Linked data and online classifications to organise mined patterns in patient data

Linked Data and Online Classifications to Organise Mined Patterns in Patient Data

Nicolas Jay, MD, PhD[1,2] and Mathieu d'Aquin, PhD[3]
[1] Université de Lorraine, LORIA, UMR 7503
Vandœvre-lès-Nancy, F-54506, France
[2] CHU de Nancy
Nancy, F-54000, France
nicolas.jay@loria.fr
[3] Knowledge Media Institute, The Open University
Walton Hall, Milton Keynes, MK7 6AA, UK
mathieu.daquin@open.ac.uk

Abstract

In this paper, we investigate the use of web data resources in medicine, especially through medical classifications made available using the principles of Linked Data, to support the interpretation of patterns mined from patient care trajectories. Interpreting such patterns is naturally a challenge for an analyst, as it requires going through large amounts of results and access to sufficient background knowledge. We employ linked data, especially as exposed through the BioPortal system, to create a navigation structure within the patterns obtained form sequential pattern mining. We show how this approach provides a flexible way to explore data about trajectories of diagnoses and treatments according to different medical classifications.

Introduction

Since the concept appeared two decades ago, knowledge discovery in databases (KDD) has became increasingly popular in the biomedical domain[1]. KDD is an iterative and interactive process of 9 steps[2] aiming at "identifying valid, novel, potentially useful, and ultimately understandable patterns in data". KDD can be broadly divided in three main stages: preprocessing of data, data mining and interpretation of results. Data mining is the application of specific algorithms for extracting patterns from massive data. While it is only a step in KDD process, large efforts have been devoted to the development of efficient and fast algorithms able to deal with huge amounts of data or complex data. Though it has received less attention, the interpretation step can in many cases be as problematic. Indeed, some methods such as association rule mining, frequent itemset search or sequential pattern mining can generate large amounts of results, overwhelming the analyst and making it difficult to extract much insight from the data. Several strategies have been employed to select interesting patterns, many of them based on quantitative metrics[3]. Subjective evaluation of the patterns is usually carried out by the domain experts who might not be familiar with KDD methods. Therefore, it is of much importance to improve the interpretability of the results to promote the adoption of KDD by medical and healthcare professionals.

In the biomedical domain, the collection of enormous amounts of data in administrative or clinical information systems, as well as in clinical and fundamental research, has stressed the need for formalised and machine processable knowledge in order to facilitate information management, interoperability and data sharing. Formally structured medical knowledge, such as ontologies, terminologies, classifications, is now widely used to annotate, retrieve or reason upon biomedical data[4]. Formalized knowledge can support the KDD process in different ways[5], especially as a way to organise patterns and therefore support domain experts in interpreting the results of the data mining step. In this context, knowledge should ideally be accessible in distributed resources, without overloading the data management task which is already very significant in KDD processes. Several initiatives led to the constitution of online repositories of biomedical knowledge[6;7]. In parallel, within the paradigm of the Semantic Web, Linked Data has emerged as a set of principles for exposing, sharing, and connecting pieces of data, information, and knowledge.

In this paper, we present a method that exploits external information available as *linked open data* to support the interpretation of data mining results, through automatically building a navigation/exploration structure in the results of a particular type of data mining tool (namely, sequence mining). We demonstrate this method through a use case based on the analysis of trajectories of care.

Patient ID	Stay Order	Item
p1	1	C34.1
p1	1	I10
p1	1	GEQE007
p1	1	GEHE001
p1	2	C34.1
p1	2	C77.1
p1	2	GFFA022
p1	3	...
...	...	...

Table 1: An exerpt of the trajectory databases

Mining sequential patterns in trajectories of care

With the increasing burden of chronic conditions, health care systems are facing the challenge of containing medical expenditures while maintaining quality of care[8]. Patients with chronic conditions use more services and a greater array of services than other consumers. Multiple encounters of chronic patients with the healthcare system define a so-called "trajectory" of care. Lack of coordination along such trajectories, bad implementation of guidelines or inappropriate organisation of the healthcare system may have a negative impact on quality and costs of care. In that context, the analysis of trajectories of care could provide useful insights about the real patient's journey in the health system, and help practitioners and deciders to better manage chronic care.

Although originally designed for billing and financing purposes, claim databases routinely collect healthcare and administrative data, potentially holding valuable information for the analysis of trajectories of care. In France, the PMSI[*] is a nationwide information system, adapted from the Diagnoses Related Groups[9], that covers all hospitalisations in the field of acute care.

In this work, we recomposed the sequences of hospitalisations of patients living in the French region of Lorraine. Patients were selected in the PMSI database if they had a stay in surgery for Lung cancer. An observation window of 2 years centred on this surgical stay was used to recompose the trajectory of care of the selected patients. The trajectory of care for each patient was then modelled as a sequence of stays, each stay being characterised by its main diagnosis, possible co-morbidities and clinical procedures delivered during the stay. In the PMSI, medical problems are coded with the 10th International Classification of Diseases (ICD10). Medical and surgical procedures are coded with a French nomenclature: the CCAM[†]. The set of trajectories, composed of these ICD10 and CCAM items, were then mined using a sequential pattern mining algorithm.

Sequential pattern mining was first introduced by Agrawal and Srikant[10]. It is a popular approach to discover patterns in ordered data and has multiple application in the biomedical domain[11;12;13;14]. The task of sequential pattern mining can be defined as follows:

Let I be a finite set of items. A subset of I is called an itemset. A sequence $s = \langle t_1 t_2 \ldots t_k \rangle$ ($t_i \subseteq I$) is an ordered list of itemsets. A sequence $\alpha = \langle a_1 a_2 \ldots a_n \rangle$ is a subsequence of a sequence $\beta = \langle b_1 b_2 \ldots b_m \rangle$ if and only if $\exists i_1, i_2, \ldots i_n$ such that $i_1 \leq i_2 \leq \ldots \leq i_n$ and $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2} \ldots a_n \subseteq b_{i_n}$. We note $\alpha \subseteq \beta$ and also say that β contains α. Let $D = \{s_1, s_2 \ldots s_n\}$ be a database of sequences. The support of a sequence α in D is the proportion of sequences of D containing α. Given a *minsup* threshold, the problem of frequent sequential pattern mining consists in finding the set of sequences whose support is greater than *minsup*.

In this work, each patient is represented by its trajectory of care, i.e. as an ordered sequence of hospital stays. Each stay is an itemset containing diagnosis codes and procedure codes. An excerpt of the trajectory database is shown in table 1 for illustration.

In this table, patient p1 had 3 stays. The first stay with diagnoses of malignant neoplasm of upper lobe, bronchus or lung (C34.1) and hypertension (I10). During this stay, the patient underwent a bronchoscopy (GEQE007) with biopsies (GEHE001). During a second stay, p1 underwent a lobectomy (GFFA022) and was diagnosed with intrathoracic lymph node metastasis (C77.1). The trajectory of care of p1 can be written as follow :

$$\langle \{\texttt{C34.1,I10,GEQE007,GEHE001}\}\{\texttt{C34.1,C77.1,GFFA022}\}\ldots \rangle$$

The trajectory database contained 828 sequences. Mining this database with a support threshold of 300 returned a set of 379 sequential patterns. Beyond the number of extracted patterns, the interpretation of the results can also be cumbersome because of their format, as sequential pattern have basically the same form as input sequences.

[*] Programme de médicalisation des Systèmes d'Information

[†] Classification Commune des Actes Médicaux

Interpretation of sequential patterns with linked data

Linked Data[15] is a set of principles and technologies that rely on the architecture of the Web (URIs and links) to share, model and integrate data. The basic idea is that data objects (e.g., a surgical procedure) are identified by web addresses (URIs), and the information attached to these objects are represented through links (themselves labeled with URIs) to values (e.g., an anatomical site) or other URIs representing other objects (e.g., a patient). Besides this simple technological model, the main novelty introduced by linked data is this idea that raw data is represented and exposed directly on the Web, making the Web a collective data space connecting contributions from any possible sources.

In the biomedical domain, large scale data generation and high-throughput technologies leveraged the development of structured knowledge and ontologies facilitating the management of data and scientific findings[6]. Accessing formalised biomedical knowledge stored into shared and public repositories has become essential for researchers who can now exploit data resources from miscellaneous domains[16]. Initiatives such as Bioportal[7] conform to the principles of Linked Data by providing access to more than 330 ontologies through a SPARQL endpoint.

Considering the potential development and availability of biomedical Linked Data, it seems natural to investigate it as a source of additional information to support the interpretation of the results of a data mining process, such as the ones presented in the previous section. Below, we describe an approach using several linked data endpoints to collect descriptive dimensions about the items that constitute the sequential patterns. Theses dimensions are used to automatically classify the extracted patterns, thus generating a structure that can support exploration and navigation into the results of the data mining step . Figure 1 gives an overview of this approach, which relies on a linked data-based description of the data mining results, on extracting from external linked data sources selected information about these items and on organising the extracted patterns in a hierarchy. This hierarchy is obtained by applying Formal Concept Analysis (FCA)[17] on the patterns and their description.

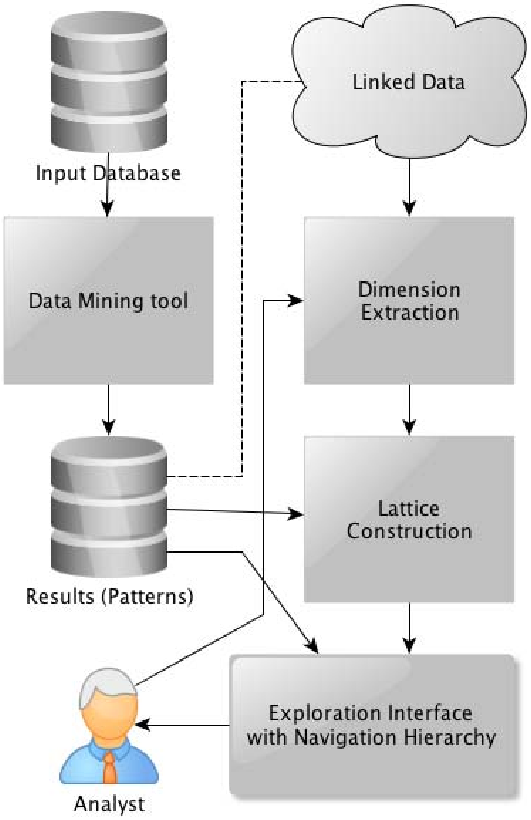


Figure 1: Overview of the approach to using Linked Data in interpreting the results of data mining.

Linking patterns to terminologies through SPARQL endpoints

Enriching extracted patterns with external information includes two aspects: 1- modelling the data mining results in RDF, in accordance with the principles of linked data and 2- ensuring that the items in the patterns are identified with URIs in reference to existing Linked Data sources.
The first step was achieved by creating a generic model of sequential patterns inspired by the "Sequence" Ontology Design Patterns[‡]. This model is represented in Figure 2. It defines three main classes (sequentialPattern, itemSet and item) and their properties.

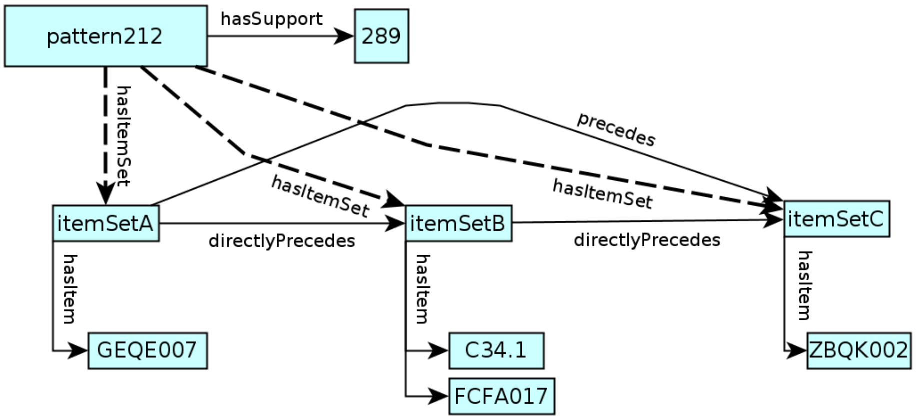


Figure 2: RDF Model of the pattern ⟨{GEQE007}{C34.1,FCFA017}{ZBQK002}⟩ with support 289.

The second step of the representation is the one that connects the generic pattern mining representation described above with Linked Data-based external information characterising the patterns. Here, we used two external sources of information relying on the controlled vocabularies employed to code diagnoses and medical procedures in the PMSI system: the ICD10 and the CCAM.
The ICD10 is available as an RDF graph on Bioportal through a SPARQL[18] query endpoint (`http://sparql.bioontology.org/sparql/`). ICD10 codes, such as C34.1 are identified by a URI of the form `http://purl.bioontology.org/ontology/ICD10/C34.1`. The hierarchical structure of the ICD10 is modelled by the `rdfs:subClassOf` property. This property allows to link a code to its ancestors at different levels: chapter, block, category. For example, the block `Extrapyramidal and movement disorders` can be reached by chaining the `rdfs:subClassOf` property from the code `G21.0:Malignant neuroleptic syndrome` through the category `G21:Secondary parkinsonism`. The Bioportal SPARQL endpoint provides additional properties about ICD10 concepts such as their label, their UMLS Concept Unique Identifier and semantic types... Among these, mappings towards other terminologies (e.g., SNOMED-CT) can constitute an interesting alternative way to explore the ICD10 items, and therefore sequential patterns made of them. From a technical point of view, ontologies and inter-terminological mappings are hosted in two separate SPARQL endpoints (`http://sparql.bioontology.org/sparql/` and `http://sparql.bioontology.org/mappings/sparql/`).
The CCAM is a nomenclature of clinical procedures that has been developed in France, following the principles established by the European Galen project[19]. CCAM concepts are identified by a semistructured 7 digits code (4 letters and 3 numbers) corresponding to a multi-axial construct. The first two letters give the topography site of procedure, the third one represents the action and the fourth one the approach or technical device. The last three (numeric) characters allow to distinguish procedures of similar type. Furthermore, the CCAM is independently organised in an arborescent structure, grouping procedures according to anatomical systems and their diagnostic/therapeutic nature. To the best of our knowledge, the CCAM is not available through Linked data technologies. However, CISMeF, a French initiative indexing health Internet resources, has developed a health multi-terminology portal[20]. The website[§] delivers HTML content including the CCAM, so we reconstructed the CCAM as a RDF graph, reusing the model and names adopted by CISMeF, and setup a local RDF store that is accessible through a SPARQL query endpoint.

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

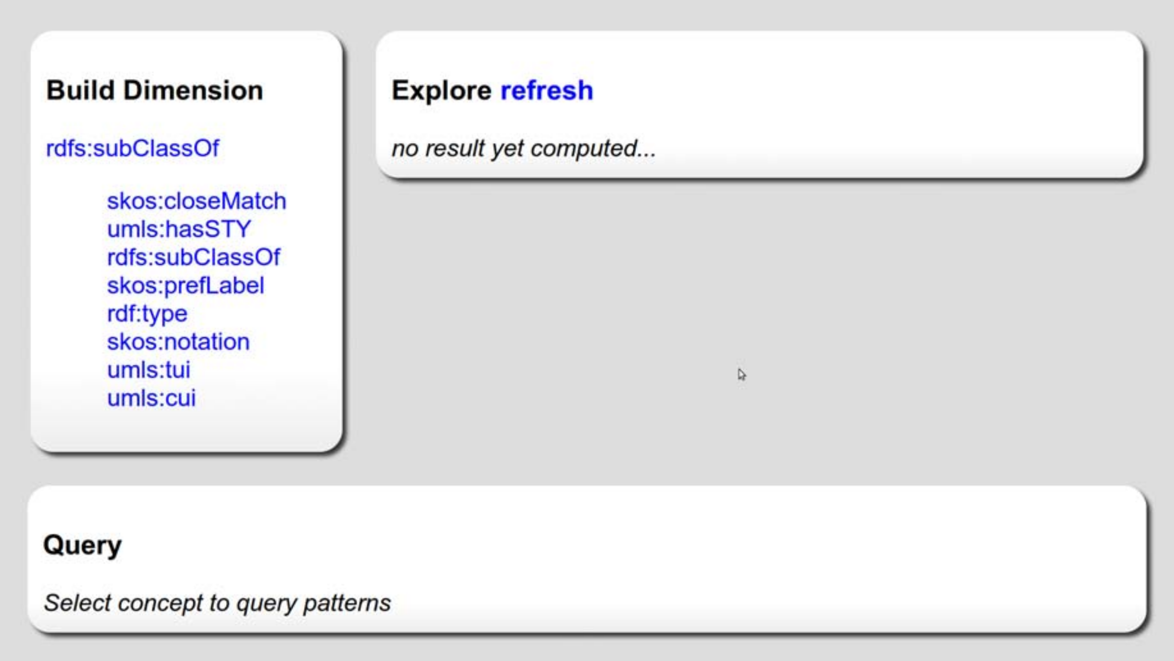
[§] `http://pts.chu-rouen.fr/`

Selection of dimensions for pattern exploration

In our system, once the patterns are linked to external data, the analyst is provided a choice among possible directions of exploration. Indeed, considering the existing data endpoints (in our case, the two BioPortal ones and the local store containing the CCAM), the properties attached to the items represent as many descriptors that enrich the initial sequences used as input of the data mining process. This applies not only to properties that are directly attached to the items, but also indirectly to any path that can be built from them in the linked data graph (i.e., any property chain) starting with the items. For example, the analyst can exploit the hierarchical structure of the ICD10 to chain the C34.1 ICD10 code to the label of its block. This operation can be achieved based on the following RDF triples :

```
ICD10:C34.1      rdfs:subClassOf   ICD10:C34
ICD10:C34        rdfs:subClassOf   ICD10:C30-C39.9
ICD10:C30-C39.9  skos:prefLabel    "Malignant neoplasms
                                   of respiratory and intrathoracic organs"
```

In order to help the analyst in selecting the property chain he wants to apply as a dimension for exploration, we built an interface that allows him/her to check what properties apply to items in the mined patterns (see Figure 3). A series of SPARQL queries are used to find out, in the available linked data endpoints, what properties first apply to any of the items forming the patterns. Subsequently, the analyst can select what properties apply to the values of these properties. By repeating this process, he/she can build a chain of properties linking to a set of new descriptors for the patterns. In the example Figure 3, the top left frame show the properties that can be selected as the next link in the property chain. On this example, the user has already selected `rdfs:subClassOf`, meaning that he/she is navigating upward in the ICD10 tree. The user then has the possibility to apply `rdfs:subClassOf` to go one level upper in the ICD10, to select the inter-terminological mappings corresponding to the actual level with the `skos:closeMatch` property, or to select another dimension included in the ICD10 (e.g., the corresponding UMLs TUI code).



Build Dimension

rdfs:subClassOf

skos:closeMatch
umls:hasSTY
rdfs:subClassOf
skos:prefLabel
rdf:type
skos:notation
umls:tui
umls:cui

Explore refresh

no result yet computed...

Query

Select concept to query patterns

Figure 3: Creating a property chain for exploring the patterns.

Once a chain of properties has been created, patterns can be reorganised and explored according to the new dimension of exploration. The new descriptors characterising the patterns can be retrieved using, in principle, a mechanism similar to executing the following SPARQL query[¶]:

```
PREFIX pat: <http://mdnj.kmi.org/pattern#>
SELECT distinct ?sp ?v WHERE {
```

[¶] Because chains of properties might span over more than one SPARQL endpoint, the actual implementation is technically more complex than this, but the query illustrates the basic underlying idea.

```
    ?sp pat:hasItemSet ?is.
    ?is pat:hasItem ?item.
    ?item p0 ?v1.
    ?v1 p1 ?v2.
    ...
    ?vn pn ?v
  }
```

Where `pat` denotes the prefix of our RDF model for sequential patterns and `p0...pn` the chain of properties selected by the analyst. The result is a set of pairs `?sp ?v` associating sequential patterns with their new descriptors.

Classification and pattern exploration with Formal Concept Analysis

At this stage, we have a set of patterns enriched with a set of new attributes that were absent from the initial data. The new attributes can be used to meaningfully organise the patterns into a hierarchy, structured according to the chosen dimension. This can be achieved with Formal Concept Analysis (FCA), a theory of data analysis introduced in[21]. FCA is tightly connected with data-mining and especially the search for frequent itemsets[22]. This method organises information into a concept lattice representing inherent structures existing in data.

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

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

The composition of the $'$ operators, noted $''$, defines a closure operator. For any $A \subseteq G$ and $B \subseteq M$, A'' and B'' are closed sets whenever $A = A''$ and $B = B''$.

A formal concept of the context $K = (G,M,I)$ is a pair $(A,B) \subseteq G \times M$ where $A' = B$ and $B' = A$. A is called the *extent* and B is called the *intent*. A concept (A_1,B_1) is a *subconcept* of a concept (A_2,B_2) if $A_1 \subseteq A_2$ (which is equivalent to $B_2 \subseteq B_1$) and we write $(A_1,B_1) \leq (A_2,B_2)$. The set of all concepts of a formal context K together with the partial order relation $\leq$ forms a lattice and is called concept lattice of K. This lattice can be represented as a Hasse diagram providing a visual support for interpretation.

In our application, the formal context represents the relation between the set of sequential patterns and the set of new descriptors. Formally, a sequential pattern p is related to a descriptor d if p contains an item that can be linked to d with the chaining process previously described.

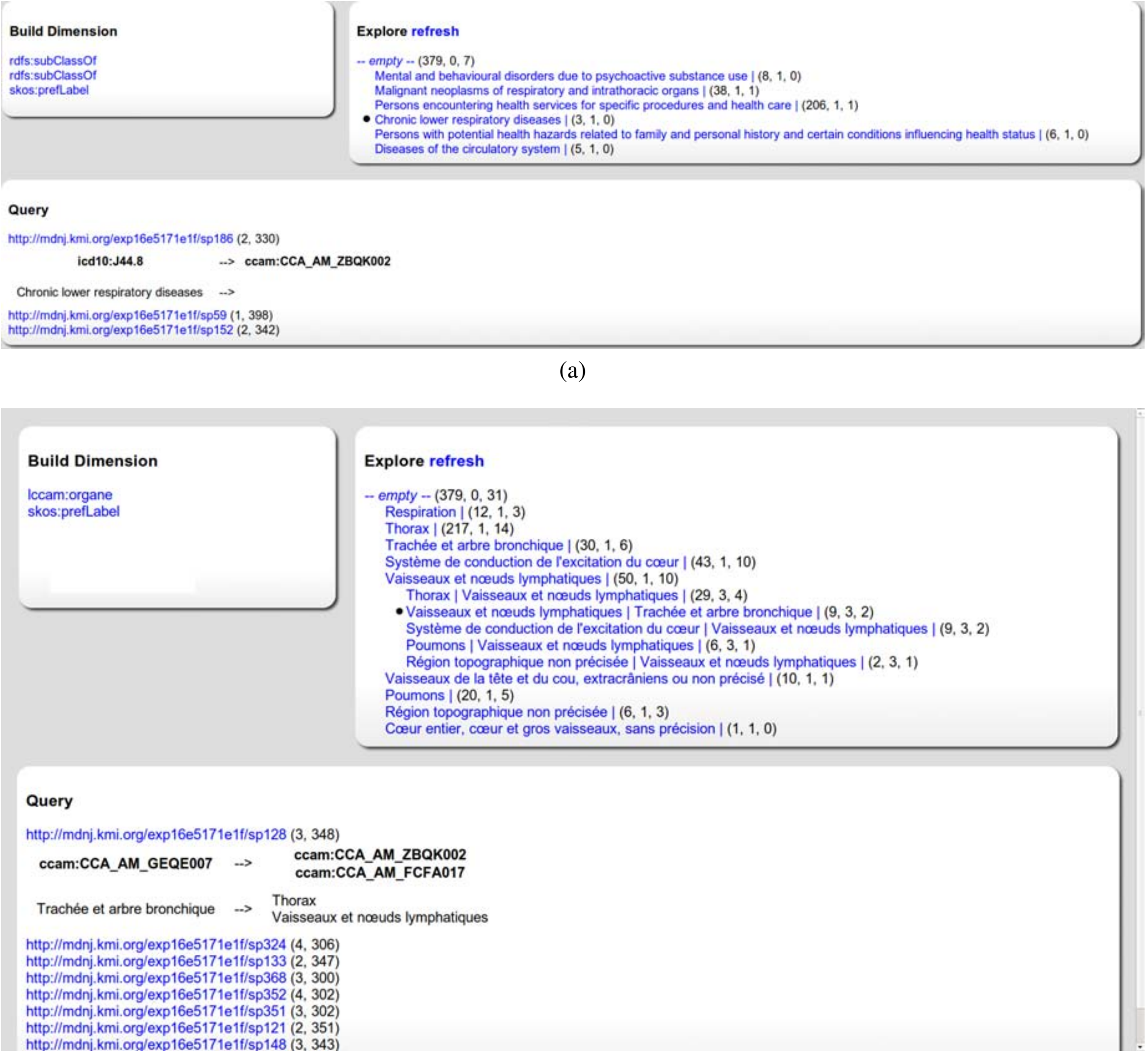
Since the built lattice constitutes a hierarchy, it is natural to use it for navigating the data upon which it was built. This has been explored before in several different domains[23]. Here however, we use the lattice to provide a further level of abstraction with respect to the original data: First, the sequential pattern mining method provides an additional structure over individual sequences and second, the lattice provides a way to classify and explore these patterns according to dimensions brought through Linked Data.

Following this idea, Figures 4(a) and 4(b) show the previously introduced interface with the navigation structure created. In (a), the chain of properties links patterns to the label of ICD10 blocks (`refs:subClassOf` twice followed by `skos:prefLabel`) as the navigation dimension. The upper-right part of the interface now shows an expandable hierarchy based on the built lattice, starting from the top concept (the one with an empty extension and all the patterns in the intension).

Next to each concept are indicators of their size/importance. The first number is the size of the concept extension (the number of patterns). For example, there are 3 sequential patterns including a diagnosis of a “Malignant neoplasm of respiratory and intrathoracic organs”. Next is the number of more general concepts, and the number of more specific concepts (e.g., the previous concept has 1 more general concept, the top of the hierarchy, and 1 more specific ones, as displayed on the interface).

Finally, each concept in the hierarchy can be selected, to show the details that relate to it. In the lower part of Figure 4(a), the concept “Chronic lower repiratory disease” has been selected, showing the 3 corresponding sequential patterns extracted in the data. The first one has the focus: it is a sequence of 2 items and is supported by 300 patients.

Similarly, Figure 4(b) shows the classification of sequential patterns along the “organe” (anatomic) axis of the CCAM. In that case, patterns including a procedure on the trachea and bronchial tree, as well as a procedure on the lymph nodes have been filtered. The selected pattern represents trajectories containing a stay with a bronchospy followed by a stay with a mediastinal lymph node resection.



Build Dimension

rdfs:subClassOf
rdfs:subClassOf
skos:prefLabel

Explore refresh

-- *empty* -- (379, 0, 7)
Mental and behavioural disorders due to psychoactive substance use | (8, 1, 0)
Malignant neoplasms of respiratory and intrathoracic organs | (38, 1, 1)
Persons encountering health services for specific procedures and health care | (206, 1, 1)
• Chronic lower respiratory diseases | (3, 1, 0)
Persons with potential health hazards related to family and personal history and certain conditions influencing health status | (6, 1, 0)
Diseases of the circulatory system | (5, 1, 0)

Query

http://mdnj.kmi.org/exp16e5171e1f/sp186 (2, 330)
icd10:J44.8 --> **ccam:CCA_AM_ZBQK002**

Chronic lower respiratory diseases -->

http://mdnj.kmi.org/exp16e5171e1f/sp59 (1, 398)
http://mdnj.kmi.org/exp16e5171e1f/sp152 (2, 342)

(a)

Build Dimension

lccam:organe
skos:prefLabel

Explore refresh

-- *empty* -- (379, 0, 31)
Respiration | (12, 1, 3)
Thorax | (217, 1, 14)
Trachée et arbre bronchique | (30, 1, 6)
Système de conduction de l'excitation du cœur | (43, 1, 10)
Vaisseaux et nœuds lymphatiques | (50, 1, 10)
Thorax | Vaisseaux et nœuds lymphatiques | (29, 3, 4)
• Vaisseaux et nœuds lymphatiques | Trachée et arbre bronchique | (9, 3, 2)
Système de conduction de l'excitation du cœur | Vaisseaux et nœuds lymphatiques | (9, 3, 2)
Poumons | Vaisseaux et nœuds lymphatiques | (6, 3, 1)
Région topographique non précisée | Vaisseaux et nœuds lymphatiques | (2, 3, 1)
Vaisseaux de la tête et du cou, extracrâniens ou non précisé | (10, 1, 1)
Poumons | (20, 1, 5)
Région topographique non précisée | (6, 1, 3)
Cœur entier, cœur et gros vaisseaux, sans précision | (1, 1, 0)

Query

http://mdnj.kmi.org/exp16e5171e1f/sp128 (3, 348)
ccam:CCA_AM_GEQE007 --> **ccam:CCA_AM_ZBQK002**
ccam:CCA_AM_FCFA017

Trachée et arbre bronchique --> Thorax
Vaisseaux et nœuds lymphatiques

http://mdnj.kmi.org/exp16e5171e1f/sp324 (4, 306)
http://mdnj.kmi.org/exp16e5171e1f/sp133 (2, 347)
http://mdnj.kmi.org/exp16e5171e1f/sp368 (3, 300)
http://mdnj.kmi.org/exp16e5171e1f/sp352 (4, 302)
http://mdnj.kmi.org/exp16e5171e1f/sp351 (3, 302)
http://mdnj.kmi.org/exp16e5171e1f/sp121 (2, 351)
http://mdnj.kmi.org/exp16e5171e1f/sp148 (3, 343)

(b)

Figure 4: Screenshots of the interface with the lattices built along (a) the ICD10, (b) the CCAM "organe" axis.

Discussion

The method presented above is generic in the sense that it only requires the results of the pattern mining method to be represented in accordance with Linked Data principles, and some relevant external Linked Data sources providing information about the considered items to function. Modelling topological relationships between items in a pattern could support the analyst for advanced querying, reasoning and classification tasks. The sequential model presented here is well suited to hospitalisation data and sufficient to express the results of the sequence mining algorithms family[10]. It allows the representation of both sequential and simultaneous facts in the form of ordered lists of itemsets. For expressing more elaborated temporal relationships such as Allen's interval algebra[24], several approach have been proposed in the field of Linked Data[25]. However, our approach is not restricted to the exploration of sequential patterns. This principle could be extended to other kinds of patterns by adapting the pattern RDF model. In our particular use case, Linked Data provides a highly customisable interface to navigate into patterns representing care processes. Different views are proposed to the analyst, e.g., classifications of patterns according to diagnoses or to procedures. In each case, the granularity of the classification can be adjusted using the hierarchical structure of the ICD10 and the CCAM. The exploration can also take advantage from the multidimensional nature of ontologies like the CCAM. Flexibility is in our view the core benefit of this approach. The possibility to easily map descriptors to patterns by traversing a chain of properties across distant sources of biomedical knowledge illustrates the vast potential of Linked Data for KDD. This stresses the importance of publicly accessible repositories exposing structured biomedical knowledge such as BioPortal. It also enables the fusion of private/local and public data and knowledge. Besides, its obvious integration capabilities of Linked Data, the proposed approach can also reduce data management costs, which are usually significant in the KDD process.

The idea of using domain knowledge in the KDD process is not new[26]. In the biomedical field, Pathak & al.[27] show the potential of Semantic Web technologies for federating heterogeneous data into an input dataset. Regarding sequential pattern mining, the seminal work of Srikant and Agrawal[10] already considered the possibility of using taxonomic data. However, these approaches focus on the incorporation of knowledge at the preprocessing or data-mining step. Our approach makes it possible to identify relevant dimensions in the data at the time of interpretation. Indeed, it is possible to select any dimension after having produced the sequential patterns, and to compare how different dimensions produce different navigational structures in the results. These dimensions could of course be included originally in the data being mined, but at the price of increased computational costs and number of patterns. Besides, it is not always possible to identify in advance what dimensions are relevant. Our method can be seen as complementary and as a way of selecting knowledge in combination with other techniques[14], thus closing the loop of the KDD process.

Of course, there are some limitations in the system presented in this paper. As we propose a classification of patterns, and not individuals, the metrics associated with patterns can be misleading when exploring dimensions that were not in the input dataset. For example, the support of patterns would change if a different ICD10 level of granularity was used to code diagnoses. Our approach was designed to suggest such adaptations but anticipating the impact of changing viewpoints on patterns metrics could clearly improve the system. Another limitation is related to technical implementation in the federation of queries on multiple SPARQL endpoints. Indeed, while a simplified illustration of the way in which multiple data endpoint can be used was described above, the actual technical implementation requires to build the results of querying each step of the chain of properties individually through querying each endpoints, therefore increasing the complexity and resource requirements of the overall approach.

Conclusion

In this paper, we have described and illustrated on a KDD process related to the mining of patients' trajectories, an approach to supporting the analyst in interpreting the results of data mining (especially, sequential pattern mining). The main contribution of this approach is not necessarily purely technical, but also in demonstrating how the availability of publicly available medical resources, exposed in a form that makes them easily connectable and integrated, can significantly impact on our ability to exploit data in a KDD process, to gain insight into the quality and effectiveness of healthcare in a particular domain.

As discussed above, the tool presented above only constitutes a first preliminary investigation into the possibilities of injecting background knowledge from Linked Data into the KDD process, with a high number of potential applications in the medical domain, but current limitations coming form the emerging field of Linked Data that will have to be tackled.

References

[1] Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, et al. Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst. 2012 Aug;36(4):2431–2448. Available from: `http://dx.doi.org/10.1007/s10916-011-9710-5`. 1

[2] Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communication of the ACM. 1996 Nov;29(11):27–34. 1

[3] Geng L, Hamilton HJ. Choosing the Right Lens: Finding What is Interesting in Data Mining. In: Quality Measures in Data Mining. Secaucus, NJ, USA: Springer-Verlag New York, Inc.; 2007. p. 3–24. 1

[4] Cimino JJ, Zhu X. The practical impact of ontologies on biomedical informatics. Yearb Med Inform. 2006;p. 124–135. 1

[5] Burgun A, Bodenreider O. Accessing and integrating data and knowledge for biomedical research. Yearb Med Inform. 2008;p. 91–101. 1

[6] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007 Nov;25(11):1251–1255. Available from: http://dx.doi.org/10.1038/nbt1346. 1, 3

[7] Noy N, Shah N, Whetzel P, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009;(37 Web Server):W170–3. 1, 3

[8] Nolte E, McKee M, editors. Caring for people with chronic conditions : A health system perspective. Open University Press; 2008. 2

[9] Fetter R, Shin Y, Freeman J, Averill R, JDThompson. Case mix definition by diagnosis-related groups. Med Care. 1980 Feb;18(2):1–53. 2

[10] Srikant R, Agrawal R. Mining Sequential Patterns: Generalizations and Performance Improvements. In: Apers PMG, Bouzeghoub M, Gardarin G, editors. Proc. 5th Int. Conf. Extending Database Technology, EDBT. vol. 1057. Springer-Verlag; 1996. p. 3–17. Available from: http://citeseer.ist.psu.edu/article/srikant96mining.html. 2, 7, 8

[11] Sallaberry A, Pecheur N, Bringay S, Roche M, Teisseire M. Sequential patterns mining and gene sequence visualization to discover novelty from microarray data. J Biomed Inform. 2011 Oct;44(5):760–774. Available from: http://dx.doi.org/10.1016/j.jbi.2011.04.002. 2

[12] Choi K, Chung S, Rhee H, Suh Y. Classification and sequential pattern analysis for improving managerial efficiency and providing better medical service in public healthcare centers. Healthc Inform Res. 2010 Jun;16(2):67–76. Available from: http://dx.doi.org/10.4258/hir.2010.16.2.67. 2

[13] Exarchos TP, Papaloukas C, Lampros C, Fotiadis DI. Protein classification using sequential pattern mining. Conf Proc IEEE Eng Med Biol Soc. 2006;1:5814–5817. Available from: http://dx.doi.org/10.1109/IEMBS.2006.260336. 2

[14] Egho E, Jay N, Raïssi C, Nuemi G, Quantin C, Napoli A. An Approach for Mining Care Trajectories for Chronic Diseases. In: Peek N, Morales RM, Peleg M, editors. AIME. vol. 7885 of Lecture Notes in Computer Science. Springer; 2013. p. 258–267. 2, 8

[15] Heath T, Bizer C. Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool; 2011. 3

[16] Rebholz-Schuhmann D, Nenadic G. Biomedical Semantics: the Hub for Biomedical Research 2.0. Journal of Biomedical Semantics. 2010;1(1):1. Available from: http://www.jbiomedsem.com/content/1/1/1. 3

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

[18] World Wide Web Consortium. SPARQL Query Language for RDF; 2008. Avaliable online from http://www.w3.org/TR/rdf-sparql-query/; last accessed 2012-08-02. 4

[19] Rodrigues JM, Rector A, Zanstra P, Baud R, Innes K, Rogers J, et al. An Ontology driven collaborative development for biomedical terminologies: from the French CCAM to the Australian ICHI coding system. Stud Health Technol Inform. 2006;124:863–868. 4

[20] Grosjean J, Merabti T, Dahamna B, Kergourlay I, Thirion B, Soualmia LF, et al. Health multi-terminology portal: a semantic added-value for patient safety. Stud Health Technol Inform. 2011;166:129–138. 4

[21] Wille R. Restructuring Lattice Theory: an approach Based on Hierarchies of concepts. In: Rival I, editor. Ordered Sets. Reidel; 1982. . 6

[22] Valtchev P, Missaoui R, Godin R. Formal Concept Analysis for Knowledge Discovery and Data Mining: The New Challenges. In: Eklund PW, editor. ICFCA. vol. 2961 of Lecture Notes in Computer Science. Springer; 2004. p. 352–371. 6

[23] Carpineto C, Romano G. Using Concept Lattices for Text Retrieval and Mining. In: Ganter B, Stumme G, Wille R, editors. Formal Concept Analysis. vol. 3626 of Lecture Notes in Computer Science. Springer; 2005. p. 161–179. 6

[24] Allen JF. Towards a general theory of action and time. Artificial Intelligence. 1984;23:123–154. 7

[25] Rula A, Palmonari M, Harth A, Stadtmüller S, Maurino A. On the Diversity and Availability of Temporal Information in Linked Open Data. In: Cudré-Mauroux P, Heflin J, Sirin E, Tudorache T, Euzenat J, Hauswirth M, et al., editors. International Semantic Web Conference (1). vol. 7649 of Lecture Notes in Computer Science. Springer; 2012. p. 492–507. 7

[26] Lieber J, Napoli A, Szathmary L, Toussaint Y. First elements on knowledge discovery guided by domain knowledge (KDDK). In: Proceedings of the 4th international conference on Concept lattices and their applications. CLA'06. Berlin, Heidelberg: Springer-Verlag; 2008. p. 22–41. Available from: `http://portal.acm.org/citation.cfm?id=1793623.1793626`. 8

[27] Pathak J, Kiefer RC, Bielinski SJ, Chute CG. Mining the human phenome using semantic web technologies: a case study for Type 2 Diabetes. AMIA Annu Symp Proc. 2012;2012:699–708. 8