# 맵리듀스를 이용한 사용자 기반 협업 필터링 추천 기법

윤소영* · 윤성대*

*부경대학교

# User-based Collaborative Filtering Recommender Technique using MapReduce

So-young Yun* · Sung-dae Youn*

*Pukyong National University

E-mail : ysmallzero@pknu.ac.kr, sdyoun@pknu.ac.kr

## 요 약

네트워크와 모바일 기기의 확산으로 데이터가 폭발적으로 증가하고 있으며 기존의 추천 기법으로는 급증하는 데이터를 효율적으로 처리하는데 문제가 있다. 따라서 가장 널리 사용되는 추천 기법인 협업 필터링 기법의 확장성 문제를 어떻게 해결할 것에 대한 연구들이 진행되고 있다. 본 논문에서는 협업 필터링 기법에 분산 병렬처리 방식인 MapReduce를 적용하여 확장성 문제를 줄이고 정확성을 높이는 기법을 제안한다. 제안하는 기법은 사용자 기반 협업 필터링 기법에 MapReduce와 색인 기법을 적용하여 유사도 계산에 사용되는 이웃의 수와 이웃의 적합성을 개선하는 방식으로 확장성과 정확성을 개선하는 효과를 기대할 수 있다.

## ABSTRACT

Data is increasing explosively with the spread of networks and mobile devices and there are problems in effectively processing the rapidly increasing data using existing recommendation techniques. Therefore, researches are being conducted on how to solve the scalability problem of the collaborative filtering technique. In this paper applies MapReduce, which is a distributed parallel process framework, to the collaborative filtering technique to reduce the scalability problem and heighten accuracy. The proposed technique applies MapReduce and the index technique to a user-based collaborative filtering technique and as a method which improves neighbor numbers which are used in similarity calculations and neighbor suitability, scalability and accuracy improvement effects can be expected.

## 키워드

Recommender Technique, Collaborative Filtering, Mapreduce, Scalability

## Ⅰ. 서 론

인터넷의 확산으로 e-commerce 기업들은 고객들이 원하는 아이템을 검색하는데 도움을 주기 위해 추천시스템을 도입하였다. 추천 시스템에서 가장 많이 사용되는 방식은 유사한 선호도를 가진 사용자들의 아이템에 대한 선호도를 바탕으로 목표 고객에게 아이템을 추천하는 협업 필터링 기법이다[1]. 그러나 협업 필터링 기법은 확장성의 문제를 가지고 있으며, 이러한 문제점을 해결하기 위한 다양한 방법들이 제안되어왔다. 최근에는 모바일 기기의 확산으로 데이터가 폭발적으로 증가하고 있는 상황에서 데이터 처리 비용과 효율성 향상을 위해 빅데이터 처리 기법들을 적용한 연구들이 이루어지고 있다.

본 논문에서는 협업 필터링 기법에 분산 병렬처리 방식인 맵리듀스를 적용한 기법을 제안하고자 한다. 제안하는 기법은 맵리듀스를 적용하여 데이터를 원하는 형태로 분산 병렬처리 한 후 이 데이터들을 사용하여 유사도와 예측값을 추출하는 단계에서도 맵리듀스를 적용하는 기법으로 이를 통해 데이터 처리 시간과 비용을 줄이고 정확성을 개선하고자 한다.

## II. 관련 연구

### 2.1 협업 필터링

협업 필터링 기법은 사용자의 선호도와 유사한 선호도를 가진 다른 사용자들의 구매 정보를 사용하여 사용자에게 아이템이나 서비스를 추천하는 기법으로 전자상거래 기업들이 가장 널리 이용하고 있는 추천기법이다[2].

협업 필터링 기법의 단계는 먼저 사용자들이 아이템에 대해 평가한 데이터를 기반으로 사용자-아이템 매트릭스 생성한다. 다음으로 목표 사용자와 가장 유사한 사용자들을 찾기 위해 사용자 간에 유사도를 계산하고 이를 기준으로 최근접 이웃을 선정한다. 마지막으로 최근접 이웃들의 평가값으로 목표 사용자가 구매하지 않은 아이템들에 대해 평가값을 예측하고 상위 N개의 아이템을 추천하는 단계로 구성된다[3].

### 2.2 MapReduce

맵리듀스는 가장 인기 있는 분산 병렬 처리 프레임워크 중의 하나이다[4]. 맵리듀스 프레임워크에서는 분산 파일 시스템을 통해 데이터를 여러 대의 컴퓨터에 나누어 저장하며, 분산 파일 시스템에서 데이터는 키와 값의 쌍의 형태로 저장된다. 맵리듀스 프레임워크에서 사용자는 메인 함수에서 맵(map) 함수와 리듀스(reduce) 함수를 한 번 또는 여러 번 반복해서 호출함으로써 원하는 연산을 병렬로 수행한다. 여러 컴퓨터에 나누어져 있는 데이터에 각각의 키-값 쌍들을 여러 개의 맵 함수가 병렬로 함께 수행되면서 데이터를 중간 단계의 키-값 쌍으로 변환시킨다. 모든 맵 함수에서 출력된 결과들은 같은 키를 갖는 값끼리 묶이고, 이를 이용하여 다수의 리듀스 함수들은 병렬로 연산을 수행하여 최종 결과를 생성한다. 최종 결과도 입력 데이터와 같이 여러 대의 컴퓨터에 분산 파일 형태로 저장된다. 사용자는 메인, 맵, 리듀스 이 세 가지 함수를 작성하여 간단히 병렬 처리 맵리듀스 알고리즘을 구현할 수 있다[5].

## III. 추천 기법

본 논문에서 제안하는 기법은 대량의 사용자-아이템 평가 데이터를 사용하여 저비용으로 효율적인 아이템 추천을 위해 기존의 사용자 기반 협업 필터링기법에 분산 병렬 처리 방식인 맵리듀스를 적용한 기법으로 사용자에게 아이템을 추천하기 위해 세 번의 맵리듀스를 거친 후 마지막으로 추천아이템을 추출하는 4단계를 거친다.

첫 번째 맵리듀스 단계에서는 사용자들의 아이템 평가 데이터에 색인 기법을 적용하여 아이템별 평가값이 높은 순으로 인덱싱한 후 사용자, 아이템, 사용자의 아이템 평가값 평균을 출력한다.

두 번째 맵리듀스 단계에서는 이전단계의 출력값을 사용하여 사용자간 유사도를 계산한다. 유사도는 피어슨 상관계수식에 정확성을 높이기 위해 두 사용자간 공통평가 아이템 임계값을 적용하여 계산한다. 임계값을 적용하여 계산한 유사도를 사용자별 유사도 리스트 결과로 출력한다.

식(1)은 유사도 계산을 위한 피어슨 상관계수식이다. $R_{i,k}$는 사용자 $i$의 아이템 k에 대한 평가값을 $\overline{R_i}$는 사용자 $i$의 아이템 평가값 평균을 나타낸다. 식(2)는 본 논문에서 제안하는 기법의 유사도 계산식으로 식(1)에 임계값을 적용한 계산식이다. $a$는 사용자, $u$는 이웃 사용자, $\gamma$은 $a$와 $u$가 공통 평가한 아이템 수의 임계값, $I(a \cap u)$는 사용자 $a$와 $u$가 공통으로 평가한 아이템의 수를 의미한다.

$$sim(i,j) = \frac{\sum_{k \in I}(R_{i,k} - \bar{R}_i)(R_{j,k} - \bar{R}_j)}{\sqrt{\sum_{k \in U}(R_{i,k} - \bar{R}_i)^2}\sqrt{\sum_{k \in U}(R_{j,k} - \bar{R}_j)^2}} \quad (1)$$

$$sim'(a,u) = \frac{Max|I(a \cap u), \gamma|}{\gamma} \cdot sim(a,u) \quad (2)$$

세 번째 맵리듀스 단계에서는 이전 두 단계의 출력값을 사용하여 사용자가 평가하지 않은 아이템에 대한 예측값을 계산하고 사용자, 아이템, 예측값을 출력한다. 식(3)은 최근접 이웃과의 유사도를 사용하여 사용자가 아직 평가하지 않은 아이템들에 대한 예측값을 계산하는 식이다. $\overline{r_a}$와 $\overline{r_u}$는 사용자 $a$와 근접 이웃 $u$의 평가값 평균을 $r_{u,i}$는 이웃 $u$의 아이템 i에 대한 평가값을 나타낸다.

$$P(a,i) = \overline{r_a} + \frac{\sum_{u \in S(a)} Sim(a,u) \cdot (r_{u,i} - \overline{r_u})}{\sum_{u \in S(a)} Sim(a,u)} \quad (3)$$

마지막으로 네 번째 단계에서는 이전 단계에서 출력된 값들 중 상위 N개의 아이템을 추출하여 사용자에게 추천한다.

## IV. 결 론

본 논문에서는 협업 필터링 기법에 분산 병렬 처리 방식인 맵리듀스를 적용하여 폭발적으로 증가하는 데이터를 효율적으로 처리하고 협업 필터링 기법의 확장성 문제도 감소시키는 기법을 제안하였다. 제안하는 기법은 사용자가 증가할수록 유사도 계산량의 증가로 인한 확장성과 비용 증가 문제를 감소시키기 위해 맵리듀스를 적용하여 데이터를 분산 처리하고, 추천의 정확성을 좀 더 높이기 위해 유사도 계산 시 공통 평가 아이템에

대한 임계값을 적용하여 이웃의 적합성을 개선하는 방식을 사용하였다. 이러한 방식을 통해 데이터 증가에 따른 확장성 문제를 개선하고 사용자에게 좀 더 정확한 아이템을 추천하는 효과를 기대할 수 있다.

## 참고문헌


[1] Manow Papagelisa, Dimitris Plexoosakis, "Qualotative analysis of user-based and item-based prediction algorithms for recommen -dation agents," *ACM Transactions on Informa -tion Systems, 22, vol 1.* pp.116-142, 2004.

[2] M. Pazzani, D. Billsus, "Learning and revising user profiles: the identification of interesting Web sites," *Machine Learning, Vol. 27, No.3*, pp. 313-331, 1997.

[3] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, “Item-based Collaborative Filtering Recommendation Algorithms," Processing of the 10th International World Wide Web Conference, ACM Press, pp. 285-295, 2001.

[4] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis, “Evaluating mapreduce for multi-core and multiprocessor systems,” in Proceedings of the 2007 IEEE 13th International Symposium on High Performance Computer Architecture, HPCA ’07, (Washington, DC, USA), pp. 13–24, IEEE Computer Society, 2007.

[5] 심규석, 김영훈, 이정훈, 김진현, 박윤재, “빅 데이터 분석을 위한 맵리듀스 알고리즘의 최근 연구 동향”, 정보과학회지 32(1), pp. 27-32, 2014.