Interpreting Data Mining Results with Linked Data for Learning Analytics: Motivation, Case Study and Directions

Mathieu d'Aquin
Knowledge Media Institute, The Open University
Walton Hall, Milton Keynes, MK7 6AA, UK
mathieu.daquin@open.ac.uk

Nicolas Jay
Université de Lorraine, LORIA, UMR 7503
Vandœvre-lès-Nancy, F-54506, France
nicolas.jay@loria.fr

ABSTRACT

Learning Analytics by nature relies on computational information processing activities intended to extract from raw data some interesting aspects that can be used to obtain insights into the behaviours of learners, the design of learning experiences, etc. There is a large variety of computational techniques that can be employed, all with interesting properties, but it is the interpretation of their results that really forms the core of the analytics process. In this paper, we look at a specific data mining method, namely sequential pattern extraction, and we demonstrate an approach that exploits available linked open data for this interpretation task. Indeed, we show through a case study relying on data about students' enrolment in course modules how linked data can be used to provide a variety of additional dimensions through which the results of the data mining method can be explored, providing, at interpretation time, new input into the analytics process.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: Applications; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Design, Experimentation

Keywords

leaning analytics, course enrolment, data mining, sequence mining, linked data, interpretation

1. INTRODUCTION

The most commonly found definition of *Learning Analytics* is given in [9] (citing the LAK 2011 call for paper) as:

> *"the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs."*

Such tasks naturally require computational tools and techniques in order to process the raw data obtained from educational systems or through other data collection method, to support the overall analysis of these data and the generation of exploitable insights from them by an expert, a teacher or directly by the learner.

Here, we are especially interested in the use of data mining to support learning analytics. The use of data mining methods on education-related data is generally studied in the area of *Educational Data Mining* (see [15] for a comparison of the two fields of Educational Data Mining and Learning Analytics). We however do not focus in this paper on the algorithmic techniques used on educational datasets, but on the aspect of using data mining which is specially relevant to Learning Analytics (as a process centred on human understanding and decision making): the interpretation of the results of data mining.

Indeed, a data mining method can extract from raw data *patterns* of interest to the application domain (because, for example, of their frequency). While these patterns are useful as the starting point of an analytics process, the challenge here is to navigate and explore these patterns in order to come up with a meaningful analysis: an interpretation or a model that can explain the patterns and be used to exploit them as useful insights for decision making. This interpretation process cannot however be purely based on the raw data and the patterns extracted from them. Indeed, in order to achieve such an understanding of the results of a data mining method, the person conducting the analysis (which we call here the *analyst*) needs to bring into the process additional information about the domain being analysed. This is especially challenging in a Learning Analytics scenario since, first, the analytics process does not pre-suppose any particular dimension in the data to be more significant than any other. In other terms, Learning Analytics is much more related to exploratory analysis than hypothesis testing or model validation. It is only once the patterns have been extracted that the analyst can explore what aspects of the domain might be relevant to interpret them, and therefore, what external information to bring into the process. Second, because of the nature of the education domain, there are a potentially infinite number of such dimensions that can explored for interpretation (from the subject of the courses to elements of the learning environment and context). Understanding and identifying what analytical dimensions to bring into the analytics process therefore cannot be done a priori, but requires an interactive process where *views* over the results of the data mining method can be created out of selected dimensions at runtime.

In this paper, we present a method that exploit external information available as *linked open data* to support the interpretation of data mining results, through automatically building a navigation/exploration structure in the results of a particular type of data mining tool (namely, sequence mining) based on *data dimensions* chosen by the analyst. We demonstrate this method through a use case based on data about students' enrolment in course modules. Through the results we obtain in this use case, we show the need and advantages of combining data mining with linked data for Learning Analytics, as providing both a way to automatically identify patterns in raw data, and to support the human interpretation and exploitation of these results through customisable views based on analytical dimensions found in external data.

We further discuss how this approach provides an initial instantiation of a more general approach to the combination of data mining and linked data (as promoted for example in [5]), and how such an approach is generally relevant to Learning Analytics.

2. MINING ENROLMENT DATA

In order to illustrate the need and general idea for linked data-based interpretation of data mining results, we rely on a concrete scenario: the analysis of student enrolment data. For many years, students could freely choose what course modules to take at the Open University[1]. There are however naturally a number of criteria they would use to choose what course to enrol to at what time: the level, number of credits, subjects of the course, how it would count towards a degree, etc. It is therefore generally interesting to look at concrete data about the enrolment of students into these modules, and to try to extract indicators of recurrent patterns of enrolment across time.

2.1 Data

The data we use is obtained from the Open University's "Course Profile" Facebook application[2], where students and prospective students can indicate what courses they are following, intend to follow or have followed in the past. The applications is useful to students as it helps identifying relevant resources to the courses of interests, as well as potential contacts following similar courses.

We use of snapshot of the database behind this application with anonymised identifiers, which contains 43,226 records on 8,806 students. Each record contains the information about the relationship between a student and a course ("currently studying", "completed" or "intend to study") at a certain date. For example, typical records are shown in Table 1.

Table 1: Example records from the student enrolment data.

Student ID	Course Code	Status	Date
112	dse212	Studying	2007
112	d315	Intend to study	2008
109	A207	Completed	2005

2.2 Mining Method: Sequential Pattern Mining

In order to analyse the type of data described above, i.e. students enrolment in course modules across time, *sequential pattern mining* [1] appears as a natural approach. It can be seen as an extension of the well known association rule problem, applied to data that can be modelled as sequences of itemsets, indexed for example by dates. It helps to discover rules such as: customers frequently buy DVDs of episodes I, II and III of Stars Wars, then buy within 6 months episodes IV, V, VI of the same famous epic space opera. Sequential pattern mining has been successfully used so far in various domains, including for example the analysis of patients' trajectories in hospitals (see [8]).

In order to apply sequential pattern mining to our enrolment data, we re-model the data as a set of *sequences* of *itemsets*. Itemsets contain *items* which in our case are the course codes mentioned in the data, and are ordered in sequences based on the dates at which the corresponding courses have been studied or completed by a given student. In other terms, each sequence represents the "trajectory" of a particular student in our data. An example of such a trajectory sequence is given below:

$$(\text{DD100}) \rightarrow (\text{D203}, \text{S180}) \rightarrow (\text{S283})$$

This means the student followed course DD100 the first year, courses D203 and S180 the second year, and course S283 the third year.

Once reformulated in this way, the data is used as input of a sequence mining tool: the Prefixspan algorithm [12]. The main parameter of the tool is the *support threshold*, which is the minimum number of sequences in which a pattern should appear to be considered frequent. In the considered example, we used 100 as a support threshold.

2.3 Results and why they need Interpretation

The result of the sequential pattern mining is a set of frequent sequential patterns, which are sequences that frequently occur in the set of sequences given as input of the process, as well as their respective *support*, which is the number of input sequences that include the pattern. Using a minimum support threshold of 100, we obtained from the data described above 126 different patterns, which include 1 to 3 itemsets each. Examples of some frequent patterns are given in Table 2.

Table 2: Examples of frequent sequential patterns.

Sequential Pattern	Support
(DD100) → (DSE212)	232
(DSE2012) → (ED209) → (DD303)	150
(B120) → (B201)	122

[1] http://www.open.ac.uk

[2] http://apps.facebook.com/courseprofiles/

We can clearly see here what would be the benefit of using such an approach if the goal was the automatic recommendation of relevant courses to certain students, or to give them options that draw on the most common trajectories (e.g., once you have done DD100, you might want to do DSE212; If you have done DSE212, ED209 and/or DD303 and/or some other frequent associations are also relevant). However, in a Learning Analytics scenario, where our goal is to better understand how students choose courses, what are the rationals behind these choices and how it could affect the design of the course modules and of proposed "standard" trajectories, these results raise more questions than they answer: What are the relationships between the topics of the course in common patterns? How do the steps relate to course levels? How are the patterns affected by the type of assessments used in each course? by the credits obtained for each course? etc.

Answering these questions, and more generally exploring the data mining results according the corresponding dimensions, requires background knowledge about all the different aspects of the items included in patterns. Here for example, to understand the second pattern in Table 2, it is useful to know that DSE212 is a level 2 course generally on psychology, and that ED209 is also a level 2 course but focusing more specifically on child development within psychology.

This illustrates our idea that interpreting the results of a data mining method within a Learning Analytics scenario requires to bring into the process external information about the various dimensions through which the items (here, the courses) included in extracted patterns can be explored. The difficulty is however that it is hard to anticipate in advance which of these dimensions will be relevant to the analyst at the time of interpretation, and more generally that we need a convenient way to integrate within these data mining results external information about a large variety of different perspectives on the analysed data, in order to leave sufficient flexibility to the analyst to explore and interpret the results according to different views. Such difficulties motivate our approach to the interpretation of data mining results through the integration with linked data. We detail this approach and the tools built to implement it in the next sections, and illustrate them on the course enrolment example presented above, using linked open data about course modules from the Open University's linked data platform[3].

3. INTERPRETING DATA MINING RESULTS WITH LINKED DATA

Linked Data [11] is a set of principles and technologies that rely on the architecture of the Web (URIs and links) to share, model and integrate data. The basic idea is that data objects (e.g., a book) are identified by web addresses (URIs), and the information attached to these objects are represented through links (themselves labeled with URIs) to values (e.g., the book title) or other URIs representing other objects (e.g., the author of the book). Besides this simple technological model, the main novelty introduced by linked data is this idea that raw data is represented and exposed directly on the Web, making the Web a collective data space connecting contributions from any possible sources. In our example, information about the book might be contributed by an organisation and information about the author by another.

This idea of Web-scale, global data integration has led to the principles of Linked Data being widely adopted especially by organisations taking advantage from the widespread dissemination of public information and open data. This includes in particular governments (see e.g., `http://data.gov.uk`), libraries (see e.g., `http://data.bnf.fr`) and, of course, educational institutions (with the Open University pioneering the use of Linked Data for education; see `http://data.open.ac.uk`, `http://linkeduniversities.org`, as well as [3]).

Considering this state of Linked Data, it seems therefore natural to investigate it as a source of additional information to support the interpretation of the results of data mining methods, such as the ones presented in the previous section. Below, we describe an approach using a linked data endpoint to collect descriptive dimensions about the items that constitute the extracted patterns, and to use these dimensions to automatically construct exploration/navigation structures into these results. Figure 1 gives an overview of this approach, which relies on a linked data-based description of the data mining results, on extracting from an external linked data source selected information about these items and on organising the extracted patterns in a hierarchy (a lattice constructed using *formal concept analysis* [17]) along the selected dimension.

3.1 Selecting a Dimension for the Exploration of the Data Mining Results

The data mining results we want to explore in our case are a set of *sequential patterns* (e.g., the ones presented in Table 2). A sequential pattern is an ordered list of *itemsets* (e.g., {DD100}, {D203, S180}), each of them being a set of items (e.g., DD100, D203, S180). In order to provide an exploration structure in these results, we need to support the analyst in selecting relevant dimensions on the items being included in sequential patterns. In other words, we want to use Linked Data to obtain simple descriptors of the items in the patterns, so that these patterns can be explored alongside these descriptors.

The first step in achieving this is to represent the results of data mining in a way compatible with a linked data representation, and that can be easily manipulated jointly with external linked data representations of the items being considered. This includes two aspects: 1- modelling the data mining results in RDF[4], in accordance with the principles of linked data and 2- ensuring that the items in the patterns are identified with URIs in reference to existing Linked Data sources.

The way we model sequential patterns in RDF/Linked Data is summarised in Figure 2. This representation is inspired by the "Sequence" Ontology Design Patterns[5] and is intended to be generic (i.e., it is independent from the course enrolment use case we are investigating here, and can be reused to represent any result of a sequential pattern mining process).

The second step of the representation is the one that connects the generic pattern mining representation described above with Linked Data-based external information about

[3] `http://data.open.ac.uk`, see [4, 19]

[4] `http://www.w3.org/RDF/`

[5] see `http://ontologydesignpatterns.org/wiki/Submissions:Sequence`

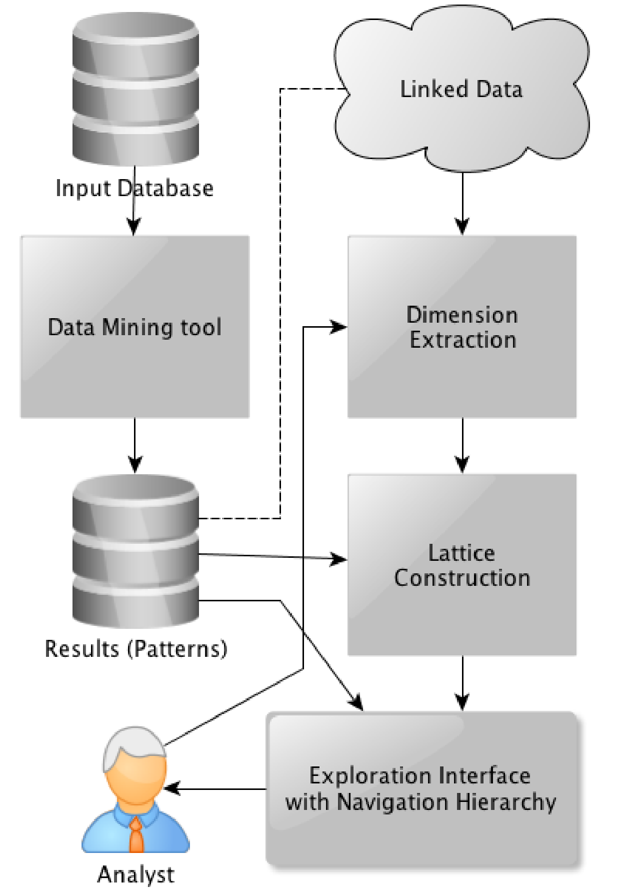


Figure 1: Overview of the approach to using Linked Data in interpreting the results of data mining.

the domain. Here, we will use information about the Open University's course catalogue, as available in `http://data.open.ac.uk`. In this Linked Data platform, a course such as DSE212 is associated with a URI of the form `http://data.open.ac.uk/course/dse212`. For courses that are currently available, `data.open.ac.uk` provides information related the subject of the course (here, for example, Psychology, represented by the URI `http://data.open.ac.uk/topic/psychology` as a *skos:Concept*), the course level, the number of credits, the modes of assessment, etc. For courses which are no longer available (e.g., DD100), the information provided includes links to the course material available at the University's library (e.g., set books), as well as any other resource that relate to the course (e.g., similar courses, units of open educational material, etc.) All this information is available as RDF through accessing the URI, or through a SPARQL [18] query endpoint (`http://data.open.ac.uk/query`).

The representation of sequential patterns connecting them to course items in `data.open.ac.uk` is loaded into a local triple store (here, we use Fuseki[6]), providing us with a SPARQL query endpoint for the extracted sequential patterns.

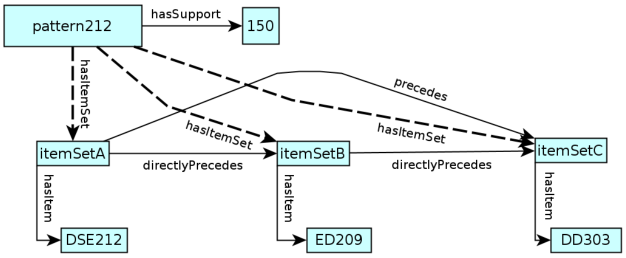


Figure 2: RDF Model of the pattern (DSE212) → (ED209) → (DD303) with support 150.

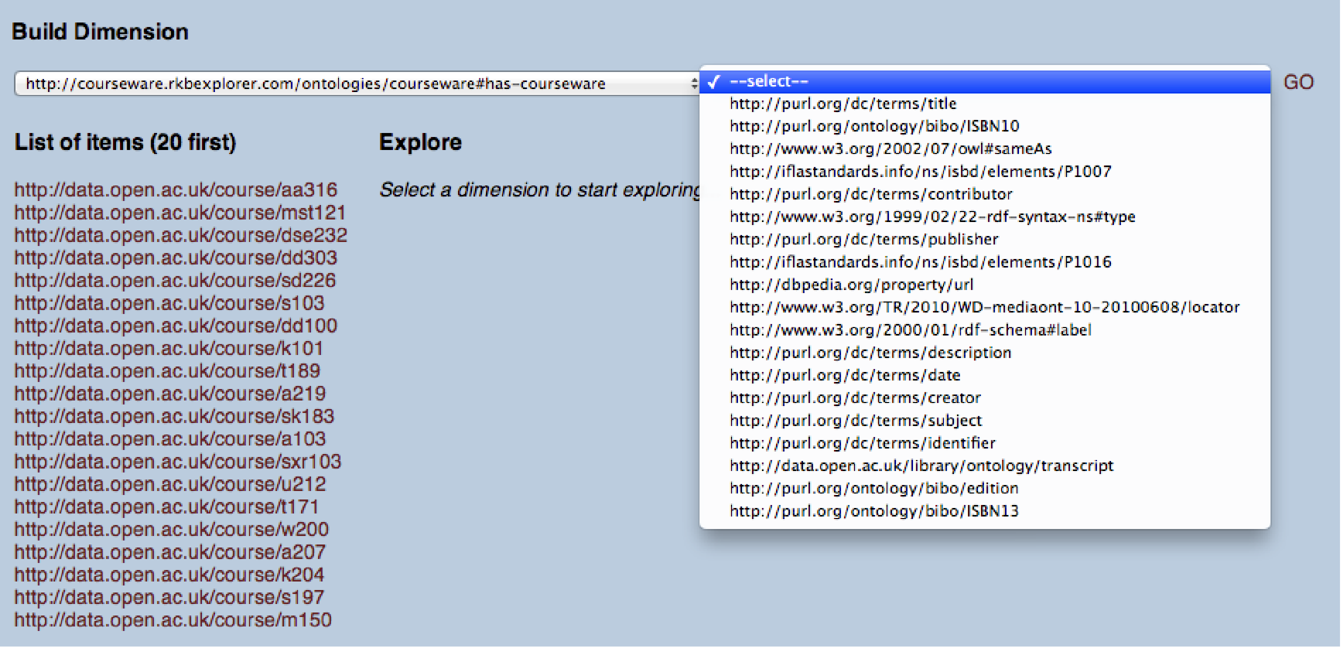
Besides making it convenient to manipulate the results of a data mining tool, the advantage of the representation above is that, following the principles of Linked Data, the use of existing URIs for the representation of the items make it possible to bring external information about them, from linked data sources that reference these URIs. As briefly discussed above, it is from these external information that we want to build additional dimensions to explore the mined patterns. Indeed, considering an existing data endpoint (in our case, `data.open.ac.uk`), the properties attached to the items represent as many descriptors that enrich the initial sequences used as input of the data mining process. This applies not only to properties that are directly attached to the items, but also indirectly to any path that can be built from them in the linked data graph (i.e., any property chain) starting with the items (e.g., the "labels of the subjects of the books used as course material for the selected course").

In order to help the analyst selecting the property chain he/she wants to apply as a dimension for exploration, we built an interface that allows him/her, given a specific linked data endpoint, to check what properties apply to typical items in the mined patterns (see Figure 3). It first allows the analyst to select a "representative item" amongst the ones present. A series of simple SPARQL queries are then used to find out, in the given linked data endpoint, what properties apply to the item, and subsequently, what properties apply to the values of these properties. In the example Figure 3, the item `http://data.open.ac.uk/course/aa316` has been selected, showing that, amongst others, the property `http://courseware.rkbexplorer.com/ontologies/courseware#has-courseware` applies to this item. It then shows that the values of these properties (books and other material used as part of the course) have a number of other properties that apply to them (including `http://purl.org/dc/terms/subject`: the subject of the considered resource) and that can be selected as part of the property chain used as a navigation dimension, as shown in the next section. For the interested reader, we show below the SPARQL query used to list these properties in `data.open.ac.uk`. If another property is selected, the properties of the values of this property are then shown.

```
SELECT distinct ?p where {
     <http://data.open.ac.uk/course/aa316}>
       <http://courseware.rkbexplorer.com/ontologies/courseware
          #has-courseware> ?v.
     ?v ?p []
  }
```

3.2 Lattice-based Classification of Mined Patterns

[6] `http://jena.apache.org/documentation/serving_data`


Build Dimension
GO
✓ --select--
http://purl.org/dc/terms/title
http://purl.org/ontology/bibo/ISBN10
http://www.w3.org/2002/07/owl#sameAs
http://iflastandards.info/ns/isbd/elements/P1007
http://purl.org/dc/terms/contributor
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://purl.org/dc/terms/publisher
http://iflastandards.info/ns/isbd/elements/P1016
http://dbpedia.org/property/url
http://www.w3.org/TR/2010/WD-mediaont-10-20100608/locator
http://www.w3.org/2000/01/rdf-schema#label
http://purl.org/dc/terms/description
http://purl.org/dc/terms/date
http://purl.org/dc/terms/creator
http://purl.org/dc/terms/subject
http://purl.org/dc/terms/identifier
http://data.open.ac.uk/library/ontology/transcript
http://purl.org/ontology/bibo/edition
http://purl.org/ontology/bibo/ISBN13

http://courseware.rkbexplorer.com/ontologies/courseware#has-courseware

List of items (20 first)

http://data.open.ac.uk/course/aa316
http://data.open.ac.uk/course/mst121
http://data.open.ac.uk/course/dse232
http://data.open.ac.uk/course/dd303
http://data.open.ac.uk/course/sd226
http://data.open.ac.uk/course/s103
http://data.open.ac.uk/course/dd100
http://data.open.ac.uk/course/k101
http://data.open.ac.uk/course/t189
http://data.open.ac.uk/course/a219
http://data.open.ac.uk/course/sk183
http://data.open.ac.uk/course/a103
http://data.open.ac.uk/course/sxr103
http://data.open.ac.uk/course/u212
http://data.open.ac.uk/course/t171
http://data.open.ac.uk/course/w200
http://data.open.ac.uk/course/a207
http://data.open.ac.uk/course/k204
http://data.open.ac.uk/course/s197
http://data.open.ac.uk/course/m150

Explore

Select a dimension to start exploring

Figure 3: Screenshot of the property chain selection tool to be used as exploration/navigation dimension in the data mining results.

The above method to select a chain of properties in a linked data source that contains information about items contained in the mined pattern has for purpose to enrich the patterns with an additional dimension for exploration, which was not present in the original data. In other terms, the values of the chain of properties that apply to the items of a pattern can be seen as a set of descriptors for the pattern. This enriched description of the patterns can be used to meaningfully organise them into a hierarchy, structured according to the chosen dimension.

To achieve this, we apply Formal Concept Analysis to the patterns and their set of new descriptors and build a concept (or Galois) lattice [16]. A *concept lattice* is a hierarchy of concepts (E, I), formed of an *extension* E (i.e., a set of objects) and an *intension* I (i.e., a set of attributes). Each concept groups together the objects (E) that have the same set of attributes (I). They are organised in a lattice according to relation $<$, which can be read as "is more specific than". For example, $(E_i, I_i) < (E_j, I_j)$ (the first concept is more specific than the second) means that $E_i \subset E_j$ and $I_j \subset I_i$.

Concept lattices are not only the product of a classification method, but are also often used as we intend to do here: to provide a navigation structure to an originally raw set of data objects. In our case, the objects are the sequential patterns mined in the data (such as the ones in Table 2), and the attributes are the values of the chain of properties obtained from Linked Data, from the items included in the patterns. For example, assuming we use the property chain [`http://purl.org/dc/terms/subject`, `http://www.w3.org/2000/01/rdf-schema#label`], the course B120 (`http://data.open.ac.uk/course/b120`) is represented by the attributes {Accounting and Finance, Business and Management}, and the course B201 (`http://data.open.ac.uk/course/b201`) by the attributes {Business and Management, Business Management Studies}. The third pattern in Table 2 is therefore represented by the set of attributes {Business and Management, Business Management Studies, Accounting and Finance}.

Building a concept lattice is a well known problem for which many tools exist. It is a computationally expensive task, requiring a lot of resources. However, after the sequential pattern mining phase, input data are already significantly reduced at this step (126 patterns in our example, out of the 8,806 original sequences). The number of attributes varies depending on the chosen property chain, but we believe that this approach is lightweight enough to be integrated within the learning analyst's workflow. This is somehow demonstrated through the "reasonable" response time we obtained from our naive javascript implementation of a concept lattice construction algorithm (a few seconds in the case of property chains with less than 100 possible values, to a couple of minutes for property chains with more than a thousand values).

Figure 4 shows an portion of the lattice built through using the previously mentioned property chain as the navigation dimension ([`http://purl.org/dc/terms/subject`, `http://www.w3.org/2000/01/rdf-schema#label`]). In this figure, the "subjects" are showed that represent the intension of the concept, together with the *support* of the concept, i.e., the number of objects (here, sequential patterns) in its extension.

3.3 Exploring Mined Patterns with the Lattice

Since the built lattice constitutes a hierarchy, it is natural to use it for navigating the data on which it was built. This

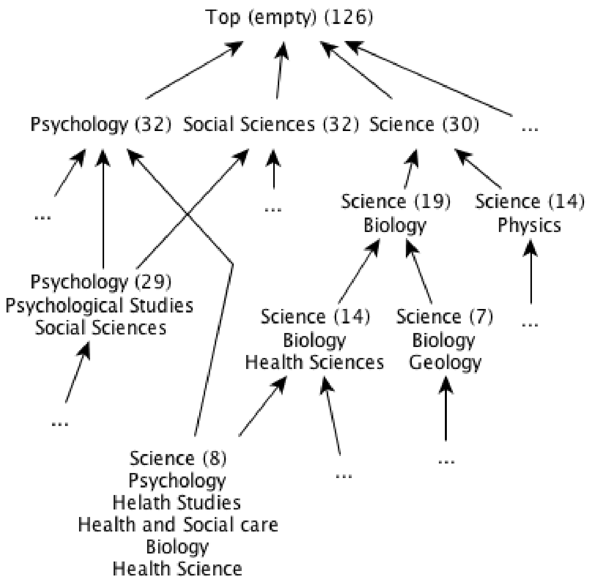

Top (empty) (126)
Psychology (32)
Social Sciences (32)
Science (30)
...
Science (19) Biology
Science (14) Physics
Psychology (29) Psychological Studies Social Sciences
Science (14) Biology Health Sciences
Science (7) Biology Geology
Science (8) Psychology Helath Studies Health and Social care Biology Health Science

Figure 4: Portion of the lattice built from the "subjects' labels" dimension.

has been explored before in several different domains [2]. Here however, we use the lattice to provide a further level of abstraction with respect to the original data: First, the sequential pattern mining method provides an additional structure over individual sequences of courses and second, the lattice provides a way to classify and explore these patterns according to dimensions brought through Linked Data.

Following this idea, Figure 5 shows the previously introduced interface with the navigation structure created. The chain of properties corresponding to the labels of subjects of courses has been selected here as the navigation dimension. The central part of the interface now shows an expandable hierarchy based on the built lattice, starting from the top concept (the one with an empty extension and all the patterns in the intension). Concepts which include subjects such as "Psychology", "Science" or "Law" are included at the first level, meaning that these concepts group sequential patterns of courses that cover these subjects, together with any number of other subjects. In the figure, the concepts "Science", "Social Sciences" and "Business and Management" have been expanded, showing concepts in which the corresponding subjects appear together with other subjects (e.g., "Biology", "Sociology"). The concept "Science | Biology" has also been expanded, showing two sub-concepts.

Next to each concept are indicators of their size/importance. The first number is the support of the concept (in terms of number of patterns). For example, there are 30 sequential patterns of courses that cover the subjects "Psychological Studies" and "Social Sciences". Next is the number of more general concepts, and the number of more specific concepts.

Using this basic hierarchy, we can already start investigating interesting elements of the extracted patterns. For example, it is very clear from the supports of the concepts at the first level that we have managed to extract much more frequent sequences in the areas of "Psychology", "Science" and "Social Sciences" than in other areas. The analyst here could therefore consider the hypothesis either that these particular topics might be transversal to many others and therefore included in a lot of frequent trajectories, or that we simply had (for unknown reasons) much more data regarding students enrolling into courses in these topics. Both hypothesis can be quickly verified by inspecting the corresponding patterns and the original data. Either ways, this represents an interesting finding, whether it is about the design of the student trajectories or about a previously unknown bias that should affect the analysis of these data.

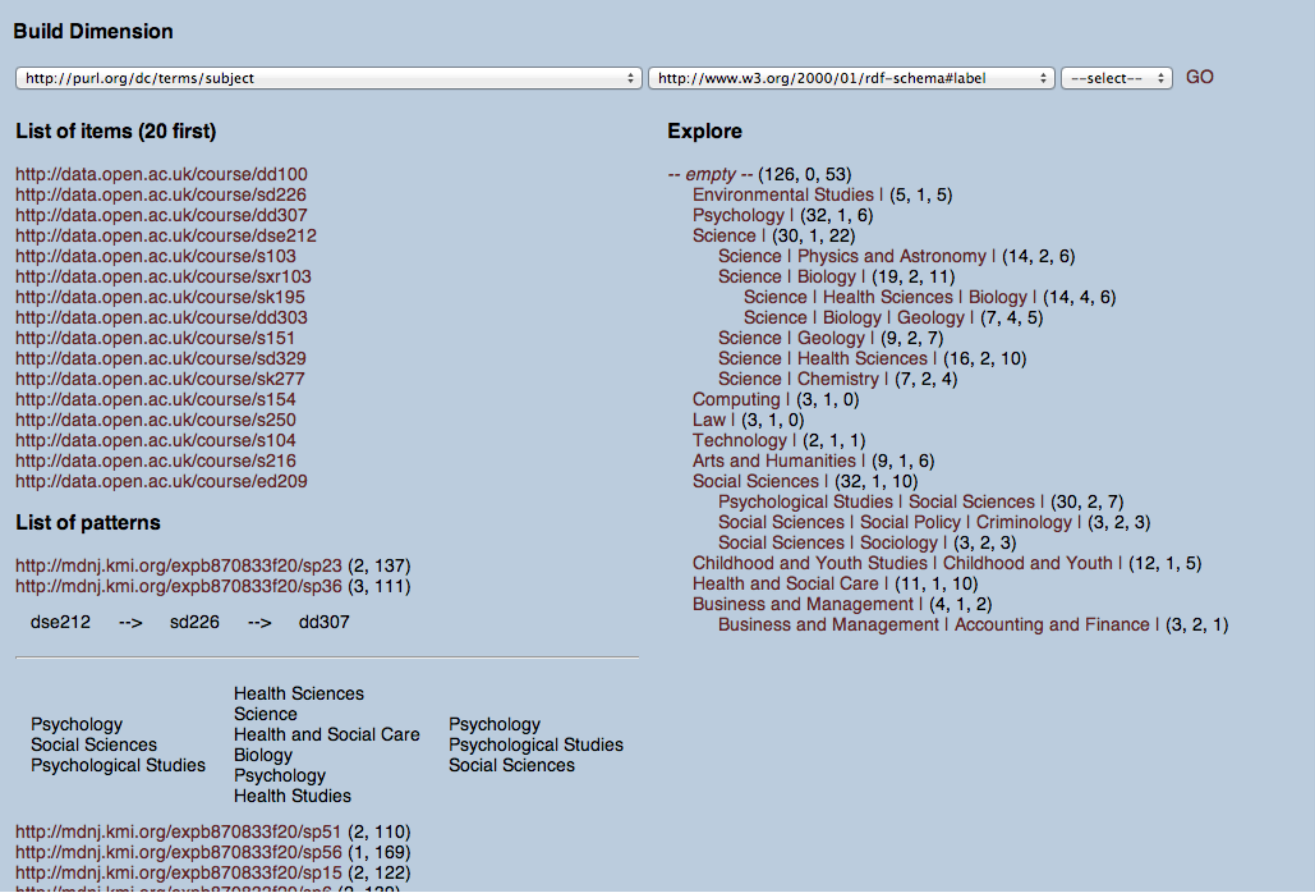
Similarly, we can see by exploring further the hierarchy that frequent sequential patterns in "Science" tend to branch into sub-topics with clear boundaries ("Physics", "Biology", "Chemistry", etc.) while "Social Sciences" quickly introduces elements from other disciplines. In particular, it appears that the "Social Sciences" branch shares a lot of patterns with the "Psychology" branch, explaining the high number of patterns in both.

Finally, each concept in the hierarchy can be selected, to show the details that relate to it. In Figure 5, the concept "Social Sciences" has been selected, showing on the left part of the interface the items (courses) that are mentioned in the corresponding patterns, as well as the patterns themselves (with their support and number of steps). The pattern sp36 (with three steps and corresponding to 111 student trajectories) has been expanded here, showing that it covers the courses DSE212, SD226 and DDS307, as well as the corresponding topics. This can allow the analyst to drill down into the details of a sub-set of the patterns, and see how it relates to the original data. It is important to notice here that the mention of subjects is due to the selection of the labels of subjects of courses as the exploration dimension, and the hierarchy would show different elements of the items if another dimension was to be selected (as discussed in the next section).

4. ADVANTAGES OF THE APPROACH

The method presented above is generic in the sense that it only requires the results of the sequential pattern mining method to be represented in accordance with the Linked Data principles, and some relevant external Linked Data sources providing information about the considered items to function. We focus here however on our case study regarding course enrolment, as it clearly demonstrates the benefits of the approach within a Learning Analytics workflow, in terms of supporting result interpretation, as well as the reuse of existing Linked Data sources.

The core benefit of the approach in our view is that, by relying on Linked Data, we give the analyst access to a large number of customisable views over the results that have been produced. Indeed, the examples above mostly focus on the subjects of courses, as it is a rather natural one to be used as a navigation dimension. However, any other characteristic of the courses could be used in exactly the same way. For example, we can quickly build a lattice that organises the courses based the course level, possibly to verify that there is no common patterns representing an unexpected type of trajectories, such as doing courses in the wrong order or courses of different levels in the same year, to see if we can identify sequential patterns that span over the whole cycle or to focus on the patterns covering a certain level. It is also interesting



Build Dimension

http://purl.org/dc/terms/subject | http://www.w3.org/2000/01/rdf-schema#label | --select-- | GO

List of items (20 first)

http://data.open.ac.uk/course/dd100
http://data.open.ac.uk/course/sd226
http://data.open.ac.uk/course/dd307
http://data.open.ac.uk/course/dse212
http://data.open.ac.uk/course/s103
http://data.open.ac.uk/course/sxr103
http://data.open.ac.uk/course/sk195
http://data.open.ac.uk/course/dd303
http://data.open.ac.uk/course/s151
http://data.open.ac.uk/course/sd329
http://data.open.ac.uk/course/sk277
http://data.open.ac.uk/course/s154
http://data.open.ac.uk/course/s250
http://data.open.ac.uk/course/s104
http://data.open.ac.uk/course/s216
http://data.open.ac.uk/course/ed209

List of patterns

http://mdnj.kmi.org/expb870833f20/sp23 (2, 137)
http://mdnj.kmi.org/expb870833f20/sp36 (3, 111)

dse212	-->	sd226	-->	dd307
Psychology Social Sciences Psychological Studies		Health Sciences Science Health and Social Care Biology Psychology Health Studies		Psychology Psychological Studies Social Sciences

http://mdnj.kmi.org/expb870833f20/sp51 (2, 110)
http://mdnj.kmi.org/expb870833f20/sp56 (1, 169)
http://mdnj.kmi.org/expb870833f20/sp15 (2, 122)
http://mdnj.kmi.org/expb870833f20/sp6 (2, 120)

Explore

- -- *empty* -- (126, 0, 53)
  - Environmental Studies | (5, 1, 5)
  - Psychology | (32, 1, 6)
  - Science | (30, 1, 22)
    - Science | Physics and Astronomy | (14, 2, 6)
    - Science | Biology | (19, 2, 11)
      - Science | Health Sciences | Biology | (14, 4, 6)
      - Science | Biology | Geology | (7, 4, 5)
    - Science | Geology | (9, 2, 7)
    - Science | Health Sciences | (16, 2, 10)
    - Science | Chemistry | (7, 2, 4)
  - Computing | (3, 1, 0)
  - Law | (3, 1, 0)
  - Technology | (2, 1, 1)
  - Arts and Humanities | (9, 1, 6)
  - Social Sciences | (32, 1, 10)
    - Psychological Studies | Social Sciences | (30, 2, 7)
    - Social Sciences | Social Policy | Criminology | (3, 2, 3)
    - Social Sciences | Sociology | (3, 2, 3)
  - Childhood and Youth Studies | Childhood and Youth | (12, 1, 5)
  - Health and Social Care | (11, 1, 10)
  - Business and Management | (4, 1, 2)
    - Business and Management | Accounting and Finance | (3, 2, 1)

Figure 5: Screenshot of the interface with the lattice built with the dimension "subjects' labels".

to check how patterns would combine different numbers of credits or different modes of assessments. To illustrate this, the lattice built for the assessment methods is represented in Figure 6. Here we can see that a large majority of the trajectories would include TMAs (tutor marked assignments), which is indeed the most common form of assessment at the Open University. We can also see that combinations that would seem to the analyst to be probably rare (such as Exam and EoCA – end-of-course assessment) still lead to a number of frequent patterns. This specific combination however, consistently with the intuition somebody with knowledge of Open University courses would have, does not appear in frequent patterns that do not also include TMAs.

Another important point here is that the approach makes it possible to identify relevant dimensions in the data at the time of interpretation. Indeed, it is possible to select any dimension after having produced the sequential patterns, and to compare how different dimensions produce different navigational structures in the results. These dimensions could of course be included originally in the data being mined, but besides the added difficulty that this would generate in terms of computational complexity for the data mining tool, it is not always possible to identify in advance what dimensions are relevant (including everything from the Linked Data source would of course be unfeasible). To give a concrete example, looking at the "subject" dimension considered before, an analyst might find it, after having ran the data mining process, not to be granular enough to allow for a meaningful analysis of frequent trajectories. Also, this dimension suffers from the issue that not all courses are associated with a subject (only the ones that are currently being taught). Having realised that, the analyst might cleverly change the dimension to use the subjects of the books that are associated with each course, providing a much more granular set of subjects available for almost every course (currently available or not). In this way for example, it would be possible to detect that several of the frequent trajectories do combine "Computer Science" with "Social Aspects", while the original subject-based lattice (see Figure 4) did not show a connection between computing and the social sciences. This idea that we can "try out" different exploration dimensions at the time of interpretation is especially powerful considering that, with Linked Data, we are not even limited to one particular source: one can easily switch from one data endpoint to the other to explore the different dimensions provided by different sources.

Finally, we argue that the approach presented here provides a suitable way of supporting the work of the (learning) analyst in interpreting the results of data mining, by providing him/her with a flexible tool to identify patterns or a sub-parts of the patterns that are especially relevant. This has been shown in some of the examples discussed above,

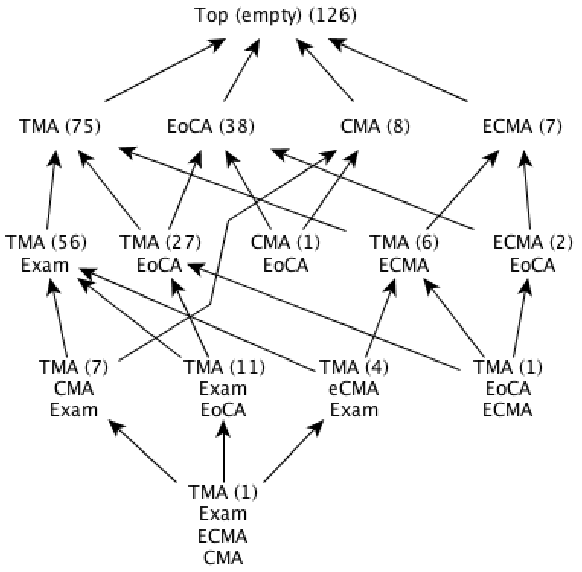


Figure 6: Lattice built on the dimension related to the mode of assessment associated with the courses. Exam means examination, TMA means tutor marked assignment, CMA mean computer marked assignment, ECMA means electronic computer marked assignment and EoCA means end-of-course assignment.

and understanding the concrete use of this tool in concrete analytics process would be outside the scope of this paper (although very interesting and one of the key points in our future work). However, we can already foresee three distinct ways in which this approach can help in the analysis:

1. *Providing an overview of the mined patterns:* by showing the relationships between the patterns, we can better understand how they distribute along a certain dimension, and how they relate to each other. This helps in understanding generally the results of the data mining process, as well as to quickly navigate to specific patterns of interest.

2. *Identify gaps and issues in the original data/process:* As shown for instance with the example related to course subjects, the approach can help (better than listing the found patterns) to identify a strong bias in the data. It would be difficult, without inspecting this specific dimension, to get any indicator that an exaggeratedly large portion of the patterns are about psychology, which, as it turns out, is due to more use of the Facebook application from which the data was collected by psychology students (for reasons that remain to be explained).

3. *Identify areas in need of further exploration:* In relation to the point above, it is also easy to see how the approach can help in identifying parts of the data that would require special attention. By part of the data, we mean either (or both) subsets of the original set of sequences that can be viewed as representing a consistent cluster (e.g., the ones about science), or (and) a specific dimension that would require further analysis.

Furthermore, considering the status of the approach as a preliminary step towards a more complete analytics methodology that would combine data mining and Linked Data, the last point above leads to the interesting idea that such a methodology can be seen as interactively iterating over the data mining interpretation processes. Indeed, starting with raw data, the analyst could obtain patterns through data mining and use the approach presented here to identify parts of the data as well as linked data-based dimensions requiring further, refined analysis. By re-injecting the selected data together with the selected dimension as input of the data mining process, refined patterns would be obtained that would directly integrate the dimension of interest, making it possible to re-run the full cycle with more and more refined data, as well as more and more refined patterns, under the control of the analyst.[7] Ultimately, combining in this way dimensions, sources of (linked data), as well as a larger variety of data mining techniques could lead to a powerful analytics environment, truly taking benefit from Linked Data.

5. RELATED WORK

The approach presented in this paper naturally relates to to the general field of *Educational Data Mining*, i.e., the application of data mining to traditional educational systems [13]. As discussed in [15], Learning Analytics is, in comparison, more concerned with the human-centric process of obtaining insight and input for decision making than with the algorithmic and technical aspects of the processing of data for extracting patterns or recommendations. In this sense, we see what we have proposed here as connecting to certain extent the two fields, being concerned with the use of data mining on educational data, while focusing on the interpretation of the results by a human analyst with a Learning Analytics purpose.

The other core aspect of the work presented in this paper relates to the use of Linked Data as part of a Learning Analytics process. Linked Data technologies seems to be naturally relevant to the Learning Analytics area, as illustrated for example by the mention of "Semantic Web and Linked Data" in the topics of the Learning Analytics and Knowledge conference's call for paper, as well as dedicated events such as the "Learning Analytics and Linked Data" workshop[8]. In these initiatives however, Linked Data is often seen as the base technology for the integration of data at the input of the learning analytics process rather than, as here, a way to enrich the results for the sake of interpretation.

We also investigate here one form of integration between data mining and linked data. Data mining has long been recognised as a potentially useful technique for achieving linked data, both as a way to extract more information from the raw available data regarding connections between various objects (see e.g., link mining [10]), as well as to support the extraction of useful information from the data (for example to provide human-friendly interfaces to linked datasets [6]). Mining with linked data sources is also an

[7] A similar, general idea was described in [5] regarding the connection between data mining and ontologies.

[8] `http://lald2012.wordpress.com/`

interesting area of research (see e.g., [14]), which remains surprisingly under investigated. Closer to the approach proposed here, another aspect concerns the combination of data mining and linked data where they both contribute to a more general knowledge discovery process (see [5] for a discussion on such a knowledge discovery methodology combining data mining and ontologies). We see the approach presented in this paper as a step towards such more general knowledge discovery cycles, especially relevant to the Learning Analytics process.

6. CONCLUSION

In this paper, we present an initial approach for exploring open Linked Data sources in interpreting the results of a data mining method, as part of a Learning Analytics process. We demonstrate on a use case relying on data about students' enrolment in course modules how results from a sequential pattern mining process can be automatically organised in a variety of dimensions, obtained from a connected Linked Data source. We also discuss the advantages of this approach in a Learning Analytics process, and how it constitutes an first step towards a methodology combining data mining and Linked Data for Learning Analytics.

Since we see this work as a first step, future work naturally revolve around closing the loop between mining and interpretation, especially through the ability to re-inject dimensions identified as relevant in Linked Data back into the data mining process, to refine this process and the obtained patterns. Also, while the case study presented in this paper provided us with valuable insight into the potential of this approach, and generally of Linked Data as a basis for the interpretation of analytics results, the concrete use of this approach in a variety of Learning Analytics scenarios is seen as a crucial next step of this work. Of course, while providing a more complete and valuable evaluation of the approach than the first case study presented in this paper, it will also require to address essential usability aspects of the tools. An interesting aspect of data mining, and more generally of knowledge discovery processes, which has not been investigate in this paper is the one of interaction with the analyst, who might not be familiar with the underlying (linked data and data mining) technologies. One approach we intend to investigate (in line with our previous work described in [6]) is to transform navigational elements, such as concepts of the lattice, into queries in pseudo-natural language that can be used to interrogate the generated sequential patterns.

Another assumption on which the idea of using linked data to interpret results in learning analytics relies is that sufficient linked data are available about the entities and concepts relevant to such interpretation. In the presented case study, linked data from the Open University (see [4]) was used. However, we can imagine scenarios in which results obtained would require additional information that might come from external, and possibly multiple sources. The availability and exploitability of linked data relevant to education-related applications is the one of the main aims of the LinkedUp support action (see `http://linkedup-project.eu`), which is in particular creating a large and widely available catalog of linked-datasets for education[9]. Through this and other preliminary initiatives[10], there is no doubt that large amounts of data will be available that could feed into our approach for data mining interpretation in Learning Analytics.

Naturally, this raises the additional issue of finding and identifying semantic data of use for interpretation, at the time of analysis. Combining the method presented here with the use of semantic web search engines (see e.g. [7]) to find such data could lead to a promising toolkit for interpretation in Learning Analytics.

Acknowledgments

This work is funded in part by the LinkedUp project (Grant Agreement: 317620) under the FP7 programme of the European Commission. See `http://linkedup-project.eu` for information.

We want thank Stuart Brown from the Open University's Communication Services for providing the students' enrolment data, supporting the authors in the exploitation of Linked Data and providing useful feedback on the process, approach and tool presented in this paper.

7. REFERENCES

[1] R. Agrawal and R. Srikant. Mining sequential patterns. In P. S. Yu and A. S. P. Chen, editors, *Eleventh International Conference on Data Engineering*, pages 3–14, Taipei, Taiwan, 1995. IEEE Computer Society Press.

[2] C. Carpineto and G. Romano. Using concept lattices for text retrieval and mining. In B. Ganter, G. Stumme, and R. Wille, editors, *Formal Concept Analysis*, volume 3626 of *Lecture Notes in Computer Science*, pages 161–179. Springer, 2005.

[3] M. d'Aquin. Linked Data for Open and Distance Learning. Commonwealth of Learning report – Available from `http://www.col.org/resources/publications/Pages/detail.aspx?PID=420`, 2012.

[4] M. d'Aquin. Putting linked data to use in a large higher-education organisation. In *Proceedings of the Interacting with Linked Data (ILD) workshop at Extended Semantic Web Conference (ESWC)*, 2012.

[5] M. d'Aquin, G. Kronberger, and M. Suárez-Figueroa. Combining data mining and ontology engineering to enrich ontologies and linked data. In *Proceedings of the Knowledge Discovery and Data Mining Meet Linked Open Data (Know@LOD) at the Extended Semantic Web Conference (ESWC)*, 2012.

[6] M. d'Aquin and E. Motta. Extracting relevant questions to an rdf dataset using formal concept analysis. In *Proceedings of the The Sixth International Conference on Knowledge Capture - K-CAP 2011*, 2011.

[7] M. d'Aquin and E. Motta. Watson, more than a semantic web search engine. *Semantic Web Journal*, 2, 2011.

[8] E. Egho, N. Jay, C. Raïssi, and A. Napoli. A FCA-based analysis of sequential care trajectories. In A. Napoli and V. Vychodil, editors, *Proceedings of The Eighth International Conference on Concept*

[9] see `http://http://datahub.io/group/linked-education`

[10] see `http://linkeduniversities.org` and `http://linkededucation.org/`

Lattices and their Applications - CLA 2011, Nancy, France, Oct. 2011. INRIA Nancy Grand Est - LORIA.

[9] R. Ferguson. The state of learning analytics in 2012: A review and future challenges. Technical Report KMI-12-01, Knowledge Media Institute, The Open University, UK, 2012.

[10] L. Getoor. Link mining: A new data mining challenge. *ACM SIGKDD Explorations newsletter*, 2005.

[11] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, 2011.

[12] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. Prefixspan: Mining sequential patterns by prefix-projected growth. In D. Georgakopoulos and A. Buchmann, editors, *ICDE*, pages 215–224. IEEE Computer Society, 2001.

[13] C. Romero and S. Ventura. Educationaldatamining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33, 2007.

[14] H. Sheng, H. Chen, T. Yu, and Y. Feng. Linked data based semantic similarity and clustering. In *Proceedings of the IEEE international conference on Information Reuse and Integration (IRI)*, 2010.

[15] G. Siemens and R. S. J. Baker. Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the Learning Analytics and Knowledge Conference*, 2012.

[16] R. Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. In I. Rival, editor, *Ordered Sets*. Reidel, 1982.

[17] R. Wille. Why can concept lattices support knowledge discovery in databases? *J. Exp. Theor. Artif. Intell.*, 14(2-3):81–92, 2002.

[18] World Wide Web Consortium. SPARQL Query Language for RDF. Avaliable online from `http://www.w3.org/TR/rdf-sparql-query/`; last accessed 2012-08-02, 2008.

[19] F. Zablith, M. d'Aquin, S. Brown, and L. Green-Hughes. Consuming linked data within a large educational organization. In *Proceedings of the second International Workshop on Consuming Linked Data (COLD) at International Semantic Web Conference (ISWC)*, 2011.