ISSN 0005-1055, Automatic Documentation and Mathematical Linguistics, 2024, Vol. 58, No. 3, pp. 208–211. © Allerton Press, Inc., 2024.
Russian Text © The Author(s), 2024, published in Nauchno-Tekhnicheskaya Informatsiya, Seriya 2: Informatsionnye Protsessy i Sistemy, 2024, No. 6, pp. 21–24.

Analysis of Detection of Empirical Regularity in Problems with a Similarity Operation Corresponding to Global Similarity

S. M. Gusakova*

Federal Research Center “Computer Science and Control,” Russian Academy of Sicences, Moscow, Russia

**e-mail: svem45@yandex.ru*

Received April 17, 2024

Abstract—This article discusses problems using the similarity operation corresponding to global similarity. Differences are noted in the carrying out of JSM-reasoning and JSM-research when solving problems using the similarity operation, corresponding to local and global similarity.

Keywords: similarity operation, global similarity, base of facts, empirical pattern, JSM-research

DOI: 10.3103/S0005105524700134

INTRODUCTION

The JSM method, which formalizes the methods of John Stuart Mill, contains two parts: JSM-reasoning and JSM-research. JSM-reasoning exhibits a synthesis of the following cognitive procedures: induction, with the help of which hypotheses are generated concerning the reasons for the manifestation of an effect or property in objects; analogy, which allows the properties an object to be determined if this is not already known; and abduction, which makes it possible to understand whether the hypotheses are based sufficient ground. In JSM-reasoning, hypotheses express empirical dependencies, such as causal, as with subobject V, the reason for the presence of the property W, written in the form $J_{(\sigma,\ n)}(V \Rightarrow {}_2W)$, where V is the body of the hypothesis, and additional definitions obtain, such as object X has a set of properties Y, $J_{(\sigma,n)}(X \Rightarrow {}_1Y)$, truth value $\sigma = \{+1, -1, 0, \tau\}$, and n represents number of the JSM-reasoning step. If the resulting hypotheses explain almost all examples of the factual basis (FB; to clarify the word *almost*, a threshold is introduced), then they are accepted to have sufficient grounds [1].

JSM-research is conducted to determine out whether the empirical dependencies obtained in the process of JSM-reasoning form empirical regularities. For this, a series of extensions to the FB are performed. Dependencies that retain their truth value under these extensions are declared preregularities. Then, the sequence of expansions of the FB changes, and if, with all possible methods of expansion, the resulting preregularity is preserved, it is declared a regularity [2].

It should be noted that, in formulating the similarity method, Mill argued that the general circumstance with respect to cases of a phenomenon under study agree is the cause (or effect) of this phenomenon [3]. Therefore, the formalization of the methods of Mill assumes the introduction of the operation of similarity (rather than relation) on a set of objects. The result of this operation forms a tentative hypothesis concerning the cause of the phenomenon or effect that is being studied. To be able to define a mathematically correct similarity operation on a set of objects, it is necessary to be able to structure these objects accordingly.

In accordance with the formulation of Mill’s similarity method, the operation that reveals a common substructure in objects is usually taken to be a similarity operation, i.e., analogous to the intersection operation. Accordingly, the theoretical justification of JSM-reasoning and JSM-research was performed for cases of the use of the similarity operation as an operation of an intersection of sets.

SIMILARITY OPERATION CORRESPONDING TO GLOBAL SIMILARITY

There are problems in which the manifestation of a property in objects is explained not by the general characteristics of the objects that have this property but by the entire set of characteristics of these objects. In these problems, the result of the similarity operation must be defined as a complete set of attributes of all objects that have a given property. This operation is analogous to the union operation. For this operation, all of the properties of the similarity operation are satisfied: idempotency, commutativity, associativity, and the presence of zero.

These two operations correspond to two types of similarities: local and global. To determine these two types of similarity, let us turn to the description of the JSM system FB.

The FB contains objects, properties (effects), and the relation that an object has many properties (exhibits many effects), which is written in the form $J_{(\sigma,\ n)}(X \Rightarrow {}_1Y)$. The FB includes examples that are positive ($\sigma = +1$), negative ($\sigma = -1$), contradictory ($\sigma = 0$), and uncertain ($\sigma = \tau$).

Thus, there are many object names in the FB N, many signs S, many properties P, and mapping variables ϕ from N to S and ψ from N to P.

The mapping variable ϕ generates local similarity λ: $n_i \lambda n_j \rightleftharpoons \phi\ (n_i) \cap \phi\ (n_j) \neq \varnothing$. This similarity is reflexive and symmetrical, therefore, it is a relation of tolerance. It can be binary or n-ary. The similarity operation Π corresponding to this similarity is an analogue of the intersection of sets.

The mapping variable ψ generates the global similarity γ: $\gamma\ \{n_1 \ldots\ n_k\} \rightleftharpoons \exists\ p\ (\psi^{-1}(p) = \{n_1 \ldots n_k\}),\ p \in P$. The global similarity is not a relation, as it has variable arity. The global similarity corresponds to the similarity operation $\coprod_{i=1}^{k} \{n1 \ldots\ nk\} = \phi\ (n_1) \cup \ldots \cup \phi\ (n_k)$, analogous to the union operation. For this operation, just like for the operation Π, the properties of the similarity operation are satisfied: idempotency, commutativity, associativity, and the presence of zero. Only the role of zero in this case is played by the universal set.

It should be noted that where objects are structured as sets, the operations that correspond to local and global similarity coincide with the operations of intersection and union of sets. In other cases this may be an analogue of these operations. For example, for objects represented by graphs, the common part of two objects corresponding to local similarity is defined as an induced subgraph on closed vertices where these objects have isomorphic clique graphs [4].

The class of problems in which the operation corresponds to global similarity is used includes attribution and identification problems. These tasks occur in the humanitarian sphere. The definition of these tasks can be formulated by associating an object with its corresponding attributes, such as the author, time or place of creation, membership in a group of documents, and others. Using this operation, the problems of dating birch bark letters [5], identifying the author of short notes [6], predicting the outcomes of neurosurgical operations [7], and categorizing text documents in natural language [8] could be solved. Thus, the Mill's basic principle "similarity of structures of objects entails the similarity of their properties," which forms the basis of the similarity predicate of the JSM method, takes on a new meaning in these tasks. However, this does not indicate the impossibility of using the JSM method in these problems and does not contradict the concept of the JSM method.

A variant of the JSM method, using the similarity operation corresponding to global similarity, is called a modified version of the JSM method. This differs somewhat from the classic version using an analogue of the intersection operation at the levels of both JSM-reasoning and JSM-research.

When using JSM-reasoning using the similarity operation corresponding to global similarity, only one cause is possible for each property. Therefore, in this variant, the JSM method with a unique cause is always used.

In problems that use the similarity operation corresponding to local similarity, when reasoning by analogy, the body of the hypothesis obtained by induction is embedded in the tested object. In problems that adopt the similarity operation corresponding to global similarity, the tested object is embedded in the body of the hypothesis.

The positive example of an object Ob with properties $(w_1,\ w_2)$ in attribution tasks is understood as that object Ob has properties w_1 or w_2. For this case, it is necessary to introduce a new truth value $\sigma = ½$. The presence of two attributes in one object only means that it is not possible to make a more accurate attribution; however, in reality the object has one attribute, while in problems with the first type of similarity operation, such an example is understood as object Ob has properties w_1 and w_2.

The body of causal hypotheses is not a subobject, but a superobject.

All causal hypotheses are obtained on sufficient grounds, because each object from examples of the form $J_{(+1,\ n)}(X \Rightarrow {}_1p)$ is embedded in the body of the hypothesis $J_{(+1,n)}(V \Rightarrow {}_1p)$.

JSM-RESEARCH FOR PROBLEMS WITH SIMILARITY OPERATION CORRESPONDING TO GLOBAL SIMILARITY

Suppose that, as a result of JSM-reasoning in a problem using an operation corresponding to global similarity, empirical dependencies of the form are obtained such that superobject V is the cause of the property p_i (in fact, in attribution tasks it would be more correct to say that superobject V characterizes the property p_i) and that to object with test X property is attributed p_i. To make sure that these empirical dependencies are empirical regularities, it is necessary to conduct a JSM study, i.e., trace the truth values of these dependencies in a sequence of expanding FBs.

However, when conducting JSM-research, it is necessary to consider that a domain model with a similarity operation corresponding to global similarity must be finely tuned to a specific task, as tasks using such an operation have different nuances.

In the modified version of the JSM method for each property p, there may be two kinds of negative examples. The first grade refers to positive examples of the species $J_{(+1,\,0)}\ (Q_j \Rightarrow {}_1p')$, written as $J_{(-1,\,0)}\ (Q_j \Rightarrow {}_1p)$. The second type refers to examples of the same type but with such objects Q_j about which it is only known that they do not have property p; however, what properties they do have is unknown, and these may be different properties. If in the process of JSM-reasoning in the presence of a positive hypothesis $J_{(+1,\,0)}$ $(V \Rightarrow {}_1p)$ from the negative examples of the first grade, a negative hypothesis is obtained $J_{(-1,\,0)}$ $(V \Rightarrow {}_1p)$; this means that the set of features that characterize the property p is not sufficient to differentiate it from the property p'. If the features of a subset of features from the set characterizing the property p occur in sets for other properties so that together they form this entire set, then the body of a negative hypothesis obtained from second-class negative examples may coincide with the body of a positive hypothesis $J_{(+1,\,0)}$ $(V \Rightarrow {}_1p)$. This indicates that in the set of features from the superobject V, there are no specific signs of a property p signs. If, in the process of expanding the FB, a positive hypothesis changes the truth value from +1 to 0, this means not only that the empirical dependence that is generated by this hypothesis is not a pattern but also that the data presentation language needs to be refined; it must be understood from the examples of what type of negative hypothesis is obtained.

Where the tested object cannot always be embedded in the body of the hypothesis, heuristics are used that replace the embedding with the maximum of intersection with the addition of the Tanimoto coefficient, as, for example, in the problem of determining the author of short notes [6], assigning this object to different properties does not entail a contradiction, and the truth value is $\sigma = ½$. This indicates an incomplete description of the tested object.

These nuances must be taken into account in JSM-research.

If in an FB$_0$ for all properties from the set P, the bodies of positive hypotheses do not intersect, all negative hypotheses are only of the first grade, and with the sequential addition of new examples, this property is preserved, then for any property p global similarity $\{n_1...n_k\} = \psi^{-1}(p)$ in the FB$_j$ database becomes $\psi^{-1}(p) = \{n_1, ..., n_k, n_{k+1}, ... n_m\}$ in FB$_{j+1}$ database. Although the content of the hypothesis changes, its truth value does not change. This situation can be described as follows: $\psi^{-1}(p)$ is the cause of the manifestation of the property p. In fact, in the process of extensions, this dependence is refined. It is clear that the tested objects, which are a property in FB$_0$, to which p was attributed, the same property will be assigned in extended bases of facts. Changing the truth value for the empirical dependence the property p_i is assigned to the tested object X. When expanding the FB, if FB$_j$ appears in the process of JSM-reasoning, an object X is not further defined, as it is not embedded in any hypothesis, but when the hypotheses were expanded, the nesting occurred. Then, the truth value of this dependence changes from τ to +1. In the considered version of the extensions, each causal empirical dependence forms a regularity. Regarding for the dependencies associated with extension, even if the truth value changes from τ to +1 during the final expansion, we can say that this is an empirical law, as this truth value will not change again. If, as the FB is gradually expanded, adding objects does not expand the superobject V from hypothesis $J_{(+1,\,0)}$ $(V \Rightarrow {}_1p)$, then this indicates the completeness of the set of features characterizing the property p.

If the condition for the intersection of sets of attributes of different hypotheses to be non-empty does not hold, then it may happen that when moving from the FB$_J$ to FB$_{J+1}$ examples are added $J_{(+1,\,0)}$ $(X \Rightarrow {}_1p_1)$ and $J_{(+1,\,0)}$ $(X \Rightarrow {}_1p_2)$ such that some of the features of the superobject $V_1 = \psi^{-1}(p_1)$ received in FB$_J$, is also included in the superobject $V_2 = \psi^{-1}(p_2)$. If the object X' from hypothesis $J_{(+1,\,0)}$ $(X' \Rightarrow {}_1p_1)$, received in FB$_J$, coincides with a subset of features common to V_1 and V_2 or is contained in this subset, then when working with FB$_{J+1}$ X' will invest not only in V_1, but also in V_2. In this case, for FB$_{J+1}$, instead of addiction $J_{(+1,\,0)}$ $(X' \Rightarrow {}_1p_1)$, the dependency $J_{(+1/2,\,0)}$ $(X' \Rightarrow {}_1p_1, p_2)$ will appear.

If, in this situation, a negative hypothesis appears at the next expansion $J_{(-1,\,0)}$ $(V \Rightarrow {}_2p_1)$ and $V = V_1$, then the hypothesis $J_{(+1,\,0)}$ $(V_1 \Rightarrow {}_2p_1)$ becomes $J_{(0,\,0)}$ $(V_1 \Rightarrow {}_2p_1)$. In essence, this means that the set of features characterizing the property p_1, is either incomplete or features that are not adequate for this property were chosen for the description. Because the object X' is invested in V_1 and in V_2, and for V_2 $J_{(+1.0)}$ $(V_2 \Rightarrow {}_2p_2)$ occurs, then it is logical to assume that the dependence $J_{(+1/2,\,0)}$ $(X' \Rightarrow {}_1p_1, p_2)$ will change to $J_{(+1,\,0)}$ $(X' \Rightarrow {}_1p_2)$. This does not coincide with the analogical inference rule for the variant of the JSM method with the similarity operation corresponding to local similarity.

CONCLUSIONS

The JSM method of automated research support is not a method in the literal sense of the word but a whole set of methods and heuristics that are united by a common ideology. Moreover, this set is open, which allows new methods and heuristics to be added, thereby creating new variants of the JSM method. The refinement of methods and heuristics is determined by the adjustment to the subject area and the problem being solved in the particular area. Thus, the development of JSM systems requires the participation of a subject matter specialist.

The JSM method, originating as a formalization of the inductive methods of John Stuart Mill, has outgrown this framework and treats similarity as a mathematical operation that essentially reflects the content of the problem being solved. This makes it possible to solve problems that use a similarity operation corresponding to both local and global similarity. At the same time, both in JSM-reasoning and in JSM-research, similar situations in these different versions of the JSM method can be interpreted differently. The main thing is that they meet the axioms of the subject area and the features of the problem being solved.

FUNDING

This work was supported by ongoing institutional funding. No additional grants to carry out or direct this particular research were obtained.

CONFLICT OF INTEREST

The author of this work declares that he has no conflicts of interest.

REFERENCES

1. Finn, V.K., JS Mill's inductive methods in artificial intelligence systems. Part I, *Sci. Tech. Inf. Process.*, 2011, vol. 38, no. 6, pp. 385–402. https://doi.org/10.3103/S0147688211060037
2. Finn, V.K. and Shesternikova, O.P., The heuristics of detection of empirical regularities by JSM-reasoning, *Autom. Doc. Math. Linguist.*, 2018, vol. 52, no. 5, pp. 215–247. https://doi.org/10.3103/s0005105518050023
3. Mill, J.S., *A System of Logic: Ratiocinative and Inductive*, Honolulu, Hawaii: Univ. Press of the Pacific Publ., 2002.
4. Gusakova, S.M., Structural similarity of objects represented by ordinary graphs, *Autom. Doc. Math. Linguist.*, 2023, vol. 57, no. 4, pp. 206–210. https://doi.org/10.3103/s0005105523040039
5. Gusakova, S.M., Logical-combinatorial methods for analyzing historical data, *Metody istoricheskogo poznaniya: Sbornik statei po materialam kruglogo stola* (Methods of Historical Cognition: Collection of Articles from the Round Table Proceedings), Moscow: Institut Vseobshchei Istorii, 2008, pp. 122–135.
6. Komarov, A.S., Intelligent data analysis in handwriting research: Software implementation, *Vestn. Ross. Gos. Gumanit. Univ., Ser.: Dokumentovedenie Arkhivovedenie. Inf. Zashch. Inf. Informatsionnaya Bezop.*, 2010, no. 12, pp. 290–298. https://elibrary.ru/nbqezf.
7. Zabezhailo, M.I. and Gavryushin, A.V., On some possibilities of application of AI methods in predicting outcomes of neurosurgery operations, *Iskusstvennyi Intellekt Prinyatie Reshenii*, 2024 (in print).
8. Lyfenko, N.D., An approach to text data categorization based on the ideas of J.S. Mill, *Autom. Doc. Math. Linguist.*, 2015, vol. 49, no. 6, pp. 202–212. https://doi.org/10.3103/s0005105515060035

Publisher's Note. Allerton Press remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.