A Review on Heterogeneous, Multi-source and Multi-dimensional data mining

Julie Bu Daher, Armelle Brun, Anne Boyer

A Review on Heterogeneous, Multi-source and Multi-dimensional data mining

Julie Bu Daher, Armelle Brun and Anne Boyer

January 15, 2018

This report discusses the general overview of the nature of data in our domain of research, the specific data that we have in METAL project and the related work concerning the types of data used in the domain of data mining and the associated algorithms

1 Introduction

Data mining is the discovery of interesting, unexpected or valuable structures in large datasets [11]. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use [5]. Data mining principles have been important for many years, and with the emergence of the domain of big data, it has become more important. Some key techniques of data mining are association (relation) when correlations are made between two or more items, classification that has a predefined set of groups or models and the values are predicted based on it, clustering that doesn't have any predefined groups, however the data itself defines the groups, prediction that is predicting some data based on previous and current data and sequential pattern mining that is a useful method for identifying regular occurrences of similar events [3].

Sequential pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence[14]. Discovering sequential patterns from a large database of sequences is an important problem in the field of knowledge discovery and data mining [10]. The problem is to discover subsequences, among a set of data sequences, that are frequent where the sequences containing them has a higher support than a user-specified minimum support [1]. Usually, sequence patterns are associated with different circumstances, and such circumstances form a multiple dimensional space. It is interesting and useful to mine sequential patterns associated with multi-dimensional information [17].

Section 2 discusses the general overview of the nature of data in our domain of research, section 3 describes the related work in this domain and section 4 describes the data of METAL project, our application domain.

2 Nature of the data in our domain of research

This section describes the general view of the nature of data in our domain of research. Data could be multidimensional, multi-source, multi-relational, background and linked data, complex data, and complex event sequences.

2.1 Multi-source, multidimensional and multi-relational data

2.1.1 Multi-source and multidimensional data

A source of information could provide data with different kinds, As discussed in [17, 23, 7, 19], different kinds of data are considered as different dimensions; thus a source of data provides one or more dimensions. Such kind of data is called multidimensional data.

In certain cases, data does not come from the same source of information; however, it comes from different sources and is gathered together in one dataset. such kind of data is called multi-source data. Data could be of the same kind or different kinds among different sources. Hence, each source of information could provide multidimensional data which makes the data complex and heterogeneous.

2.1.2 Multi-relational data

There could be relations between the dimensions that come from the same or different sources. Each dimension could have a relation between one or more other dimensions, the dimensions in this case are interrelated. This kind of data is called multi-relational data that can be represented in multi-relational databases as described in [6, 15]. Thus, multi-relational data mining is used for this kind of data. Multi-relational data mining approaches look for patterns that involve multiple tables (relations) from a relational database [6].

2.2 Background and linked data

Using background knowledge in the domain of frequent pattern mining can help discovering patterns, as well as finding completely new patterns that originate from combining the original data with additional background data [16]. Thus, adding background and linked data as additional information to the core data that already exists in the dataset helps in obtaining more efficient results or better explaining the results obtained. Additional data could be one or more dimensions from the multidimensional data, and thus it could be from one or more sources that are already existing or new.

2.3 Complex data, sequences and events

Complex datasets are data collections in which the individual data items are no longer “simple” (atomic in database terminology) values but are (semi-

)structured collections of data themselves[20].

A sequence is a flow of events occurring consecutively, where an event is either an item or an item-set (ordered or unordered) occurring at a certain time interval. A sequence is complex when the elements in each time-stamp are complex, which means that there are more than one item where there can be some characteristics between items such as order and other possible relationships between them[8]. A complex sequence could also be different events occurring simultaneously [22].

Complex events can be in the form of several events occurring (multi-variables) at one time slot in terms of various intervals (e.g., hours, days and weeks).[13]

There could be additional data coming from external sources attached to each event in a sequence. This data provides additional information about the items or item-sets. The data could be one or more dimensions and from one or more data sources.

2.4 Data privacy and ethics

Certain research domains require treating users' data which could contain some personal information about users; they are more specifically the domains that provides results that are personalized for each user. However, when dealing with such kind of data, certain measures of confidentiality and privacy should be taken into consideration because this data is subject to some privacy policies and regulations and should respect data ethics. As a result, when treating this kind of data, the real identity of the user is hidden and could not be identified, and this is done either by anonymization or pseudo-anonymization.

2.4.1 Data anonymization

When data is anonymized, the identity and the personal information of the user can't be identified. In this case, when there are sequential activities that are not performed during the same period of time such as sessions, days or specific intervals, the flow of activities of the user among the different periods can't be known as it is not possible to identify that these activities are performed by the same user.

2.4.2 Data pseudonymization

When data is pseudonymized, the identity of the user is replaced by "pseudonym" that is another identifier of the user different from his real identity and information, where additional information is required to re-identify the user. In the case of pseudonymization, the real identity of the user is not known, but the flow of activities performed by the same user across different periods of time can be identified based on the same pseudonym.

2.5 Mining process

As previously described. The nature of the data is considered as sequential, multidimensional, multi-relational, heterogeneous data and complex sequences. In order to treat such kind of data using data mining techniques, the data needs to be combined, structured and preprocessed so that it can be treated in efficient ways to obtain the best results.

Globally, data can be divided into sequential data that constitutes the sequences of events and multidimensional data that is all the data that could be attached to the sequence of events. Most of the studies performed in this domain are concerned with one-dimensional data. However, there exist some research studies that are concerned with d >2 dimensional data. The mining process of such kind of data becomes more complicated.

Sequential data is usually mined through traditional sequential pattern mining algorithms to obtain frequent sequential patterns. Regarding the nature of data in our domain of research, sequential data and multidimensional data are strictly attached together. Mining sequential data separately without taking into consideration the multidimensional data attached to it would lead to inaccurate resulting frequent patterns. Thus, the multidimensional data attached to the items or to the event in the sequence should be taken into consideration accordingly.

Each event in a sequence could have additional descriptive information attached to it. In addition, each single item of an event could have specific information related to it. The attached information to the sequential items could be one or more dimensions from the multidimensional data.

Considering each dimension separately, a dimension is one element that has different attributes. For each sequential item, the relevant attributes of each dimension is attached to it. Figure 1 demonstrates the multi-source and multi-dimensional information attached to event sequences.

Dimensions could have relations among each other and relations among their different attributes. The relation among the different attributes of a single dimension could help in building a well-defined structure of these attributes together which consequently simplifies the mining process. If the attributes of one dimension have relations of higher to lower importance, more general to more specific, or certain attributes leading to others, thus they can be represented in the form of hierarchies of attributes for each dimension where the attributes are distributed among different levels of granularities

In this case, each dimension could be represented in the form of hierarchy ranging from more general to less general items. This representation is beneficial in the mining process. When there aren't enough specific data, the representation of the data in this form allows to replace it by more general data to be able to provide results even though with less accuracy.

It is also interesting to understand the relations between one or more different dimensions that affects the results.

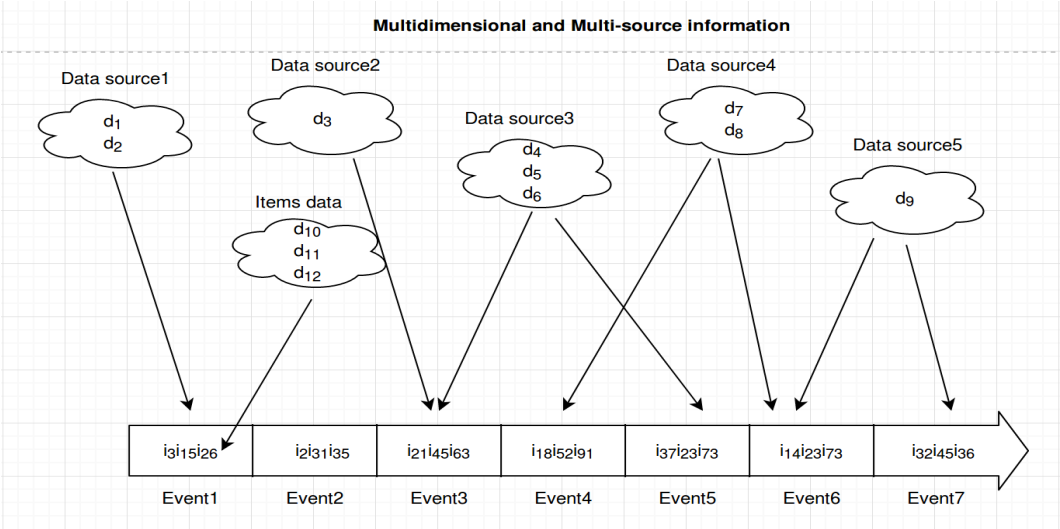

Figure 1: Data sources

3 Related Work

The mining of several data types has been studied in the literature, and many algorithms were designed. In this section, we draw a clear view of these types and the algorithms.

The approach of multidimensional sequential pattern mining was discussed in [17]. In their work, they proposed 3 different algorithms to mine such kind of data. The first algorithm, Uni-seq, treats all dimensional values as sequential items, and it finds all patterns using sequential pattern mining algorithm prefixspan. The second algorithm, Dim-seq, find frequent dimensional value combinations,and then mines sequential patterns from the set of sequences that satisfy each of these combinations. The third algorithm, Seq-dim, mines the sequential patterns for the whole dataset only once, and then mines the corresponding frequent dimensional patterns related to each sequential pattern. The following subsection provides some comparison measures among the three algorithms

3.0.1 Comparison among the three algorithms

Dimensionality When the dimensionality is low, Uni-seq outperforms the other 2 methods as the major cost is mining sequential information and the elements containing multidimensional information are short and prefixspan can handle them easily. However, when the dimensionality is high, the major cost is mining multidimensional information, so Seq-dim and Dim-seq are faster than Uni-seq because Uni-seq has to deal with longer sequences and patterns when many dimensional values are included.

Cardinality When the cardinality is high, the database becomes sparse; all methods have similar performance. However, when cardinality is low, the database becomes dense. Uni-seq and Dim-seq encounter difficulties in dealing with many patterns and Seq-Dim outperforms the other 2 algorithms. The main problem in Dim-seq algorithm is that it must explore all frequent multidimensional combinations before finding any sequential patterns even though some of them may not lead to any multidimensional sequential pattern. Seq-dim avoids these costs as it only explores multidimensional combination under the condition of some sequential patterns.

Scalability When database is large, there could be many frequent multidimensional values which lead to no multidimensional sequential patterns. That is why Dim-seq in this case has poor scalability while Seq-dim is the most scalable one.

The three algorithms proposed in [17] were found useful and were used by other related research works across time.

In Fournier et al., [9] sequential pattern mining is used to extract a partial problem space from logged user interactions, and it supports tutoring services during problem-solving exercises. They use the seq-dim approach proposed in Pinto et al.,[17] because original prefixspan algorithm doesn't consider the context of each sequence. They use Hirate-Yamana algorithm, which is a generalized sequential pattern mining algorithms that deals with item intervals; in other words a sequence with 1-day or 1-year interval), and they modify the algorithm to find only closed sequences; a frequent closed sequential pattern is a frequent sequential pattern such that it is not included in another sequential pattern having exactly the same support. For MD-Patterns mining, they applied the AprioriClose algorithm. They chose AprioriClose as it allows mining the set of closed MD-Patterns, and thus to eliminate some redundancy in the mined sequences.

In Hu et al. [12], they provide an algorithm to solve the problem of mining the multidimensional sequential patterns for large databases in the distributed environment. They use the Uni-seq approach provided by Pinto et al. [17] where multidimensional information is embedded into the corresponding sequences in order to convert the mining on the multidimensional sequential patterns to sequential patterns. Then the sequences are clustered, summarized and analyzed on the distributed sites, and the local patterns could be obtained by the effective approximate sequential pattern mining method. And then the global multidimensional sequential patterns could be mined by high vote sequential pattern model after collecting all the local sequential patterns on one site.

In Songram et al. [21], they discuss the closed multidimensional sequential pattern mining approach. They discuss that dim-seq and seq-dim algorithms provided by Pinto et al. [17] may generate large number of redundant patterns, that's why closed multidimensional sequential pattern mining was proposed to solve this problem. They make a combination of closed item-set pattern mining with a closed sequential pattern mining. The set of closed multidimensional

sequential patterns can cover the set of multidimensional sequential patterns. In our research domain, it would be interesting to keep all multidimensional patterns to have higher possibilities of matching students profiles as this step is followed by recommendations.

In Boonjing et al. [2], they improved the algorithms of Songram et al., [21] where all sequences that occur with mined multidimensional information are mined, or all the dimensional information that occur with the mined sequences are mined. Closed sequential pattern mining method must be called many times, and it leads to a large cost of mining. It also leads to large number of database scans being required because all sequences occurring with mined multidimensional information are scanned. Thus, they proposed two new algorithms of closed sequential patterns associated with lower-level closed item-set patterns or the inverse. They discuss that it is unnecessary to mine sequences at lower-levels of the branch. Although these algorithms can reduce computation time and the number of database scans, the redundant patterns still are generated from both combinations. However, these algorithms do not outperform existing algorithms without existing of super-set patterns. In our research domain, we are concerned with all the items of the sequence associated with the multidimensional information, or with all the multidimensional information associated with the sequence, so we interested in keeping all of this information.

In Plantevit et al. [18], they present a method of mining sequential patterns from multidimensional databases, at the same time taking advantage of the different dimensions and levels of granularity. The way they consider multidimensionality is generalized in the sense that patterns contain several dimensions combined over time. In their approach, they aim at building rules that combines two dimensions and also combines them over time and at different levels of granularity. They define a sequential pattern in their work as a sequence of sets of tuples defined over the analysis dimensions at different levels of granularity.

In Zhang et al. [24], they present the algorithm ApproxMGMSP (approximate mining of global multidimensional sequential sequential patterns) to solve the problem of mining multidimensional sequential patterns for large databases in the distributed environment. They convert mining of multidimensional sequential patterns to sequential patterns. The multidimensional information is embedded into the corresponding sequences. Then, the sequences are clustered, summarized and analyzed on the distributed sites, and the local patterns could be obtained by the effective approximate sequential pattern mining method. Then the global multidimensional sequential patterns could be quickly mined by high vote sequential pattern model after collecting all the local patterns on one site. They use Uniseq approach provided by Pinto et al. [17] where multidimensional information could be embedded into the corresponding sequence through introducing a special element.

In Buzmakov et al. [4], they focus on the analysis of complex sequential data by means of interesting sequential patterns. They use formal concept analysis approach and its extension based on pattern structures. Their aim is to develop a framework for enumerating only relevant patterns based on data lattices and its associated measures. They explain the usage of projections, which are mathe-

matical mappings for defining approximations. Projections for sequences allows the reduction of computational costs and decreasing the number of patterns and preserving the most interesting ones. They apply their method for discovering and analyzing interesting patient patterns from a French Health-care dataset on cancer. They describe Pinto's approach,[17] and they provide an example of multi-dimensional sequence patterns for health information of patients. They discuss that the dimensions in this approach remain constant, so it is not possible to have a pattern having two different values for one dimension like different cities, in which each one of them is attached to a sequence.

In Egho et al. [7], they proposed an approach called MMISP (mining multidimensional item-set sequential patterns) to extract pattern from complex sequences including both multi-dimensional items and item-sets. This approach incorporates background knowledge in the form of hierarchies over attributes. Their dataset is in the health-care domain where hospitalizations are defined as sequences of multi-dimensional attributes (eg hospital or diagnosis) associated with two sets, set of medical procedures and set of medical drugs. They objective is to mine health-care patients trajectories and give potential interesting patterns for health-care specialists.

Their approach is a bottom-up approach. It focuses on extracting frequent multidimensional items that can exist at different level of granularity, then it considers the item-sets parts of the events and compute the support of every item in the database for each item-set. Then, frequent multidimensional items and frequent item-sets are combined to generate events. Finally, the frequent events are mapped to a new representation and a standard sequential mining algorithm is applied to enumerate multidimensional item-set sequential patterns. First, they generate frequent multi-dimensional items, then frequent item-sets, then frequent events by combining frequent multi-dimensional components with frequent item-sets, and the last step is extracting frequent multidimensional item-sets pattern.

4 METAL project-Educational barometer

The general aim of METAL project is to improve the quality of learning process of students. The aim of our work is to design an educational barometer intended for school students that traces the digital activities of each student. Through this barometer, each student can know his academic level in details, and he will be provided with recommendations of additional pedagogic resources, as exercises, lectures and exams to increase his motivation and to help him improve his academic level and position in class. The recommendations will be calculated based on his pedagogic needs, certain preferences, previous digital traces of activity and his student profile.They be provided through pattern mining techniques that will be used to extract frequent patterns to be recommended to students in a personalized manner.

4.1 Data

This subsection describes the data of METAL project with its different sources and dimensions. The dataset contains information about students, courses, additional resources, teachers and schools.

- **Student Data**
  - Student's demographic information
    - Name
    - Age
    - Gender
    - Parents' social and marital statuses
  - Student's traces of activities performed on the dashboard
  - Class level
  - current academic year
  - grades and averages of previous years
  - Student's time schedule
  - Student's absences Student's engagement in the forum
- **Resources Data**
  - ID
  - Name
  - Type
  - Author
  - Description
  - Keywords
  - Level
- **Course Data**
  - ID
  - Name
  - Description
  - Academic year
  - Class level
  - Time schedule
  - Number of registered students
- **Teacher Data**
  - ID

- Name
- Set of courses taught by the teacher
- List of students taught by the teacher in a specific course

As previously detailed, the data is various and contains various attributes. The data comes from different schools in the region of Lorraine and thus from different sources. Each school has its own data about the students, programs, teachers and pedagogical resources. In France, all schools have the same educational program for the same class level; thus, data of different schools can be combined together when needed. For example we can recommend students from a certain school an additional academic resource provided by a teacher in another school.

In our dataset, we have different data for students, teachers, courses and pedagogical resources. Data is multi-relational as it is interlinked and can be distributed among multiple relational tables. The frequent patterns can then be discovered among multiple relational tables which are called relational patterns. In addition, data is multidimensional as we have different dimensions of data that come from various sources. Therefore, the challenge is to design a model that is able to treat this heterogeneous data taking into consideration the algorithm complexity.

5 Conclusion

The goal of this report was to discuss the nature of data in our domain of research as well as in our project and the related work concerning the types of data used in the domain of data mining and the associated algorithms.

References

[1] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pages 3–14. IEEE, 1995.

[2] Veera Boonjing and Panida Songram. Efficient algorithms for mining closed multidimensional sequential patterns. In *Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on*, volume 2, pages 749–753. IEEE, 2007.

[3] Martin Brown. Data mining techniques. *Retrieved from website: https://www. ibm. com/developerworks/library/badata-mining-techniques*, 2012.

[4] Aleksey Buzmakov, Elias Egho, Nicolas Jay, Sergei O Kuznetsov, Amedeo Napoli, and Chedy Raïssi. On mining complex sequential data by means of fca and pattern structures. *International Journal of General Systems*, 45(2):135–159, 2016.

[5] Soumen Chakrabarti, Martin Ester, Usama Fayyad, Johannes Gehrke, Jiawei Han, Shinichi Morishita, Gregory Piatetsky-Shapiro, and Wei Wang. Data mining curriculum: A proposal (version 1.0). *Intensive Working Group of ACM SIGKDD Curriculum Committee*, 140, 2006.

[6] Sašo Džeroski. Multi-relational data mining: an introduction. *ACM SIGKDD Explorations Newsletter*, 5(1):1–16, 2003.

[7] Elias Egho, Chedy Raïssi, Dino Ienco, Nicolas Jay, Amedeo Napoli, Pascal Poncelet, Catherine Quantin, and Maguelonne Teisseire. Healthcare trajectory mining by combining multidimensional component and itemsets. In *International Workshop on New Frontiers in Mining Complex Patterns*, pages 109–123. Springer, 2012.

[8] Lina Fahed, Armelle Brun, and Anne Boyer. Extraction de règles d'épisodes minimales dans des séquences complexes. In *EGC*, pages 545–548, 2014.

[9] Philippe Fournier-Viger, Roger Nkambou, and Engelbert Mephu Nguifo. A knowledge discovery framework for learning task models from user interactions in intelligent tutoring systems. In *Mexican International Conference on Artificial Intelligence*, pages 765–778. Springer, 2008.

[10] Minos N Garofalakis, Rajeev Rastogi, and Kyuseok Shim. Spirit: Sequential pattern mining with regular expression constraints. In *VLDB*, volume 99, pages 7–10, 1999.

[11] David J Hand. Principles of data mining. *Drug safety*, 30(7):621–622, 2007.

[12] Kong-Fa Hu, Chang-Hai Zhang, and Ling Chen. A scalable method of mining approximate multidimensional sequential patterns on distributed systemts. In *Machine Learning and Cybernetics, 2007 International Conference on*, volume 2, pages 762–766. IEEE, 2007.

[13] Kuo-Yu Huang and Chia-Hui Chang. Efficient mining of frequent episodes from complex sequences. *Information Systems*, 33(1):96–114, 2008.

[14] Nizar R Mabroukeh and Christie I Ezeife. A taxonomy of sequential pattern mining algorithms. *ACM Computing Surveys (CSUR)*, 43(1):3, 2010.

[15] Neelamadhab Padhy and Rasmita Panigrahi. Multi relational data mining approaches: A data mining technique. *arXiv preprint arXiv:1211.3871*, 2012.

[16] Heiko Paulheim. Exploiting linked open data as background knowledge in data mining. *DMoLD*, 1082, 2013.

[17] Helen Pinto, Jiawei Han, Jian Pei, Ke Wang, Qiming Chen, and Umeshwar Dayal. Multi-dimensional sequential pattern mining. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 81–88. ACM, 2001.

[18] Marc Plantevit, Anne Laurent, Dominique Laurent, Maguelonne Teisseire, and Yeow Wei Choong. Mining multidimensional and multilevel sequential patterns. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(1):4, 2010.

[19] Chedy Raïssi and Marc Plantevit. Mining multidimensional sequential patterns over data streams. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 263–272. Springer, 2008.

[20] Arno Siebes and Zbyszek Struzik. Complex data: Mining using patterns. In *Pattern Detection and Discovery*, pages 24–35. Springer, 2002.

[21] Panida Songram and Veera Boonjing. Closed multidimensional sequential pattern mining. *International Journal of Knowledge Management Studies*, 2(4):460–479, 2008.

[22] Cheng-Wei Wu, Yu-Feng Lin, Philip S Yu, and Vincent S Tseng. Mining high utility episodes in complex event sequences. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 536–544. ACM, 2013.

[23] Chung-Ching Yu and Yen-Liang Chen. Mining sequential patterns from multidimensional sequence data. *IEEE Transactions on Knowledge and Data Engineering*, 17(1):136–140, 2005.

[24] Changhai Zhang, Kongfa Hu, Zhuxi Chen, Ling Chen, and Yisheng Dong. Approxmgmsp: A scalable method of mining approximate multidimensional sequential patterns on distributed system. In *Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on*, volume 2, pages 730–734. IEEE, 2007.