Using Formal Concept Analysis for mining and interpreting patient flows within a healthcare network

Nicolas Jay[1,2], François Kohler[2], and Amedeo Napoli[1]

[1] Équipe Orpailleur, LORIA, Vandoeuvre-lès-Nancy, France
[2] Laboratoire SPI-EAO, Faculté de Médecine, Vandoeuvre-lès-Nancy, France

Abstract. This paper presents an original experiment based on frequent itemset search and lattice based classification. This work focuses on the ability of iceberg-lattices to discover and represent flows of patient within a healthcare network. We give examples of analysis of real medical data showing how Formal Concept Analysis techniques can be helpful in the interpretation step of the knowledge discovery in databases process. This combined approach has been successfully used to assist public health managers in designing healthcare networks and planning medical resources.

Key words: Formal Concept Analysis, frequent itemsets, network.

1 Introduction

Knowledge Discovery in Databases (KDD) is an iterative and interactive process for identifying valid, novel, and potentially useful patterns in data [1]. KDD is usually divided into three main steps: data preparation, data mining, and interpretation of the extracted units. Data mining, often considered as the central step in this process, is still an active field of research. The success key in KDD practice relies also on ability of easily producing units understandable as knowledge units. One way of achieving such a goal relies on an adapted visualization of the extracted units.

In this paper, we present an original experiment based on both frequent itemset search and lattice-based classification. This experiment holds on medical data and is aimed at showing the interactions and collaborations between hospitals in the French Region of Lorraine. This experiment may be regarded from two points of view: on the one hand, it is based on frequent itemset search on a medico-economic database, and on the other hand, the visualization of extracted units is based on Formal Concept Analysis (FCA) techniques [2], organizing the extracted units into a lattice for medical analysis and interpretation. At our knowledge, this is an original combination of data mining and FCA techniques that has been rarely carried on until now. Indeed, this is one of the main feature of this paper to show how FCA techniques can be very helpful in the interpretation step of KDD process. The results of this experiment have been used by healthcare administration in Lorraine for planning and evaluation purposes [3].

2 Health networks and collaborations

Healthcare networks are sets of healthcare actors working in cooperation, sharing information, and providing care for the same patients. In France, some networks are formally structured but others are still in an implicit existence. Thus, healthcare policy should be based on this current state of things to plan new networks or optimize existing ones. However, for both structured and implicit networks, knowledge on the degree of collaboration between hospitals is poor, because no information system is dedicated to this type of monitoring. Such an information system could help measuring collaboration by analyzing the flow of patients being treated in more than one hospital.

This issue is close to the problem of cartographying a communication network [4]. A healthcare network can be represented by an undirected graph where hospitals are the nodes, and edges represent patients flows, i.e. sets of patients shared by two hospitals. In our context, healthcare networks can involve hundreds of hospitals and tens of thousands of patients. It is a challenge to visualize a network with such a volume of data.

Furthermore, this problem goes beyond simple cartography. Patient flows depend on several constraints: geography, location of high technology devices and specialized medical teams, personal affinities between physicians, regulations, type of disease... According to these constraints, hospitals do not have the same role within a healthcare network. There exists high level relations that cannot be represented in usual network maps. In the domain of social network analysis, Freeman [5] has proposed to use FCA to produce useful insights about structural properties of relationships between social actors. This approach could be extended to our problem. Nevertheless, a lattice-based representation does not always support the size of data. A way to deal with that issue is to only represent the most significant flows.

The analysis of patient flows can also be seen as a consumer behavior problem. Consumer behavior and market basket analysis are well-known problems in data mining and can be solved using frequent itemset search and association rule extraction [6]. In our application domain, a formal context can be built with patients as objects and the hospitals in which they have been treated as attributes. Discovering significant flows of patient between hospitals can be achieved by mining this context for searching for frequent itemsets of hospitals sharing the same patients. However, it may be difficult to exploit the results because of the large number of extracted units, and because of the lack of visualization support.

The links between the frequent itemset search and FCA have been studied by several research groups [7–9]. Stumme [10] has introduced iceberg lattices, which are concept lattices of frequent closed itemsets. The approach combining visualization and frequent itemset search is a feature of first importance in our research work. Firstly, it is a top-down method for gradually discovering and representing significant patient flows. Secondly, it provides easily understandable results, especially for novice users. In a similar way, Duquenne [11] has studied associations of psychological handicaps of children. Using filters on a weighted lattice, he has shown the ability of FCA to describe profiles of patients. Further-

more, due to their ability to encode dualities [12], concept lattices can provide two points of view for interpreting patient flows: an intensional one in which flows result from interaction and collaboration of healthcare providers within a network, and an extensional one where flows can be regarded as groups of patients sharing a common medical profile.

3 Iceberg-lattices

Let $\mathbb{K} := (G, M, I)$ be a formal context where G is a set of objects, M a set of attributes and I a binary relation between G an M.

Definition 1. *Let $B \subseteq M$ and let minsupp be a threshold $\in [0, 1]$. The support count of the attribute set B in $\mathbb{K}$ is $supp(B) := \frac{|B'|}{|G|}$. B is said to be a frequent attribute set if $supp(B) \geqslant minsupp$.*

A concept is called frequent concept if its intent is frequent. The set of all frequent concepts of a context $\mathbb{K}$ is called iceberg lattice of the context $\mathbb{K}$.

4 Discovery process of patient flows

In France, the PMSI[3] database is a national information system used to describe hospital activity with both an economical and medical point of view. We have worked on two years of PMSI data of the Lorraine Region in France. Data preparation consists in building a formal context where objects are patients and attributes are hospitals lying in the database. A patient is related to a hospital whenever the patient has been treated in that hospital. An iceberg-lattice is then built from this context using the Titanic algorithm [10] implemented in Galicia 3.0 [13]. Hasse diagrams are drawn with the Graphiz [14] tools.

5 Results

We present here an example of cancer network analysis. In this experiment, the formal context holds 28009 patients and 158 hospitals. Figure 1 shows the resulting iceberg for a minsupp=0.017, i.e. 50 patients. For clarity, $\perp$ was removed and right and leftmost part of the lattice are not drawn. A first comment can be made about its general shape. It is more wide than deep because the context is sparse and data are poorly correlated. This means that patient flows are most of the time tightly partitioned, and that patients are rarely hospitalized in more than two hospitals. The intent of co-atoms, i.e. immediate descendants of $\top$, is always a singleton. This means a hospital never shares all of its patients with another one, or if it is so, less than 50 patients are involved in the interaction. The intent of atoms, i.e. the immediate ascendant of $\perp$, is always a pair. The extent of atoms gives an idea of the strength of the collaboration between the

[3] Programme de Médicalisation des Systèmes d'information.

two hospitals: the larger is the cardinal of the extent, the higher is the strength of the collaboration(i.e. the more patients are shared between the two hospitals). The iceberg can be divided in two parts:

- on the right, concepts that are both atoms and co-atoms. They represent institutions that share a few patients with others. This is that either they treat a few patients, or they work in a relative autonomy, or collaboration is split with many other hospitals.
- on the left, concepts that have at least a sub-concept (different from $\bot$). They represent a hospital receiving a significant number of patients, and having collaborations with at least one other establishment.

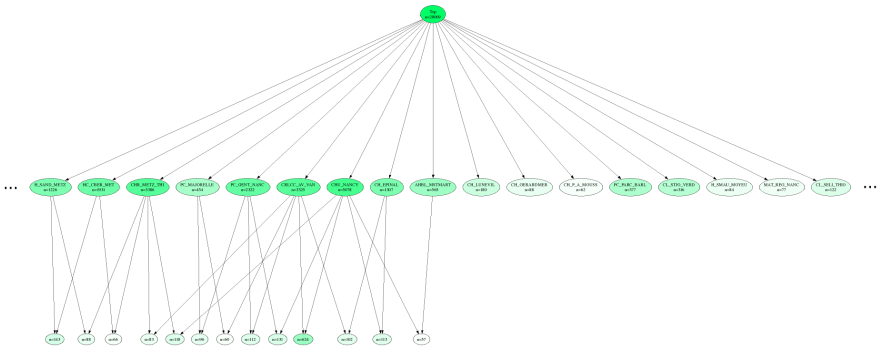

Fig. 1. Iceberg lattice for all type of cancer

The left part of the iceberg may be seen as the backbone of collaborations for cancer treatment, in the Lorraine region. This sub-lattice can be re-drawn removing both $\top$ and $\bot$ as shown on figure 2, along with a map of the hospitals in the Lorraine region. Co-atoms are then represented by ellipses. Their label shows the name of the hospital in their intent and the size of their extent. Diamonds are the second rank concepts (i.e. the atoms). Their label shows the size of their extent. Arrows represent the super-concept/sub-concept relation. These diamonds can be seen as cooperation between several hospitals. For example,CHU_NANCY and CRLCC_AV_VAN hospitals share 624 patients.

This figure contains a lot of information for the domain expert. First of all, three concepts have a large number of patients and many sub-concepts: CHU-NANCY, CHR-METZ-THI, and CRLCC-AV-VAN. They are located in Nancy and Metz, the two largest cities in Lorraine . The large number of sub-concepts related to these institutions precisely shows that they are reference centers. They employ highly skilled and specialized personnel. Treatments given there rely on state-of-art technology. Furthermore, they actively participate in anti-cancer research programs.

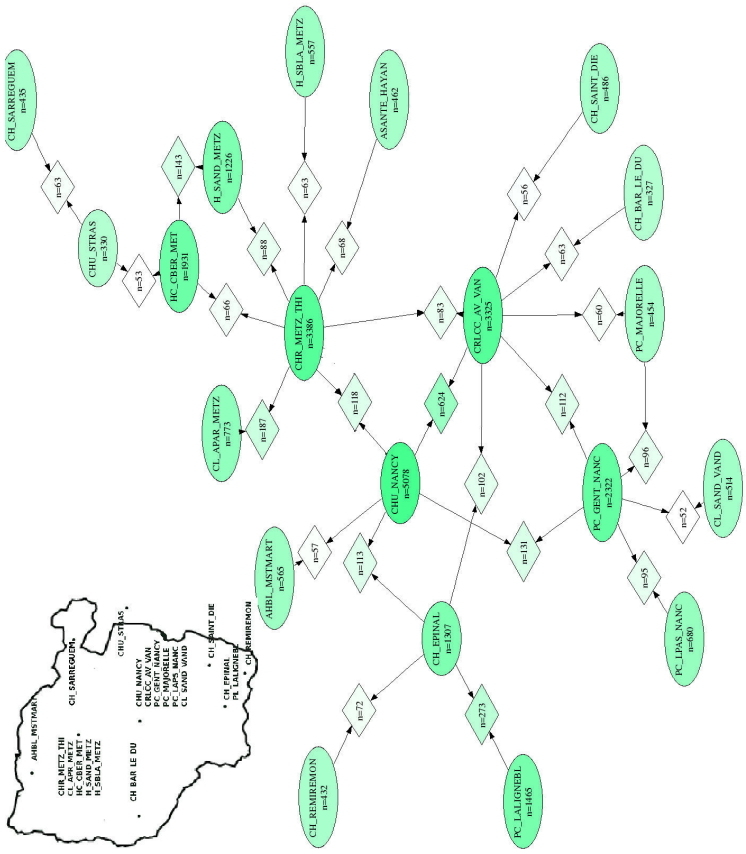

Fig. 2. Cooperation between hospitals for the treatment of cancer

The aspect of the lattice reflects geographical constraints that shapes the network structure. An East-West separation line clearly appears. Many flows are concentrated in the north around the CHR-METZ-TH hospital. By contrast, concepts sharing sub-concepts with CHU-NANCY and CRLCC-AV-VAN concepts concern most of the time hospitals located in the southern Lorraine. Let us also notice that the CH-SARRGUEMIN hospital on the top right of the figure has a trans-border cooperation with the CHU-STRAS hospital in the next region of Alsace.

The lattice also illustrates the influence of statutory constraints The PC-GENT-NANCY concept has common sub-concepts with PC-LPAS-NANCY, CL-SAND-VAND and PC-MAJORELLE. This makes a sub-network gathering private hospitals in the city of Nancy.

Finally, the lattice allows to visualize two important sets of knowledge units: one is on the most important centers for the treatment of cancer, the other on the

geographical locations of centers and the patient flows between these locations. Indeed, this can be seen as the concrete result of a working KDD system.

6 Conclusion

We have presented here a combined approach relying on data mining and FCA for representing patient flows in a healthcare system. This method takes advantage of iceberg-lattices to discover and to display in a simple way the backbone of healthcare networks.

References

1. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communication of the ACM **29-11**(1996), 27–34.
2. Ganter, B., Wille, R.: Formal Concept Analysis: mathematical foundations. Springer. Heidelberg 1999.
3. Jay, N., Napoli, A., Kohler, F.: Cancer Patient Flows Discovery in DRG Databases. Proc. MIE 2006 Conf. to appear.
4. Becker, R.A., Eick, S.G., Wilks, A.R. Visualizing Network Data IEEE Transactions on Visualization and Computer Graphics,**1**(1995), 16-28
5. Freeman, L.C., White D.R.: Using Galois Lattices to Represent Network Data. Sociological Methodology, **23**(1993), 127–146.
6. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In Proc. of the ACM SIGMOD Conference on Management of Data, Washington, D.C. (May 1993) 207–216.
7. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Closed set based discovery of small covers for association rules. Proc. BDA conf., (1999) 361–381.
8. Stumme, G.: Conceptual Knowledge Discovery with Frequent Concept Lattices. FB4-Preprint 2043, TU Darmstadt (1999)
9. Zaki, M.J., and Hsiao, C.: CHARM: An Efficient Algorithm for Closed Itemset Mining. SDM, 2002
10. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with TITANIC Data Knowl. Eng., Elsevier Science Publishers B. V.**42** (2002), 189–222
11. Duquenne, V.: Lattice analysis and the representation of handicap associations. Social Networks **18** (1996), 217–230.
12. Duquenne, V.: Latticial structures in data analysis. Theorical Computer Science. **217** (1999), 407–436.
13. Valtchev, P., Grosser, D., Roume, C., Hacene, M.R.: GALICIA: an open platform for lattices. in Using Conceptual Structures: Contributions to the 11th Intl. Conference on Conceptual Structures (ICCS'03), 241–254, Dresde (DE), Shaker Verlag, (21–25 July) 2003.
14. Gansner, E. R. and North, S. C. An open graph visualization system and its applications to software engineering. Softw. Pract. Exper. **30, 11** (Sep. 2000), 1203–1233.