Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 50 (2015) 229 – 234

Procedia
Computer Science

2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Performance and Quality Assessment of Similarity Measures in Collaborative Filtering Using Mahout

Saikat Bagchi*
Indian Institute of Technology
Kharagpur, India
sbagchi1982@gmail.com

ABSTRACT

Recommendation systems use knowledge discovery and statistical methods for recommending items to users. In any recommendation system that uses collaborative filtering methods, computation of similarity metrics is a primary step to find out similar users or items. Different similarity measuring techniques follow different mathematical approaches for computation of similarity. In this paper, we have analyzed performance and quality aspects of different similarity measures used in collaborative filtering. We have used Apache Mahout in the experiment. In past few years, Mahout has emerged as a very effective and important tool in the area of machine learning. We have collected the statistics from different test conditions to evaluate the performance and quality of different similarity measures.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Measurement Techniques, Performance attributes

General Terms

Performance, Measurement

Keywords

Performance and Quality of Similarity Measures, Performance of Mahout-based Recommendation, Performance of User-based Recommendation, Analysis of Similarity Measures, Similarity Measures in Collaborative Filtering

1. INTRODUCTION

Recommendation systems use knowledge discovery and statistical methods for recommending different kind of items to users. At present e-commerce systems offer millions of products for sale. Customers of e-commerce systems often have very little or no knowledge about all offerings provided by those. e-Commerce systems have to predict preferences of customers and recommend products to them to optimize sales. A recommendation system may collect preferences of a customer for different items and recommend new products to him/her predicting his/her preferences for those products. Recommendation techniques play very important role in social networking and other online services like online news service, music/movie service etc. where presentation of personalized items to users is a very important aspect of business. There are various types of techniques for recommendation. Collaborative filtering, content-based recommendation, hybrid recommendation etc. are well-known approaches for generating recommendations. In collaborative filtering approaches of recommendation, items are recommended to a customer by assessing preferences of other customers who are in the neighborhood based on their historically similar taste to the first customer, so similarity-measure is a significant aspect of collaborative filtering.

In this paper we are going to analyze performance and quality aspects of recommendation using different types of similarity measures provided by Apache Mahout. Apache Mahout is an open-source project, which provides scalable implementations of machine learning techniques like collaborative filtering, clustering, classification etc. We will use Movie Lens data from Group Lens dataset for the experiment.

In sections 2 and 3 we will mention summary about related work, overview about recommendation system, definitions of similarity measures and Mahout, which are used in our assessment. In section 4 we will explain our work on performance and quality assessment of similarity measures used in recommendation system.

2. RELATED WORK

Recommender systems have emerged to help users to navigate through large volume of online content. Many online search systems, e-commerce websites, online news services, online multimedia services etc. are exploiting the benefits of recommendation systems in providing extra mileage to their business. Works on evaluation of recommendation systems include Herlocker et al.'s [8] survey and Shani and

[*] Student in Master of Technology at Indian Institute of Technology, Kharagpur, India.

doi:10.1016/j.procs.2015.04.055

Gunawardana's book [13]. There have been several other works on this topic. In almost 50% of the studies on benchmarking of recommendation systems, open data sets have been used; almost similar amount of studies presented information on test/training splits. Very less number of studies used open dataset, open framework, and provided all necessary details for replication of experiments and results. Algorithmic details have been disclosed in almost 25% of the studies. Said et al. [12] have performed comparative study and benchmarking of recommendation systems implemented using separate open source frameworks and open data sets and tried to address the issues related to replication of experimental results. Owen et al. has provided some details about comparative analysis of different similarity measures using Mahout in their book [1], but that is not complete w.r.t. the above mentioned parameters for replication. There is a need of comparative analysis of similarity measure algorithms with open dataset, open framework with disclosure of full details about algorithms and environments for facilitating future study and validation on benchmarking of recommendation systems.

3. OVERVIEW ABOUT CONCEPTS & TOOLS USED IN THE ASSESSMENT

3.1 Recommendation System

Recommendation System, a sub-class of information filtering system, helps in predicting top-N preferred items for a user. Recommendation techniques follow mainly following approaches: collaborative filtering, content-based recommendation and hybrid recommendation. Collaborative filtering methods build a model using information about past purchases or ratings provided by users. A model may also be created based on decisions (preference ratings or selection of items) taken by similar users. This model may be used for prediction of preference rating for a given item. In content-based methods, features of an item are compared against features of other items to recommend items. In collaborative filtering process a large amount of information on a user is required to make accurate predictions (cold-start problem), where as content-based recommendation needs very little information to get started. Following subsection gives a summary about collaborative filtering method.

3.1.1 Collaborative Filtering

Collaborative filtering methods analyze large amount of information about preferences of users and predict preferences of similar users for recommending items. In collaborative filtering method an accurate prediction of preferences of a user and recommendation of items is possible without any need for detailed analysis of item features. A basic assumption in collaborative filtering is that users would like similar kinds of items as they have liked in past.

Collaborative filtering methods suffer from issues like – cold start, scalability and sparsity.

Following section describes about similarity measurement techniques, which are used in collaborative filtering methods.

3.2 Similarity Measures

A similarity measure or similarity function is a real-valued function that quantifies the similarity between two objects. Although no single definition of a similarity measure exists, usually similarity measures are in some sense the inverse of distance metrics: they take on large values for similar objects and either zero or a negative value for very dissimilar objects.

One of the preferred approaches to collaborative filtering (CF) recommenders is to use k-Nearest-Neighborhood (kNN) classifier, which is dependent on defining an appropriate similarity or distance measure. Definitions[1] of some popular similarity measures, which are used in our experiment, are given below:

3.2.1 Euclidean distance

Mathematical definition of Euclidean distance measure is given below for two objects x and y:

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Here n is number of dimensions (attributes) and x_k and y_k are k^{th} attributes (components) of data objects x and y

3.2.2 Minkowski distance

Minkowski distance is a generalized distance measure and is represented mathematically as below:

$$d(x,y) = (\sum_{k=1}^{n}|x_k - y_k|^r)^{\frac{1}{r}}$$

Here r is degree of distance. Depending on the value of r, generic Minkowski distance is known with specific names:

- For r = 1, City block (Manhattan, taxicab or L_1 norm) distance
- For r = 2, Euclidean distance
- For $r \to \infty$, Supremum (L_{max} norm or L_∞ norm) distance, which corresponds to computing the maximum difference between any dimensions of k objects.

3.2.3 Cosine similarity or L_2 Norm

Cosine similarity is the measure of similarity between two vectors of an inner product space that measures the cosine of angle between them.

$$\cos(x,y) = \frac{(x \bullet y)}{\|x\|\|y\|}$$

Here • indicates vector dot product and ||x|| is the norm of vector x.

3.2.4 Pearson correlation

Pearson correlation score checks how highly 2 variables are correlated. A Pearson correlation coefficient is represented as below:

$$Pearson(x,y) = \frac{\sum(x,y)}{\sigma_x \times \sigma_y}$$

Here ∑ is the covariance of data points x and y and σ is the standard deviation.

[1] http://en.wikipedia.org