**Towards Enhancing Performance of Collaborative Filtering through Advanced Similarity Measures**

Ali A. Amer, Taiz University, Yemen

Loc Nguyen, Loc Nguyen’s Academic Network, Vietnam

**Abstract**

In collaborative filtering (CF), how to provide favorite to users depends on analyses of users’ preferences and existing correlation between their preferences. One of popular algorithms in CF is nearest neighbors (NN) algorithm on which we focus when doing this research. Accuracy of NN is mainly affected by both the similarity measure, which is a critical factor, and the approach used to find neighbors of each user/item. Some frequently used tradition measures like cosine and Pearson correlation coefficient fail to reach the desired accuracy as they place a complete emphasize on the co-related ratings and disregard the non-co-rated items when users’ correlations are investigated. Even though they give an acceptable accuracy, they fail to address data sparsity problem, known as cold-start problem, in an effective manner. In this research project, therefore, some approaches to effectively tackle CF under the sparsity problem are presented. Several similarity measures either combined or newly proposed are developed in an attempt to efficiently increase CF accuracy while considering the problem of data sparsity. These measures seek to address the lined-above faults so that the data sparsity is significantly tackled and CF accuracy is enhanced. The novelty of these measures represented in their power to solve the problem of similarity across finding the relationship among the correlated and non-correlated users at the same time. These measures will be later integrated with semantic and singularity concepts to provide more powerful results. Moreover, a new variation of NN algorithm is going to be introduced. The comparison with the state of art in each phase of project is going to be held. For example, to indisputably prove our claims, these measures along with the state-of-art measures are extensively tested, using the most widely used metrics of evaluation like precision, recall, F-measure, MAE, and MSE, on several benchmarked dataset including MovieLens 100K, MovieLens 1M, and Netflix.

**Keywords**: recommendation system, collaborative filtering, similarity measure, k-nearest neighbor (KNN) algorithm.

**Project ID:** **ASIM** which is abbreviation of advanced similarity measures.

**Time period:** Two-year period.

**Ethical clearance:** We confirm to obey ethical criteria when doing ASIM project.

**Key goal:** Providingadvanced similarities of NN algorithm for enhancing CF.

**1. Statement of research problem**

CF is an important approach for recommendation system, which recommends an item to a user if her/his neighbors are interested in such item. One of popular algorithms in CF is NN algorithm. Finding neighbors of a user is the heart of NN algorithm, which in turn depends on similarity measures between users. Therefore, the main problem of NN algorithm is how to calculate / select good similarity measures when accuracy of a similarity measure is affected by sparsity of dataset in which many items are not rated. Some measures improve their accuracy by concerning both rated items and non-rated items whereas other ones concern singularity of ratings and user agreement in rating items. Even some hybrid measures try to combine other measures with expectation that the accuracy will be improved. Especially, statistical techniques are applied into calculating measures. In general, ASIM tries to solve the problem of NN algorithm by two following tasks:

* Testing and evaluating many current measures in order to draw their strong points and drawbacks.
* Improving current measures and proposing new measures by two approaches such as combining rated items and non-rated items in calculating similarity of two users and combining different measures in order to take advantages of their strong points.

**2. Literature review**

(*Ali Amer refers to our papers in order to compose this section*)

**3. Goals of the research**

ASIM project has two main goals:

1. Despite the fact that the similarity and distance measures used in recommendation system are broadly available, their efficiency and effectiveness in CF has not yet paid the sufficient attention to broadly benchmark their impact on CF process. Driven by this fact, a comprehensive experimentally done thoroughly conducted evaluation for the most widely used similarity measures in CF literature along with proposing a new similarity measures, is going to be extensively introduced. The performance of all measures (including newly proposed ones) will be compared with each other in terms of efficiency and effectiveness. For each measure, efficiency (time and complexity) and effectiveness (precision, recall, F-measure, and MSE) are drawn under both user-based and item-based models. The ultimate purpose is to find measures which are meant to be universally best for most CF cases including data sparsity problems. Moreover, the drawbacks and shortcoming for the concerned measures will be defined by performing a thorough analysis to carefully identify their effects on CF process. This work will therefore help scholars as a reliable framework to study the behavior of the most common similarity measures and apply the selected measures according to their needs and conditions. Of course, the testing summary of concerned measures along with the reliable framework for CF and NN algorithm regarding similarity measures is provided as derived result of the first goal.
2. We will improve current measures and propose new measures by two combined approaches such as combining rated items and non-rated items in calculating similarity of two users and combining different measures in order to take advantages of their strong points. In other words, combination of measures is the second goal. It is expected that combined (hybrid) measures as derived result of the second goal will improve accuracy of NN algorithm for CF.

**4. Research methodology**

**4.1. Experimental design**

Because NN algorithm uses a similarity measure to find out nearest neighbors, we evaluate measures by executing NN algorithm on datasets like Movielens and Netflix. For example, Movielens dataset has two popular versions: MovieLens 100K and MovieLens 1M. MovieLens 100K has 100,000 ratings from 943 users on 1682 movies (items). MovieLens 1M has 1,000,209 ratings from 6,040 users on 3,900 movies. Every rating in Movielens dataset ranges from 1 to 5. In the experiments, dataset Movielens is divided into 5 folders and each folder includes training set and testing set. Training set and testing set in the same folder are disjoint sets. The ratio of testing set over the whole dataset depends on the testing parameter *r*. For instance, if *r* = 0.1, the testing set covers 10% the dataset, which means that the testing set has 10,000 = 10%\*100,000 ratings and of course the training set has 90,000 ratings. In our experimental design, parameter *r* has nine values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9.

Four main metrics to assess NN algorithms are mean absolute error (MAE), mean squared error MSE), recall, and precision. Quality of a CF algorithm like NN algorithm depends on both estimation and recommendation. Estimation ability is ability to estimate or predict exactly missing values. Recommendation is ability to provide list of recommended items which is as suitable as possible to users. Hence, we do not follow previous researches to focus on recommendation tasks with metrics MAE, precision, and recall. Instead we divide our tests into two processes such as estimation and recommendation as follows:

* In estimation process, we use MAE and MSE to evaluate accuracy of NN algorithm.
* In recommendation process, we use precision and recall to evaluate quality of NN algorithm.

Hence, different metrics (MAE, MSE, recall, precision) are used for different evaluation processes (estimation and recommendation). This independent evaluation allows use to test measures more objectively, in which estimation process focused on accuracy of NN algorithm and recommendation process focuses on quality of NN algorithm.

Moreover, the problem in recommendation is how to determine the number of recommended items denoted *C* which is the length of recommended vector. As a convention, *C* is called recommendation count. The count *C* cannot be too small or too large. If it is too small, evaluation is inaccurate. Otherwise, if it is too large, evaluation task will run slowly. Some researches set fixed number whereas other researches changed such number over some values such as 10, 20, and 100. We proposed a method to determine *C* based on dataset with purpose that *N* will be more accurate and objective. The proposed method is dynamic and takes advantages of the sparse ratio of dataset.

**4.2. Techniques and tools**

As aforementioned, two main goals of ASIM project are to test concerned measures and to propose new measures. For proposing new measures, we focus on combination techniques to derive new measures by taking advantages of strong points of current measures. Moreover, we concern statistical techniques and theories related to information retrieval, vector manipulation, collaborative filtering and recommendation system. Especially, we also research, implement, and improve NN algorithm. User-based rating matrix is the matrix in which rows indicate users and columns indicate items and each cell is a rating which a user gave to an item. When user-based rating matrix is transposed into item-based rating matrix in which every vector is item rating vector, equations for these measures are not changed in semantics. NN algorithm for user-based rating matrix becomes user-based NN algorithm and NN algorithm for item-based rating matrix becomes item-based NN algorithm. In our current implementation, item-based NN algorithm splits one execution of user-based NN algorithm into *n* executions where *n* is equal to the number of missing values.

The environment for testing similarity measures regarding NN algorithm is Hudup framework. Hudup framework available at http://www.hudup.net is the recommender framework dedicated to scientists and software developers who create or deploy recommendation solutions and algorithms in e-commerce and e-learning. Hudup is composed of three modules:

* The infrastructure to set up recommendation algorithms including NN algorithm.
* The evaluation system to measure recommendation algorithms according to metrics.
* The simulation environment to execute and test recommendation algorithms before deploying them in real-time applications.

**5. Research schedule and deliverables**

This project consists of three prime phases. Each phase has its own outcomes and expected outputs as follows:

1. In the first phase, a comprehensive experimentally driven study for the most widely used similarity measures in CF, in terms of all performance factors including efficiency (time and complexity) and effectiveness (MAE, MSE, precision, recall, F-measure) will be produced. As a result, a research paper containing this study will be published. We can propose new measures in the first phase as our research effort. The first phase corresponds with the first goal.
2. In the second phase, our proposed similarity measures will be introduced in this phase. The proposed measures are based on combination of other measures. To show supremacy of our measures, a thoroughly made comparison with the state of art measures that are studied in phase 1 will be drawn. The second research paper will reflect the outputs of this phase. The second phase corresponds with the second goal.
3. The third phase is auxiliary phase as our research effort because two previous phases focus on completing two our main goals. So, the third phase has two options. The first option is to publish a tool for developing and testing similarity measures for CF. Such tool is named ASIM tool which is based on Hudup framework. The second option is much more ambitious when we try to combine NN algorithm (with the best measure produced from phase 2) and singular vector decomposition (SVD) technique to derive a combined algorithm called NN+SVD as a group of researchers who won Netflix prize in recommendation system did. However, we expect that our NN+SVD algorithm is better with support of the best similarity measure produced from phase 2.