

Healtcare Trajectory Mining by Combining Multi-dimensional Component and Itemsets

Elias Egho, Dino Ienco, Nicolas Jay, Amedeo Napoli, Pascal Poncelet, Catherine Quantin, Chedy Raïssi, Maguelonne Teisseire

▶ **To cite this version:**

Elias Egho, Dino Ienco, Nicolas Jay, Amedeo Napoli, Pascal Poncelet, et al.. Healtcare Trajectory Mining by Combining Multi-dimensional Component and Itemsets. NFMCP: New Frontiers in Mining Complex Patterns, Sep 2012, Bristol, United Kingdom. 10.1007/978-3-642-37382-4_8 . lirmm-00732661

HAL Id: lirmm-00732661

https://hal-lirmm.ccsd.cnrs.fr/lirmm-00732661

Submitted on 16 Sep 2012

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

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

Healthcare Trajectory Mining by Combining Multi-dimensional Component and Itemsets

Elias Egho[1], Dino Ienco[2,3], Nicolas Jay[1], Amedeo Napoli[1], Pascal Poncelet[2,3], Catherine Quantin[4], Chedy Raïssi[1] and Maguelonne Teisseire[2,3]

[1] Orpailleur Team, LORIA, Vandoeuvre-les-Nancy, France
{firstname.lastname}@loria.fr
[2] Irstea, Montpellier, France
{firstname.lastname}@teledetection.fr
[3] LIRMM, Univ. Montpellier 2, Montpellier, France
{firstname.lastname}@lirmm.fr
[4] Department of Biostatistics and Medical Information
CHU of Dijon, Dijon, France

Abstract. Sequential pattern mining is an approach to extract correlations among temporal data. Many different methods were proposed to either enumerate sequences of set valued data (i.e., itemsets) or sequences containing multidimensional items. However, in many real-world scenarios, data sequences are described as events of both multi-dimensional and set valued informations. These rich heterogeneous descriptions cannot be exploited by traditional approaches. For example, in healthcare domain, hospitalizations are defined as sequences of multi-dimensional attributes (e.g. Hospital or Diagnosis) associated with sets of medical procedures (e.g. { Radiography, Appendectomy }). In this paper we propose a new approach called MMISP (*Mining Multi-dimensional-Itemset Sequential Patterns*) to extract patterns from sequences including both multi-dimensional and set valued data. The novelties of the proposal lies in: (i) the way in which the data can be efficiently compressed; (ii) the ability to reuse a state-of-the-art sequential pattern mining algorithm and (iii) the extraction of new kind of patterns. We introduce as a case-study, experiments on real data aggregated from a regional healthcare system and we point out the usefulness of the extracted patterns. Additional experiments on synthetic data highlights the efficiency and scalability of our approach.

Keywords: Sequential Patterns, Multi-dimensional Sequential Patterns, Data Mining

1 Introduction

Data warehouses are constituting a large source of informations that are used and exploited to extract useful knowledge for expert analysis and decision makers [3]. In temporal data warehouses, every bit of information is associated with a timeline describing a total order over events. This particular ordering introduces more complexity to the extraction process and more precisely to mining

processes that enumerate patterns that encompass interesting transient events. Many efficient approaches were developed to mine these patterns (i.e., sequential patterns) like PrefixSpan [5], SPADE [12], SPAM [1], PSP [4], DISC [2], PAID [10], FAST [8]. However, all these techniques and algorithms, without any exception, focus solely on sequences of set valued data (i.e., *itemsets*) and contrast with real-world data that have multiple dimensions. To overcome this problem, Pinto et al. [6] introduced the notion of multi-dimensionality in sequences and proposed an efficient algorithm. Later works, like Zhang et al. [13] or Yu et al. [11] extended the initial Pinto's approach for different scenarios and use-cases. While in set valued approaches the events are represented by itemsets, in multi-dimensional temporal databases the events are defined over a fixed schema in which all the attributes are mandatory in the extracted patterns. Furthermore, and this is particularly true in the data warehouse environment, background knowledge is usually available and can be represented as a hierarchy over the values of the attributes. Following this logic, Plantevit et al. introduced *M3SP* [7], an efficient algorithm that is able to incorporate different dimensions and their ordering (organization) in the sequential pattern mining process. The benefit of this approach is to extract patterns with the most appropriate level of granularity. Still, this idyllic representation of uniform data is very uncommon in real-world applications where heterogeneity is usually elevated to a foundational concept. In this study, we focus on extracting knowledge from medical data warehouse representing information about patients in different hospitals. The successive hospitalizations of a patient can be expressed as a sequence of multi-dimensional attributes associated with a set of medical procedures. Our goal is to be able to extract patterns that express patients stays along with combinations of procedures over time. This type of pattern is very useful to healthcare professionals to better understand the global behavior of patients over time. Unfortunately this full richness and complexity of the data cannot be exploited by any of the traditional sequential pattern mining techniques. In this paper, we propose a new approach to extract patterns from sequences which include multi-dimensional and set valued data at the same time. In addition, the proposed approach incorporates background knowledge in the form of hierarchies over attributes.

The remainder of this paper is organized as follows, Section 2 introduces the problem statement as well as a running example. The method for extracting multi-dimensional-itemset frequent patterns is described in Section 3. Section 4 presents experimental results from both quantitative and qualitative point of views and Section 5 concludes the paper.

2 Problem Statement

In this section we list some preliminary definitions needed to formalize and present our problem. First of all, we introduce a motivating example from a real data set related to the PMSI (Program of medical information systems). This French nationwide information system describes hospital activities from

both economical and medical points of view. In this system, each hospitalization is related to the recording of administrative, demographical and medical data.

Patients	Trajectories
P_1	$\langle((UH_{Paris}, C_1), \{p_1, p_2\}), ((UH_{Paris}, C_1), \{p_1\}), ((GH_{Lyon}, R_1), \{p_2\})\rangle$
P_2	$\langle((UH_{Paris}, C_1), \{p_1\}), ((UH_{Paris}, C_1), \{p_1, p_2\}), ((GH_{Lyon}, R_1), \{p_2\})\rangle$
P_3	$\langle((UH_{Paris}, C_1), \{p_1, p_2\}), ((GH_{Lyon}, R_1), \{p_2\})\rangle$
P_4	$\langle((UH_{Paris}, C_1), \{p_2\}), ((UH_{Paris}, R_2), \{p_3\}), ((GH_{Lyon}, R_2), \{p_2\})\rangle$

Table 1. An example of a database of patient trajectories

Let S_{DB} be a database of multi-dimensional-itemset data sequences. Table 1 illustrates such a database.

Definition 1. *(Dimensions and specialization down(d)) A dimension $(D, \leqslant)$ is a partially ordered set. For a given d in D, down(d) (resp. up(d)) denotes the set of all specializations $\{x \in D | x \leqslant d\}$ (resp. generalizations $\{x \in D | d \leqslant x\}$) of d.*

Example 1. Figure 1 shows two hierarchical taxonomies characterizing both the hospital and diagnosis dimensions. For hospital dimension, $UH_{Paris} \in down(UH)$ as UH_{Paris} is a direct descendant of UH.

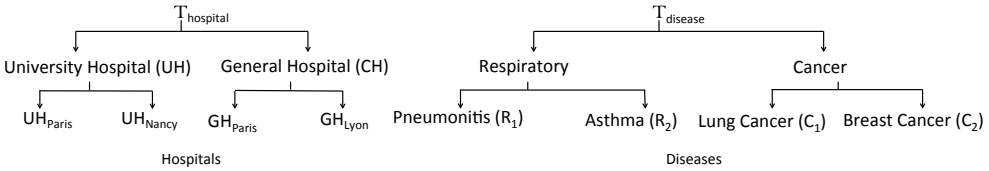


Fig. 1. Hospital and diagnoses taxonomies

By taking into account taxonomies, we define a multi-dimensional component as follows:

Definition 2. *(Multi-dimensional component) Given a dimension $(D, \leqslant)$, a multi-dimensional component over D, denoted $(mdc, \leqslant_{mdc})$, is a tuple $(d_1, ..., d_m)$ where $d_i \in D$, $i = 1, \cdots, m$. For two given multidimensional components $mdc = (d_1, ..., d_m)$ and $mdc' = (d'_1, ..., d'_m)$, $mdc' \leqslant_{mdc} mdc$ denotes that mdc' is more specific than mdc, if for every $i = 1, ..., m$, $d'_i \in down(d_i)$.*

Example 2. Let (UH_{Paris}, *Lung Cancer*) and (UH, *Cancer*) be two multidimensional components. (UH_{Paris}, *Lung Cancer*) $\leqslant_{mdc}$ (*UH, Cancer*) because $UH_{Paris} \in down(UH)$ and *Lung Cancer* $\in$down(*Cancer*).

Definition 3. *(Event) An event* $e = ((d_1, ..., d_m), \{p_1, ..., p_n\})$ *is a pair including a multidimensional component and an associated itemset. Given two events* $e = ((d_1, ..., d_m), \{p_1, ..., p_n\})$ *and* $e' = ((d_1^{'}, ..., d_m^{'}), \{p_1^{'}, ..., p_{n'}^{'}\})$*, e is included in e′, denoted by* $e \subseteq_e e'$*, if and only if* $(d_1, ..., d_m) \leqslant_{mdc} (d_1^{'}, ..., d_m^{'})$ *and* $\{p_1, ..., p_n\} \subseteq \{p_1^{'}, ..., p_{n'}^{'}\}$.

Example 3. $e = ((UH, T_{disease}), \{p_1, p_2, p_3\})$ is an event, where $(UH, T_{disease})$ is a multidimensional component with two dimensions representing hospital and diagnosis. $\{p_1, p_2, p_3\}$ denotes the set of medical procedures. An event $e' = ((UH_{Paris}, C_1), \{p_1, p_2\})$ is included in e, $e' \subseteq_e e$, because $(UH_{Paris}, C_1) \leqslant_{mdc} (UH, T_{disease})$ and $\{p_1, p_2\} \subseteq \{p_1, p_2, p_3\}$.

A multi-dimensional-itemset data sequence is composed of events.

Definition 4. *(Multi-dimensional-itemset Sequence) A multi-dimensional-itemset sequence* $s =< e_1, e_2, ..., e_l >$ *is an ordered list of events* e_i*. Given two Multi-dimensional-itemset Sequences* $s =< e_1, e_2, ..., e_l >$ *and* $s' =< e'_1, e'_2, ..., e'_{l'} >$*, s is included in s′, denoted by* $s \subseteq_s s'$*, if there exist indices* $1 \leq i_1 < i_2 < ... < i_l \leq l'$ *such that* $e_j \subseteq_e e'_{i_j}$ *for all* $j = 1 \dots l$ *and* $l \leqslant l'$.

Example 4. The sequence $s = \langle((UH_{Paris}, Cancer), \{p_1, p_2\}), ((GH_{Lyon}, R_1), \{p_2\})\rangle$ is a sequence of two events. It expresses the fact that a patient was admitted to the University Hospital of Paris UH_{Paris} for a cancer disease $Cancer$ and underwent procedures p_1 and p_2, then he went to the General Hospital of Lyon GH_{Lyon} for pneumonitis R_1 and underwent procedure p_2. A sequence $s = \langle((UH_{Paris}, C_1), \{p_1\})\rangle$ is included in s, $s' \subseteq_s s$, because $((UH_{Paris}, C_1), \{p_1\}) \subseteq_e ((UH_{Paris}, Cancer), \{p_1, p_2\})$.

Definition 5. *(Patient Trajectory) A patient trajectory is defined as* a multi-dimensional-itemset sequence.

Example 5. In Table 1, the sequence $s = \langle((UH_{Paris}, C_1), \{p_1, p_2\}), ((UH_{Paris}, C_1), \{p_1\}), ((GH_{Lyon}, R_1), \{p_2\})\rangle$ represents the trajectory for the patient P_1.

Let $supp(s)$ be the number of sequences that includes s in S_{DB}. Furthermore σ be a minimum support threshold specified by the end-user.

Definition 6. *(Most Specific Frequent Sequence) Let s be multi-dimensional-itemset sequences, we can say that, s is the most specific frequent sequences in* S_{DB}*, if and onl if:* $supp(s) \geq \sigma$ *and* $\nexists s' \in S_{DB}$*, where* $supp(s) = supp(s')$ *and* $s' \subseteq_s s$.

Problem 1. The problem of mining multi-dimensional-itemset sequences is to extract the set of all most specific frequent sequences in S_{DB} such as $supp(s) \geq \sigma$. By using the taxonomies we can extract more or less general or specific patterns and overcome problems of excessive granularity and low support.

Example 6. Let $\sigma = 0.75$ (i.e. a sequence is frequent if it appears at least three times in S_{DB}). The sequence $s_1 = \langle(UH_{Paris}, C_1), \{p_1, p_2\}), ((GH_{Lyon}, R_1), \{p_2\}\rangle$ is frequent. $s_2 = \langle(UH, Cancer), \{p_1, p_2\}), ((GH, Respiratory), \{p_2\}\}\rangle$ is also frequent. Nevertheless, s_2 is not kept since it is too general compared to s_1.

3 Mining Multi-dimensional-Itemset Sequential Patterns

In this section, we present the MMISP (*Mining Multi-dimensional-Itemset Sequential Patterns*) algorithm for extracting multi-dimensional-Itemset sequential patterns with different levels of granularity over each dimension. MMISP follows a bottom-up approach by first focusing on extracting *mdc* that can exist at different level of granularity, then it considers the itemset part of the events and compute the support of every item is S_{DB}. After these two steps, frequent multi-dimensional components and frequent items are combined to generate events. In the final step, the frequent events are mapped to a new representation and a standard sequential mining algorithm is applied to enumerate multi-dimensional itemset sequential patterns.

In the next subections, we provide the details of each step of our work and discuss the different challenges.

3.1 Generating Multi-dimensional Components

MMISP starts by processing the multi-dimensional components of the sequences. Basically it considers three types of dimensions: a temporal dimension D_t, a set of analysis dimensions D_A and a set of reference dimensions D_R. MMISP splits S_{DB} into blocks according to reference dimension D_R. Then, MMISP sorts each block according to the temporal dimension D_t. The tuples of multi-dimensional component appearing in an event are defined w.r.t. analysis dimensions D_A. The support of the multi-dimensional component is computed according to dimensions of D_R. It is the ratio of the number of blocks supporting the multi-dimensional component over the total number of blocks. This a classic way of partitioning the database and was introduced in [7].

Date	Hospital	Diagnosis
1	UH_{Paris}	C_1
2	UH_{Paris}	C_1
3	GH_{Lyon}	R_1

Block: $Patient_1$

Date	Hospital	Diagnosis
1	UH_{Paris}	C_1
2	GH_{Lyon}	R_1

Block: $Patient_3$

Date	Hospital	Diagnosis
1	UH_{Paris}	C_1
2	UH_{Paris}	C_1
3	GH_{Lyon}	R_1

Block: $Patient_2$

Date	Hospital	Diagnosis
1	UH_{Paris}	C_1
2	UH_{Paris}	R_2
3	GH_{Lyon}	R_2

Block: $Patient_4$

Fig. 2. Block partition of the database according to D_R={Patient}

Example 7. In our example, H (hospitals) and D (diseases) are the analysis dimensions, $Date$ is the temporal dimension, and P (patients) is the reference dimension. By using P (patients) to split the dataset, we obtain four blocks defined by $Patient_1$, $Patient_2$, $Patient_3$ and $Patient_4$ as shown in Figure 2.

Following this partitioning step, MMISP generates all the frequent multi-dimensional components. Firstly, we generate the most general multi-dimensional component, that is $(T_1, ..., T_m)$. In our running example, we have two dimensions (hospital and disease), so the most general multi-dimensional component is $(T_{hospital}, T_{disease})$. Then, our approach generates all multi-dimensional components of the form $(T_1, ...T_{i-1}, d_i, T_{i+1}, ..., T_m)$ where $d_i \in down(T_i)$. We take only the frequent multi-dimensional component which has support greater than σ. In our running example and for $\sigma = 75\%$ (3 blocks from 4), we have four new frequent multidimensional components: $(UH, T_{disease})$, $(GH, T_{disease})$, $(T_{hospital}, Respiratory)$ and $(T_{hospital}, Cancer)$.

We continue the recursive generation of the new multidimensional components by using each previously generated frequent multidimensional component (a). This is done with a pivot method that identifies an integer z which is the position of the last dimension in a and is not top T. For example if a=$(UH, T_{Disease})$, z is the first dimension (hospital) because the value for the hospital dimension (UH) and the second dimension (disease) has the value $T_{disease}$.

For each dimension d_k in a, where $k \in [z, m]$, we replace d_k with one of its specialization from the set $down(d_k)$. For example, if a=$(UH, T_{Disease})$, we have z=1 and we can generate four new mdc_s: $\{(UH_{Paris}, T_{Disease}), (UH_{Nancy}, T_{Disease}),$ $(UH, Respiratory), (UH, Cancer)\}$. The first and the second multidimensional components are generated by replacing UH by $down(UH) = \{UH_{Paris}, UH_{Nancy}\}$, the third and the forth multidimensional components are generated by replacing $T_{Disease}$ by $down(T_{Disease}) = \{Respiratory, Cancer\}$.

We select only the frequent multidimensional components. For our previously example with $\sigma = 75\%$, $\{(UH_{Paris}, T_{Disease}), (UH, Cancer)\}$ are the new frequent multidimensional components generated by $(UH, T_{Disease})$.

Finally, from all frequent multi-dimensional components generated, we select only the most specific multi-dimensional component.

Definition 7. *(Most Specific Multi-dimensional Component) Let a be multi-dimensional component, we can say that, a is the most specific multi-dimensional component, if and only if* $\nexists$ *a′ multi-dimensional component, where supp*$(a) =$ *supp*(a') *and* $a' \leqslant_{mdc} a$.

Frequent multi-dimensional component
(UH_{Paris}, C_1)
(GH_{Lyon}, R_1)

Table 2. The most specific frequent multi-dimensional components

Example 8. Fiqure 3 illustrates the mechanism work of generation all frequent multi-dimensional components on our running example with $\sigma = 0.75$. We can notice that the most specific components are (UH_{Paris}, C_1) and (GH_{Lyon}, R_1).

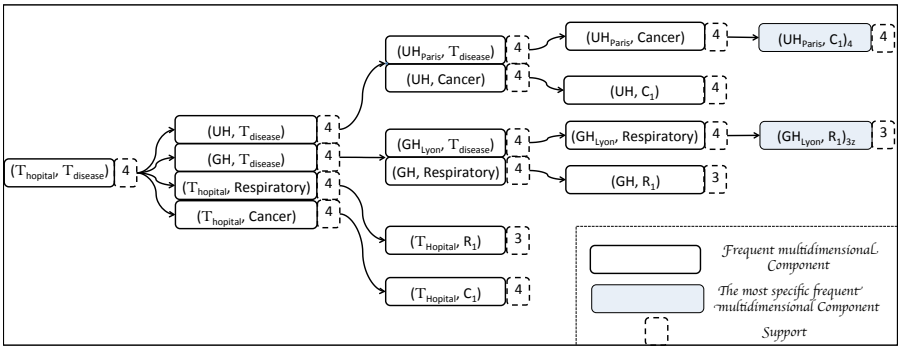


Fig. 3. Frequent multi-dimensional components generation

3.2 Generating Frequent Itemset

In this step, MMISP focuses on the itemset part of the sequences. Basically, this step aims at extracting the set of all items that are frequent in a sequence of length 1. Let us remind that usually, in level-wise approaches, either itemset-extension or sequence-extension can be considered. For example, if we have a sequence $s_1 = \langle\{1,2,3\}\rangle$, then $s_2 = \langle\{1,2,3\}\{4\}\rangle$ is a sequence-extended sequence of s_1 and $s_3 = \langle\{1,2,3,4\}\rangle$ is an itemset-extended sequence of s_1. In our context focusing on sequence of length 1 we will only consider itemset-extension. Such an operation can be easily done by using any standard sequential pattern algorithm.

Patients	Sequences of procedures
P_1	$\langle\{p_1,p_2\}\{p_1\}\{p_2\}\rangle$
P_2	$\langle\{p_1\}\{p_1,p_2\}\{p_2\}\rangle$
P_3	$\langle\{p_1,p_2\}\{p_2\}\rangle$
P_4	$\langle\{p_2\}\{p_3\}\{p_2\}\rangle$

Sequences of procedures

Frequent Itemset Candidates
$\{p_1\}$
$\{p_2\}$
$\{p_1,p_2\}$

The frequent itemset

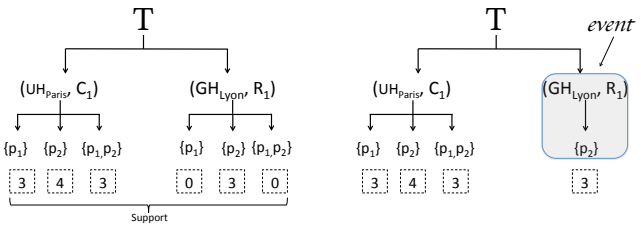
Fig. 4. The frequent itemset generated

Example 9. Figure 4 shows the sequence of medical procedures for patients and the frequent itemset candidates for $\sigma = 0.75$.

3.3 Generating Frequent Events

Generating frequent events is achieved by combining frequent multi-dimensional components with frequent itemsets. This task is achieved by building a prefix tree such that the first level in this tree is composed of the frequent multidimensional components and the second level is composed the frequent itemsets. More precisely, each branch in the tree represents an event. Then a scan is performed

over the database to prune irrelevant events from the tree. For example, Figure 5 illustrates the tree before and after pruning infrequent events for $\sigma = 0.75$.



Combination tree before pruning Combination tree after pruning

Fig. 5. An example of the tree for generating frequent events before and after the pruning

3.4 Extracting Frequent Multi-dimensional Itemset Patterns

Frequent sequences can then be mined by using any standard sequential pattern mining algorithm. As these algorithms require that the dataset to be mined is composed of pairs in the form (id, seq), where id is a sequence identifier and seq is a sequence of itemsets, we transform the initial dataset as follows:

- Each branch in the combination tree after pruning is assigned a unique id which will be used during the mining operation. This is illustrated in Table 3 .
- Each block (patient) is assigned a unique id of the form P_i.
- Every block b is transformed into a pair $(P_i, \mathbb{S}(p_i))$, where $\mathbb{S}(P_i)$ is built according to the date and the content of the blocks. The final result is reported in Table 4.

A standard sequence mining algorithm can be applied on the transformed database.

event-id	Frequent Event
e_1	(UH_{Paris}, C_1), $\{p_1\}$
e_2	(UH_{Paris}, C_1), $\{p_2\}$
e_3	(UH_{Paris}, C_1), $\{p_1, p_2\}$
e_4	(GH_{Lyon}, R_1), $\{p_2\}$

Table 3. Identification each branch (Event) in T

id	Sequence data
P_1	$\langle\{e_1, e_2, e_3\}\{e_1\}\{e_4\}\rangle$
P_2	$\langle\{e_1\}\{e_1, e_2, e_3\}\{e_4\}\rangle$
P_3	$\langle\{e_1, e_2, e_3\}\{e_4\}\rangle$
P_4	$\langle\{e_2\}\rangle$

Table 4. Transformed database

Then, the extraction of frequent sequences can be carried out. With $\sigma = 0.75$, the pattern $\langle\{e_3\}\{e_2\}\rangle$ is frequent. This sequence transforms to $\langle((UH_{Paris}, C_1), \{p_1, p_2\}), ((GH_{Lyon}, R_1), \{p_2\})\rangle$ by using the identification in Table 3.

4 Experiments

We conduct experiments on both real and synthetic datasets. The algorithms is implemented in Java and the experiments are carried out on a MacBook Pro with a 2.5GHz Intel Core i5, 4GB of RAM Memory running OS X 10.6.8. The extraction of sequential patterns is based on the public implementation of CloSpan algorithm [9]. We use the implementation supplied by the IlliMine[5] toolkit.

In order to assess the effectiveness of our approach, we run several experiments on the PMSI dataset. This database includes the following informations: *Patients* (id, gender ...), *Stays* (id, hospital, principal diagnosis, ...) and *Medical Procedures* (id, date,...). Our dataset contains 486 patients suffering from lung cancer and living in the eastern region of France. The average length of data sequences is 27. The data is encoded using controlled vocabularies. In particular, diagnoses are encoded with the International Classification of Diseases (ICD10)[6]. This classification is used as an input taxonomy for MMISP. The ICD10 can be seen as a tree with two levels. As illustrated in Figure 6, 3-characters codes such as C34 (Lung cancer) have specializations: C340 is cancer of the main bronchus, C341 is cancer of upper lobe etc.

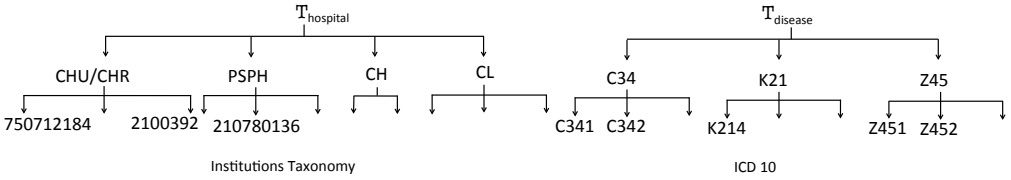


Fig. 6. Examples of taxonomies used in multilevel sequential pattern mining

Patients	Trajectories
P_1	$\langle((C341, 750712184), \{ZBQK002\}), ((Z452, 580780138), \{ZZQK002\}), \ldots\rangle$
P_2	$\langle((C770, 100000017), \{ZBQK002\}), ((C770, 210780581), \{ZZQK002, YYYY030\}), \ldots\rangle$
P_3	$\langle((H259, 210780110), \{YYYY030\}), ((H259, 210780110), \{ZZQK002\}), \ldots\rangle$
P_4	$\langle((R91, 210780136), \{YYYY030\}), ((C07, 210780136), \{ZBQK002\}), \ldots\rangle$

Table 5. Care trajectories of 4 patients

Table 5 is an example of care trajectories described over two dimensions (diagnosis, hospital ID) coupled with a set of medical procedures. For example (C341, 750712184), $\{ZBQK002\}$ represents the stay of a patient in the Univer-

[5] http://illimine.cs.uiuc.edu/
[6] http://apps.who.int/classifications/apps/icd/icd10online/

sity Hospital of Dijon (coded as 210780581) treated for a lung cancer (C341), where the patient underwent chest radiography (coded as ZBQK002).

The experiments are designed to extract some multi-dimensional sequential patterns for helping the medical experts to describe some healthcare patients trajectories. For this experiment the support value is set to 15 (i.e. $\sigma = 0.03$). MMISP generates 121 different patients trajectories. Table 6 shows some patients trajectories obtained by our approach. *Pattern 2* can be interpreted as follows: 42% of patients have a hospitalization in the University Hospital of Dijon for a lung cancer (210780581,C341), where they underwent supplement procedures (coded as YYYY030) for passing the chest radiography (coded as ZBQK002). Then, the same patients go to any university hospital for doing chemotherapy (CHU/CHR,Z511), where they underwent only the chest radiography (coded as ZBQK002).

id	Support	Trajectory Patterns
1	70%	$\langle((CH, Z515), \{ZBQK002, YYYY030\})\rangle$
2	42%	$\langle((210780581, C341), \{ZBQK002, YYYY030\}), ((CHU/CHR, Z511), \{ZBQK002\})\rangle$
3	38%	$\langle((210780581, Z511), \{ZBQK002, YYYY030\}), ((210780581, Z511), \{ZBQK002, YYYY030\})\rangle$

Table 6. Some healthcare patients trajectories obtained by MMISP

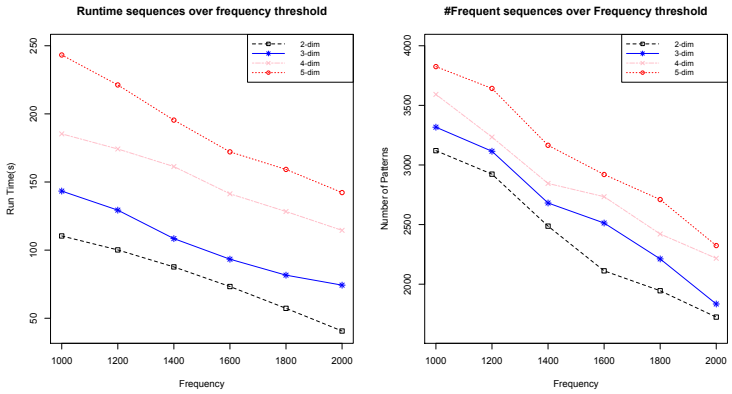


Fig. 7. Running Time (left) and Number of extracted pattern (right) obtained by MMISP with varying in the number of dimension

In the second experiments, we study the scalability of our approach. We consider the number of extracted patterns and the running time with respect to two different parameters, the number of the dimension in the multidimensional

components and the average length of the itemsets in the data sequences. The first batch of synthetic data generated contains 10 000 sequences defined over (2, 3, 4 and 5) analysis' dimensions. Each sequence contains 30 events and each event is described, in average, by 15 items in the itemset. Hierarchical relations are defined over 5 levels of granularity between elements of each analysis dimension. Figure 7 reports the results according to different values of support threshold for different number of analysis dimension in multidimensional component. We can notice that the running time increases for each newly added analysis dimension. The second batch of synthetic data generated contains 10 000 sequences with varying number of items 5, 10, 15 and 20. The sequences in the four generated data sets have an average cardinality of 30 events,by 3 analysis dimensions. Hierarchical relations are defined over 5 levels of granularity between elements of each analysis dimension. Figure 8 reports the results according to different values of support threshold for different lengths of itemsets.

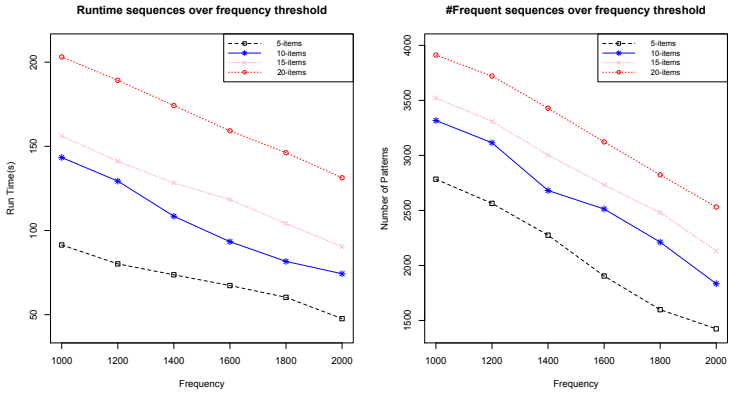


Fig. 8. Running Time (left) and Number of extracted pattern (right) obtained by MMISP with varying itemsets' cardinalities

5 Conclusion

In this paper, we propose a new approach to mine multi-dimensional itemset sequential patterns. Our approach is able to capture knowledge from dataset represented over both multi-dimensional component and itemsets. We provide formal definitions and propose a new algorithm MMISP to mine this new kind of pattern. We conduct experiments on both real and synthetic datasets. The method was applied on real-world data where the problem was to mine healthcare patients trajectories. According to medical experts, new patterns are easier to understand.

References

1. Jay Ayres, Jason Flannick, Johannes Gehrke, and Tomi Yiu. Sequential pattern mining using a bitmap representation. In *KDD*, pages 429–435, 2002.
2. Ding-Ying Chiu, Yi-Hung Wu, and Arbee L. P. Chen. An efficient algorithm for mining frequent sequences by a new strategy without support counting. In *ICDE*, pages 375–386, 2004.
3. Jeffrey Cohen, John Eshleman, Brian Hagenbuch, Joy Kent, Christopher Pedrotti, Gavin Sherry, and Florian Waas. Online expansion of largescale data warehouses. *PVLDB*, 4(12):1249–1259, 2011.
4. Florent Masseglia, Fabienne Cathala, and Pascal Poncelet. The psp approach for mining sequential patterns. In *PKDD*, pages 176–184, 1998.
5. Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Meichun Hsu. Prefixspan: Mining sequential patterns by prefix-projected growth. In *ICDE*, pages 215–224, 2001.
6. Helen Pinto, Jiawei Han, Jian Pei, Ke Wang, Qiming Chen, and Umeshwar Dayal. Multi-dimensional sequential pattern mining. In *CIKM*, pages 81–88, 2001.
7. Marc Plantevit, Anne Laurent, Dominique Laurent, Maguelonne Teisseire, and Yeow Wei Choong. Mining multidimensional and multilevel sequential patterns. *TKDD*, 4(1):1–37, 2010.
8. Eliana Salvemini, Fabio Fumarola, Donato Malerba, and Jiawei Han. Fast sequence mining based on sparse id-lists. In *Proceedings of the 19th international conference on Foundations of intelligent systems*, ISMIS'11, pages 316–325, Berlin, Heidelberg, 2011. Springer-Verlag.
9. Xifeng Yan, Jiawei Han, and Ramin Afshar. Clospan: Mining closed sequential patterns in large datasets. In *In SDM*, pages 166–177, 2003.
10. Zhenglu Yang, Masaru Kitsuregawa, and Yitong Wang. Paid: Mining sequential patterns by passed item deduction in large databases. In *IDEAS*, pages 113–120, 2006.
11. Chung-Ching Yu and Yen-Liang Chen. Mining sequential patterns from multidimensional sequence data. *IEEE Trans. Knowl. Data Eng.*, 17(1):136–140, 2005.
12. Mohammed J. Zaki. Spade: An efficient algorithm for mining frequent sequences. *Mach. Learn.*, 42(1-2):31–60, January 2001.
13. Changhai Zhang, Kongfa Hu, Zhuxi Chen, Ling Chen, and Yisheng Dong. Approx-mgmsp: A scalable method of mining approximate multidimensional sequential patterns on distributed system. In *FSKD (2)*, pages 730–734, 2007.