HAL
open science

Can we Take Advantage of Time-Interval Pattern Mining to Model Students Activity?

Oriane Dermy, Armelle Brun

To cite this version:

Oriane Dermy, Armelle Brun. Can we Take Advantage of Time-Interval Pattern Mining to Model Students Activity?. Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020), Jul 2020, Ifrane, Morocco. hal-02974678

HAL Id: hal-02974678

https://inria.hal.science/hal-02974678v1

Submitted on 22 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Can we Take Advantage of Time-Interval Pattern Mining to Model Students Activity?

Oriane Dermy, Armelle Brun
Université de Lorraine, CNRS, Loria
Campus Scientifique
54506 Vandœuvre-lès-Nancy, France
name.surname@loria.fr

ABSTRACT

Analyzing students' activities in their learning process is an issue that has received significant attention in the educational data mining research field. Many approaches have been proposed, including the popular sequential pattern mining. However, the vast majority of the works do not focus on the time of occurrence of the events within the activities. This paper relies on the hypothesis that we can get a better understanding of students' activities, as well as design more accurate models, if time is considered. With this in mind, we propose to study time-interval patterns.

To highlight the benefits of managing time, we analyze the data collected about 113 first-year university students interacting with their LMS. Experiments reveal that frequent time-interval patterns are actually identified, which means that some students' activities are regulated not only by the order of learning resources but also by time. In addition, the experiments emphasize that the sets of intervals highly influence the patterns mined and that the set of intervals that represents the human natural time (minute, hour, day, etc.) seems to be the most appropriate one to represent time gap between resources.

Finally, we show that time-interval pattern mining brings additional information compared to sequential pattern mining. Indeed, not only the view of students' possible future activities is less uncertain (in terms of learning resources and their temporal gap) but also, as soon as two students differ in their time-intervals, this difference indicates that their following activities are likely to diverge.

Keywords

Students behavioral patterns, time-interval pattern mining, interval granularities, sequential pattern mining.

1. INTRODUCTION

The wealth of data that can be collected from a Learning Management System (LMS), mainly the logs of students' interactions with learning resources, provide opportunities to get a more comprehensive understanding of students learning process: point out engaged or at-risk students, identify the most commonly studied or the most difficult resources, highlight recurrent students' activities, etc. In addition to this thorough understanding, inferences or decisions can be drawn: estimate students outcome, predict students future behavior (including dropout), personalize learning by providing students with information or recommendations, etc.

To carry out such understanding, inference or decision, data mining methods have been applied. Pattern mining, that discovers frequent patterns of events in data, is one of these methods and is also used in a large number of application fields. Sequential Pattern Mining (SPM) consists of discovering patterns when data is sequential in nature. These patterns, named sequential patterns, are frequent ordered sequences of events.

In the educational field, a sequential pattern often represents a recurrent sequence of learning resources, that we call an activity [30, 5].

The time of occurrence of events is often part of the data to be mined. However, in most of the cases, the patterns mined do not contain temporal information. Nevertheless, the literature has introduced different ways of including such information in patterns. We can, for example, cite temporal patterns, made of events that are associated with their time of occurrence [36], their duration [8], or the time gap between the events. In [9], gaps between events are grouped into intervals, resulting in time-interval sequential patterns. Since a time-interval pattern conveys more information than its corresponding sequential pattern, they are still the focus of research works [33]. In the rest of the paper, time-interval patterns will be referred to as *ti*-patterns and sequential patterns to as *s*-patterns.

We think that *ti*-patterns are adequate to represent students' activities. Indeed, it is rare that two students perform exactly the same activities, in both learning resources and time, even though they share underlying sequential activities. To the best of our knowledge, no work in the field of educational data mining has focused on the mining of *ti*-patterns.

In this work, we thus rely on the hypothesis that mining *ti*-patterns will contribute to a better view and understanding of students' learning activities. These patterns do not only indicate in which order students interact with learning resources, but provide also information about the temporal relationship between these resources. For example, let us

consider that students tend to interact sequentially with two resources, each of them being lecture slides. The sequence of both resources represents a sequential activity. Suppose that mining *ti*-patterns highlights that the time gap between both resources tends to be less than 1 minute for some students and between 2 and 4 hours for others. We can thus deduce beyond this sequential activity that there are two typical behaviors.
To support our hypothesis, we will conduct a study to evaluate if *ti*-patterns can be actually identified from students' activity data and evaluate to what extent *ti*-patterns provide additional information about students' activities.

In the following sections, we will first present an overview of related works on sequential and temporal pattern mining (Section 2). We then present the methodology we adopt to support the hypothesis that we draw (Section 3). Section 4 details the experiments we conduct on a real dataset and presents some *ti*-patterns. The last sections discuss the results (Section 5), then conclude the work and present our expected future work (Section 6).

2. LITERATURE REVIEW

2.1 Sequential Pattern Mining (SPM)

Sequential Pattern Mining is a popular task in Data Mining, introduced by Agrawal and Srikant in [1]. SPM aims to discover frequent sequential patterns in sequential databases. A sequential database D is a set of tuples $D = \{(sid_i, d_i)\}$, where sid_i is the unique identifier of a sequence, and d_i an input sequence. A sequence is an ordered list of events: $s = \langle E_1 E_2 \dots E_x \rangle$, with $E_i \in E$ the set of events. To understand what a frequent sequential pattern is, let us first define what a sub-sequence is. $\alpha = \langle E_1 \dots E_n \rangle$ is a sub-sequence of $\beta = \langle E'_1 \dots E'_m \rangle$ if:

$$\exists [1 \leq j_1 \leq \dots \leq j_n \leq m] | \{E_1 = E'_{j_1}, \dots, E_n = E'_{j_n}\}.$$

We also say that β is a super-sequence of α, or that it contains α. Let us now define what the support of a sequence is. The support of α, noted $supp(\alpha)$, is the number of sequences in the sequential database D that contain α. Based on both definitions, we can now define that a sub-sequence α is a frequent sequential pattern, if $supp(\alpha) \geqslant \delta$, for a defined minimum support threshold δ. We define SP as the set of frequent sequential patterns.

Many SPM algorithms have been proposed in the literature. The most commonly cited ones are *GSP* [32], *PrefixSpan* [26], *SPADE* [37]. All these algorithms use the "*apriori property*": "*If a sequence s is not frequent, then non of the super-sequences of s is frequent.*". Thus, when one pattern is infrequent, it is not extended. Algorithms can be divided into two main approaches. Apriori-like algorithms (also called *breadth-first* search algorithms), such as *Generalized Sequential Pattern Mining* (GSP) algorithm [32], are the first algorithms that have been proposed. However, these algorithms suffer from scalability problems, mainly due to memory requirements. *Depth-first* search algorithms, which include pattern-growth algorithms, do not suffer from memory complexity, which explains their popularity.
For a couple of years, the most common SPM algorithm is the *Prefix-Projected Sequential Pattern Growth* (*PrefixSpan*) algorithm [26], which is a pattern-growth algorithm, that relies on projected databases. Projected databases generally reduce the research space as the size of the projected databases decreases at each iteration. However, the main cost is linked to the generation of these projected databases. The pseudo-code of *PrefixSpan* is presented in Algorithm 1.

Algorithm 1 PrefixSpan (α,l,D)

Inputs:
α: a sequential pattern and l its length.
D: a sequential database, or a projected database.
Outputs:
SP: the set of all frequent sequential patterns.
Method:
Scan S to find all frequent items b.
for all b **do**
  add $\alpha' = \langle \alpha b \rangle$ to SP as a new sequential pattern.
end for
for all α' **do**
  create the α'-projected database $D|'_\alpha$
  call PrefixSpan(α',$l+1$,$D|_{\alpha'}$)
end for

In the works mentioned below, both the database and the patterns are sequential. However, in some cases, the database can be temporal, i.e. contain information about the time of occurrence of the events. In these cases a sequence is defined as: $s = \langle (t_1, E_1), (t_2, E_2), \dots, (t_n, E_n) \rangle$. where (t_i, E_i) represents an event E_i and its time of occurrence t_i.

When sequential patterns are mined from these databases, time can be either used as an information or order between events, such as in SPADE [37]. The time of appearance of events can also be used as a constraint. For example, in [18] the authors consider that when two consecutive items in a sequence are separated by a time gap bigger than a predefined threshold, they are temporally too distant to represent an association that makes sense. In the same context, [31] discards uninteresting patterns by introducing an interval constraint between items.

2.2 Sequential Patterns Mining in EDM

Sequential Pattern Mining has been extensively used in Educational Data Mining. They are mainly used to identify frequent patterns of students' activities [16, 28], including those that maximize the student learning performance [10]. In [21] SPM is used to study the differences in students' productive and unproductive learning behaviors and thus identify high versus low performing students. A similar objective has been studied on group work systems to understand the success factors in groups behavior [27, 25].
SPM is also used to detect learning problems early, such as in [20] where frequent sequential patterns and flag interaction sequences that are indicative of problems are mined.

One step further, SPM can act as a first step in decision making. In [7], the prerequisite structure of skills is find out, by identifying relations between variables from data. The algorithms developed in [28, 34, 11] provide students with personalized recommendations of learning resources according to their current activity or their learning style.
A complete view of various approaches used in educational data mining is presented in [2].

2.3 Temporal Pattern Mining

Temporal information appears to be fundamental in many contexts, hence the number of works interested in the mining of patterns that contain temporal information. Here again, time information can be used in several ways and for different goals: gaps, duration, intervals, etc.

Time information is often used as a gap between events of a pattern. For example in [36], the author considers that each occurrence of a sequential pattern may have different, but close, temporal elements. So, they propose to associate each pair of events of a pattern with a minimal, a mean and a maximal gap values between these events. The resulting model is made of sequential patterns enriched with temporal information, called delta patterns. Similarly, [15] proposes to add temporal information to each pair of events in a sequential pattern. This information, referred to as an annotation, represents a typical gap value between each pair of events of a pattern. In this work, the acceptance of the variation around this typical gap value is automatically evaluated. At the opposite of the previous works, [35] predefines a maximal gap value between events of a pattern, which results in temporal patterns called chronicles. [22] introduces an even more constraining frame, the exact gap interval value is imposed. This approach results in a decrease in the support of each pattern. Thus, the number of extracted patterns decreases.
In addition to the gap value, [17] exploits the duration of events. Each element of a pattern is composed of the event, associated with its begin and end timestamps. They propose an Apriori-like algorithm, that uses a hypercube representation of temporal sequences.
More recently, [13] introduces an Apriori-like temporal pattern mining algorithm on multi-modal data streams. At the opposite of the previous works, they do not only use the time gap between events (that represents the duration of the event), but also use the exact starting time of each event.

In line with the works presented above, [6] also manages gap values between events, that are grouped into intervals. At the opposite of other works, the intervals of gap values are predefined, and form "time-interval sequential patterns". A time-interval sequence is defined as:

$$\alpha = \langle E_1 \tau_1 E_2 \tau_2 \ldots \tau_{l-1} E_l \rangle$$

where $E_i \in E$ is the set of events for $1 \leq i \leq l$ and $\tau_i \in TI$ the set of time-intervals. The sequence α is a time-interval pattern if $supp(\alpha) \geq \delta$. We note TP the set of frequent time-interval patterns of a database D. In their article, the authors propose two algorithms called *I-Apriori* and *I-prefixSpan*, and results show that *I-PrefixSpan* outperforms *I-Apriori* both in computing time and scalability. The pseudo-code of the *I-PrefixSpan* algorithm is presented in Section 2.

A few years later, [19] pointed out that most algorithms of the literature use time information only as a time constraint or to represent the time-interval between successive items [9]. The novelty of this work is that not only the delay between successive items is taken into account, but also between distant items. The "multi time-interval (MI) sequential pattern" models the time-intervals between all pairs of items within a pattern. Two algorithms have been proposed,

Algorithm 2 I-PrefixSpan (α,l,D)

Inputs:
$\alpha = \langle E_1 \tau_1 \ldots \tau_{l-1} E_l \rangle$: a temporal pattern.
l: the length of α.
D: a sequential database, or a projected database.
Outputs:
TP: the set of all frequent temporal patterns.
Method:
Scan D to find each frequent pair (τ_l, E_{l+1}), where $\tau_l \in TI$ is the gap interval between items E_{l-1} and E_{l+1}.
for all (τ_l, E_{l+1}) **do**
  add $\alpha' = \langle E_1 \ldots \tau_{l-1} E_l \tau_l E_{l+1} \rangle$ to TP, as a new temporal pattern.
end for
for all α' **do**
  create the α'-projected database $D|'_\alpha$
  call I-PrefixSpan(α',$l+1$,$D|_{\alpha'}$)
end for

MI-Apriori and *MI-prefixSpan*, that are highly similar to the *I-PrefixSPan* and *I-Apriori* algorithms.

Discovering time-interval patterns has attracted considerable efforts, due to its widespread applications. However, several challenges remain, such as the definition of the adequate set of intervals (whether manual or automatic), including the problem of the granularity of the intervals.

2.4 Temporal Granularities

As soon as intervals are introduced, an issue arises: how to choose these intervals?
[3] proposes to manage different temporal granularities. An algorithm composed of Timed Automata with Granularities (TAGs), associated with heuristics is proposed. TAGs test whether a candidate time pattern appears frequently in a time sequence. The heuristic allows to reduce the number of candidates. [29] focuses on mining periodic patterns, where interesting periods cannot be defined in advance. Two temporal granules are proposed: a fine-grained granule for hourly periods and a coarse-grained granule for daily periods. The time distribution of different time granularities is then estimated by using a combination of Gaussian distribution.

2.5 Temporal Data Mining in EDM

To the best of our knowledge, little use has been made of Temporal Pattern Mining in the EDM field. [23] takes time into account by evaluating the rate at which students change the learning resources of interest. They progressively improve "when" resources have to be recommended to the student. In a learning context, where students can choose both which and when courses and exams to take, the research work presented in [4] uses time information that corresponds either to the "semester in which the exam was taken" or to the "delay with which it was taken". Using this time information, they then study the course and exam schedule that the students take and understand better students' behaviors. Using clustering and comparison, they are then able to suggest improvements to the scheduling of courses and exams of students.

3. DEFINITIONS AND METHODOLOGY

The previous literature review highlights that time-intervals are mainly adopted to model temporal patterns. The algorithm proposed in [6], *I-PrefixSpan*, has the main advantage to consider intervals as a core element of the patterns and the mining process. Intervals are considered as a constraint about the patterns, not as supplementary information about the patterns. It is the main reason why we choose to adopt this algorithm in our work.
We start by introducing definitions that will be used in the following methodology and in the experiments.

3.1 Definitions

Let $p = \langle E_1 \tau_1 E_2 \ldots \tau_{n-1} E_n \rangle$ and $p' = \langle E'_1 \tau'_1 E'_2 \ldots \tau'_{m-1} E'_m \rangle$ be two *ti*-patterns and $s = \langle E''_1 E''_2 \ldots E''_l \rangle$ be a *s*-pattern, with n (resp. m and l), the length of the pattern p (resp. p' and s). Given these patterns, we put the following definitions. Recall that TP is the set of frequent *ti*-patterns and SP the set of *s*-patterns.

Definition 3.1. ***ti*-form of an *s*-pattern**
p is a ti-form of s, denoted by $isform(s, p)$ if and only if:
$(n = m) \wedge (E_i = E'_i)$, $\forall i \in [1 : n]$.
ti-form(s) *is the set of* ti-forms*, in TP, of s.*

Definition 3.2. ***s*-form of a *ti*-pattern**
s is a s-form of p if and only if: $(n = s) \wedge (E_i = E'_i)$, $\forall i \in [1 : n]$. s-form(p) is the (unique) frequent s-form, in SP, of p.
s-form(P) is the set of frequent s-forms (in SP) of the set of ti-patterns $p \in P$.

Definition 3.3. ***s*-equivalence of *ti*-patterns**
p and p' are s-equivalent, denoted s-eq(p, p') if and only if:
$(n = m) \wedge (E_i = E'_i)$, $\forall i \in [1 : n]$.
In other words, s-form(p)=s-form(p').

Definition 3.4. **Prefix of a *ti*-pattern**
p' is a prefix of p if and only if:
$(m < n) \wedge (E_i = E'_i) \wedge (\tau_i = \tau'_i), \forall i \in [1 : m]$.

Definition 3.5. **Extension of a pattern**
p' is an extension of p if p is a prefix of p'. We note $ext(p)$ the set of extensions of p that belong to TP.
A similar definition can be put for s-patterns.

Definition 3.6. **Extended part of a pattern**
Let p' be an extension of p. The extended part of p, with respect to p', is the pattern p'', where $concat(p, p'') = p'$. Thus, $p'' = \langle E_n \tau'_n E'_{n+1} \ldots \tau'_{m-1} E'_m \rangle$.
We note $extPart(p)$ the set of extended parts of p, i.e. the set of patterns that, when concatenated with p, result in a pattern that belongs to TP.
A similar definition can be given for s-patterns.
Example: Let $p = \langle e_1 I_1 e_0 \rangle$, and $p' = \langle e_1 I_1 e_0 I_2 e_1 \rangle$ be two ti-patterns. $p'' = \langle e_0 I_2 e_1 \rangle$ is an extended part of p.

Definition 3.7. **Pseudo-equivalence of *ti*-patterns**
p and p' are said to be pseudo-equivalent, if and only if: $s\text{-}eq(p, p') \wedge (\tau_n \neq \tau'_n) \wedge (\tau_i = \tau'_i), \forall i \in [1 : n-1]$., i.e. they differ only in their last time-interval.

3.2 Methodology

To support our hypothesis and identify the actual value of a *ti*-pattern model, we define a methodology. More precisely, this methodology aims at identifying if there actually are temporal regularities between students' activities, if managing temporal activities allows to have a better view of students' future activities, and concretely what type of activities are mined. Recall that mining *ti*-patterns is quite new in educational data mining.

We intend to mine *ti*-patterns in a temporal database D, which is a database made up of temporal sequences. A temporal sequence is an ordered list of events (concretely a list of resources students interacted with) and their associated timestamp. Each temporal sequence represents one student's temporal activities and each student is represented by a unique (and long) sequence.
Our methodology relies on four steps, described hereafter.

3.2.1 Determining the set of time-intervals

Recall that although timestamps are discrete values, their precision is so high that relying on time-point (or gap) patterns will probably only lead to infrequent patterns. For example, two sequences that only differ by one second: $\langle (0, E_1)(3, E_2) \rangle$ and $\langle (0, E_1)(4, E_2) \rangle$ will correspond to two different patterns. Grouping gaps to form *ti*-patterns, will increase the support of patterns. In addition, if the intervals are appropriate, the loss of precision about temporal activities will be limited.
So, before assessing the relevance of mining *t*i-patterns, we have to choose the adequate set of time-intervals. Indeed, this set influences the information conveyed.

Let $TI = \{I_0, I_1, ...I_t\}$ be a set of time-intervals, where $I_j = [gapmin_j; gapmax_j[$ is an interval that contains all gap values between $gapmin_j$ and $gapmax_j$. Notice that the set of intervals should represent a continuum of gap values from $gapmin_0$ to $gapmax_t$.

We propose to evaluate the quality of a set of intervals TI with 3 criteria:
The fitting ratio. It is the ratio between the number of non-empty intervals and the total number of intervals. A non-empty interval is an interval that is part of frequent patterns. The higher the ratio, the better the set of intervals, as the number of "useless" intervals is low.
The number of intervals. On the one hand, the more intervals, the higher the potential of the model. Notice that when $TI = \{I_0\} = [0 : +\infty[$, it comes down to *PrefixSpan*. On the other hand, using too many intervals increases the complexity of the model. In addition, as there are many intervals, the *ti*-patterns discovered will probably be infrequent. Thus, a good set is a set that has an in-between number of patterns.
The horizon. It is represented by TI, the upper bound of the last interval (the maximal time value of the set of intervals). The larger the horizon, the more complete the model, as it is able to represent long-term recurrences.

From our point of view, the best set of intervals is the one that maximizes the fitting ratio while having a large horizon, with a limited number of intervals.

3.2.2 Comparing sets of s-patterns and ti-patterns

After having fixed the set of intervals, the set TP of *ti*-patterns can be mined. In this second step, we aim at comparing the set TP with the set SP set of *s*-patterns, and propose some measures to perform this comparison.

First of all, we propose to study the number of patterns, and their average length, to get a coarse-grained view of the set of patterns. Of course, this measure cannot be used alone, as the goal is definitely not to mine the highest number of patterns.
Second, we study the correspondence between both sets of patterns. Let us start by noticing that the number of *ti*-patterns cannot be deduced (not even approximately) from the number of *s*-patterns. A brief explanation follows.

Let s be a frequent *s*-pattern and $ti\text{-}cand(s) = \{ts_1, ts_2, \ldots, ts_k\}$ the set of the candidate-*ti*-forms of s. Note that ts_i may be infrequent. Two cases arise:
– $|\mathbf{ti\text{-}cand(s)}| = \mathbf{1}$. This case occurs when all the occurrences of s have the same candidate *ti*-form ts. Here, $supp(ts) = supp(s)$, thus ts is also frequent. The corresponding set of *s*-patterns is noted S^1.
– $|\mathbf{ti\text{-}cand(s)}| > \mathbf{1}$. This case occurs when some occurrences of s have different candidate *ti*-forms. As a consequence, $\forall i$, $(supp(ts_i) < supp(s)) \wedge (\sum_{i=1}^{k} supp(ts_i) = supp(s))$. Here, come three possibilities:
- $\nexists\ ts_i, supp(ts_i) > \delta$: there exists no frequent *ti*-form of s, thus the number of frequent patterns decreases. The associated set of patterns is noted S^0.
-$\exists!\ ts_i$, $supp(ts_i) > \delta$, thus: $\forall j | \{(1 \leq j \leq |ti\text{-}seq(s)|) \wedge (j \neq i)\}$, $supp(ts_j) < \delta$. In this case, there exists a unique frequent *ti*-form of s, the number of patterns remains stable.
- $\exists (i, j)$, $(i \neq j) \wedge (supp(ts_i) > \delta) \wedge (supp(ts_j) > \delta)$. In this case, there exist several frequent *ti*-forms of s, the number of patterns increases. The set of patterns associated with both last cases is noted S^{1+}. Based on this, we first introduce the pattern loss measure, that represents the ratio of *s*-patterns that have no *ti*-form in TP ($s \in S^0$).

$$pLoss(SP) = \frac{|SP| - |\bigcup_{p \in TP} \text{s-form}(p)|}{|SP|} \tag{1}$$

To complete the pattern loss measure, we define the support loss measure, which applies for any *s*-pattern that has at least one frequent *ti*-pattern ($s \in S^{1+}$). The support loss measure evaluates the proportion of "lost" occurrences of s, i.e. that have no correspondence in TP.
Let s be a *s*-pattern and $P = \{p_1, p_2, \cdots, p_k\}$ be a set of *ti*-patterns, where $isform(s, p_i), \forall p_i \in P$. The support loss of s is defined in equation (2).

$$sLoss(s) = \frac{supp(s) - supp^*(P)}{supp(s)} \tag{2}$$

where $supp^*(\cdot)$ is the support of a set of patterns, defined in equation (3).

$$supp^*(P) = |\bigcup_{p \in P} Seq_id(p)| \leq \sum_{p \in P} |supp(p)| \tag{3}$$

where $Seq_id(p)$ is the set of sequence ids in D, where p is a subsequence. We can see that the support of P is not defined as the sum of the supports of the patterns in P. To explain this, let us consider $P = \{p_1, p_2\}$, with p_1 and p_2 two *s*-equivalent *ti*-patterns.

By definition, the *s*-form of p_1 (which is the same as the *s*-form of p_2) occurs at most once in each sequence of D. Similarly, p_1 and p_2 occur at most once in each sequence, but both can occur in the same sequence. As a consequence, the support of P may be lower than the sum of the supports of p_1 and p_2.

The support loss defined above applies for a *s*-pattern. If the support loss has to be evaluated on a set of patterns, the average support loss and the associated standard deviation can be used.

3.2.3 Evaluating the impact of time on the set of possible future activities of students

In the following third and fourth steps, we aim to evaluate the benefit brought by time in patterns (through *ti*-patterns) about the possible future activities of students. To perform this evaluation, we adopt a two-stage approach.
Let p be a *ti*-pattern and $extPart(p)$ the set of extended parts of p (see Def. 3.6). From the educational point of view, the set of extended parts of a *ti*-pattern p represents the *ti*-activities that students frequently do after p.

In this third step, we aim at discovering if managing time allows to reduce the uncertainty about the future activities of students. We compare the set of extended parts of *s*-patterns and the set of extended parts of their *ti*-forms.
To conduct this comparison, we propose to use the well-known entropy measure. The entropy of a pattern p represents the "degree of disorder" of the set of its extended parts. From the educational view, given an activity performed by students, the entropy measures the uncertainty of its following activities. The higher the entropy, the more uncertain the following activities. Relying on the entropy is not new in the educational field [38]. Equation (4) presents the way the entropy of a *ti*-pattern p is evaluated.

$$Ent(p) = -\sum_{j=1}^{m} prob(p_j) log_2(prob(p_j)), \tag{4}$$

with $prob(p_j) = \frac{supp(p_j)}{\sum_{k=1}^{m} supp(p_k)}$ and p_j is one of the m extended parts of p. The same equation stands for *s*-patterns. Given a *s*-pattern s, we thus propose to evaluate the benefit of considering time-intervals in this pattern, by evaluating the entropy loss (see Equation 5). Entropy loss of an *s*-pattern s considers the entropy of s ($Ent(s)$) and the maximum entropy of its *ti*-form.

$$eLoss(s) = \frac{Ent(s) - max_{p \in \text{ti-form}(s)}\{Ent(p)\}}{Ent(s)} \tag{5}$$

Several cases may arise. First, $eLoss = 1$. This represents the best case: each of the *ti*-forms of s has exactly one extension. This means that when managing time in patterns, the future activities are totally certain.
Second, $eLoss = 0.0$. This case represents one of the worst cases: at least one *ti*-form of s has the same entropy as s. Here, we cannot say that managing time makes the possible future activity less uncertain.

Last, $eLoss < 0.0$. This case represents the other worst case: all the *ti*-form of s has an entropy higher than s. In this case, considering time decreases the quality of the model. Notice here that the term *Loss* is a misnomer as it may theoretically be < 0.0. However, this term has been chosen to be coherent with previous measures.
As a consequence, the higher the entropy loss ratio the more managing time in patterns contributes to better estimate students' future activities.

3.2.4 Evaluating the impact of a specific time-interval on students' future activities

This fourth and last step is dedicated to the evaluation of the impact of a specific time-interval of a *ti*-pattern on its extended parts. More precisely, we are interested in the impact of the last time-interval of a pattern. We focus on the following situation: given two pseudo-equivalent *ti*-patterns (*cf.,* 3.7), to what extent do their set of extended parts differ?
This evaluation allows to study to what extent two students, who perform the same temporal activity, except about the time of their last activity, do have identical future activities. In other words, is a temporal difference between two activities an indicator of activities that are beginning to diverge?

To perform this evaluation, we first evaluate the proportion of identical *ti*-patterns between pairs of sets of extended parts, as defined in Equation (6).

$$idExt(PQ) = \frac{\sum_{(P,Q)\in PQ} \frac{|P\cap Q|}{|P\cup Q|}}{|PQ|} \tag{6}$$

with $PQ = \{(extPart(p), extPart(q))\}|psd\text{-}eq(p,q)$ the pairs of sets of extended parts of all pseudo-equivalent pairs of *ti*-patterns. The higher this proportion, the lower the impact of the last time-interval.

Second, we rely on the proportion of s-equivalent extended parts. This measure also evaluates the impact of the last time-interval on the set of extended parts, but by considering only their sequential nature. The proportion of s-equivalent extended parts is defined in Equation (7).

$$sidExt(PQ) = \frac{\sum_{(P,Q)\in PQ} \frac{|\text{s-form}(P) \cap \text{s-form}(Q)|}{|\text{s-form}(P) \cup \text{s-form}(Q)|}}{|\text{s-form}(P,Q)|} \tag{7}$$

This proportion represents if students tend to share their following sequential activities, even though they differ in their last time-interval. Here also, the higher this proportion, the lower the impact of the time-interval.

Notice that for reasons of readability, s-form$(\cdot)$ is used here to represent the sequential form of a set of *ti*-patterns and a set of pairs of *ti*-patterns.

4. EXPERIMENTS

We apply the methodology described in the previous section to evaluate to what extent mining *ti*-patterns increases the knowledge about students' activities. We first present the dataset on which the experiments are conducted, then use the 4 steps of the methodology and draw conclusions for each of them. Finally, some mined *ti*-patterns are displayed.

4.1 Dataset overview and implementation

We collected data from 113 first-year university students, enrolled in a Mathematics and Computer Science Bachelor program and who interact with learning resources on their LMS. We focus on one specific course: algorithms and programming from the Fall semester in 2018. This course is a core course of this program. Diverse online materials are available: slides, exercises for lab sessions, tests, etc.
Most of the students own a personal computer, so they can access the course both during teaching hours (lectures or lab sessions) and after official teaching hours.
The set of events E is made of 35 learning resources, that students can consult. About 50% of these resources are studied during the teaching hours (lectures or lab). The dataset is made up of about 6,300 actions and each student sees on average 56 resources. The dataset spans almost one year, as it includes actions performed not only during the teaching period but also during revisions for the final examination and actions conducted for the retake examination (for the subset of students who failed the final examination).

In the experiments conducted, we use a relative minimum support $\delta = 0.1$. Two algorithms are studied: *PrefixSpan*, to mine sequential patterns and *I-PrefixSpan*, to mine *ti*-patterns. The source code used for *I-PrefixSpan* algorithm is the one available in [12] (we have slightly adapted the code to our needs). The source code used for the classical *PrefixSpan* algorithm is the one proposed by Gao [14].

4.2 Determining the set of time-intervals

We propose to study two types of intervals: Linear intervals, where each interval has an equal duration, and granular intervals, where the duration of intervals grows with the gap value.

Table 1 presents various sets of intervals studied. For each of them, the number of intervals, the maximal horizon, the fitting of the set, the frequency of each frequent interval, as well as the number of frequent patterns, are displayed. To avoid an artificially high fitting value, we consider that an interval is frequent if its frequency is no less than 10. The frequency of an interval is evaluated as the number of times the interval is used in the frequent patterns.

Before going into the details of the analysis of the set of intervals, we would like to mention that the sets do not all have the same number of intervals, so these values in Table 1 are not directly comparable. In addition, two contiguous granular intervals represent a totally different duration (for example up to 1 hour and up to 1 day), the frequencies are therefore not comparable. Last, notice that the total number of patterns in one set of intervals cannot be explained by the number of patterns of another set. Let us for example consider two sets of intervals and their associated number of patterns. Suppose that the first interval has an average duration twice longer than the second one. A pattern that is frequent in the first set may correspond to either two frequent patterns in the second set, or only one frequent pattern, or no frequent pattern at all (see section 3.2.2).

Let us first consider the three sets of linear intervals. For the two first sets (30 min and 1 hour), the fitting measure is quite low: 8%, which means that the vast majority of in-

Type	Duration	Number of intervals	Horizon	Fitting	Used intervals & associated frequency	# patterns
Linear	30 min	25	12h.	8%	$I_0 : 350,000$; $I_{24} : 1,770,000$	550,000
Linear	1 hour	25	12h.	8%	$I_0 : 549,380$; $I_{24} : 1,843,660$	356,811
Linear	1 day	25	24d.	72%	$I_0 : 90,976$; $I_1 : 42$; $I_2 : 25$; $I_3 : 54$; $I_4 : 119$; $I_5 : 36$; $I_6 : 239$; $I_7 : 189$; $I_8 : 17$; $I_{11} : 14$; $I_{12} : 44$; $I_{13} : 80$; $I_{14} : 17$; $I_{17} : 14$; $I_{18} : 13$; $I_{20} : 36$; $I_{21} : 45$; $I_{24} : 33,298$	37,764
Granular	expon.	16 $I_0 = [0$ sec. ; 10mn.[	8mt.	56%	$I_0 : 17,739$; $I_1 : 79$; $I_8 : 82$; $I_9 : 269$; $I_{10} : 4,278$; $I_{11} : 5,126$; $I_{12} : 6,403$; $I_{13} : 3,693$; $I_{14} : 1,159$	15,754
Granular	human	6 $I_0 = [0$ sec. ; 1 mn.[	1y.	100%	$I_0(sec) : 7,706$; $I_1(min) : 10,551$; $I_2(hour) : 1,615$; $I_3(day) : 30,925$; $I_4(week) : 68,614$; $I_3(month) : 22,479$	51,025

Table 1: Fitting, examples of intervals and number of patterns for several sets of intervals

tervals are not found in frequent patterns. For example, in "30 min", only the first interval (between 0 and 30 minutes) and the last interval (more than 12 hours) are not empty. We can conclude that both sets of intervals are not good candidates. Caution must be exercised in interpreting this result. It might mean that students do not regularly switch from one resource to another, with a time gap between 30 minutes and 12 hours. It can also mean that the 30 min. time-interval is not relevant. Despite the lack of relevance of these intervals, the number of patterns discovered is important. As only two interval patterns are used, we can consider that *I-PrefixSpan* behaves here almost as PrefixSpan.
The fitting value of the "1 day" set is quite larger: 72%, which means that most of the 25 intervals are frequent. However, the total number of frequent patterns in this set is highly decreased, compared to the "30 min" and "1h" sets (by about 10 times). In addition, many interval frequencies are not so high, some of them being close to the minimal threshold, except the first and last one. This tends to mean that many intervals are not that representative of the data. Moreover, although the number of intervals is quite large (25), the maximal horizon represented by this set remains limited (all together, except the last one, represent a horizon of smaller than a month). Recall that the dataset spans almost one year. Obviously, the horizon can be extended, but it will be at the cost of an even larger number of intervals, as well as an increase in the space and computation time.
These results tend to suggest that the set of intervals should contain small intervals for close events (such as suggested by the frequency of I_0 in the 30 min set), and larger intervals for furthest gaps (such as suggested by the frequency of I_{24} in the 1 day set). Thus, a granular set of intervals should better fit the dataset.

We propose to study now two sets of granular intervals. In the first set, the duration of intervals grows exponentially: the duration of an interval is twice larger than the duration of the preceding interval. The fitting of this set is greater than for the two first ones, but smaller than the third one. Nevertheless, the horizon is larger than for all the previous ones (about 4 months), and the number of intervals is decreased. The empty intervals (from I_2 to I_7) tend to represent a gap between 20 min and 10 hours 40 min.

The second set of granular intervals is referred to as "human", the intervals are designed to represent the human natural time: minute, hour, day, week, etc. This set of intervals has a maximal fitting (100%). At the opposite of the "1 day" intervals, that has the highest fitting value till then, the frequency of each interval is quite large (greater than 1,600) and the number of intervals is reduced (only 6 intervals). Besides, the total number of patterns is larger than both the "1 day" and the "exponential" sets.

All these elements contribute to consider the "human" set as the best set of intervals. In this set, time is represented by the {minute, hour, day, week, month, year} intervals. This set has a maximal fitting (100%), covers a large horizon (till a year, which corresponds to the span of the dataset), with a limited number of intervals (6 intervals) and provides a quite large number of frequent temporal patterns. Therefore, in the following experiments, this set of intervals will be used.

Given these elements, we would like to highlight that this set of intervals intrinsically represents the classical rhythm of courses, for example one lecture (or one lab session) is planned each week. The human set of intervals thus allows to mine patterns that represent natural students temporal activities: some students tend to work immediately following a lab session (or a lecture) represented by I_0 or I_1; other students wait for some hours in the same $24h$, and others work during the week, or even the week after (before the next session) represented by I_4. It is typically the type of information that we expect to get when we aim at modeling students' activities.

4.3 Comparing s-patterns and ti-patterns

This second experiment aims at comparing sets of *s*-patterns and *ti*-patterns. Table 2 presents both sets of patterns, associated with measures introduced in the methodology. Let us first focus on the number of frequent patterns (line 1). The total number of frequent *ti*-patterns is dramatically smaller than the number of frequent *s*-patterns. The pattern loss is larger than 0.99. This means that the great majority of *s*-patterns has no frequent *ti*-forms, probably due to the spread of occurrences of *s*-patterns over numerous *ti*-patterns. These findings are in line with [22]. In addition,

$\delta = 0.1 (= 11)$	*PrefixSpan*	*I-PrefixSpan*
Number of patterns	$\|SP\| = 12,826,760$	$\|TP\| = 51,025$ - $pLoss(SP) = 0.998$
Average	8	3.8
Max. length	17	8
Example of pattern s, $frequEnt(s)$ $\wedge\{\nexists p \mid (isform(s,p) \wedge frequEnt(p))\}$	$s = \langle e_{31}, e_{29} \rangle$ $supp(s) = 26$	$ts_1 = \langle e_{31} I_0 e_{29} \rangle,\ supp(ts_1) = 1$ $ts_2 = \langle e_{31}\ I_1\ e_{29} \rangle,\ supp(ts_2) = 10$ $ts_3 = \langle e_{31}\ I_2\ e_{29} \rangle,\ supp(ts_3) = 6$ $ts_4 = \langle e_{31}\ I_3\ e_{29} \rangle,\ supp(ts_4) = 3$ $ts_5 = \langle e_{31}\ I_4\ e_{29} \rangle,\ supp(ts_5) = 4$ $ts_6 = \langle e_{31}\ I_5\ e_{29} \rangle,\ supp(ts_6) = 7$
Example of pattern s, $frequEnt(s)$ $\wedge\{\exists (p_i, p_j) \mid (frequEnt(p_i) \wedge frequEnt(p_j))\}$	$s = \langle e_{22}, e_{33} \rangle$ $supp(s) = 53$	$p_1 = \langle e_{22} I_0 e_{33} \rangle,\ supp(p_1) = 25$ $p_2 = \langle e_{22} I_1 e_{33} \rangle,\ supp(p_2) = 22$
Support loss		sLoss(SP^{1+})= 0.33 ; std(sLoss(SP^{1+}))=0.10

Table 2: Comparison of sets of patterns mined with *PrefixSpan* and I-PrefixSpan

we can see in Line 2 that the average length of *ti*-patterns is about twice smaller than the length of *s*-patterns, the same for their maximal length. A first conclusion that can be drawn here is that most of frequent sequential patterns have no recurrences in their time-intervals. This means that students tend to have numerous recurrent sequential activities, and quite less recurrent time-interval activities. However, even though the average length of patterns is divided by 2, *ti*-patterns have a significant length, which means that they do represent a meaningful students' activities.
Moreover, a tens of thousands *s*-patterns (about 34,000) have one or more frequent *ti*-forms (about 51,000). This means that for these sequential activities, there are actually temporal regularities. These activities will be studied in more detail in the following section.

Lines 4 and 5 in Table 2 illustrate some examples of *s*-patterns and their candidate or frequent *ti*-forms. Line 4 presents one of the 99.8% *s*-patterns that has no *ti*-form (thus, from SP^0). This pattern ($s = \langle e_{31}, e_{29} \rangle$) has 6 candidate *ti*-forms, but none of them is frequent. We can conclude that no obvious time-interval regularity is observed for this activity. Thus, this activity does not seem to be guided by temporal constraints. We can also observe here that the sum of the support of the *ti*-patterns is greater than the support of their *s*-form s. This was mentioned in section 3.2.2.

In the remaining sequential patterns (SP^{1+}) made up of about 34,000 *s*-patterns, 65% of the *s*-patterns have exactly 1 frequent *ti*-form and 91% have 1 or 2 frequent *ti*-forms. The highest number of frequent *ti*-forms of an *s*-pattern is 9, which is quite high. Let us now consider line 5 in Table 2, that presents a *s*-pattern that has several frequent *ti*-forms. This *s*-pattern ($s = \langle e_{22}, e_{33} \rangle$) has a support equal to 53 and two frequent *ti*-forms. Such a pattern occurs with two temporal recurrences, and most of its occurrences have a time gap between 1 minute and 1 day. Such patterns are highly interesting and will also be further studied.

Based on these findings, it is legitimate to ask whether a *ti*-pattern-based model can replace a *s*-pattern-based model. Line 1 gives first indications. Many sequential patterns "disappear" with such a model (more than 99% of sequential patterns have no frequent *ti*-pattern). If the objective is to replace traditional *s*-patterns by *ti*-patterns, a problem of coverage of the model arises. However, if the goal is to identify which activities (sequential) have temporal regularities, *ti*-patterns are of the highest interest.

Let us now focus on the support loss associated with the complete set SP^{1+} of *s*-patterns that have at least one frequent *ti*-form. $sLoss(SP^{1+}) = 0.33$, with a standard deviation equal to 0.1. This means that on average 1/3 of the occurrences of an *s*-pattern "disappear", i.e. they do not belong to any frequent *ti*-form. We can conclude that among patterns with identified temporal regularities, 33% of the occurrences do not follow this regularity, which may be high.

4.4 Evaluating the impact of time on the set of possible future activities of students

Following our methodology, we evaluate now if *ti*-patterns carry more information than *s*-patterns about future activities of students. As a preliminary remark, we would like to mention that $\nexists s, eLoss(s) \leq 0$. We mentioned previously that this case would occur rarely, in practice here it does not occur.

In the set S^{1+}, 71% of the patterns have at least one extension in SP (see Def. 3.5). Let us first consider the 66% of these patterns that have a unique extension. By definition for these patterns, $Ent(s) = 0$ and $Ent(p) \geq 0, \forall p \in$ ti-form(s). The first Line of Table 3 is an example of such a case. The s-pattern $\langle e_{24}\ e_{27}\ e_{14} \rangle$ has only one extended part, so its entropy equals zero. It has three ti-forms, but only one has an extended part. So, all these ti-forms have an entropy equals to zero.
In this case, even if the entropy loss is null, the information about the future activities of students is increased, as only one *ti*-pattern has a frequent extended part.
Let us now consider the 34% remaining patterns, which have more than one extension in SP. The average entropy is 0.84 with a maximal entropy of 7.71. When focusing on the set of their *ti*-forms, the average entropy is 0.35 and the maximal entropy is 6.22. To make entropies as comparable as possible, the average entropy for *s*-patterns has been evaluated only on the set of *s*-patterns that have at least one *ti*-form. We can first notice that entropy of *s*-patterns is globally higher than the one of *ti*-patterns (for both maximal and average values). More precisely, the average entropy of *s*-patterns is 2.4 times bigger than the one of *ti*-patterns. We can thus draw a first global conclusion: managing time in

s-pattern	Examples of extPart(s)	Ent(s)	Examples of p $\in$ *ti*-form(s)	nbExt(p)	Examples of extPart(p)	max-Ent(p)	mean-Ent(p)
$\langle e_{24}\ e_{27}\ e_{14}\rangle$	$\langle e_{12}\rangle$	0.0	$\langle e_{24}\ I_2\ e_{27}\ I_3\ e_{14}\rangle$ $\langle e_{24}\ I_3\ e_{27}\ I_3\ e_{14}\rangle$	0 1	$\langle e_{14}\ I_3\ e_{12}\rangle^{(*)}$	0.0	0.0
$\langle e_1\ e_{10}\ e_{12}\rangle$	$\langle e_{\{3,13\}}\rangle$ $\langle e_{19}\ e_{\{3,22\}}\rangle$ $\langle e_{12}\ e_{\{12,19\}}\rangle$	5.49	$\langle e_1\ I_5\ e_{10}\ I_1\ e_{12}\rangle$ $\langle e_1\ I_5\ e_{10}\ I_0\ e_{12}\rangle$ $\langle e_1\ I_5\ e_{10}\ I_2\ e_{12}\rangle$ $\langle e_1\ I_2\ e_{10}\ I_2\ e_{12}\rangle$	0 1 13 24	$\langle e_{12}\ I_4\ e_3\rangle$ $\langle e_{12}\ I_5\ e_{\{19,13,12\}}\rangle$ $\langle e_{12}\ I_5\ e_{\{19,13,15\}}\rangle$	4.55	0.78

Table 3: Examples s-patterns with ti-forms**, extended parts and entropy values.** (*) **The corresponding extension pattern is** $p'' = \langle e_{24}\ I_3\ e_{27}\ I_3\ e_{14}\ I_3\ e_{12}\rangle$**.**

patterns allows to decrease the uncertainty of students' future activities.
We will now compare the entropy of each *s*-pattern, with the entropy of its *ti*-forms (through *eLoss*). In 68% of the cases, the entropy loss between the *s*-patterns and their *ti*-forms is higher than 0. This means that when considering a temporal student activity, in 2 cases out of 3, the future activity of this student is less uncertain than when managing his/her sequential activity. These 68% are divided into 51% with a loss equal to 1, which means that future activities become certain. 17% of the cases have a loss between 0 and 1. The average entropy loss on all s-patterns is quite high: $eLoss = 0.4$. Roughly speaking, the future activities of students are on average 40% less uncertain when managing time in patterns, which is highly promising.

Thanks to these experiments, we confirm that managing time-interval patterns allows, in most cases, to have a better view of the following activities of students. In addition, for a significant number of activities, future activities are now totally certain.

Let us now focus on an example presented in the second Line of Table 3. The *s*-pattern $s = \langle e_1 e_{10} e_{12}\rangle$ has many extensions in SP and many *ti*-forms, among which many of them have extensions. Notice that although the entropy loss is low (the maximal entropy of the *ti*-forms is 4.55), on average it is significantly lower (0.78). In this specific case, *eLoss* measure is not that representative of the difference in entropy, the entropy decrease is probably higher than the *eLoss* value.

4.5 Evaluating the impact of a specific time-interval on students' future activities

The experiments conducted here fall within the scope of the last step of our methodology. They aim at evaluating to what extent two students who perform a similar activity (both in terms of resources and time-interval) and who only differ in their last time-interval, have the same future activities. In the experiments conducted, we will only focus on patterns made up of at least 3 events (and 2 time-intervals) to ensure that the patterns can be considered as activities.

In the set PQ composed of $|PQ| = 9,510$ of pseudo-equivalent pairs of patterns (*cf.,* Definition 3.7), 25% of the extended parts of a pattern of any pair are also part of the extended parts of the other pattern (sequentially and temporally identical). 11% additional pairs have sequentially identical extended parts. This highlights that even when two *ti*-patterns differ in their last time-interval only, this small difference leads to a significant difference in their sets of extended parts. In terms of students' activities, this means that when two students make exactly the same activity, except on the last time-interval, their following activities mainly differ: not only in terms of temporal activities but also in terms of their sequential activities. We can conclude that the last time-interval highly influences students' future activities and that it may be viewed as an indicator of activities that are beginning to diverge.

Experiments conducted in both previous sections confirm that *ti*-patterns contribute to the increase of the information about students' future activities whereby the uncertainty of this future is reduced. As a consequence, we can say that time is an important information in students' activities.

4.6 Interpretation of ti-patterns

In this section, we present examples of frequent *ti*-patterns, in an understandable format to better analyze and understand students' activities.
The events ids in patterns are replaced by their type and an id. Lec_n will refer to the slides associated with the nth lecture; $Glos_n$ will be the nth glossary resource; Stx_n a syntax resource; Sum_n will be a summary resource ; Lab_n a resource that contains exercises that are studied during lab sessions (exercise sheets); FA_n are facultative additional exercises; finally Ad is the advise resource. The time-intervals are noted $(I_s, I_{mn}, I_h, I_d, I_w, I_{mt})$, which refer to seconds, minutes, hours, days, weeks and months.
Given that the longer an activity, the more information it contains, we will preferably focus on the longest *ti*-patterns.

Activities made up of temporally close events
Let us start by studying activities that contain only the "seconds" time-interval (i.e. events with a maximal gap of 1 minute). This will allow us to have a better view of the type of activities that are performed on the spot. First, the corresponding activities tend to be made up of specific types of events: they are a mix of glossary, syntax, advertisement and lab resources. Second, the maximum length here is 7, which means that there are actually long recurrent "quick" activities made by students. Third, when analyzing the activities, we can remark that they all have a similar skeleton: students generally start by looking at the following resources (in any order): $\{Sum_5, Stx_3, Glos_3\}$, then study one or more *Lab* exercises and finally consult an advice page. Let us for example present a *ti*-pattern of length 7:
$\langle Sum_3\ I_s\ Stx_3\ I_s\ Glos_3\ I_s\ Lab_1\ I_s\ Lab_2\ I_s\ Lab_3\ I_s\ Ad\rangle$
Such patterns can be interpreted as follows: they represent

typical activities performed when preparing an exam. Not only several *Lab* sheets are studied, but also before these resources, students have a quick look at the syntax, glossary, and summary of the lectures. They finally consult the advice page. In such patterns, students interact with resources within a short time, including with the lab sheets.

Activities made up of I_h time-intervals

Let us now focus on patterns that use the "hours" time-interval, where patterns are made up of events with a gap value between 1 hour and 1 day. Here again, we identify a skeleton shared by most of the *ti*-patterns:
$\langle Lab_{\{1,3\}}\ I_h\ Lab_{\{2,3\}}\ I_h\ Lab_{\{2,3,4\}}\rangle$, where $Lab_{\{1,3\}}$ means either Lab_1 or Lab_3.
These patterns highlight that some students tend to work sequentially on several exercise sheets. The gap being between 1 hour and 1 day, tends to mean that students dig deep into their works: they spend some hours to perform each exercise sheet.

Other intervals

We have performed a similar study on other time-intervals. For each of them, we also identify skeletons shared by almost all the patterns.

An interesting conclusion that can be drawn from these findings is that for any given time-interval, typical long activities are made by students, that do all have the same skeleton.
More importantly, when comparing skeletons between time-intervals, they are totally different. We can thus conclude that the type of activity performed is strongly linked with the "rhythm" of the activity. Here, "rhythm" means a time-interval granularity shared by all gap between all events of the activity.
Last, when studying the timestamps associated with each occurrence of the activities presented above, there is no specific period associated: they are performed at any moment in the semester. For example, when considering the first example given, that is mainly related to the 3rd lecture, we found similar patterns for the 1st, 2nd, etc. lecture resources.

5. DISCUSSION

While traditional studies emphasize that students have typical sequential learning behaviors (identified by frequent sequential patterns), this study further emphasizes that for specific activities students work with temporal regularities. Based on the experiments conducted in the previous sections, we initiate a discussion.

The results have highlighted that among the sets of intervals tested (linear and granular), the one that represents the human natural time is the most relevant one, at least for the dataset used in the experiments (see section 4.2). In addition to outperforming other sets of intervals according to predefined measures, this set conforms to the scope of application: the duration of most of the lectures or lab sessions is about one hour, two successive lectures tend to occur each week, etc. So, the interpretation of the discovered patterns is enhanced. Of course, many other sets of intervals remain untested and may be more adequate. Besides, an automatic approach that learns the optimal set of intervals could be tested, as in [24]. However, this would be at a significant additional computational cost, without any guarantee of applicative interpretability of these intervals.

As expected, a high number of sequential patterns have no frequent *ti*-form. In the experiments conducted, we have even highlighted that most of the sequential activities have no temporal regularities. This results in a high number of "lost" patterns, which can be problematic, in case we are interested in both frequent *ti*-patterns and *s*-patterns. A solution could manage both types of patterns: sequential students' activities mixed with temporal students' activities. This solution would not only maintain the coverage of the model, thanks to sequential patterns but also manage time, thanks to temporal patterns, when suitable. Here is an example of such a pattern: $\langle E_2\ E_{27}\ I_1\ E_{13}\rangle$. This pattern means that many students consult E_2 then E_{27} (whatever is the time-interval), then between 1 minute and 1 hour later they do consult E_{13}.

Focusing on *s*-patterns and their various frequent *ti*-forms can help to highlight different learning approaches adopted by students. For example, an activity done with a gap lower than 1 minute between its events may represent the fact that the associated students are used to first download all the resources and then work offline. The same activity with a time gap between 1 minute and 1 hour may reflect that students do work online, they do not access a resource before finishing the previous one. So, in addition to highlighting the diversity of activities of students, *ti*-patterns are also a way to identify students' learning practices. One can foresee that these patterns could be used as input information for many works such as those that focus on students' engagement.

6. CONCLUSION AND FUTURE WORKS

The study presented in this paper highlights the relevance of using time information when mining patterns of students' activities. A time-interval pattern mining approach, through the *I-PrefixSpan* state-of-the-art algorithm, has been adopted to conduct this study.

The experiments conducted have pointed out that the nature of the set of intervals used highly impacts the representativity of the model and that the set of intervals that represents the human natural time is adequate. We also found that most of the sequential students' activities do not correspond to any time-interval activity. However, for other cases, managing this time-interval provides a better view of the future possible students' activities, thanks to temporal indicators. Moreover, results show that a single time-interval difference between two events of two patterns sequentially equivalent results in significantly different subsequent activities.
We thus confirm our hypothesis: temporal information is highly promising for a more precise modeling of students' activities. One additional experiment has illustrated some frequent students' activities both temporal and sequential. It has put forward that, by looking at some specific time-intervals, we can understand what activities students often perform instantly or throughout a longer period.
The work we have conducted provides a first step towards longer-term research. One of our future goals is to provide students with recommendations of educational resources. By relying on *ti*-patterns, we are confident that not only the accuracy of the recommendations provided to students will be increased but also that these patterns will give indications about the right time to propose recommendations to students.

7. REFERENCES

[1] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proc. 11th Int. Conf. on Data Engineering*, volume 95, pages 3–14, 1995.

[2] N. Anjum and S. Badugu. A study of different techniques in educational data mining. In *ICETE*, pages 562–571. Springer, 2020.

[3] C. Bettini, X. S. Wang, S. Jajodia, and J. Lin. Discovering frequent event patterns with multiple granularities in time sequences. *Trans. Knowledge Data Engineering*, 10(2):222–237, 1998.

[4] R. Campagni, D. Merlini, R. Sprugnoli, and M. C. Verri. Data mining models for student careers. *Expert Systems with Applications*, 42(13):5508–5521, 2015.

[5] R. Cerezo, M. Sanchez-Santillan, J. Nunez, and M. P. Paule. Different patterns of students interaction with mooodle and their relationship with achievement. *Computer Science*, 2015.

[6] Y. Chen, M. Chiang, and M. Ko. Discovering time-interval sequential patterns in sequence databases. *Expert Systems with Applications*, 25(3):343–354, 2003.

[7] Y. Chen, P.-H. Wuillemin, and J.-M. Labat. Discovering prerequisite structure of skills through probabilistic association rules mining. *International Educational Data Mining Society*, 2015.

[8] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee. Mining temporal patterns in time interval-based data. *TKDE*, 27(12):3318–3331, 2015.

[9] Y.-L. Chen, M.-C. Chiang, and M.-T. Ko. Discovering time-interval sequential patterns in sequence databases. *Expert Systems with Applications*, 25(3):343–354, 2003.

[10] S. Doroudi, K. Holstein, V. Aleven, and E. Brunskill. Sequence matters, but how exactly? A method for evaluating activity sequences from data. In *Proc. 9th IEDMS, 2016*, pages 70–77, 2016.

[11] S. Fatahi, F. Shabanali-Fami, and H. Moradi. An empirical study of using sequential behavior pattern mining approach to predict learning styles. *Education and Information Technologies*, 23(4):1427–1445, 2018.

[12] P. Fournier-Viger, J. C. Lin, A. Gomariz, T. Gueniche, A. Soltani, Z. Deng, and H. T. Lam. The SPMF open-source data mining library version 2. In *Proc. Int. Conf. ECML PKDD, Riva del Garda, Italy*, volume 9853, pages 36–40. Springer, 2016.

[13] D. Fricker, H. Zhang, and C. Yu. Sequential pattern mining of multimodal data streams in dyadic interactions. In *1st ICDL-EPIROB,*, pages 1–6, 2011.

[14] C. Gao. Prefixspan-py, 2015-2020.

[15] F. Giannotti, M. Nanni, D. Pedreschi, and F. Pinelli. Mining sequences with temporal annotations. In *Proc ACM SAC, Dijon, France*, pages 593–597, 2006.

[16] S. Gutierrez-Santos, M. Mavrikis, and G. Magoulas. Sequence detection for adaptive feedback generation in an exploratory environment for mathematical generalisation. In *Int. Conf. on AIMSA*, pages 181–190. Springer, 2010.

[17] T. Guyet and R. Quiniou. Mining temporal patterns with quantitative intervals. In *Proc. Int. Conf. on Data Mining Workshops*, pages 218–227, 2008.

[18] Y. Hirate and H. Yamana. Generalized sequential pattern mining with item intervals. *JCP*, 1(3):51–60, 2006.

[19] Y. Hu, T. C. Huang, H. Yang, and Y. Chen. On mining multi-time-interval sequential patterns. *Data Knowledge Engineering*, 68(10):1112–1127, 2009.

[20] J. Kay, N. Maisonneuve, K. Yacef, and O. Zaïane. Mining patterns of events in students' teamwork data. In *Proc. Workshop on EDM at the 8th Int. Conf. on Intelligent Tutoring Systems*, pages 45–52, 2006.

[21] J. S. Kinnebrew, K. M. Loretz, and G. Biswas. A contextualized, differential sequence mining method to derive students' learning behavior patterns. *J. of Educational Data Mining*, 5(1):190–219, 2013.

[22] H. Kitakami, T. Kanbara, Y. Mori, S. Kuroki, and Y. Yamazaki. Modified prefixspan method for motif discovery in sequence databases. In *Proc. Pacific Rim International Conference on Artificial Intelligence*, pages 482–491. Springer, 2002.

[23] C. Krauss, A. Merceron, and S. Arbanowski. The timeliness deviation: A novel approach to evaluate educational recommender systems for closed-courses. In *Proc. 9th Int. Conf. on LAK, Tempe, USA*, pages 195–204, 2019.

[24] S. Mahajan and A. Reshamwala. An approach to optimize fuzzy time-interval sequential patterns using multi-objective genetic algorithm. In *Technology systems and management*, pages 115–120. Springer, 2011.

[25] R. Martínez Maldonado, K. Yacef, J. Kay, A. Kharrufa, and A. Al-Qaraghuli. Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop. In *Proc. 4th Int. Conf. on Educational Data Mining, Eindhoven, The Netherlands*, pages 111–120, 2011.

[26] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proc. 17th Int. Conf. on Data Engineering*, pages 215–224, 2001.

[27] D. Perera, J. Kay, I. Koprinska, K. Yacef, and O. R. Zaïane. Clustering and sequential pattern mining of online collaborative learning data. *IEEE Trans. Knowl. Data Eng.*, 21(6):759–772, 2009.

[28] L. K. Poon, S.-C. Kong, M. Y. Wong, and T. S. Yau. Mining sequential patterns of students' access on learning management system. In *Int. conf. on data mining and big data*, pages 191–198. Springer, 2017.

[29] P. Rashidi and D. J. Cook. Keeping the resident in the loop: Adapting the smart home to the user. *IEEE Transactions on systems, man, and cybernetics-part A: systems and humans*, 39(5):949–959, 2009.

[30] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1):135–146, 2007.

[31] I. Sato, Y. Hirate, and H. Yamana. Text mining using prefixspan constrained by item interval and item attribute. In *Proc. 22nd ICDE, Atlanta, USA*, page 118, 2006.

[32] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *Int. Conf. on Extending Database*

Technology, pages 1–17. Springer, 1996.

[33] S. Sumalatha and R. Subramanyam. Distributed mining of high utility time interval sequential patterns using mapreduce approach. *Expert Systems with Applications*, 141:112967, 2020.

[34] J. K. Tarus, Z. Niu, and A. Yousif. A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining. *Future Generation Computer Systems*, 72, 2017.

[35] A. Vautier, M. Cordier, and R. Quiniou. An inductive database for mining temporal patterns in event sequences. In *International Joint Conference On Artificial Intelligence*, pages 1640–1641, 2005.

[36] M. Yoshida, T. Iizuka, H. Shiohara, and M. Ishiguro. Mining sequential patterns including time intervals. In *Data Mining and Knowledge Discovery: Theory, Tools, and Technology II, Orlando, USA*, volume 4057, pages 213–220, 2000.

[37] M. J. Zaki. SPADE: an efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1/2):31–60, 2001.

[38] G. Zhu, W. Xing, and V. Popov. Uncovering the sequential patterns in transformative and non-transformative discourse during collaborative inquiry learning. *The Internet and Higher Education*, 2019.