협업 필터링 및 세션 키워드 추천 알고리즘 기반 데이터 활용도 향상 시스템에 관한 연구

박성민, 최규홍, 이동욱, 천승태
㈜데이터스트림즈

{sminpark, gyuhong.choi, dwlee, stchun}@datastreams.co.kr

A study on data utilization enhancement system based on collaborative filtering and session keyword recommendation algorithm

Park Seong Min, Choi Gyu Hong, Lee Dong Wook, Chun Seung Tae
DataStreams Corp.

요 약

본 논문에서는 데이터 활용 시스템에서 사용자의 데이터 활용도를 향상 시킬 수 있도록 키워드를 추천하는 알고리즘과 이를 마이크로서비스 아키텍처로 구현한 결과물을 설명한다. 추천 알고리즘은 사용자 기반 협업 필터링, 키워드 기반 협업 필터링과 세션 키워드 패턴 3가지 알고리즘으로 구성된다. 각 알고리즘의 추천 결과가 다음 알고리즘의 후보가 되는 하이브리드 방식을 적용하였다. 전체 추천 알고리즘을 거쳐 최종적으로 사용자에게 적절한 키워드를 추천한다. 이후 추천 알고리즘을 기반으로 마이크로서비스 아키텍처로 구성하여 REST API로 기존 시스템과 연계하여 활용한 방법을 설명한다. 이전 시스템에서 사용되던 모델 및 알고리즘과 비교하여 성능을 평가하였고, 이전 모델 대비 10%의 성능향상을 보였다.

Ⅰ. 서 론

데이터의 활용이 비즈니스 성패를 결정하는 현대 기업환경에서 데이터 거버넌스[1]의 중요성이 증대되고 있다. 데이터 거버넌스 시스템은 사용자가 필요로 하는 데이터에 대한 접근성을 향상시킬 수 있는 방법을 제공해야 한다. 데이터 카탈로그[2], 데이터맵[3] 등이 이러한 목표를 달성하기 위해 연구되고 있으며, 사용자는 검색을 통해 이를 활용하게 된다.

따라서 사용자가 필요한 데이터에 접근할 수 있도록 검색과정에서 키워드를 추천하여 데이터의 활용도를 높이는 것은 매우 중요하다. 기존 연구에서는 이러한 문제를 해결하고자 사용자 기반 협업 필터링[4], 아이템 기반 협업 필터링[5]을 단독으로 적용하여 추천하는 연구가 이루어졌다. 하지만 기존 협업 필터링 기반의 추천은 사용자의 만족도 측면에서 한계를 보인다. 이에 따라 더 정확한 추천을 제공하여 시스템 활용도를 높이는 것에 대한 필요성은 계속해서 증가하고 있다. 본 논문에서는 더 정밀한 추천을 위해 사용자, 아이템 기반 협업 필터링과 세션 키워드 패턴을 하이브리드 방식[6]으로 적용하는 것을 제안하고, 이를 구현한 시스템을 제안한다.

Ⅱ. 본론

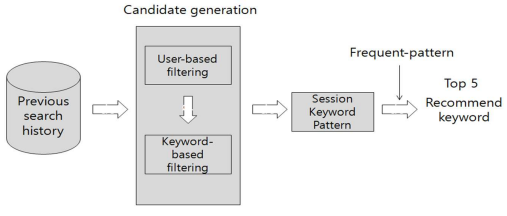

그림 1 협업 필터링 및 세션 키워드 기반 추천

제안 시스템은 <그림 1>과 같이 사용자 기반 협업 필터링을 적용한 후, 그 결과를 키워드 기반 협업 필터링의 입력으로 사용하여 추천 후보 패턴을 생성한다. 생성된 후보패턴을 세션 키워드 패턴의 입력으로 하여 최종적인 추천 결과를 제시한다.

Ⅱ-1. 사용자 기반 협업 필터링

사용자 기반 협업 필터링은 검색 기록을 기반으로 유사한 검색 기록을 가진 사용자들을 추천하도록 하는 방법이다. 먼저 검색 기록을 각 사용자의 키워드별 검색 횟수를 나타낸 형태로 변환한다 <표 1>.

표 1. 사용자 별 키워드 검색 이력

	키워드1	*키워드2*	...	*키워드 n*
사용자 1	*10*	*5*	...	*2*
...	...	...	...	...
사용자 i	*3*	*2*	...	*1*

이후 각 사용자 행의 값으로 사용자 간의 유사도를 계산한다. 유사도 계산 시 코사인 유사도에 보정된 상관계수를 적용하는 피어슨 유사도를 사용한다. 수식(1)은 특정 사용자 행 X,Y간의 피어슨 유사도를 구하는 식이다.

$$r_{XY} = \frac{\sum_{i}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i}^{n}(Y_i - \overline{Y})^2}} \quad (1)$$

이와 같이 각 행을 $X_{1,2,3,\ldots,n}$ 으로 하는 사용자 행 X집합에서 모든 사용자 쌍에 대하여 피어슨 유사도를 계산한 뒤 유사도 값을 사용자 기반 협업 필터링 테이블에 저장한다.

Ⅱ-2. 키워드 기반 협업 필터링

키워드 기반 협업 필터링은 유사한 키워드들을 추천하도록 하는 방법으로 검색 기록을 기반으로 키워드 간 유사도를 계산한다. 검색 기록에서 키워드를 검색한 사용자 패턴(<표 2>)을 산출하여, 유사도 계산에 사용한다.

표 2. 키워드 별 검색 이력 (사용자)

	사용자1	*사용자2*	...	*사용자 i*
키워드 1	*10*	*2*	...	*3*
...	...	...	...	...
키워드 k	*2*	*1*	...	*1*

유사한 키워드를 찾기 위해 자카드 유사도를 사용한다. 키워드 K_i, K_j가 있을 때 자카드 유사도를 계산하는 식은 다음과 같다.

$$J(K_i, K_j) = \frac{|K_i \cap K_j|}{|K_i \cup K_j|} = \frac{|K_i \cap K_j|}{|K_i| + |K_j| - |K_i \cap K_j|} \quad (2)$$

이와 같이 전체 키워드 간 유사도 계산을 위해 각 행을 $K_{1,2,3,\ldots,n}$ 으로 하는 키워드 행 K집합에서 K_i , K_j 키워드간의 유사도를 계산한다. 계산한 유사도는 키워드 기반 협업 필터링 테이블에 저장한다. 사용자 기반 협업 필터링 테이블과 키워드 기반 협업 필터링 테이블은 다음과 같이 활용된다.

키워드를 검색한 사용자에 대하여 다음 키워드 추천을 제공하기 위해 사용자 협업 필터링 테이블에서 유사도가 높은 사용자를 찾아낸다. 높은 유사도의 사용자의 검색 키워드 목록 중 사용자가 검색한 키워드와 유사도가 높은 키워드를 키워드 기반 협업 필터링 테이블에서 찾아낸다. 높은 유사도의 키워드들은 세션 키워드 패턴 단계로 넘겨진다.

Ⅱ-3. 세션 키워드 패턴

유사도 키워드 패턴을 생성하기 위해 추천 대상 사용자가 검색한 키워드와 사용자 기반 협업 필터링과 키워드 기반 협업 필터링까지 거쳐 나온 키워드 후보들을 묶는다.

위와 같은 방법으로 유사도 키워드 패턴을 키워드 후보들의 개수만큼 생성한다. 세션 키워드 패턴은 세션 단위별 키워드들의 나열이다. 다음 <표 3>과 같이 시스템 검색기록에서 세션별 검색 이력을 생성한다.

표 3. 세션별 검색 이력

	키워드	*시간*
세션 1	*키워드 1*	2023.04.21 13:58:42
	키워드 2	2023.04.21 14:00:01
	키워드 3	2023.04.21 14:02:01
세션 2	*키워드 4*	2023.04.21 17:58:42
	키워드 5	2023.04.21 18:01:23
	키워드 6	2023.04.21 18:03:11
...	...	...
세션 k	*키워드k*	2023.05.01 10:23:51

<표 3>으로부터 세션별 키워드 패턴을 다음과 같이 생성한다.

세션1 키워드 패턴: (키워드1,키워드2,키워드3)
세션2 키워드 패턴: (키워드4,키워드5,키워드6)

*k*개의 세션 키워드 패턴을 순회하며 각 세션 키워드 패턴 내에서 유사도 키워드 패턴의 키워드들이 모두 등장하면 유사도 키워드 패턴의 추천 가중치를 증가시킨다. 이 과정을 모든 유사도 키워드 패턴에 대하여 실시한다. 이후 추천 가중치가 높은 순서대로 유사도 키워드 패턴을 정렬한다. 상위 5개 유사도 키워드 패턴의 사용자 협업 필터링과 키워드 기반 협업 필터링을 통해 나온 키워드 후보를 추천 키워드로 제공한다.

Ⅲ. 시스템 구현

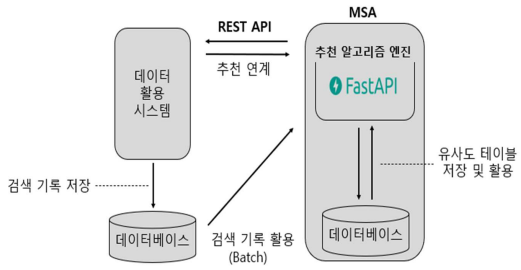

그림 2 추천 시스템 연계 구성도

본 논문의 추천 알고리즘을 마이크로 서비스 아키텍처(MSA)[7]로 구성하였다. 마이크로 서비스 아키텍처 내부 추천 알고리즘 엔진의 프레임워크는 Starlette[8]를 사용해 빠른 성능을 지원하는 Python FastAPI[9]를 사용하였다. 추천 알고리즘을 FastAPI 엔진에서 REST API로 설계하여 데이터 활용 시스템과 통신하도록 하였다. 이와 같은 마이크로 서비스 아키텍처 추천 시스템을 서버에서 편리하게 빌드하고 실행할 수 있도록 Dockerfile, Shell script 와 Command guide를 제공하였다. <그림 2>는 마이크로 서비스 아키텍처 추천시스템이 데이터 활용 시스템과 REST API로 통신하는 전체 시스템의 연계 구성도이다.

Ⅳ. 평가

데이터 거버넌스 시스템의 검색 기록을 활용하여 추천을 하고 타 모델과 비교하여 평가하였다. 정량 평가는 다음과 같이 실시하였다. 먼저 본 논문의 추천 알고리즘 모델을 사용하였다. 비교 모델로 평가업체를 통해 시행한 성능 평가에서 90%의 정확도를 달성했던 Bidirectional-lstm[10] 모델로 세션 키워드 패턴과 패턴 다음에 입력 되는 키워드를 레이블로 하여 학습시켜 모델을 만들었다. 또한 기존의 협업 필터링을 각각 적용한 모델과 본 논문의 알고리즘을 평가하였다. 평가지표로는 Hit@10를 사용하였다.

표 4. 모델별 평가지표

모델 \ 지표	Hit@10
사용자 기반 협업 필터링	0.3041
키워드 기반 협업 필터링	0.3152
Bidirectional-LSTM	0.3439
collaborative filtering and session data hybrid recommendation	0.3812

이후 정성평가 결과에서 본 논문에서 적용한 추천 알고리즘 결과가 타 모델의 결과보다 유의미한 결과를 나타내는 것을 확인하였다.

Ⅴ. 결론

본 논문에서는 추천 시 협업 필터링을 단순하게 적용하지 않고 다른 알고리즘과 하이브리드한 방식으로 적용하여 더 나은 추천을 제공하게 했다. 이를 위해 사용자 협업 필터링과 키워드 기반 협업 필터링으로 후보군을 생성하고 세션 키워드 패턴을 활용하여 하이브리드 방식으로 추천하는 알고리즘 연구를 진행하였다. 연구를 통한 추천 알고리즘의 결과가 협업 필터링 모델, Bidirectional-LSTM 모델과 비교하였을 때 더 높은 성능을 나타내었다. 향후 연구에서는 이러한 추천 알고리즘을 적용한 시스템을 확장성 및 가용성 등의 장점이 있는 Kubernetes 시스템에 적용하기 위한 연구를 진행한다.

ACKNOWLEDGMENT

본 연구는 2022년도 중소벤처기업부의 기술개발사업 지원에 의한 연구임 [RS-2022-00140586]

참 고 문 헌

[1]"데이터거버넌스",위키백과,2022년 2월 28일 ,https://ko.wikipedia.org/wiki/%EB%8D%B0%EC%9D%B4%ED%84%B0_%EA%B1%B0%EB%B2%84%EB%84%8C%EC%8A%A4

[2] "tibco cloud",tibco,n.d, https://www.tibco.com/ko/reference-center/what-is-a-data-catalog

[3]"통합데이터지도 FAQ",통합데이터지도,.n.d, http://www.bigdata-map.kr/board/faq#

[4]Zhi-Dan Zhao, Ming Sheng Shang, "User-Based Collaborative-Filtering Recommendation Algorithms on Hadoop", 2010.

[5]Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms", 2001, https://www.ra.ethz.ch/cdstore/www10/papers/pdf/p519.pdf

[6]손지은 외 , "추천시스템 기법 연구동향 분석," Journal of the Korean Institute of Industrial Engineer Vol. 41 No. 2, Apr. 2015, pp. 185-208

[7] "마이크로서비스",위키백과,2022년 11월 21일 ,https://ko.wikipedia.org/wiki/%EB%A7%88%EC%9D%B4%ED%81%AC%EB%A1%9C%EC%84%9C%EB%B9%84%EC%8A%A4

[8] "Starlette",Starlette,n.d, https://www.starlette.io/

[9] "FastAPI",FastAPI,n.d, https://fastapi.tiangolo.com/ko/

[10] Mike Schuster,Kuldip K. Paliwal. "Bidirectional recurrent neural networks",1997.