A FCA-based analysis of sequential care trajectories

Elias EGHO, Nicolas Jay, Chedy Raissi and Amedeo Napoli

Orpailleur Team, LORIA, Vandoeuvre-les-Nancy, France
elias.egho,nicolas.jay,chedy.raissi,amedeo.napoli@loria.fr

Abstract. This paper presents a research work in the domains of sequential pattern mining and formal concept analysis. Using a combined method, we show how concept lattices and interestingness measures such as stability can improve the task of discovering knowledge in symbolic sequential data. We give example of a real medical application to illustrate how this approach can be useful to discover patterns of trajectories of care in a french medico-economical database.

Keywords: Data-Mining, Formal Concept Analysis, Sequential patterns, stability

1 Introduction

Sequential pattern mining, introduced by Agrawal et al [2], is a popular approach to discover patterns in ordered data. 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 first 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 : DNA sequencing, customer behavior, web mining ... [2].

Many scalable methods and algorithms have been published so far to efficiently mine sequential patterns. However few of them deal with the multidimensional aspect of databases. Multidimensionality conveys two notions:

- items can be of different intrinsic nature. While the common approach considers objects of the same dimension, for example articles bought by customers, databases can hold much more information such as article price, gender of the customer, location of the store and so on.
- a dimension can be considered at different levels of granularity. For example, apples in a basket market analysis can be either described as fruits, fresh food or food following a hierarchical taxonomy.

Plantevit et al. [13] address this problem as mining multidimensional and multilevel sequential patterns and propose a method to achieve this task. They rely on the support measure to efficiently discover relevant sequential patterns. Support

indicates to what extent a pattern is frequent in a database. Many (sequential and non sequential) itemset mining methods use support as measure for finding interesting correlations in databases. However, the most relevant patterns may not be the most frequent ones. Moreover, discovering interesting patterns with low support leads generally to overwhelming results that need to be further processed in order to be analyzed by human experts.

Formal Concept Analysis (FCA) is a theory of data analysis introduced in [17], that is tightly connected with data-mining and especially the search of frequent itemsets [16]. FCA organizes information into a concept lattice representing inherent structures existing in data. Recently, some authors proposed new interest measures to reduce complex concept lattices and thus find interesting patterns. In [9], Kuznetsov introduces stability, successfully used in social network and social community analysis [7, 6].

To our knowledge, there are no similar approaches to find interesting sequential patterns. In this paper, we present an original experiment based on both multilevel and multidimensional sequential patterns and lattice-based classification. This experiment may be regarded from two points of view: on the one hand, it is based on multilevel and multidimensional sequential patterns search, and on the other hand, visualization and classification of extracted sequences is based on Formal Concept Analysis (FCA) techniques, organizing them into a lattice for analysis and interpretation. It has been motivated by the problem of mining care trajectories in a regional healthcare system, using data from the PMSI, the so called French hospital information system. The remaining of the paper is organized as follows. In Section 2, we present the problem of mining care trajectories. Section 3 presents the methods proposed in domains of multilevel and multidimensional sequential patterns and Formal Concept Analysis. In Section 4, we present some of the results we achieved.

2 Mining healthcare trajectories

The PMSI (Programme de Médicalisation des Systèmes d' information) database is a national information system used in France to describe hospital activity with both an economical and medical point of view. The PMSI is based on the systematic collection of administrative and medical data. In this system, every hospitalization leads to the collection of administrative, demographical and medical data. This information is mainly used for billing and planning purposes. Its structure can be described (and voluntarily simplified) as follows:

- Entities (attributes):
  - Patients (id, gender ...)
  - Stays (id, hospital, principal diagnosis, ...)
  - Associated Diagnoses (id)
  - Procedures (id, date,...)
- Relationships
  - a patient has 1 or more stays
  - a stay may have several procedures

- a stay may have several associated diagnoses

The collection of data is done with a minimum recordset using controlled vocabularies and classifications. For example, all diagnoses are coded with the International Classification of Diseases (ICD10)[1]. Theses classifications can be used as taxonomies to feed the process of multilevel sequential pattern mining as shown in figure 1.

ICD 10 Institutions taxonomy

Fig. 1. Examples of taxonomies used in multilevel sequential pattern mining

Healthcare management and planning play a key role for improving the overall health level of the population. From a population point of view, even the best and state-of-the-art therapy is not effective if it cannot be delivered in the right conditions. Actually, many determinants affect the effective delivery of healthcare services: availability of trained personnel, availability of equipment, security constraints, costs, proximity .... All of these should meet economics, demographics, and epidemiological needs in a given area. This issue is especially acute in the field of cancer care where many institutions and professionals must cooperate to deliver high level, long term, and costly care. Therefore, it is crucial for healthcare managers and decision makers to be assisted by decision support systems that give strategic insights about the intrinsic behavior of the healthcare system.

On the one hand, healthcare systems can be considered as rich in data as they produce massive amounts of data such as electronic medical records, clinical trial data, hospital records, administrative data, and so on. On the other hand, they can be regarded as poor in knowledge as these data are rarely embedded into a strategic decision-support resource [1]. We used the PMSI system as a source of data to study patient movements between several institutions. By organizing themselves into groups of sequences representing trajectories of care, we aim at discovering patterns describing the whole course of treatments for a given population. This global approach contrasts with the usual statistical exploitations of the PMSI data that focus mainly on single hospitalizations.

In this experiment, we have worked on four years (2006 – 2009) of the PMSI data of the Burgundy region related to patient suffering from lung cancer.

[1] http://apps.who.int/classifications/apps/icd/icd10online/

3 Related work

3.1 Sequential Pattern Mining

Let $\mathtt{I}$ be a finite set of items. A subset of $\mathtt{I}$ is called an itemset. A sequence $\mathtt{s} = \langle \mathtt{s}_1\mathtt{s}_2 \dots \mathtt{s}_k \rangle$ ($\mathtt{s}_i \subseteq \mathtt{I}$) is an ordered list of itemsets. A sequence $\mathtt{s} = \langle \mathtt{s}_1\mathtt{s}_2 \dots \mathtt{s}_n \rangle$ is a subsequence of a sequence $\mathtt{s}' = \langle \mathtt{s}'_1\mathtt{s}'_2 \dots \mathtt{s}'_m \rangle$ if and only if $\exists \mathtt{i}_1, \mathtt{i}_2, \dots \mathtt{i}_n$ such that $\mathtt{i}_1 \leq \mathtt{i}_2 \leq \dots \leq \mathtt{i}_n$ and $\mathtt{s}_1 \subseteq \mathtt{s}'_{\mathtt{i}_1}, \mathtt{s}_2 \subseteq \mathtt{s}'_{\mathtt{i}_2} \dots \mathtt{a}_n \subseteq \mathtt{s}'_{\mathtt{i}_n}$. We note $\mathtt{s} \subseteq \mathtt{s}'$ and also say that $\mathtt{s}'$ contains $\mathtt{s}$. Let $\mathtt{D} = \{\mathtt{s}_1, \mathtt{s}_2 \dots \mathtt{s}_n\}$ be a database of sequences. The support of a sequence $\mathtt{s}$ in $\mathtt{D}$ is the proportion of sequences of $\mathtt{D}$ containing $\mathtt{s}$. Given a `minsup` threshold, the problem of frequent sequential pattern mining consists in finding the set $\mathtt{FS}$ of sequences whose support is not less than `minsup`. Following the seminal work of Agrawal and Srikant [2] and the Apriori algorithm, many studies have contributed to the efficient mining of sequential pattern. The main approaches are PrefixSpan [11], SPADE [20], SPAM [3], PSP [10], DISC [4] and PAID [18].

Much work has been done in the area of single-dimensional sequential patterns, i.e, all the items in a sequence have the same nature like the sequence of products sold in a certain store. But in many cases, the information in a sequence can be based on several dimensions. For example: a male patient had a surgical operation in Hospital A and then received chemotherapy in Hospital B. In this case, we have 3 dimensions: gender, type of treatment (chemotherapy, surgery) and location (Hospitals A and B). Pinto et al [12] is the first work giving solutions for mining multidimensional sequential patterns. They propose to include some dimensions in the first or the last itemset in the sequence. But this works only for dimensions that remain constant over time, such as gender in our previous example. Among other proposals addressed in this area, Yu et al [19] consider multidimensional sequential pattern mining in the web domain. In their approach, dimensions are pages, sessions and days. They present two algorithms AprioriMD and PrefixMDSpan by modifying the Apriori and PrefixSpan algorithms. Zhang et al [21] propose the mining of multidimensional sequential patterns in distributed system.

Moreover, each dimension can be represented by different levels of granularity, using a taxonomy which defines the hierarchical relations between items. Figure 2 shows an example of a diseases taxonomy. Including knowledge contained in the taxonomy leads to the problem of multilevel sequential pattern mining. Its interest resides in the capacity to extract more or less general/specific sequential patterns and overcome problems of excessive granularity and low support. For example, using the diseases taxonomy in Figure 2, sequences such as $\langle \mathtt{HeartDisease}, \mathtt{BrainDisease} >$ could be extracted while $\langle \mathtt{Arryth.}, \mathtt{BrainDisease} >$ and $\langle \mathtt{Myoc.Inf.}, \mathtt{BrainDisease} >$ may have a too low support.

Although Srikant and Agrawal [14] early introduced hierarchy management in the extraction of association rules and sequential patterns, their approach was not scalable in a multidimensional context. Han et al [5] proposed a method for mining multiple level association rules in large databases. But their approach could not extract patterns containing items from different levels in the taxonomy. Plantevit et al [13] proposed M3SP, a method taking both multilevel and multidimensional aspects into account. M3SP is able to find sequential patterns with the most appropriate level of granularity.

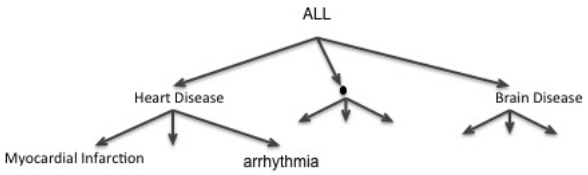

Fig. 2. disease's taxonomy

The PMSI is a multidimensional database holding information coded with controled vocabularies and taxonomies. Therefore, we relied on M3SP to extract multilevel and multidimensional sequential patterns. Nevertheless, the M3SP paradigm is still the search of frequent patterns. As our objective is to discover interesting patterns that may be infrequent, we ran M3SP iteratively until very low support thresholds. (See appendix for more details about M3SP and how we used it). This produced massive amounts of patterns requiring further processing for a practical interpretation by a domain expert. This next phase was conducted with a lattice-based classification of sequential patterns described in the following section.

3.2 Formal Concept Analysis

Introduced by Wille [17], Formal Concept Analysis is based on the mathematical order theory. FCA has successfully been applied to many fields, such as medicine and psychology, musicology, linguistic databases, information science, software engineering .... A strong feature of Formal Concept Analysis is its capability of producing graphical visualizations of the inherent structures among data.

FCA starts with a formal context $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ where $\mathtt{G}$ is a set of objects, $\mathtt{M}$ is a set of attributes, and the binary relation $\mathtt{I} = \mathtt{G} \times \mathtt{M}$ specifies which objects have which attributes. Two operators, both denoted by $'$, connect the power sets of objects $2^{\mathtt{G}}$ and attributes $2^{\mathtt{M}}$ as follows:

$$' : 2^{\mathtt{G}} \rightarrow 2^{\mathtt{M}}, \mathtt{X}' = \{\mathtt{m} \in \mathtt{M} | \forall \mathtt{g} \in \mathtt{X}, \mathtt{gIm}\}$$

$$' : 2^{\mathtt{M}} \rightarrow 2^{\mathtt{G}}, \mathtt{Y}' = \{\mathtt{g} \in \mathtt{G} | \forall \mathtt{m} \in \mathtt{Y}, \mathtt{gIm}\}$$

The operator $'$ is dually defined on attributes. The pair of $'$ operators induces a Galois connection between $2^{\mathtt{G}}$ and $2^{\mathtt{M}}$. The composition operators $''$ are closure operators: they are idempotent, extensive and monotonous. For any $\mathtt{A} \subseteq \mathtt{G}$ and $\mathtt{B} \subseteq \mathtt{M}$, A'' and B'' are closed sets whenever $\mathtt{A} = \mathtt{A}''$ and $\mathtt{B} = \mathtt{B}''$.

A formal concept of the context $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ is a pair $(\mathtt{A}, \mathtt{B}) \subseteq \mathtt{G} \times \mathtt{M}$ where $\mathtt{A}' = \mathtt{B}$ and $\mathtt{B}' = \mathtt{A}$. A is called the *extent* and $\mathtt{B}$ is called the *intent*. A concept $(\mathtt{A}_1, \mathtt{B}_1)$ is a *subconcept* of a concept $(\mathtt{A}_2, \mathtt{B}_2)$ if $\mathtt{A}_1 \subseteq \mathtt{A}_2$ (which is equivalent to $\mathtt{B}_2 \subseteq \mathtt{B}_1$) and we write $(\mathtt{A}_1, \mathtt{B}_1) \leq (\mathtt{A}_2, \mathtt{B}_2)$. The set $\mathfrak{B}$ of all concepts of a formal context $\mathbb{K}$ together with the partial order relation $\leq$ forms a lattice and is called concept lattice of $\mathbb{K}$. This lattice can be represented as a Hasse diagram providing a visual support for interpretation.

4 Classification and selection of interesting care trajectories

We use FCA to classify and filter the results of the sequential mining step. The formal context is built by taking patients as objects, and sequential patterns as attributes. A patient `p`, considered as a sequence, is related to a sequential pattern `s` if `p` contains `s`. Table 4 shows a formal context K_{PS} representing the binary relation between the patients and the sequences. The cross indicates that the patient has passed completely in the sequence of the health facilities. Thus, we achieve a classification of patients according to their trajectories of care.

	Seq_1	Seq_2	Seq_3	Seq_4
P_1	x	x	x	
P_2			x	
P_3		x	x	x
P_4		x		

Table 1.formal context K_{PS}

In order to choose the most important concepts, we rely on stability, a measure of interest introduced in [8] and revisited in [9].

Let $(\mathtt{A}, \mathtt{B})$ be a formal concept of $\mathfrak{B}$. Stability of $(\mathtt{A}, \mathtt{B})$ is defined as:

$$\gamma(\mathtt{A}, \mathtt{B}) = \frac{|\{\mathtt{C} \subseteq \mathtt{A} | \mathtt{C}' = \mathtt{A}' = \mathtt{B}\}|}{2^{|\mathtt{A}|}}$$

The stability index of a concept indicates how much the concept intent depends on particular objects of the extent. It indicates the probability of preserving concept intent while removing some objects of its extent. A stable concept continues to be a concept even if a few members stop being members. This means also that a stable concept is resistant to noise and will not collapse when some members are removed from its extent.

Stability offers an alternative point of view on concepts compared to the well known metric of support based on frequency, which is noticely used to build iceberg lattices [15]. Actually, combining support and stability allows a more subtle interpretation, as shown in a previous work in the same application domain [6].

5 Results

5.1 Patient healthcare trajectories

The PMSI is a relational database holding informations for any hospitalization in France. We reconstituted patient care trajectories from PMSI data considering each stay as an itemset. The sequence of stays for a same patient defines his care trajectory. In our experiment, itemsets could be made of various combinations of dimensions. Table 2 shows the trajectories of care obtained using two dimension (principal diagnosis, hospital ID). For example (C341,210780581) represents one hospitalization for a patient

Patient	Sequence
p1	$\langle(C341, 750712184)(Z452, 580780138)(D122, 030785430)\ldots\rangle$
p2	$\langle(C770, 100000017)(C770, 210780581)(Z080, 210780581)\ldots\rangle$
p3	$\langle(H259, 210780110)(H259, 210780110)(K804, 210010070)\ldots >$
p4	$\langle(R91, 210780136)(C07, 210780136)(C341, 210780136)\ldots\rangle$

Table 2. Care trajectories of 4 patients showing principal diagnoses and hospital IDs

in the University Hospital of Dijon (coded as 210780581) treated for a lung cancer (C341). Our dataset contained 486 patients suffering from lung cancer and living in the French region of Burgundy.

Table 3 shows some of the patterns generated by M3SP with the data presented in Table 2 using taxonomies of Figure 1. Pattern 3 can be interpreted as follows: 36% of patients have a hospitalization in a private institution (CL), for any kind of principal diagnosis (ALL). Then, 3 hospitalizations follow with the same principal diagnosis (Z511 coding for chemotherapy). That kind of pattern demonstrates the interest of multilevel and multidimensional sequential pattern mining: though principal diagnosis are the same in the third last stays, hospitals can be different. Mining at the lowest level of granularity, without taxonomies, would generate many different patterns with lower support.

ID	Support	Pattern
1	100%	$\langle(All, All)\rangle$
2	65%	$\langle(Z511, All)(Z511, All)(Z511, All)\rangle$
3	36%	$\langle(All, CL)(Z511, All)(Z511, All)(Z511, All)\rangle$
4	21%	$\langle(Z511, CH)(Z511, CH)(Z511, CH)(Z511, CH)(Z511, CH)\rangle$

Table 3. Example of sequential patterns generated by M3SP

However, for low support thresholds, the number of extracted patterns dramatically grows with the size of the database, depending on the number of patients, the size of the taxonomies and the number of dimensions as shown in Table 4.

Dimensions used	Number of patterns
`Institutions`	1529
`Principal Diagnosis, Institution`	4051
`All diagnoses`	50546
`Institutions, Medical Procedures`	293402

Table 4. Number of patterns generated by M3SP (`minsup`=5%)

The next step consists in building a lattice with the resulting sequential patterns in order to facilitate interpretation and selection of interesting care trajectories.

5.2 Lattice-based classification of sequential patterns

We illustrate this approach with patterns representing the sequences of institutions that are frequent in the patients set. We built a formal context relating 486 patients and 1529 sequential patterns. These sequences are generated in the first experimental by considering only one dimension (healthcare institutions). It is characterized with a taxonomy with two levels of granularity. We iteratively applied M3SP, decreasing threshold by one patient at each step. The resulting lattice has 10145 concepts organized on 48 different levels. Figure 3 shows the upper part of the lattice. Concepts intents are sets of one or more sequential patterns. From the lowest right concept, we can see that 37 patients support 3 sequential patterns:

- at least one hospitalization in the hospital 690781810
- they were hospitalized at least once in a University Hospital (CHU/CHR)
- they had at least 2 hospitalization, for simplicity, 2*(ALL) is the contraction of (ALL)(ALL).

The intent of top concept is $\langle\langle(ALL)\rangle\rangle$, because all patients have at least one hospitalization during their treatment. The intent of co-atoms (i.e. immediate descendant of top) is always a sequence of length one, holding items of high level of granularity.

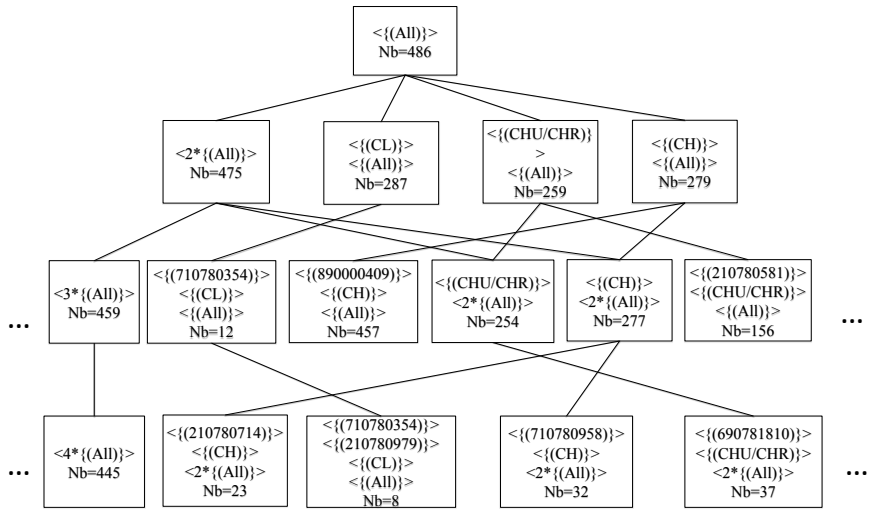

Fig. 3. Lattice of sequences of healthcare institutions

Filtering concepts can be achieved using both support and stability. In order to highlight the interesting properties of stability, we try to answer the question "is there a number of hospitalizations that characterizes care trajectories for lung cancer?". A basic scheme in lung cancer treatment consists generally in a sequence of 4 chemotherapy sessions possibly following a surgical operation. Due to noise in data or variability in

practices, we may observe sequences of 4, 5, 6 or more stays in the PMSI database. Mining such data with an *a priori* fixed support threshold may not discover the most interesting patterns. If the threshold is too high, we simply miss the good pattern. If it is too low, similar patterns, differing only in length, with close values of support can be extracted. Figure 4 shows the power of stability in discriminating such patterns. The concept with intent $\langle (CL)\rangle\langle 2*(ALL)\rangle$ is the most frequent. It represents patients with at least a stay in a private organization, and at least 2 stays in hospital. Similar concepts have a relatively close support, and differ only in the total number of stays. The concept with 5 stays has the highest stability. This probably matches the basic treatment scheme of lung cancer. Our interpretation relies on the power of stability to point out noisy concepts. Actually, only a few patients in concept $\langle (CL)\rangle\langle 2*(ALL)\rangle$ have only 2 stays.

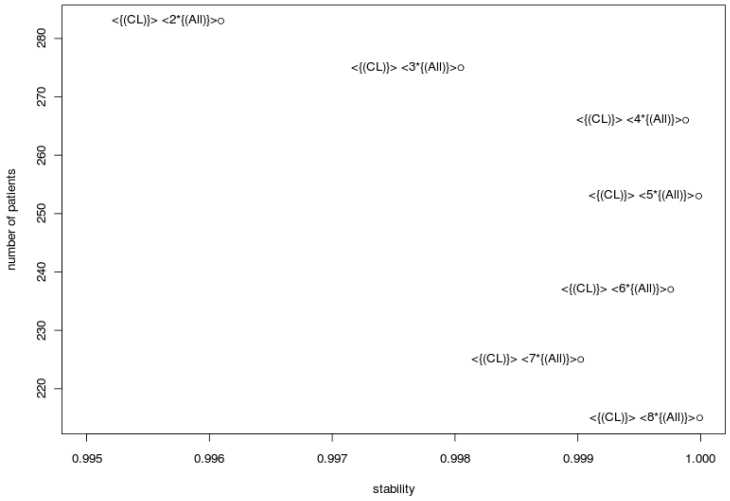

Fig. 4. Discriminating power of stability: scatter plot of support and stability of concepts (represented by their intent)

Another interesting feature of lattice-based classification of sequential patterns lies in its ability to characterize objects by several patterns. Let consider the minimal database of sequences $\mathtt{D} = \{\mathtt{s}_1 = \langle(\mathtt{a})(\mathtt{b})(\mathtt{c})\rangle; \mathtt{s}_2 = \langle(\mathtt{a})(\mathtt{c})(\mathtt{b})\rangle : \mathtt{s}_3 = \langle(\mathtt{d})\rangle\}$. With a 2/3 threshold, $\langle(\mathtt{a})(\mathtt{b})\rangle$ and $\langle(\mathtt{a})(\mathtt{c})\rangle$ are considered as frequent sequential patterns, but sequential pattern mining will give no information about the fact that all sequences containing the pattern $\langle(\mathtt{a})(\mathtt{b})\rangle$ contain also the pattern $\langle(\mathtt{a})(\mathtt{c})\rangle$. However this information can be obtained by classifying sequential patterns with FCA.

6 Conclusion

In this paper we propose an original combination of sequential pattern mining and FCA to explore a database of multidimensional sequences. We show some interesting properties of concept lattices and stability index to classify and select interesting sequential patterns. This work is in a early step. Further developments can be made in several axes. First, other measures of interest could be investigated to qualify sequential patterns. Furthermore, connexions between FCA and the sequential mining problem could be explored in a more integrative approach, especially by studying closure operators on sequences.

7 Acknowledgments

The authors wish to thank the TRAJCAN project for its financial support and Mrs. Catherine QUANTIN, the responsible of TRAJCAN project at university hospital of Dijon.

References

1. Abidi, S.S.: Knowledge management in healthcare: towards 'knowledge-driven' decision-support services. Int J Med Inform 63(1-2), 5–18 (Sep 2001)
2. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Yu, P.S., Chen, A.S.P. (eds.) Eleventh International Conference on Data Engineering. pp. 3–14. IEEE Computer Society Press, Taipei, Taiwan (1995), citeseer.ist.psu.edu/agrawal95mining.html
3. Ayres, J., Gehrke, J., Yiu, T., Flannick, J.: Sequential pattern mining using a bitmap representation. pp. 429–435. ACM Press (2002)
4. ying Chiu, D., hung Wu, Y., Chen, A.L.P.: An efficient algorithm for mining frequent sequences by a new strategy without support counting. In: In Proceedings of the 20th International Conference on Data Engineering (ICDE'04. pp. 375–386. IEEE Computer Society (2004)
5. Han, J., Fu, Y.: Mining multiple-level association rules in large databases. Knowledge and Data Engineering, IEEE Transactions on 11(5), 798 –805 (sep/oct 1999)
6. Jay, N., Kohler, F., Napoli, A.: Analysis of social communities with iceberg and stability-based concept lattices. In: Medina, R., Obiedkov, S.A. (eds.) International Conference on Formal Concept Analysis (ICFCA'08). LNAI, vol. 4923, pp. 258–272. Springer (2008)
7. Kuznetsov, S., Obiedkov, S., Roth, C.: Reducing the representation complexity of lattice-based taxonomies. In: Priss, U., Polovina, S., Hill, R. (eds.) Proc. of ICCS 15th Intl Conf Conceptual Structures. LNCS/LNAI, vol. 4604, pp. 241–254. Springer (2007)
8. Kuznetsov, S.O.: Stability as an estimate of the degree of substantiation of hypotheses derived on the basis of operational similarity. Nauchn. Tekh. Inf., Ser.2 (Automat. Document. Math. Linguist.) 12, 21–29 (1990)
9. Kuznetsov, S.O.: On stability of a formal concept. Annals of Mathematics and Artificial Intelligence 49, 101–115 (2007), http://www.springerlink.com/content/fk1414v361277475/

10. Masseglia, F., Cathala, F., Poncelet, P.: The psp approach for mining sequential patterns. pp. 176–184 (1998)
11. Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M.: Prefixspan: Mining sequential pattern by prefix-projected growth. In: ICDE. pp. 215–224 (2001)
12. Pinto, H., Han, J., Pei, J., Wang, K., Chen, Q., Dayal, U.: Multi-dimensional sequential pattern mining. In: CIKM '01: Proceedings of the tenth international conference on Information and knowledge management. pp. 81–88. ACM Press, New York, NY, USA (2001)
13. Plantevit, M., Laurent, A., Laurent, D., Teisseire, M., Choong, Y.W.: Mining multidimensional and multilevel sequential patterns. ACM Trans. Knowl. Discov. Data 4(1), 1–37 (2010)
14. Srikant, R., Agrawal, R.: Mining sequential patterns: Generalizations and performance improvements. In: Apers, P.M.G., Bouzeghoub, M., Gardarin, G. (eds.) Proc. 5th Int. Conf. Extending Database Technology, EDBT. vol. 1057, pp. 3–17. Springer-Verlag (25–29 1996), http://citeseer.ist.psu.edu/article/srikant96mining.html
15. Stumme, G.: Efficient data mining based on formal concept analysis. In: Lecture Notes in Computer Science, vol. 2453, p. 534. Springer (Jan 2002)
16. Valtchev, P., Missaoui, R., Godin, R.: Formal concept analysis for knowledge discovery and data mining: The new challenges. In: Eklund, P.W. (ed.) ICFCA. Lecture Notes in Computer Science, vol. 2961, pp. 352–371. Springer (2004)
17. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival, I. (ed.) Ordered Sets. Reidel (1982)
18. Yang, Z., Kitsuregawa, M., Wang, Y.: Paid: Mining sequential patterns by passed item deduction in large databases. In: IDEAS'06. pp. 113–120 (2006)
19. Yu, C.C., Chen, Y.L.: Mining sequential patterns from multidimensional sequence data. Knowledge and Data Engineering, IEEE Transactions on 17(1), 136 – 140 (jan 2005)
20. Zaki, M.J.: Spade: An efficient algorithm for mining frequent sequences. Machine Learning 42(1-2), 31–60 (January 2001), http://www.springerlink.com/link.asp?id=n3t642725v615427
21. Zhang, C., Hu, K., Chen, Z., Chen, L., Dong, Y.: 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. vol. 2, pp. 730 –734 (aug 2007)

Appendix

The M3SP algorithm is able to extract sequential patterns characterized by several dimensions with different levels of granularity for each dimension [13]. Each dimension has a taxonomy which defines the hierarchical relations between items. M3SP runs in three steps: data pre-processing , MAF-item generation and sequence mining.

In Figure 5, we present an example to illustrate the mechanism of M3SP. Table 5b shows a dataset of hospitalizations relating patients (P) with attributes from three dimensions

- T, the date of stay,
- H, the healthcare setting in which the hospitalization takes place,
- D, the disease of the patient.

For instance, the first tuple means that, at date 1, the patient 1 has been treated for the disease $\mathtt{D}_{11}$ in hospital $\mathtt{H}_{11}$. Let us now assume that we want to extract all multidimensional sequences that deal with hospitals and diseases that are frequent in the patients set. Figure 5a displays a taxonomy for dimensions H and D.

Pre-processing step

M3SP considers three types of dimensions: a temporal dimension $\mathtt{D}_\mathtt{t}$, a set of analysis dimensions $\mathtt{D}_\mathtt{A}$, and a set of reference dimensions $\mathtt{D}_\mathtt{R}$. M3SP orders the dataset according to $\mathtt{D}_\mathtt{t}$. The tuples appearing in a sequence are defined over the dimensions of $\mathtt{D}_\mathtt{A}$. The support of the sequences is computed according to dimensions of $\mathtt{D}_\mathtt{R}$. M3SP splits the dataset into blocks according to distinct tuple values over reference dimensions. The support of a given multidimensional sequence is the ratio of the number of blocks supporting the sequence over the total number of blocks. In our example, H (hospitals) and D (diseases) are the analysis dimensions, T is the temporal dimension and P (patients) is the only reference dimension. We obtain two blocks defined by $\mathtt{Patient}_1$ and $\mathtt{Patient}_2$. as shown in table 5c.

MAF-item generation step

In this step, M3SP generates all the Maximal Atomic Frequent items or MAF-items. In order to define MAF-items, we fisrt define the specificity relation between items.

Specificity relation. Given two multidimensional items $\mathtt{a} = (\mathtt{d}_1, ..., \mathtt{d}_\mathtt{m})$ and $\mathtt{a}' = (\mathtt{d}_1', ..., \mathtt{d}_\mathtt{m}')$, $\mathtt{a}'$ is said to be more specific than $\mathtt{a}$, denoted by $\mathtt{a} \preccurlyeq_\mathtt{I} \mathtt{a}'$, if for every $\mathtt{i} = 1, ..., \mathtt{m}$, $\mathtt{d}_\mathtt{i}' \in \mathtt{d}_\mathtt{i} \downarrow$. Where $\mathtt{d}_\mathtt{i} \downarrow$ is the set of all direct specializations of $\mathtt{d}_\mathtt{i}$ according to the dimension taxonomy of d_i. In our example, we have $(\mathtt{H}_1, \mathtt{D}_1) \preccurlyeq_\mathtt{I} (\mathtt{H}_1, \mathtt{D}_{11})$, because $\mathtt{H}_1 \in \mathtt{H}_1 \downarrow$ and $\mathtt{D}_1 \in \mathtt{D}_{11} \downarrow$.

MAF-item. An atomic item $\mathtt{a}$ is said to be a Maximal Atomic Frequent item, or a MAF-item, if $\mathtt{a}$ is frequent and if for every $\mathtt{a}'$ such that $\mathtt{a} \preccurlyeq_\mathtt{I} \mathtt{a}'$, the item $\mathtt{a}'$ is not frequent. In our example, if we consider $\mathtt{minsup} = 100\%$, $b = (H_1, D_1)$ is a MAF-item, because it is frequent and there is not another item as frequent and more specific than b.

The computation of MAF-items is represented by a tree in which the nodes are of the form $(\mathtt{d}_1, \mathtt{d}_2)_\mathtt{s}$, meaning that $(\mathtt{d}_1, \mathtt{d}_2)_s$ is an atomic item with support s as we

show in Figure 5d. In this tree, MAF-items are displayed as boxed nodes. We note that all leaves are not necessarily MAF-items. For example, $(\mathtt{H_2}, \mathtt{D_{21}})_{100\%}$ is a leaf, but not a MAF-item. This is because $(\mathtt{H_2}, \mathtt{D_{21}})_{100\%} \preccurlyeq_{\mathtt{I}} (\mathtt{H_{21}}, \mathtt{D_{21}})_{100\%}$ and $(\mathtt{H_{21}}, \mathtt{D_{21}})$ has been identified as being an MAF-item.

Sequence mining step

Frequent sequences can be mined using any standard sequential pattern-mining algorithm (PrefixSpan in this work). Since in such algorithms, the dataset to be mined is a set-pairs of the form (id, seq), where id is a sequence identifier and seq is a sequence of itemsets, our example dataset is transformed as follows :

- every MAF-item is associated with a unique identifier denoted by $ID(a)$ (table 5e), playing the role of the items in standard algorithms.
- every block b is assigned a patient identifier $ID(p)$, playing the role of the sequence identifiers in standard algorithms,
- every block b transformed into a pair $(ID(b), \zeta(b))$, where $\zeta(b)$ is a sequence. (table 5f)

PrefixSpan is run over table 5f. By considering a support threshold `minsup` =50%, table 5g displays all the frequent sequences in their transformed format as well in their multidimensional format in which identifiers are replaced with their actual values.

The basic step in M3SP method is MAF-item generation, because it provides all multidimensional items that occur in sequences to be mined. If the set of MAF-items is changed, the sequence will be changed. M3SP always extracts the most specific multidimensional items.

For example $(\mathtt{H_1}, \mathtt{D_1})$ is frequent according to `minsup`=50%, but another item, $(\mathtt{H_{11}}, \mathtt{D_{11}})$ is more specific and still frequent. As a result, $(\mathtt{H_1}, \mathtt{D_1})$ is not a MAF-item and consenquently not used to build squences. Finaly the frequent sequence $\langle\{(\mathtt{H_1}, \mathtt{D_1}), (\mathtt{H_{21}}, \mathtt{D_{21}})\}\rangle$ does not appear in the results of M3SP. However, tables 5 and 6 show the MAF-items set and the frequent sequences extracted by M3SP at a 100% threshold. It can be noticed that (H_1, D_1) is a MAF-item and that the sequence $\langle\{(\mathtt{H_1}, \mathtt{D_1}), (\mathtt{H_{21}}, \mathtt{D_{21}})\}\rangle$ is generated.

Thus, given two `minsup` thresholds $\sigma^{'} < \sigma$. The set of frequent sequences obtained for σ' may not always contain the set of sequences obtained for σ.

Considering this as a limit in our approach as we wanted to extract both general and specific sequences, we iteratively applied M3SP, decreasing threshold by one patient at each step. This allowed us to extract more potentially interesting sequences than by using a single low `minsup` threshold.

MAF-item
$(\mathtt{H_1}, \mathtt{D_1})$
$(\mathtt{H_{21}}, \mathtt{D_{21}})$

Table 5. maf-item,`minsup` =100%

Frequent Multidimensional Sequences
$\langle\{(\mathtt{H_1}, \mathtt{D_1})\}\rangle$
$\langle\{(\mathtt{H_{21}}, \mathtt{D_{21}})\}\rangle$
$\langle\{(\mathtt{H_1}, \mathtt{D_1}), (\mathtt{H_{21}}, \mathtt{D_{21}})\}\rangle$

Table 6. Sequences for `minsup` =100%

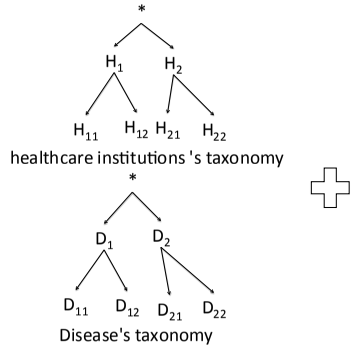

Figure 5(a). Dimensione 's taxonomy

P	T	H	D
1	1	H_{11}	D_{11}
1	2	H_{21}	D_{21}
2	1	H_{12}	D_{12}
2	2	H_{21}	D_{21}

Table 5(b). DataSet

Input

T	H	D
1	H_{11}	D_{11}
2	H_{21}	D_{21}

Patient$_1$

T	H	D
1	H_{12}	D_{12}
2	H_{21}	D_{21}

Patient$_2$

Figure 5(c). Block partition of Table I according to D_R = {Patient}.

I-Pre-processing

II-maf-item generation

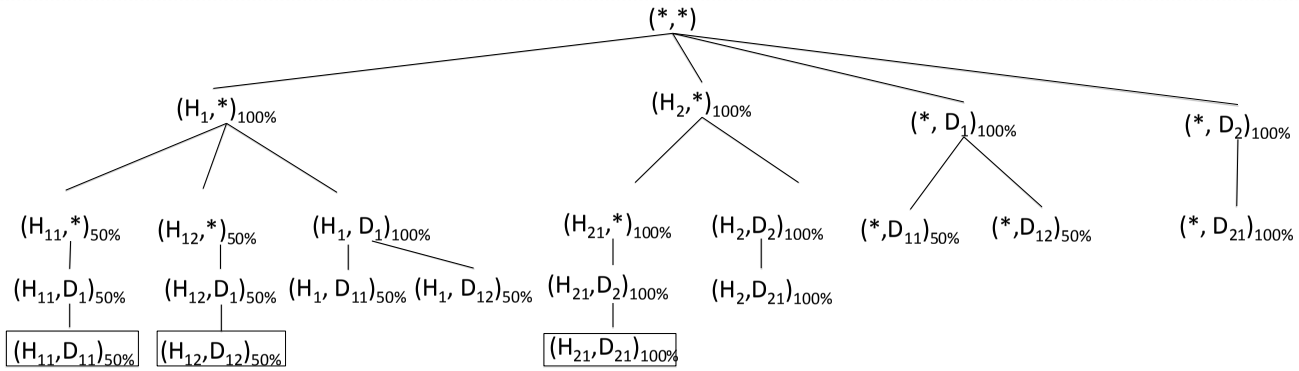

Figure 5(d). Tree of frequent atomic item

Maf-item	ID(a)
(H_{11},D_{11})	1
(H_{12},D_{12})	2
(H_{21},D_{21})	3

Table 5(e). Maf-item

ID(p)	ς (b)
1	<(1),(3)>
2	<(2),(3)>

Table 5(f). Transformed Database

PrefixSpan
minsup=50%

Frequent Multidimensional Sequences	
<(1) >	<{(H_{11},D_{11})}>
<(2) >	<{(H_{12},D_{12})}>
<(3)>	<{(H_{21},D_{21})}>
<(1),(3)>	<{(H_{11},D_{11}), (H_{21},D_{21})}>
<(2),(3)>	<{(H_{12},D_{12}), (H_{21},D_{21})}>

Table 5(g). Frequent Multidimensional Sequences

III-Sequences mining

Fig. 5. example for M3SP method, minsup =50%