Combining Clustering and Regression Models for Recommending Researchers

Jaimel de Oliveira Lima[1], Elias de Oliveira[1]

[1]Laboratório de Computação de Alto Desempenho
Universidade Federal do Espírito Santo
Vitória, Brasil
jaimel.lima@ufes.br, elias@inf.ufes.br

Abstract. Due to the increase in scientific production, especially in recent years, management and decision support challenge also increase significantly. The task of recommending researchers, for example, to a project is not simple. Even with the proper amount of data, ranking and recommending researchers becomes a challenging process. Despite the different methods, what can happen is that the datasets of an institution or research areas do not have a ranking value, that is, a value that can be used to assess the position of a researcher. Even having a necessary dataset, there is no ranking information for these researchers, and this process of obtaining data for training a model can be costly. We propose to use clustering techniques to support the ranking process, reducing the human effort to obtain examples for models training. Then, we used this dataset to train the regression models and Mean Squared Error (MSE) and Normalized Discounted Cumulative Gain(nDCG) to evaluate them. Tests demonstrate that our solution can support the researchers' recommendation process in an adaptive process to the needs of an organization.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: recommender systems, academic analytics, machine learning, support vector machine, linear regression

1. INTRODUCTION

Recommender Systems (RecSys) can be understood as systems capable of recommending items to users, or users to users, based on previous information of these users or characteristics they have in common [Shah et al. 2017; Sharma and Mann 2013]. In general, models can be content-based (CB), collaborative filtering (CF), or a hybrid model, which, in this case, seeks to avoid the problems that occur in other types. CB systems use information based on the items a user consumes the most. On the other hand, CF systems use information from other users to recommend an item to a specific user [Sharma and Mann 2013; Liang et al. 2016].

Different studies already use the RecSys concepts for topics related to scientific production, researchers [Tang et al. 2012; Liang et al. 2016; Chaiwanarom and Lursinsap 2015], papers [Son and Kim 2018; Sebastian et al. 2017; Wang et al. 2018], citations [Huang et al. 2015; Tang et al. 2012], and academic locations [Yang et al. 2014; Beel et al. 2016; Alhoori and Furuta 2017; Yu et al. 2018]. Among the possibilities of using a RecSys are the processes of ranking and recommending researchers. These processes are essential to select researchers to form a team for a particular project, for a particular activity or even to finance a research project, among other examples.

Many ranking and recommendation models use bibliometric indexes already known in the decision-making process. [Rost and Frey 2011; Dorogovtsev and Mendes 2015; Lima et al. 2013; Bidgoli et al. 2019; Gao and Kumar 2019; Ghani et al. 2019; Maqsood et al. 2020]. These indicators can be used directly for measuring and evaluating the performance of a research area, institution, or a researcher [Davison and Price 2009]. However, one of the difficulties found, in general, is that these indexes do not always reflect the situation of researchers [Sabour 2019]. In addition, scientific production has been increasing, with faster dissemination, the challenges at the time of decision-making and selection of researchers and projects, for example [Alhoori and Furuta 2017].

The authors [Ghani et al. 2019] found that "most of these measures are divergent in nature and follow their pattern for classifying authors" and, in general, the indices are not as significant as one would like. The authors state the fact that the scientific community does not adopt a single index as an ideal is why there is a divergence between specific ranking indexes and datasets.

Still, on the h-index, two specific points need to be considered: (a) the index is directly linked to the research period, and (b) even without publications in a period, the index increases as the author receives citations. These points make the h-index an unfair index when there is a need to compare a junior and a senior researcher [Sabour 2019]. The various propositions of indexes and metrics for ranking researchers demonstrate the challenges in the process. There are several areas of research and different types of metrics, bringing complexity to the process [Jin et al. 2007]. Although indexes do not reflect the needs of a group or institution, it is not feasible to create an index for each new ranking need that occurs.

Another challenge is related to the ranking of researchers is in obtaining data for model training. Despite the amount of data available, these data are not annotated; that is, there is no ranking information available. Furthermore, systems are not always adaptive to allow the ranking, selection, or ranking processes to be adapted to each institution's needs. In general, obtaining data for training machine learning models can be costly, requiring significant human effort in the process. In their studies, [Oliveira et al. 2014] raises this issue and proposes a clustering process to reduce annotation effort to classify opinions tweets about approving a Brazilian law.

Among these initial problems, there is still a challenge inherent to data mining processes: obtaining adequate data for the process. Since the indices proposed in the literature do not reflect the reality and need of an institution at the time of decision-making, we propose a solution that reflects the reality of a group of researchers or institutions. Our work aims to evaluate the combination of clustering and regression for ranking and recommending researchers, in particular, to reduce human effort in this process. This perspective assumes that a recommendation system that uses model-based techniques can have greater scalability and shorter forecasting time, despite being more complex to implement [Tatiya and Vaidya 2014; Zriaa and Amali 2021].

In this sense, the main contributions of this study are: (a) evaluation of the use of unsupervised learning (clustering) for annotation of researchers' ranking data; (b) Understanding the adaptability of models for dynamic processing of researchers; (c) evaluation of regression models for the evaluation of researchers; (d) analysis of the ranking data of an institution for the selection of research projects.

This work is organized as follows. In section 2, there are works related to this study, in addition to a brief review of internal issues. In the section 3, the methodological aspects of the models obtained and experiments carried out are discussed, and also the main results. Finally, in Section 4, we present the work contributions and possibilities for future work.

2. RELATED WORK

In this section, we present work related to our study. The works are related by using the methods or domain of interest to ours.

In their studies, [Oliveira et al. 2014] combine clustering and classification to reduce the effort to classify tweets according to user needs. The authors use an unsupervised learning algorithm (K-Means) in conjunction with a classification algorithm (KNN). The authors' primary objective is to demonstrate the possibility of reducing human effort in the classification process. Documents are initially clustered. If the user with the same label annotates the first and last element of a cluster, all cluster members receive the same label. If they are different, the elements of the cluster are again clustered. The process is iterative, automating part of the annotation process, seeking to reduce effort.

In their work, [Khanam and Alkhaldi 2020] propose a system for recommending undergraduate courses for students who wish to participate in university selection processes. The system analyzes the profile of the students and compares it with the profile of the courses, making the process of choosing the course more straightforward. The solution has three main components: data analyzer, classifier, and visualization. The authors use Random Forest, combined with clustering algorithms. The authors state the results demonstrate that the system is reliable, faster, wiser, and more suitable for selecting colleges for admission. Results for accuracy are greater than 90%.

Another work focused on the academic area is [Pradhan and Pal 2020], that propose CNAVER, a content-based and network-based SR for recommending vehicles for publications. The researchers state that this type of recommendation can help find places more related to the publication to be made, improving the impact and avoiding rejection of good articles due to lack of alignment with the publication vehicle. The authors propose a four-tier architecture. The first layer performs data processing and centrality calculation. The second layer deals with the title (word2vec) and abstract (LDA) process. The third layer has two modules that handle the processing of item pairs and vehicle pairs. In the last module, the calculations of similarities between the abstracts and the vehicle of interest occur. Finally, in the fourth layer, the ranking calculations and the recommendation of the top N vehicles occur. The authors conclude that the proposal presents better solutions for accuracy, precision, nDCG metrics, and measures of diversity and stability of the model.

In their studies, [Ströele et al. 2017] use central measures for clustering and analysis in social networks formed by Brazilian researchers. The study was carried out in two stages: (1) data extraction, transformation, and modeling and (2) analysis of social networks formed by the relationships between researchers. In the analysis phase, the authors propose techniques for visualizing the relationships, representing the groups, and the possibility of filtering, besides considering the local and global importance of the researchers.

Using techniques based on multilevel graphs and big data [Rathore et al. 2018] try to identify reviewers and assess the impact factor and ranking of journals and researchers. The authors propose graphs representing the relationships "authors and co-authors", "author and citations", "author and research area", "author and journal", "author and organization", journals, and citations, among others. The proposed architecture involves collecting data from bibliometric information sources, such as IEEE, ACM, Spring, among others. The construction processes and the processing of the graphs are carried out by the Hadoop and Spark tools. After processing, selecting reviewers is proposed, considering conflicts of interest, a ranking of researchers, and ranking of journals. The authors conclude that the proposed architecture is quite efficient and meets the needs of the academic community.

In their work, [Ghani et al. 2019] propose an empirical investigation comparing the h-index with M-quotient indexes, hl-index, hm-index , hc-index, hw-index, "fractional counting in journals" and "fractional counting of citations", which are extensions of the same index. The authors use the research area of mathematics, understanding that the area is the basis for several other sciences. In addition to comparisons between indexes, the authors also compare these indexes with data from awards in the area. The authors conclude that the indexes have their own characteristics, since, compared to the ranking of awarded researchers, they present different behaviors.

Considering researchers in the field of evolutionary computing, [Bidgoli et al. 2019] propose an index for ranking researchers based on the Pareto principle and the multi-criteria decision-making method. The authors justify that the h-index harms younger researchers, regardless of the number of academic research years. The authors state that the proposed methodology allows a fairer comparison between researchers with different years of academic career, considering that the achievement of particular value for the h-index in a shorter time is an essential factor.

3. EXPERIMENTS AND RESULTS

In this section, we describe how the proposed solution works and the experiments conducted to evaluate it. Our proposal combines clustering with regression models. This combination aims to obtain data for ranking researchers considering the human in the loop. In other words, as the human being presents examples to the solution, the solution returns the result allowing the human to decide which other examples may be necessary. This approach tries to contribute to reducing the cost of obtaining data for training.

3.1 Proposed Solution

Figure 1 presents an overview of the proposed solution. We apply clustering concepts to obtain data that will serve as training for the regression models. This approach helps to get the training data from the process, putting the human in the loop.

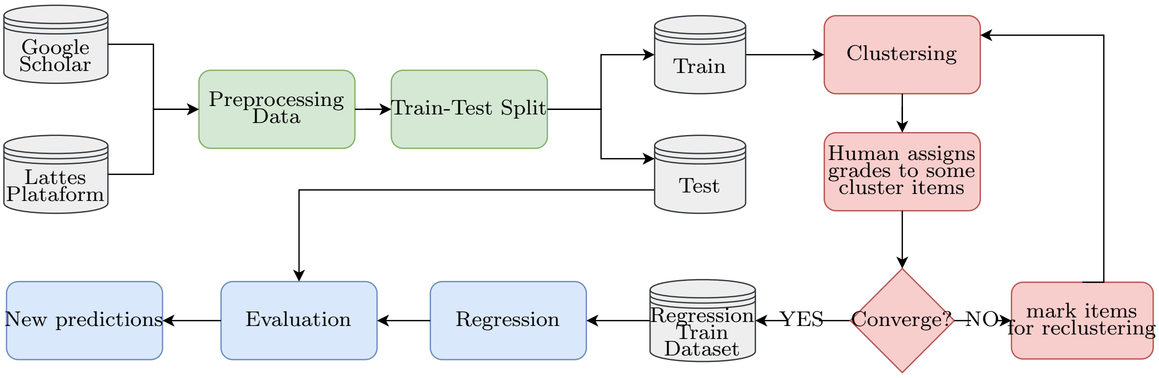

Fig. 1. We obtained data from Google Scholar and Lattes platforms. Then, we preprocessed and split the dataset into training and test sub-datasets, considering 1/3 for testing. The training data is then used in the clustering process. Each obtained cluster is evaluated, and two items in each one of them and annotated by a human. If the annotation values are within the expected (threshold), the data is added for training. If not, the data is again clustered until convergence or the maximum number of iterations. The data obtained are used to train a regression model, which allows obtaining the weights of each feature and new predictions.

First, we obtained curriculum data as well as bibliometric data from each researcher. We then perform data preprocessing. We collected data from a Brazilian institution regarding a researcher selection process in a research grant award process to evaluate the proposed solution. The data used for selection refer to data from the curriculum of 194 researchers.

From that point on, we were concerned with creating a clustering process capable of obtaining data for the regression. In the clustering step, the characteristics chosen by the user are combined to obtain clusters. Then, the first and last elements of each cluster are annotated by the user. If the difference between the given notes is more significant than a previously chosen threshold, the cluster elements are again clustered until the threshold or the number of iterations is reached, whichever comes first.

The data used are from published papers, published books, published chapters, among other information. All characteristics to be used are calculated in this process, and researchers are manually ranked. In our work, for dimensionality reduction, we use Principal Component Analysis (PCA), selecting the two main characteristics.

Our solution proposal allows the user to select the characteristics manually as well. Thus, particularities of the process are considered when creating the model, including the human in the process from the beginning. For the experiments presented in Section 3.2, we selected the characteristics listed in the Table I, which provides descriptive statistics.

The difference regarding the bibliometric indexes presented above is that recent publications are counted, which do not yet count for the indexes, and publications older than 3 or 5 years are not used, depending on the conditions in which the researcher is. Thus, the ranking is done considering the specifics of the organization. On the other hand, known bibliometric indexes are not used. The use of these indexes can help to reduce the human effort in the ranking process. In our experiments, we used published book chapters, papers, scientific initiation supervisions and h-index5y.

3.2 Results

In this section, we present the results found in the carried-out experiments. For each hyperparameter, we performed 30 repetitions. Figure 2 shows an example of cluster formation, using K-Means and considering k=7. The main idea is to show that clustering with the PCA is efficient in separating researchers into groups.

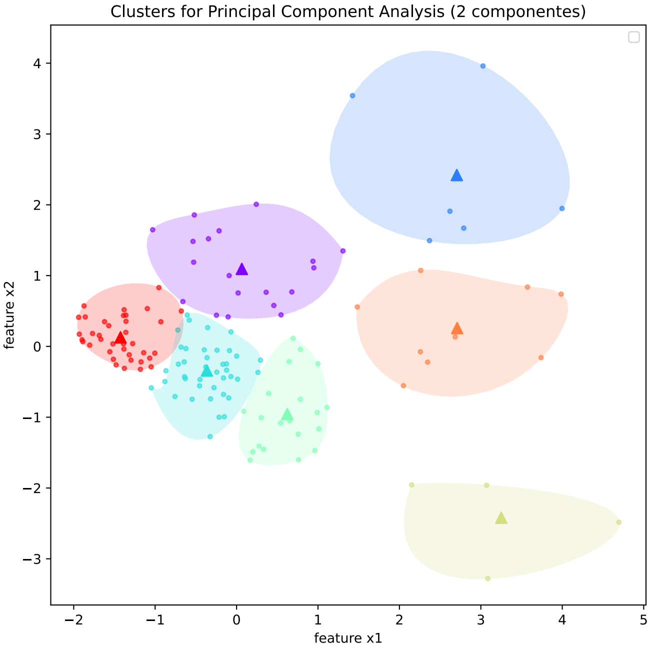

Fig. 2. An example with clusters interpolation, using K-Means with k=7.

Then, knowing that clustering can contribute to the human insertion process in the loop, we study the ideal number of clusters for the dataset. In this case, we use Within Cluster Sum of Squares (WCSS) for evaluation, varying the k value from 3 to 35. Figure 3 shows (left) that the ideal k value for the dataset (or the Elbow point) would be in all of 11. We performed tests for clustering with k=11 without reclustering. The results are showed inf Figure 3 (right) and in Table I.

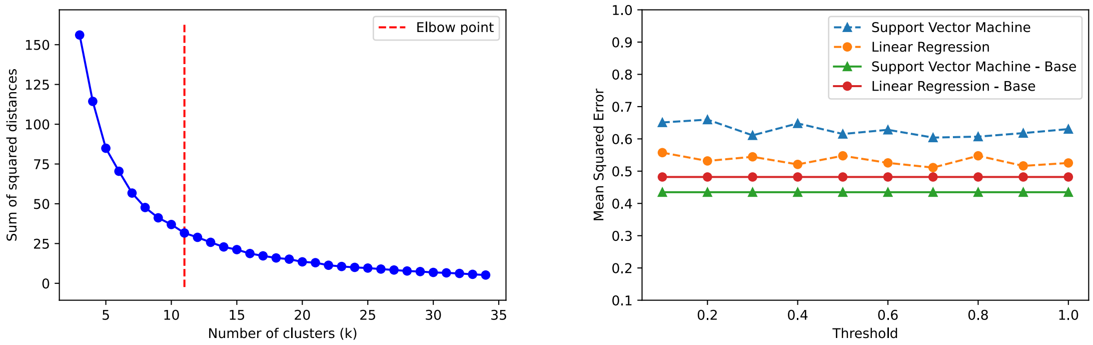

Fig. 3. Graphs for the WCSS results (left) to obtain the Elbow point, varying k from 5 to 35. The figure also shows the variation for the MSE considering different threshold values. It is possible to observe that the two tested models present specific stability when the threshold values are varied. The average MSE for the models was 0.8550 for the SVM and 0.6483 for the Linear Regression.

Figure 4 shows comparison for Mean Squared Error varying the number of clusters (left). It is possible to verify that the Linear Regression behaved better considering the cluster value between 10 and 12. However, the Support Vector Machine presented less variation for the hyperparameter.

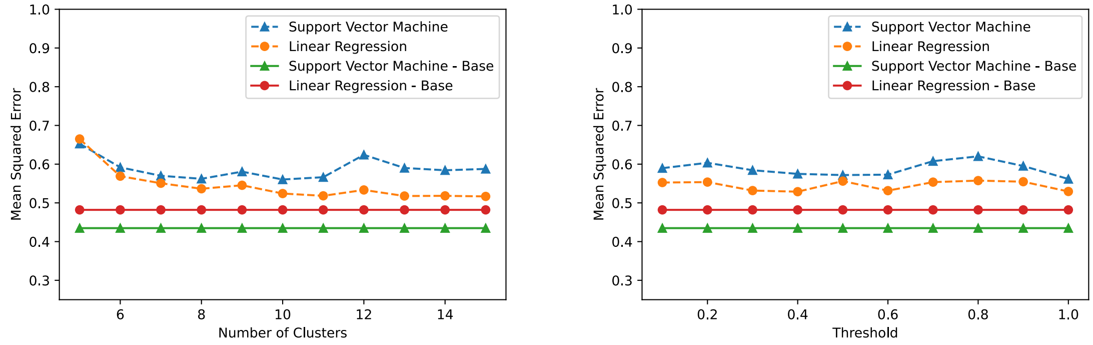

Fig. 4. Graph for comparison between models, considering MSE. It is possible to observe (left) that, with the increase in the number of clusters, the SVM model has a worse performance when compared to the LR. The variation is not the same for threshold comparison. However, it is possible to observe that, in both cases, LR presents better results when clustering occurs. The values for the base-model tests are also presented in the two images, where 2/3 of the dataset is used for training and 1/3 of the dataset for testing.

Still in Figure 4 shows the comparison for Mean Squared Error by varying the threshold values (right). Again, it is possible to observe that the SVM remains more stable while the linear model presents significant variations, even with higher thresholds. We also presented the comparison between the models in Table I. When comparing the rankings generated using the nDCG metric, our solution is promising, and in the four evaluations (nDCG@5 to nDCG@20), one of our models was among the two best solutions.The values in bold refer to the solution, which obtained lower results than the models used, which was expected, but better results concerning clustering and annotation in a single time, considering the number of clusters for the Elbow point. Figure 5 presents the average of data needed for each k value.

Table I. Comparison of regression models. The SVR and LR models represent the tests carried out considering the entire dataset and served as a basis for comparing the other models. For the SVR-11 and LR-11 models, the results are presented without reclustering. The SVR-11 and LR-11 models represent the best models obtained with the reclustering solution, with threshold = 0.4. It is possible to verify that the two models present very similar results.

Modelo	Clustering	Reclustering	MSE	nDCG@5	nDCG@10	nDCG@15	nDCG@20
SVR–11	yes	yes	**0.5185**	0.7677	0.7899	0.7559	0.7203
LR–11	yes	yes	**0.5182**	0.8136	0.7941	0.7764	0.7450
SVR	no	no	0.4193	0.8161	**0.8148**	**0.8148**	0.7621
LR	no	no	0.4458	0.8181	0.7861	0.7861	**0.7738**
SVR-11	yes	no	0.8550	**0.8312**	**0.8048**	0.7868	**0.7707**
LR–11	yes	no	0.6483	**0.8233**	0.7846	**0.7937**	0.7449

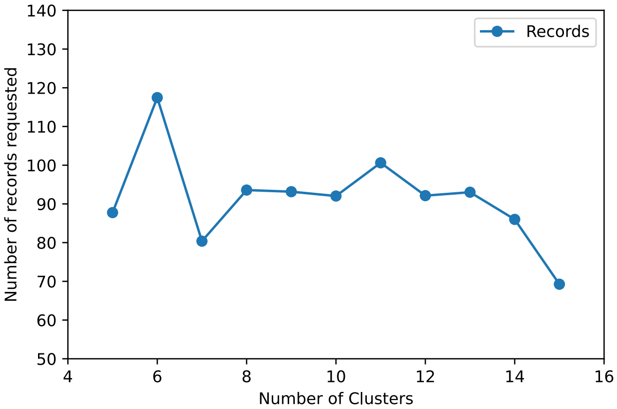

Fig. 5. The average number of annotations (records), varying the number of clusters. On average, 91 iterations were required for the training base. For k=11, an average of 100 grades were required.

4. CONCLUSION

The main objective of our work was to evaluate how the combination of clustering and regression models can support the ranking process of researchers, in particular, in reducing the human effort for annotation. We use K-Means combined with SVM and LR. With the two techniques for regression, it was possible to verify that our proposal can support the researchers' recommendation process in a scenario in which data cannot be easily obtained and including the human in the loop. Considering the MSE values for the models trained using entire dataset (SVM = 0.4193 and LR = 0.4458) compared to our model (SVM = 0.5185 and LR = 0.5182), it is possible to verify a reduction in the MSE and a better stability between the models. Furthermore, when analyzing the ranking of the top 5 researchers (nDCG@5), we also verified that our solution with only one clustering step has the best performance. In future work, we intend to study the strategies to find the most representative items of each cluster for the regression, reducing the amount of data needed for training. In addition, we intend to study how Natural Language Processing can contribute to ranking and recommending researchers.

REFERENCES

Alhoori, H. and Furuta, R. Recommendation of scholarly venues based on dynamic user interests. *Journal of Informetrics* 11 (2): 553563, 2017.

Beel, J., Gipp, B., Langer, S., and Breitinger, C. paper recommender systems: A literature survey. *International Journal on Digital Libraries* 17 (4): 305338, 2016. Publisher: Springer.

Bidgoli, A., Rahnamayan, S., Mahdavi, S., and Deb, K. A Novel ParetoVIKOR Index for Ranking Scientists' Publication Impacts: A Case Study on Evolutionary Computation Researchers. pp. 24582465, 2019.

Chaiwanarom, P. and Lursinsap, C. Collaborator recommendation in interdisciplinary computer science using degrees of collaborative forces, temporal evolution of research interest, and comparative seniority status. *Knowledge-Based Systems* vol. 75, pp. 161172, 2015. Publisher: Elsevier.

Davison, E. and Price, J. How do we rate? An evaluation of online student evaluations. *Assessment & Evaluation in Higher Education* 34 (1): 5165, 2009. Publisher: Taylor & Francis.

Dorogovtsev, S. N. and Mendes, J. F. Ranking scientists. *Nature Physics* 11 (11): 882883, 2015. Publisher: Nature Publishing Group.

Gao, B. and Kumar, G. CoRank: Simultaneously Ranking Publication Venues and Researchers. pp. 60556057, 2019.

Ghani, R., Qayyum, F., Afzal, M., and Maurer, H. Comprehensive evaluation of hindex and its extensions in the domain of mathematics. *Scientometrics* 118 (3): 809822, 2019.

Huang, W., Wu, Z., Liang, C., Mitra, P., and Giles, C. A neural probabilistic model for context based citation recommendation. Vol. 29, 2015. Issue: 1.

Jin, B., Liang, L., Rousseau, R., and Egghe, L. The Rand ARindices: Complementing the hindex. *Chinese science bulletin* 52 (6): 855863, 2007. Publisher: Springer.

Khanam, Z. and Alkhaldi, S. An Intelligent Recommendation Engine for Selecting the University for Graduate Courses in KSA: SARS Student Admission Recommender System. *Lecture Notes in Networks and Systems* vol. 98, pp. 711722, 2020.

Liang, D., Charlin, L., McInerney, J., and Blei, D. M. Modeling user exposure in recommendation. pp. 951961, 2016.

Lima, H., Silva, T. H., Moro, M. M., Santos, R. L., Meira Jr, W., and Laender, A. H. Aggregating productivity indices for ranking researchers across multiple areas. pp. 97106, 2013.

Maqsood, S., Islam, M., Afzal, M., and Masood, N. A comprehensive author ranking evaluation of network and bibliographic indices. *Malaysian Journal of Library and Information Science* 25 (1): 3145, 2020.

Oliveira, E., Gomes Basoni, H., Rodrigues Saúde, M., and Marques Ciarelli, P. Combining Clustering and Classification Approaches for Reducing the Effort of Automatic Tweets Classification:. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*. SCITEPRESS Science and and Technology Publications, Rome, Italy, pp. 465472, 2014.

Pradhan, T. and Pal, S. CNAVER: A Content and Networkbased Academic VEnue Recommender system. *KnowledgeBased Systems* vol. 189, 2020.

Rathore, M. M. U., Gul, M. J. J., Paul, A., Khan, A. A., Ahmad, R. W., Rodrigues, J., and Bakiras, S. Multilevel graphbased decision making in big scholarly data: An approach to identify expert reviewer, finding quality impact factor, ranking journals and researchers. *IEEE Transactions on Emerging Topics in Computing*, 2018.

Rost, K. and Frey, B. S. Quantitative and qualitative rankings of scholars. *Schmalenbach Business Review* 63 (1): 6391, 2011. Publisher: Springer.

Sabour, S. H. Index, an ugly truth. *Shiraz E Medical Journal* 20 (5), 2019.

Sebastian, Y., Siew, E., and Orimaye, S. O. Learning the heterogeneous bibliographic information network for literaturebased discovery. *KnowledgeBased Systems* vol. 115, pp. 6679, 2017. Publisher: Elsevier.

Shah, K., Salunke, A., Dongare, S., and Antala, K. Recommender systems: An overview of different approaches to recommendations. IEEE, pp. 14, 2017.

Sharma, M. and Mann, S. A survey of recommender systems: approaches and limitations. *International Journal of Innovations in Engineering and Technology* 2 (2): 814, 2013. Publisher: Citeseer.

Son, J. and Kim, S. B. Academic paper recommender system using multilevel simultaneous citation networks. *Decision Support Systems* vol. 105, pp. 2433, 2018. Publisher: Elsevier.

Ströele, V., Campos, F., David, J. M. N., Braga, R., Abdalla, A., Lancellotta, P. I., Zimbrão, G., and Souza, J. Data abstraction and centrality measures to scientific social network analysis. In *2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, pp. 281286, 2017.

Tang, J., Wu, S., Sun, J., and Su, H. Crossdomain collaboration recommendation. pp. 12851293, 2012.

Tatiya, R. V. and Vaidya, A. S. A survey of recommendation algorithms. *IOSR J. Comput. Eng* 16 (6): 1619, 2014.

Wang, G., He, X., and Ishuga, C. I. HARSI: A novel hybrid article recommendation approach integrating with social information in scientific social network. *KnowledgeBased Systems* vol. 148, pp. 8599, 2018. Publisher: Elsevier.

Yang, Z., Yin, D., and Davison, B. D. Recommendation in academia: A joint multirelational model. IEEE, pp. 566571, 2014.

Yu, S., Liu, J., Yang, Z., Chen, Z., Jiang, H., Tolba, A., and Xia, F. PAVE: Personalized Academic Venue recommendation Exploiting copublication networks. *Journal of Network and Computer Applications* vol. 104, pp. 3847, 2018.

Zriaa, R. and Amali, S. A Comparative Study Between KNearest Neighbors and KMeans Clustering Techniques of Collaborative Filtering in eLearning Environment. *Lecture Notes in Networks and Systems* vol. 183, pp. 268282, 2021.