Galois Connections in Data Analysis: Contributions from the Soviet Era and Modern Russian Research

Sergei O. Kuznetsov

All-Russia Institute for Scientific and Technical Information (VINITI)
Usievicha 20, 125190 Moscow, Russia
serge@viniti.ru

Abstract A retrospective survey of several research directions at the All-Soviet (now All-Russia) Institute for Scientific and Technical Information (VINITI), as well as research represented in several VINITI editions, is proposed. In a number of papers of the 1970-1980s, taxonomies (classifications) were naturally considered as lattices. Several problems of classification required consideration of tolerance relations as a model of similarity of objects. Such relations define symmetric formal contexts. A JSM-method of inductive plausible reasoning, which has been developed at VINITI since the early 1980s, is considered in terms of Galois connections and concept lattices. Mathematical research around the JSM-method and its applications is discussed.

1 Introduction

The research on Galois connection in the classification school at the All-Soviet (now All-Russia) Institute for Scientific and Technical Information (VINITI), Moscow, was first motivated by problems of classification and storage of documents, which needed formal models of similarity of objects. Later motivations came from problems of data analysis in various applied domains and their solution by means of the JSM-method of hypothesis generation.

In this article we give a review of the research activity in VINITI and/or in its journals, mainly *Nauchno-Tekhnicheskaya Informatsiya* (NTI), series[1], translated to English by Allerton Press under the name *Automated Documentation and Mathematical Linguistics*, and also in *Semiotika i Informatika* (Semiotics and Computer Science), *Itogi Nauki i Tekhniki* (Reviews in Science and Technology).

Around the mid 1960s, Yulii A. Shreider (1927-1998), one of the leading researchers of VINITI, considered the problem of automatic classification of documents and their retrieval by means of a model consisting of a triple of sets (M, L, f), where M is a set of objects (documents), L is a set of attributes and $f: M \to \mathcal{P}(\mathcal{L})$ is a mapping taking each object to a set attributes from L [77]. Similarity of documents x and y was defined by nonemptyness of the set of their common attributes: $f(x) \cap f(y) \neq \emptyset$. Defined in this way similarity is reflexive and symmetric, i.e., similarity is a tolerance relation on the set of objects.

Shreider mentioned the relevance of lattices to problems of classification and mathematical retrieval in his early paper [77], where he also cited the work of Soergel [82] on this issue. In [80] Shreider wrote about classifications of objects described by attributes, where each classification is given by an idempotent commutative semigroup (which is actually a semilattice) uniquely specified by bases (actually, by sets of irreducible elements). Implication between single attributes, analogous

[1] with Prof. Ruggero S. Gilyarevsky being the editor-in-chief for four decades, first *de facto*, and later also *de jure*

to that in Formal Concept Analysis (FCA), was defined. Together with Sergei V. Meien, a biological methodologist from St. Petersburg, he wrote on the duality of taxonomies and meronomies (the latter term, denoting a hierarchy of parts, was coined by Meien from the Greek word $\mu\epsilon\rho o\zeta$, part) [59] [2]. This was almost like the starting point of FCA [89], however, no systematic theory appeared. Two directions of thought, the one of them related to the (semi)lattice nature of classification and the other one, which considered tolerance relations on the set of objects and their classes given by Galois correspondences, developed independently. An analogue of concept lattice theory appeared later, in mid 1980s, in works by O.M. Polyakov and V.V. Dunaev [72, 73, 14, 74].

To provide a general framework for the overview of research in different groups, we will use the standard definitions of Formal Concept Analysis [89, 27], which we will briefly recall below.

Let G and M be sets and $I \subseteq G \times M$ be a relation. The elements of G and M are called the sets of objects and attributes, respectively, and gIm (i.e., $(g, m) \in I$) is read: the object g has the attribute m. The triple $\mathbb{K} = (G, M, I)$ is called *formal context.* The *derivation operators*, defined for any $A \subseteq G$ and $B \subseteq M$ by

$$A^I := \{m \in M \mid gIm \text{ for all } g \in A\}, \quad B^I := \{g \in G \mid gIm \text{ for all } m \in B\}$$

induce a *Galois connection* between the ordered powersets $(\mathcal{P}(G), \subseteq)$ and $(\mathcal{P}(M), \subseteq)$. In the case of a fixed relation I one usually writes $(\cdot)'$ instead of $(\cdot)^I$. Any pair of sets (A, B) such that $A \subseteq G$, $B \subseteq M$, $A' = B$ and $B' = A$ is called a *formal concept* of the context $\mathbb{K}$ with *(formal) extent* A and *(formal) intent* B. The set of all formal concepts of a formal context $\mathbb{K}$ forms a complete lattice called *(formal) concept lattice* $\mathfrak{B}(\mathbb{K})$. Moreover, for an arbitrary complete lattice L, there is a concept lattice isomorphic to it. For $A, C \subseteq M$ the *implication* $A \rightarrow C$ holds if $A' \subseteq C'$ (or $C \subseteq A''$), i.e., if all objects from G that have the set of attributes A also have the set of attributes C.

The issues around Galois connections and functional dependencies motivated many other researchers. For example, in [11] the equivalence of the category of functional dependencies to the category of closure systems (Moore families) [6] was considered. In [84] the author studies implications between binary attributes. The definition of implication was not extended to sets of attributes as it had been done previously in FCA. The author considers the equivalence relation on objects (described by same sets of attributes) and studies the cases where the order defined by implications induces a Boolean algebra, as well as the possibility of embedding elements of a binary relation with implication in Boolean algebras by mappings called strict homomorphisms by the author (more known under the name *order embeddings*): for two ordered sets A and B a mapping $f: A \rightarrow B$ is called a *strict homomorphism* if $f(a_1) \leq f(a_2)$ iff $a_1 \leq a_2$ for any $a_1, a_2 \in A$.

The rest of the paper is organized as follows. In the second section we consider models of taxonomies and their relation to certain type of dependencies in databases. In the third section we consider models of similarity based on tolerance relations, classes of tolerance, and their relation to FCA. In the fourth section we give a review of the research around the JSM-method of hypothesis generation, a machine learning method naturally formalized in terms of Galois connections and FCA.

2 Taxonomies and Dependencies

After [80] the next step in the development of Shreider's classification model was made in [76], where objects are described by attributes, taxonomies are defined as

[2] See also www.ento.vt.edu/~sharov/biosem/shreidr/shreidr.html

refining sequences of coverings of object sets, meronomies are defined as refining sequences of coverings of attribute sets (see exact definitions below). An archetype was defined as a description common to all objects from a taxon (i.e., a member of classification). The authors relate this construction to the notion of a concept, its intent and extent, noticing that the size of the former decreases with the growth of size of the latter. However, no further mathematical theory was proposed. A theory that would pass completely to these methodological considerations is exactly that of FCA. Together with [59], where the duality of taxonomies and meronomies was underlined, the paper [76] is actually a prolegomena to FCA. So it is not surprising that a counterpart of FCA notions appeared later (and few years later than similar work in French and German schools) in the classification theory by Polyakov and Dunaev [72, 73, 14, 74].

They started from object-attribute representation, defined Galois correspondence as it is done in FCA and obtained two antiisomorphic complete lattices on sets of objects (called taxonomy) and sets of attributes (called meronomy). They stated that both lattices can be generated by the sets of corresponding irreducibles. Polyakov and Dunaev also considered relations between sets of objects, i.e., of the form $I \subseteq G \times G$, since to their minds, relations between objects induce taxonomies (see below) and often come before attributes. Moreover, attributes often result from the observation of relations between objects, for example the Mendeleev periodic table of chemical elements, discussed in [14], resulted from ordering of the objects (chemical elements) with respect to their atomic weights. This ordering motivated further study of properties of classes of chemical elements.

Several results from FCA were repeated by Dunaev and Polyakov: Representing concept lattices by products of concept lattices of subcontexts ([73, Theorem 2]) which allowed them to draw lattices in the way as it is done with nested line diagrams in FCA; describing morphisms of concept lattices specified by subsets of the set of all attributes ([73, Theorem 3]).

An aspect of their research that has not been previously covered by research in FCA is related to the study of multivalued and mutual dependencies, which are generalizations of functional dependencies. They showed how dependencies of this kind allow for decomposition of taxonomies into products. Below we present some definitions and results from [74].

Let $D(U)$ be a finite set of objects (U is the *name* of this set). A *taxonomy* $\mathcal{T}$ consisting of a system of *taxons*,

which are subsets $D(U)$, is given as follows: $D(U)$ is a taxon; and if $T_1 \in \mathcal{T}$, $T_2 \in \mathcal{T}$ are taxons, then $T_1 \cap T_2$ is a taxon.

Obviously, taxonomy defined in this way is a *closure system* [27] (or, equivalently, a *Moore family* [6]) and the set of all taxons induces a lattice. In terms of FCA this is the lattice of extents. When objects are described in terms of attributes, then the dual lattice, called *meronomy*, on closed sets of attribute arises. In terms of FCA this is the lattice of intents.

A *product of taxonomies* [74] $\mathcal{T}_1$ and $\mathcal{T}_2$ on the same set of objects, denoted by $\mathcal{T}_1 \cdot \mathcal{T}_2$, is a taxonomy such that $T \in \mathcal{T}_1 \cdot \mathcal{T}_2$ iff $T = T_1 \cap T_2$ for some $T_1 \in \mathcal{T}_1$, $T_2 \in \mathcal{T}_2$. The order on taxonomies $\leq$ defined as $\mathcal{T}_1 \leq \mathcal{T}_2$ iff $\mathcal{T}_1 \cdot \mathcal{T}_2 = \mathcal{T}_1$ induces a lattice on the set K of all taxonomies of the set $D(U)$. This order on taxonomies is obviously related to refinement order on closure systems [27]

An *attribute* X is given in [74] as a pair $\langle D(X), sX \rangle$, where $D(X)$ is the set of attribute values, sX is the "object-attribute value" relation ($sX \subseteq D(U) \times D(X)$). In terms of FCA, X is a *many-valued attribute* with the set of values $D(X)$. The relation sX defines its *scaling* already at the many-valued level: in contrast to FCA, there is no implicit dependencies of values that are specified by the choice of a scaling, i.e., a method of reduction to one-valued attributes. All possible (object, attribute value) pairs are given explicitly.

Criteria of decomposition of a taxonomy lattice resulting from a set of many-valued attributes into products of taxonomy lattices arising from single attributes are given in [74]. These criteria were given in terms of multi-valued and mutual dependencies.

Recall from [57] that a *functional dependency* $U \to X$ holds if for any $u \in D(U)$, $x, \tilde{x} \in D(X)$ the relations $(u, x) \in sX$, $(u, \tilde{x}) \in sX$ imply $x = \tilde{x}$. *Multivalued dependency* $U \twoheadrightarrow X$ holds if for any $u \in U$ the facts $(u, x, y) \in sV$ and $(u, x', y') \in sV$ imply $(u, x, y') \in sV$. The functional dependency $U \to X$ obviously implies $U \twoheadrightarrow X$. *Mutual dependency* [66] $U \simeq X$ holds if for every $u \in D(U)$ the facts $(u, x, y) \in sV$, $(u, x', y') \in sV$ and $(x, y') \in sV[XY]$ imply $(u, x, y') \in sV$. Multivalued dependency is obviously a particular type of mutual dependency.

For two attributes X and Y their (*natural*) *join* $sX \bowtie sY$ is defined as follows: $(u, x, y) \in sX \bowtie sY$ iff $(u, x) \in sX$ and $(u, y) \in sY$. By a theorem from [57], the decomposition $sV = \bowtie_{i=1}^{n} sX_i$ is possible iff there exists a set of multivalued dependencies $U \twoheadrightarrow X_i$, $i = \{1, \ldots, n\}$.

The decomposition $sV = sX \bowtie sY \bowtie sV[XY]$ is possible iff there exists mutual dependency $U \simeq X$ [66]. The following propositions from [74] give criteria of decomposition of a taxonomy lattice arising from the whole set of attributes into products of taxonomy lattices arising from single attributes.

Proposition 1 *Let $sX_i = sV[UX_i]$, where $V = X_1 \ldots X_n$, $X_i \cap X_j = \emptyset$ for any $i \neq j; i, j = 1, \ldots, n$. Mutual dependencies $U \simeq X_i$ hold for all $i = \{1, \ldots, n\}$ iff for any $(x_1, \ldots, x_n) \in sV[V]$ the relation*

$$\{u \in D(U) \mid (u, x_1, \ldots, x_n) \in sV\} = \bigcap_{i=1}^{n} \{u \in D(U) \mid (u, x_i) \in sX_i\}$$

holds.

In terms of taxonomies this result can be recast in the following form.

Proposition 2 *Let $sX_i = sV[UX_i]$, where $V = X_1 \ldots X_n$, $X_i \cap X_j = \emptyset$ for all $i \neq j$, $i, j = 1, \ldots, n$. If mutual dependencies $U \simeq X_i$ hold for all $i = \{1, \ldots, n\}$, then $\mathcal{T}(V) \subseteq \mathcal{T}(X_1) \cdot \ldots \cdot \mathcal{T}(X_n)$.*

As a corollary one has the following

Proposition 3 *If $sX_i = sV[UX_i]$, where $V = X_1 \ldots X_n$, $X_i \cap X_j = \emptyset$ for $i \neq j$ and multivalued dependencies $U \twoheadrightarrow X_i$ hold for all $i = \{1, \ldots, n\}$, then $\mathcal{T}(V) = \mathcal{T}(X_1) \cdot \ldots \cdot \mathcal{T}(X_n)$.*

In fact, if multivalued dependencies $U \twoheadrightarrow X_i$ hold for all $i = \{1, \ldots, n\}$, then $\mathcal{T}(V) \supseteq \mathcal{T}(X_1) \cdot \ldots \cdot \mathcal{T}(X_n)$.

3 Tolerance relation: symmetric contexts

In the works of V.Ya. Gusakov and S.M. Gusakova (Yakubovich) classes of a tolerance relation were studied. This study of tolerance was motivated first by modeling similarity of documents in document retrieval systems [91, 92, 34, 35].

It was Zeeman [98] who proposed first to formalize similarity as a tolerance (reflexive and symmetric) relation. The relation of similarity, being naturally reflexive and symmetric, should not be transitive: e.g., children are often similar to both their parents, the latter being very different. Although some authors, like Tversky [85] doubt that similarity is naturally symmetric and reflexive, this seems to be adequate to model similarity between documents.

Definition 3.1 For a set G a binary relation $T \subseteq G \times G$ is called *tolerance* if
(1) $\forall x \in G\ xTx$ (reflexivity)
(2) $\forall x, y \in G\ xTy \to yTx$ (symmetry)
A set G with tolerance T is called the *space of tolerance* and denoted by G_T.

Definition 3.2 A subset $K \subseteq G$ is called a *class of tolerance* if
(1) $\forall x, y \in K,\ xTy,$
(2) $\forall z \notin K\ \exists u \in K \neg(zTu)$
An arbitrary subset of a class of tolerance is called a preclass.

Definition 3.3 A set $\mathcal{A} = \{A_j\}_{j \in J}$ of preclasses is called a *system of preclasses preserving* T if

$$T = \bigcup_{j \in J} A_j \times A_j.$$

The most important preserving system of preclasses for the tolerance T is the system of all classes, which is denoted by $\mathcal{K}(G_T)$.

Tolerance classes defined by a tolerance relation are cliques (inclusion-maximal complete subgraphs) of the graph (G, T). On the other hand, a tolerance relation can be considered as an origin of formal context representation. First, some objects are observed to be pairwise similar. Then all pairs of the tolerance relation, and further on, the set all of maximal classes of similarity (classes of tolerance) is constructed. Eventually, the classes are given names, which are further used as attributes that describe objects.

By symmetry of the tolerance relation T, the Galois connection associated with the context (G, G, T) is given by a single mapping $(\cdot)^T$, where x^T is a set of all elements from G tolerant to x and X^T is the set of all elements from G tolerant to each $x \in X$.

Let $\mathcal{L}$ be a system of preclasses preserving tolerance T on the set G, then the context $(G, \mathcal{L}, \mathcal{I})$ is defined as usual: for an object $g \in G$ and a preclass $L \in \mathcal{L}$ one has gIL iff $g \in L$. The Galois connection given by the derivation operator $(\cdot)^I$ is called the *Galois connection that agrees with the tolerance* T *by the preserving system* $\mathcal{L}$.

The following relation from [37] recast in FCA terms gives so called *canonical representation of similarity.*

Proposition 4 *Let* G *be a set and* $T \subseteq G \times G$ *a tolerance relation and let* $\mathcal{A}$ *be a system of preclasses preserving* T. *Then* $(G, \mathcal{A}, \in)$ *is a formal context satisfying*

$$(g, h) \in T \iff g' \cap h' \neq \emptyset \text{ for all } g, h \in G.$$

Conversely, if (G, M, I) *is a context with* $g' \neq \emptyset$ *for all* $g \in G$, *then* $T := \{(g, h) \mid g' \cap h' \neq \emptyset\}$ *is a tolerance relation and* $\mathcal{A} := \{m' \mid m \in M\}$ *is a system of preclasses preserving* T.

Thus, each tolerance can be obtained from some formal context and in turn, an arbitrary tolerance gives rise to a formal context: Starting from a tolerance relation, one finds classes of tolerance, which, after being named, can be used further used as attributes.

The results obtained for tolerances were partially extended to the case of n-ary relations in [36], where the notions of n-ary tolerance relation and the corresponding definitions of a class, preclass, preserving system of preclasses, and basis are introduced.

In the 1980s a motivation for the further study of tolerance [36–39] came from the theory of plausible reasoning [17] based on similarity operation (see next section about the JSM-method). In [36, 37] the relationship between so-called global and local similarities was studied.

In [36] the following two definitions of similarity arising from formal contexts (called *karta*, map) were considered.

Definition 3.5. Objects $g_1, \dots, g_n$ are *n-locally similar* in the context $\mathbb{K} = (G, M, I)$ if $g_i \in G$ $(i = 1, \dots, n)$ and $\{g_1, \dots, g_n\}^I \neq \emptyset$.

In terms of FCA, locally similar objects are exactly those that occur together in a formal extent of the formal context $\mathbb{K} = (G, M, I)$. The sets of n-locally similar objects induce a tolerance on $G \times G$, these sets being preclasses of the tolerance.

Definition 3.6. Objects $g_1, \dots, g_k$ are *globally similar* in the context $\mathbb{K} = (G, M, I)$ if $m' = \{g_1, \dots, g_k\}$ for some $m \in M$.

So, a set of globally similar objects is an attribute extent of the formal context $\mathbb{K} = (G, M, I)$. Global similarity is not a relation, because it involves tuples of varying length. A global similarity on the set G can be represented by a covering $\pi = \{\pi_j\}_{j \in J}$ of G, where each π_j is a set of globally similar elements. A global similarity $\langle G, \pi \rangle$ *is represented by n-local similarity* if $\pi = \mathcal{K}$, where $\mathcal{K}$ is the set of all classes of the tolerance T induced by the n-local similarity [36].

To satisfy this condition, the global similarity should not give rise to new classes of the tolerance induced by the n-local similarity, and the maximal sets of n-locally similar objects should be globally similar.

Example 1. Consider the following context (G, M, I) with $G = \{g_1 =$ **whale**, $g_2 =$ **duck**, $g_3 =$ **robin**, $g_4 =$ **tree frog**$\}$, $M = \{$**warm blood**, **can fly**, **can swim**$\}$.

	warm blood	**can swim**	**can fly**
whale	×	×	
duck	×	×	×
robin	×		×
tree frog		×	×

Consider the global similarity $\pi = \{\pi_1, \pi_2, \pi_3\}$, where $\pi_1 = \{g_1, g_2, g_3\}$, $\pi_2 = \{g_1, g_2, g_4\}$, $\pi_3 = \{g_2, g_3, g_4\}$. This global similarity generates the binary local similarity R_1 with the set of classes $K_{R_1} = \{\{g_1, g_2, g_3, g_4\}\}$ and ternary similarity R_2 with the set of classes $K_{R_2} = \{\{g_1, g_2, g_3\}, \{g_1, g_2, g_4\}, \{g_2, g_3, g_4\}\}$. Obviously, $K_{R_1} \neq \pi$ and, therefore, the global similarity is not representable by the binary local similarity R_1, whereas $K_{R_2} = \pi$ and the global similarity is representable by the ternary local similarity K_{R_2}.

In [92] the notion of a conjugate tolerance to a tolerance $T \subseteq G \times G$ was defined as a tolerance relation on the set $\mathcal{K}(G_T)$ of classes of T: a pair of classes belong to the conjugate tolerance if they are not disjoint. The relation between conjugated spaces to the initial tolerances, as well as sequences of conjugations, were studied in [92].

A further generalization of the similarity models was nonsymmetric similarity relation considered in [38], where criteria for canonical representation of nonsymmetric relation was given in terms of preclasses preserving relation.

4 JSM-method

The initial motivation for the first version of the JSM-method proposed by Viktor K. Finn in late 1970s was the intention to describe induction in purely deductive form and give at least partial justification of induction. The method was named in honor of the English philosopher John Stuart Mill, who proposed schemes of inductive reasoning in the 19th century. Most well-known are the first and second canons of inductive logic [63].

The first canon, also called *Method of Agreement*, was formulated as follows: "If two or more instances of the phenomenon under investigation have only one circumstance in common, ... [it] is the cause (or effect) of the given phenomenon."

The second canon or *Method of Difference* sounds like: "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon."

To formalize the Mill's methods, Finn and colleagues used the principle of two-layered logics of Dmitrii A. Bochvar [3] [10]: Several truth types, including "empirical contradiction" between generalizations of data, were allowed at the internal logical level and classical logical values are used at the external level. More precisely, the JSM-method was described by means of a many-valued many-sorted extension of the First-Order Predicate Logic

with quantifiers over tuples of variable length (this logic is a proper part of the second order logic, often called *weak second order logic*).

The motivation for the use of quantifiers over tuples of variable length is as follows: Induction is based on the observation of similarity of objects. Since the number of objects with a particular similarity is not known in advance, quantification over tuples of variable length is necessary in the case of infinite number of objects, to express their similarity.

For example, the Mill's Method of Agreement is the formalized by the following predicate $\mathcal{M}^+_{a,n}(V,W)$ (some other Mill's canons, e.g., the method of differences, as well as new methods of inductive reasoning were described in similar way):

$$\mathcal{M}^+_{a,n}(V,W) := \exists k\, \widetilde{\mathcal{M}}^+_{a,n}(V,W,k),$$

$$\widetilde{\mathcal{M}}^+_{a,n}(V,W,k) := \exists Z_1 \ldots \exists Z_k\, \exists U_1 \ldots U_k\, (\mathop{\&}_{i=1}^{k} J_{\langle 1,n\rangle}(Z_i \Rightarrow_1 U_i)\, \&$$

$$\&\, \forall U(J_{\langle 1,n\rangle}(Z_i \Rightarrow_1 U) \rightarrow \mathcal{U} \subset U_i))\, \&(Z_1 \cap \ldots \cap Z_k) = V\, \&\, V \neq \emptyset\, \&\, W \neq \emptyset\, \&$$

$$\&\, \forall i\, \forall j((i \neq j)\, \&\, 1 \leq i,j \leq k) \rightarrow Z_i \neq Z_j)\, \&\, \forall X\, \forall Y((J_{\langle 1,n\rangle}(X \Rightarrow_1 Y)\, \&$$

$$\&\, \forall \mathcal{U}(J_{\langle 1,n\rangle}(X \Rightarrow_1 U) \rightarrow \mathcal{U} \subseteq Y)\, \&\, \&\, V \subset X) \rightarrow (W \subseteq Y\, \&(\bigvee_{i=1}^{k}(X = Z_i))))\, \&\, k \geq 2).$$

Here, $J_{\langle \varepsilon,n\rangle}$ is a Rosser-Turquette operator taking formulas of many-valued "internal" logic to two classical logic values: $\varepsilon \in \{-1,0,1,\tau\}$, -1 denotes "empirically false," 1 denotes "empirically true," 0 denotes "empirical contradiction", and τ denotes "empirically undeterminate". n denotes the number of iteration step, which is an important feature of the JSM-method.

From the agreement (similarity) predicate one can construct some other predicates imposing additional conditions, like the following one:

$$\forall X\, \forall Y((V \subseteq X\, \&\, W \subseteq Y) \rightarrow (J_{\langle 1,n\rangle}(X \Rightarrow_1 Y) \vee (J_{\langle \tau,n\rangle}(X \Rightarrow_1 U)),$$

which is called "no counterexample" or "counterexample forbidding". Additional conditions (conjunctively added to the main agreement predicate) make the "lattice of methods".

[3] In his 1938 paper Bochvar proposed one of the first many-valued logics for the treatment of the liar paradox, where there were two types of logical values: the inner values were "true", "false" and "contradiction", whereas the external values were classical "true" and "false".

Upon construction of all pairs (V, W) by a certain method, one uses them for classification of new examples. When the latter are classified, they are added (now, as new positive or negative examples) to the initial sets of positive and negative examples, and the whole procedure is iterated.

Algebraic redefinitions of inductive methods started from the observation that the agreement predicate defines a Moore family w.r.t.

$\bigcap$ with the set of generators given by sets of attributes each of which describes a positive example (the operation $\bigcap$ is a means of expressing "similarity" of objects described by attribute sets). This observation allowed redefinition of hypotheses [45] as pairs of the form

$$\langle V, \{Z_1, \ldots, Z_k\}\rangle : V = Z_1 \cap \ldots \cap Z_k, \quad \forall Z \in D \setminus \{Z_1, \ldots, Z_k\} \quad V \not\subseteq X, \qquad (1)$$

where $V, Z_1, \ldots, Z_k \subseteq U$ for some set of attributes U and $D = \{Z_1, \ldots, Z_n\}$ is the set of all positive examples (given as sets of attributes that describe them) of the phenomenon W.

In [45] the equivalence of pairs $\langle V, \{Z_1, \ldots, Z_k\}\rangle$ to bicliques (inclusion-maximal complete bipartite graphs) of a bipartite graph was shown. Some years later the equivalence between pairs of this form (with components interchanged) and formal concepts was recognized [47].

The following definition of a hypothesis ("no counterexample-hypothesis") in FCA terms was given in [24]:

Let a context $\mathbb{K} = (G, M, I)$ be given. In addition to attributes of M, a *target attribute* $\omega \notin M$ is considered. This partitions the set G of all objects into three subsets: The set G_+ of those objects that are known to have the property ω (these are the *positive examples*), the set G_- of those objects of which it is known that they do not have ω (the *negative examples*) and the set G_τ of *undetermined examples*, i.e., of those objects, of which it is unknown if they have property ω or not. This gives three subcontexts of $\mathbb{K} = (G, M, I)$, the first two staying for the training sample:

$$\mathbb{K}_+ := (G_+, M, I_+), \quad \mathbb{K}_- := (G_-, M, I_-), \quad \text{and } \mathbb{K}_\tau := (G_\tau, M, I_\tau),$$

where for $\varepsilon \in \{+, -, \tau\}$ we have $I_\epsilon := I \cap (G_\varepsilon \times M)$ and the corresponding derivation operators are denoted by $(\cdot)^+$, $(\cdot)^-$, $(\cdot)^\tau$, respectively.

A subset $h \subseteq M$ is a *simple positive hypothesis* for ω if it satisfies the positive agreement predicate (see above) and does not satisfy the (symmetrically formulated) negative predicate). In terms of FCA,

$$h^{++} = h \quad \text{and} \quad h^{--} \neq h.$$

Another type of hypothesis which are mostly used in practice, namely "no counterexample hypothesis" [17, 18] (in what follows, we call it just a *positive hypothesis*), is an intent of $\mathbb{K}_+$ such that $h^+ \neq \emptyset$ and $h \not\subseteq g^- := \{m \in M \mid (g, m) \in I_-\}$ for any negative example $g \in G_-$. Equivalently,

$$h^{++} = h \quad \text{and} \quad h' \cap G_- \neq \emptyset,$$

where $(\cdot)'$ is taken in the whole context $\mathbb{K} = (G, M, I)$. An intent of $\mathbb{K}_+$ that is contained in the intent of a negative example is called a *falsified (+)-generalization. Negative hypotheses* and falsified generalizations are defined similarly. Hypotheses can be used to classify undetermined examples: If the intent

$$g^\tau := \{m \in M \mid (g, m) \in I_\tau\}$$

of an object $g \in G_\tau$ contains a positive, but no negative hypothesis, then g^τ is *classified positively*. Negative classifications are defined similarly. If g^τ contains hypotheses of both kinds, or if g^τ contains no hypothesis at all, then the classification is contradictory or undetermined, respectively. In this case one can apply standard probabilistic techniques known in machine learning and data mining (majority vote, Bayesian approach, etc.). Obviously, for classification purposes it suffices to have only *minimal* (w.r.t. inclusion $\subseteq$) hypotheses, positive as well as negative.

Example 2. Consider the following data table

G \ M	color	firm	smooth	form	**target**
1 apple	yellow	no	yes	round	+
2 grapefruit	yellow	no	no	round	+
3 kiwi	green	no	no	oval	+
4 plum	blue	no	yes	oval	+
5 toy cube	green	yes	yes	cubic	−
6 egg	white	yes	yes	oval	−
7 tennis ball	white	no	no	round	−

This dataset or *multivalued context* can be reduced to a context of the form presented above by *scaling* [27], e.g., as follows (scaling 1):

G \ M	w	y	g	b	f	$\bar{f}$	s	$\bar{s}$	r	$\bar{r}$	**target**
1 apple		×				×	×		×		+
2 grapefruit		×				×		×	×		+
3 kiwi			×			×		×		×	+
4 plum				×		×	×			×	+
5 toy cube			×		×		×			×	−
6 egg	×				×		×			×	−
7 tennis ball	×					×		×	×		−

Here we use the following abbreviations: "w" for white, "y" for yellow, "g" for green, "b" for blue, "s" for smooth, "f" for firm, "r" for round, "o" for oval, and "$\overline{m}$" for $m \in \{$w, y, g, b, s, f, r, o,$\}$.

This context gives rise to the positive concept lattice in Fig. 1, where we marked minimal (+)-hypotheses and falsified (+)-generalizations. If we have an undetermined example mango with $\texttt{mango}^\tau = \{$y, $\bar{\text{f}}$, s, $\bar{\text{r}}\}$ then it is classified positively, since $\texttt{mango}^\tau$ contains the minimal hypothesis $\{\bar{\text{f}}, \bar{\text{r}}\}$ and does not contain any negative hypothesis. For this scaling we have two minimal negative hypotheses: {w} (supported by examples `egg` and `tennis ball` and {f, s, $\bar{\text{r}}$} (supported by examples `toy cube` and `egg`.

The context can be scaled differently, e.g. in this way (scaling 2):

G \ M	w	y	g	b	$\bar{w}$	$\bar{y}$	$\bar{g}$	$\bar{b}$	f	$\bar{f}$	s	$\bar{s}$	r	o	$\bar{r}$	$\bar{o}$	**target**
1 apple		×			×		×	×		×	×		×	×			+
2 grapefruit		×			×		×	×		×		×	×	×			+
3 kiwi			×		×	×		×		×		×		×	×		+
4 plum				×	×	×	×			×	×			×	×		+
5 toy cube			×		×	×		×	×		×				×	×	−
6 egg	×					×	×	×	×		×			×	×		−
7 tennis ball	×					×	×	×		×		×	×	×			−

This scaling gives rise to another positive concept lattice, all intents of which are (+)-hypotheses. The unique minimal hypothesis (corresponding to the top element

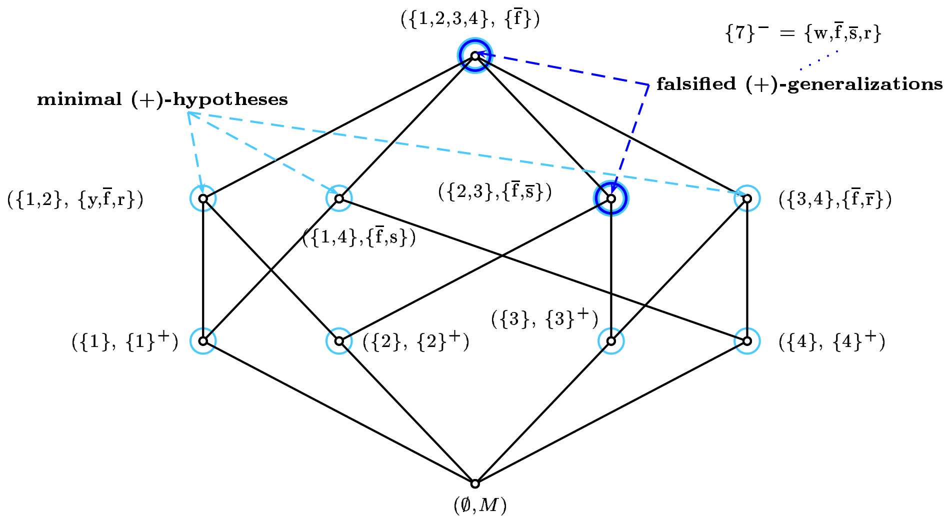


Fig. 1. Positive concept lattice for scaling 1

of the concept lattice) is $\{\overline{\mathrm{w}}, \bar{\mathrm{f}}, \mathrm{o}\}$. Two minimal negative hypotheses are $\{\bar{\mathrm{y}}, \bar{\mathrm{b}}, \bar{\mathrm{r}}, \mathrm{f}, \mathrm{s}\}$ (supported by examples 5 and 6) and $\{\bar{\mathrm{y}}, \bar{\mathrm{g}}, \bar{\mathrm{b}}, \mathrm{w}, \mathrm{o}\}$ (supported by examples 6 and 7).

The definitions of JSM-hypotheses can be varied, e.g., as follows:

- by imposing other logical conditions (e.g. of the "Difference method" of J.S. Mill), which gives rise to the "lattice of methods" [18],
- by allowance for $\alpha\%$ of counterexamples (for hypotheses and/or classifications) [29, 30],
- by using nonsymmetric classification (e.g., $(-)$-hypotheses are selected by stronger conditions than $(+)$-hypotheses) [18, 20],
- by varying "similarity operation" (see Section 4.2)

4.1 Various hypotheses of the JSM-method in terms of Galois connections

Various types of hypotheses expressed via respective plausible reasoning predicates of the JSM-method were supposed to capture different aspects of the relationship between structural and functional (target) attributes of objects. In the previous sections we considered representation of JSM-hypotheses by means of Galois connections for the case with a single target attribute. Here we give a description of various types of JSM-hypotheses from [17, 3, 18, 20, 40, 21], assuming that there are several target attributes.

Here, for simplicity sake we also assume that each example (object) in the training dataset is either a positive or a negative example with respect to each attribute. Then the situation can be represented by two formal contexts: a *structural* context $\mathbb{K}_M = (G, M, I)$ and a *target* context $\mathbb{K}_P = (G, P, J)$, where P is the set of target attributes (properties) and M is the set of structural attributes. The resulting context (called the *apposition* of the two contexts) can be given by the following matrix:

	M	P
G	I	J

Derivation operators $(\cdot)^I$ and $(\cdot)^J$ are defined in the usual way. Complements of relations I and J are defined naturally as $\bar{I} : g\bar{I}m \iff \neg gIm$, $\bar{J} : g\bar{J}m \iff \neg gJm$. These relations also define derivation operators $(\cdot)^{\bar{I}}$ and $(\cdot)^{\bar{J}}$. Now the definitions of various hypothesis types of the JSM-method can be represented by the following table (here $V \subseteq M$ and $W \subseteq P$):

hypothesis name	expression for (+)	expression for (-)
agreement	$V = (V^I \cap W^+)^I$, $W = V^{I+}$	$V = W^{-I}$, $W = V^{I-}$
no counterexample	$V = V^{II}$, $V^I \subseteq W^+$	$V = V^{II}$, $V^I \subseteq W^-$
inverse	$W = W^{++}$, $V = W^{+I}$	$W = W^{--}$, $V = W^{-I}$
situational	$V = ((V \cup S)^I \cap W^+)^I$	$W^{-I} = V \cup S$
Mill's difference	$(\bigcup_{j=0}^{m} V_j)^{\bar{I}} \subseteq W^{\{-,\tau\}}$	symmetric
generalized	$\mathcal{X} = \min\{B \mid V \subset B = (B^I \cap W^-)^I\}$	symmetric

Here "symmetric" means that the expression for (-) is obtained from the expression for (+) by replacing "+" with "-". Each hypothesis type defines the set of all pairs (V, W) such that the set of structural attributes V is a hypothetical cause of the set of target attributes W. Above we have considered the methods of agreement (also with additional "no counterexample condition") and difference (as formulated by J.S. Mill). Its JSM-formalization requires that the effect W does not occur in the absence of causes from $\{V_j\}_j$ (determined by other methods, e.g., by agreement). The intuitive meaning of other methods in this table is as follows. The *inverse method* is applied for "effect-cause reasoning" [20], usually when the number of attributes in P is larger than that in M. In the *situational method* [21] the importance of situation S for establishing relation between cause V and effect W is underlined. In the *generalized method* [20] it is assumed that each hypothetical cause V of effect W can have specific hindrances from the set $\mathcal{X}$, so V plausibly causes W only in the absence of elements of $\mathcal{X}$. Note that for "no counterexample" hypotheses there can be no hindrances as defined in the table.

Another specific feature of the JSM-method is the so-called *condition of causal completeness* [20], which states that for chosen methods and a dataset the generated positive and negative hypotheses should classify the initial data correctly:

$$\bigcup_{M_x^+(V,W)} V^I = W^+, \qquad \bigcup_{M_y^-(V,W)} V^I = W^-,$$

where $M_x^+(V,W)$ and $M_y^-(V,W)$ denote some positive and negative methods, respectively. The condition is supposed to be tested each time hypotheses are generated.

The invariant feature of hypotheses w.r.t. different types of predicates is that they are sought among closed subsets of attributes.

4.2 Similarity operation

Initially, similarity of object descriptions was defined in the JSM-method by means of set-intersection $\cap$. However, this definition suggested an obvious generalization: defining similarity as an idempotent, commutative and associative operation, i.e., as a meet operator in a semilattice. So, for each application domain with its specific data structure, a "similarity" operation was to be defined. This approach is equivalent to *scaling* in FCA where each many-valued attribute is turned to a set of related binary atributes.

An example of a similarity operation different from $\cap$ is the following interval algebra on real numbers. For two intervals $[a, b]$ and $[c, d]$ with $a, b, c, d \in \mathbb{R}$ and $a \leq c$ their meet can be defined as

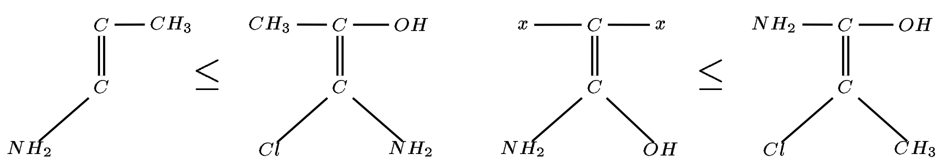
$$[a, b] \wedge [c, d] = [max(a, c), min(b, d)] \text{ if } b \geq c, \text{ otherwise } \Lambda,$$

where Λ denotes the empty interval with $\Lambda \wedge [a, b]$ for any $a, b \in \mathbb{R}$. This operation on intervals is often used in life-science applications, where, e.g., a number stays for a dose of a substance introduced [69] or a characteristic activation energy of a substance [55]. From the very beginning, the most important application of the JSM-method was the study of "chemical structure - biological activity" relationship. For this problem, adequate representation of chemical structure is essential. A special encoding scheme, called Fragmentary Code of Substructure Superposition (FCSS) (see, e.g., [9]), which turns molecular graphs to sets of binary attributes, was used. This encoding scheme allows efficient search for molecular similarities, however it leads also to the loss of information on connection between molecular parts. This problem motivated the search for mathematical means that would help dealing directly with graph representation of molecules. A solution was proposed in the form of a semilattice of graph sets [43, 44, 52, 46].

This semilattice is based on the following ordered set P of graphs with labels from the set $\mathcal{L}$ with partial order $\preceq$. Each labeled graph Γ from P is a triple of the form $((V, l), E)$, where V is a set of vertices, E is a set of edges and $l: V \to \mathcal{L}$ is a label assignment function, taking a vertex to its label.
For two graphs $\Gamma_1 := ((V_1, l_1), E_1)$ and $\Gamma_2 := ((V_2, l_2), E_2)$ from P Γ_1 *dominates* Γ_2 or $\Gamma_2 \leq \Gamma_1$ if there exists a one-to-one mapping
$\varphi: V_2 \to V_1$ such that it

- respects edges: $(v, w) \in E_2 \Rightarrow (\varphi(v), \varphi(w)) \in E_1$,
- fits under labels: $l_2(v) \preceq l_1(\varphi(v))$.

Example 3. Let $\mathcal{L} = \{$C, NH_2, CH_3, OH, $x\}$, then we have the following domination relations:



vertex labels are unordered

$x \preceq A$ for any vertex label $A \in \mathcal{L}$

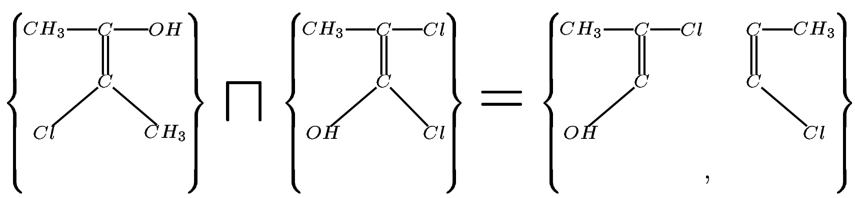
A meet operation $\sqcap$ on graph sets can then be defined as follows: For two graphs X and Y from P

$$\{X\} \sqcap \{Y\} := \{Z \mid Z \leq X, Y, \quad \forall Z_* \leq X, Y \quad Z_* \not\geq Z\},$$

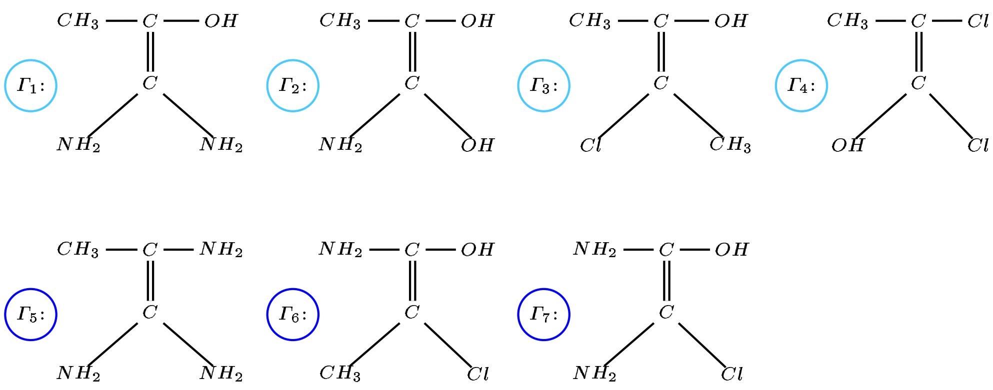
i.e., $\{X\} \sqcap \{Y\}$ is the set of all maximal common subgraphs of X and Y up to substitution of a vertex label by a vertex label smaller w.r.t. $\preceq$. The meet of nonsingleton sets of graphs is defined as

$$\{X_1, \ldots, X_k\} \sqcap \{Y_1, \ldots, Y_m\} := \mathrm{MAX}_{\leq}(\bigcup_{i,j}(\{X_i\} \sqcap \{Y_j\}))$$

for details see [46, 48, 25]. Here is an example of applying $\sqcap$ defined above:



Let positive examples be described by graphs $\Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4$ and negative examples be described by graphs $\Gamma_5, \Gamma_6, \Gamma_7$:



then the lattice of graph sets generated by positive examples and their graph descriptions is given in Fig. 2, where (+)-hypotheses and falsified (+)-generalizations are highlighted:

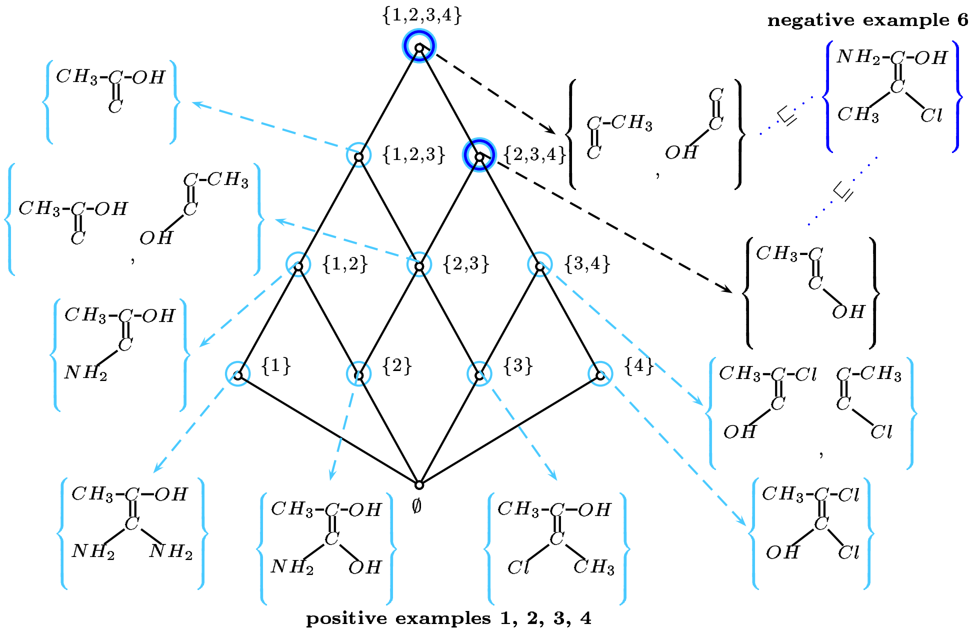


Fig. 2. The lattice of the positive pattern structure

The same approach is realizable for arbitrary data descriptions with generality (subsumption) order $\leq$. The general idea is to consider the (distributive) lattice of order ideals of $\leq$, distinguish the elements of it that correspond to descriptions of examples (objects) and consider these elements as generators of a meet-subsemilattice of the lattice of order ideals. Being supplied with a dummy top element (which is feasible, e.g., in the case when the number of objects is finite) this subsemilattice turns into a sublattice of the lattice of order ideals of $\leq$ (which is not necessarily distributive). These ideas were proposed and developed in [25], where semilattices of patterns were considered. Let G be a set (elements of which are called objects), let $(D, \sqcap)$ be a meet-semilattice and let $\delta : G \to D$ be a mapping. Then $(G, \underline{D}, \delta)$ with $\underline{D} = (D, \sqcap)$ is called a *pattern structure*, provided that the set

$$\delta(G) := \{\delta(g) \mid g \in G\}$$

generates a complete subsemilattice $(D_\delta, \sqcap)$ of $(D, \sqcap)$, i.e., every subset X of $\delta(G)$ has an infimum $\sqcap X$ in $(D, \sqcap)$

and D_δ is the set of these infima. Each such complete semilattice has lower and upper bounds, which we denote by $\mathbf{0}$ and $\mathbf{1}$, respectively. There are two natural situations where the condition on the complete subsemilattice is automatically satisfied: when $(D, \sqcap)$ is complete, and when G is finite. If $(G, \underline{D}, \delta)$ is a pattern structure, we define the derivation operators as

$$A^\diamond := \bigsqcap_{g \in A} \delta(g) \qquad \text{for } A \subseteq G$$

and

$$d^\diamond := \{g \in G \mid d \sqsubseteq \delta(g)\} \qquad \text{for } d \in D.$$

The elements of D are called *patterns*. The natural order on them is given, as usual, by

$$c \sqsubseteq d :\Longleftrightarrow c \sqcap d = c,$$

and is called the *subsumption* order.

An algebraic model of approximation was proposed in [25] in the form of projection (or kernel, i.e., idempotent, monotone and contracting) operator and the reduction to standard concept lattices was discussed. Since projections (kernels) preserve meet operator, hypotheses in projected data have preimages in the original data that are hypotheses too. The research of pattern structures and their approximations led to further practical applications in chemistry with the approximation level being controlled by a parameter [23].

4.3 Mathematical activity around concept-based hypotheses

Here we give a partial list of references to some research around JSM-method, concept-based hypotheses and implication bases.
Logics. Construction of quasi-axiomatic theories of plausible JSM-reasoning, completeness problem of the theory of plausible reasoning based on rules that generate hypotheses were considered in [3, 5]. Argumentation logic (where a proof of a statement takes into account arguments for and against the statement) were considered in [19]. In [86, 88] the author studies (partial) expressibility of plausible reasoning rules in Prolog and the expressibility in first-order predicate logic was studied in [87]. Logics of causal reasoning in the JSM-method were studied in [2]. A modal logic of incomplete contexts was studied in [67].

Algebraic issues. Similarity operation on sets of labeled graphs, which is an infimum (meet) operation in a corresponding semilattice, was defined and studied in [43, 44, 46, 1, 48, 25]. A distributive lattice (of order ideals) of data for JSM-method was studied in [1]. Further generalization of the graph set semilattice and its translation in FCA terms was realized in [25], where general pattern semilattices were studied (see the previous subsection).
Algorithmic issues. First algorithms for computing JSM-hypotheses were proposed in [94, 61, 95], a recent review which includes theoretical and experimental comparison of various algorithms for computing closed sets and concept lattices is found in [54]. Polynomial tractability and intractability of certain decision problems related to generation of hypotheses was considered in [96, 45, 46]. In [45] it was proved that the problem of computing the number of hypotheses is #P-complete, In [46, 51] same result was proved for the number of minimal hypotheses. In [49, 51] similar results were demonstrated for concepts. In particular, it was shown that the problems of computing the number all concepts is #P-complete. A very efficient incremental algorithm for computing concept lattices was proposed in [60]. A fast incremental algorithm for computing Duquenne-Guigues implication bases (with the best known experimental performance) was proposed in [68].

4.4 Applications of JSM-method

Starting from the early 1980s JSM-hypotheses were used in several applied domains, including bioscience analysis of biological activity of chemicals (see reviews [7, 97]) predicting metabolic pathways [15, 58]), medical diagnostics, technical diagnostics, sociology, document dating, spam filtering, and so on. JSM-hypotheses were used successfully for making predictions at two international competitions: that for predictive toxicology [9] (where JSM-hypotheses resulted in optimal classifications in all test groups) and that for spam filtering [13]. A freeware system QuDA [32, 33], which incorporates several data mining techniques also presents a possibility of generating JSM-hypotheses.
Life sciences. Most numerous experiments were carried out in applied pharmacology or Structure-Activity Relationship domain, which deals with predicting biological activity of chemical compounds with known molecular structure. JSM-hypotheses were generated for antitumor [71], antibacterial, antileprous, hepatoprotective [12], plant growth-stimulating, cholesterase-inhibitine, toxic and carcino-

genic activities, see reviews [7, 97]. JSM-method was many times applied to problems of medical diagnostics, e.g., the results of the study of human papilloma are found in [69]. Recent results in the study of toxicity of different substances, including alcohols and halogen-substituted hydrocarbons by means of learning in pattern structures on graph sets are found in [55].

Sociology and Humanities. In [21] strike readiness at joint-stock factories in St. Petersburg and Elets was analysed. The advantages of the JSM-based approach as compared to statistical methods resided in the fact that the former allowed for creating taxonomies of socio-psychological types and enabled creating "social portraits". In paleography [41] the JSM-method was applied to dating birch-bark documents of 10–16 centuries of the Novgorod republic. Here there were five types of attributes describing individual letter features, features common to several letters, handwriting, language features: morphology, syntax, and typical errors, style: letter format, addressing formulas and their key words. Time was considered as many-valued target attribute, with 20 nointersecting time intervals as attribute values. A model for analyzing human conflicts that uses for similarity of labeled graphs was studied in [22].

Spam filtering. A first successful application of the JSM-like (concept-based) hypotheses for filtering spam was reported in [16]. In April-May 2003 Technical University Chemnitz, European Knowledge Discovery Network, and PrudSys AG organized the Data Mining Cup (DMC) competition for students specializing in Machine Learning [13]. Among 514 participants from 199 universities of 38 countries the sixth place was taken by a model that combined "Naive Bayes" approach with JSM-hypotheses.

5 Machine Learning in terms of Galois connection and FCA

In recent years some progress was done in describing various learning models like version spaces, decision trees in terms of Galois connection and concept lattices [26, 50].

6 Decision trees embedded in concept lattices

As input, a system constructing a decision tree (see, e.g., [75]) receives descriptions of positive and negative examples (or positive and negative contexts, in terms of the previous section). The root of the tree corresponds to the beginning of the process and is not labeled. Other vertices of the decision tree are labeled by attributes and edges are labeled by values of the attributes (e.g., 0 or 1 in case of binary contexts), each leaf is additionally labeled by a class $+$ or $-$, meaning that all examples with attribute values from the path leading from the root to the leaf belong to a certain class, either $+$ or $-$.

Systems like ID3 [75] (see also [65]) compute the value of the *information gain* (or negentropy) for each vertex and each attribute not chosen in the branch above. The attribute with the greatest value of the information gain (with the smallest entropy, respectively) "most strongly separates" objects from classes $+$ and $-$. The algorithm sequentially extends branches of the tree by choosing attributes with the highest information gain. The extension of a branch stops when a next attribute value together with attributes above in the branch uniquely classify examples with this value combination in one of classes $+$ or $-$. In some algorithms, the process of extending a branch stops before this in order to avoid *overfitting*, i.e., the situation where all or almost all examples from the training sample are classified correctly by the resulting decision tree, but objects from test datasets are classified with many errors.

Now we consider decision trees more formally. Let the training data be described by the context $\mathbb{K}_{+-} = (G_+ \cup G_-, M, I_+ \cup I_-)$ with the derivation operator denoted by $(\cdot)'$. In FCA terms this context is called the *subposition* of $\mathbb{K}_+$ and $\mathbb{K}_-$. Assume for simplicity sake that for each attribute $m \in M$ there is an attribute $\overline{m} \in M$, a "negation" of m: $\overline{m} \in g'$ iff $m \notin g'$. A set of attributes M with this property is called *dichotomized* in FCA. We call a subset of attributes $A \subseteq M$ *noncontradictory* if either $m \notin A$ or $\overline{m} \notin A$. We call a subset of attributes $A \subseteq M$ *complete* if for every $m \in M$ one has $m \in A$ or $\overline{m} \in A$.

First no optimization functional (like information gain) for selecting attributes is involved and construction of all possible decision trees is considered. The construction of an arbitrary decision tree proceeds by sequentially choosing attributes. If different attributes $m_1, \ldots, m_k$ were chosen one after another, then the sequence $\langle m_1, \ldots, m_k \rangle$ is called a *decision path* if $\{m_1, \ldots, m_k\}$ is noncontradictory and there exists an object $g \in G_+ \cup G_-$ such that $\{m_1, \ldots, m_k\}' \subseteq g'$ (i.e., there is an example with this set of attributes). A decision path $\langle m_1, \ldots, m_i \rangle$ is a (proper) subpath of a decision path $\langle m_1, \ldots, m_k \rangle$ if $i \leq k$ ($i < k$, respectively). A decision path $\langle m_1, \ldots, m_k \rangle$ is called *full* if all objects having attributes $\{m_1, \ldots, m_k\}$ are either positive or negative examples (i.e., have either $+$ or $-$ value of the target attribute).

We call a full decision path *irredundant* if none of its subpaths is a full decision path. The set of all chosen attributes in a full decision path can be considered as a sufficient condition for an object to belong to a class $\varepsilon \in \{+, -\}$. A decision tree is then a set of full decision paths.

In what follows, we use the one-to-one correspondence between vertices of a decision tree and the related decision paths, representing the latter, when this does not lead to ambiguity, by their last chosen attributes. By *closure of a decision path* $\langle m_1, \ldots, m_k \rangle$ we mean the closure of the corresponding set of attributes, i.e., $\{m_1, \ldots, m_k\}''$. Now we relate decision trees with the covering relation graph of the concept lattice of the context $\mathbb{K} = (G, M, I)$, where the set of objects G is of size $2^{|M|/2}$ and the relation I is such that the set of object intents is exactly the set of complete noncontradictory subsets of attributes. In terms of FCA [27] the context $\mathbb{K}$ is the *semiproduct* of $|M|/2$ *dichotomic scales* or $\mathbb{K} = D_1 \between \ldots \between D_{|M|/2}$ (denoted by $\between_M D$ for short), where each dichotomic scale D_i stays for the pair of attributes (m,$\overline{\text{m}}$).

In a concept lattice a sequence of concepts with decreasing extents we call a *descending chain*. If the chain starts at the top element of the lattice, we call it *rooted*.

Proposition 5 *Every decision path is a rooted descending chain in $\underline{\mathfrak{B}}(\between_M D)$ and every rooted descending chain consisting of concepts with nonempty extents in $\underline{\mathfrak{B}}(\between_M D)$ is a decision path.*

To relate decision trees to hypotheses introduced above we consider again the contexts $\mathbb{K}_+ = (G_+, M, I_+)$, $\mathbb{K}_- = (G_-, M, I_-)$, and $\mathbb{K}_{+-} = (G_+ \cup G_-, M, I_+ \cup I_-)$. The context $\mathbb{K}_{+-}$ can be much smaller than $\between_M D$ because the latter always has $2^{|M|/2}$ objects while the number of objects in the former is the number of examples. Also the lattice $\underline{\mathfrak{B}}(\mathbb{K}_{+-})$ can be much smaller than $\underline{\mathfrak{B}}(\between_M D)$.

The relation between decision trees and (minimal "no counterexample") hypotheses from the previous section is given by the following

Proposition 6 *A full decision path $\langle m_1, \ldots, m_k \rangle$ corresponds to a rooted descending chain $\langle (m_1'', m_1'), \ldots, (\{m_1, \ldots, m_k\}'', \{m_1, \ldots, m_k\}') \rangle$ of the line diagram of $\underline{\mathfrak{B}}(\mathbb{K}_{+-})$ and the closure of each full decision path $\langle m_1, \ldots, m_k \rangle$ is a hypothesis, either positive or negative. Moreover, for each minimal hypothesis h, there is a full irredundant path $\langle m_1, \ldots, m_k \rangle$ such that $\{m_1, \ldots, m_k\}'' = h$.*

By the proposition, hypotheses correspond to the "most cautious" (most specific) learning strategy in the sense that they are least general generalizations of descriptions of positive examples (or object intents, in terms of FCA). The shortest decision paths (for which in no decision tree there exist full paths with proper subsets of attribute values) correspond to the "most courageous" ("most discriminating") learning strategy: being the shortest possible rules, they are most general generalizations of positive example descriptions. However, it is not guaranteed that for a given training set resulting in a certain set of minimal hypothesis there is a decision tree such that minimal hypotheses are among closures of its paths (see Example 4 below). In general, to obtain all minimal hypotheses as closures of decision paths one needs to consider several decision trees, not all of them being optimal w.r.t. a procedure based on the information gain functional (like ID3 or C4.5). The issues of generality of generalizations and, in particular, the relation between most specific and most general generalizations, are naturally captured in terms of version spaces, which we consider in the next section.

In real systems for the construction of decision trees like ID3 or C4.5 the process of constructing a decision path is driven by the information gain functional: a next chosen attribute should have maximal information gain. For dichotomized attributes the information gain is defined for a pair of attributes $m, \overline{m} \in M$.

Given a decision path $\langle m_1, \ldots, m_k \rangle$

$$\mathrm{IG}(m) := -\frac{|A'_m|}{|G|}\mathrm{Ent}(A_m) - \frac{|A'_{\overline{m}}|}{|G|}\mathrm{Ent}(A_{\overline{m}}),$$

where $A_m := \{m_1, \ldots, m_k, m\}$, $A_{\overline{m}} := \{m_1, \ldots, m_k, \overline{m}\}$, and for $A \subseteq M$

$$\mathrm{Ent}(A) := -\sum_{\varepsilon \in \{+,-\}} p(\varepsilon \mid A) \cdot \log_2 p(\varepsilon \mid A),$$

$\{+,-\}$ are values of the target attribute and $p(\varepsilon \mid A)$ is the conditional sample probability (for the training set) that an object having a set of attributes A belongs to a class $\varepsilon \in \{+,-\}$.

If the derivation operator $(\cdot)'$ is associated with the context $(G_+ \cup G_-, M, I_+ \cup I_-)$, then, by definition of the conditional probability, we have

$$p(\varepsilon \mid A) = \frac{|A' \cap G_\varepsilon|}{|A'|} = \frac{|(A'')' \cap G_\varepsilon|}{|(A'')'|} = p(\varepsilon \mid A'')$$

by the property of the derivation operator $(\cdot)'$: $(A'')' = A'$. This observation implies that instead of considering decision paths, one can consider their closures without affecting the values of the information gain. In terms of lattices this means that instead of the concept lattice $\mathfrak{B}(\mathbb{X}_M\ D)$ one can consider the concept lattice of the context $\mathbb{K}_{+-} = (G_+ \cup G_-, M, I_+ \cup I_-)$. Another consequence of the invariance of IG w.r.t. closure is the following fact: If implication $m \to n$ holds in the context $\mathbb{K}_{+-} = (G_+ \cup G_-, M, I_+ \cup I_-)$, then an IG-based algorithm will not choose attribute n in the branch below chosen m and will not choose m in the branch below chosen $\overline{n}$.

Example 4. Consider the training set from Example 2. The decision tree obtained by the IG-based algorithm is given in Fig. 3. Note that attributes f and w has the same IG value (a similar tree with f at the root is also optimal), the IG-based algorithms usually take the first attribute with the same value of IG.

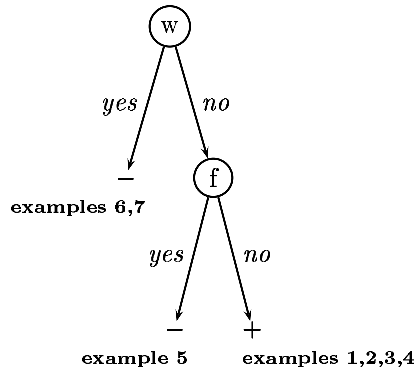


Fig. 3. A decision tree for Example 2

The decision tree in Fig. 3 corresponds to three implications $\{\mathrm{w}\} \to -$, $\{\overline{\mathrm{w}}, \mathrm{f}\} \to -$, $\{\overline{\mathrm{w}}, \overline{\mathrm{f}}\} \to +$, such that closures of their premises make the corresponding negative and positive hypotheses for the second scaling from Example 2. Note that the hypothesis $\{\overline{\mathrm{w}}, \mathrm{f}\}''$ is not minimal, since there is a minimal hypothesis $\{\mathrm{f}\}''$ contained in it. The minimal hypothesis $\{\mathrm{f}\}''$ corresponds to a decision path of the mentioned IG-based tree with the attribute f at the root.

6.1 Version spaces vs. concept-based hypotheses

The term *version space* was proposed by T. Mitchell [64, 65] to denote a variety of models compatible with the training sample of positive and negative examples. Version spaces can be defined in different ways. Here they are described in terms somewhat different to those in [64], in order to avoid collision with FCA terminology.

- An *example language* L_e (elsewhere also called *instance language*) by means of which the examples (instances) are described. This language describes a *set* E *of examples.*
- A *classifier language* L_c describing the possible classifiers (elsewhere called *concepts*). This language describes a set C of *classifiers.*
- A *matching predicate* $M(c, e)$ that defines if a classifier c does or does not *match* an example e: We have $M(c, e)$ iff e is an example of classifier c. The set of classifiers is (partially) ordered by a *subsumption order*: for $c_1, c_2 \in L_c$ the classifier c_1 subsumes c_2 or $c_1 \sqsupseteq c_2$ if c_1 corresponds to a more specific description and thus, covers less objects than c_2:

$$c_1 \sqsupseteq c_2 :\Longleftrightarrow \forall_{e \in E}\ M(c_1, e) \to M(c_2, e).$$

  The corresponding strict order $\sqsupset$ is called *proper subsumption.*
- Sets E_+ and E_- of *positive* and *negative examples* of a *target attribute* with $E_+ \cap E_- = \emptyset$. The target attribute is not explicitly given.
- *consistency predicate* $\mathrm{cons}(c)$:
  $\mathrm{cons}(c)$ holds if for every $e \in E_+$ the matching predicate $M(c, e)$ holds and for every $e \in E_-$ the negation $\neg M(c, e)$ holds. The set of all consistent classifiers is called the *version space*

$$\mathrm{VS}(L_c, L_e, M(c, e), E_+, E_-).$$

The learning problem is then defined as follows:

Given $L_c, L_e, M(c, e), E_+, E_-$.
Find the version space $\mathrm{VS}(L_c, L_e, M(c, e), E_+, E_-)$.

In the sequel, we shall usually fix L_c, L_e, and $M(c,e)$ and write $\mathrm{VS}(E_+, E_-)$ or even just VS for short. Version spaces are often considered in terms of *boundary sets* proposed in [64]. They can be defined if the language L_c is *admissible*, i.e., if every chain in the subsumption order has a minimal and a maximal element. In this case,

$$\mathrm{GVS}(L_c, L_e, M(c,e), E_+, E_-) := \mathrm{MIN}_{\sqsubseteq}(\mathrm{VS}) := \{c \in \mathrm{VS} \mid \neg\exists c_1 \in \mathrm{VS}\, c_1 \sqsubset c\},$$
$$\mathrm{SVS}(L_c, L_e, M(c,e), E_+, E_-) := \mathrm{MAX}_{\sqsubseteq}(\mathrm{VS}) := \{c \in \mathrm{VS} \mid \neg\exists c_1 \in \mathrm{VS}\, c \sqsubset c_1\}.$$

If a version space VS is fixed, we also use notation G(VS) and S(VS) for short.

The elements of the version space can be used as potential classifiers for the target attribute: A classifier $c \in \mathrm{VS}$ *classifies* an example positively if c matches e and negatively else. Then, all positive examples are classified positively, all negative examples are classified negatively, and undetermined examples may be classified either way. If it is assumed that E_+ and E_- carry sufficient information about the target attribute, we may expect that an undetermined example is likely to have the target attribute if it is classified positively by a large percentage of the version space (cf. [65]). We say that an example e is $p\%$-classified (for $0 \leq p \leq 100$) if no less than $p\%$ classifiers of the version space classify it positively. This means, e.g., that 100%-classification of e takes place if e is matched by all elements of SVS and negative classification of e (0%-classification) takes place if e is not matched by any element of GVS.

As showed in [26] the basic properties of general version spaces can easily be expressed with Galois connections. Consider the formal context (E, C, I), where E is the set of examples containing the disjoint sets of observed positive and negative examples: $E \supseteq E_+ \cup E_-$, $E_+ \cap E_- = \emptyset$, C is the set of classifiers and the relation I corresponds to the matching predicate $M(c,e)$: for $c \in C$, $e \in E$ the relation eIc holds iff $M(c,e) = 1$. The complementary relation, $\bar{I}$, corresponds to the negation: $e\bar{I}c$ holds iff $M(c,e) = 0$. As shown in [26]

$$\mathrm{VS}(E_+, E_-) = {E_+}^I \cap {E_-}^{\bar{I}}.$$

This characterization of version spaces implies immediately the property of *merging version spaces*, proved in [42]: For fixed L_c, L_e, $M(c,e)$ and two sets E_{+1}, E_{-1} and E_{+2}, E_{-2} of positive and negative examples one has

$$\mathrm{VS}(E_{+1} \cup E_{+2}, E_{-1} \cup E_{-2}) = \mathrm{VS}(E_{+1}, E_{-1}) \cap \mathrm{VS}(E_{+2}, E_{-2}).$$

This follows from the relation $(A \cup B)' = A' \cap B'$, which holds for a derivation operator $(\cdot)'$ of an arbitrary context.

The classifications produced by classifiers from the version space are characterized as follows. The set of all 100%-classified examples w.r.t the version space $\mathrm{VS}(E_+, E_-)$ is given by

$$({E_+}^I \cap {E_-}^{\bar{I}})^I.$$

In particular, if one of the following conditions is satisfied, then there cannot be any 100%-classified undetermined example:

1. $E_- = \emptyset$ and ${E_+}^{II} = E_+$,
2. $({E_+}^I \cap {E_-}^{\bar{I}})^I = E_+$.

The set of examples that are classified positively by at least one element of the version space $\mathrm{VS}(E_+, E_-)$ is given by

$$E \setminus ({E_+}^I \cap {E_-}^{\bar{I}})^{\bar{I}}.$$

Consider a very important special case where the ordered set $(C, \leq)$ of classifiers given in terms of some language L_c makes a meet-semilattice w.r.t. $\wedge$ meet operation, like in Section 4.2. This also covers the case of attributes with values.

In [26] it was shown that in the case where the classifiers, ordered by subsumption, form a complete semilattice, the version space is a complete subsemilattice for any sets of examples E_+ and E_-. If the set of classifiers C makes a complete semilattice $(C, \sqcap)$, we can consider a *pattern structure* $(E, (C, \sqcap), \delta)$, where E is a set (of "examples"), δ is a mapping $\delta : E \to C$, $\delta(E) := \{\delta(e) \mid e \in E\}$. The subsumption order can be reconstructed from the semilattice operation: $c \sqsubseteq d \iff c \sqcap d = c$.

The version space may be empty, in which case there are no classifiers separating positive examples from negative ones. This happens, e.g., if there is a *hopeless* positive example (an outlier), by which we mean an element $e_+ \in E_+$ having a negative counterpart $e_- \in E_-$ such that every classifier which matches e_+ also matches e_-. An equivalent formulation of the hopelessness of e_+ is that $(e_+)^{\diamond\diamond} \cap E_- \neq \emptyset$. The following relation between the version space with latticeordered classifiers and minimal hypotheses was shown in [26]:

Suppose that the classifiers, ordered by subsumption, form a complete meet-semilattice $(C, \sqcap)$, and let $(E, (C, \sqcap), \delta)$ denote the corresponding pattern structure.

Proposition 7 *The following statements are equivalent:*

1. *The version space* $\mathrm{VS}(E_+, E_-)$ *is not empty.*
2. $(E_+)^{\diamond\diamond} \cap E_- = \emptyset$.
3. *There are no hopeless positive examples and there is a unique minimal positive hypothesis* h_{min}.

In this case, $h_{\min} = (E_+)^{\diamond}$, and the version space is a convex set in the lattice of all pattern intents, ordered by subsumption, with maximal element $h_{\min}$.

In case where conditions 1-3 are satisfied, the set of training examples is often referred to as *separable* in machine learning. The theorem gives access to generation of the version space, e.g., with the use of a standard Next Closure [27] algorithm.

According to [27] a subset $A \subseteq M$ can be defined as a *proper premise of an attribute* $m \in M$ if $m \notin A$, $m \in A''$ and for any $A_1 \subset A$ one has $m \notin A_1''$. In particular we can define a *positive proper premise* as a proper premise of the target attribute ω. In [26] we generalized this notion to include the possibility of the unknown value of a target attribute (for an undetermined example): $d \in L_c$ is a *positive proper predictor with respect to examples* E_+, E_-, and E_τ if the following conditions 1-3 are satisfied:

1 $d^\diamond \subseteq E_+ \cup E_\tau$,
2 $\exists g \in E_+ : g \in d^\diamond$ (or $d^\diamond \cap E_+ \neq \emptyset$),
3 $\forall d_1$ such that $d \sqsubseteq d_1$ and $d \neq d_1$, the relation $d_1^\diamond \not\subseteq E_+ \cup E_\tau$ holds.

In the case where $E_\tau = \emptyset$, condition 2 of the definition follows from condition 1 and a proper predictor is just a *proper premise* [27] of the target attribute.

The proper predictors and hypotheses are related to the boundaries of the version space as follows [26]:

Proposition 8 *Let* $PP_+(\Pi, E_+, E_-)$ *denote the set of all positive proper predictors for the pattern structure* $\Pi = (E, (C, \sqcap), \delta)$ *and sets of positive and negative examples* E_+ *and* E_-. *Let* $H_+(\Pi, E_+, E_-)$ *denote the set of positive hypotheses and* $VS(\Pi, E_+, E_-)$ *denote the version space for the pattern structure* $\Pi = (E, (C, \sqcap), \delta)$ *and sets of examples* E_+ *and* E_-. *Then the following holds:*

1. $PP_+(\Pi, E_+, E_-) = MAX_{\sqsubseteq}(\bigcup_{F_+ \subseteq E_+} GVS(\Pi, F_+, E_-))$,
2. $H_+(\Pi, E_+, E_-) = \bigcup_{F_+ \subseteq E_+} SVS(\Pi, F_+, E_-)$.

In contrast to version spaces with purely conjunctive classifiers, hypotheses propose a sort of "context-restricted" disjunction (which, hence is not so "loose" as purely syntactical disjunction over conjunction of attribute values): not all disjunctions are possible, but only those of minimal hypotheses (that are equivalent to certain conjunctions of attributes), which express similarities of examples.

7 Conclusion

We considered activity in classification, data analysis, and machine learning around VINITI Institute in Moscow and its NTI journal that used models naturally described in terms of Galois connections and FCA. Early research was related to the models of taxonomies and meronomies, which are naturally recast in terms of concept lattices. Galois connection are very helpful in modeling similarity given by tolerance relation and its classes.

Recasting the JSM-method in FCA terms motivated further activity in describing well-known models of machine learning and knowledge discovery, such as version spaces and decision trees, in terms of Galois connections and concept lattices. Translations of this kind often provide with a unified view, simpler definitions,and simpler proofs of the results. Further work in this direction will be related to other widely used models of learning, such as Naive Bayes, induction of ripple-down rules, support vector machines, and so on. The language of Galois connections provide with standard algorithmic machinery from FCA and new developments from Data Mining related to finding (closed) frequent itemsets.

Concept lattices that seem from the first glance to be a tool for processing binary tables, actually provide with means for dealing with complex structure such as logical formulas, labeled graphs (e.g., concept graphs, molecular graphs), texts, 3D-structures. This aspect indicates yet another direction of further study: fast algorithms for models with complex and/or large data. Successful applications in chemistry and conflict modelling give hope for future results.

Acknowledgments

This work was supported by the Russian Foundation for Humanities, project no. 05-03-03019a, by the project *Problem solver for the analysis of causal dependencies* of the Russian Academy of Science, and by the Alexander-von-Humboldt Foundation. The author thanks Bernhard Ganter for helpful discussions.

References

1. O.M. Anshakov, On a data lattice for the JSM-method of automated hypothesis generation, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1996, No. 5-6, pp. 33-36 [in Russian].
2. O.M. Anshakov, Causal models of subject domains, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 2000, No. 3, pp. 3-16 [in Russian], 2000.
3. O.M. Anshakov, D.P. Skvortsov, V.K. Finn, Logical Means of Expert Systems of JSM-type, *Semiotika i Informatika*, **28**, 1986, pp. 65-101 [in Russian].
4. O.M. Anshakov, D.P. Skvortsov, V.G. Ivashko, and V.K. Finn, Logical Means of the JSM-method of Automated Hypothesis Generation: Main Notions and System of Inference Rules, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1987, No. 9, pp. 10-18 [in Russian].
5. O.M. Anshakov, V.K. Finn, and D.P. Skvortsov, On axiomatization of many-valued logics associated with the formalization of plausible reasonings, *Stud. Log.*, **25**, No. 4, 23-47 (1989).
6. G.D. Birkhoff, *Lattice Theory*, Amer. Math. Soc. (1979).
7. V.G. Blinova, Results of Application of the JSM-method of Hypothesis Generation to Problems of Analyzing the Relation "Structure of a Chemical Compound - Biological Activity," *Autom. Docum. Math. Ling.*, vol. 29, no. 3, pp. 26-33, 1995.
8. V.G. Blinova, D.A. Dobrynin, Languages for Representing Chemical Structures in Intelligent Systems of Drug Design *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 2000, No. 6, pp. 14-21 [in Russian].

9. V.G. Blinova, D.A. Dobrynin, V.K. Finn, S.O. Kuznetsov, and E.S. Pankratova, Toxicology analysis by means of the JSM-method, *Bioinformatics* 2003, vol. 19, pp. 1201-1207.
10. D.A. Bochvar, On a three-valued calculus and its application to the analysis of paradoxes of the classical extended functional calculus, *Matematicheskii sbornik*, 1938, No. 2, pp. 287-308 [in Russian].
11. V.B. Borshev, V.A. Brudno, M.V. Khomyakov, Algebraic Description of the Structure of Dependencies in a Database, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1977, No. 3, pp. 17-18 [in Russian].
12. A.P. Budunova, V.V. Poroikov, V.G. Blinova, and V.K. Finn, The JSM-method of hypothesis generation: Application for analysis of the relation "Structure - hepatoprotective detoxifying activity", Nauchno-Tekhnicheskaya Informatsiya, no. 7, pp.12-15, 1993 [in Russian].
13. Data Mining Cup (DMC), http://www.data-mining-cup.de
14. V.V. Dunaev, O.M. Polyakov, Methodological Aspects of Relational Classification Theory, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1987, No. 4, pp. 21-27 [in Russian].
15. E.F. Fabrikantova, Problems of computer modeling of metabolic transformations of xenobiotics in a human organism *Itogi Nauki i Tekhniki, Seriya Informatika*, **15**, 115-135, 1991 [in Russian].
16. S. Férré and O. Ridoux, The Use of Associative Concepts in the Incremental Building of a Logical Context in *Proc. 10th Int. Conf. on Conceptual Structures, ICCS'2002*, U. Priss, D. Corbet, G. Angelova, Eds., Lecture Notes in Artificial Intelligence, **2393**, 2002, 299-313.
17. V.K. Finn, On Machine-oriented Formalization of Plausible Reasoning in F.Bacon-J.S.Mill Style, *Semiotika i Informatika*, 1983, No.20, pp. 35-101 [in Russian].
18. V.K. Finn, Plausible Reasoning in Systems of JSM Type, *Itogi Nauki i Tekhniki, Seriya Informatika*, **15**, 54-101, 1991 [in Russian].
19. V.K. Finn, On a variant of argumentation logic, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1996, No. 5-6, pp. 3-19 [in Russian].
20. V.K. Finn, Synthesis of cognitive procedures and the problem of induction, *Nauchno-Tekhnicheskaya Informatsiya*, ser. 2, 1999, No. 1-2, pp. 8-44 [in Russian].
21. V.K Finn, M. A. Mikheyenkova, On logical means of conceptualization of opinion analysis, *Nauchno-Tekhnicheskaya Informatsiya*, ser. 2, 2002, No.6, pp. 4-21 [in Russian].
22. B.A. Galitsky, S.O. Kuznetsov, and M. V. Samokhin, Analyzing Conflicts with Concept-Based Learning, Proc. 11th Int. Conf. on Conceptual Structures, ICCS'04, F. Dau, M.-L. Mugnier, Eds., Lecture Notes in Artificial Intelligence (2005).
23. B.Ganter, P.A. Grigoriev, S.O. Kuznetsov, M.V. Samokhin, Concept-Based Data-Mining with Scaled Labeled Graphs, *Proc. 12th Int. Conf. on Conceptual Structures (ICCS'04)*, Lecture Notes in Artificial Intelligence, **3127**, 2004, pp. 94-108.
24. B. Ganter and S.O. Kuznetsov, Formalizing Hypotheses with Concepts, *Proc. 8th Int. Conf. on Conceptual Structures, ICCS'98*, G. Mineau and B. Ganter, Eds., Lecture Notes in Artificial Intelligence, **1867**, 2000, pp. 342-356.
25. B. Ganter and S. Kuznetsov, Pattern Structures and Their Projections, *Proc. 9th Int. Conf. on Conceptual Structures, ICCS'01*, G. Stumme and H. Delugach, Eds., Lecture Notes in Artificial Intelligence, **2120** 2001, pp. 129-142.
26. B. Ganter and S.O. Kuznetsov, Hypotheses and Version Spaces, *Proc. 10th Int. Conf. on Conceptual Structures, ICCS'01*, A.de Moor, W. Lex, and B.Ganter, Eds., Lecture Notes in Artificial Intelligence, **2746** 2003, pp. 83-95.
27. B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*, Springer, 1999.
28. T. Gergely, V.K. Finn, On Solver of "Plausible Inference + Deduction" Type, in Intelligent Information-Computing Systems, *Artif. Intel. IFAC Ser. No. 9*, London Pergamon Press, 1984.
29. P.A. Grigoriev, SWORD-systems or JSM-systems for strings employing statistical considerations, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1996, No. 5-6, pp. 45-51 [in Russian].

30. P.A. Grigoriev, On computer forecast of repeated hypophysis adenoma, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1999, No. 1-2, pp. 83-88 [in Russian].
31. P.A. Grigoriev, S.O. Kuznetsov, S.A. Obiedkov, S.A. Yevtushenko, On a Version of Mill's Method of Difference, *Proc. ECAI 2002 Int. Workshop on Advances in Formal Concept Analysis for Knowledge Discovery in Databases*, Lyon, 2002, pp. 26-31.
32. P.A. Grigoriev and S.A. Yevtushenko, Elements of an Agile Discovery Environment, *Proc. 6th International Conference on Discovery Science (DS 2003)*, G. Grieser, Y. Tanaka and A. Yamamoto, Eds., Lecture Notes in Artificial Intelligence, **2843**, 2003, pp. 309–316.
33. P.A. Grigoriev, S.A. Yevtushenko and G.Grieser, QuDA, a data miner's discovery enviornment, Tech. report, FG Intellektik, FB Informatik, Technische Universität Darmstadt, 2003, AIDA 03 06, http://www.intellektik.informatik.tu-darmstadt.de/~peter/QuDA.pdf,
34. V.Ya. Gusakov, S.M. Yakubovich (Gusakova), Galois Connection and Some Theorems on Representing Binary Relations, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1974, No. 7, pp. 3-6 [in Russian].
35. V.Ya. Gusakov, S.M. Yakubovich (Gusakova), On Classification Algorithms, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1976, No. 12, pp. 17-22 [in Russian].
36. S.M. Gusakova, V.K. Finn, On Formalization of Local and Global Similarities, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1986, No. 6, pp. 16-19 [in Russian].
37. S.M. Gusakova, Canonical Representation of Similarities, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1987, No. 9, pp. 19-22 [in Russian].
38. S.M. Gusakova, V.K. Finn, On New Means for Formalization of Local and Global Similarities, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1987, No. 10, pp. 14-22 [in Russian].
39. S.M. Gusakova, V.K. Finn, Similarity and Plausible Reasoning, *Izvestia Akademii Nauk (Tekhnicheskaya Kibernetika)*, 1987, No. 5, pp. 42-63 [in Russian].
40. S.M. Gusakova, M.A. Mikheenkova, V.K. Finn, On Logical Means for Automated Analysis of Opinions, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 2001, No. 5, pp. 4-24 [in Russian].
41. S.M. Gusakova, Paleography with JSM-method. Technical Report, VINITI, 2001.
42. H. Hirsh, Generalizing Version Spaces, *Machine Learning* **17**, 5-46, 1994.
43. S.O. Kuznetsov, On the lattice on graph sets for graphs with ordered vertex labels, Proc. Workshop on Semiotical Aspects of Formalization of Intelligent Activity, Borzhomi (Georgia, USSR), 1988, vol. 1, pp. 204-207 [in Russian].
44. S.O. Kuznetsov, Similarity operation on hypergraphs as as a Basis of Plausible Inference, Proc. 1st Soviet Conference on Artificial Intelligence, 1988, vol. 1, 442-448 [in Russian].
45. S.O. Kuznetsov, Interpretation on Graphs and Complexity Characteristics of a Search for Specific Patterns, *Nauchn. Tekh. Inf., Ser. 2 (Automat. Document. Math. Linguist.)* no. 1, 23-27, 1989 [in Russian].
46. S.O. Kuznetsov, JSM-method as a machine learning method, *Itogi Nauki i Tekhniki, ser. Informatika*, vol. 15, pp.17-50, 1991 [in Russian].
47. S.O. Kuznetsov, Mathematical aspects of concept analysis, Journal of Mathematical Science, Ser. Contemporary Mathematics and Its Applications, **18**, pp. 1654-1698, 1996 [in Russian].
48. S.O. Kuznetsov, Learning of Simple Conceptual Graphs from Positive and Negative Examples. In: J. Zytkow, J. Rauch (eds.), Proc. *Principles of Data Mining and Knowledge Discovery, Third European Conference, PKDD'99*, Lecture Notes in Artificial Intelligence, **1704**, pp. 384-392, 1999.
49. S.O. Kuznetsov, On Computing the Size of a Lattice and Related Decision Problems, *Order*, 2001, **18**(4), pp. 313-321.
50. S.O. Kuznetsov, Machine Learning and Formal Concept Analysis, *Proc. 2nd Int. Conf. on Formal Concept Analysis (ICFCA'04)*, P. Eklund, Ed., Lecture Notes in Artificial Intelligence, **2961**, 2004, pp. 287-312.
51. S.O. Kuznetsov, Learning in Concept Lattices from Positive and Negative Examples, *Discrete Applied Mathematics*, 2004, **142**, pp. 111-125.
52. S.O. Kuznetsov and V.K. Finn, Extension of Expert Systems of JSM-type to Graphs, *Izvestia AN SSSR, ser. Tekhn. Kibern.*, 1988, No. 5, p. 4-11.

53. S.O. Kuznetsov and V.K. Finn, On a model of learning and classification based on similarity operation, *Obozrenie Prikladnoi i Promyshlennoi Matematiki* **3**, no. 1, 66-90, 1996 [in Russian].
54. S.O. Kuznetsov, S.A. Obiedkov, Comparing performance of algorithms for generating concept lattices, *J. Exp. Theor. Artif. Intell.*, 2002, vol. 14, nos. 2-3, pp. 189-216.
55. S.O. Kuznetsov, M.V. Samokhin, Learning Closed Sets of Labeled Graphs for Chemical Applications, Proc. 15th Conference on Inductive Logic Programming, ILP'05, S. Kramer, B. Pfahringer, Eds., Lecture Notes in Artificial Intelligence (2005).
56. A.E. Leibov, Some methods of realization of the similarity operations for chemically oriented expert systems of JSM type, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1996, No. 5-6, pp. 20-32 [in Russian].
57. D. Maier, The Theory of Relational Databases, Computer Science Press, 1983.
58. A.A. Matveev, E.F. Fabrikantova, Algorithmic and Programming Means for Metabolism Forecasting, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 2002, No. 6, pp. 26-34 [in Russian].
59. S.V. Meien, Yu.A. Shreider, Methodological Aspects of Classification Theory, *Problemy Filosofii*, 1976, No. 12, pp. 67-69 [in Russian].
60. D. van der Merwe, S. A. Obiedkov, D. G. Kourie, AddIntent: A New Incremental Algorithm for Constructing Concept Lattices, *Proc. 2nd Int. Conf. on Formal Concept Analysis (ICFCA'04)*, P. Eklund, Ed., Lecture Notes in Artificial Intelligence, **2961**, 2004, pp. 372-385.
61. M.A. Mikheenkova, V.V. Avidon, and S.A. Sukhanova, On Program Realization of the JSM-method of Automated Hypothesis Generation with Nonelement set of attributes, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1984, No. 11, pp. 20-26 [in Russian].
62. M.A. Mikheenkova, V.K. Finn, On a Class of Expert Systems with Incomplete Information, *Izvestia AN SSSR, ser. Tekhn. Kibern.*, 1986, No. 5, p. 82-103 [in Russian].
63. J. S. Mill, *A System of Logic, Ratiocinative and Inductive*, London, 1843.
64. T. Mitchell, Generalization as Search, *Artificial Intelligence* **18**, no. 2, 1982.
65. T. Mitchell, *Machine Learning*, The McGraw-Hill Companies, 1997.
66. J.M. Nicolas, Mutual Dependencies and Some Results on Undecomposable Relations, *Proc. 4th Int. Conf. on Very Large Data Bases*, West Berlin, 1978, p.360-376.
67. S. A. Obiedkov, Modal Logic for Evaluating Formulas in Incomplete Contexts, *Proc. 10th Int. Conf. on Conceptual Structures, ICCS'02*, Lecture Notes in Artificial Intelligence, **2393**, 2002, pp.314-325.
68. S.A. Obiedkov, V. Duquenne, Incremental Construction of the Canonical Implication Basis, Proc. International Conference Journee de lInformatique Messine (JIM03), Metz (2003). To appear in Discrete Applied Mathematics (2005).
69. E.S. Pankratova, D.V. Pankratov, V.K. Finn, I.P. Shabalova, Application of the JSM-method for forecasting high pathogenicity viruses of human papilloma, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 2002, No. 6, pp. 22-25 [in Russian].
70. N.S. Panova, Yu.A. Shreider, On Symbolic Nature of Classification, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1974, No. 12, pp. 3-10 [in Russian].
71. D.V. Popov, V.G. Blinova, E.S. Pankratova, Drug Design: JSM-method of Hypothesis Generation for Antitumor Activity and Toxic Effects Forecast with Respect to Plant Products, FECS 5th Int. Conf. Chem. and Biotechnol. Biologica. Act. Nat. Prod. Sept. 18-23, 1989, Varna, Bulgaria, pp. 437-440.
72. O.M. Polyakov, V.V. Dunaev, Classification Schemes: Synthesis through Relations, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1985, No. 6, pp. 15-21 [in Russian].
73. O.M. Polyakov, Classification Data Model, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1986, No. 9, pp. 13-20 [in Russian].
74. O.M. Polyakov, On Systematization, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1988, No. 12, pp. 21-28 [in Russian].
75. J.R. Quinlan, Induction on Decision Trees, *Machine Learning*, **1**, No. 1, 81-106 (1986).
76. A.A. Raskina, I.S. Sidorov, and Yu.A. Shreider, Semantical Foundations of Object-Attribute Languages, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1976, No. 5, pp. 18-25 [in Russian].
77. Yu.A. Shreider, Mathematical Model of Classification Theory, VINITI, Moscow,1968, pp.1-36 [in Russian].

78. Yu.A. Shreider, Equality, Similarity, Order, Moscow, Nauka, 1971 [in Russian].
79. Yu.A. Shreider, Logic of Classification, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1973, No. 5, pp. 3-7 [in Russian].
80. Yu.A. Shreider, Algebra of Classification, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1974, No. 9, pp. 3-6 [in Russian].
81. Yu.A. Shreider, Typology as a Base of Classification, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1981, No. 11, pp. 1-5 [in Russian].
82. D. Soergel, Mathematical Analysis of Documentation Systems, *Inf. Stor. Retr.*, 1967, No. 3, pp. 129-173.
83. M.Sh. Tsalenko, Semantical and Mathematical Models of Databases, *Itogi Nauki i Tekhniki*, Ser. Informatika, vol. 9, 1985, pp. 3-207 [in Russian].
84. M.Sh. Tsalenko, Canonical Representation of Irreducible Systems and Classification Schemes *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1985, No. 1, pp. 30-34 [in Russian].
85. A. Tversky, Features of Similarity, *Psychological Review*, vol. 84, no. 4, pp. 327-352, 1977.
86. D.V. Vinogradov, Logical Programms for Quasi-Axiomatic Theories, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1999, No. 1-2, pp. 61-64 [in Russian].
87. D.V. Vinogradov, Formalization of Plausible Reasoning in FOPL, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 2000, No. 11, pp. 17-20 [in Russian].
88. D.V. Vinogradov, Correct Logical Programms for Plausible Reasoning, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 2001, No. 5, pp. 25-27 [in Russian].
89. R. Wille, Restructuring Lattice Theory: an Approach Based on Hierarchies of Concepts, In: *Ordered Sets* (I. Rival, ed.), Reidel, Dordrecht–Boston (1982), 445-470
90. R. Wille, Conceptual Structures of multicontexts. In: P. W. Eklund, G. Ellis, G. Mann (eds.): *Conceptual Structure Representation as Interlingua.* Springer, Berlin-Heidelberg-New York (1996), 23-39
91. S.M. Yakubovich (Gusakova), Axiomatic Theory of Similarity, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1968, No. 10, pp. 15-19 [in Russian].
92. S.M. Yakubovich (Gusakova), On Properties of Conjugated Tolerance Spaces, *Information Problems of Semiotics, Linguistics, and Machine Translation*, 1971, vol. 1, pp. 116-123 [in Russian].
93. S.M. Yakubovich (Gusakova), Decomposition of Tolerance Spaces, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1973, No. 3, pp. 26-28 [in Russian].
94. M.I. Zabezhailo, V.K. Finn, A.V. Avidon, V.G. Blinova et al., On experiments with a database using JSM-method of hypothesis generation, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1983, No. 2, pp. 28-32.
95. M.I. Zabezhailo, V.G. Ivashko, S.O. Kuznetsov, M.A. Mikheenkova, K.P. Khazanovskii, O.M. Anshakov, Algorithmic and Program Means of JSM-method Of Automatic Hypothesis Generation, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1987, No. 10, pp. 1-14.
96. M.I. Zabezhailo, On search problems arising in automatic hypothesis generation with JSM-method, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1988, No.1, pp. 28-31 [in Russian].
97. M.I. Zabezhailo, Formal models of reasoning in decision making: Applications of the JSM-method in systems of intelligent control and automation of scientific research, *Nauchno-Tekhnicheskaya Informatsiya*, Ser. 2, 1996, No. 5-6, pp. 20-32 [in Russian].
98. E.C. Zeeman, The Topology of Brain and Visual Perception, In *The Topology of 3-Manifolds and Related Topics*, K.M. Ford, ed., pp. 240-256, Prentice Hall, 1965.