New Measures for Offline Evaluation of Learning Path Recommenders

Zhao Zhang, Armelle Brun, Anne Boyer

New Measures for Offline Evaluation of Learning Path Recommenders

Zhao Zhang, Armelle Brun, Anne Boyer

Université de Lorraine, CNRS, Loria
Campus Scientifique 54506 Vandoeuvre-lès-Nancy, France
{name.surname@loria.fr}

Abstract. Recommending students useful and effective learning paths is highly valuable to improve their learning experience. The evaluation of the effectiveness of this recommendation is a challenging task that can be performed online or offline. Online evaluation is highly popular but it relies on actual path recommendations to students, which may have dramatic implications. Offline evaluation relies on static datasets of students' learning activities and simulates paths recommendations. Although easier to run, it is difficult to accurately evaluate offline the effectiveness of a learning path recommendation. To tackle this issue, this work proposes simple offline evaluation measures. We show that they actually allow to characterise and differentiate the algorithms.

Keywords: Learning path recommendation · offline evaluation.

1 Introduction

Over the last decade, there has been an increasing adoption of technology in our everyday life. Nevertheless, this general adoption faces a main limit: students can be drowned in the large number of learning resources they can access. Educational recommender systems, that aim at selecting and recommending students learning resources, rely on the mining of the traces of interaction of students [18, 7], and are one solution to this problem.

The goal of recommender systems can be classified into two time horizons: short-term goal and long-term goal recommender systems. Short-term goal recommenders suggest users a set of resources to consult next. They are traditional recommenders and are still highly studied. In education, they recommend the next-step learning resource. Long-term goal recommenders recommend users sequences of resources. More and more attention is being paid to these recommenders, that are also called sequence-aware or path recommender systems. In education, recommending a sequence of learning resources is referred to as a Learning Path (LP) recommendation [8, 10, 22]. When students adopt a recommended LP, they may improve their level of knowledge, increase their chance of success at exams, maintain their motivation, etc.

Obviously, managing the accuracy of such a recommendation is of the highest importance. However, it remains relatively unstudied. In line with traditional

evaluation methodologies, the evaluation of the recommendation of a sequence or a LP can be performed either online or offline. Online evaluation is used in a live environment and focuses on the impact of the recommendations on the users. Although highly informative, online evaluation is highly time consuming, requires the availability of real users and is often not completely reproducible. At the opposite, offline evaluation focuses on the accuracy of the recommendations by relying on existing offline datasets. It is thus less costly, which justifies its popularity, but provides a limited estimation of the accuracy of the recommendation model, it is thus more challenging. Although offline evaluation has been highly studied for traditional recommender systems, few evaluation measures and frameworks have been proposed for the evaluation of sequences of recommendations, especially for offline evaluation [9].

The educational domain is highly sensitive when conducting online evaluation, as it relies on real recommendations to students. Indeed, low quality recommendations may negatively impact the students' learning experience, especially for LP recommendations, and can even ruin students' whole academic year. To the best of our knowledge, although offline evaluation of LP recommendations is attracting more and more attention, it still remains a challenge [21] and there is no commonly adopted evaluation measures. Several measures are simple adaptations of those used in traditional recommender systems, but they do not totally fit the characteristics of the educational context. Other measures, specifically designed for LP, tend to be highly complex and rely on specific data.

From this observation, this paper proposes to supplement the set of LP offline evaluation measures, to contribute to a more complete set of measures. These measures are designed to be simple, generic and easy to implement.

This paper is organised as follows. Section 2 discusses related work. Section 3 proposes new offline evaluation measures for LP recommendations, which are assessed in section 4. Finally, section 5 concludes this work.

2 Related Work

2.1 Learning Path Recommender Systems

Recommender systems are traditionally designed to recommend a user, referred to as the target user, one resource or a set of resources that fits his/her profile, preferences, etc. These short-term goal recommenders usually focus on the immediate value of the recommendations. Recently, the problem of recommending personalised sequences of resources to reach a long-term goal has emerged, but remains less widespread [12, 13]. Sequence recommendation is highly different from traditional recommendation [8] as it does not only exploit the temporal ordering for the generation of sequences of resources, but also contribute to reach a goal along the sequence.

In education, LP recommenders aim at identifying the most appropriate sequence of learning resources that fits the target student's preferences, capability or increases the chance of reaching a long-term goal (success at exam) [2, 21].

Several approaches have been proposed for sequence recommendations. For example, [5] relies on Markov chains to manage transition probabilities between and generate a sequence. However, this may not guarantee to reach the long-term goal, as the LP is made up of a series of "local" recommendations.

In the educational context, several approaches have been proposed, that differ not only in the data and information they exploit, but also in the way the paths are formed. Recently, [21] relied on deep learning, precisely LSTM associated with clustering. [3] focused on students learning styles and knowledge levels with a variable length genetic algorithm to recommend personalised LPs. Few years ago, Bayes theorem, Association Link Network, Item Response Theory, graph theory were studied to form and recommend LPs.

When developing a recommender system, an important dimension lies in its evaluation. As previously mentioned, two scenarios are highlighted: online and offline evaluation. The two following sections introduce some works conducted for both settings for LP recommendations.

2.2 Online Evaluation of LP Recommendations

Generally, online evaluation aims at quantifying the actual effect of the recommended LP on the students learning experience. Concretely, such an evaluation exploits the students' level of knowledge: results at exams, tests, quizzes, etc. at different times. Notice that commitment, motivation, etc. are rarely considered.

[10] proposed to use A/B testing, one of the most popular and simple online evaluation form. Students are divided into two groups: one group gets personalised LP recommendations, while the other group gets non personalised LP recommendations. After getting recommendations, the students in each group take the final exam. The average score in each group are compared. The main advantage of this approach is that it is easy to implement and the measure is very simple as it does not require additional elements However, from our point of view, the fact that the second group gets recommendations (non personalised) introduces a bias in the evaluation. [20] went a step further in the evaluation approach. The students' performance is not only compared between groups, but their performance before and after adopting a LP is also considered. Concretely, a single test is used, which is taken twice by the students. A pre-test score is associated to the first time the test is taken and a post-test score is associated to the second time the test is taken, i.e. after adopting a LP. The students are randomly divided into three groups: those who follow a self-decided LP, those who follow a non personalised LP, and those who follow a personalised recommended LP. Self-decided LPs are determined by students themselves (obtained by questionnaires for example) as proposed in [22]. In our view, the evaluation conducted here is more fine grained.

Although online evaluation is accurate, we can see from these works that conducting such an evaluation is costly, mainly in terms of time required and stakeholders availability (including students availability). We identify two additional limitations. First, if the recommendation model is not accurate, the recommendations proposed to students may negatively impact their learning experience.

Second, it cannot guarantee student equity, especially when students are split into groups and get different types of recommendations. This is the reason why offline evaluation is generally used before conducting online evaluation.

2.3 Offline Evaluation of LP Recommendations

Offline evaluation relies on predefined static datasets. In education, these datasets are mainly made up of students' activities, i.e. students' LP. At the opposite of online evaluation, no real LP recommendations and no test can be actually proposed to students. So, to evaluate the accuracy of LP recommendations, the ground-truth is thus required. This ground-truth can be inferred from the data, obtained from experts, or estimated from external knowledge [14]. Inferring ground-truth from the data is the most simple technique as it requires only data. Nevertheless, as it relies on data obtained independently of the recommender system under evaluation, it is always uncertain.

The literature assumes that for each target student one ground-truth LP exists, i.e. a LP that conducts to the expected goal, for example success at exams. The challenge is thus twofold: (1) identify the ground-truth LP, that is supposed to exist in the dataset, (2) compare this LP to the recommended LP. The measures proposed in the literature differ in these two elements.

However, although many measures have been proposed for traditional recommender systems, evaluating LP is far less popular. We can still identify two types of measures: those commonly used in many other domains and adapted to LP recommendations, and those specifically designed for LP recommendations.

Measures Used in Traditional Settings Precision is the most commonly used evaluation measure in information retrieval, classification, prediction and recommendation [13]. It represents the fraction of recommendations that are relevant. In the frame of LP recommendations, precision has been adapted to fit the complex nature of the LPs. It is often assumed that one LP is a series of resource recommendations, where each resource recommendation is viewed as either a good or a bad resource. Precision identifies the resources in the estimated LP that are also part of the reference LP: one recommended resource that is part of the reference LP is a good resource [6]. One main drawback of this technique is that it does consider not consider the order between the resources.

Conversely, [15, 21] considered a LP as a whole: good LP or a bad LP. A LP is considered as good if its estimated learning effect on the target student is greater than the expected student's learning effect when no recommendation is proposed. The associated precision measure thus represents the proportion of good LPs. In this work, the evaluation of the effect relies on both student and resource descriptions, as well as on results at tests. The main disadvantage of this measure lies in its complexity: it requires much information and the estimation of several elements: each resource degree of difficulty, each student's learning ability, the effect of a resource on a student, and the effect of a LP on a student. [9] has recently proposed a framework for offline evaluation of sequence-aware recommenders, that relies on the combination of 8 traditional measures used in recommendation, among which precision, coverage, diversity.

Another part of the literature proposes to rely on the evaluation of the distance between the recommended LP and the ground-truth LP. Concretely, the well-known Edit Distance (ED) is the most commonly used measure [22]. ED simply counts the minimum number of operations required to transform one of the sequences into the other. In [22], ground-truth LPs are determined by asking the students self-organized LP (through questionnaires), which is the only interaction with students. One main drawback in this work lies in the fact that students have to be able to self-determine their LP.

Measures Designed for LP The measures presented above are not specifically designed for sequence or path evaluation, nor for the educational domain. We will focus now on some of the rare measures recently proposed in the specific frame of LP recommendations.

[8] proposed two measures: promotion and logicality. Promotion represents the increase in the knowledge of students. This measure is close to the one presented in [20, 10], mentioned in the online evaluation section. The second measure proposed is more original, it represents the logicality of a LP, with respect to the knowledge structure in the recommended LP. This evaluation is not only performed automatically, by relying on the knowledge structure of the resources, but more originally, it is also performed by experts in education.

[11] proposed to automatically associate each resource with a score and a learning time, that can be estimated by several approaches such as the well known item response theory (IRT) [4]. Given a recommended LP, its accuracy is estimated by using the traditional Mean Absolute Error (MAE) between the real LP score and time, and the recommended path's estimated score and time.

3 New Measures for Offline Evaluation of LP

The previous section has highlighted that the literature lacks of well-adopted and well-defined offline evaluation measures for LP recommendations. In addition, we have seen that most of them requires additional information. For example, they require either expert knowledge, learning resource information: content, difficulty level, structure, or student knowledge: learning ability, level of knowledge, etc. They are thus highly complex to implement and may not be adapted to many contexts. This may partially justify their lack of success.

In this section, we propose new measures that cope with these limits. They are designed to be simple so that they can be easily and widely adopted, whatever are the contexts, datasets, even goals. Concretely, they rely only on students' traces of activities (their LP) and students' performance (results at exams or quiz for example). No content information or difficulty level about learning resources or about students is required. In line with the literature, they rely on a ground-truth LP. These measures do not require that this ground-truth LP is defined by experts. It is automatically inferred from the dataset. They vary not only in the way the ground-truth LP is defined, but also in the way the link between the recommended LP and the ground-truth LP is used.

As for online evaluation, two focus elements have guided the design of these measures. First, a recommended LP should be **in line with the target student's learning behavior**. Second, a recommended LP should **improve the student's level of knowledge**. A downside of the simplicity of the measures is that they do not fit all the students' level. Thus, for each measure, we will indicate the student level(s) it is mainly dedicated to. In addition, these measures may be combined during evaluation.

3.1 Problem definition

The problem can be defined as follows. Let t be a target student, known by (1) the sequence of resources he has consulted so far: his starting LP ($LP_{start}(t)$), (2) his estimated level of knowledge, from the grades obtained for the evaluations he took ($level(t)$). $RLP(t)$ is the recommended LP, i.e. the estimated LP that t should adopt as his ending LP ($LP_{end}(t)$).

An evaluation measure evaluates the accuracy of the LP recommended to t ($RLP(t)$). To reach this goal, it relies on a set of students S from the dataset and the associated traces of activity (their LP) ($LP(s),\ s \in S$), as well as their estimated level ($level(s),\ s \in S$).

The problem is thus: how to determine the ground-truth LP from the dataset and how to compare it to $RLP(t)$?

3.2 Mentor Students Based Measures

The first two measures introduced in this section rely on the hypothesis that students who have a high level of knowledge (i.e. get good grades at exams, quizzes, etc.) have an adequate learning behavior, i.e. they adopt an adequate LP. Their ending LP ($LP_{end}(s), s \in S$) is thus considered as the ground-truth LP. These students are considered as "mentor" students.

The exact way mentor students are chosen and the way the comparison of both LPs is taken into account varies between measures.

The GSLP Measure The Good Students LP based measure (GSLP) simply defines a mentor student m as a student with a high level of knowledge ($level(m)$) (i.e. has good grades), whatever is his LP. The set of mentor students in the set S of students ($m(S)$) is thus shared by all target students and can be defined beforehand the evaluation of GSLP. GSLP considers that if the recommended LP of a target student ($RLP(t)$) is close to the LP of a mentor student ($LP_{end}(s), s \in S$), the recommendation is accurate.

The main focus element of GSLP is thus the increase in knowledge of the students. Indeed, considering that the paths adopted by good students lead to a high level of knowledge (as the students who adopt them get good grades at exams), if the recommended LP is close to the one adopted by a good student, we can say that this path leads to an increase in the student's level of knowledge. The recommendation is thus accurate. $GSLP$ represents the proportion of target

students with an accurate recommendation, as presented in Equation (1).

$$GSLP = \frac{|t, \exists m \in m(S) \wedge d(RLP(t), LP_{end}(m)) < c|}{|S|} \quad (1)$$

where m is a mentor student, $d(L_1, L_2)$ is a distance measure between two LPs L_1 and L_2 and c is a maximal distance threshold. The larger GSLP, the better the recommender system.
Notice that the distance measure can simply rely on the number of resources shared in both LP (as in traditional metrics), or on the traditional Edit Distance.

Obviously, GSLP has several limits due to its simplicity. First, it does not consider the fact that several mentor students may have a LP close to the recommended LP of a target user, which could improve the accuracy of the measure. In addition, as in the traditional precision measure, the distance value is not explicitly taken into account in the evaluation of GSLP.

Notice that $GSLP$ has not been designed for students with a low $level(t)$ value. Indeed, it may not be adequate to consider that the LP adopted by good students is a good LP for other students, due to the large difference in their levels of knowledge, their LPs and their learning ability. Thus the values of GSLP for students with a low $level(t)$ value may not make sense.

The SSLP Measure The Similar Students LP based measure (SSLP) refines GSLP. A mentor student m is first of all a student who adopted a LP ($LP_{start}(m)$) close to the one of the target student ($LP_{start}(t)$). Second, a mentor student should be a successful student (as for GSLP).

At the opposite of GSLP, SSLP ensures that a mentor student ($m(t)$) has a learning behavior (LP) close to the one of the target student, so the main focus element of this measure is the learning behavior of the target student. The improvement in level of knowledge ranks second. Another difference with GSLP is that SSLP uses the distance value between the recommended LP of a target student ($RLP(t)$) and the actual LP adopted by his mentor student ($LP_{end}(m(t))$), not only the fact that a close path exists.

$SSLP$ represents the average distance between the recommended LPs and the ground-truth LPs. $SSLP$ is evaluated as defined in Equation (2). The smaller SSLP, the better the recommender system.

$$SSLP = \frac{\sum_{t=1}^{|S|} d(RLP(t), LP_{end}(m(t)))}{|S|} \quad (2)$$

As for GSLP, several distance measures can be adopted, as ED.

3.3 Difference in LP Based Measure (DLP)

The DLP measure addresses the limitation highlighted in GSLP: it is probably not adequate to consider that the recommended LP of a student who does not comply with the expected level (a promising student) should be close to the one adopted by a student who does (a good student). Intuitively, it should even be a highly different LP. DLP is in line with this intuition: it evaluates to what

extent the recommended LP of a target student, especially a promising student, is different from the ones adopted by good students. To avoid that a non-sense recommended LP is considered as a good recommended path (as it is highly different from any LP adopted by good students), this distance is also compared to the one with average students (with $level(s)$ close to the average level). Assuming that average students tend to have LPs less different from the ones of promising students, the paths recommended to average students should be closer to the ones recommended to promising students. Concretely, DLP represents to what extent the distance between the recommended LP of a target student and the actual LP of a good student is larger than the distance between the recommended LP of a target student and the actual LP of an average student. The highest this value, the better the recommendation.

The main focus element of DLP is thus the increase in the level of knowledge and at the second rank the learning behavior. Equation (3) presents the exact way this measure is evaluated.

$$DLP = \frac{\sum_{t=1}^{|S_p \cup S_a|} d(LP_{end}(m_g(t)), RLP(t)) - d(LP_{end}(m_a(t)), RLP(t))}{|S_p \cup S_a|} \quad (3)$$

where S_p and S_a are respectively the set of promising and average students, $m_g(t))$ is the mentor student of t in the set of good students and $m_a(t))$ is the mentor student of t in the set of average students. A mentor student $(m(t))$ is chosen as the student with the closest starting LP with t. At the opposite of the previous measures, DLP is dedicated to promising and average students.

3.4 Performance in LP based Measure (PLP)

As for the three first measures, PLP considers the ending LP of a mentor student as the ground-truth. PLP requires that at least two temporally distant evaluations have been taken by students, noted E_1 and E_2 (E_1 being taken before E_2).

Given a target student t, PLP relies on the following assumption: if there exists a student m with a level of performance associated to E_1 similar to the one of t, and with a level of performance associated to E_2 higher than the one of t, this means that the LP adopted by m is more adequate than the one adopted by t (as it leads to a higher increase in performance). m can thus be considered as the mentor of t. A similar assumption is made about m's ending LP. As a consequence, the closer the recommended LP of t to the ending LP of m, the better the recommendation. m is chosen as the student that meets a trade off between minimum distance in the levels of knowledge for E_1 and maximal distance in the levels of knowledge for E_2.

The main focus element of PLP is thus the increase in the level of knowledge of students, the learning behavior of t is not considered at all.

$$PLP = \frac{\sum_{t=1}^{|S|} d(RLP(t), LP_{end}(m(t))}{|S|} \quad (4)$$
$$\text{with } level(m(t), E_1) = level(t, E_1) \wedge level(m(t), E_2) > level(t, E_2)$$

$level(t, E_1)$ represents the grade obtained by t at the evaluation E_1. Notice that at the opposite of the previous measures, PLP can be used for each student, from promising to good students.

PLP and SSLP both evaluate the average distance between target students and their mentor student. They differ in the way their mentor student is selected: SSLP focuses on the learning behavior of students (LP), and PLP focuses on their level of knowledge.

4 EVALUATION

In this section we describe the experiments we conducted to assess the evaluation measures that we introduced into the previous section.

4.1 Dataset and Experimental Setup

The dataset used is made up of the traces of interactions of students with learning resources on their LMS. These students are first-year university students and are enrolled in a Mathematics and Computer Science Bachelor program. The traces represent the interactions for one specific course: "algorithms and programming" from the Fall Semester in 2018, which is a core course of this program. The students' LP are made up of the sequence of all accessed resources. Diverse learning materials are available: slides, exercises for lab sessions, quizzes and final exam. Each learning record represents one student activity, and includes (1) student ID, (2) timestamp of the interaction (3) type of access (view, submit), (4) resource ID (5) score (for quizzes and exam).

In this dataset, we identify two periods. The *course period*, where students access resources after they are uploaded by the teacher. Some activities are imposed by the teachers, other activities are optional and are mainly adopted on the students' initiative. The *review period*, where all learning resources are available and students review on their own initiative for the final exam.

Before the course starts, each student's level of knowledge is assessed on an initial quiz. At the end of the semester, each student takes a final exam. Besides, an optional mid-term quiz is proposed to students in the middle of the course period. Some of the students do not take the initial or mid-term quiz, but everyone took the final exam.

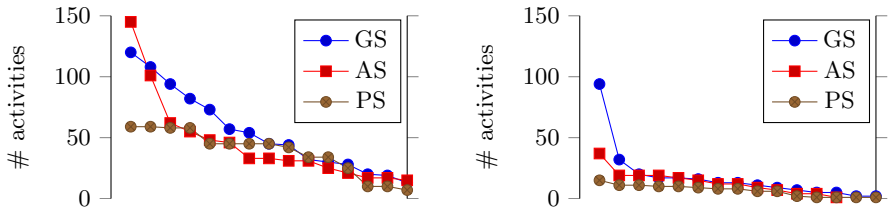
To perform a detailed analysis of the measures, we propose to divide students into three groups, based on their estimated level of knowledge. This level of knowledge is estimated with a weighted average of the quizzes and exam scores. The three groups are good students (GS), average students (AS) and promising

Table 1: Overview of dataset

Indicators		Values
	Number of students/resources	104/39
Course period	Number of learning records	3,279
	Median number of students' learning records	46
	GS/AS/PS average number of learning records	60/49/35
Review period	Number of learning records	693
	Median number of students' learning records	10
	GS/AS/PS average number of learning records	17/13/11

students (PS). The split criterion between groups has been determined by the teacher of the course, on the basis of his personal experience. As a result, GS represents the top 30% students, AS represents the 40% following students, and the remaining students are PS. As the number of students in each group is almost similar, we can consider that the dataset is well-balanced and that students' level will not bias the models trained.

Table 1 presents an overview of the dataset. First of all, we can see that this dataset is not a large dataset, which will be considered in the experiments.
We can also notice that students work more during the course period than during the review period, about 3.5 times more for each group. This can be due to the duration of the periods (course is 1.5 times longer than review), but it also means that students work less by themselves, which justifies the need of recommendations. In addition, good students tend to have a larger amount of activity than other students, but they do not tend to review neither more, nor less than promising students, w.r.t. their initial amount of activity. To refine this analysis, Figure 1, displays the number of activities (learning records) for each group of students. Students are sorted in descending order of their number of activities (1 point every 3 students). This figure confirms that the amount of activity of students differs between groups. Although we can see that the students who work the more during the course period are good students, we cannot say that students that have a low activity are promising students. GS and AS tend to have a similar number of activities during the review period.



(a) Course period (b) Review period

Fig. 1: Students' activities by group during both course and review periods

We start by explaining the way we propose to define training and test sets. Since the dataset is quite small in terms of number of students, number of resources and number of activities, we propose to adopt the leaving one out cross-validation method. One student (and his associated activities, whether the course of review period) is used as test (called the target student) and the remaining students form the training set. The traces of activities of this target student are split into two: course and review periods. The traces from the course period will be used for computing the recommended LP. The traces from the review period will not be used at all. Let us notice here that we decide to not compare the recommended LP to the actual LP of the students, as from our view this does not make sense for PS, and only a little bit for AS.

As mentioned in [8], the length of the RLP is required by any algorithm. In accordance with the literature, it will be set as the median length of the students' LP in the review period. It is thus fixed to 10 (see Table 1).

We propose to rely on two well-adopted LP recommendation models: Sequential Pattern Mining (SPM) [16, 19] and Markov Chain (MC) [17]. Similarly to [19], the recommended LP of the target student is formed by exploiting the patterns from the model, that fit his traces during the course period. The MC algorithm considers the recommendation as a decision making problem, and forms the recommended LP by exploiting Markovian properties [1]. The recommended LPs are formed by a naive optimal policy calculated by MC.

Recall that the goal of the evaluation is not to identify the best recommendation algorithm, but to study the evaluation measures. That is the reason why we did not choose the most evolved recommendation algorithms, but simple algorithms that we can thoroughly understand. As mainly proposed by the literature, the distance measure we use is the simple Edit Distance.

4.2 Experimental Results and Analysis

Each recommendation algorithm (MC and SPM) is trained and used to recommend each target student, in turn, a LP. As previously mentioned, to perform a detailed analysis, the set of target students is split into the three previously introduced groups (GS, AS, PS) and the four measures are evaluated for each group. Table 2 presents the associated values.

Table 2: Evaluation measures for each recommendation algorithm and group

Evaluation Measure	Student group	Recommendation	
		MC	SPM
GSLP	PS / AS / GS	0.5 / 0.6 / 0.7	0.4 / 0.5 / 0.6
SSLP	PS / AS / GS	6.9 / 7.3 / 7.9	6.8 / 5.4 / 5.8
PLP	PS / AS / GS	5.4 / 5.6 / 6.8	5.7 / 6.3 / 7.9
DLP	PS	3.5	1.4

GSLP measures to what extent the students' recommended path is close to at least one (ending) LP of good students. As GSLP does not perform any strict

selection of mentor students (except the fact that they are good students), it can be seen as the upper bound of the accuracy. It is expected that the values for AS and GS are high, for both algorithms. It is actually the case: for AS and GS, both algorithms tend to recommend LPs adopted by good students. The highest values are obtained for GS (0.7 and 0.6), which means that from 60% to 70% of good students receive a LP that is a LP close to the one adopted by good students. Notice that we expected that this value would be higher. This may be due to the fact the training set is made up of traces of all students' levels and that both MC and SPM do not consider the students' level during training. So, AS and PS influence the recommended LP. A simple experiment conducted on GS only (for train and test) shows a small improvement (> 0.8), which confirms the impact of promising students. SPM has a lower value than MC, whatever is the group. This was expected as SPM may suffer from overfitting in case of lack of data, which is the case of the dataset used. Notice that the values for PS are the lowest ones. As mentioned in section 3.2, this value may have non sense as recommending a promising student the path adopted by good students may not be adequate. We confirm that GSLP is an interesting measure. It highlights differences between both algorithms, and shows that MC tends to recommend paths close to those adopted by good students.

SSLP focuses on students' learning behavior, it evaluates to what extent the recommended LPs are close to the ones adopted by students with a similar learning behavior, and who also succeed in their exam. Obviously, none of both recommendation algorithms recommend LPs with this explicit purpose. First, we can see that SPM recommends LPs that better fit the students' learning behavior, which can be explained by the fact that SPM manages long patterns that better reflect students' learning characteristics. Average students are those who get recommendations that respect more their learning behavior, for both algorithms. This is probably explained by the number of possible students who adopt a close LP (some may be part of GS and others of PS). We confirm that SSLP is also an interesting measure, that shows the ability of the algorithms to recommend LP that are in accordance with students' LP. Here also, a difference between both algorithms is highlighted, for SSLP it is in favor of SPM.

PLP focuses on the increase in the level of knowledge of students, while respecting the students' learning behavior. As for SSLP, this information is not explicitly considered by both algorithms. First, we can see that, at the opposite of SSLP, MC is the most adequate recommendation algorithm in the sense of PLP: the paths recommended by MC are closer to the paths adopted by the students who increase more their level. For both algorithms, PS tend to get recommendations that actually correspond to an increase in their level, which can be justified by the fact that the training set is mainly made up of traces of activity of students that have better grades. AS for SSLP, the values associated to GS are not that good. They tend to not get recommendations that increase their level. A thorough study explains that this is due to a lack of 'candidate' mentors for good students. An additional experiment confirms that promising students negatively influence the recommendations made to good students.

DLP is totally different from the other three measures and is used for promising students only. It identifies to what extent the recommendation algorithms tend to recommend PS LPs that are more different from the ones adopted by good students than the ones adopted by average students. As previously mentioned, all groups are almost equally represented in the training dataset, which guarantees that there is no bias due to their weight. For both algorithms, the DLP value is positive, it confirms that the paths recommended to PS are closer to those adopted by AS, which is a positive finding. As mentioned in section 3.2, promising students should not receive LPs that are too similar to those adopted by good students. In addition, MC has the largest value, which shows that the recommendations proposed by MC to promising students are highly different from the ones adopted by good students, so they may fit the target students better.

Throughout these experiments we confirmed the usefulness of the measures proposed. They contribute to highlight some specificities of each of the recommendations algorithm studied. As mentioned previously, the algorithms used are simple algorithms, but we can analyse their recommendations and assess the relevance of the measures. Thanks to these measures, we can say that MC tends to recommend paths that are close to the ones adopted by average students, and that it recommends LPs that increase the level of knowledge of students. As for SPM, we can conclude that it is more adequate to recommend LPs that fit students' learning behavior.
Let us notice that if only one recommendation algorithm is used, the values of the measures, especially SSLP and PLP can be compared to a baseline that represents the average distance between the recommended LP and the actual ending LP of the target students.

5 Conclusion

Based on the fact that the literature lacks of offline evaluation measures for LP recommendation algorithms, this work has proposed new measures to contribute to such an evaluation. These measures are designed to be simple measures. First, they rely on few information: only students traces of activity and results at exams, so that they can be used in many datasets. Second, they are designed to fit the educational domain, at the opposite of several measures of the literature. Indeed, they are designed to represent elements such as the increase in knowledge level and the fitting to students' learning behavior. These measures have been assessed on recommendation algorithms of the literature. These algorithms have been chosen for their simplicity and the exact understanding of the recommendations performed.

In a future work, we plan to conduct a thorough analysis on more and larger datasets, as well as a more evolved distance measure to draw stronger conclusions. We also plan to propose additional measures to conduct a thorough analysis of the recommendation algorithms of the literature.

References

1. Durand, G., Laplante, F., Kop, R.: A learning design recommendation system based on markov decision processes. In: 17th ACM KDD (2011)
2. Durand, G., Belacel, N., LaPlante, F.: Graph theory based model for learning path recommendation. Information Sciences **251**, 10–21 (2013)
3. Dwivedi, P., Kant, V., Bharadwaj, K.: Learning path reco. based on modified variable length genetic algorithm. Education & Inf. Techno. **23**(2), 819–836 (2018)
4. Embretson, S.E., Reise, S.P.: Item response theory. Psychology Press (2013)
5. Feng, S., Li, X., Zeng, Y., Cong, G., Chee, Y.M., Yuan, Q.: Personalized ranking metric embedding for next new poi recommendation. In: 24th IJCAI (2015)
6. Hsieh, T., Wang, T.: A mining-based approach on discovering courses pattern for constructing suitable learning path. Exp. Syst. with App. **37**(6), 4156–4167 (2010)
7. Léonard, M., Peter, Y., Secq, Y.: Patterns and loops: Early computational thinking. In: Eur. Conf. on Technology Enhanced Learning. pp. 280–293. Springer (2019)
8. Liu, Q., Tong, S., Liu, C., Zhao, H., Chen, E., Ma, H., Wang, S.: Exploiting cognitive structure for adaptive learning. In: Proc. 25th KDD. pp. 627–635 (2019)
9. Monti, D., Palumbo, E., Rizzo, G., Morisio, M.: Sequeval: An offline evaluation framework for sequence-based rs. Information **10**(5), 174 (2019)
10. Nabizadeh, A., Gonçalves, D., Gama, S., Jorge, J., Rafsanjani, H.: Adaptive lp recommender approach using auxiliary learning objects. Comp. & Educ. **147** (2020)
11. Nabizadeh, A., Jorge, A., Leal, J.: Estimating time and score uncertainty in generating successful learning paths under time constraints. Exp. Syst. **36**(2) (2019)
12. Nabizadeh, A., Jorge, A., Leal, J.P.: Long term goal oriented recommender systems. In: Proc. of the 11th Webist. pp. 552–557 (2015)
13. Quadrana, M., Cremonesi, P., Jannach, D.: Sequence-aware recommender systems. ACM Computing Surveys (CSUR) **51**(4), 1–36 (2018)
14. Rossetti, M., Stella, F., Zanker, M.: Contrasting offline and online results when evaluating recommendation algorithms. In: Proc. 10th RecSys. pp. 31–34 (2016)
15. Su, C.: Designing and developing a novel hybrid adaptive learning path recommendation system (alprs) for gamification mathematics geometry course. Eurasia J. of Mathematics, Science and Technology Education **13**(6), 2275–2298 (2017)
16. Su, J., Tseng, S., Wang, W., Weng, J., Yang, J., Tsai, W.: Learning portfolio analysis and mining for scorm compliant environment. J. of Educational Technology & Society **9**(1), 262–275 (2006)
17. Taraghi, B., Saranti, A., Ebner, M., Schön, M.: Markov chain and classification of difficulty levels enhances the learning path in one digit multiplication. In: Int. Conf. on Learning and Collaboration Technologies. pp. 322–333. Springer (2014)
18. Venant, R., Sharma, K., Vidal, P., Dillenbourg, P., Broisin, J.: Using sequential pattern mining to explore learners behaviors and evaluate their correlation with performance in inquiry-based learning. In: EC-TEL. pp. 286–299. Springer (2017)
19. Vesin, B., Klašnja-Milićević, A., Ivanović, M., Budimac, Z.: Applying recommender systems and adaptive hypermedia for e-learning personalizatio. Computing and informatics **32**(3), 629–659 (2013)
20. Xie, H., Zou, D., Wang, F.L., Wong, T.L., Rao, Y., Wang, S.H.: Discover learning path for group users: A profile-based approach. Neurocomputing **254**, 59–70 (2017)
21. Zhou, Y., Huang, C., Hu, Q., Zhu, J., Tang, Y.: Personalized learning full-path recommendation model based on lstm neural networks. Inf. Sci. **444**, 135–152 (2018)
22. Zhu, H., Tian, F., Wu, K., Shah, N., Chen, Y., Ni, Y., Zhang, X., Chao, K.M., Zheng, Q.: A multi-constraint learning path recommendation algorithm based on knowledge map. Knowledge-Based Systems **143**, 102–114 (2018)