한국컴퓨터정보학회 동계학술대회 논문집 제24권 제1호 (2016. 1)

하둡 및 스파크 기반의 협력 필터링 추천 알고리즘 연구

정영교*, 김상영*, 이정준O, 윤희용*
*O성균관대학교 정보통신대학
e-mail:{joung0491, impsoft. jungjune86, youn7147}@skku.edu*O

A Study on Collaborative Filtering Recommendation Algorithm base on Hadoop and Spark

Young Gyo Jung*, Sang Young Kim*, Jung-June LeeO, Hee Yong Youn*
*ODept. of Electrical and Computer Engineering, Sungkyunkwan University

● 요 약 ●

최근 사용자들의 추천 서비스를 위해 다른 사용자들의 평가값을 이용하여 특정 사용자에게 서비스를 추천해주는 추천 시스템은 협력 필터링 방법을 널리 사용되고 있다. 하지만 이러한 추천 시스템은 클러스터링 과정에서 이미 분류된 그룹에 특정 사용자가 분류되어 정확히 분류되지 못하고, 사용자들의 평가값 오차가 클 경우 정확하지 못한 결과를 추천하는 문제점이 있다. 본 논문에서는 이러한 문제점을 해결하기 위하여 협력 필터링 알고리즘을 클러스터링 기반으로 분산 환경에서 구현하여, 추천의 효과를 최적화 하는 기법을 제안하며 하둡 및 스파크 기반으로 시스템을 구성하여 협력 필터링 추천 알고리즘을 비교 하였다.

키워드: 협력 필터링 알고리즘(cooperative filtering algorithm), 하둡(Hadoop), 스파크(Spark)

I. Introduction

협업 필터링(collaborative filtering)은 데이터를 사용자의 속성에 맞게 추천해 주는 알고리즘 중 가장 널리 사용되는 알고리즘 중 하나이다.[1] 협업 필터링은 비슷한 행동을 둘 이상의 사용자들은 수행 할 다음 행동 역시 비슷할 것이라는 전제를 바탕으로 기존 사용자들의 행동을 분석하여 새로운 사용자에게 아이템을 추천해 주거나, 새로운 물품을 구매하고자 하는 사용자에게 그 사용자가 원할만한 아이템을 추천해 준다. 하지만 클러스터링 과정에서 이미 분류된 그룹에 특정 사용자가 분류되어 정확히 분류되지 못하고, 사용자들의 평가값 오차가 클 경우 정확하지 못한 결과를 추천하는 문제점이 있다.

본 논문에서는 협업 필터링 알고리즘을 클러스터링 기반으로 분산 환경에서 구현하여, 추천의 효과를 최적화 하는 기법을 제안 한다. 또한 하둡 및 스파크 기반으로 시스템을 구성하여 협력 필터링 추천 알고리즘을 비교 하였다.

II. Preliminaries

1. Related works

1.1 협력 필터링

협력 필터링에는 기존에 사람들이 아이템에 대한 평가를 기준으로 추천을 해주는 Memory-based방식, 데이터를 특정 확률적 모델에 적용하여 아이템을 추천해 주는 Model-based방식, 사용자가 아이템에 대해 평가한 데이터뿐만 아니라 다른 정보도 같이 활용하는 Hybrid 방식이 있다.

Memory-based 협력 필터링에는 비슷한 행동을 한 사용자를 클러스터링 하여 목표 사용자가 속하는 군집에서 다른 사람들이 높은 점수의 평가를 부여한 아이템을 추천 해 주는 사용자 기반의 협력 필터링이 있으며, 사용자 기반이 아니라 아이템 사아의 연관성을 파악하여 비슷한 아이템의 군집을 생성하고, 그 군집 안에서 목표 사용자의 행동 기록을 기반으로 다른 아이템을 추천해 주는 아이템 기반 협력 필터링 기법이 있다.

Model-based 협력 필터링은 데이터를 확률적 모델 등에 적용하여 추천을 수행하는 방식으로, 주로 활용되는 시법으로는 Bayesian-belief Network 기반의 협력 필터링, 클러스터링 기반의 협력 필터링, Markov Decision Process기반의 협력 필터링 등이 존재한다.

Hybrid 방식은 앞서 설명한 Memory-based 방식과, Model-based

방식을 혼합하여 구성하거나, 다른 알고리즘을 결합한 경우를 일컫는다. 표 1은 협력 필터링의 종류를 나타낸다.

Table 1. Type of collaborative filtering algorithms

Type	Implementation
Memory-based CF	Neighbor-based Item/User-based
Model-based CF	Bayesian belief nets CF Clustering CF Regression-based CF Latent CF Dimensionality reduction
Hybrid Recommender	Content-based CF Combining algorithms

본 논문에서는 협력 필터링 알고리즘을 이용하여 제품이 자신의 이전 구매내역에 따라 사용자에 권장하고 항목 간의 유사성과 보유하고 있는 데이터들의 정확한 정밀도를 계산하기 위하여 타니모토(tanimoto) 계수를 사용하는 알고리즘을 제안한다.[2]

III. The Proposed Scheme

본 논문에서는 하나의 마스터 노드와 두 개의 데이터 노드 총 3개의 노드가 있는 클러스터를 사용했다. 또한 Apache의 하둡의 Cloudera 버전 4.1.3(라이브러리 Mahout 버전 0.7)과 스파크 버전 1.5.1에서 실행하여 실험을 실시했다. 제안한 기법은 Apache 스파크 플랫폼과 Scala 프로그래밍 언어를 사용하여 병렬 아이템 기반의 협력 필터링 알고리즘을 구현하였다.

그림 1은 아이템의 유사도 행렬을 생성하고 Mahout 라이브러리에서 아이템 유사 프로그램과 Apache 스파크로 구현한 시스템에 대한 비교 연구 결과를 보여준다.

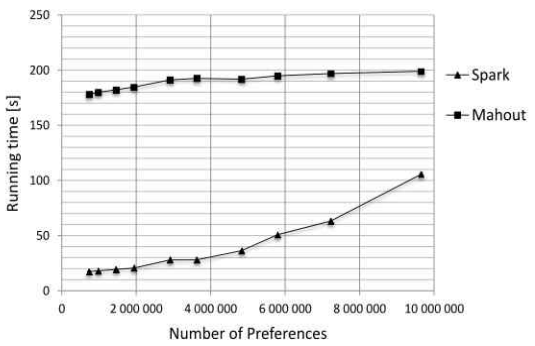

Fig. 1. Item Similarity Job based on Mahout and Spark

아이템 유사 프로그램에서의 평균 계산 시간은 3분 19초이며, 스파크에서는 1분 45초로 거의 절반을 단축 시켰다. 또한 입력 데이터의 양이 낮을수록 더 큰 차이가 있었다.

IV. Conclusions

본 논문에서는 클러스터링 과정에서 이미 분류된 그룹에 특정 사용자가 분류되어 정확히 분류되지 못하고, 사용자들의 평가값 오차가 클 경우 정확하지 못한 결과를 추천하는 문제점을 해결하기 위하여 협업 필터링 알고리즘을 클러스터링을 이용한 분산 환경에서 추천의 효과를 최적화 하는 기법을 제안 했다.

또한 하둡 및 스파크 기반으로 시스템을 구성하여 협력 필터링 추천 알고리즘을 비교를 통하여 스파크가 하둡보다 더 효율적이라는 것을 보여 주었다.

향후 연구로는 기존의 데이터 보다 대량의 데이터를 사용하여 완벽한 신뢰성을 제공하는 연구가 필요하다.

Acknowledgment

본 연구는 BK21Plus 사업, 한국연구재단 기초연구사업 (2013R1A1A2060398), 삼성전자, 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업 (1391105003), 미래부/정보통신방송연구개발사업의 일환으로 수행하였음.

References

[1] Su, X, & K, T. M. “A Survey of Collaborative Filtering Techniques”. Advances in Artificial Intelligence, 2009.

[2] L. K, P. J. “Rousseeuw, Finding groups in data: an introduction to cluster analysis”.