2.1. 제목 단어 빈도 증가 모델

제목 단어 빈도 증가 모델의 경우, 텍스트 컨텐츠 문서의 본문 대비 제목 단어 집합의 비중 α_{title} 이 증가함에 따라 점수 기반 모델의 정밀도가 증가하는 것을 확인할 수 있다.

제목 단어 빈도 증가 모델은 제목 문서 분리 모델과 달리 제목과 본문을 하나의 문서로 같이 활용하되 제목을 구성하는 단어의 빈도만 늘려주는 방식이다. 따라서 제목에 컨텐츠 본문의 내용이 압축적으로 담겨 있으므로 이의 비중을 높이는 것이 모델 성능을 향상시킬 것이라는 가정이 옳다면, 제목의 단어들이 본문의 단어 집합과 연계되어 텍스트 컨텐츠 문서의 핵심 주제가 나타날 확률을 높여줄 것이다. 점수 기반 학습 시 α_{title} 증가에 따른 정밀도 향상을 볼 때 이러한 가정이 의미 있음을 확인할 수 있다.

다만 순위 기반 학습의 경우 α_{title} 의 변화가 모델의 성능에 큰 영향을 끼치지 않거나, 오히려 악영향을 끼친다. 이는 순위 기반 학습과 점수 기반 학습의 학습 방식의 차이 때문으로 보인다. 점수 기반 학습은 사용자가 선호하는 컨텐츠의 주제 방향으로만 모델의 변수를 학습하지만, 순위 기반 학습은 선호하지 않는 컨텐츠에 대하여도 모델 변수에 부정적인 방향으로 학습을 수행한다. 따라서 4.3.1.1에서 언급한 것처럼 순위 기반 학습의 부정적인 학습 데이터를 정확하게 선정하지 못할 경우, 이로 인한 모델 성능의 하락이 더 크게 나타날 수 있다.

4.3.3. 성능 비교

본 연구에서 제시한 모델이 효과적으로 텍스트 컨텐츠를 추천하는지 확인하기 위하여 비교 대상 모델을 선정하였다. 비교 대상으로 사용할 모델은 교차 최소 제곱법 기반 행렬 분해 모델과 3.3.4에서 기존 모델로 제시한 Chen 등의 트위터 컨텐츠 추천 방법론[6]이다.

교차 최소 제곱법 행렬 분해 모델은 사용자와 컨텐츠 간의 행렬을 잠재 변수를 통해 분해하여 결과를 예측하는 모델로, 협업 필터링 기법을 사용한 기본적인 추천 모델이다. 실험에는 넷플릭스 프라이즈 데이터를 분석하는 데 사용된 ALS-WR(Alternating Least Squares with Weighted-λ-Regularization) 행렬 분해 모델[28]과 그 구현체[26]를 사용하였다.

Chen 등의 기존 모델의 경우 γ_{bias}를 1로 고정하여 성향의 영향력을 조절할 수 없도록 하였으며, 주제 필터링을 수행하지 않고 해당 논문에서 제시한대로 정규화 변수만을 활용하였다. 또한 해당 논문에서 제시한 순위 기반 방법론만으로 학습을 수행하였다.

4.3.3.1. 정밀도 비교

ALS-WR 모델과 Chen 등의 모델을 주어진 텍스트 컨텐츠 데이터를 통해 추천 정밀도를 측정한 결과는 그림 7과 같다. 해당 모델도 점수 기반 학습 및 순위 기반 학습과 동일하게 100개의 잠재 변수에 대하여 학습한 모델 3개를 합산하는 방식을 사용하였다. 각 실험은 3회씩 반복하여 이루어졌으며, 결과는 정밀도의 평균 값이다.

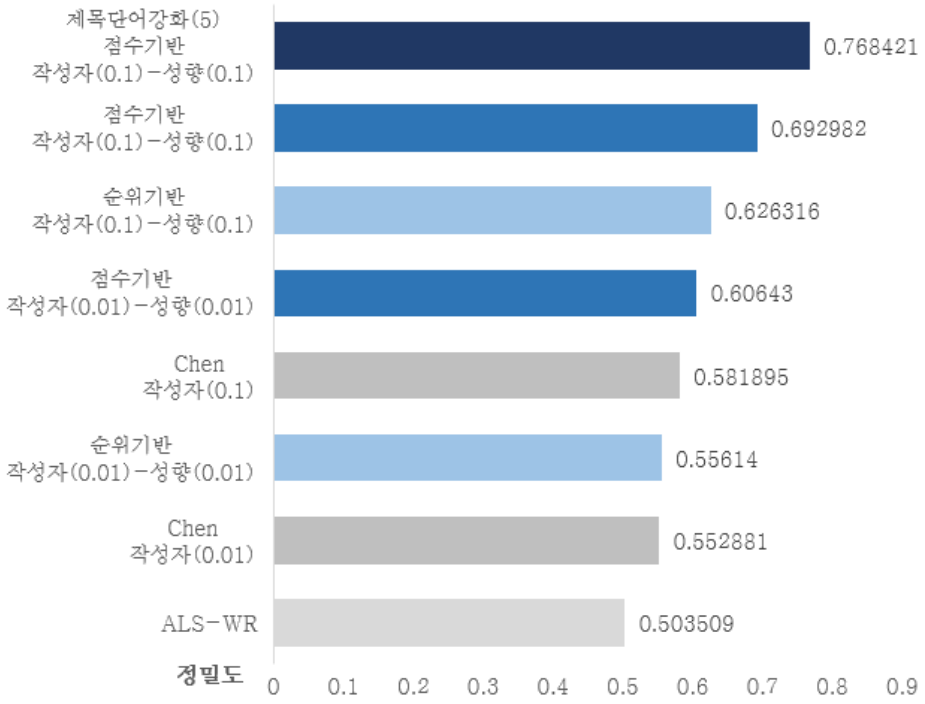

그림 7. 모델 간 정밀도 비교

제안한 모델 중 가장 성능이 좋았던 점수 기반 제목 단어 빈도 증가 모델은 물론, 기본 점수 기반 모델 및 순위 기반 모델도 ALS-WR 모델에 비해 더 높은 정밀도를 보였다. 이를 통해서 본 논문에서 제안한 행렬 분해 기반의 협업 필터링 추천 모델이 일반적인 컨텐츠 선호만을 고려하여 이루어지는 기존의 협업 필터링 추천 방식보다 텍스트 컨텐츠 추천에 더 효과적임을 확인할 수 있다.

또한 Chen 등이 제시했던 기존 모델의 경우 Chen과 동일한 방법으로 학습한 제안 모델의 순위 기반 모델과는 비슷하거나 약간 더 좋은 모습을 보였으며, 점수 기반 모델의 경우에는 확실한

정확도 향상을 보였다. 이를 통해 본 논문에서 제시한 주제 필터링 및 성향의 영향력 조절 변수가 모델의 정확도 개선에 영향을 미치고 있으며, 기존 모델에 비해 추천에 더 효과적임을 확인할 수 있다.

4.3.3.2. 결과 크기 비교

행렬 분해 기반의 잠재 변수 모델에서 추천 결과를 얻기 위해서는 사용자와 컨텐츠의 잠재 변수 행렬을 사용하여야 한다. 사용자 수를 n_u, 컨텐츠 수를 n_i라 할 때, k개의 잠재 변수를 사용하여 모델을 학습할 경우, 일반적인 사용자-컨텐츠 간 잠재 변수 기반 행렬 분해 모델에서 최종적으로 얻는 결과 행렬의 크기는 $(n_u + n_i) * k$이다.

그러나 3.3에서 제안한 행렬 분해 모델의 경우 사용자 잠재 변수 행렬 크기는 $n_u * k$로 동일하지만, 컨텐츠 내용과 작성자 잠재 변수의 행렬 크기는 주제 수를 n_z, 컨텐츠 작성자의 수를 n_p라고 할 때 $(n_z + n_p) * k$이다. 따라서 모델의 최종 결과 행렬 크기는 $(n_u + n_z + n_p) * k$이다.

웹 커뮤니티 또는 소셜 네트워크는 사용자들이 다수의 텍스트 컨텐츠를 빠른 속도로 생산하기 때문에, 컨텐츠의 수에 비해 사용자의 수가 적고 사용자 중에서도 실제 컨텐츠 작성자의 수는 더 적다. n_p와 n_i가 충분히 큰 커뮤니티라면 $n_z + n_p < n_i$로 볼 수 있으므로, 본 논문에서 제시한 행렬 분해 모델의 결과 행렬 크기는 일반적인 협업 필터링 기반 행렬 분해 모델보다 작다. 또한 n_p의 증가 속도가 n_i의 증가 속도보다 느리므로, 제안한 모델의 행렬 크기의 증가 속도도 더 느리다.

제 5 장 결론

5.1. 결론

본 논문은 온라인 커뮤니티의 텍스트 컨텐츠에 대한 사용자의 상호 작용을 기반으로 사용자의 컨텐츠에 대한 선호를 선호도 값으로 계량화하여 추정하고, 이를 기반으로 사용자에게 다른 텍스트 컨텐츠를 추천할 수 있는 협업 필터링 기반의 행렬 분해 모델을 제시하였다. 계량화한 선호 정도는 실제 점수 또는 상대적인 선호도 두 가지 방법으로 해석하여 각각 점수 기반 학습과 순위 기반 학습을 수행하였다.

제시한 두 모델은 사용자와 컨텐츠 주제, 컨텐츠 작성자 간의 관계를 잠재 변수를 사용하여 사용자의 컨텐츠에 대한 선호를 계산하도록 하였다. 또한 추가적으로 사용자와 컨텐츠 데이터에서 얻을 수 있는 성향 정보를 반영하여 간접적으로 선호 정도를 추론한 결과를 반영할 수 있도록 하였다. 이러한 요소들을 통해 데이터가 적은 상황에서 발생하는 콜드 스타트 문제를 어느 정도 해결할 수 있도록 하였다.

제시한 모델을 통해 텍스트 컨텐츠 추천을 한 결과, 점수 기반 학습 시 최대 0.69, 순위 기반 학습 시 최대 0.62의 정밀도를 보였다. 이를 통해 본 논문에서 사용한 선호도 추정이 추천 시스템에 활용 가능함을 알 수 있었다.

전체적으로 점수 기반 학습이 순위 기반 학습에 비해 더 좋은 모습을 모였다. 이는 순위 기반 학습의 학습 데이터 선정으로 인해 발생하는 문제인 것으로 생각된다. 그러나 점수 기반 학습은 선호도

추정 계산을 어떻게 하는 가에 따라서 모델의 안정성이 떨어질 수 있으므로, 상황에 따라 적합한 모델을 선택하여 추천 시스템을 만들 필요가 있다.

모델에 사용된 잠재 변수 및 성향 요소들 중에서는 사용자-컨텐츠 주제 잠재 변수가 가장 큰 영향력을 보였다. 사용자-컨텐츠 작성자 잠재 변수와 사용자 및 컨텐츠는 단일로 사용하거나 그 비중을 높일 경우 모델 정밀도에 악영향을 미치지만, 사용자-컨텐츠 주제 잠재 변수와 적당한 비중으로 결합하여 사용할 경우 주제 잠재 변수를 통한 컨텐츠 추천 결과 정밀도를 더 향상시키는 것을 알 수 있었다.

한편 제목과 본문으로 이루어진 텍스트 컨텐츠의 경우 제목이 본문의 내용 또는 주제를 함축적으로 표현하고 있으므로, 제목의 비중을 조절하여 모델을 학습할 경우 더 좋은 결과를 얻을 것이라는 가정 하에 제목과 본문을 분리한 두 가지 모델을 제시하였다. 이 중 잠재 디리클레 할당을 위한 주제 모델링 과정에서 텍스트 컨텐츠 문서의 제목 단어 집합의 비중을 높여주는 제목 단어 빈도 증가 모델의 성능이 약 0.77의 정밀도를 보였다.

최종적으로 제시한 모델 모두 일반적인 추천 시스템에 사용되는 모델에 비해서는 상대적으로 좋은 결과를 보였으며, Chen 등이 제시한 기존 모델에 비해 향상된 결과를 보였다. 이를 통해 제시한 모델이 텍스트 컨텐츠 추천에 적합하다는 결론을 내릴 수 있었다.

따라서 본 모델을 활용하여 사용자들의 선호도를 추정하고 이를 통해 텍스트 컨텐츠를 추천할 경우, 기존의 추천 방법론에 비해 더 효과적으로 컨텐츠를 추천해줄 수 있을 것이다. 또한 이를 통해 커뮤니티를 이용하는 사용자들의 사용자 경험을 증진하는 효과를 불러 올 것이다.

5.2. 향후 발전 방향

대부분의 협업 필터링 모델을 이용한 추천은 초기 데이터가 부족한 상황에서 상당히 부족한 정밀도를 보여준다. 실제로 본 논문의 경우도 일반적인 데이터 사이즈에 비하면 상당히 작은 편에 속한다. 물론 본 논문에서 사용한 잠재 변수는 기존의 사용자-컨텐츠 간 잠재 변수를 사용한 모델에 비하면 더 적은 데이터에서도 사용이 가능한 장점이 있는 것은 사실이나, 초기 데이터 부족 문제를 완전히 해결하지는 못한다.

따라서 모델에 반영하는 성향 데이터를 더욱 풍부하게 늘리거나, 데이터 양에 따라서 주제어 수 등의 파라미터를 조절하는 방식으로 초기 데이터 부족 문제를 해결할 방법을 고민 할 필요성이 있다.

위에서 언급한 협업 필터링 모델의 최적화와 더불어 제시한 모델의 추천 결과를 다른 알고리즘을 활용한 추천 방법론들과 비교할 필요도 있다. 다양한 모델의 비교를 통해 알고리즘의 시간 및 공간 복잡도와 정확도 간의 트레이드-오프 등을 확인하고 제시한 방법론이 효과적인지를 확인하는 것이 필요하다. 특히 사용자 또는 컨텐츠 등의 인기도를 반영하여 추천을 수행하는 MP(most popular) 방법론은 단순하면서도 높은 정확도를 보여주므로 좋은 비교 대상이 될 것이다.

한편 본 논문의 결과는 잠재 디리클레 할당을 통한 주제 추출에 상당부분 의존하고 있다. 따라서 주제를 효과적으로 얼마나 효과적으로 추출해내는가에 따라 정밀도가 크게 달라질 수 있다. 또한 주제 추출에 한글의 형태소 분석이 사용되기 때문에, 한글을

얼마나 효과적으로 분석하느냐도 중요한 요소가 될 수 있다.

따라서 단순한 정밀도 향상의 측면에서는 한글의 형태소 분석을 효과적으로 하고 이를 활용하여 잠재 디리클레 할당을 통한 주제 추출을 어떻게 효과적으로 할 것인가를 연구하는 것도 매우 중요하다. 이는 용어 사전 도입을 통한 형태소 분석 효율의 증대, 동의어 및 유의어 사전 도입을 통한 주제 묶음 등을 통해 이룰 수 있을 것이다.

참고문헌

[1] X. Amatriain, J. Basilico (2012), Netflix Recommendations: Beyond the 5 stars (Part 1), The Netflix Tech Blog, Retrieved from http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html

[2] M. Balabanović, Y. Shoham (1997), Fab: content-based, collaborative recommendation, Communications of the ACM, 40(3), pp. 66-72, doi:10.1145/245108.245124

[3] R. Bell, Y. Koren, C. Volinsky (2007), Modeling relationships at multiple scales to improve accuracy of large recommender systems, Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 95–104, doi: 10.1145/1281192.1281206

[4] D. Blei, A. Ng, M. Jordan (2003), Latent Dirichlet Allocation, Journal of Machine Learning Research, 3, pp. 993–1022

[5] J. Boyd-Graber, J. Chang, S. Gerrish, C. Wang, D. Blei (2009), Reading Tea Leaves: How Humans Interpret Topic Models, Proceedings of the 23rd Annual Conference on Neural Information Processing Systems

[6] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, Y. Yu (2012), Collaborative personalized tweet recommendation, Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 661-670, doi:10.1145/2348283.2348372

[7] T. Chen, Z. Zheng, Q. Lu, W. Zhang, Y. Yu (2011), Feature-Based Matrix Factorization, APEX Technical Report, Retrieved from http://svdfeature.apexlab.org/wiki/images/7/76/APEX-TR-2011-07-11.pdf

[8] A. Das, M. Datar, A. Garg, S. Rajaram (2007), Google news personalization: Scalable online collaborative filtering, Proceedings of the 16th international conference on World Wide Web, pp. 271-280, doi:10.1145/1242572.1242610

[9] D. Fleder, K. Hosanagar (2007), Recommender systems and their impact on sales diversity, Proceedings of the 8th ACM conference on Electronic commerce, pp. 192-199, doi:10.1145/1250910.1250939

[10] A. Jeffries (2013), The man behind Flickr on making the service awesome again, The Verge, Retrieved from http://www.theverge.com/2013/3/20/4121574/flickr-chief-markus-spiering-talks-photos-and-marissa-mayer

[11] Y. Koren (2008), Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 426–434, doi:10.1145/1401890.1401944

[12] Y. Koren, R. Bell (2011), Advances in Collaborative Filtering, Recommender Systems Handbook, pp. 145–186, doi: 10.1007/978-0-387-85820-3_5

[13] Y. Koren, R. Bell, C. Volinsky (2009), Matrix Factorization

Techniques for Recommender Systems, Computer, 42(8), pp. 30-37, doi:10.1109/MC.2009.263

[14] V. Kostyukov (2013), La4j – Linear Algebra for Java, Retrieved from https://github.com/vkostyukov/la4j

[15] Youtube (2013), Youtube Statistics, Retrieved from http://www.youtube.com/yt/press/statistics.html

[16] S. Lee (2013), Arirang Analyzer for Lucene/Sorl 4.4+, Retrieved from http://cafe.naver.com/korlucene

[17] G. Linden, B. Smith, J. York (2003), Amazon.com Recommendations: Item-to-Item Collaborative Filtering, IEEE Internet Computing, 7(1), pp. 76-80, doi:10.1109/MIC.2003.1167344

[18] B. Marlin (2004), Modeling User Rating Profiles For Collaborative Filtering, Advances in Neural Information Processing Systems 16, MIT Press.

[19] M. Marshall (2006), Aggregate Knowledge raises $5M from Kleiner on a roll, Venture Beat, Retrieved from http://venturebeat.com/2006/12/10/aggregate-knowledge-raises-5m-from-kleiner-on-a-roll/

[20] A. McCallum (2002), MALLET: A Machine Learning for Language Toolkit, Retrieved from http://mallet.cs.umass.edu

[21] Y. Moshfeghi, B. Piwowarski, J. Jose (2011), Handling Data Sparsity in Collaborative Filtering using Emotion and Semantic-based Features, Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 625–634,

doi:10.1145/2009916.2010001

[22] D. Ramage, S. Dumais, D. Libeling (2010), Characterizing microblogs with topic models, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp. 130-137.

[23] R. Salakhutdinov, A. Mnih, G. E. Hinton (2007), Restricted Boltzmann Machines for collaborative filtering, Proceedings of the 24th international conference on Machine learning, pp. 791–798, doi: 10.1145/1273496.1273596

[24] A. Schein, A. Popescul, L. Ungar, D. Pennock (2002), Methods and Metrics for Cold-Start Recommendations, Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 253–260, doi: 10.1145/564376.564421

[25] D. Shin, J. Lee, J. Yeon, S. Lee (2009), Context-Aware Recommendation by Aggregating User Context, IEEE Conference on Commerce and Enterprise Computing, pp. 423–430, doi: 10.1109/CEC.2009.38

[26] S. Zhang (2011), Cis Recommender, Retrieved from https://github.com/zhangsi/CisRec

[27] R. Zhou, S. Khemmarat, L. Gao (2010), The impact of YouTube recommendation system on video views, Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, pp. 404-410, doi:10.1145/1879141.1879193

[28] Y. Zhou, D. Wilkinson, R. Schreiber, R. Pan (2008), Large-

scale Parallel Collaborative Filtering for the Netflix Prize, Proceedings of the 4th international conference on Algorithmic Aspects in Information and Management, pp. 337-348, doi:10.1007/978-3-540-68880-8_32

Abstract

A Collaborative Filtering Approach to Content Recommendation using Matrix Factorization and Textual and Social Latent Variables

Joo-shik Yoon
Department of Industrial Engineering
The Graduate School
Seoul National University

Online shopping service and multimedia contents service utilize recommendation system to let users find their desired contents easily. Portal community site or social network service can also use recommendation system such as collaborative filtering to recommend various text content to user.

But in web community, text contents are rapidly produced and consumed because of their ease of use and accessibility. So user's preferences are not accumulated on single text content. Even more, user's preference of text contents are not directly visible. Therefore existing basic collaborative filtering based recommendation system is not suitable for text contents.

This paper proposes collaborative filtering based method to

effectively recommend text contents. Proposed model estimate user's preference of text content using interaction log such as content view or comment creation, and learn matrix factorization model based on those estimation value. Model's matrix factorization model use text content's topics (based on title and contents) extracted from topic model and content's publisher as latent variable. Utilizing various latent variables and bias value complement the cold start problem.

Experiments show that proposed basic score-based, rand-based model and title word term reinforcement model effectively recommend text contents, and both models outperform basic matrix factorization model based on user-item latent variable. Further experiment verified the contribution of topic, social latent variables and bias information to recommendation

Keywords : Collaborative Filtering, Matrix Factorization, Latent Variable, Text Content Recommendation, Topic Modeling
Student Number : 2012-21065

감사의 글

부족한 논문을 마무리하고 마지막으로 감사의 글을 쓰려고 하니 참 부끄럽습니다. 학위 논문을 조금 더 잘 쓸 수 있지 않았을까 하는 생각부터 시작해서, 이런 공부를 조금 더 해봤더라면, 프로젝트에 저런 걸 조금 더 실험해봤더라면 등등 많은 생각을 하게 되네요. 짧다면 짧고 길다면 긴 2년간의 연구실 생활 동안 아쉬운 점이 참 많습니다만, 그래도 참 보람차고 재미있었습니다.

2년간 거둬주시고 가르쳐주시고 먹여주신 지도 교수님, 박종헌 교수님께 무엇보다도 먼저 감사 인사를 드리고 싶습니다. 학부 2학년때 상담을 부탁 드린다고 찾아가서는 말도 안되는 상담을 했던 부끄러운 기억(아마 기억 못하시겠지요)이 있는데, 그 바보 같은 학생이 교수님 덕분에 학위도 받고 또 한번 졸업 하게 되었습니다. 교수님의 가르침이 없었다면 딱히 잘 하는 것 없는 평범한 학생에 불과했을 텐데, 교수님 덕분에 연구실에 들어와 좋은 가르침도 받고 다양한 프로젝트도 해보며 정말 많은 것을 배울 수 있었습니다. 다시 한번 감사 드립니다.

7년간 산업공학과 다니는 동안 가르침을 주셨던 모든 교수님들께도 감사 드립니다. 재미있었던 수업도 있고 정말 어렵고 힘들었던 수업도 있었지만, 그 어떤 수업이던 많은 도움이 되었고 시야를 넓히는데 큰 역할을 해주었습니다. 특히 부족한 졸업 논문의 심사위원 부탁을 흔쾌히 수락해주시고 조언해주신 조성준 교수님과 이재욱 교수님 감사합니다.

2년간 (아니면 그보다는 조금 짧은 시간 동안) 함께 한 연구실 사람들에게도 고마운 마음 전달하고자 합니다. 연구실 사람 많지도 않으니 한 사람 한 사람 따로 이야기 쓰겠습니다.

대학원 들어와서 ‘아 연구하는 사람은 이런 마음으로 해야 하는구나’라는 생각을 하게 주신 관호 형님. 2년동안 그런 자세의 발 끝만큼도 못 좇아간 듯 하지만, 앞으로 마음속에서라도 그 모습 본받고 싶습니다.

승준이형. 대학원 와서 정말 많은 도움 받았고 정말 많이 배웠습니다. 프로젝트는 물론이고 논문 쓰면서 받은 도움만 해도 셀 수 없을 정도네요. 가끔 하기 싫어하고 짜증 낸 적도 많았는데 잘 받아주셔서 감사하고, 앞으로도 다른 친구들도 많이 이끌어주시길 바랍니다. 그리고 조만간에 꼭 좋은 짝을... 아닙니다.

엄밀하게 계산하면 연구실에서 가장 대장인 범석이형. 항상 옆자리에 계셔서 듬직하고 마음 편했습니다. 논문 준비 하시면서 대규모 팀 이끄느라 정말 고생이 많으신데, 좋은 결과 기대하고 있겠습니다.

유일한 석사 석배님이자 방장님인 용석이형. 덕분에 연구실 생활이 항상 유쾌하고 즐거웠습니다. 디펜스는 물론이고 2년 동안 잡다하게 도움 안 받은게 없는 것 같네요. 결혼식 기다겠습니다.

컴퓨터 때문에 얼떨결에 유일하게 생일을 외우게 된 준석이형. 아마도 연구실에서 가장 재능러가 아닐까 생각합니다. 꼭 재능을 꽃피운 좋은 결과 얻으시길 바라고요. BK 담당 할 때 좌충우돌하면서 참 폐 많이 끼쳤었는데 죄송하고 감사합니다.

예림누나.. 아니 누님.. 아니 최쌤. 덕분에 학회도 많이 다니고 발표도 해보고 좋은 경험 많이 해보고 즐거웠습니다. 부디 건강을 위하여 술과 야식을 좀 줄이시고, 앞으로 새로운 노예를 키워 성공적인 연구를 진행하시길 빕니다.

이런 저런 상담은 물론이고 좋은 이야기 참 많이 해주신 성호형님. 얻어 먹은게 너무 많아서 그 만큼 앞으로 베풀고 살기도

힘들 것 같네요. 요즘 뵙기가 힘든데 일이든 학교든 연구 잘 되셨으면 좋겠습니다.

성민. 참 의견 안 맞고 참 많이 투닥대고 참 많이 타박받았지만, 그래도 동기 좋은 건 이유가 있다. 난 아직도 OA 수업에서 네가 손으로 스케쥴을 짜냈을때의 그 희열을 잊지 못하고 있다. 업적 축하하고 부디 그 악마의 재능 조심해서 잘 다루길.

선배이자 동기이고 무엇보다도 좋은 친구인 남연아. 졸업이니 뭐니 하면서 힘들었던 마지막 학기에 함께 할 수 있었던 사람이 있다는 것 만으로도 참 많은 위안이 되었다. 헛소리 잘 받아줘서 고맙게 생각하고, 앞으로도 꼭 1분 고민하길 바란다.

연구실에 와서야 그 실체를 알게 된 인범이. 너의 놀라운 언어 유희와 유우-머를 들을때마다 정말 깜짝깜짝 놀라곤 한다. 몸 걱정할 일 없도록 앞으로 운동도 열심히 하길 바란다.

막판에 최쌤과 나의 갈굼을 받아주느라 고생한 승욱아. 이젠 일요일 밤에 만날 수 없게 되어 아쉬움을 금할 수 없다. 앞으로도 최쌤 잘 도와주고, 여분의 베X카 1통을 준비해놓길 바란다.

연구실의 우렁각시인 희웅이형에게는 고맙고 미안한 마음. 형 부디 내년에는 청소왕 타이틀을 내려놓고 후계자를 양성하길 바라겠습니다.

잘생긴 보형아. 전자레인지 그만 돌리고 건강식 했으면 좋겠다. 내가 없더라도 피자는 도미노피자로, 도우는 꼭 씬으로 하길 바라고, 담당자가 바뀌게 되면 꼭 전달해주길.

성실한 재석이형. 앞으로 열심히 잘 할 것 같아서 사실 걱정은 별로 안되고, 다만 아마도 수업 조교하게 될 텐데 그게 좀 걱정이긴 하네. 고생하겠지만 잘 해낼거라 믿음.

근 10년을 똑같이 해오던 불량 기도를 최신으로 업데이트 하게

만들어준 윤정이. 디펜스때 많이 도와줘서 정말 고맙고, 남은 1년 반도 잘 마치길 바랄게.

언급한 사람들 외에도 2년 동안 같이 연구실에서 함께하지는 못했지만 많은 조언과 도움을 주신 연구실 선배님들께도 감사드립니다.

연구실 외에도 2년간 석사 생활을 하면서 도움을 주신 많은 분들에게도 감사 인사를 전하고 싶습니다. 가장 먼저 많은 친구들과 선후배님들. 항상 고맙습니다. 앞으로도 오랫동안 많이 고마워하고 싶고, 고마운 사람이 되고 싶습니다.

와플스튜디오는 제 인생의 방향을 크게 바꿔놓았고 아마 와플이 아니었으면 지금 이 자리에 있지 않았을지도 모르겠습니다. 정말 고맙고 앞으로도 고마울 일이 많을 것 같습니다.

석사 생활 내내 많은 즐거움을 준 실론극회 사람들. 일일이 이름은 언급 못하겠지만 참 좋은 추억이 되었습니다. 앞으로도 협곡에서 계속 만나게 되겠지요. 또 봅시다.

마지막으로 문과도 안가고 뜬금없이 이과에 가고, 고시도 안보겠다 하고, 군대는 안가고 대학원에 가겠다 하고, 대학원은 가면서 박사도 안하겠다 하는 청개구리 같은 아들임에도, 어떤 길을 선택하던 조언을 아끼지 않고 끝까지 믿어주시는 아버지와 어머니, 그리고 제대를 앞두고 미래에 대한 걱정에 마음고생 하고 있을 동생에게 진심으로 감사한다는 말과 사랑한다는 말을 전하고 싶습니다.