Evaluating Educational Recommendation Systems: A systematic mapping

Yelco Antonio Marante Cañizales[1,2], **Vinicius Alberto Alves da Silva**[1], **Jorão Gomes Jr.**[1,2],**Marluce Aparecida Vitor**[1], **André Ferreira Martins**[1,2], **Jairo Francisco de Souza**[1,2]

[1] LApIC Research Group

[2]Graduate Program in Computer Science
Federal University of Juiz de Fora (UFJF)
36.360-900 – Juiz de Fora – MG – Brazil

{yelcomarante,viniciusalves,joraojunior}@ice.ufjf.br

{marlucevitor,andre.martins,jairo.souza}@ice.ufjf.br

Abstract. *Recommendation systems (RS) have been used in many scenarios, from entertainment to health. Inside the RS area, Educational Recommendation Systems (ERS) are becoming popular, been used for different types of recommendations such as recommending materials, exercises, and learning paths. As ERS works in a different scenario of classics RS, ERS requires specific evaluation metrics. However, the task of evaluating ERS is difficult since the educational field has its features to be analyzed. To help other researchers in this field, this work presents a systematic mapping on methods used for evaluating ERS. This study analyzed 91 papers of the last five years and provide an overview of the main methodologies, subject, metrics, and trends in the evaluation of ERS.*
Keywords: Educational Recommendation System, System assessment, System evaluation

1. Introduction

Recommendation systems (RS) have been used in many scenarios, from entertainment [Andjelkovic et al. 2019] to health [Gyrard and Sheth 2020]. Such systems typically provide to the user a recommended list with items they might prefer or forecast how much they might prefer each item [Shani and Gunawardana 2011]. Therefore, RS aims to generate meaningful recommendations for a collection of users for items or products that are of interest to them [Melville and Sindhwani 2010].

Inside the RS area, Educational Recommendation Systems (ERS) are becoming popular [Dwivedi and Roshni 2017]. The RS used in the educational field is known as Enhanced Technological Learning (TEL). According to [Erdt et al. 2015], this term is used to describe a technological application in teaching and learning. The TEL approach has been used for different types of recommendations, such as recommending materials [De Medio et al. 2020, Machado et al. 2019], exercises [Lv et al. 2018], learning paths [Machado et al. 2020, Nabizadeh et al. 2020], partners for study groups [Khosravifar et al. 2018], etc. For these works, classic techniques used in recommendation systems are used, such as content recommendation, collaborative filtering, and others.

DOI: 10.5753/cbie.sbie.2020.912

ERS has different purposes compared to RS used in other scenarios, such as streaming videos, music, or e-commerce. While other systems intend to keep active user participation within the system to increase the number of views of the recommended items for purely commercial purposes, ERS should have pedagogical objectives, such as improving learning, personalizing teaching practices, reducing the information overload for teachers and students, or increasing student engagement to reduce anxiety, drop out rates, among other factors [Santos et al. 2014]. These characteristics, however, make the evaluation of ERS approaches require specific evaluation metrics once that educational field has its own features to be analyzed.

The task of evaluating ERS is difficult since the educational scenario: (i) does not have enough open datasets to do it; (ii) have many ways to evaluate the RS algorithm once exists a lot of metrics that can be used to this propose, such as accuracy and f-measure, making exhaustive to know the best one; (iii) have to measure if the recommendations were good to the users and if represent their needs; (iv) once all the above were solved, it is not easy to perform real tests with students or teachers and the results achieved with synthetic bases may not represent the needs in a real scenario [Peralta et al. 2018]. Therefore, the best evaluation of ERS will be a combination of the algorithms results, user behavior, dataset features, and a final application on a real case test to get user feedback/experience.

To help other researchers in this field, this work presents a systematic mapping of methods used for evaluating ERS. This study analyzed 91 papers of the last five years and provide an overview of the main methodologies, subject, metrics, and trends in the evaluation of ERS. The article is structured as follows: section 2 discusses the related works, comparing this work with other reviews. Section 3 describes the systematic mapping protocol. Section 4 presents a discussion of the results. Section 5 presents the concluding remarks and future works.

2. Related Work

Educational Recommendation Systems are a growing trend in technology today. Over the years, many articles have been published, constructing a deep and dense field of study. Therefore, attempts to review and map the literature have already been made, but these studies vary considerably concerning to the discussions on the evaluation of ERS approaches. In [Tarus et al. 2018], for instance, is presented a review of ontology-based ERS and is discussed the different techniques, knowledge representation, ontology type, and ontology representation that are used in literature. The authors mention that ontology can help improve the recommendation although it is difficult to evaluate ontology-based ERS due to a lack of a standard e-learning database. In [Truong 2016], is reviewed different strategies to integrate learning styles in ERS, and it is noticed that not all articles contain an evaluation section. Among those who did evaluate their approaches, some have used statistical evaluation tests based on pre-and post-test performance, time spent on the task, level of completeness, engagement, and cognitive loads level.

In [Yu et al. 2018, Cui et al. 2018], different recommendation models used in ERS are presented, comparing their advantages and disadvantages. Some common problems are mentioned, such as the difficulty of offering recommendations to new users or low levels of personalization. In [Yu et al. 2018], the authors stated that the presented works use conventional recommendation techniques, without addressing the specificities of the

area of education. However, the evaluation process used in those papers is not considered in the review. In [Cui et al. 2018], the authors discuss the difficulty of effectively evaluating recommendation systems, but they are limited to the conventional metrics, such as accuracy and diversity, and did not consider specific characteristics of the educational area. Besides, several questions about ERS applications were analyzed in [Rivera et al. 2018], such as educational areas covered, approaches used to generate recommendations, and which platform is used. The evaluation method was also investigated, describing whether they used surveys, case studies, or experiments. The authors concluded that most of the papers did not present any kind of evaluation. Of the works that presented some evaluation, the use of experiment technique was the most common kind. Also, they highlight the lack of data considering different learner profiles and personal characteristics.

Finally, the evaluation of ERS approaches was the main topic in the following works: [Erdt et al. 2015, Drachsler et al. 2015]. In [Erdt et al. 2015], the authors present a survey on ERS and describe the most common evaluation methods used. This study defines three methodologies that are used for evaluating ERS approaches: Offline Experiment (or dataset driven evaluation), User Study (or user experiment), and Real Life Testing which is also called Online Evaluation. Several effects measured by the evaluations were considered in order to classify the reviewed works. The authors conclude the need for evolving evaluation methods, the need for more explicit discussion on how the evaluation is done, and the difficulty of comparing work in the area due to the fact that there is no standardization of evaluation methods. In [Drachsler et al. 2015], the need to standardize the evaluation of ERS is also highlighted, besides the review, a framework for ERS evaluation is also proposed in an attempt to satisfy that demand.

In this paper, we present a systematic mapping of the evaluation methods used in ERS literature in the last five years. This mapping is an update of the last work [Erdt et al. 2015] in this topic, which collected papers until 2014. We show that there has been a significant increase in new research since then, and we analyze how the ERS evaluation methods have changed over these years.

3. Mapping Protocol

This mapping follows from the procedures proposed by [Erdt et al. 2015] and the following research questions were defined: **(Q1)** What is the most common type of evaluation methodology used?; **(Q2)** Who is the subject of evaluation?; **(Q3)** What are the effects measured by evaluation of ERS?

This review sought to answer these questions based on evidence and the construction of a well-designed search string. The scope of the review is shown in Table 1.

Table 1. Scope of the current research defined using PICO.

Population (P)	Articles based on education
Intervention (I)	Recommendation Systems
Comparison (C)	-
Outcome (O)	Evaluation

Finally, we defined the search string as follows and it was used in the Scopus[1] repository:

> *("education" OR "web-based learning" OR "learning system" OR "educational" OR "learning environment" OR "course" OR "mooc" OR "Massive Open Online Course" OR "intelligent tutoring system") AND ("recommender system" OR "recommendation system" OR "recommendation approach" OR "recommendation tool" OR "recommendation framework") AND ("experimentation" OR "evaluating" OR "evaluation" OR "accuracy" OR "metrics" OR "experiment" OR "results")*

The search string is a Boolean expression concatenating with an AND statement the three main concepts from Table 1. Synonyms and alternative spellings were added with an OR statement. The result set was limited to articles published after 2014.

To include only studies relevant to the scope defined, we list the exclusion criteria: **(EC1)** Articles that do not present an ERS algorithm or tool for TEL (Technology Enhanced Learning); **(EC2)** Articles that do not contain an evaluation section; **(EC3)** Articles that do not present a learning material recommendation approach, such as course recommendation or grade prediction; **(EC4)** Article not written in English; **(EC5)** Grey literature; **(EC6)** Articles that full text is not available.

The survey conducted on October 31, 2019, returned 1116 articles. Based on this result, a first reading of the titles and abstracts was carried out. After this refinement, get 235 articles that were read in detail. In the end, 91 articles were included in the mapping. The number of articles excluded in each criterion is **EC1**: 52; **EC2**: 26; **EC3**: 61; **EC4**: 2; **EC5**: 2; **EC6**:1; to guarantee the quality of the mapping, the excluded articles were peer-reviewed. The full accept articles list can be found here[2].

4. Results

The 91 selected articles were categorized according to the classification criteria explained in Section 3. The results of the survey are presented and discussed in the following sections.

4.1. Evaluation Methodologies and Subject of Evaluation

Evaluation methodologies applied for the evaluation of the recommender systems can be classified into three categories:

1. *Offline Experiment*: use datasets consisting of user interactions to evaluate recommender systems. Two types of datasets are used: Historical datasets consisting of real interactions of actual users in a real system over some time. The second type are synthetically constructed datasets normally used to test how recommender algorithms perform in constructed scenarios and under specified conditions [Erdt et al. 2015].
2. *User Study*: used to find out how a recommender system influences the user's experience, perception, and interactions with a system [Knijnenburg 2012].

[1] `https://www.scopus.com`
[2] `https://github.com/lapic-ufjf/ERS-systematic-mapping`

3. *Real Life Testing*: real users using the system under normal conditions for a long period.

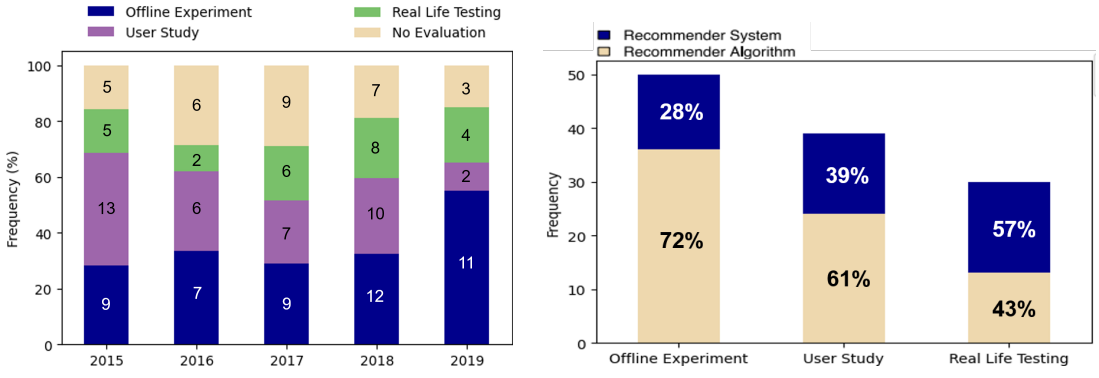
In the evaluations of recommendation systems, two types of evaluation subjects are used:

1. *Recommender Algorithm*: the evaluation is based on how well algorithms predict or rank recommendations [Erdt et al. 2015].
2. *Recommender System*: the focus is on the entire recommendation system, including aspects such as user interface or system usage [Erdt et al. 2015].

Figure 1a presents an overview of the 91 articles classified according to the evaluation methodology: offline experiments (48), user studies (38), real-life testing (25), and no evaluation (30) distributed over the last five years. The map shows that the ERS evaluation has become increasingly important over the years: 2015 (20), 2016 (13), 2017 (19), 2018 (25), and 2019 (14), having in 2018 the maximum number of publications with evaluation. The distribution of the evaluation methods has remained relatively stable over the last five years with 34.04% offline experiments, 26.95% user studies, 17.73% real life testing, and 21.28% no evaluation. Offline experiments predominate over the years since many works are published with results of prototypes evaluations using simulations on historic and synthetic data. That is, few studies evolve to evaluate properly the effects of recommendations on user learning and satisfaction.

On the other hand, some researchers have improved their evaluation by combining other methods with offline experiments, for instance, offline experiment and user study, offline experiment, and real life testing, user study and real life testing. 35 articles used only offline experiments and 13 articles used offline experiments together with others.

Distribution of the subject of evaluation across evaluation methodologies is shown in Figure 1b. Offline experiments are used mainly to evaluate recommendation algorithms (72%) because the algorithms can be evaluated with a low effort [Erdt et al. 2015]. There is a tendency towards an increase in the number of studies that use User Study Methodology to evaluate the recommender algorithm (61%). The recommender system is normally the focus of the evaluation (57%) in real life testing, because there is a better comprehension of how the system should properly work.



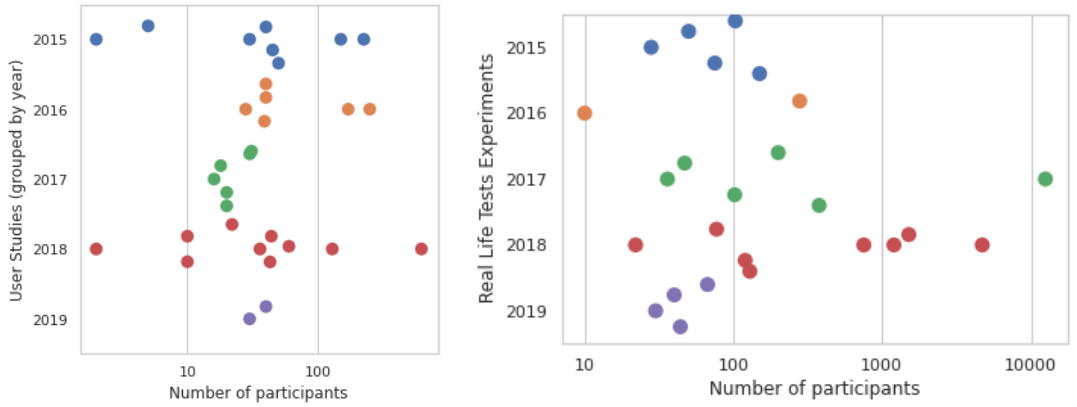
(a) Evaluation Methodologies (b) Subject of Evaluation

Figure 1. Evaluation Methodologies and Subject of Evaluation

In [Erdt et al. 2015], the authors state that the average of the number of participants in User Studies is 53, but the median is only 24, and few studies have a large number of

participants. In our review, the average is 71 and the median is 37, then we can see that researchers are carrying out experiments with a larger number of participants on average. User Study methodology uses techniques to obtain information about user behavior and feedback. The most common is a questionnaire where users answer about their perceived value of the system, 21 works present questionnaires, 11 works used pre-tests and post-tests to discover the influence of the ERS in the learning. Other techniques involve Experts Opinions, User Observation, which appeared four times, and Interviews, which were used in only two publications.

The number of participants is one of the main differences between User Studies and Real Life Tests. Real Life Tests have a median of 102 participants and a mean of 902 participants across the works. Figure 2 shows the number of participants distribution in User Studies and Real Life Tests methodologies over the years (using a log scale). An exclusive characteristic of Real Life Tests is the long test period which ranges from 10 to 1640 days, with an average of 265 and a median of 278 days. Although ERS is a prolific area of study with new approaches every year, it is worth noting how many approaches are evaluated using a small group of participants. Sample calculation, significance analysis, and other basic statistical tools are rarely used in most works. Moreover, few authors make the dataset available for further research. Thus, it is challenging to compare and identify novel valuable contributions in ERS literature.



(a) Participants in User Studies

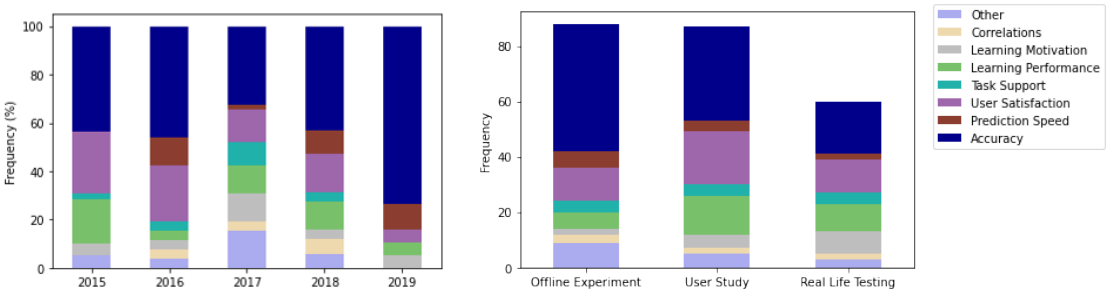
(b) Participants in Real Life Tests

Figure 2. Number of participants in user studies and real life tests per year

4.2. Effects Measured

The unique requirements of ERS demand authors to measure specific effects on the recommendation process evaluation. These effects were assembled in [Erdt et al. 2015] and used in the classification of our survey. Each methodology described in section 4.1 tends to measure an effect. The occurrence of the effects in each methodology is presented in Figure 3b. The [Erdt et al. 2015] review presented an increase in the variety of effects measured throughout the years. In Figure 3a is shown that the measured effects are stable over the years, indicating a point of maturity in the literature on evaluation of ERS.

To estimate the efficiency of the recommender system, *Accuracy* is the most assessed effect. It includes metrics to measure the relevance and exactness of the recommender system, such as precision, error rate, recall, Top-N, Mean Absolute Error (MAE),



(a) By Years (b) By Evaluation Methodologies

Figure 3. Effects Measured by Years and Evaluation Methodologies

and Root Mean Squared Error (RMSE). Researchers are always concerned with the *Accuracy* of their work, which makes it widely used regardless of the methodology. Another effect related is the *Efficiency*, which is the time spent by a recommender system to produce the recommendations or the time taken for the user to notice the recommendation, defined as *Prediction Speed*. Response time and execution time are metrics used to measure this effect, which is easier to measure in Offline Experiments.

The objective of any recommender system is to fulfill the user's needs, then User-Centric effects are often presented in the literature. The user's perceived value of the recommendations and user's comfort within the system is measured as *User Satisfaction*, which is the second most measured effect in our review. Authors usually use questionnaires with Likert Scales to obtain user feedback. Few authors have used advanced methods as ResQue [Pu et al. 2011]. Some works present statistical evaluation with user's answers applying Anova and Tukey tests. The helpfulness provided by the recommender system in intention to support the user's task is labeled as *Task Support*. This influence can be measured using A/B tests, that is, comparing groups of learners who used the system with those who did not, and, then, analyzing which group performs a task better. A minority of works presented an analysis of this effect, perhaps because this kind of evaluation is more costly.

Measuring how the recommendation process affect user learning is a requirement of a reasonable ERS evaluation. In current researches, Effects on Learning metrics are more regularly presented than in the past. *Learning Motivation* estimate how much the Recommender System influence the Learner engagement. Real Life Tests evaluations cover more often this effect, because of the difficulty of measuring motivation in the short periods of user studies. The perceived success of learning using the recommender system and improvements in the learning is covered by *Learning Performance* where it is mainly evaluated by the improvement in the performance of learners after they received recommendations. The main techniques are applying exams, questionnaires, challenges, pre-tests, and post-tests. *Correlations* are an effect that quantifies associations between the learner activities and other measured effects on learning, such as co-occurrences between different activities. An example is the use of Pearson correlation to find co-occurrences of an item in the user's logs or to find a correlation between actions or recommendations to the performance of the user.

The effects that do not belong to any of the specific topics above are classified as

Others. This category consists of metrics such as serendipity, network bandwidth usage, emotion, novelty, variety, diversity, learning styles preference [de Almeida et al. 2019], and the variety of learning paths. Offline Experiments have more effects of this category by the use of simulation models, which are more flexible and less time-consuming than other methods.

Considering that Offline Experiments are fast, easy to conduct, and the most popular methodology applied in the evaluation of the algorithms, it is remarkable that all effects were measured in this methodology, as seen in Figure 3b. Hence *User-centered* and *Effects on Learning* metrics can be measured in the simulation of the recommendation process with synthetic or historical datasets. This result presents a change in ERS evaluation in the last years compared to the study of [Erdt et al. 2015], where the authors showed that Offline Experiments were only used to measure accuracy, prediction speed, learning performance, and correlations. Offline Experiments are now more concerned with presenting results with measures that consider the learned, aligned with the unique requirements of Educational Recommendation Systems evaluations.

Despite the advantages of using simulation models in Offline Experiments, they are simplistic models and some effects may not be noticed or overlooked. Therefore, there is still a need for Offline Experiment projects to evolve to the point that it is possible to accomplish evaluations with users whether in controlled environments or real tests. This is the way to find out if the results of the simulations are in by the learner's perception and discover new effects.

5. Concluding remarks and Future works

To help other researchers in this field, this work presented a systematic mapping of the evaluation methods of ERS. 91 ERS papers of the last five years were collected and an overview of the main methodologies, subject, metrics, and trends in the evaluation of the ERS was provided. The mapping showed that the ERS evaluation has become increasingly important over the last years, having in 2018 almost 25 publications with evaluation sections. Few papers evolve to evaluate properly the effects of recommendations on user learning and satisfaction since that *Offline Experiments* are predominant over the years. ERS researchers still need to perform more real tests with students or teachers to a better comprehension of the value of their approaches, since the results achieved with synthetic datasets may not represent the user needs in a real scenario. On the other hand, we noticed an improvement in the *Offline Experiments* in some papers, where authors have combined *Offline Experiments* with other methods to perform better analyses of their contributions.

It was noticed an increase in the number of participants in studies using *User Study* and *Real Life Tests* methodologies, showing that there is an increase in concern about the effects on learning and user-centered effects. Due to unique ERS requirements, authors should give preference to effects on learning metrics, such as *learning performance*, *learning motivations*, and learner's log *correlations*. Besides that, some popular topics in intelligent tutoring systems, such as *learning styles preference* or *learning paths*, are almost ignored in the ERS literature. Also, some important user-centered metrics as novelty, diversity, privacy, and serendipity still not being frequently measured. This shows that it is still necessary for authors to give a preference for measuring learning effects, evaluating if their systems are really increasing students learning.

This study has some limitations. Due to space limitations, it was not possible to have a detailed discussion on the ERS approaches found in this mapping. Although the characteristics of each approach influence the kind of evaluation carried out, this map tried to discuss general aspects of ERS evaluation and to present an overview of metrics and methodologies used in recent years. In addition, it is possible that relevant articles have not been indexed by Scopus. Although other repositories can be used, Scopus is one of the largest abstract and citation databases of peer-reviewed literature and has high coverage of the Computer Science literature.

Acknowledgments

We thank RNP (Brazilian National Education and Research Network) and CAPES (Coordination for the Improvement of Higher Education Personnel) for the financing of this work.

References

Andjelkovic, I., Parra, D., and O'Donovan, J. (2019). Moodplay: Interactive music recommendation based on artists' mood similarity. *International Journal of Human-Computer Studies*, 121:142–159.

Cui, L.-Z., Guo, F.-L., and Liang, Y.-j. (2018). Research overview of educational recommender systems. In *Proceedings of the 2nd International Conference on Computer Science and Application Engineering*, pages 1–7.

de Almeida, M. A., Souza, J., and Barrére, E. (2019). Learning style identification and usage in academia: A systematic mapping. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, page 1341.

De Medio, C., Limongelli, C., Sciarrone, F., and Temperini, M. (2020). Moodlerec: A recommendation system for creating courses using the moodle e-learning platform. *Computers in Human Behavior*, 104:106168.

Drachsler, H., Verbert, K., Santos, O. C., and Manouselis, N. (2015). *Panorama of Recommender Systems to Support Learning*, pages 421–451. Springer US, Boston, MA.

Dwivedi, S. and Roshni, V. K. (2017). Recommender system for big data in education. In *2017 5th National Conference on E-Learning & E-Learning Technologies (ELELTECH)*, pages 1–4. IEEE.

Erdt, M., Fernandez, A., and Rensing, C. (2015). Evaluating recommender systems for technology enhanced learning: a quantitative survey. *IEEE Transactions on Learning Technologies*, 8(4):326–344.

Gyrard, A. and Sheth, A. (2020). Iamhappy: Towards an iot knowledge-based cross-domain well-being recommendation system for everyday happiness. *Smart Health*, 15:100083.

Khosravifar, B., Cuevas, C., and Bentahar, J. (2018). Dynamic peer-to-peer amplifier system used in agent-based intelligent tutoring system. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, pages 352–358. The Steering Committee of The World Congress in Computer Science, Computer ....

Knijnenburg, B. P. (2012). Conducting user experiments in recommender systems. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 3–4.

Lv, P., Wang, X., Xu, J., and Wang, J. (2018). Utilizing knowledge graph and student testing behavior data for personalized exercise recommendation. In *Proceedings of ACM Turing Celebration Conference-China*, pages 53–59.

Machado, M. d. O. C., Barrére, E., and Souza, J. (2019). Solving the adaptive curriculum sequencing problem with prey-predator algorithm. *International Journal of Distance Education Technologies (IJDET)*, 17(4):71–93.

Machado, M. d. O. C., Bravo, N. F. S., Martins, A. F., Bernardino, H. S., Barrere, E., and de Souza, J. F. (2020). Metaheuristic-based adaptive curriculum sequencing approaches: a systematic review and mapping of the literature. *Artificial Intelligence Review*, pages 1–44.

Melville, P. and Sindhwani, V. (2010). *Recommender Systems*, pages 829–838. Springer US, Boston, MA.

Nabizadeh, A. H., Leal, J. P., Rafsanjani, H. N., and Shah, R. R. (2020). Learning path personalization and recommendation methods: A survey of the state-of-the-art. *Expert Systems with Applications*, page 113596.

Peralta, M., Alarcon, R., Pichara, K., Mery, T., Cano, F., and Bozo, J. (2018). Understanding learning resources metadata for primary and secondary education. *IEEE Transactions on Learning Technologies*, 11(4):456–467.

Pu, P., Chen, L., and Hu, R. (2011). A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 157–164.

Rivera, A. C., Tapia-Leon, M., and Lujan-Mora, S. (2018). Recommendation systems in education: A systematic mapping study. In *International Conference on Information Theoretic Security*, pages 937–947. Springer.

Santos, O. C., Boticario, J. G., and Manjarrés-Riesco, Á. (2014). An approach for an affective educational recommendation model. In *Recommender Systems for Technology Enhanced Learning*, pages 123–143. Springer.

Shani, G. and Gunawardana, A. (2011). *Evaluating Recommendation Systems*, pages 257–297. Springer US, Boston, MA.

Tarus, J. K., Niu, Z., and Mustafa, G. (2018). Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning. *Artificial Intelligence Review*, 50(1):21–48.

Truong, H. M. (2016). Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities. *Computers in Human Behavior*, 55:1185 – 1193.

Yu, X., Wei, D., Chu, Q., and Wang, H. (2018). The personalized recommendation algorithms in educational application. In *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pages 664–668. IEEE.