Chapter 6
Collaborative Filtering Based Data Mining for Large Data

Amrit Pal
Indian Institute of Information Technology Allahabad, India

Manish Kumar
Indian Institute of Information Technology Allahabad, India

ABSTRACT

Size of data is increasing, it is creating challenges for its processing and storage. There are cluster based techniques available for storage and processing of this huge amount of data. Map Reduce provides an effective programming framework for developing distributed program for performing tasks which results in terms of key value pair. Collaborative filtering is the process of performing recommendation based on the previous rating of the user for a particular item or service. There are challenges while implementing collaborative filtering techniques using these distributed models. Some techniques are available for implementing collaborative filtering techniques using these models. Cluster based collaborative filtering, map reduce based collaborative filtering are some of these techniques. Chapter addresses these techniques and some basics of collaborative filtering.

INTRODUCTION

In this big technological environment, the amount of data generated is increasing at a very high rate. Computer Engineers at European Council for Nuclear Research (CERN) announced that the amount of data recorded by them for CERN Data Centre has crossed 100 Petabytes of physics data in the last 20 years (CERN, 2015). Experiments in the Large Hadron Collider (LHC) generates huge amount of data, more than 75 Petabytes of this data is generated in last four years. Amazon has about 270 million accounts of active users worldwide (Amazon, 2015). Recommendation for this huge amount of users requires extra efforts. For finding information from this huge and distributed data parallel processing can be used, Google's Map Reduce provides an effective framework for finding information from this data. Hadoop distributed file system for storage of the data and the MapReduce for the retrieval of the relevant information from this data. It is known that the Hadoop framework works well on large file size.

DOI: 10.4018/978-1-5225-0489-4.ch006

Collaborative filtering (CF) is used in recommender system which involves a collection of agents, different viewpoints and data sources. CF main challenges (Su, 2009) are data sparsity, scalability, synonymy, gray sheep, shilling attacks, privacy protection etc. (Linden, 2003). There are three types of collaborative techniques available Memory-based CF, Model-based CF and Hybrid recommenders. Chapter will address, the scalability challenges in performing the collaborative filtering on large datasets, clustering based collaborative approach available for collaborative filtering on large datasets, Prediction algorithms which can be used for a parallel analysis of the datasets using collaborative filtering techniques, challenges in the algorithm design for collaborative filtering on large datasets, real time approach for collaborative filtering of data.

COLLABORATIVE FILTERING

It's a rating system where a user provides his/her response in a specific domain, these responded values by the user helps in recommending the next items to the similar users. There are two basic methods neighborhood and model-based for selecting the users and find similarity among them (Resnick, 1994).

There are two types of user information in system active users and passive users. The users which are currently using the system are active users and the information stored about the activity and their response for the items is stored in a database act as a passive user or passive user information. The process of neighborhood based filtering (Herlocker, 2002) starts with selection of a sample of users from the set of passive users based on their response to a particular item, basically similarity in their response for that item.

The prediction process for an item from item set to an active user can be described as:

- Select a set of passive users based on their similarity with the active user.
- Calculate the mean rating for the active and passive users.
- To measure the similarity Pearson correlation coefficient can be used.

$$w_{a,u} = \frac{\sum_{i \in I} \left(ur_{a,i} - \overline{ur_a} \right) \left(ur_{u,i} - \overline{ur_u} \right)}{\sqrt{\sum_{i \in I} \left(ur_{a,i} - \overline{ur}_a \right)^2 \sum_{i \in I} \left(ur_{u,i} - \overline{ur}_u \right)^2}}$$

- Select users which are having high similarity value corresponding to an active user.
- Use this weight for calculating the weighted average of the deviations from the neighbor's mean as:

$$p_{a,i} = \overline{ur_a} + \frac{\sum_{u \in K} \left(\mathrm{u}r_{u,i} - \overline{ur_u} \right) \times w_{a,u}}{\sum_{u \in K} w_{a,u}}$$

here: