

기능 조합 하이브리드 시스템 : 매트릭스 분해와 결합된 컨텐츠 기반

응웬 시 띨*, 류 호 균*, 최 은 진*, 김 광 용^o

Feature Combination Hybrid System : Content-Base Combined With Matrix Factorizations

Si-Thin Nguyen*, Ho-Gun Rou*, Eun-Jin Choi*, Gwang-Yong Gim^o

요 약

권장 시스템에서 모든 기본 모델은 특히 다양한 데이터 소스를 사용할 수 있는 경우 다소 제한적으로 보인다. 하이브리드 추천 시스템은 이러한 가능성을 탐색하도록 설계되었다. 하이브리드 시스템의 전통적인 연구는 새로운 알고리즘의 개발에 한정되었고, 다른 접근법을 결합하는 연구는 거의 없었다. 또한 빅 데이터 시대에는 데이터 스트림과 증분 모델 평가가 필요하다. 본 연구에서는 컨텐츠 기반과 매트릭스 인수 분해를 기반으로 기능 조합 하이브리드 시스템을 제안한다. 또한 스트리밍 데이터 환경에 적용할 수 있는 추천 시스템에 대한 사전 평가 프로토콜도 보여준다. 본 연구에서 제안하는 알고리즘은 다른 최신 모델과 비교하여 정확성과 시간 업데이트 모두에서 더 나은 결과를 얻는다.

키워드 : 한글 키워드 추천 시스템, 행렬 분해, 확률적 경사 하강, 하이브리드 시스템, 협업 필터링

Key Words : Recommendation System, Matrix Factorization, Stochastic Gradient Decent, Hybrid System, Collaboration Filter

ABSTRACT

In recommendation system, all basis models seem rather restrictive in isolation, especially when various sources of data are available. Hybrid recommender systems have been designed to explore these possibilities. Conventional research in hybrid system has been limited to the development of new algorithms and little research combines the different approaches. Moreover, data stream and evaluating incremental model are necessary in big data era. In this research, we propose a feature combination hybrid system based on conjunction content-base and Matrix Factorization. In addition, we also show a prequential evaluation protocol for recommender systems, applicable for streaming data environments. Comparing with other state-of-the-art models, our algorithm gains better results in both accuracy and time update.

◆ First Author : Soongsil University, Department of IT Policy Management, thinns168@gmail.com

◦ Corresponding Author : Soongsil University, Department of Business Administration, gygim@ssu.ac.kr

* Soongsil University, Department of IT Policy Management, har1203@soongsil.ac.kr; chansol@kosmes.or.kr

논문번호 : 20190-238-0-SE, Received October 22, 2019; Revised November 4, 2019; Accepted November 16, 2019

I. 서론

추천 시스템은 일련의 항목 (예 : 책, 노래, 영화, 농담, 가제트, 웹 사이트, 응용 프로그램, 여행 목적지 및 전자 학습 자료)에 대한 사용자 우선순위에 대한 정보를 집계한다¹¹. 정보는 명시적으로 (일반적으로 사용자의 등급을 수집하여) 또는 내재적으로 (일반적으로 다운로드를 한 응용 프로그램, 들었던 노래, 읽은 책 및 방문한 웹 사이트와 같은 사용자의 행동을 제어하여) 얻을 수 있다^{2,41}.

추천 평가에 대해서는 오프라인 평가가 추천자 성능을 비교하고 다양한 알고리즘의 수행 방식에 대한 이해를 얻는 데 효과적이지만 다양한 방법으로 제한된다⁵⁻⁷. 시간 기반 방법은 이를 다소 커밋(Commit) 하지만 오프라인 평가에서 사용자가 찾지 못한 적절한 항목을 추천하는 추천 기능을 측정하는 데 비효율적이다⁸.

온라인 평가는 사용자 유지 또는 장기 이익과 같은 포괄적인 시스템 목표를 직접 측정한다고 가정한다. 그러나 이러한 속성을 변경하는 것은 독립적으로 쉽지 않으며 온라인 시험을 통해 여러 알고리즘을 비교하는 것은 비용이 많이 들기 때문에 이러한 관계에 대한 철저한 이해를 어렵게 만들 수 있다⁹.

일반적으로 콘텐츠 기반 (Content-Base-CB), 공동 작업 (collaborative filtering -CF) 및 지식 기반의 세 가지 추천 방법이 있다. 이러한 방법은 서로 다른 데이터 소스를 사용하며 각각의 장단점이 있다. 예를 들어, 지식 기반 시스템은 등급이 필요하지 않기 때문에 콘텐츠 기반 또는 협업 시스템보다 콜드 스타트 (cold-start) 문제를 훨씬 더 쉽게 해결할 수 있다. 반면에, 과거 데이터로부터 지속해서 개인화를 사용한다는 점에서 콘텐츠 기반 및 협업 시스템보다 약하다. 다른 사용자가 지식 기반 대화식 인터페이스에서 동일한 요구 사항과 데이터를 입력하면 이에 대한 같은 결과를 얻을 수 있다. 하이브리드 추천 시스템은 이러한 가능성을 탐구할 수 있도록 설계된 것이다. 그러나 하이브리드 시스템에 대한 선행연구가 부족하여, 본 연구에서는 CB와 인 크리에이션 매트릭스 요인화 (Incremental Matrix Factorization)을 결합한 항목 평가 예측 문제에 중점을 둔다.

첫째, CB와 CF는 서로 다른 데이터 소스를 가지고 있다. CB의 항목에 대한 텍스트 설명과 CF의 커뮤니티 등급을 함께 사용하면 교육 데이터를 더 빠르게 작성할 수 있다.

두 번째 거의 온라인 시스템에는 구현된 수치 등급

기능이 있다. 증분 매트릭스 인자의 정격 예측 알고리즘은 유익성에 대한 거의 연구에서 고려된다. 셋째, 발표된 연구의 대다수가 기본 모델의 알고리즘에 초점을 맞추고 있지만, 하이브리드 시스템과 평가 스트림 환경의 특정 속성과 관련된 일부 이슈는 발견되지 않았다. 그리고 이 연구는 다음의 두 가지 연구 문제에 초점을 맞추고 있다.

1) 콘텐츠 기반과 모델 기반 협업 필터링의 조합에 하이브리드 시스템이 적합한가?

2) 기능 조합 하이브리드 시스템이 다른 모델에 비해 정확도 및 시간 업데이트에서 더 나은 결과를 얻는가?

우리의 모델은 최신 알고리즘과 비교했을 때 정확성과 예측 등급 향상 모두에서 더 나은 결과를 얻는다.

본 문서의 나머지 부분은 다음과 같이 설계된다. 추천자 시스템에서 기본 모델의 간략한 개요를 보여준다. 또한 본 문서의 진화를 설명하며, 섹션 2에 이러한 시스템에 대한 원래의 분류를 제공한다. 섹션 3은 증분 학습 알고리즘과 그에 대한 관련 작업에 대한 자세한 논의를 제공한다. 우리는 4절에서 제안된 모델에 대해 구체적으로 설명한다. 자료와 평가 문제는 제5절에 명시되어 있다. 섹션 7에서 간단한 결론을 내리기 전에, 우리는 섹션 6의 결과에 대해 토론할 뿐만 아니라 사전 필수 평가 프로토콜에 대해 제시한다.

II. 추천 시스템의 기본 모델

2.1 추천 시스템에서의 모델 개발

추천 시스템의 기본 모델은 (i) 텍스트 프로파일 또는 관련 키워드와 같은 사용자 및 항목에 대한 속성 정보, (ii) 등급 또는 구매 행위와 같은 사용자-항목 상호 작용과 같은 두 가지 종류의 데이터로 작업한다. 전자를 사용하는 방법은 콘텐츠 기반 필터링 방법, 후자를 사용하는 방법은 협업 필터링 추천 방법이라고 한다¹⁰⁻¹³. 콘텐츠 기반 시스템은 대부분의 경우 등급 매트릭스를 사용하지만, 모델은 일반적으로 모든 사용자의 등급이 아닌 단일 사용자의 등급에 중점을 둔다.

Breese 등은 CF 알고리즘을 메모리 기반 방법 또는 NCF (neighbourhood-base CF) 및 모델 기반 방법의 두 그룹으로 분류할 수 있다¹⁴. 메모리 기반 CF는 항목 또는 사용자 간의 유사성을 사용하여 권장 사항을 만든다^{15,16}. 이 방법은 채택하기 쉽고 매우 효과적이기 때문에 널리 사용되지만 권장 시스템의 규모가 발전함에 따라 유사성 계산이 점점 어려워지고 있다.

또한, 데이터 희소성이 높으면 이 방법의 성능이 제한된다.

위에서 언급 한 문제를 해결하기 위해 잠재 시맨틱 모델, 회귀 기반 모델, 베이저란 모델, 군집 모델 및 행렬 인수 분해 모델과 같은 많은 모델 기반 권장 알고리즘이 제안되었다^[17-21]. 여러 CF 기술 중에서 매트릭스 인수 분해가 가장 일반적인 방법이다. 이 방법은 항목과 사용자를 유사한 차원의 벡터에 매핑하여 항목 또는 사용자의 잠재 된 기능을 표시한다. 이 방법의 대표적인 작업에는 NPCA (Nonparametric Probabilistic Principal Component Analysis), PMF (Probabilistic Matrix Factorization) 및 SVD (Singular Value Decomposition) 가 있^[22-24]. 그러나 행렬 인수 분해 방법으로 학습 한 잠재 특징은 주로 등급 행렬이 매우 희박한 경우 일반적으로 크게 효과적이지 않다.

잠재 요인 모델에 대한 가장 성공적인 연구 중 일부는 행렬 인수 분해를 기반으로 한다. 원래의 형태로, 행렬 인수 분해는 항목 평가 패턴에서 해석되는 요소 벡터로 사용자와 항목을 모두 특성화한다. 사용자와 항목 요소 사이의 높은 상관관계는 권장 사항으로 이어진다. 따라서 이러한 방법은 예측 정확도와 우수한 확장성을 통합함으로써 최근에 일반화되었다. 또한 다양한 실제 상황을 모델링 할 때 많은 유연성을 제한한다^[25].

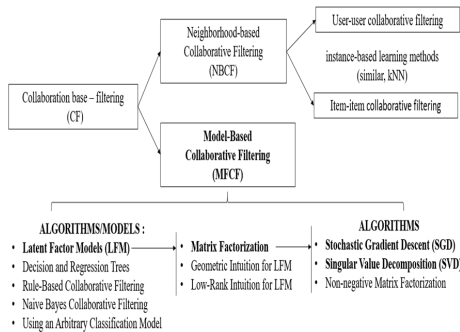


그림 1. 협업 기반 필터링 알고리즘 개발
Fig. 1. The development of collaboration base-filtering algorithms

2.2 CF 를위한 Matrix Factorization

초기 행렬 인수 분해 방법론 등급에서 행렬 $R(m \times n)$ 은 대략 행렬 $W(m \times k)$ 와 행렬 $X(n \times k)$ 로 $R \approx WX^T$ 로 인수 분해된다.

등급 예측은 사용자 $U = \{u_1, u_2, \dots\}$ 및 항목 $I = \{i_1, i_2, \dots\}$ 에 대한 회귀 모델로 알려져 있다. 여기서 대상 함수 $y: U \times I \rightarrow R$ 을 계산해야한다. 타겟 기

능은 예를 들어 $y(u, i)$ 는 사용자 u 의 항목 i 에 대한 등급이다^[1].

매트릭스 인수 분해의 한 가지 장점은 추가 정보의 융합을 허용한다는 것이다. 추천 시스템은 명시적 피드백을 사용할 수 없는 경우 암시적 피드백을 사용하여 사용자 기본 설정을 유추 할 수 있으며, 구매 내역, 검색 패턴, 인터넷 사용 기록 또는 마우스 움직임에 포함하여 사용자 행동에 따라 간접적으로 의견을 제시한다. 암시적 피드백은 일반적으로 이벤트의 유무를 나타내므로 일반적으로 밀도가 높은 매트릭스로 표시된다^[25].

CF에 대한 행렬 분해는 정보 검색 영역에서 사용되는 대규모 텍스트 문서 모음을 식별하는 데 널리 사용되는 기술인 LSI (Latent Semantic Indexing) 에서 비롯되었다. LSI는 큰 문서 용어 행렬의 SVD를 수행한다. 이와 유사하게, CF 문제에서 동일한 기술이 사용자-항목 등급 매트릭스에 적용될 수 있으며, 아이템 및 사용자 모두에게 인기 있는 잠재 특징 공간을 발견한다. 대안으로 최적화 방법은 희소 등급 행렬을 해결하기 위해 제안되었다^[26-30].

앞에서 언급했듯이 매트릭스 요인화 (Matrix Factorization) 모델은 항목과 사용자를 차원 차원의 공동 잠재 요인 공간 f에 매핑하여 사용자 항목 상호 작용이 해당 공간의 내부 제품으로 계산되도록 한다. 따라서 각각의 사용자 u 는 벡터 $W_u \in R^f$ 와 연결되고 각각의 아이템 i 는 벡터 $X_i \in R^f$ 와 연결된다. 예를 들어, 사용자 u , W_u 의 요소는 사용자가 핵심 요인에서 높은 항목, 즉 부정적 또는 긍정적 인 항목의 흥분 정도를 추정한다. 또한 항목 i , X_i 의 요소는 항목이 해당 요소를 소유하는 정도를 음수 또는 양수로 추정한다. 결과적인 내적 제품인 $W_u X_i^T$ 는, 품목 i 와 사용자 u 간의 상호 작용, 즉 품목의 기능에 대한 사용자의 전반적인 관심을 포착한다. 이것은 사용자 u 의 평가 항목 i 에 근사하며, 이는 R_{ui} 로 표시되며 추정치로 이어진다^[25].

$$R_{u,i} = W_u X_i^T \tag{1}$$

R 의 알려진 값과 해당 예측 등급에 대해 $L2$ 정규화 된 제품 오차를 최소화하여 교육을 수행할 수 있다.

$$\min_{W, X} \sum_{(u,i) \in D} (R_{u,i} - W_u \cdot X_i^T)^2 + \lambda (\|W_u\|^2 + \|X_i\|^2) \tag{2}$$

위의 방정식에서 우리는 D 가 정규화의 양을 포착하고 등급 값이 알려진 매개 변수인 λ 가 사용자 항목 조합 세트임을 식별해야 한다. 정규화 용어 $\lambda(\|W_u\|^{2+} \|X_i\|^2)$ 는 과적합을 피하는 데 사용되는 중요한 의미를 갖는다. 이 용어는 일반적으로 낮은 일반화 전력으로 지나치게 복잡한 모델을 초래하는 높은 크기의 매개 변수에 불이익을 줄 수 있다. 이 최적화 문제를 해결할 수 있는 가장 성공적인 방법은 SGD (Stochastic Gradient Descent)와 ALS (Alternating Least Squares)이다^[27,28].

Bell and Koren은 행렬 계수의 모델 클래스에 대한 모든 사용자 요소와 모든 항목 요소를 최적화하는 것을 변경하는 ALS 방법을 제안했다^[31,32]. SGD 알고리즘은 표준 ALS의 복잡한 문제가 존재하기 때문에 ALS보다 추천 시스템 문헌에서 더 많이 사용하고 있다. SGD 기반 최적화는 일반적으로 모델 정확도와 런타임 성능 측면에서 최소 데이터를 사용할 때 ALS보다 성능이 우수하다고 가정되었다^[8,18].

2.3 하이브리드 시스템

이전 부분에서 세 가지 종류의 추천 시스템에 대해 논의했다. 공동 작업 방법은 권장 사항을 만들기 위해 사용자 커뮤니티의 등급을 사용하는 반면, 콘텐츠 기반 방법은 속성 중심 항목 설명과 함께 단일 사용자의 등급을 사용하여 권장 사항을 만든다. 일부 추천 시스템은 이러한 다양한 측면을 결합하여 하이브리드 시스템을 만든다. 하이브리드 시스템은 다양한 유형의 추천 시스템의 장점을 결합하여 다양한 설정에서보다 강력하게 수행 할 수 있는 기술을 만들 수 있다. 그림 2는 하이브리드 추천 시스템을 만드는 세 가지 주요 방법을 보여준다.

1) 앙상블 설계: 이 설계에서는 상용 알고리즘의 결과가 보다 강력한 출력으로 결합 된다. 예를 들어, 콘텐츠 기반 및 협업 추천자의 평가 출력을 단일 출력으로 결합할 수 있다. 조합 프로세스에 사용되는 특정 방법론 측면에서 상당한 차이가 있다. 작업의 기본 원리는 클러스터링, 분류 및 이상치 분석과 같은 많은 데이터 마이닝 응용 프로그램의 앙상블 설계 방법과 크게 다르지 않다.

2) 모 놀리 식 디자인: 이 경우 다양한 데이터 유형을 사용하여 통합 추천 알고리즘이 생성된다. 알고리즘의 다양한 부분들 (예를 들어, 콘텐츠 및 협력) 사이에 명확한 구별이 존재하지 않을 수 있다. 다른 경우에는 기존 협업 또는 콘텐츠 기반 추천 알고리즘을 수정하여 콘텐츠 기반 단계와 협업 단계 사이에 명확한

구분이 있는 경우에도 전체 접근 방식에서 사용되도록 해야 할 수도 있다. 따라서 이 접근 방식은 다양한 데이터 원본을 보다 밀접하게 통합하는 경향이 있으며 개별 구성 요소를 기성품 블랙박스로 쉽게 볼 수 없다.

3) 혼동 시스템: 앙상블과 마찬가지로 이러한 시스템은 여러 권장 알고리즘을 블랙박스로 사용하지만 다양한 시스템에서 권장하는 항목은 나란히 표시된다. 예를 들어, 텔레비전 프로그램은 여러 항목을 포함하는 복합 엔티티이다. 단일 항목의 권장 사항을 개별적으로 보는 것은 의미가 없다. 오히려 권장 사항을 작성하는 항목의 조합이다.

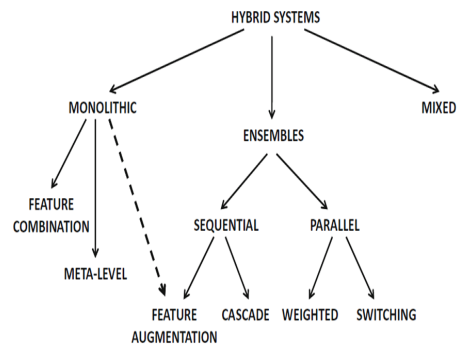


그림 2. 하이브리드 시스템의 분류
Fig. 2. The taxonomy of hybrid systems

III. 증분 학습 알고리즘

3.1 증분 학습

실제 시스템에서 등급 데이터를 살펴보면 데이터 스트림으로 접근해야 한다. 등급 데이터가 지속적으로 생성되고 있으며 새로운 등급의 도착순서 또는 데이터 속도를 제어 할 수 없다^[33]. 이상적으로, 데이터 스트림에서 학습하는 알고리즘은 예측 모델을 점진적으로 유지하여 단일 통과 데이터를 요청한다.

증분 CF에 대한 선행연구는 많지 않다. 증분 이웃 기반 CF는 노출되었고 이전 연구에서 다른 연구가 수행되었다^[34-36]. 증분 행렬 인수 분해의 경우, 작성자가 폴드-인 (Fold-In) 방법을 사용하여 인수 행렬을 증분 업데이트하는 첫 번째 솔루션이 제안된다^[37,38]. 선택적 샘플링 전략을 사용하는 데이터 순위를 지정하는 증분 알고리즘이 가정되어 있다^[39]. 배치 프로세스의 단순화 된 버전을 사용하여 사용자 요소를 업데이트하는 증분 방법이 노출된다^[30]. SGD를 사용하는 두 가지 증분 방법이 서로 비교된다^[40].

본 연구에서는 앞서 언급한 선행연구와의 차별점은 첫째, 우리는 평가 항목 예측 문제를 해결하고 둘째, 제안된 평가 프로토콜이 알고리즘의 진화하는 정확성을 평가하기 위해 순차적인 방법론을 사용하기 때문에 기존의 선행연구와 차별화된다고 할 수 있다.

3.2 SGD 연산

SGD 알고리즘은 그림 3에 표시되어있다. 등급 행렬의 관측 된 항목에서 주요 관측은 업데이트가 오류의 선형 함수라는 것이다. 업데이트는 모든 항목이 아닌 개별 관측 항목의 오류와 관련된 작은 부분으로 분할 하여 다른 방식으로 수행할 수 있다.

SGD는 $\langle user, item, rating \rangle$ 형식의 튜플로 구성된 훈련 데이터 세트를 고려할 때 일부 중지 기준이 충족 될 때까지 데이터 세트를 통해 여러 번의 패스를 수행한다. 일반적으로 최대 반복 횟수 또는 수렴 범위이다. 반복 할 때마다 SGD는 알려진 모든 등급 R_{ui} 를 스윙하고 해당 행 W_u 및 X_i^T 를 업데이트하여 학습률 또는 단계 크기로 알려진 $\eta \leq 1$ 의 계수로 오류의 기울기의 반대 방향으로 수정한다. 해당하는 오류는 알려진 각 등급에 대해 $err_{ui} = R_{ui} - \widehat{R}_{ui}$ 로 계산되며 다음 업데이트 작업이 실행된다.

$$\begin{aligned} W_u &\leftarrow W_u + \eta(err_{ui}X_i - \lambda W_u) \\ X_i &\leftarrow X_i + \eta(err_{ui}W_u - \lambda X_i) \end{aligned} \quad (3)$$

또는 $\phi = (W; X)$,인 경우 SGD는 대조적으로 각 학습 예 $x^{(i)}$ 및 레이블 $y^{(i)}$ 에 대한 매개 변수 업데이트를 실행한다.

$$\phi = \phi - \eta \cdot \nabla \phi L(\phi; x^{(i)}; y^{(i)}) \quad (4)$$

훈련 세트에서 SGD의 명백한 이점 중 하나는 알려진 등급의 수에 따라 복잡성이 증가하고 실제로 R 의 희소성이 높은 이점을 얻을 수 있다는 것이다.

랜덤 트레이닝 포인트 $(i, j) \in D$ 를 선택한 후 W_i 와 X_j 만 업데이트해야하며 다른 트레이닝 포인트에 대한 요소는 업데이트 할 필요가 없다. 이 계산 비용의 절감은 전체 손실을 지역 손실의 합으로 표현한 결과이다^[41].

위와 같은 이유로 SGD는 종종 순차적 경사 하강 또는 온라인 학습으로 간주 된다. 여러 지역 손실이 평균인 배치 버전도 적합하지만 실제로는 성능이 낮다.

정확한 그래디언트 (GD)를 노이즈(noise)가 있는

```

Algorithm 1. SGD
Data:  $D = \{ \langle u, i, r \rangle \}$ 
Input:  $feat, \lambda, \eta$ 
for  $\langle u, i \rangle \in D$  do
    {
        if  $u \notin Rows(W)$  then  $W_u \leftarrow Vector(size: feat);$ 
        if  $i \notin Rows(X^T)$  then  $X_i^T \leftarrow Vector(size: feat);$ 
         $err_{ui} \leftarrow 1 - W_u X_i^T$ 
         $W_u \leftarrow W_u + \eta(err_{ui}X_i - \lambda W_u)$ 
         $X_i \leftarrow X_i + \eta(err_{ui}W_u - \lambda X_i)$ 
    }
    
```

그림 3. 알고리즘 1- SGD
Fig. 3. Algorithm 1- SGD

추정치 (SGD)로 대체하는 것이 왜 유리할 수 있는지에 대한 이해를 위해 식(2)를 고려할 수 있다. 이러한 이유는 정 한 기울기 계산(gradient computation)에 비용이 많이 드는 반면 노이즈(noisy)가 많은 추정치를 쉽고 빠르게 얻을 수 있기 때문이다. 일정 시간이 지나면 신중하게 계획된 GD 단계 대신 몇 가지 빠르고 간편하게 SGD 업데이트를 실행할 수 있다. 노이즈(noise) 과정은 또한 국소 최소 점을 피하는 데 도움이 된다. (특히 학습 속도가 클 때 처음에는 매력이 적은 분지가 있는 것들) 또한 SGD는 데이터 내에서 반복을 사용할 수 있다.

IV. 제안 된 모델

4.1 CB와 CF

앞서 언급했듯이 두 시스템은 서로 다른 입력 소스를 활용하며 서로 다른 시나리오에서 잘 작동 할 수 있다. 표 1에 두 가지 모델의 장단점을 요약 하였다.

콘텐츠 기반 추천 시스템에서 항목의 설명 속성은 권장 사항을 작성하는 데 사용된다. “콘텐츠”라는 용어는 이러한 설명을 나타낸다. 콘텐츠 기반 방법에서 사용자의 등급 및 구매 행동은 항목에서 사용 가능한 콘텐츠 정보와 결합된다. 예를 들어, 사용자가 영화 터미네이터(Terminator)를 높은 등급으로 평가했지만 다른 사용자의 등급에 액세스 할 수없는 상황을 예로 들 수 있다. 따라서 협업 필터링 방법이 배제된다. 그러나 터미네이터(Terminator)의 항목 설명에는 에어리언 (Alien) 및 프리데이터(Predator)와 같은 다른 공상 과학 영화와 유사한 장르 키워드가 포함되어 있다. 그러한 경우, 이 영화들은 사용자에게 추천 될 수 있다. 콘텐츠 기반 방법에서 등급으로 레이블이 지정된

표 1. 권고 시스템에서 기본 모델의 장단점
Table 1. Advantage and Disadvantage of Basis Models in Recommendation System

카테고리		협동		입력 소스	이점	불리
		User	Item			
Content-Base		-	-	항목의 텍스트 설명	- 아이템 특징 추출	협업 불필요
CF	NBCF	User-User	✓	커뮤니티 등급	- 단순하다 - 설명하기 쉽다	작은 데이터
		Item-Item	-		- Big data	
	Model	✓	✓			

항목 설명은 사용자 별 분류 또는 회귀 모델링 문제점을 작성하기 위한 학습 데이터로 사용된다.

각 사용자에게 대해 교육 문서는 구매 또는 평가 한 항목에 대한 설명과 일치한다. 클래스 (또는 종속) 변수는 지정된 등급 또는 구매 행동에 해당한다.

이 교육 문서는 현재 사용자 (또는 활성 사용자)에게 특정한 분류 또는 회귀 모델을 작성하는 데 사용된다. 이 사용자 별 모델은 해당 개인이 자신의 등급이나 구매 행동을 알 수 없는 항목을 좋아하는지 여부를 예측하는 데 사용된다. 보다 다양한 입력으로 하이브리드 시스템에는 최상의 결과를 얻을 수 있는 많은 기회가 있다.

4.2 기능 결합 하이브리드

정확도와 시간 업데이트 모두에서 최상의 결과를 얻기 위해 모델을 결합하는 방법을 자세히 고려하고 있다. 표 2는 세 가지 유형의 하이브리드 시스템에 대한 일반적인 평가를 보여준다.

피쳐 조합 하이브리드에서 아이디어는 예측 알고리즘을 적용하기 전에 다양한 소스 (예 : 콘텐츠 및 협업)의 입력 데이터를 통합 된 표현으로 결합하는 것이다. 대부분의 경우 이 예측 알고리즘은 협업 정보를 추가 기능으로 사용하는 콘텐츠 기반 알고리즘이다.

그러한 접근법의 예는 RIPPER 분류기가 증강 데

이터 세트에 적용되는 제시되었다. 이 방법론은 순전히 협력적인 접근 방식에 비해 상당한 개선을 이루었^[44]. 그러나 이 결과를 얻으려면 콘텐츠 기능을 직접 선택해야 한다. 따라서 이 접근 방식은 데이터 세트 및 기능 표현 선택에 민감 할 수 있다. 이 접근 방식은 항목을 평가 한 사용자 수에 대한 시스템 감도를 감소 시킨다.

조합은 다양한 유형의 배경 지식으로 다양한 방식으로 수행 될 수 있음에 유의한다. 예를 들어, 각 항목이 항목의 장르를 나타내는 상위 분류 체계와 연관된 경우를 고려해야 한다. 사용자 및 항목의 표현 프로파일은 계층 구조의 관련 장르로 향상 될 수 있다. 등급 매트릭스는 아이템이 아닌 장르의 관점에서 구성 될 수 있다. 희소 행렬에서 이러한 접근 방식은 열 수를 줄이고 대 부분의 항목이 압축 된 행렬에 채워질 수 있기 때문에 보다 효과적인 결과를 제공 할 수 있다.

다른 방법은 평가 매트릭스를 늘리고 항목 외에 키워드에 대한 열을 추가하는 것이다. 따라서 등급 행렬은 $m \times (n+d)$ 행렬이 된다. 여기서 n 은 항목 수이고 d 는 키워드 수이다. “키워드 항목”의 가중치는 사용자가 액세스, 구매 또는 평가 한 항목에 대한 설명의 가중치 집계를 기반으로 한다. 이 확장 된 매트릭스에는 전통적인 인접 또는 매트릭스 인수 분해 방식을 사용할 수 있다. 두 가지 최적화 모델의 이러한 유형의 조합은 하이브리드 설정에서 일반적이며, 목적 함수는 파라미터 벡터 Θ 와 관련하여 다음과 같이 설정된다.

$$J = CF\ Objective(\theta) + \beta.CB\ Objective(\theta) + Regularization \quad (5)$$

그런 다음 목적 함수는 매개 변수 벡터 Θ 에 대해 최적화된다. 아래에서 논의되는 특정 예는 부가 정보를 가진 희소 선형 모델의 일반화이다.

본 연구의 모델에서, 우리는 R 을 $m \times n$ 의 암시 적

표 2. 하이브리드 시스템의 3가지 유형에 대한 일반적인 평가
Table 2. Generally Evaluation about Three Types of Hybrid System

	Improve accurate	Improve time	Hadoop-Map/Reduce	Level execute
Monolithic Algorithm 1 + Algorithm 2 Input → Algorithm 12 → Output	Good	Good	Good	Complex
Ensembles/ Sequential Input → Algorithm 1 → Output 1 Output ← Algorithm 2 →	Good	X	X	Normal
Ensembles/ Parallel Input → Algorithm 1 → Output 1 Input → Algorithm 2 → Output 2 → Output	Good	X	X	Normal

피드백 등급 행렬로 설정하고, C 를 $d \times n$ 의 내용 행렬로 설정한다. 여기서 각 항목은 음수가 아닌 d 단어의 주파수로 설명된다. 예로는 품목에 대한 설명이나 품목에 대한 간단한 리뷰가 있다. R 은 암시 적 피드백 매트릭스이므로 누락 된 항목은 0으로 가정된다. 그 후, W 는 등급이 $\hat{R} = R \cdot W$ 로 예측되는 $n \times n$ 개의 항목-항목 계수 행렬이며 이 경우 등급을 $\hat{R} = C \cdot W$ 로 예측할 수도 있다. 따라서 $\|R - R \cdot W\|^2$ 만 최적화하는 대신 콘텐츠 기반 용어 $\|R - C \cdot W\|^2$ 를 추가한다. Elastic-net 정규화 및 비-음성 / 대각 제약과 함께 향상된 최적화 모델은 다음과 같다.

$$\text{Minimize } J = \|R - RW\|^2 + \beta \cdot \|R - CW\|^2 + \lambda \|W\|^2 + \lambda_1 \cdot \|W\|_1 \quad (6)$$

에 따라:

$$W \geq 0$$

대각선 (W) = 0

가중치 파라미터 β 는 튜닝 단계에서 결정될 수 있다. 등급을 $\hat{R} = R \cdot W$ 또는 $\hat{R} = C \cdot W$ 로 예측할 수 있지만 이전 예측 함수 만 사용된다. 따라서 $\|R - C \cdot W\|^2$ 라는 용어는 목적 함수를 추가 정규화 도구로 세분화하는 데만 사용된다. 즉, 추가 용어의 목표는 사용자의 미래 (아직 알려지지 않은) 행동을 예측하기 위해 모델의 일반화 능력을 향상시키는 것이다.

이 유형의 접근 방식은 다른 유형의 협업 필터링 (최적화) 모델을 콘텐츠 기반 방법과 결합하는 데 사용할 수 있다. 예를 들어, 행렬 인수 분해의 경우, $m \times k$ 사용자 계수 행렬 U , $n \times k$ 공유 항목 계수 행렬 X 및 $d \times k$ 콘텐츠 계수 행렬 Z 를 사용하여 다음과 같이 최적화 모델을 설정할 수 있다.

$$\text{Minimize } J = \|R - UX^T\|^2 + \beta \cdot \|C - ZX^T\|^2 + \lambda (\|U\|^2 + \|X\|^2 + \|Z\|^2) \quad (7)$$

항목 계수 행렬 X 는 등급 행렬과 콘텐츠 행렬의 인수 분해 간에 공유된다. 이러한 공유 매트릭스 인수 분해 모델은 사회적 신뢰 데이터와 같은 다른 유형의 부가 정보를 통합하는 데에도 사용한다.

4.3 KMSGD 연산

4.3.1 SGD의 한계

수렴은 행렬 인수 분해에서 관측된 항목을 반복적

으로 순환하여 인수 행렬을 업데이트할 때 최종적으로 달성할 수 있다. 일반적으로 전역 방법은 일반적으로 로컬 방법보다 느리지만 수렴을 보장한다고 할 수 있다. 학습 속도 (또는 단계 크기)의 일반적인 값은 $\alpha = 0.005$ 와 같은 작은 상수 값이다. 따라서 수렴 속도를 높이고 로컬 최솟값을 피하기 위함보다 적합한 방법은 볼드 드라이버 알고리즘(bold driver algorithm)을 사용하여 각 반복에서 α 를 적응적으로 식별하는 것이다^{45,46}. 원칙적으로 다양한 요인에 대해 서로 다른 학습률을 사용하는 것도 가능하다⁴⁷. 이러한 다양한 모델에 대한 인상적인 관찰은 많은 반복에 대한 수렴이 수행 될 때까지 모델을 수행하면 일반적으로 관찰되지 않은 항목의 솔루션 품질이 약간 저하 될 수 있다는 것이다. 따라서 수렴 기준을 너무 엄격하게 설정하지 않는 것이 일반적으로 적합하다고 할 수 있다.

4.3.2 아이디어 제안

미니 배치 그룹: 미니 배치 그라디언트 디센트 (Gradient descent)는 모델 데이터를 업데이트하고 모델 오류를 식별하는 데 사용되는 훈련 데이터 세트를 더 작은 배치로 나누는 그라디언트 디센트(Gradient descent) 알고리즘의 변형이다. 실행은 미니 배치에 대한 그라디언트(Gradient)를 계산하거나 그라디언트(Gradient)의 분산을 추가로 감소시키는 그라디언트(Gradient)의 평균을 취하도록 선택할 수 있다.

이 알고리즘은 증분 프로세스로 작동하도록 구성되어 현재 관측치 만 기반으로 계수 행렬 W 및 X 를 업데이트 한다. 이 알고리즘은 알고리즘 1과의 형식적 유사성에도 불구하고 두 가지 중요한 차이점이 있다. 먼저 학습 절차는 사용 가능한 데이터에 대한 단일 패스를 요구한다. 알고리즘 2에서, 각각의 관측 값 $\langle u, i \rangle$ 에서, 인수 행렬 W 및 X 에 대한 변경은 단일 단계로 이루어진다. 또 다른 가능한 접근 방법은 다시 반복하는 데 필요한 추가 시간을 소비하면서 정확도를 향상 시키면서 각각의 새로운 관측에 대해 여러 번 반복하는 것이다. 둘째, 다른 전 처리 또는 데이터 셔플 링이 실행되지 않는다. 등급 피드백을 처리한다고 가정하면 등급 행렬 R 에 접근하게 된다. 따라서 오류를 $err_{ui} = 1 - \hat{R}_{ui}$ 로 측정하고 (3)의 업데이트 작업을 사용하여 W 및 X^T 의 행을 업데이트 하고 이러한 알고리즘을 알고리즘 2라고 한다.

미니 배치 그라디언트 디센트는 결국 두 차원을 모두 활용하고 n 개의 훈련 예제의 각 미니 배치에 대한

업데이트를 실행한다.

$$\phi = \phi - \eta \cdot \nabla \phi L(\phi; x^{(i:i+k)}, y^{(i:i+k)}) \quad (8)$$

이 방법으로 보다 안정적인 수렴을 목표로 매개 변수 업데이트의 편차를 줄일 수 있으며, 그라디언트 $W.X.R$ 을 계산하는 최첨단 기계 학습 라이브러리에 일반적으로 고도로 최적화 된 매트릭스 최적화를 사용하여 미니 배치가 효과적이라는 장점을 가지고 있다.

모멘텀(momentum)은 앞서 언급 한 과제를 해결하기 위해 고차원 데이터 세트에 대해 실험에서 실행 가능한 알고리즘을 적용한다. 네임 모멘텀(name momentum)물리학의 운동량과 유사성에서 시작한다. 매개 변수 공간을 통해 이동하는 입자로 생각되는 무게 벡터 W 는 손실 함수의 기울기 (“force”)로부터 가속을 받는다. 전통적인 확률적 경사하강과는 달리, 같은 방향으로 계속 움직이면서 진동을 방지하는 것을 목표로 한다. 모멘텀(momentum)은 컴퓨터 과학자들에 의해 수 십년 동안 인공지능망에 관한 연구에서 효과적으로 적용되었다.

모멘텀(momentum)은 감쇠 진동과 관련 방향으로 SGD를 가속 시키는 방법이다⁴⁸⁾. 과거 시간 단계의 업데이트 벡터의 일부를 최근 업데이트 벡터에 보완하여 이를 수행한다.

$$v_t = \gamma \cdot v_{t-1} + \eta \nabla \phi L(\phi) \quad (9)$$

$$\phi = \phi - v_t$$

운동량 항 γ 는 종종 0 : 9 또는 유사한 값으로 설정된다. 기본적으로 운동량이 사용될 때, 우리는 언덕 아래로 공이 하강한다. 그리고 공은 내리막길에서 운동 속도가 누적되어 도중에 더 빨라진다. (공기 저항이 예를 들어 $\gamma < 1$ 인 경우, 끝 속도에 도달 할 때까지). 매개 변수 업데이트와 비슷한 일이 발생한다. 그라디언트(Gradient) 가 비슷한 방향을 가리키는 치수의 운동량은 증가하고 그라디언트(Gradient)의 방향이 바뀌는 치수의 업데이트는 감소한다. 따라서 진동과 수렴이 줄어든다.

위의 분석에서 우리는 KMSGD를 각 반복에 v_t 를 추가하여 그림 4와 같이 제안한다.

각 반복에 v_t 를 추가함으로써 제안 된 알고리즘은 더 나은 수렴을 시도한다. 더욱이, mini bath k 데이터 포인트를 사용한 훈련은 알고리즘의 빠른 학습을 가

```

Algorithm 3. KMSGD
Data:  $D = \{ \langle u, i, r \rangle \}$ 
Input:  $feat, \lambda, \eta, k$ 
for count  $\leftarrow 1$  to k
{
  for  $\langle u, i \rangle \in D$  do
  {
    if  $u \notin Rows(W)$  then  $W_u \leftarrow$  Vector (size:  $feat$ );
    if  $i \notin Rows(X^T)$  then  $X_i^T \leftarrow$  Vector (size:  $feat$ );
     $err_{ui} \leftarrow 1 - W_u X_i^T - v_t$ 
     $W_u \leftarrow W_u + \eta(err_{ui} X_i - \lambda W_u) - v_t$ 
     $X_i \leftarrow X_i + \eta(err_{ui} W_u - \lambda X_i) - v_t$ 
  }
}
    
```

그림 4. KMSGD 알고리즘
Fig. 4. Algorithm for KMSGD

져올 것으로 예상된다. 제안 된 방법의 수행은 다음 절에 제시하였다.

V. 평가 문제 및 데이터

5.1 데이터

이 실험에서는 MovieLens-100K 및 MovieLens-1M이라는 두 가지 실제 데이터 세트를 사용하여 모델의 성능을 테스트한다.

전자 인 MovieLens-100K 데이터 세트에는 약 943 명의 사용자와 약 1,682 개의 평가 기록이 있다. 데이터 세트는 Movie-Lens 웹 사이트에서 수집되며 모든 등급 척도는 1에서 5까지이다.

후자 인 MovieLens-1M 데이터 세트는 Movie-Lens 웹 사이트에서도 수집되지만 6,040 명의 사용자로부터 1,000,209 개의 등급과 3,952 개의 영화가 포함된

표 3. 두 데이터 세트의 통계
Table 3. The Statistics of the Two Data sets

Features	Data	
	MovieLens-100K	MovieLens-1M
# of users	943	6,040
# of items	1,682	3,952
# of ratings	100,000	1,000,209
# of ratings per user	106.4	165.6
# of ratings per item	59.45	253.09
Rating Sparsity	93.70%	95.81%

다. 또한 각 사용자는 최소 20 개의 영화를 평가했으며 이전 데이터베이스보다 늦게 개봉하였다.

5.2 평가 측정

우리는 모델의 정확성을 평가하기 위해 권장 시스템을 포함한 많은 분야에서 널리 사용되는 RMSE (Root Mean Squared Error) 방법을 채택하였으며 측정값은 다음과 같다.

$$SE = \sqrt{\frac{1}{N} \sum_{ij} (R_{ij} - \hat{R}_{ij})^2} \quad (10)$$

RMSE의 값이 낮을수록 예측 모델의 성능이 더 우수함을 나타낸다.

실험에 적용하는 두 번째 방법은 시간 업데이트이다. 정확성에 대한 테스트 외에도 평가하고자 하는 매우 중요한 값은 알고리즘 데이터 학습 속도이다.

5.3 비교 방법

본 연구의 실험은 두 가지 부분으로 나눌 수 있다. 정확도를 확인하고 다른 권장 알고리즘과 비교하여 제안된 예측 모델의 업데이트 시간 성능을 비교한다. 실험을 시작할 때 제안된 모델의 정확성을 테스트하였다. 본 실험에서는 RMSE 측정 방법만 사용하였다.

일반적인 평가를 위해, 우리는 알고리즘과 비교하기 위해 기본부터 복잡성까지 권장 알고리즘을 선택했다. 모든 알고리즘은 CB 알고리즘(1), NBCF 알고리즘의 사용자 및 항목별 CF 알고리즘과 모델 기반 CF(2)의 ALS, SGD, KMSGD, 순차, 병렬 및 제안 모델(3)의 3가지 그룹으로 나뉜다. 이 범주를 보유하는 이유는 다음과 같은 상태인지 여부를 테스트하기 때문이다.

- (1) 정확성 : 복잡한 알고리즘 일수록 예측 정확도가 높아질 것이다.
- (2) 시간 업데이트 : 복잡한 알고리즘 일수록 업데이트 속도가 느려질 것이다.

5.4 평가 문제

5.4.1 오프라인 평가

평가 방법과 표준화 된 프레임 워크가 이러한 경우를 위해 개발되었으므로 오프라인 방법은 권장 시스템 알고리즘을 테스트하는 가장 일반적인 방법 중 한 가지이다. 오프라인 평가의 주된 약점은 향후 추천 시스템에 반영 할 사용자의 실제 성향을 평가하지 않는다는 것이다. 예를 들어, 데이터는 시간이 지남에 따

라 발전 할 수 있으며 현재 예측은 미래에 가장 적합한 예측에 반응하지 않을 수 있다. 더욱이 정확성과 같은 수단은 참신함과 세렌디피티(serendipity)와 같은 권장 사항의 주요 특성을 포착하지 못한다. 이러한 권고는 권고의 환율에 장기적인 영향을 미친다. 그러나 이러한 약점에도 불구하고 오프라인 방법은 여전히 권장 시스템 평가에 가장 널리 사용되는 방법이다. 그 이유는 이러한 테스트 방법을 통해 이해하기 쉬운 정량화와 통계적으로 강력하기 때문이다⁴⁹⁾.

5.4.2 온라인 평가

온라인 평가는 또한 사용자가 일반적으로 상용 또는 완전히 배포 된 시스템의 실제 사용자라는 점을 제외하고 사용자 조사를 유도한다. 이 방법은 일반적으로 사용자가 자연스럽게 업무를 수행함에 있어 시스템을 직접 사용하기 때문에 채용 프로세스에서 편견에 덜 민감하기 때문이다. 이 시스템들은 보통 다른 알고리즘의 성능을 평가하기 위해 적용될 수 있다⁵⁰⁾.

온라인 평가 방법하기 위한 방법은 다음과 같이 두 가지의 알고리즘을 비교하는 것이다 :

- (1) 사용자를 두 그룹 A와 B로 나눈다.
- (2) 일정 기간 동안 각각 다른 알고리즘을 사용하고 두 그룹에 걸쳐 다른 모든 조건 (예 : 사용자 선택 프로세스)을 가능한 유사하게 유지한다.
- (3) 절차가 끝날 때 두 그룹의 전환율을 비교한다.

이 일반적인 접근 방식은 증분 학습 개념과 관련이 있으며 일반적으로 온라인 시스템과 쌍을 이룰 수 있다. 증분 학습이 회귀 모델링 및 분류 및 문헌에서 광범위하게 조사되었지만, 권장 영역에서의 관련 연구는 보다 제한적이다⁵¹⁻⁵⁴⁾. 이러한 알고리즘의 추가 개선을 위한 중요한 연구 기회가 있다.

이 방법의 주된 약점은 많은 수의 사용자가 이미 참여하지 않으면 그러한 시스템을 현실적으로 채택 할 수 없다는 것이다. 따라서 시작 단계에서 이 방법을 사용하기가 어렵다. 더욱이 이러한 시스템은 대부분 공개적으로 접근 할 수 없으며, 일반적으로 현재 이용 가능한 상용 시스템의 소유자만 접근 할 수 있다. 따라서 이러한 테스트는 상업적 실체와 해당 시스템에 의해 제한된 수의 시나리오에서만 실행할 수 있다⁵⁵⁾.

VI. 결과 및 토론

6.1 제안된 평가 프로토콜

4절에서 언급한 문제를 감안할 때, 우리는 데이터

스트림(data stream)을 처리하는 알고리즘의 평가에 적합한 사전 방법을 설계한다⁵⁶⁾. 항목 i 와 사용자 u 사이의 등급 상호 작용을 나타내는 관찰된 각 이벤트 $\langle u, i \rangle$ 에 대해 다음 단계를 실행한다.

- (1) u 가 알려진 사용자 인 경우 현재 모델을 사용하여 항목 등급 값을 u 로 예측한다. 그렇지 않으면 3 단계로 이동한다.
- (2) 관찰 된 항목 i 가 주어진 등급 목록과 비교하여 평가한다.
- (3) 관찰 된 이벤트로 모델을 업데이트 한다.
- (4) 데이터 세트에서 다음 이벤트로 진행한다.

이 프로토콜은 몇 가지 이점을 제공한다. 추천 시스템의 예측 능력에 대한 지속적인 온라인 모니터링이 가능하며 온라인 통계는 알고리즘의 논리에 통합 될 수 있다. (예 : 배치 재훈련 드리프트 감지를 트리거하는 자동 파라미터 조정) 또한 전체적으로 개별 알고리즘 파라미터의 상대적 가중치를 조정할 수 있다.

또한 이 프로토콜은 등급 예측 및 항목 예측에 모두 적합하다. 실험 오프라인 설정에서는 데이터 세트가 사실상 유한하기 때문에 개별 점수의 전체 평균을 최종적으로 계산할 수 있다. 실제 환경에서 실행되는

추천 시스템의 경우 이 프로토콜을 사용하면 점수 통계를 추적하여 실험 전체 시간 동안 점수의 진행 상황을 추적 할 수 있다. 이에 따라 알고리즘의 처리 과정이 시간이 지나면서 어떻게 진화하는지 이해할 수 있다. 표 3에 설명 된 데이터 세트를 사용하여 기초에서 복잡성까지 5 개의 알고리즘의 전체 정확도를 비교한다. 연구 대상이 아닌 콜드 스타트 문제를 피하기 위해 각 데이터 세트의 처음 20% 데이터 포인트를 사용하여 알고리즘의 원래 배치 교육을 실행하고 왼쪽 데이터는 증분 훈련을 위해 처리된다.

모든 알고리즘의 모든 매개 변수는 각 데이터 세트에 대해 독립적으로 조정된다. 교차 검증은 스트리밍 데이터에 적합하지 않으므로 제시된 실험에 채택 된 유사한 평가 프로토콜을 사용하지만 각 데이터 세트의 데이터의 초기 20%에 대해서만 평가한다.

6.2 결과 및 토론

정확성 및 시간 업데이트 결과는 각각 MovieLen-100k 데이터 및 MovieLen-1M과 함께 표 4 및 표 5에 표시한다. 또한 정확도와 시간 업데이트의 평균값을 가진 모델 간의 비교는 그림 6에 제시하였다.

표 4 및 표 5의 실험 결과를 관찰하여 다음과 같은

표 4. MOVIELEN-100K 데이터의 전체 결과
Table 4 Overall results of Movielen-100k Data

대책	(1) Content-Base (CB)	(2) Collaboration Base - Filtering (CF)					(3) 하이브리드 시스템		
		Neighborhood-Base CF (NBCF)		Model-Based CF (MFCF)			앙상블		모 놀리 식 디자인
		User-User	Item-Item	ALS	SGD	KMSGD	잇달아 일어나는	평행	제안 된 모델
RMSA	0.914	0.976	0.968	0.951	0.948	0.949	0.894	0.896	0.890
Time update	0.886ms	1.186ms	1.166ms	1.986ms	0.118ms	0.101ms	0.108ms	0.106ms	0.104ms

표 5. MOVIELEN-1M 데이터의 전체 결과
Table 5. Overall results of Movielen-1M Data

대책	(1) Content-Base (CB)	(2) Collaboration Base - Filtering (CF)					(3) 하이브리드 시스템		
		Neighborhood-Base CF (NBCF)		Model-Based CF (MFCF)			앙상블		모 놀리 식 디자인
		User-User	Item-Item	ALS	SGD	KMSGD	잇달아 일어나는	평행	제안 된 모델
RMSA	0.916	0.975	0.966	0.944	0.940	0.936	0.912	0.918	0.910
Time update	3.016ms	3.180ms	3.166ms	4.168ms	0.201ms	0.198ms	0.204ms	0.201ms	0.199ms

논의를 도출 할 수 있다.

6.2.1 정확한 정보

이 부분에서는 더 큰 모델 유연성이 더 정확한지 경험적으로 고려한다.

그룹 1과 2를 비교할 때 가장 흥미로운 알고리즘은 가장 간단한 알고리즘이 CB 알고리즘에 대해 RMSA 0.914에서 최상의 결과를 보여준다.

CF 알고리즘에서 그 차이는 눈에 띄지 않지만 모델 알고리즘은 다른 알고리즘보다 약간 더 큰 값을 가지며 특히 0.949의 KMSGD이다.

그룹 2 (NBCF 및 MFCF)에만 중점을 두면 KMSGD 알고리즘이 SGD를 제외한 최상의 결과를 보여준다.

0.914 RMSA 값으로 인상적인 CB 알고리즘으로 돌아온다. 그 결과를 설명하기 위해 입력 데이터를 고려하였다. 복소수 모델은 복소수 데이터에 적합하다. MovieLen-100k 및 MovieLen-1M에서 항목 (동영상)의 기능은 매우 명확하고 간단하다.

그로부터 영화의 특징은 그 영화의 종류에 대해 1개의 값을 갖는 19개의 이진수 (19가지의 영화)를 보여준다. 이는 CB에서 TF-IDF 알고리즘을 용이하게 하며 이것은 CB에서 매우 낮은 RMSA 값에 대한 합리적인 해답이다.

이러한 이유로, 데이터가 더 큰 MovieLens-1M일 때 0.916의 CB에서 가장 낮은 RMSA 값으로 유사한 것이 반복된다. 그러나 KMSGD 알고리즘은 그룹 2 중 0.936으로 최고의 가치를 달성하였다. 변형이 있는 SGD 알고리즘이 여전히 행렬 분해에 가장 적합한 것으로 가정한다.

앞서 언급했듯이 CB 알고리즘은 정확도가 가장 높았다. 이를 위해 우리의 모델은 CB와 CF의 기능을 단일 모델로 결합하였다.

이 결합을 통해 두 모델 모두에서 0.890 및 0.910 RMSE 값으로 최상의 결과를 얻었다.

연구 질문에 따르면, 콘텐츠 기반과 모델 기반 협업 필터링을 결합하는 데 하이브리드 시스템이 적합하고 기능 조합 하이브리드 시스템은 다른 모델에 비해 정확성과 시간 업데이트에서 더 나은 결과를 얻었다.

6.2.2 시간 업데이트

본 연구에서 시간 업데이트(time-update)는 복잡한 모델은 빅 데이터 처리 속도가 느릴 것으로 가정한다.

이 상태는 사용자-사용자, 항목-항목, 특히 ALS의 시간 업데이트 값으로 각각 1.186ms, 1.166ms,

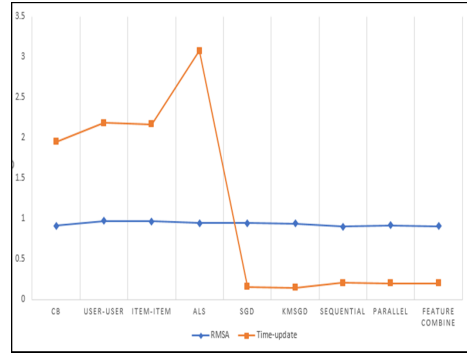


그림 5. 모델 간 비교 평균 RMS 및 시간 업데이트 값
Fig. 5. Comparison average RMSA and time update value among models

1.986ms로 입증되었다. 0.886ms의 덜 복잡한 모델 CB보다 두 배나 높다. 향상된 알고리즘 (SGD, KMSGD 및 모델)의 값은 각각 0.118ms, 0.101ms 및 0.104ms 로 상당히 낮다. 이러한 결과는 본 연구 모델이 최상의 결과를 가져온다는 것을 보여준다.

결과는 MovieLen-1M 데이터와 거의 유사하였다. 이러한 결과는 “데이터를 처리하는 데 더 복잡한 모델이 느릴 것”이라는 주장과 다르다는 것을 의미한다. 그러나 알고리즘의 시간 업데이트는 큰 문제이다. Movie-Len 1M을 사용하면 ALS의 속도 업데이트 모델이 4.168ms로 상당히 커진다. 다른 모드와 비교하면서 증분 모델은 계속 인상적인 값을 보여준다.

제안 모델의 효과를 평가하기 위해 그림 5와 같이 오실 레이트를 그래프로 볼 수 있다. 하이브리드 시스템 및 증분 알고리즘 SGD 및 KMSGD는 향후 매트릭스 인수 분해에 미치는 이점에 대해 중요한 조사가 가능하다는 결론을 내릴 수 있다.

VII. 결 론

간단한 SGD 알고리즘과 통합 기능 CB 및 CF로 시작하여 현재 사용 가능한 관측 소규모 그룹만 사용하여 향상 훈련을 실행하도록 역할을 조정한다. 본 연구의 알고리즘 성능에 대한 실험 결과는 본 연구에서 제안한 모델의 예측 능력과 시간 업데이트 부분에서 더 나은 결과를 낸 것을 확인할 수 있다. 데이터를 학습하고 업데이트하는 속도는 그룹 2에서 경쟁력이 있었다. 또한 사용자 피드백 스트리밍을 위해 설계된 새로운 실험 프로토콜(protocol)도 제안한다.

이 연구의 목적 추천 시스템에 하이브리드 접근법의 유용성과 적합성 조사였다. 결과를 본 이번 연구

건물 추천 시스템에 새로운 하이브리드 접근법의 잠재력을 실증적 증거 제공했다. 이번에 하이브리드 시스템은 불투명한“블랙박스”로 운영되고 다재다능하고 튼튼한 아이템들과 CF의 특징을 제공하는 두 개의 접근법CB의 특징을 담고 있다.

전반적으로 우리의 모델의 성능 그룹 1에서 모델과 2보다 더 나은 것이었다. 이 일에 관한 최선의 실행은 그 시간 시스템을 배우기를 원한다. 이러한 결과는 하이브리드 모델의 권장 방법 추천 시스템에 적절한 것을 의미한다. 특히 영화제, 도서제 등 사용자의 등급이 분명한 텍스트나 코멘트가 있는 상품을 추천한다.

향후 연구에서 우리는 추천 시스템에서의 컨볼루션 신경망 (convolutional neural network) 방법과 같은 다른 딥 러닝(deep-learning) 기법의 적용을 연구하고 그 성과를 더욱 발전시키기 위한 노력이 필요하다. 본 연구를 통해 맵-레듀스 (Map-reduce) 메카니즘으로 분산 데이터에 적용하려고 시도하였다.

References

- [1] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, “Recommender systems survey,” *Knowledge-based Syst.*, vol. 1, no. 46, pp. 109-132, Jul. 2013.
- [2] K. Choi, D. Yoo, G. Kim, and Y. Suh, “A hybrid online-product recommendation system: combining implicit rating-based collaborative filtering and sequential pattern analysis,” *Electron. Commerce Res. and Appl.*, vol. 11, no. 4, pp. 309-317, 2012.
- [3] S. K. Lee, Y. H. Cho, and S. H. Kim, “Collaborative filtering with ordinal scale-based implicit ratings for mobile music recommendations,” *Inf. Sci.*, vol. 180, no. 1, pp. 2142-2155, 2010.
- [4] E. R. Núñez-Valdéz, J. M. C. Lovelle, O. S. Martínez, V. GarcíaDíaz, P. O. Pablos, and C. E. M. Marín, “Implicit feedback techniques on recommender systems applied to electronic books,” *Computers in Human Behavior*, vol. 28, no. 4, pp. 1186-1193, 2012.
- [5] B. M. Marlin and R. S. Zemel, “Collaborative prediction and ranking with non-random missing data,” *ACM RecSys '09*, pp. 5-12, 2009.
- [6] B. M. Marlin, R. S. Zemel, S. Roweis, and M. Slaney, “Collaborative filtering and the missing at random assumption,” in *UAI '07*, pp. 50-54, AUAI, 2007.
- [7] S. M. McNee, J. Riedl, and J. A. Konstan, “Being accurate is not enough: How accuracy metrics have hurt recommender systems,” in *ACM CHI '06 Extended Abstracts*, pp. 1097-1101, ACM, 2006.
- [8] M. D. Ekstrand, J. T. Riedl, and J. A. Konstan, “Collaborative filtering recommender systems,” *Foundations and Trends® in Human-Computer Interaction*, vol. 4, no. 2, pp. 81-173, May 2011.
- [9] F. Ricci, L. Rokach, and B. Shapira, “Introduction to recommender systems handbook,” *Recommender systems handbook*, Springer, pp. 1-35, Boston, MA, 2011.
- [10] M. J. Pazzani and D. Billsus, “Content-based recommendation systems,” *The Adaptive Web*, Springer, pp. 325-341, 2007.
- [11] G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” *IEEE Trans. Knowledge and Data Eng.*, vol. 17, no. 6, pp. 734-749, 2005.
- [12] Z. Huang, D. Zeng, and H. Chen, “A comparison of collaborative-filtering recommendation algorithms for ecommerce,” *IEEE Intell. Syst.*, vol. 22, no. 5, 2007.
- [13] X. Su and T. M. Khoshgoftaar, “A survey of collaborative filtering techniques,” *Advances in Artificial Intell.*, vol. 2009, 2009.
- [14] J. S. Breese, D. Heckerman, and C. Kadie, “Empirical analysis of predictive algorithms for collaborative filtering,” in *Proc. Fourteenth Conf. Uncertainty in Artificial Intell.*, pp. 43-52, Morgan Kaufmann Publishers Inc., 1998.
- [15] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Item based collaborative filtering recommendation algorithms,” in *Proce. 10th Int. Conf. World Wide Web*, ACM, pp. 285-295, 2001.
- [16] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, “An algorithmic framework for

- performing collaborative filtering,” in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. and Develop. in Inf. Retrieval*, pp. 230-237, ACM, 1999.
- [17] T. Hofmann and J. Puzicha, “Latent class models for collaborative filtering,” in *IJCAI*, vol. 99, no. 1999, Jul. 1999.
- [18] S. Vucetic and Z. Obradovic, “Collaborative filtering using a regression-based approach,” *Knowledge and Inf. Syst.*, vol. 7, no. 1, pp. 1-22, 2005.
- [19] K. Miyahara and M. Pazzani, “Collaborative filtering with the simple bayesian classifier,” *PRICAI 2000 Topics in Artificial Intell.*, pp. 679-689, 2000.
- [20] J. Liu, Y. Jiang, Z. Li, X. Zhang, and H. Lu, “Domain-sensitive recommendation with user-item subgroup analysis,” *IEEE Trans. Knowledge and Data Eng.*, vol. 28, no. 4, pp. 939-950, 2016.
- [21] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” *Computer*, vol. 42, no. 8, Aug. 2009.
- [22] K. Yu, S. Zhu, J. Lafferty, and Y. Gong, “Fast nonparametric matrix factorization for large-scale collaborative filtering,” in *Proc. 32nd Int. ACM SIGIR Conf. Res. and Develop. in Inf. Retrieval*, pp. 211-218, ACM, 2009.
- [23] A. Mnih and R. R. Salakhutdinov, “Probabilistic matrix factorization,” in *Advances in Neural Inf. Process. Syst.*, pp. 1257-1264, 2008.
- [24] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Application of dimensionality reduction in recommender system-a case study,” DTIC Document, Tech. Rep., 2000.
- [25] R. M. Bell and Y. Koren, “Scalable collaborative filtering with jointly derived neighborhood interpolation weights,” in *ICDM*, pp. 43-52, IEEE Computer Society, 2007.
- [26] S. Funk, from <http://sifter.org/~simon/journal/20061211.html>, 2006.
- [27] A. Paterek, “Improving regularized singular value decomposition for collaborative filtering,” in *Proc. KDD Cup and Workshop*, vol. 2007, pp. 5-8, 2007.
- [28] G. Takács, I. Pilászy, B. Németh, and D. Tikk, “Scalable collaborative filtering approaches for large recommender systems,” *J. Machine Learning Res.*, vol. 10, pp. 623-656, 2009.
- [29] R. M. Bell and Y. Koren, “Scalable collaborative filtering with jointly derived neighborhood interpolation weights,” in *Proc. 7th IEEE Int. Conf. Data Mining*, pp. 43-52, IEEE Computer Society, Omaha, Nebraska, USA, 2007.
- [30] I. Pilászy, D. Zibriczky, and D. Tikk, “Fast als-based matrix factorization for explicit and implicit feedback datasets,” in *RecSys '10: Proc. Fourth ACM Conf. Recommender Syst.*, pp. 71-78, ACM, New York, NY, USA, 2010.
- [31] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” in *Proc. COMPSTAT'2010*, pp. 177-186, Physica-Verlag HD, 2010.
- [32] R. Burke, “Hybrid recommender systems: Survey and experiments. User modeling and user-adapted interaction,” vol. 12, no. 4, pp. 331-370, 2002.
- [33] E. Eaton, editor, “*Lifelong Machine Learning*,” AAAI Spring Symposium, vol. SS-13-05 of AAAI Technical Report, AAAI, 2013.
- [34] C. Miranda and A. M. Jorge, “Incremental collaborative filtering for binary ratings,” *IEEE Comput. Soc.*, vol. 01, pp. 389-392, 2008.
- [35] J. Vinagre and A. M. Jorge, “Forgetting mechanisms for scalable collaborative filtering,” *J. Brazilian Comput. Soc.*, vol. 18, no. 4, pp. 271-282, 2012.
- [36] B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Incremental SVD-based algorithms for highly scalable recommender systems,” in *Fifth Int. Conf. Comput. and Inf. Technol.*, pp. 27-28, 2002.
- [37] M. Berry, S. Dumais, and G. O'Brien, “Using linear algebra for intelligent information retrieval,” *SIAM Rev.*, pp. 573-595, 1995.
- [38] E. Diaz-Aviles, L. Drumond, L. Schmidt-Thieme, and W. Nejdl, “Real-time top-n recommendation in social streams,” in *Proc.*

- Sixth ACM Conf. Recommender Syst.*, ACM, pp. 59-66, 2012.
- [39] G. Takács, I. Pilászy, B. Németh, and D. Tikk, "Scalable collaborative filtering approaches for large recommender systems," *J. Machine Learning Res.*, vol. 10, pp. 623-656, 2009.
- [40] G. Ling, H. Yang, I. King, and M. R. Lyu, "Online learning for collaborative filtering," in *IEEE 2012 Int. Joint Conf. Neural Netw.*, pp. 1-8, 2012.
- [41] R. Gemulla, E. Nijkamp, P. J. Haas, and Y. Sismanis, "Large-scale matrix factorization with distributed stochastic gradient descent," in *Proc. 17th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 69-77, ACM, Aug. 2011.
- [42] L. Bottou and O. Bousquet, "The tradeoffs of large-scale learning," in *Advances in Neural Inf. Process. Syst.*, vol. 20, pp. 161-168, 2008.
- [43] C. Basu, H. Hirsh, and W. Cohen, "Recommendation as classification: using social and content-based information in recommendation," *AAAI*, pp. 714-720, 1998.
- [44] S. Ruder, "An overview of gradient descent optimization algorithms," arXiv preprint arXiv:1609.04747, Sep. 2016.
- [45] R. Battiti, "Accelerated backpropagation learning: Two optimization methods," *Complex Syst.*, vol. 3, no. 4, pp. 331-342, 1989.
- [46] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Math. Statist.*, vol. 22, no. 3, pp. 400-407, 1951.
- [47] C. Darken, J. Chang, and J. Moody, "Learning rate schedules for faster stochastic gradient search," *Neural Netw. for Sign. Process. II Proc. 1992 IEEE Wrkshp.*, pp. 1-11, Sep. 1992.
- [48] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional nonconvex optimization," *Advances in Neural Inf. Process. Syst.*, pp. 2933-2941, 2014.
- [49] Q. Ning, "On the momentum term in gradient descent learning algorithms," *Neural Netw.*, vol. 12, no. 1, pp. 145-151, 1999.
- [50] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne, "Controlled experiments on the web: survey and practical guide," *Data Mining and Knowledge Discovery*, vol. 18, no. 1, pp. 140-181, 2009.
- [51] G. Shani and A. Gunawardana, "Evaluating recommendation systems," in *Recommender systems handbook*, Springer, pp. 257-297, Boston, MA, 2011.
- [52] R. Kohavi, R. Longbotham, D. Sommerfield, and R. Henne, "Controlled experiments on the web: Survey and practical guide," *Data Mining and Knowledge Discovery*, vol. 18, no. 1, pp. 140-181, 2009.
- [53] R. Sutton and A. Barto, "Reinforcement learning: An introduction," MIT Press, Cambridge, 1998.
- [54] T. Mahmood and F. Ricci, "Learning and adaptivity in interactive recommender systems," *Int. Conf. Electron. Commerce*, pp. 75-84, 2007.
- [55] T. Mahmood and F. Ricci, "Improving recommender systems with adaptive conversational strategies," *ACM Conf. Hypertext and Hypermedia*, pp. 73-82, 2009.
- [56] N. Taghipour, A. Kardan, and S. Ghidary, "Usage-based web recommendations: A reinforcement learning approach," *ACM Conf. Recommender Syst.*, pp. 113-120, 2007.

응웬시띤 (Si-Thin Nguyen)



2006년 8월: 베트남 다낭 기술
대학교 정보 기술 졸업
2012년 8월: 베트남 다낭 대학
교 컴퓨터 과학 석사
2017년 9월~현재: 송실대학교
IT정책경영학과 박사과정
<관심분야> 머신 러닝, 추천 시

스템전자공학.

[ORCID:0000-0002-2583-1033]

최은진 (Eun-Jin Choi)



2001년: 동국대학교 컴퓨터공
학과 학사
2017년: 송실대학교 대학원 IT
정책경영학과 석사
2019년~현재: 송실대학교 IT정
책경영학과 박사과정

<관심분야> 중소기업정보화, 정보보호정책, 개인정
보보호, 정보시스템 개발 및 운영

[ORCID:0000-0002-7175-6441]

류호균 (Ho-Gun Rou)



2005년 2월: 고려대학교 학사
2008년 2월: 고려대학교 경영
학 석사학위
2019년 9월~현재: 송실대학교
IT정책경영학과 박사과정
<관심분야> IT보안, 빅데이터,
인공지능, IoT, IT정책

[ORCID:0000-0002-1407-9317]

김광용 (Gwang-Yong Gim)



1984년 2월: 고려대학교 학사
1991년 8월: 조지아 주립대학
보험수리학 석사학위
1995년 8월: 조지아 주립대학
경영학 박사학위
<관심분야> 4차산업혁명, 빅데
이터 및 인공지능, 플랫폼, 블

록체인, 혁신적 비즈니스모델링

[ORCID:0000-0002-6921-1071]