Enhancing Performance of Recommendation Systems through Intelligent Collaborative Filtering Techniques: Future Work Proposal

Ali A. Amer
Computer Science Department, Taiz University, Yemen, aliaaa2004@yahoo.com

Abstract

This project introduces advanced AI techniques to promote collaborative filtering (CF) performance. Using the K-nearest neighbor algorithm (KNN), a comprehensive CF framework for almost 60-90 similarity measures will be presented. Considering the data sparsity problem, new similarity measures are developed to promote CF performance. The novelty of these measures is embedded in their simplistic design, efficiency, and capability to overcome the data sparsity by addressing the relationships among the correlated and non-correlated users/items. To achieve highly sustainable performance, word2vector (C-BOW and Skip Gram) is utilized. Furthermore, advanced variations of KNN, support vector machine (SVM), multi nominal Bayesian (MNB), and Convolutional Neural Network (CNN) are evolved, and integrated with the top-performer similarity measures to further improve CF performance. Finally, sentimental analysis is leveraged to boost CF rendering. All proposed techniques are evaluated against the state-of-the-art rivals regarding effectiveness and efficiency on several benchmarked datasets, like MovieLens (100K, 1M), and Netflix.

Problem Statement

In collaborative filtering (CF), offering favorites to users is essentially depending on the analyses of users' preferences and the existing correlation between their preferences using the K-nearest neighbors (KNN) algorithm. KNN performance is mainly affected by both: similarity measure and dataset sparseness. According to CF literature, some frequently used traditional measures like Cosine and PCC do not reach the desired performance as they place a complete emphasize on the co-related ratings and disregard the non-co-rated items when users' correlations are investigated. Although acceptable accuracy, they fail to effectively address the data sparsity problem (cold-start problem). Consequently, a few studies have been proposed to solve this problem. Nevertheless, these studies still suffer from data sparsity problem [1,31]. Moreover, the vast majority of these studied evade testing measures on item-based model due to the task complexity. Finally, till recently, there has been no efficient solution benchmarked to be universal and stable for CF performance. On the other hand, besides the way in which KNN is being designed, the key problem of KNN is the similarity computation when CF performance is affected by dataset's sparsity, in which the majority of items are not rated. Therefore, this project comes to effectively address these problems through the following:

1. Introducing comprehensive framework for CF similarity measures (sixty "60-90" measures). Relevant CF earlier studies will be investigated deeply. A thoroughly made "unprecedented" experimental analysis is made for the considered studies (including the newly-proposed ones) regarding their effectiveness and efficiency. Both users-based and item-based models are thoroughly examined. All similarity measures are tested and evaluated on several benchmarked CF datasets. Moreover, the top performer similarity measures will be evaluated using word2vector models (C-BOW and Skip Gram) to benchmark their performance using neural networks.

2. Presenting effective similarity measures to solve the sparsity problem, which influence CF performance significantly. Moreover, as appropriate, the similarity measures are combined by two approaches either by

implicitly combining rated items and non-rated items in calculating the similarity of two users, or by combining different measures in order to take advantages of measures' combination.

3. Enhancing CF performance through: (1) Presenting a sophisticated KNN variation, (2) Consolidating KNN with neural networks [28-29], (3) Using Sentimental analysis to address the sparseness of datasets using natural language processing (NLP) [30], (4) Evolving advanced variations of SVM and MNB, and (5) Using CBOW and Skip Gram models, w2v-based CNN will be applied and compared with baseline CNN.

Goals of the research

The main goal of the proposed research is to take a leading role in producing new knowledge and innovative technologies in the domain of Artificial Intelligence that have direct impact on the knowledge economy especially the recommendation-based online retails, by:

1. The efficiency and effectiveness of the similarity/distance metrics used in recommendation system have not yet received the convincing attention to broadly benchmark their impact on CF process. Consequently, an expansively investigation of the widely used similarity measures in CF literature along with proposals of new similarity measures is rigorously highlighted in this work. 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 (accuracy, precision, recall, F1 measure, and MSE and RMSE) will be examined 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 (that would contain almost 60-90 measures) to study the behavior of the similarity measures and select measures according to their needs/conditions.
2. Developing new similarity measures by implicitly combining the rated items and non-rated items in calculating similarity of users and/or combining different measures to take advantages of measures' combination. It is expected that combined (hybrid) measures will highly improve accuracy of KNN algorithm, and CF performance as well.\
3. Using word2vector models, performance of top-performer measures will be benchmarked on several datasets.
4. Evolving an advanced KNN variation using the newly proposed dominant set to enhance RS quality and expedite classification and prediction rate while maintaining higher accuracy. By this technique, online prediction is expected to be more accurate and reliable.

Enhancing CF performance through: (1) Presenting a sophisticated KNN variation, (2) Consolidating KNN with neural networks, (3) Using Sentimental analysis to address the sparseness of datasets using natural language processing [30], (4) To analyze CF behavior using different classifiers, we seek to evolve advanced variations of SVM and MNB, and, later combine top-performer similarity measures with them, and finally, (5) Using CBOW and Skip Gram models, w2v-based CNN will be applied and compared with baseline CNN.

. Literature review

In CF literature, dozens of similarity measures and machine learning models are separately applied on CF-based recommendation Systems (RS), and their effects recorded either in user-based or item-based models [1-2, 3]. For instance, [4-7] utilized the contextual information of users to present similarity measures based on the singularity factor. While [6] proposed Context Based Rating Prediction (CBRP) to find the context score of each nominated user for the considered pair "user-item". In [4] the Mean-Jaccard-Differences, MJD, was presented to improve the

traditional similarity measures using the singularity of user ratings. In [8], several traditional similarity measures (PCC, cosine and some distance metrics) were combined to introduce a combined similarity measures. A similarity measure called (IPWR), short for improved PCC weighted with user rating preference behavior (RPB), was presented in [9] by combining PCC with RPB. IPWR was seen superior over the state-of-art measures. PIWR was shown better than traditional measures.

Similarly, [10, 11] produced new measures to select neighbors based on the neighborhood union and intersection. Like jaccard, these measures was reliant on the shared items when finding the neighbors of user of interest. The similarity between items would have zero value when there were non-shared items between the intended users, making these measures faulty in this case. On the other extreme, using Bhattacharyya coefficient, several studies has been dedicated to solve the dilemma of data sparsity [12-14]. Meanwhile, [15] presented subspace clustering-driven measure to tackle both problems of the high dimensionality and the data sparsity. A fast neighbor user searching (FNUS) approach was proposed in [16] to enhance RS performance via generating the item subspaces into: interested item, neither interested nor uninterested (NINU) item, and uninterested item subspaces. Then, the co-rated item numbers between a target user and other users were computed, and used to find the three subsets of neighbor users for the target user. Through the union of the three neighbor user subsets, the final neighbor user set is drawn. The FNUS was seen having a competitive rendering on huge datasets.

In [17], a linear combination was proposed. Using PSS, Bhattacharya Coefficient, and Jaccard, the preferences and local context of users' behavior and the percentage of shared ratings between each user pair were considered. In [18-19], new measures, based on the global user preference and context information, were proposed to enhance RS accuracy. In [20], a Wasserstein Collaborative Filtering (WCF) technique was proposed. Under user embedding constraint, WCF predicted user preference on cold-start items by minimizing the Wasserstein distance. Ahn [21] investigated the deficits of traditional similarity measures in CF. Using the specific meanings of co-ratings and the explanation of user ratings, author then introduced a heuristic similarity measure called PIP which stands for three semantic heuristics, namely, Proximity, Impact and Popularity. In [22], however, PIP was shown faulty in some cases of data sparsity. So, a PIP-based heuristic similarity model called (NHSM) was introduced to tackle the limitation of PIP effectively.

On the same page, in [23], the data sparsity was demonstrated to have a devastating impact on the performance of recommendation systems. Even though the PIP and NHSM measures [21] provided an improved solution for sparsity problem, the range of values for each component in PIP, in particular, is very high. So, a modified proximity-impact-popularity (MPIP) similarity measure was introduced in [23]. The MPIP expression was designed to close the Gab of PIP measure whose range of values for each component was very high. MPIP was shown better than PIP and other competitive measures. In [24], using cross validation, authors designed three efficient similarity measures to effectively tackle the data sparsity problem, namely, difference-based similarity measure (SMD), hybrid difference-based similarity measure (HSMD), and, triangle-based cosine measure (TA). SMD and TA were proven to be superior comparing with their rivals. In [25], the latent semantic integrated explicit rating (LSIER) scheme was presented to enhance RS performance. The LSIER scheme was designed by combining the probabilistic latent semantic index (PLSI) model which was used to train user's access records, and the probabilistic matrix factorization (PMF) model which was used to give the user feature and service feature matrices. Finally, using the domain sensitivity, in [26], a sentiment-based model with contextual information was developed for RS,

In parallel with the presented-above measures, our proposed project comes to meticulously address all the drawn-above limitations of CF. The cold start "data sparsity" and high dimension problems will be thoroughly analyzed and carefully tackled using the proposed measures which would also consider the contextual information of users, the dilemma of high dimensionality, and implicit user's feedback. Most importantly, a comprehensive framework for CF similarity measures, that would include an almost "60-90" measures, is set to be introduced. Moreover, to promote performance of CF, new variations of KNN, CNN, SVM and MNB will be developed, and later merged with top-performer similarity measures.

Research methodology

5.1. Data Collections and Experimental Design

All similarity measures, top performers in particular, and their combinations will be evaluated and examined using the ML and DL models, on the most broadly-used datasets including (but not limited) Movielens (100K, 1000K), Tencent, Epinions, MovieTweetings, DePaulMovie, InCarMusic. In doing so, this work is comparable and reproducible with the previous/ future works.

5.2. Similarity Measures

Investigating the most widely utilized similarity measures including the newly-proposed (roughly 60-90 similarity measures) in a concrete step to build the whole framework. Their strength and weaknesses theoretically and empirically is analyzed so a comprehensive experimental guide (in terms of effectiveness and efficiency) is introduced. This will help researchers/Scholars to: (1) find out that which similarity measure is better and under which conditions, and (2) identify the measures that could be applied in any circumstances like data sparsity conditions.

5.3. Techniques and Tools

As aforementioned, more goals of this project are decided. Undertaking a thorough testing of the concerned measures and the new proposed measures. Moreover, we involve statistical techniques and theories related to information retrieval, vector manipulation, collaborative filtering, and recommendation system. The KNN algorithm is expected to be significantly enhanced by either merging it with new measures or combining it with neural networks. Moreover, CNN, SVM and MNB are expected to significantly promote CF performance, especially when their advanced variations are developed. Moreover, they will be integrated with the top perform similarity measures to produce a new variations of CNN, SVM and MNB.

The environment for performing this project - project including similarity measures and other classifiers including KNN algorithm and other machine learning models, is Java and Python. While Java (using Hadoop) is used to form the entire framework of similarity measures, Python has several libraries that support building and analysis of recommendation systems using ML and DL models. Scikit-learn surprise library of python provides platform for building recommendation system. Other libraries like numpy and pandas will be used for data processing and analysis. In addition to Python, R will be used for further statistical analysis as appropriately necessary. The combinations of machine learning models and deep learning approaches are expected to maximally increase RS performance.

5.4. Semantic Based Similarity

To further overcome the sparseness problem, a semantic based similarity measure/algorithm will be utilized using principle component analysis (PCA) and Latent semantic analysis (LSA) to consider CF performance before and after reduction. The sentimental analysis technique along with KNN will also be leveraged as annotation will be introduced to define the relationship between ratings.

1.5. Machine learning Models

Besides using KNN< our work will use three key ML models, namely, support vector machine (SVM), multi nominal Bayesian (MNB), and Convolutional Neural Network (CNN). We will use baseline and seek to present advanced versions of these models.

5.5. Evaluation Metrics

Six main metrics to assess CF performance will be used for different evaluation processes (estimation and recommendation), namely, mean absolute error (MAE), mean squared error MSE), RMSE, recall, precision and F1 measure. Quality of a CF algorithm like KNN algorithm depends on both estimation and recommendation. This independent evaluation allows us to test measures more objectively, in which estimation process focused on accuracy of NN algorithm and recommendation process focuses on quality of NN algorithm. Unlike previous studies, in this work each process is assessed severally according to its best fit metrics. Run time and complexity of each measure are involved as well to assess techniques' efficiency.

Deliverables

Five main phases, and each phase consists of several challenging tasks. Each of the four phases has its essential requirements to be completed including tasks like data collection, data pre-processing, features extraction, data sampling, implementation of data clustering and classification, predictions, and recommendation, reproducing literature studies, experiments conduction, data entering, comparisons drawing, results collection and analysis, results' visualization, papers' draft preparation, and finally project documentation and results dissemination. While the main achievement of this project represents in producing an effective tool for collaborative filtering, this project consists of four prime phases to build this tool. 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, RMSE, precision, recall, F-measure) will be produced. The evaluation will be made using both user base and item-based models. As a result, a review research paper containing this study will be published.

2. In the second phase, our proposed similarity measures will be introduced. To show the improvement brought by 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 framework of CF similarity measures will come into fruition in this phase.

3. The third phase concentrates on enhancing KNN algorithm through either presenting a new variation of this algorithm based on dominant set theory or integrating KNN algorithm with the best measure produced from phase (2).

4. The fourth phase will focus on integrating singular vector decomposition (SVD) and neural network techniques with machine learning techniques to derive a robust combined algorithm called (ML+SVD). As matter of fact, we are driven by the CF group of researchers [27] who won Netflix prize in recommendation system. As consequence, we expect that our proposed ML with SVD algorithm will outperform their algorithm significantly.

5. The fifth phase will focus on drawing CF performance using word2vector using top-performer similarity measures and CNN model.

References

[1] Khojamli, H. & Razmara, J. Survey of similarity functions on neighborhood-based collaborative filtering. Expert Systems with Applications 185, 115482 (2021).

[2] Portugal, I., Alencar, P. & Cowan, D. The use of machine learning algorithms in recommender systems: A systematic review. Expert Systems with Applications 97, 205–227 (2018).

[3] Widiyaningtyas, T., Hidayah, I. & Adji, T. B. User profile correlation-based similarity (UPCSim) algorithm in movie recommendation system. Journal of Big Data 8, (2021).

[4] Bobadilla, J., Ortega, F., Hernando, A. & Bernal, J. A collaborative filtering approach to mitigate the new user cold start problem. Knowledge-Based Systems 26, 225–238 (2012).
[5] .Jin, Q., Zhang, Y., Cai, W. & Zhang, Y. A new similarity computing model of collaborative filtering. IEEE Access 8, 17594–17604 (2020).
[6] Ali, W. et al. Context-aware collaborative filtering framework for rating prediction based on novel similarity estimation. Computers, Materials and Continua 63, 1065–1078 (2020).
[7] Huynh, H. X. et al. Context-Similarity Collaborative Filtering Recommendation. IEEE Access 8, 33342–33351 (2020).
[8] Choi, K. & Suh, Y. A new similarity function for selecting neighbors for each target item in collaborative filtering. Knowledge-Based Systems 37, 146–153 (2013).
[9] Ayub, M. et al. Modeling user rating preference behavior to improve the performance of the collaborative filtering based recommender systems. PLoS ONE 14, (2019).
[10] Wang, D., Yih, Y. & Ventresca, M. Improving neighbor-based collaborative filtering by using a hybrid similarity measurement. Expert Systems with Applications 160, (2020).
[11] Bag, S., Kumar, S. K. & Tiwari, M. K. An efficient recommendation generation using relevant Jaccard similarity. Information Sciences 483, 53–64 (2019).
[12] Koohi, H. & Kiani, K. Two new collaborative filtering approaches to solve the sparsity problem. Cluster Computing 24, 753–765 (2021).
[13] Cao, H., Deng, J., Guo, H., He, B. & Wang, Y. An improved recommendation algorithm based on Bhattacharyya Coefficient. IEEE International Conference on Knowledge Engineering and Applications, ICKEA (2016) 241–244.
[14] Patra, B. K., Launonen, R., Ollikainen, V. & Nandi, S. A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data. Knowledge-Based Systems 82, 163–177 (2015).
[15] Koohi, H. & Kiani, K. A new method to find neighbor users that improves the performance of Collaborative Filtering. Expert Systems with Applications 83, 30–39 (2017).
[16] Li, Z. & Zhang, L. Fast neighbor user searching for neighborhood-based collaborative filtering with hybrid user similarity measures. Soft Computing 25, 5323–5338 (2021).
[17] Saranya, K. G. & Sudha Sadasivam, G. Modified heuristic similarity measure for personalization using collaborative filtering technique. Applied Mathematics and Information Sciences 11, 307–315 (2017).
[18] Al-bashiri, H., Abdulgabber, M. A., Romli, A. & Salehudin, N. B. A developed collaborative filtering similarity method to improve the accuracy of recommendations under data sparsity. International Journal of Advanced Computer Science and Applications 9, 135–142 (2018).
[19] Gazdar, A. & Hidri, L. A new similarity measure for collaborative filtering based recommender systems. Knowledge-Based Systems 188, (2020).
[20] Meng, Y., Yan, X., Liu, W., Wu, H. & Cheng, J. Wasserstein Collaborative Filtering for Item Cold-start Recommendation. Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization 318–322. (2020).
[21] Ahn, H. J. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. Information Sciences 178, 37–51 (2008).
[22] Liu, H., Hu, Z., Mian, A., Tian, H. & Zhu, X. A new user similarity model to improve the accuracy of collaborative filtering. Knowledge-Based Systems 56, 156–166 (2014).
[23] Manochandar, S. & Punniyamoorthy, M. A new user similarity measure in a new prediction model for collaborative filtering. Applied Intelligence 51, 586–615 (2021).
[24] Amer, A. A., Abdalla, H. I. & Nguyen, L. Enhancing recommendation systems performance using highly-effective similarity measures. Knowledge-Based Systems 217, (2021).
[25] Duan, L., Gao, T., Ni, W. & Wang, W. A hybrid intelligent service recommendation by latent semantics and explicit ratings. International Journal of Intelligent Systems (2021).
[26] Osman, N. A., Noah, S. A. M., Darwich, M. & Mohd, M. Integrating contextual sentiment analysis in collaborative recommender systems. PLoS ONE 16, (2021).