

Using Domain Knowledge to Guide Lattice-based Complex Data Exploration

Nizar Messai¹ and Marie-Dominique Devignes and Amedeo Napoli and Malika Smail-Tabbone²

Abstract. In this paper we propose an approach which combines semantic resources and formal concept analysis to deal with heterogeneous data sets represented as many-valued (MV) formal contexts. We define a new Galois connection considering the semantic relationships between attribute values in a MV context. The semantic relationships are used to calculate the similarity between attribute values to decide whether an attribute is shared by a set of objects or not. Then, based on this Galois connection, we define MV formal concepts and MV concept lattices. Depending on a chosen similarity threshold, MV concept lattices may have different levels of precision. We take advantage of this feature to browse the content of a biological databases repository in a dynamic and progressive way. The browsing process combines the navigation in several MV concept lattices and allows zooming operations by switching between MV concept lattices with higher or lower precision.

1 Introduction

With the important growth of Semantic Web, many reliable semantic resources such as ontologies and thesauri are made available for public use. At the same time, Web data resources such as databases continue expansion and have more and more complex and heterogeneous contents. Both evolutions are not fully independent. Indeed, on the one hand, the development of Web data resources relies more than ever on semantic resources in almost all steps including structures design, content organization, etc. On the other hand, semantic resources are extended in parallel with the evolution of the content of data resources. The challenge now is how to take advantage of semantic resources for an optimal and efficient exploitation, i.e. search and reuse, of data resource contents. The present research work addresses this challenge in identifying relevant biological resources w.r.t. a repository of public biological databases (DBs), BioRegistry³ [4].

In BioRegistry, more than one thousand biological DBs are annotated with a set of metadata w.r.t. a defined annotation model based on the Dublin Core Metadata Initiative (DCMI⁴) model. Annotation metadata are terms belonging to semantic resources such as the Medical Subject Headings (MeSH⁵) thesaurus and the National Center for Biotechnology Information (NCBI⁶) taxonomy describing existing semantic relationships between these metadata.

Browsing BioRegistry for retrieving relevant biological DBs was first introduced in [10]. The approach consists of classifying biological DBs into a concept lattice built thanks to Formal Concept Analysis (FCA) [8]. The use of a concept lattice is motivated by flexibility and navigation capabilities provided by concept lattices [3]. The classification of biological DBs into a concept lattice requires an encoding into a formal context, i.e. a binary table where lines correspond to objects (DBs) and columns correspond to attributes (annotation metadata). However, in this way, the relationships between attributes when they exist, and this is often the case with real-world data, are lost. In particular, encoding BioRegistry content as a formal context does not allow to take into account semantic relationships between annotation metadata.

In this paper, we propose an approach taking into account semantic relationships between attributes during the concept lattice construction. These semantic relationships, as this is the case for BioRegistry, can be taken into account within a many-valued (MV) context, where attributes may have sets of elements or intervals of numbers as values. Then, a specific Galois connection is defined for attribute sharing. Here, two objects g_1 and g_2 share an attribute m when the values of m for g_1 and g_2 are “sufficiently similar”. Similarity is checked with respect to domain knowledge possibly lying in a domain ontology. The present research work completes and extends the ideas proposed in [11, 12]. The objective is to build an organizational structure for navigation and retrieval of resources. Domain knowledge is directly reused in the navigation and search process, and in the concept lattice construction. Thus, this research work participates in the design of concept lattices guided by domain knowledge, or in what could be called *Similarity-based Formal Concept Analysis*.

The paper is organized as follows. Section 2 recall main FCA definitions. Section 3 gives a detailed formalization of the FCA extension to consider semantic relationships while processing MV contexts. Due to space limitation, all the proofs are removed from this formalization. Section 4 details the application of the FCA extension for progressive and dynamic exploration of BioRegistry content and gives details on implementation and experimentations. Finally, Section 5 concludes the present work.

2 Formal Concept Analysis

2.1 Basics

In this section we recall basic definitions related to FCA. More detailed definitions and results can be found in [8]. FCA takes as input a data set represented as a formal context and produces the set of all the formal concepts which form a concept lattice. A formal context is denoted by $\mathbb{K} = (G, M, I)$ where G is a set of objects, M is a set of attributes, and I is a binary relation between G and M

¹ Procton Labs, 180 rue de Vaugirard 75015, Paris, FRANCE. (Nizar Messai was involved in the presented research work as member of LORIA and part-time assistant professor at Nancy University before August 2009)

² LORIA - CNRS - INRIA - Nancy University, 615 rue du Jardin Botanique 54600, Villers-lès-Nancy, FRANCE, email: surname.name@loria.fr

³ <http://bioregistry.loria.fr>

⁴ <http://dublincore.org/>

⁵ <http://www.nlm.nih.gov/mesh/meshhome.html>

⁶ <http://www.ncbi.nlm.nih.gov/>

($I \subseteq G \times M$). (g, m) $\in I$ denotes the fact that the object $g \in G$ is in relation through I with the attribute $m \in M$ (also read as g has m). A formal context can be encoded as a table where rows correspond to objects and columns to attributes. When an object g has an attribute m , the corresponding table entry contains a cross ($\times$). Table 1 shows an example of a formal context. The objects are supposed to be biological databases and the attributes are terms used to annotate these databases with respect to some available semantic resources (*NCBI taxonomy*, *MeSH thesaurus* and *Nucleic Acid Research (NAR) hierarchy of categories*⁷). The attributes corresponding to the DB Category are read as follows: *category.subcategory*. For example “1.2” corresponds to the second subcategory, “*Gene structure, introns and exons, splice sites*”, of the first category, “*Nucleotide Sequence Databases*”. In this example, the database *HSD* has the attributes *Human*, *Proteins* and 7.3.

Table 1. An example of formal context.

	DB Subject (MeSH)				Organism of interest (NCBI)							Category (NAR)			
	Genome	Genome Components	Proteins	Transcription factors	Eukaryotes	Human	Mammals	Plants	Prokaryotes	Rice	Vertebrates	1.2	5.2	7.3	13
ExInt		$\times$			$\times$							$\times$			
HSD			$\times$			$\times$								$\times$	
rRNDB	$\times$								$\times$				$\times$		
SpliceDB		$\times$					$\times$					$\times$			
CropNet								$\times$							$\times$
GOLD	$\times$												$\times$		
INE									$\times$						$\times$
TRANSCompel				$\times$							$\times$	$\times$			

The formal concepts are computed based on the relation I as maximal sets of objects having in common maximal sets of attributes. These maximal sets are obtained by the means of two dual derivation operators, $' : \mathfrak{P}(G) \rightarrow \mathfrak{P}(M)$ and $' : \mathfrak{P}(M) \rightarrow \mathfrak{P}(G)$, defined as follows for any $A \subseteq G$ and $B \subseteq M$:

$$A' = \{m \in M \mid (g, m) \in I \forall g \in A\}$$

$$B' = \{g \in G \mid (g, m) \in I \forall m \in B\}$$

$\mathfrak{P}(G)$ and $\mathfrak{P}(M)$ denote the powersets of G and M , respectively. A' is the set of all objects which have all attributes in B and B' is the set of all attributes common to all objects in A . For example, $\{rRNDB, GOLD\}' = \{Genome, 5.2\}$ and $\{Genome, 5.2\}' = \{rRNDB, GOLD\}$.

Formal concepts are then defined as pairs (A, B) such that $A \subseteq G$, $B \subseteq M$, $A' = B$, and $B' = A$. A and B are respectively called the *extent* and *intent* of the concept (A, B) . For example, $(\{rRNDB, GOLD\}, \{Genome, 5.2\})$ is a formal concept. A set $A \subseteq G$ of objects (respectively $B \subseteq M$ of attributes) is the extent (respectively the intent) of a concept if and only if $A'' = A$ (respectively $B'' = B$). In this case, A (respectively B) is said to be closed. For example $\{rRNDB, GOLD\}$ is closed whereas $\{GOLD\}$ is

not closed and $\{Genome, 5.2\}$ is closed whereas $\{Genome\}$ is not closed. In this way $'$ defines two dual closures systems on $\mathfrak{P}(G)$ and $\mathfrak{P}(M)$ and the derivation operators define a Galois connection between the two closure systems [8].

A concept $(A1, B1)$ is called a subconcept of $(A2, B2)$ when $A1 \subseteq A2$ (or equivalently $B2 \subseteq B1$). In this case, $(A2, B2)$ is called a superconcept of $(A1, B1)$ and we write $(A1, B1) \leq (A2, B2)$. For example, $(\{rRNDB\}, \{Prokaryotes, Genome, 5.2\}) \leq (\{rRNDB, GOLD\}, \{Genome, 5.2\})$. The set of all concepts in a context (G, M, I) ordered using the partial order “ $\leq$ ” forms a concept lattice denoted by $\mathfrak{B}(G, M, I)$. The concept lattice of the context given in Table 1 is represented by the so-called line diagram (or Hasse diagram) shown in Figure 1. This representation follows *reduced labelling* where objects and attributes are mentioned only once. The extent of a concept is obtained by considering the objects that can be reached by descending line paths from this concept. Correspondingly, the intent is obtained by considering the attributes that can be reached by following all line paths going upwards from this concept.

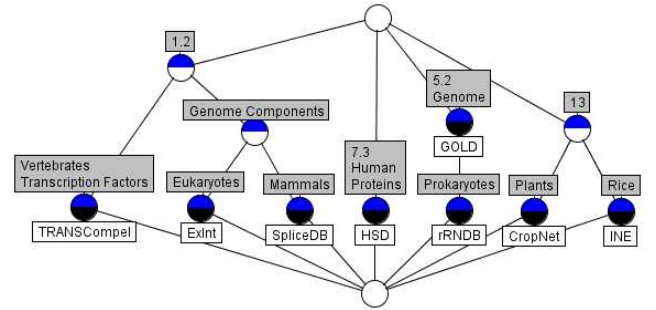


Figure 1. The concept lattice corresponding to the formal context given in Table 1.

2.2 Many-valued contexts and scaling

FCA provides appropriate tools for processing data sets representable as formal contexts. However, in real-world data sets, objects are often described by attributes which can take on several values. In this case, the attributes are said to be *many-valued*, contrasting to *one-valued* attributes considered above. Let us consider for example the context given in Table 1. Originally the biological databases in this context are described by three MV attributes: *DB Subject*, *Organism of interest*, and *Category* as shown in Table 2. This representation is called *many-valued context* [8]. Formally, a MV context is denoted by (G, M, W, I) where G is a set of objects, M is a set of attributes, W is a set of attribute values, and I is a ternary relation between G , M and W (i.e., $I \subseteq G \times M \times W$). (g, m, w) $\in I$ denotes the fact that “the MV attribute m takes the value w for the object g ”. This fact is also denoted by $m(g) = w$ (this notation will be preferred in the present paper). For example, $MeSH(HSD) = Proteins$.

The basic FCA setting detailed so far cannot be immediately applied to MV contexts. Therefore, a transformation mechanism called *conceptual scaling* [8] is proposed. It consists of transforming a MV context into a formal context (called “*scale*”) by replacing each MV attribute by a set of one-valued attributes. In this way, the context given in Table 1 results from a particular scaling, called *plain scaling*, of the MV context given Table 2. Although there are some rules that

⁷ <http://www.oxfordjournals.org/nar/database/c>

Table 2. An example of MV context.

	DB Subject (MeSH)	Organism of interest (NCBI)	DB category (NAR)
ExInt	Genome components	Eukaryotes	1,2
HSD	Proteins	Human	7,3
rRNDB	Genome	Prokaryotes	5,2
SpliceDB	Genome components	Mammals	1,2
CropNet		Plants	13
GOLD	Genome		5,2
INE		Rice	13
TRANSCompel	Transcription factors	Vertebrates	1,2

help to obtain scales [8], conceptual scaling does not give a unique result. It is possible to define different scales for the same MV context. The choice of a scale depends on attribute interpretations and makes conceptual scaling a user-dependent task which can hardly be automatized in the case of large data sets.

Some approaches have been proposed to directly process MV contexts without computing scales. They consist in defining new Galois connections adapted to the type of MV attributes. In [2], an approach based on *Symbolic Data Analysis* originally proposed to work with attributes valued in domains, intervals and histograms is extended to probabilistic data. In [1], approaches based on *Fuzzy Logic* are defined w.r.t. *Fuzzy Formal Contexts*, where values of attributes are truth degrees (of the relation existing between an object and the attribute). These two families of approaches have produced interesting results and allow to extend FCA to non binary contexts. However, they remain dependant on specific data formats. By contrast, the approaches proposed in [5] and [7] present an effective generalization of FCA and can be applied to general forms of MV contexts. These generalizations are similar as they both describe objects using formal descriptions instead of attributes. Both approaches provide a general framework which must be instantiated in practice by defining and implementing appropriate operators to handle the considered MV context. In [12], a new approach based on similarity intervals has been proposed to deal with MV contexts where attribute values can be totally ordered. This form includes MV contexts containing numerical attributes or categorical attributes with totally ordered values (e.g. tiny, small, medium, large). In the following, we extend this approach to deal with more general MV contexts. This extension takes advantage of available semantic resources to compute similarity between attribute values to efficiently process MV contexts of complex and heterogenous data sets.

3 MV contexts w.r.t. similarity between attribute values

3.1 Computing similarity between attribute values based on semantic relationships

Dealing with the similarities between attribute values in a MV context is the key idea of the approach developed hereafter. In the case of complex data, we consider available semantic resources to compute such similarities. For example, in the case of the MV context given in Table 2, we consider the MeSH thesaurus, the NCBI taxonomy, and the NAR hierarchy of categories to compute similarities between the values of the attributes DB Subject, Organism of interest, and Category, respectively. Fragments of these resources involving the terms used in the considered MV context are shown in Figure 2.

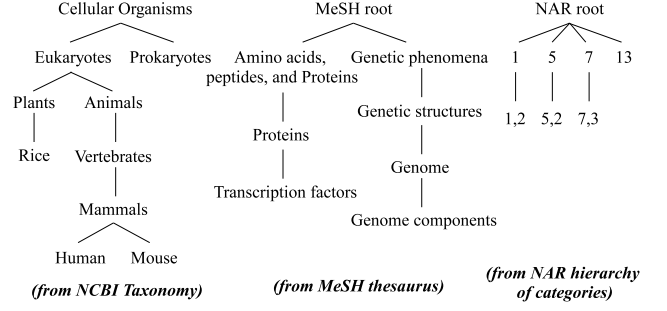


Figure 2. Fragments of MeSH thesaurus, NCBI taxonomy, and NAR hierarchy of categories expressing the semantic relationships between annotation metadata of DBs in Table 2.

Several approaches have been proposed to calculate the similarity between two or more terms in an ontology. A detailed comparison between the best known approaches can be found in [6] and in [9]. In the following we consider the Jaccard Coefficient together with hierarchies of terms to compute the similarity between any two terms of the same hierarchy. Therefore, semantic resources such as taxonomies and thesauri are seen as hierarchies of terms (rooted directed acyclic graphs) denoted by $\mathcal{H} = (\mathcal{V}, E)$, where $\mathcal{V}$ is the set of vertices and E is the set of edges. For any term $v \in \mathcal{V}$, $Ancestors(v)$ denotes the set of vertexes in the path from v to the root of $\mathcal{H}$ including both extremities. For example, $Ancestors(Rice) = \{Rice, Plants, Eukaryotes, Cellular Organisms\}$. Then we define the similarity between any two terms v_i and $v_j \in \mathcal{V}$ as follows:

$$sim(v_i, v_j) = \frac{|Ancestors(v_i) \cap Ancestors(v_j)|}{|Ancestors(v_i) \cup Ancestors(v_j)|}$$

For example, $sim(Rice, Plants) = 0.75$.

This similarity measure considers the proximity of terms in the hierarchy and their precision level (the depth in the hierarchy). Considering this measure, two terms v_i, v_j are said to be similar, and we write $v_i \simeq v_j$, if and only if $sim(v_i, v_j) \geq \theta$ where $\theta \in [0, 1]$ a chosen similarity threshold. For example, for $\theta = 0.6$, $Rice \simeq Plants$ whereas for $\theta = 0.8$, $Rice \not\simeq Plants$. It should be noticed here that the similarity operator “ $\simeq$ ” is not transitive.

The similarity threshold θ expresses the maximal variation allowed between two similar terms. The choice of θ reflects the precision requirements to be considered during the analysis process of data. Higher values of θ mean that only the closest terms will be considered as similar in the hierarchy whereas lower values of θ mean that more distant terms can be considered as similar. It is possible to choose a threshold for each hierarchy of terms separately. However, in the following, we consider the same threshold for all the hierarchies.

3.2 Attribute sharing between objects

In FCA, two objects g_i, g_j share an attribute m whenever they both have m . This definition is extended to MV contexts based on similarity as follows: g_i and g_j share m whenever $m(g_i) \simeq m(g_j)$. Objects g_i and g_j are said to share m w.r.t to the values $\{w_i, w_j\}$ written as $(m, \{w_i, w_j\})$. More generally, a set of objects A shares (m, W_m^A) where $W_m^A = \{m(g) \in W, g \in A\}$ whenever $\forall g_i, g_j \in A, m(g_i) \simeq m(g_j)$. In this case we say that A is valid w.r.t. m and

W_m^A is a set of similar values of m for A . In the same way, A shares a set of attributes B whenever A shares all attributes in B .

For example, for $\theta = 0.5$, *HSD* and *SpliceDB* share $(NCBI, \{Human, Mammals\})$ since $Human \simeq Mammals$ in the NCBI taxonomy. These two objects do not share the attribute *MeSH* since $Proteins \not\simeq Genome\ components$.

3.3 Resigning MV concepts and MV concept lattice

In FCA, formal concepts are defined as maximal sets of objects having in common maximal sets of attributes. Therefore, we need to compute (i) for any attribute its corresponding maximal valid set of objects and (ii) for any set of objects the maximal sets of similar values of their common attributes.

(i) *Maximal valid sets of objects*: Given an attribute m and a set of objects A valid w.r.t. m . We define the set of reachable objects from A w.r.t. m as follows:

$$\mathfrak{R}(A, m) = \{g_i \in G \mid m(g_i) \simeq m(g), \forall g \in A\}$$

$\mathfrak{R}(A, m)$ is the maximal set containing all objects similar to those in A w.r.t. m . This set may not be valid w.r.t. m because of the non transitivity of “ $\simeq$ ”. The maximal valid set of objects containing A is the subset of $\mathfrak{R}(A, m)$ obtained by removing from $\mathfrak{R}(A, m)$ all pairs of objects which do not share m . Formally this set is defined as follows:

$$\mathfrak{R}_v(A, m) = \mathfrak{R}(A, m) \setminus \{g_i, g_j \mid m(g_i) \not\simeq m(g_j)\}$$

More generally, the maximal valid set containing A with respect to $B \subseteq M$ is: $\mathfrak{R}_v(A, B) = \bigcap_{m \in B} \mathfrak{R}_v(A, m)$.

(ii) *Maximal sets of similar attribute values*: When $A \subseteq G$ shares an attribute $m \in M$, the largest set of similar values of m for A is the set of values of m for the objects in $\mathfrak{R}_v(A, m)$ obtained as follows:

$$\gamma(A, m) = \{m(g) \in W, g \in \mathfrak{R}_v(A, m)\}$$

Then we say that A shares $(m, \gamma(A, m))$. For example, for $\theta = 0.5$, $\{ExInt, CropNet, INE\}$ shares $(NCBI, \{Eukaryotes, Plants, Rice\})$.

Based on these maximal sets we define the following derivation operators for $A \subseteq G$ and $B \subseteq M \times \mathfrak{P}(W)$:

$$A^\uparrow = \{(m, \gamma(A, m)) \in M \times \mathfrak{P}(W) \mid \gamma(A, m) \neq \emptyset\}$$

$$B^\downarrow = \mathfrak{R}_v(\{g \in G \mid \forall (m, V_m) \in B, m(g) \simeq w, \forall w \in V_m\}, B)$$

$A^\uparrow$ is the maximal set of MV attributes shared by all objects in A and $B^\downarrow$ is the maximal set of objects sharing all MV attributes in B .

In the example of MV context in Table 2, and for $\theta = 0.5$, we have: $\{INE, CropNet\}^\uparrow = \{(NCBI, \{Plants, Rice\}), (NAR, \{13\})\}$ and $\{(NCBI, \{Plants, Rice\}), (NAR, \{13\})\}^\downarrow = \{INE, CropNet\}$.

Pairs of the form (m, V_m) are elements of $M \times \mathfrak{P}(W)$. Subsets of $M \times \mathfrak{P}(W)$ can be partially ordered by the means of the combination (product) of the inclusion relations on $\mathfrak{P}(M)$ and on $\mathfrak{P}(W)$. Given $B_1 = \{(m_1^i, V_1^i)\}_{i \in I}$ and $B_2 = \{(m_2^j, V_2^j)\}_{j \in J}$ in $\mathfrak{P}(M \times \mathfrak{P}(W))$, $B_1 \subseteq_\theta B_2$ if and only if $\forall (m_1^i, V_1^i) \in B_1, \exists (m_2^j, V_2^j) \in B_2$ such that $m_1^i = m_2^j$ and $V_2^j \subseteq V_1^i$. For example, for $\theta = 0.5$, $\{(NCBI, \{Human, Mammal, Vertebrates\})\} \subseteq_\theta \{(NCBI, \{Mammals, Vertebrates\}), (NAR, \{1.2\})\}$.

We can check that $A^{\uparrow\downarrow} = \mathfrak{R}_v(A, M)$ which is the maximal set of objects containing A and sharing the same MV attributes as A . Similarly, we can check that $B^{\downarrow\uparrow} = \{(m, \gamma(B^\downarrow, m))\}$ which is the maximal set of MV attributes containing B and shared by the same

objects as B . Then it follows that “ $\uparrow\downarrow$ ” and “ $\downarrow\uparrow$ ” are closure operators on $\mathfrak{P}(G)$ and $\mathfrak{P}(M \times \mathfrak{P}(W))$ respectively. The closed sets of objects and the closed sets of attributes formed by both operators respectively are dual and the operators $^\uparrow$ and $^\downarrow$ form a Galois connection between $(\mathfrak{P}(G), \subseteq)$ and $(\mathfrak{P}(M \times \mathfrak{P}(W)), \subseteq_\theta)$.

Based on the derivation operators defined above we define MV concepts as (A, B) where $A \subseteq G$ and $B \subseteq M \times \mathfrak{P}(W)$ such that $A^\uparrow = B$ and $B^\downarrow = A$. A and B are respectively the extent and the intent of (A, B) . For $\theta = 0.5$, $\{(INE, CropNet), (NCBI, \{Plants, Rice\}), (NAR, \{13\})\}$ is an example of MV concept.

The MV concepts are ordered as follows:

$$(A_1, B_1) \leq_\theta (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 (\Leftrightarrow B_2 \subseteq_\theta B_1)$$

(A_1, B_1) is a subconcept of (A_2, B_2) and (A_2, B_2) is a superconcept of (A_1, B_1) . The set of all MV concepts of a MV context ordered by $\leq_\theta$ is denoted by $\mathfrak{B}_\theta(G, M, W, I)$ and called the MV concept lattice of (G, M, W, I) for the similarity threshold θ . The MV concept lattice of the MV context represented in Table 2 for $\theta = 0.5$ is shown in Figure 3.

4 Exploration of Complex Data in Biology

4.1 Multi-grained classification

The MV concept lattice design can be used as a “tuning” classification tool, from a coarse to a fine-grained classification, depending on the value of θ . The variation of θ has a direct influence on attribute sharing, i.e. restricting or relaxing sharing. Decreasing θ corresponds to enlarging similarity of attribute values and implies the formation of MV concepts whose intents contain MV attributes with large sets of similar values. Conversely, increasing θ corresponds to a restriction of similarity and implies the formation of MV concepts whose intents contain MV attributes with reduced sets of similar values. As a consequence, MV concept lattice design can be viewed as a multi-grained classification process, depending on the values of θ (from 0 to 1). This feature is particularly interesting for large contexts. A first general MV concept lattice can be built (choosing a low value of θ) to provide a coarse classification. Then, according to the desired precision, more specific MV concept lattices can be built by increasing values of θ .

The two extreme MV concept lattices, for $\theta = 0$ and for $\theta = 1$, correspond to the concept lattices of two particular conceptual scaling. The upper lattice is obtained when $\theta = 0$ which means that all the values of an attribute are similar. This case is equivalent to conceptual scaling which consists in replacing each non empty table entry by a cross ($\times$) in the MV context. The lower lattice is obtained when $\theta = 1$ (i.e. no similarity between attribute values). This case is equivalent to *plain scaling* [8].

4.2 Dynamic navigation

Once a satisfactory classification grain is reached, the corresponding MV concept lattice can be used for navigation support. The navigation process is performed in the same way as in standard FCA [3], starting from a given concept and following generalization or specialization links in the MV concept lattice.

An additional interesting feature offered by MV concept lattices is the possibility of zooming w.r.t. the concept precision while exploring a MV concept lattice. The idea consists in dynamically moving from a MV concept lattice to another depending on a higher or lower grain level. More precisely, while navigating in a MV concept lattice,

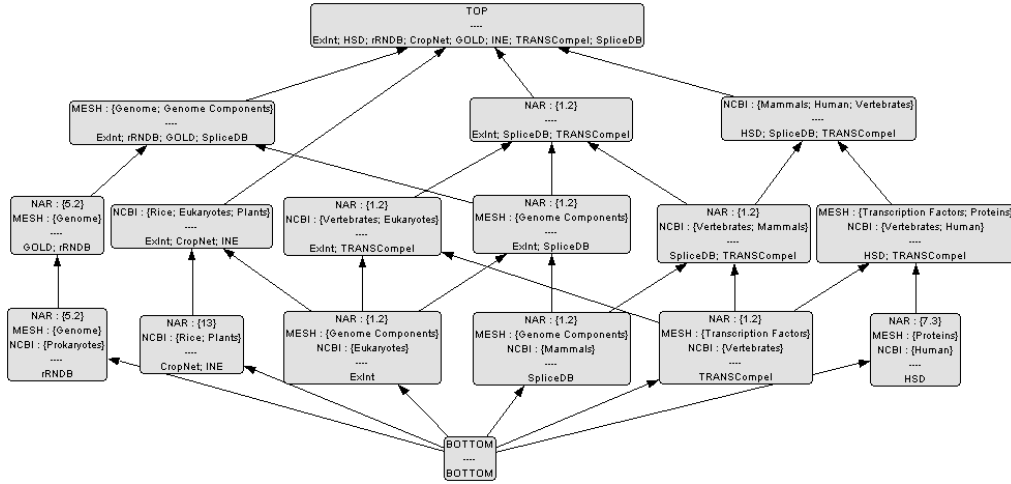


Figure 3. The MV concept lattice $\mathfrak{B}_{0.5}(G, M, W, I)$.

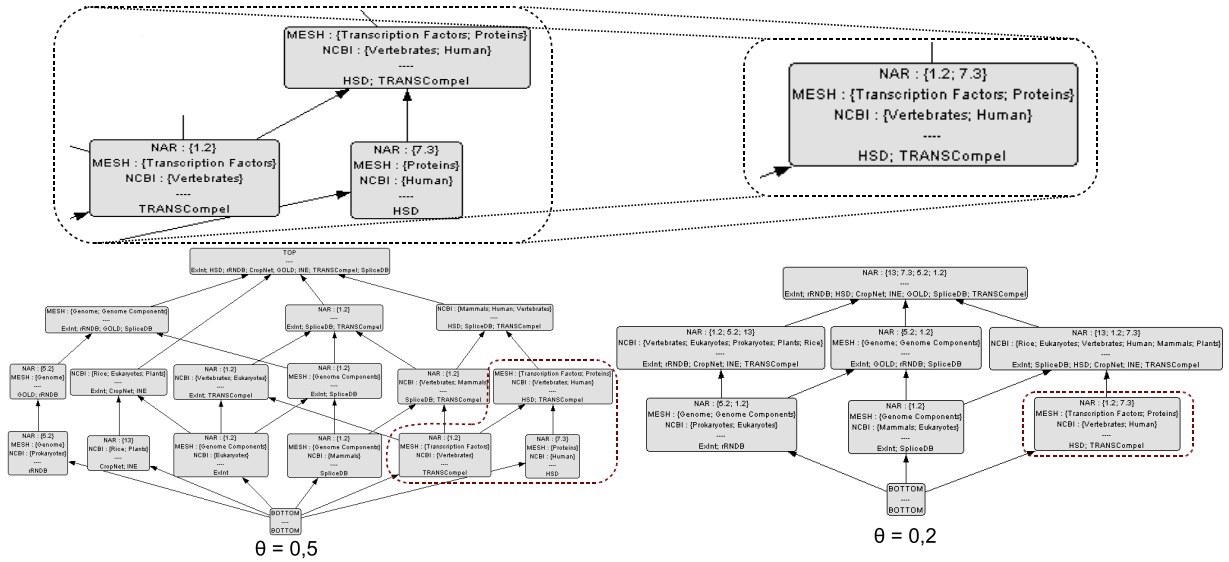


Figure 4. An example of MV concept zooming in a dynamic navigation process.

one may need to “zoom-in” to get detailed object organization without being obliged to restart the navigation process and considering a more specific MV concept lattice. “Zoom-in” a MV concept, say C , consists in showing the part corresponding to C in a more specific MV concept lattice. In another case the need may be to “zoom-out” a MV concept C to see the variation of the extent of C when the similarity threshold is increased. The illustration of MV concept zooming is given in Figure 4. The considered MV concept lattices are $\mathfrak{B}_{0.5}(G, M, W, I)$ (left) and $\mathfrak{B}_{0.2}(G, M, W, I)$ (right). Zoom-in and zoom-out examples are shown respectively from right to left and from left to right.

As the zooming operation needs to dynamically move from a MV concept lattice to another, it requires the computation and the

cache of several MV concept lattices depending on different similarity thresholds. In practice it may be difficult to ensure an efficient management especially for large MV contexts. One way to avoid the problem consists in dynamically designing a MV concept lattice for a reduced MV context, involving only the objects and attributes under study.

4.3 Experimentation

The approach proposed in this paper is implemented in the *SimBA* system (*Similarity-Based concept Analysis*⁸). The MV lattice construction is based on an adapted levelwise algorithm based on a pair-

⁸ <http://www.loria.fr/~messai/en/SimBA.htm>

wise intersection of sets of MV attributes which takes into account similarities between attribute values.

Experimentations of SimBA were mainly performed on two MV contexts extracted from BioRegistry repository. The first MV context contains 48 DBs, 5 MV attributes related to the DBs subjects and corresponding to 5 different trees in the MeSH thesaurus structures (namely “Organisms [B]”, “Chemicals and Drugs [D]”, “Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]”, “Biological Sciences [G]” and “Natural Sciences [H]”), and 39 different MeSH terms as attribute values distributed in the context with a density of 0.34. The second MV context contains all the DBs annotated with at least one MeSH term. This MV context contains 585 DBs, 9 attributes corresponding to 9 different trees in the MeSH thesaurus and 410 different MeSH terms distributed in the context with a density of 0.3. In this MV context a table entry can contain more than one MeSH term. The average number of MeSH terms per non empty table entry is 1.7.

The similarity between two arbitrary MeSH terms is computed w.r.t. the MeSH thesaurus in the same way as introduced a Section 3.1. The position of any MeSH term in the corresponding MeSH tree and the set of its ancestors are immediately given by its associated tree number in the MeSH thesaurus which facilitates similarity computing. For example, the tree numbers of Mammals and Vertebrates are “B01.150.900.649” and “B01.150.900” respectively. These two terms have 3 common nodes in their path to the root (B01, B01.150, and B01.150.900). The similarity between these two terms is 0.75.

The evolution of MV concept lattices in terms of number of MV concepts with respect to the similarity threshold for the two considered MV contexts as well as for the example given in Table 2 is shown on Figure 5. Although the MV contexts are different from each other, the overall evolution of MV concepts in their corresponding MV concept lattices has almost the same behavior. These results confirm the usefulness of MV concept lattices for progressive clustering as well as for dynamic exploration of complex data.

The navigation into the obtained MV concept lattices is supported by the visualization module which generates the Hasse diagrams of the MV lattices based on prefuse API⁹. The Hasse diagrams shown in Figures 3 and 4 are samples visualized with SimBA visualization module. The navigation process is facilitated by visual aspects such that highlighting concept neighbors, focusing on parts of lattices rather than on the whole lattice, and showing different information about concepts while visualizing lattices (intents only, extents only, supports of the intents).

5 Conclusion

In this paper, we presented an approach adapting FCA to take advantage of available semantic resources to deal with complex data represented as MV contexts. The basic idea is that two objects share an attribute whenever the values taken by this attribute for these objects are similar with respect to a considered semantic resource and a similarity threshold. We defined a Galois connection which considers similarities between attribute values. Then we derived MV concepts and MV concept lattices. The resulting MV concept lattices are used to facilitate the exploration of complex data sets in a progressive and dynamic manner. The approach was implemented and successfully applied for browsing the contents of the biological databases repository BioRegistry. In the future, we will study the potential of the approach to mining complex data based on semantic relationships.

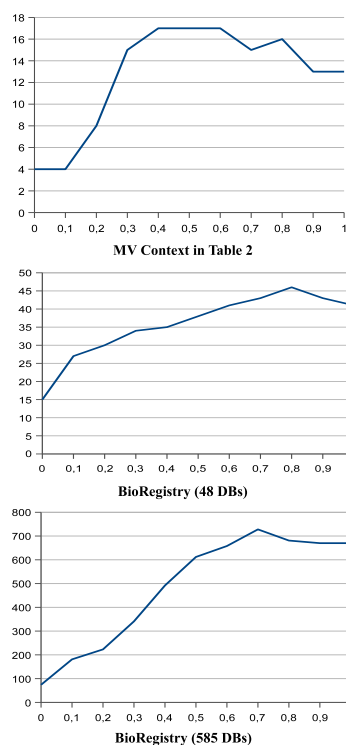


Figure 5. Evolution of the number of MV concepts in the MV concept lattices.

REFERENCES

- [1] R. Belohlavek and V. Vychodil, ‘What is a fuzzy concept lattice?’, in *CLA’05, Olomouc, Czech Republic*, pp. 34–45, (2005).
- [2] P. Brito and G. Poliaillon, ‘Structuring Probabilistic Data by Galois Lattices’, *Mathematics and Social Sciences*, **169**(1), 77–104, (2005).
- [3] C. Carpineto and G. Romano, ‘A lattice conceptual clustering system and its application to browsing retrieval’, *Machine Learning*, **24**(2), 95–122, (1996).
- [4] M-D. Devignes, P. Franiatte, N. Messai, A. Napoli, and M. Smail-Tabbone, ‘BioRegistry: Automatic Extraction of Metadata for Biological Database Retrieval and Discovery’, in *RED’08, Joint to iiWAS’08*, pp. 456 – 461, Linz Austria. ACM.
- [5] S. Ferré and Olivier Ridoux, ‘A logical generalization of formal concept analysis’, in *ICCS’00*, volume 1867 of *LNCS*, pp. 371–384, (2000).
- [6] P. Ganesan, H. Garcia-Molina, and J. Widom, ‘Exploiting hierarchical domain structure to compute similarity’, *ACM Transactions on Information Systems (TOIS)*, **21**(1), 64–93, (2003).
- [7] B. Ganter and S. Kuznetsov, ‘Pattern structures and their projections’, in *ICCS’01, Stanford, CA, USA*, volume 2120 of *LNCS*, pp. 129–142.
- [8] B. Ganter and R. Wille, *Formal Concept Analysis*, Springer, mathematical foundations edn., 1999.
- [9] A. Hliaoutakis, G. Varelakis, E. Voutsakis, E. G. M. Petrakis, and E. E. Milios, ‘Information retrieval by semantic similarity’, *International Journal on Semantic Web and Information Systems*, **2**(3), 55–73, (2006).
- [10] N. Messai, M-D. Devignes, A. Napoli, and M. Smail-Tabbone, ‘Querying a bioinformatic data sources registry with concept lattices’, in *ICCS’05, Kassel, Germany*, volume 3596 of *LNAI*, pp. 323–336.
- [11] N. Messai, M-D. Devignes, A. Napoli, and M. Smail-Tabbone, ‘Extending attribute dependencies for lattice-based querying and navigation’, in *ICCS’08, Toulouse, France, LNAI*. Springer, (2008).
- [12] N. Messai, M-D. Devignes, A. Napoli, and M. Smail-Tabbone, ‘Many-valued concept lattices for conceptual clustering and information retrieval’, in *ECAI’08, Patras, Greece*, (2008).

⁹ <http://prefuse.org/>