Automatic Documentation and Mathematical Linguistics
Vol. 29, No. 3, pp. 11-20, 1995

Nauchno-Tekhnicheskaya Informatsiya, Seriya 2
No. 5, pp. 8-13, 1995
UDC 007:510.64:159.955

SIMILARITY IN THE GENERALIZED JSM METHOD AND ITS GENERATION ALGORITHMS

S. M. Gusakova and S. O. Kuznetsov

The article considers generation of hypotheses in the "generalized" JSM method, which consists in generation of "conditional" hypotheses about cause-and-effect relations for positive and negative examples. The conditionality of the hypotheses is that the causes about which the hypotheses are advanced "operate" (cause the manifestation of a property) in the absence of certain substructures called "brakes" of these causes. In contrast to "ordinary" negative causes expressing similarities of negative examples, the brakes possess some "cause-and-effect structure," bearing their own opposite: conditional positive causes. An algebraic description of similarity in the generalized JSM method is suggested, and also algorithms for generating generalized hypotheses (together with analysis of computational complexity).

1. THE GENERALIZED METHOD AND SIMILARITY

The basic idea of the JSM method [1,2], to establish cause-and-effect relations between substructures of structured objects and a subset of properties of these objects on the basis of determination of the objects' essential similarity, is naturally expanded in the generalized JSM method, the principle of which consists in taking into account the context (in the structure of the object) when establishing the cause-and-effect relations.

Such consideration of context is due to realities of the problems to be solved with the help of the JSM method from the field of chemistry, pharmacology, and sociology.

A formal definition of a generalized JSM predicate was suggested in [3]. Its informal meaning amounts to the following: in the presence of a description of some positive and negative examples for some class of objects (characterized by a particular property), "conditional" hypotheses are generated. Their conditionality is that the causes "operate" (cause manifestation of the property) in the absence of certain substructures called "brakes" of these causes. In contrast to "ordinary" (see [1,2]) negative causes expressing similarities of negative examples, the brakes possess some "causal (cause-and-effect) structure," bearing their own opposite: conditional positive causes. Thus, the binary relation $V \Rightarrow_2 W$, which is read as "the subobject V is a cause of the presence of a set of properties W," of the simple JSM method is replaced by the ternary relation $T(V, \varkappa, W)$, which is read as "V is a cause of the presence of a set of properties W in the absence of 'brakes' $\varkappa$," in the generalized JSM method.

Restrictions on the length of the article and the wish to avoid repeating the material set forth forces the authors to refer readers unfamiliar with the JSM method who want to get to know it to [1-4], which contain all of the necessary information.

However, the minimum information, concepts, and notations necessary for reading the present work are given below.

The JSM method of automatic hypothesis generation is used to analyze and process data represented in a data base with incomplete information. The class of problems to be solved by the JSM method is characterized by the following conditions:

- the subject area's data are well structured;
- there is a set of objects and a set of properties in each of which certain operations are assigned;
- in the set of objects, the relation "to be a subobject of an object" is determined;
- in the product of sets of objects and properties, the partially determined relations "an object possesses

a set of properties" (represented by $\overset{*}{\Rightarrow}_1$) and "a subobject is the cause of a set of properties" (represented by $\overset{*}{\Rightarrow}_2$) are assigned;

- there is a set of positive and negative examples of the first relation, i.e., a set of pairs of the type (x, y), where x is an object and y is a property, that satisfy and do not satisfy the relation;

The examples are obtained empirically.

The presence of properties of the objects or their absence is determined by "positive" (+) and "negative" (−) causes, i.e., the presence of some phenomenon is due to some set of (+) causes, and the absence of a phenomenon is due either to the presence of (−) causes or the absence of corresponding (+) causes.

As a rule, problems of this class arise in poorly formalized fields of knowledge.

The initial data for the JSM method are matrices of partially determined relations: $\overset{*}{\Rightarrow}_1$ and $\overset{*}{\Rightarrow}_2$. The rows of the matrices correspond to objects in the former case and subobjects in the latter; and the columns, to elementary properties. At the intersection of the i-th row and the j-th column, there is +1 if $x_i \overset{*}{\underset{e}{\Rightarrow}} y_j$ is fulfilled, −1 if it is not fulfilled, and τ if it is not known whether or not $x_i \overset{*}{\underset{e}{\Rightarrow}} y_j$ (l = 1, 2) takes place. The subrow of the matrix corresponding to all +1 (−1) is called a positive (negative) example.

With the help of logical combinatorial algorithms, the simple JSM method generates hypotheses of the type

$$J_{(\nu,n)}(C \Rightarrow_1 A),\ J_{(\nu,n)}(\overline{C} \Rightarrow_2 A),$$
$$J_{(\tau,n)}(C \Rightarrow_1 A) \text{ and } J_{(\tau,n)}(\overline{C} \Rightarrow_2 A),$$

where J is a one-place operator; and ν = (1, −1, 0) and τ are types of truth values denoting actual truth (+1), actual falsehood (−1), actual contradiction (0), and indeterminacy (τ).

With the help of significantly more complex algorithms, the generalized method, which reflects deeper ideas about the nature of cause-and-effect relations, generates hypotheses of the type

$$J_{(\nu,n)} T(\overline{C}, \overline{\varkappa}, A) \text{ and } J_{(\tau,n)} T(\overline{C}, \overline{\varkappa}, A),$$

where C and A are an object and a set of properties, respectively; $\overline{C}$ is a subobject; and $\overline{\varkappa}$ is a set of brakes.

In various complicated versions, the logical combinatorial algorithms of the JSM method realize a simple idea of J. S. Mill (hence the name JSM method): the cause of similarity of objects' properties is similarity of their structures.

Therefore, the concept of similarity is one of the central concepts of JSM theory, and determination of essential similarity in specific applied problems and for specific data structures is one of the central problems of creating an applied JSM system.

JSM systems use the operation of similarity to distinguish the causes of subobjects, and local (relation) and global (family of sets) similarities to construct algorithms.

In JSM methods, the similarity operation is understood as an idempotent, commutative, and associative operation on pairs of objects. These properties make it possible to unambiguously express similarity of a set of objects in terms of pair similarities, regardless of the order of the objects' arrangement in the data base.

The operation of similarity is represented as Π. A strict definition of it can be found, for example, in [5].

Definition 1. Objects $X_{i_1}, \ldots, X_{i_n}$ are locally similar if $\prod_{j=1}^{n} X_{i_j} \neq \varnothing$.

Local similarity is an n-ary tolerance relation (see [6]).

Definition 2. Objects $X_{i_1}, \ldots, X_{i_k}$ are globally similar if

$$\left(\prod_{j=1}^{k} X_{i_j} = h\right) \& \forall X_{i_m}\left(\left(\prod_{j=1}^{k} X_{i_j}\right) \Pi X_{i_m} = h\right) \&$$
$$\& \left(\&_{r=1}^{k} \neg (m = r)\right).$$

Global similarity links all objects that contain a certain subobject (result of the similarity operation) and therefore is not a relation.

The structure of global similarity is described by a family of sets (M, G), where M is the set of objects under consideration; G = $\{g_i\}$; and g_i is the subset of globally similar objects. We recommend [7,8] to readers who want to familiarize themselves with the theory of global and local similarity.

For further exposition, it will be convenient for us to use the operational interpretation of local and global similarity (see [5]), in accordance with which:

Definition 1'. h is local similarity of $X_1, ..., X_n$ from S (S is a set of objects), if

$$\prod_{i=1}^{n} X_i = h.$$

Definition 2'. $\langle h, \{X_1, ..., X_k\}\rangle$ is global similarity, if $X_1, ..., X_k \in S$, $\prod_{i=1}^{k} X_i = h$, and $\forall Y \in S\backslash\{X_1, ..., X_k\}$, $Y \sqcap h \neq h$ takes place.

The logical combinatorial algorithms realizing the simple JSM method are based on finding the global similarity of objects possessing some property.

Algorithms for realizing the generalized JSM method seek triads in the relation T(V, ×, W). The set of triads of this relation is the region of truth of a generalized JSM similarity predicate, which is represented as $M^{+}_{ag,n}$ (we are considering the case of a positive predicate, remembering that all of the reasoning can be transferred to the case of a negative predicate $M^{-}_{ag,n}$ with no difficulty).

On the strength of the capaciousness of the indicated predicate's formulation, we will give it here in a simplified form. For an exact formulation, we refer the reader to [9], which is devoted to study of this predicate.

We will represent the predicate $M^{+}_{ag,n}(V, \times, W)$ in the form of a conjunction of several parts, to which we will give mnemonic notations for the sake of convenience:

$$M^{+}_{ag,n}(V, \times, W) = EX\&ED\&CE\&B\mathit{l}.$$

EX describes a set of examples of the type $J_{(1,\ n)}(X_i \underset{e}{\Rightarrow} Y_j)$ which are the basis of a plausible deduction;

ED is an empirical dependence between V, W, and × expressing the fact that "V is a cause of W in the absence of brakes × = $\{V_1, ..., V_r\}$.

This dependence has the form

$$\forall X \forall Y((J_{(1,\ n)}(X \underset{e}{\Rightarrow} Y) \& \forall U(J_{(1,\ n)}(X \underset{e}{\Rightarrow} Y) \rightarrow U \subseteq Y) \& V \subset$$
$$\subset X) \& \neg(V_1 \subset X \vee \ldots \vee V_r \subset X)) \rightarrow W \subseteq Y \& W \neq \varnothing).$$

CE is the condition of exhaustibility, which provides for consideration of all suitable examples from the data base:

$$\bigvee_{i=1}^{k} (X = Z_i).$$

B*l* describes the set of examples that generate brakes and includes two parts, Bl_1 and Bl_2, which describe sets of positive and negative examples, respectively:

$$Bl_1: \underset{j=1}{\overset{s}{\&}} W_j ((\neg J_{(1,n)}(Z_j \Rightarrow_t W_j) \& W \subseteq W_j) \& V \subset Z_j,$$
$$Bl_2: (J_{(1,n)}(Z_j \Rightarrow_t W_j) \& \neg(W \subseteq W_j)) \& V \subset Z_j,$$
$$Bl: (Bl_1 \vee Bl_2) \& \left(V_q = \bigcap_{p=1}^{l} Z_{j_p}\right) \& V \subset V_q.$$

The rest of the fragments of the predicate $M^{+}_{ag,n}$ characterize the conditions of exhaustibility and imaginariness for the brakes.

The essence of the predicate $M^{+}_{ag,n}$ is that a subobject found as operational similarity of objects $X_1, ..., X_k$ is a cause of the property W, if none of the indicated objects contains brakes from × = $\{V_1, ..., V_r\}$, $V \subset V_i$, i = 1, ..., r. The brakes, in turn, are found from objects $Z_1, ..., Z_s$, for negative, as well as positive examples containing V,

but not possessing the property W.

If brakes for V cannot be extracted from the examples available in the data base, then V is not a cause of the property W in the generalized sense.

At the algorithmic level, the main difficulties are due to the fact that when actions checking the conditions of the predicate are performed in succession a subobject V found as a candidate for a cause of the property W may be thrown out. The next step in this case will be conducted on a set of objects $X_1, ..., X_n$ without the objects that contain brakes, and this can lead to the appearance of a new candidate for a cause, which, in turn, entails a search for new brakes. The presence in the predicate $M^*_{a,n}$ of a condition according to which brakes are also sought among positive examples that do not possess the property W complicates the algorithm even more.

Investigation of the properties determined in a set of objects by a predicate of generalized similarity provides the opportunity to distinguish different situations that arise depending on the structure of the data base, and among these situations to find those for which the algorithm realizing the generalized JSM method is simpler than in the general case.

Therefore, we will start this investigation.

Analyzing the predicate $M^*_{ag,n}$ of generalized similarity from the point of view of the nature of similarities generated by it in a set of objects (so-called procedural similarities), we must note the following significant points:

1) procedural similarities in a set of objects are global similarities, on the strength of the condition of exhaustibility;

2) procedural similarity determining a cause is some composite of simple similarities. This follows from the trinariness of the predicate and the relations between v and ×.

We will introduce the following notations:

$W = \{w_1, ..., w_a\}$ is a set of elementary properties.

$\bar{W} = 2^W = \{\bar{w}_1, ..., \bar{w}_{\bar{a}}\}$ is the set of all subsets of the set W.

$$\overset{\sigma}{G}{}^l = \{\overset{\sigma}{g}_k\}, \quad k = 1, \ldots, \bar{k}; \quad l = 1, \ldots, \bar{a}; \quad \sigma = (+, -, \pm);$$

$\overset{\sigma}{g}_k = (\overset{\sigma}{h}_k{}^l, \overset{\sigma}{S}_k{}^l)$ is global similarity, with $\overset{+}{S}_k{}^l$ consisting of objects taken from the subset of positive examples of the type $J_{\langle 1,0\rangle}(X \Rightarrow_1 \bar{W}_i)$, $\overline{S}_k{}^l$ of objects from the subset of negative examples also for the property $\bar{w}_i$, and $\overset{\pm}{S}_k{}^l$ of objects from the subset of positive, as well as negative examples of this property.

Thus, $\{\overset{+}{h}_i\}$ is the set of all maximum positive "intersections" (or rather, the results of the similarity operation); and $\{\overline{h}^l\}$, of all negative ones, i.e.,

$$\overset{+}{h}_k{}^l = \prod_{x_n \in \overset{+}{S}_k^l} X_n, \quad \overset{+}{S}_k{}^l = \{X_n \mid (J_{\langle 1,0\rangle}(X_n \Rightarrow_1 Y) \,\&\, (\bar{W}_l \subseteq Y)\};$$

$$D^{l,j} = \{d_m^{l,j}\}, \quad m = 1, \ldots, \bar{m}; \quad l, j = 1, \ldots, \bar{a};$$

$$l \neq j; \quad \bar{W}_l \subseteq \bar{W}_j;$$

$$d_m^{l,j} = \{\overset{+}{h}_m^{l,j}, \overset{+}{S}_m^{l,j}\}; \quad \overset{+}{h}_m^{l,j} = \prod_{x_n \in \overset{+}{S}_m^{l,j}} X_n;$$

$$\overset{+}{S}_m^{l,j} = \{X_n \mid ((J_{\langle 1,0\rangle}(X_n \Rightarrow_1 Y_1) \,\&\, (\bar{W}_l \subseteq Y_1)) \vee$$

$$\vee ((J_{\langle 1,0\rangle}(X_n \Rightarrow_1 Y_2) \,\&\, (\bar{W}_j \subseteq Y_2))\},$$

i.e., $\overset{+}{S}_m^{l,j}$ consists of objects taken from positive examples for properties $\bar{w}_i$ and $\bar{w}_j$, with $\bar{w}_i \not\subseteq \bar{w}_j$.

The need to consider the property D^{ij} is determined by the part of the predicate Bl_2 for seeking brakes.

We will define two ternary predicates:

$$\zeta_1(\overset{+}{g}_k{}^l, \overline{g}_n{}^l, \overset{\pm}{g}_m{}^l) =$$

$$= 1 \Rightarrow (\overset{+}{h}_k{}^l \subset \overline{h}_n{}^l) \,\&\, (\overline{h}_n{}^l = \overset{\pm}{h}_m{}^l) \,\&\, \forall l \neg (\overline{h}_l{}^l \subset \overline{h}_n{}^l);$$

$$\zeta_2(\overset{+}{g}_k^l, \overset{+}{g}_{n'}^j, d_m^{l,j}) =$$

$$=1\Leftrightarrow\left(\overset{+}{h}{}_k^l\subset\overset{+}{h}{}_{n'}^{l'}\right)\&\left(\overset{+}{h}{}_{n'}^{l'}=h_m^{l,l'}\right)\&\forall l'\neg\left(\overset{+}{h}{}_i^{l'}\subset\overset{+}{h}{}_{n'}^{l'}\right).$$

It is clear that ζ_1 and ζ_2 are equal to 0 if even one of the terms of the conjunction is equal to zero (1 and 0 are the truth values "truth" and "falsehood").

We will define global similarity $\overset{+}{P}{}^l=\{\overset{+}{p}{}_k^l\}$, where $\overset{+}{p}{}_k^l=(\overset{+}{v}{}_k^l,\ \overset{+}{\Sigma}{}_k^l)$ as follows:

$$\overset{+}{v}{}_k^l=\overset{+}{h}{}_k^l,$$
$$\overset{+}{\Sigma}{}_k^l=\overset{+}{S}{}_k^l\setminus\left[\left(\bigcup_{m,n}(\overset{\pm}{S}{}_m^l\setminus\overline{S}{}_n^l)\mid\zeta_1(\overset{+}{g}{}_k^l,\ \overline{g}{}_n^l,\ \overset{\pm}{g}{}_m^l)=1\right)\cup\right.$$
$$\left.\left(\bigcup_{m',n',l'}(S_m^{l,l'}\setminus\overset{+}{S}{}_{n'}^{l'})\mid\zeta_2\left(\overset{+}{g}{}_k^l,\ \overset{+}{g}{}_{n'}^{l'},\ d_m^{l,l'}\right)=1\right)\right];$$
$$\overset{+}{p}{}_k^l=\varnothing,\quad\text{if}\quad(\forall n\forall m\zeta_1(\overset{+}{g}{}_k^l,\ \overline{g}{}_n^l,\ \overset{\pm}{g}{}_m^l)=0)\&$$
$$\&\left(\forall n'\forall m'\zeta_2\left(\overset{+}{g}{}_k^l,\ \overline{g}{}_n^{l'},\ d_m^{l,l'}\right)=0\right)\vee\exists r\,(\overset{+}{\Sigma}{}_k^l=\overset{+}{S}{}_r^l).$$

Proposition 1. For any $\overset{+}{p}{}_k^l=\{\overset{+}{v}{}_k^l,\ \overset{+}{\Sigma}{}_k^l\}$, $T\,(\overset{+}{v}{}_k^l,\ \varkappa,\ \tilde{w}_t)=1$ takes place, where

$$\varkappa=\{(\overline{h}_n^l\mid\exists m\zeta_1(\overset{+}{g}{}_k^l,\ \overline{g}{}_n^l,\ \overset{\pm}{g}{}_m^l)=1\}\cup$$
$$\cup\{\overline{h}_{n'}^{l'}\mid\exists m'\zeta_2\left(\overset{+}{g}{}_k^l,\ \overset{+}{g}{}_{n'}^{l'},\ d_m^{l,l'}\right)=1)\}\}.$$

Proof.

1. From the definition of similarity $\overset{+}{G}{}^l$, it follows that for each i and k the set $\overset{+}{S}{}_k^l$ constitutes exactly the set of examples described in the part *Ex* of the predicate $M^+_{ag,n}$. The condition of exhaustibility (CE) is provided by the globalness of the similarity $\overset{+}{g}{}_k^l$. Consequently, $\overset{+}{v}{}_k^l=\overset{+}{h}{}_k^l$ satisfies the conditions of $M^+_{ag,n}$, since it is the maximum operational similarity of positive examples.

2. From the definition of $\overline{g}_n^i$, it follows that $\overline{h}_n^i$ is the maximum operational similarity of negative examples. The condition of minimality $\overline{h}_n^i$ with respect to embedding, which is imposed on the brakes, is provided by fulfillment of the predicate ζ_1.

3. From the definition of d_m^{ij} it follows that $\overset{+}{h}{}_n^l$ is the maximum operational similarity of positive examples that do not possess the property $\bar{w}_1$.

4. From the definition of $\overset{+}{\Sigma}{}^l$ it follows that

$$\forall n\forall X\in\overset{+}{\Sigma}{}_k^l\neg(\overline{h}_n^l\subset X)\&\forall n'\forall X'\in\overset{+}{\Sigma}{}_k^l\neg\left(\overset{+}{h}{}_{n'}^{l'}\subset X'\right).$$

Fulfillment of conditions (1-4) provides for $T\,(\overset{+}{v}{}_k^l,\ \varkappa,\ \tilde{w}_t)=1$ with $\varkappa$ described in the formulation of the proposition.

Now we will show that if $T(v',\ \varkappa',\ \bar{w}_i)=1$, then $\exists k\ v'=\overset{+}{h}{}_k^l$.

$$\exists n\forall m\exists n'\exists m'\varkappa'=(\{\overline{h}_n^l\mid\zeta_1(\overset{+}{g}{}_k^l,\ \overline{g}{}_n^l,\ \overset{\pm}{g}{}_m^l)=1\}\cup$$
$$\cup\{\overline{h}_{n'}^{l'}\mid\zeta_2\left(\overset{+}{g}{}_k^l,\ \overset{+}{g}{}_{n'}^{l'},\ d_m^{l,l'}\right)=1\}\}).$$

By the definition of the relation T, v is the maximum operational similarity of some subset of positive examples, but all such similarities are included in the global similarity $\overset{+}{G}{}^l$, consequently $\exists k\ v=\overset{+}{h}{}_k^l$.

Reasoning analogously, we find that

$$\forall v_q\in\varkappa'\,(\exists n\,(v_q=\overline{h}_n^l))\vee\left(\exists l'\exists n'\,\left(v_q=\overset{+}{h}{}_{n'}^{l'}\right)\right).$$

Fulfillment of the predicates ζ_1 and ζ_2 follows from the definition of similarities $\overset{0}{G}$ and $\overset{+}{P}$.

Strictly speaking, proposition 1 follows directly from the definition of the relation T, the predicate $M^+_{ag,n}$

and the similarities $\overset{\delta}{G}$ and $\overset{-}{P}$, which we can see directly from the proof.

Thus, the global similarity $\overset{+}{P}{}^{l}$ is generated by all those and only those objects the operational similarity of which is expressed by the cause $\overset{+}{v}{}^{l}$ of the property $\bar{w}^{l}$ in the absence of brakes $\varkappa$.

As we can see from the definition, the global similarity $\overset{+}{P}{}^{l}$ depends significantly on the property $\bar{w}^{l}$, therefore the structure of the set of properties in the data base largely determines the complexity of the algorithmic procedure for finding $\overset{+}{P}{}^{l}$. So, it is obvious that for the case $W = \{w\}$, $\bar{W} = W$ this procedure is significantly simplified. In fact, the predicate of similarity $M^{+}_{\text{sg,n}}$ is primarily simplified, since the addition Bl_2 is removed from it.

For this case, the similarity $\overset{+}{P} = (\overset{+}{p}_j)$ takes the form

$$\overset{+}{P}_j = \{\overset{+}{v} = \overset{+}{h}_j;\ \overset{+}{\Sigma}_j = \overset{+}{S}_j \setminus$$
$$\setminus \left(\bigcup_{m,n} (\overset{\pm}{S}_m{}^{l} \setminus \bar{S}_n{}^{l}) \mid \zeta_1 (\overset{+}{g}_j{}^{l},\ \bar{g}_n{}^{l},\ \overset{\pm}{g}_m{}^{l}) = 1 \right).$$

$\overset{+}{P}$ is formed with the help of similarities $\overset{+}{G} = (\overset{+}{g}_j)$, $\bar{G} = (\bar{g}_n)$ and $\overset{\pm}{G} = (\overset{\pm}{g}_m)$.

If the similarity P is found, then all of the triads $\{v, \varkappa, w\}$ satisfying the predicate T are found.

2. ALGORITHMS FOR CONSTRUCTING GENERALIZED HYPOTHESES

In essence, the way the definition of similarity $\overset{+}{P}$ is constructed incorporates algorithm 1 for finding it, which consists in a sequence of the following actions:

1. Finding similarities $\overset{+}{G}$ and $\bar{G}$, which is done with the help of the so-called algorithm of maximum intersections realized in the version of the JSM system for the simple method.
2. Verification of the predicate ζ_1.
3. Construction of the similarity $\overset{+}{P}$.
4. Verification of the condition $\forall r\ (\overset{+}{\Sigma}{}^{l}_{k} \neq \overset{+}{S}_r{}^{l})$.

We will demonstrate the algorithm's operation for the case of one property on the following example:

A set of objects: $\{x_1, \ldots, x_{11}\}$: $\Omega^{+} = \{x_1, x_2, x_3, x_4, x_5\}$, $\Omega^{-} = \{x_6, x_7, \ldots, x_{11}\}$; x_1 = vaa′, x_2 = vbdd′, x_3 = vdb′d″m, x_4 = vd′cm, x_5 = vd″c′m, x_6 = vae, x_7 = vae′, x_8 = vbdf, x_9 = vbdf′, x_{10} = vd′gg′, x_{11} = vd′gg″.

We find the similarity $\overset{+}{G}$ from positive examples:

$$\overset{+}{g}_1 = \{v;\ x_1, x_2, x_3, x_4, x_5\};$$
$$\overset{+}{g}_2 = \{vbd';\ x_2, x_3\};$$
$$\overset{+}{g}_3 = \{vd';\ x_2, x_3, x_4\};$$
$$\overset{+}{g}_4 = \{vm;\ x_3, x_4, x_5\};$$
$$\overset{+}{g}_5 = \{vd''m;\ x_3, x_5\}.$$

We find the similarity $\bar{G}$ from negative examples:

$$\bar{g}_1 = \{v;\ x_6, x_7, x_8, x_9, x_{10}, x_{11}\};$$
$$\bar{g}_2 = \{vae;\ x_6, x_7\};$$
$$\bar{g}_3 = \{vbd;\ x_8, x_9\};$$
$$\bar{g}_4 = \{vd'g;\ x_{10}, x_{11}\}.$$

The similarity $\overset{\pm}{G}$ is represented in the following way:

$$\overset{\pm}{g}_1=\{v;\ x_1, \ldots, x_{11}\};$$
$$\overset{\pm}{g}_2=\{va;\ x_1, x_6, x_7\};$$
$$\overset{\pm}{g}_3=\{vb;\ x_2, x_3, x_9, x_8\};$$
$$\overset{\pm}{g}_4=\{vbd';\ x_3, x_5\};$$
$$\overset{\pm}{g}_5=\{vbd;\ x_2, x_8, x_9\};$$
$$\overset{\pm}{g}_6=\{vd';\ x_2, x_3, x_4, x_{10}, x_{11}\};$$
$$\overset{\pm}{g}_7=\{vd''m;\ x_9, x_6\};$$
$$\overset{\pm}{g}_8=\{vd'g;\ x_{10}, x_{11}\}.$$

We will let Th_ζ represent the set of triads $\{\overset{+}{g}_i, \overline{g}_j, \overset{\pm}{g}_m\}$ for which $\overset{+}{\zeta}$ is true. In the given example, it is the following:

$$Th_\zeta=\{(\overset{+}{g}_1, \overline{g}_2, \overset{\pm}{g}_2),\ (\overset{+}{g}_1, \overline{g}_3, \overset{\pm}{g}_3),\ (\overset{+}{g}_1, \overline{g}_4, \overset{\pm}{g}_4),$$
$$(\overset{+}{g}_3, \overline{g}_4, \overset{\pm}{g}_4)\}.$$

Now we can construct the similarities $\overset{+}{P}$:

$$\overset{+}{P}_1=\{v;\ \overset{+}{\Sigma}_1\{\{x_1, x_2, x_3, x_4, x_5\}\setminus[\{x_1, x_6, x_7\setminus x_6, x_7\}\cup$$
$$\cup\{x_2, x_8, x_9\setminus x_8, x_9\}\cup\{x_{10}, x_{11}\setminus x_{10}, x_{11}\}]\}=$$
$$=\{x_2, x_4, x_5\}\}.$$

But since $\overset{+}{\Sigma}_1=\overset{+}{S}_4$, then $\overset{+}{P}_1=\varnothing$.

$\overset{+}{P}_2=\varnothing$, since $\forall j\neg(\overset{+}{h}_1\subset\overline{h}_j)$ $j=1, \ldots, 4$.

$\overset{+}{P}_3=\{vd';\ \overset{+}{\Sigma}_3\{\{x_2, x_3, x_4\}\setminus\{x_{10}, x_{11}\setminus x_{10}, x_{11}\}\}=\{x_2, x_3, x_4\}$.

$\forall l\ \overset{+}{\Sigma}_3\neq\overset{+}{S}_l$ $(l=1, \ldots, 5)$, consequently $\overset{+}{P}_3\neq 0$.

$\overset{+}{P}_4=\varnothing$ and $\overset{+}{P}_5=\varnothing$ for the same reason that $\overset{+}{P}_2=\varnothing$.

As a result of the algorithm's operation, we find one triad: {vd', x = vd'g, w} that satisfies the predicate $M^+_{zg,n}$ and, consequently, belongs to the relation T.

An increase in the set of elementary properties even to two elements entails a significant increase in the diversity of situations and, consequently, complication of the algorithm.

We will figure an upper estimate of the number of computer operations necessary to generate generalized hypotheses with the help of the indicated procedure for an arbitrary set of positive examples Ω^+, set of negative examples Ω^-, and set of elementary structures U (subsets of which are positive and negative examples). In [5], the ZO algorithm was suggested for seeking intersections (similarities), the time requirement of which is linear in relation to the number of intersections generated. Supposing that similarities (intersections) are sought with the help of the ZO algorithm, we will determine that finding the set $\overset{+}{G}$ of all positive similarities requires spending

$$O(|\overset{+}{G}|\cdot|U|\cdot|\Omega^+|)$$

computer operations; finding the set $\overline{G}$ of all negative similarities requires spending

$$O(|\overline{G}|\cdot|U|\cdot|\Omega^-|)$$

computer operations; and finding the set of similarities $\overset{\pm}{G}$ of all examples (positive and negative) requires spending

$$O(|\overset{\pm}{G}|\cdot|U|\cdot(|\Omega^+|+|\Omega^-|))=$$
$$=O(|\overset{+}{G}|\cdot|\overline{G}|\cdot|U|\cdot(|\Omega^+|+|\Omega^-|))$$

computer operations. In the worst case, generating each set $\overset{+}{\Sigma}_j$ requires

$$O(|\overset{+}{G}_j|\cdot|\overset{+}{\overline{G}}|\cdot\mathrm{pol}_1(|U|, |\Omega^+|, |\Omega^-|))=$$
$$=(|\overset{+}{G}|^2\cdot|\overline{G}|^2\,\mathrm{pol}_1(|U|, |\Omega^+|, |\Omega^-|))$$

computer operations, where $\mathrm{pol}_1(\cdot)$ is some small-degree polynomial depending on $|U|$, $|\Omega^+|$, and $|\Omega^-|$ and determined by a computer model.

We can suggest more efficient algorithms for constructing generalized hypotheses. The speed of these algorithms depends significantly on the type of similarity operation (even if one considers the time required to perform such operations to be the same). First we will give a description of such an algorithm for constructing generalized positive hypotheses in the case of representation of data by sets and the presence of just one property.

Algorithm 2

1. All (+) intersections are constructed with the help of the ZO algorithm [5].
2. For each (+) intersection V:

2.1. Negative examples containing it are found.

2.2. All intersections of negative examples that are minimal with respect to embedding and contain V (i.e., brakes) are sought: Let N be the union of all negative examples containing V. Then the arbitrary minimal intersection of negative examples containing V has the form $X^t=\bigcap_j X_{tj}$, where $X_{ij}\in X_i=\{X|X\in\Omega^-, (V\cup\{q_i\})\subseteq X\}$, where $q_i\in N$. Naturally, $\bigcup_j X_{tj}$ is determined, if $j\geq 2$.

In order to rule out generation of the same intersections and thus to accelerate the process of generating them, in constructing the next set $\bigcup_j X_{tj}$, the corresponding "addition" to v can be chosen not from N, but from N minus all of the preceding "additions" and generated intersections, or formally, $N_{k+1}=N_k\setminus(\{q_{k+1}\}\cup\bigcap_j X_{k_j})$.

2.3. From the set of (+) examples containing V, all of those that contain any brakes are removed, and the remaining (+) examples are reintersected. If the result coincides with V, then V, together with the minimal intersections found, is the generalized hypothesis; if not, then no (+) generalized hypothesis exists in relation to the intersection V, and we move on to the next (+) intersection.

The time complexity for constructing one generalized positive hypothesis for a fixed (+) intersection V in steps 2.1-2.3 of this algorithm is

$$O(N\cdot|\Omega^-|).$$

Thus, the complexity of constructing all positive generalized hypotheses is

$$O(|\overset{+}{G}|\cdot N\cdot|\Omega^-|).$$

We will consider the algorithm's action for the sets of positive and negative examples given above. In this case, the set N is N = {a, b, d, d', e, f, f', g, g', g'', v}. Suppose that in the second stage of the algorithm's action we chose the (+) intersection v. Then the minimal intersections of (−) examples containing v are found as

$$X^a=\bigcap_j X_j^a, \text{ where } X_j^a\in X_a=\{X\in\Omega^-|\{v, a\}\subset X\}=\{v, a\},$$
$$N_3=N\setminus\{a\}=\{b, d, d', e, e'f, f', g, g', g''\};$$
$$X^b=\bigcap_j X_j^b, \text{ where } X_j^b\in X_b=\{X\in\Omega^-|\{v, b\}\subset X\}=\{v, b, d\},$$
$$N_4=N_3\setminus\{b, d\}=\{d', e, e'f, f', g, g', g''\}.$$

The intersection corresponding to {v, d} is not generated, since in this step $d\notin N$.

$$X^{d'}=\bigcap_j X_j^{d'}, \text{ where } X_j^{d'}\in X_{d'}=\{X\in\Omega^-|\{v, d'\}\subset X\}=$$
$$=\{v, d', g\},\ N_4=N_3\setminus\{d', g\}=\{e, e'f, f', g', g''\};$$
$$X^e=\bigcap_j X_j^e, \text{ where } X_j^e\in X_e=\{X\in\Omega^-|\{v, e\}\subset X\},$$
$$X^e \text{ is not determined},\quad N_5=N_4\setminus\{e\}=\{e'f, f', g', g''\};$$

$$X^{e'} = \bigcap_j X_j^{e'}, \text{ where } X_j^{e'} \in X_{e'} = \{X \in \Omega^- \mid \{v,\ e'\} \subset X\},$$

$X^{e'}$ is not determined, $N_6 = N_5 \setminus \{e'\} = \{f,\ f',\ g',\ g''\}$.

$$X^f = \bigcap_j X_j^f, \text{ where } X_j^f \in X_f = \{X \in \Omega^- \mid \{v,\ f\} \subset X\},$$

X^f is not determined, $N_7 = N_6 \setminus \{f\} = \{f',\ g',\ g''\}$;

$$X^{f'} = \bigcap_j X_j^{f'}, \text{ where } X_j^{f'} \in X_{f'} = \{X \in \Omega^- \mid \{v,\ f'\} \subset X\},$$

$X^{f'}$ is not determined, $N_8 = N_7 \setminus \{f'\} = \{g',\ g''\}$.

The minimal intersection, which corresponds to {v, g}, is not generated, since $g \notin N_8$,

$$X^{g'} = \bigcap_j X_j^{g'}, \text{ where } X_j^{g'} \in X_{g'} = \{X \in \Omega^- \mid \{v,\ g'\} \subset X\},$$

$X^{g'}$ is not determined, $N_9 = N_8 \setminus \{g'\} = \{g''\}$;

$$X^{g''} = \bigcap_j X_j^{g''}, \text{ where } X_j^{g''} \in X_{g''} = \{X \in \Omega^- \mid \{v,\ g''\} \subset X_j\},$$

$X^{g''}$ is not determined, $N_{10} = N_9 \setminus \{g''\} = \varnothing$.

Actions relating to step 2.3 remain to be carried out. From the set of (+) examples containing the intersection v, i.e., from $\{x_1, x_2, x_3, x_4, x_5\}$, we will remove all examples containing the minimal intersections (−): examples containing v, i.e., va, vbd, and vd′gg′. Such (+) examples will be x_1 = v*aa*′ and x_2 = vbdd′. We will intersect the remaining (+) examples, i.e., x_3, x_4, and x_5, and get vm ≠ v. Hence, there will be no generalized (+) hypothesis corresponding to the similarity of (+) examples v.

We will carry out the computations of steps 2.1-2.3 for the (+) intersection vd′. The (−) examples containing it will be x_{10} = vd′gg′ and x_{11} = vd′gg″. Hence, the sole minimal intersection of (−) examples containing vd′ will be vd′g′. Since none of the (+) examples contain it, in step 2.3 none of the (+) examples will be thrown out, and the triad {vd′, k = vd′g, w} will be accepted as a generalized (+) hypothesis.

As we already said above, algorithm 2 is intended for data represented by sets. This circumstance makes it possible to accomplish step 2.2 very efficiently. When the data are represented by objects from an arbitrary semilattice with similarity operation Π, the situation is worse: we cannot form an intersection of (−) examples that is minimal with respect to embedding and contains the given intersection of (+) examples v by "building up" v with atoms of the lattice of (−) intersections (for information about the lattice of intersections, see, for example, [5]) - we simply cannot know these atoms. Thus, for an arbitrary similarity operation, instead of step 2.2 of algorithm 2 we must seek the minimal intersections "from the top down:" by taking the intersections of a growing number of (−) examples containing v. As soon as the intersection of (−) examples becomes equal to v in the next step, we fall back on the preceding intersection of (−) examples: it will be minimal among those containing v. In this case, the upper estimate of the number of computer operations necessary to generate all of the minimal intersections containing v will be $O(|\overline{G}|\,\mathrm{pol}_2(|U|, |\Omega^+|, |\Omega^-|))$; consequently, the algorithm's overall complexity will be $O(|G| \cdot |\overline{G}|\,\mathrm{pol}_2(|U\Omega|, |\Omega^+|, |\Omega^-|))$, where $\mathrm{pol}_2(\cdot)$ is some small-degree polynomial depending on $|U|$, $|\Omega^+|$, and $|\Omega^-|$ and determined by a computer model.

REFERENCES

1. V. K. Finn, "Plausible deductions and plausible reasoning," in: Results of Science and Technology. Probability Theory, Mathematical Statistics, and Theoretical Cybernetics Series [in Russian], vol. 28, pp. 3-34, VINITI, Moscow, 1988.

2. S. O. Kuznetsov, "The JSM method as an automatic teaching system," in: Results of Science and Technology, Information Science Series [in Russian], vol. 15, pp. 17-54, VINITI, Moscow, 1991.

3. V. K. Finn, "The generalized JSM method of automatic hypothesis generation," Semiotika i Informatika, no. 29, pp. 93-123, 1989.

4. S. M. Gusakova and V. K. Finn, "Similarity and plausible deduction," Izv. AN SSSR, Ser. Tekhn. Kibernetika, vol. 5, pp. 42-63, 1987.

5. S. O. Kuznetsov, "Quick algorithm for constructing all intersections of objects from a finite semilattice," Nauchno-Tekhnicheskaya Informatsiya, Ser. 2, no. 1, pp. 17-20, 1993.

6. S. M. Gusakova, Canonical representations of similarities," Nauchno-Tekhnicheskaya Informatsiya, Ser. 2, no. 9, pp. 19-21, 1987.

7. S. M. Gusakova and V. K. Finn, "New means of formalizing local and global similarities," Nauchno-Tekhnicheskaya Informatsiya, Ser. 2, no. 10, pp. 14-22, 1987.

8. S. M. Gusakova, "Formalization of the concept of similarity and its application in intelligent information systems," Author's abstract of candidate's dissertation, VINITI, Moscow, 1988.

9. V. K. Finn and M. A. Mikheenkova, "Some problems of the generalized JSM method of automatic hypothesis generation," Semiotika i Informatika, no. 33, pp. 136-163, 1983.

5 May 1995

THE ALLERTON PRESS JOURNAL PROGRAI

AUTOMATIC DOCUMENTATION & MATHEMATICAL LINGUISTICS

Selected major articles from

NAUCHNO-TEKHNICHESKAYA INFORMATSIYA

Seriya 2. Informatsionnye Protsessy i Sistemy

Published six times a year

ALLERTON PRESS, INC.
150 Fifth Avenue New York, N.Y. 10011

ISSN 0005-1055

AUTOMATIC DOCUMENTATION AND MATHEMATICAL LINGUISTICS

(Nauchno-Tekhnicheskaya Informatsiya, Seriya 2)

Vol. 29, No. 3

ALLERTON PRESS, INC.