INTELLIGENT ANALYSIS OF INCOMPLETE DATA FOR BUILDINGFORMALONTOLOGIES

V.A. Semenova, S.V. Smirnov

Institute for the Control of Complex Systems, Russian Academy of Sciences

Abstract. The article presents models and methods of ontological data analysis aimed at identifying conceptual structure, or formal ontology, of the studied knowledge domain (KD). The realities of the accumulation of empirical data are reflected in model of generalized “object-properties” table and the incompleteness of this information implies the need for models of multi-valued logic. For the first time detailed models and method of accounting for “existence constraints”, which can be known priori by the researcher of KD in the context of measured attributes of the objects’ learning sample.

Keywords: formal ontology, ontological data analysis, formal concept analysis, multi-valued logic, properties existence constraints

Citation: Semenova VA, Smirnov SV. Intelligent analysis of incomplete data to building formal ontologies. CEUR Workshop Proceedings, 2016; 1638: 796-805. DOI: 10.18287/1613-0073-2016-1638-796-805

Introduction

Complex informational systems (in the broadest coverage, including the internet, Big Data etc.) are effective only at a safe and consistent representation of their subject matter. Systematization, development and use of such information models make up contemporary content of the ontological approach in computational science.

In contrast to the philosophy ontology in computer science describes some limited sphere of knowledge, knowledge domain (KD). Therefore, by virtue of the multiplicity of Sciences and KD, where each of them has its own or even several competing terminologies, here, in contrast to philosophy the use of the plural term for ontology makes sense. Moreover, the distinction between linguistic and formal ontology is made. Formal ontologies inherit the paradigm of models and methods of knowledge representation developed in artificial intelligence.

Formal ontology describes a KD by means of standardizing terminology (vocabulary) and heterogeneous relations between concepts determined in it. Classes, relations, functions and axioms are modeling primitives of ontological specification, in a manner, that brings together these knowledge representation structures and appropriate computer resources with A.I. Maltsev algebraic systems [1].

There are three main ways of developing ontologies [2]:

- the most common way is related to the direct formalization of experience and knowledge of experts, who formalize their views on KD by means of human-oriented language or fix them with the help of knowledge engineer;
- actual ontology can be synthesized as a result of human-machine procedures of composition / decomposition of approved formal ontologies (patterns) of different levels and direction [3, 4];
- the third way is connected with the automatic "output" of the formal ontology from available data. This data is considered as the result of measurements of objects of investigated KD and filled in standardized "object-properties" table, the analysis of which leads to the identification of conceptual KD structure. The most effective methods in this direction are based on the branch of theory of lattices - formal concept analysis (FCA) [4-8].

Considering the third way of formal ontology developing so-called intelligent ontological data analysis (ODA), - it is important to pay attention to the different role of the researcher. In fact, his main task is offering hypotheses concerning object properties and then completing priori of arsenal of measuring procedures (sensory organs, verbal possibilities experts, artificial sensors, instruments, systems, etc.), by means of that the interests of its KD will be probed.
In the present report the attention is concentrated on two aspects of the ODA. Firstly, it is a reflection of the realities of KD data collection, leading to a need to use models of multi-valued logic to represent the empirical data, and secondly (and in connection with the first), it is modeling and the method of accounting a priori known by researcher dependencies between measurable properties in the work.

Empirical data formation

ODA is based on the assumption that any measurement of the object properties can give special result "**None**", which may demonstrates a finding of a measured property value outside of sensitivity threshold and the dynamic range of a measuring instrument; it shows a "semantic mismatch" of the object and the measuring procedure etc. In a fundamental FCA such effects are achieved by performing a cognitive procedure called conceptual scaling [4, 9]. The researcher can a priori subjectively to "cover" the actual range of properties measurement procedure, forming a set of new object properties, actually measured afterwards in the binary items scale {**X**, **None**}, where **X** replaces any symbol of scales of dynamic ranges of measurement procedures.
Anyway, "**None**-concept" allows us to change the analysis paradigm of experimental data and naturally convert "object-properties" table into the set of truth values of basic semantic proposition (BSP) b_{xy} = "*x* object has *y* property": $\|b_{ij}\|$ = **True**, if the *y* property measurement result of *x* object is **X**, **False** – otherwise.
FCA is focused on processing of such data. The tuple $\boldsymbol{K} = (G^*, M, I)$ – Formal Context– puts together initial empirical data for FCA. $G^* = \{g_i\}_{i = 1,\ldots, r}$, $r = |G^*| \geq 1$ – set of objects of investigated KD (i.e. a set of objects' learning sample: $G^* \subseteq G$, where G - the set of all conceivable objects of KD), $M = \{m_j\}_{j = 1,\ldots, s}$, $s = |M| \geq 1$ - set of meas-

ured objects properties, I – binary incidence "object-properties" i.e. set of estimates $\|b_{ij}\| \in \{\textbf{True}, \textbf{False}\}$.

For evaluation of the part of BSP truth is vague because of the quality of the original empirical material (e.g., formed by an expert based on experience or intuition), and for the evaluation of such statements use natural truth values, input multi-valued logic. But then arises the question about the model explaining the origin and the method of calculating of these estimates.

In general it is clear that multivaluedness of BSP truth values is consequence of incompleteness of the data concerning KD (inaccuracy, inconsistency, etc.). However, this incompleteness is not reflected in the standard structure of "object-properties" table. The reasons for incompleteness are caused by realities of the empirical data collection. These realities include performing of multiple independent measurements of the property $m_j \in M$ of the object $g_i \in G^*$, using of several different procedures for the measurement of the same property m_j (congruent sources), differentiation of trust to different measurement procedures. Therefore, as an adequate model of the original data we offer generalized "object-properties" table, described by the tuple

$$(G^*, M, Se, Pr, A), \quad (1)$$

where:

- $Se = \bigcup_{i=1}^{r} Se_{(i)}$ - the set of all performed series of measurements in probing of KD, $|Se| = \sum_{i=1}^{r} |Se_{(i)}| = m$, и$Se_{(i)} = \{se_{(i)k}\}_{k=1,\dots,q_{(i)}}$, $q_{(i)} \geq 1$, $i = 1,\dots, r$ - set of series of measurements, applied to object $g_i \in G^*$;
- $Pr = \bigcup_{j=1}^{s} Pr_{(j)}$ - arsenal of all used measurement procedures for KD probing, $|Pr| = \sum_{j=1}^{sr} |Pr_{(j)}| = n$, and$Pr_{(j)} = \{pr_{(ji)k}\}_{k=1,\dots,p_{(j)}}$, $p_{(j)} \geq 1$, $j = 1,\dots, s$ - set of congruent procedures for measuring property $m_j \in M$, and every procedure $pr_{(j)k}$ is characterized by a degree of confidence in its results $t_{(j)k}$;
- $A = (a_{ij})_{i=1,\dots, m;\ j=1,\dots, n}$ – matrix of measurements series results *Se*of properties M of objects from sample G^*, made by means of measurement procedures *Pr*. The elements of this matrix can be linguistic constants **NM**, **None**, **Failure** and **X**. **Failure** – a result that records measurement failure (denial, measurement means malfunction, abstention, etc), **NM** (*not measured*) – a result indicating that as a matter of fact in this series of measurements particular procedure was not used (introduction of this formal result is necessary to save the two-dimensional nature of the generalized "object-properties" table).

Model (1) makes it possible to calculate the "soft" truth values of the BSP about KD. For example, this can be done on the basis of fuzzy logic. In [10] a non-strict formal context $\boldsymbol{K}$ with a non-strict incidence I is constructed on the basis of (1) and by the means of a more adequate multi-valued vector logic V^{TF} [11]. Non-strict incidence I is formed with BSP $\|b_{ij}\| \in \langle b^+_{ij}, b^-_{ij}\rangle$, $b^+_{ij}, b^-_{ij} \in [0, 1]$, wherein the component (true aspect) b^+_{ij} - *Truth* – is formed by evidences confirming the BSP, and the component (false aspect) b^-_{ij} - *False* - is formed by denying BSP.

Unfortunately, today there are no effective methods of output of conceptual structure of KD directly from the "soft" formal contexts. For example, theoretically and com-

putationally complex method using a closure operator of fuzzy set [12] represents only academic interest because generates a huge amount of fuzzy concepts, even for small-sized "sparse" fuzzy contexts. Therefore effective methods are based on preliminary α-approximation of "soft" correspondences "objects-properties" when a researcher sets a threshold of confidence in the initial data [13, 14]. Then proven concept output methods of FCA are applied to produce binary approximations.
Nevertheless, this approach generally proves to be incorrect, because receiving α-section does not account dependencies between measured properties that are known priori by the researcher.

Dependencies between measured properties

Standard procedure α-section is "blind" to any links between sets of elements that are involved in non-strict binary relations. Since the FCA follows the paradigm that all relations between KD objects are manifestations of intrinsic properties of an object, then this "blindness" of approximation is indifferent regarding objects' learning sample. The situation is different with the set of measured properties. The binary approximation is able to "reveal" the contradictions between BSP, which are shown for the researcher as a priori inadmissible combination of objects' properties. In the general case these differences cannot be detected in a model with soft truth values of BSP.
Conceptually, the source of the problem is FCA's intrinsic cognitive asymmetry of "objects" and "properties". Formally the objects G^* are independent from the KD researcher, while the properties are the result of KD hypotheses production made by the subject of research which is based on his worldview or on the combination of a priori knowledge available to him. The subject is the "owner" of the properties measurement procedures arsenal and possesses sufficient information about them. In particular he may know about certain dependencies between results of executing different measurement procedures applied to the same object, i.e. dependencies between objects properties.
The common models of typical types of dependencies between properties are proposed in [15] in the form of binary relations "existence constraints". In particular, a pair of properties m_j, $m_k \in M$, $j \neq k$ for any KD object (and hence $\forall g_i \in G^*$) can be:

- conditional if possessing m_j property, g_i object immutably has the m_k property (though the reverse may not be true), i.e. $C(m_j, m_k) \leftrightarrow \forall g_i \in G^*$: $m_j \in \{g_i\}' \rightarrow m_k \in \{g_i\}'$ where, according to FCA notation $\{g_i\}'$ - the set of g_i object properties;
- incompatible if possessing the m_j property, g_i object certainly has not the m_k property, and vice versa, i.e. $E(m_j, m_k) \leftrightarrow \forall g_i \in G^*$: $m_j \in \{g_i\}' \rightarrow m_k \notin \{g_i\}'$.

We note that for such dependencies the object of learning sample can have only a normal subset of the set's measured properties [15, 16]. A subset of the measured properties $Z \subseteq M$ is normal iff it is closed and compatible: Z is closed if it contains all of the properties conditioned by any element of Z, i.e. $\forall m_j \in Z$ ($\exists m_k \in M$:

$C(m_j, m_k) \rightarrow m_k \in Z$); Z is compatible, if any two elements of Z are not connected by the relation of incompatibility, i.e. $\forall m_j \in Z\ (\exists m_k \in M: E(m_j, m_k) \rightarrow m_k \notin Z)$.
Specific tasks motivate to consider characteristic groups of dependent properties.
For example, in [15, 16] during ontology building the pair of interdependent properties are focused on the base of knowledge about KD object properties and dependencies between them.
In [17] the occurrence of properties existence constraints as a result of FCA conceptual scaling is analyzed. The most common technique of scaling is based on the using of nominal scales. This method generates groups of properties with paired incompatibility in a set of measurable properties. We call such important for us groups conceptually conjugated. In fact, each of these group generally represents a proto-property, whose values domain is disjunctive splitted as a result of scaling.
Each group of conceptually conjugate properties (GCCP) sets one of two possibilities: either all measured object properties belonging to the group should not exist, or the learning sample's object must have some one and only one measurable property from given group.
It is obvious that any measurable property not included in any of the GCCP can be considered as a single GCCP ("non-splitted" proto-property). Therefore it is natural to proceed from the fact that existence constraints are set on the set of proto-properties. Thereby a two-tier model of property existence constraints is formed and it is appropriate to use this model in further analysis.
The proposed model of properties existence constraints will be determined by the tuple (M_P, E_P, C_P), where:

- M_P - set of actual for researcher proto-properties of KD objects, $1 \le |M_P| \le |M|$, $M_P = M_{P1} \cup M_{P2}$, $M_{P1} \cap M_{P2} = \varnothing$; M_{P1} - a subset of single proto-properties (i.e. "non-splitted" proto-properties); M_{P2} - a subset of proto-properties subjected to conceptual scaling or the set of all groups of conceptual conjugate properties (GCCP): $M_{P2} = \{Gr_1, \ldots, Gr_{|M_{P2}|}\}$, - and its constituent measured properties are incompatible in every GCCP, i.e. $(\forall m_j, m_k \in Gr_i, j \ne k) \rightarrow E(m_j, m_k) =$ **True**;
- E_P - pair of incompatible proto-properties, $E_P \subseteq M_P \times M_P$, $|E_P| \le C^2_{|M_P|}$(number of combinations);
- C_P – pair of conditional proto-properties, $C_P \subseteq M_P \times M_P$, $|C_P| \le A^2_{|M_P|}$ (number of arrangements).

Rational binary approximation of non-strict formal context

Formally the problem of correct α-section of non-strict formal context can be reduced to the construction of single predicate "α-section is correct" with a vector argument $\alpha = (\alpha^+, \alpha^-)$, $\alpha^+, \alpha^- \in [0, 1]$, where the condition of truth confirmation of every empirical b_{ij} BSP

$\|b_{ij}\| \ge \alpha^+ \wedge \|b_{ij}\| \le \alpha^-$,

must be combined with the implementation of all the properties existence constraints of KD objects. Next, find an existence domain (perhaps, it will be empty) of confidence thresholds α, delivering **Truth** value to such predicate.

Of course, in general case to build such a predicate and to identify the target area is very difficult. However, even allowing the possibility of such a solution, binding the researcher with need to select the threshold from only this region is highly not practical. For example, with this approach for the subject of research the intuitive response expectations to easing and tightening of the threshold of confidence in the source data will not be executed.

Instead, we propose the following heuristics: the subject is free to choose the threshold. And the corresponding to the threshold in the general case the unacceptable set of properties for each object of learning sample sequentially reduced by a cut at each step of the property that violates existence constraints. The main criterion for selecting property for the cut-off is minimum tightening of threshold established by the researcher.

The proposed heuristic method of obtaining the correct binary approximation of non-strict formal context is effective.

Indeed, the method is separately implemented for each object of learning sample. And at some stage there will appear one of the following conditions:

- the set of object properties will satisfy the existence constraints (note that an empty set of properties satisfies them, but then the object has to be qualified as unidentified);
- will be stated that in the initial "soft" BSP truth values there is ineradicable contradiction: about the object there are initially true in the classical (Aristotelian) logic BSP that violate priori properties existence constraints.

The implementation of the proposed method is appropriate as the two-stage procedure. First, all multiple GCCP should be processed by the proposed method. At the same time we put a principal condition for the preservation of not more than one property of each group for every object of learning sample. Then the results of the first stage is processed by the same method, together with the source data about "non-splitted" proto-properties.

The reason for this recommendation is "relative transitivity" of incompatibility relation of properties [15]:

$C(m_j, m_k) \,\&\, E(m_k, m_l) \rightarrow E(m_j, m_l), j \neq k, j \neq l, k \neq l.$

The implementation of the method in one step is not correct because of the unique situation when a proto-property is conditional and "splitted" at the same time. Each combination produces n original versions of binary conditionality relationship on the set of measurable properties, where n is the number of measurable properties into which the conditional proto-property is "splitted". Therefore one-step strategy for implementing the proposed method must by perforce include the constructing and processing of the direct product of all variants of the measured properties existence constraints. These results must be compared that is not rational.

Building a formal ontology of knowledge domain

For building a formal ontology of studied KD derived by means of basic FCA algorithms formal concepts lattice should be transformed into an object classes taxonomy. All non-taxonomic structural relationships between the concepts-classes will be realized by measuring the properties-valences of KD [18]. This transformation becomes a non-trivial task, given the prospect of generating databases on the basis of a formal ontology for storing denotative object models of KD.
Formal concepts according to the formation of their extensions are divided into three types:

- The concepts of the first type describe objects really exist in the analyzed domain. These concepts define a class of objects that deserve the naming of "*fundamental*".
- The concepts of the second kind - only generalize other notions. In software design these classes are known as "*virtual*".
- The third type of concepts is characterized by combining these features concepts first and second kinds.

The above-mentioned pragmatic considerations require to restrict at the building formal ontology with only fundamental and virtual object classes, and generally based on the following principles of transformation of the formal concepts lattice into the classes taxonomy [19]:

- all the concepts of the lattice are candidates for fundamental classes of the model;
- the fundamental class becomes the minimum (in the terminology of lattices) concept containing the object in its extension;
- attribute is preserved to the maximum of the concepts contained this attribute in its intension;
- the highest concept lattice (his sign - power extension equal to the of objects) is certainly excluded from the model, if its intention is empty;
- the smallest concept lattice (his sign - the power intention equal to the power set of attributes) are known to be excluded from the model if its extension is empty;
- analysis of candidates in the fundamental classes begins with the smallest concept, and conducted by levels nearest super-concepts.

Method which follows those principles includes the following steps:

1. The original version of the model is formed as *a copy* of the formal concept lattice.
2. In the model is searched *the greatest concept*.
   If *the intension of this concept is empty*, it is excluded from the model with break his ties with sub-concepts.
3. In the model is searched *the smallest concept*.
   If *extension* of the smallest concept is *empty*:
   - this concept is excluded from the model with the breaking its ties with super-concept;
   - a set of candidates in fundamental classes is formed of his closest super-concepts.

If extension of the smallest concept is not empty, then a set of candidates in fundamental classes is formed of *one* smallest concept.

4. Loop through a set of candidates.
   - For each super-concept of the candidate under consideration excludes objects from extension that are within the extension of this candidate (the extension super-concept is always not less than the extension sub-concept).
   - In consideration of the candidate from the intension excludes any attribute that is part of the intension of at least one super-concept (a combination of all super-concept's intension is always not more than concept intent, which they are).
   - If the candidate has no sub-concepts, it is recorded as the fundamental class. In such case one of two alternatives is implemented:
     - if the candidate has no sub-concepts, it is recorded straight as a fundamental;
     - otherwise for this candidate creates a new sub-concept, in which the extension is transferred (and only extension) of the candidate. This new sub-concept is fixed as the fundamental class of objects. The intension of such fundamental class is empty. The candidate is retained in the model as a virtual class with an empty extension.
   - Promising set of concepts-candidates is unalterably filling with super-concepts of a current candidate.
5. Promising set of candidates is being reduced: remains only root concepts of generalization relationship, which is determined in a promising set of concepts-candidates.
6. If a set of promising candidates is not empty, then algorithm repeats from Step 4.
7. Classes with an empty extent and intent are excluded from a formed set. These could be only intermediate (i.e. not root or node class) classes of developed taxonomy.

Conclusion

Presented in the article models and methods show a new stage of the development of methodology for the identification of conceptual structure and ultimately formal ontology of experimentally studied KD. The basis of the technique is still formal concept analysis, in which retained the classical understanding of the concept as a fundamental notional element defined by scope and content.

The focus was on the problem of reflection of the realities (incompleteness, inaccuracy, inconsistency of source data) of the accumulation of empirical information about the KD. To solve this problem we had to generalize a standard model of representation of object-attributive data and apply models and apparatus of multi-valued vector logic for its processing.

However increasing the adequacy of initial data models given rise to new problems of formal concept analysis. Needed to develop an intelligent method of converting

source data of the new format into binary formal context for which effective algorithms of derivation of formal concepts are known. At the same time constructed models for accounting priori known dependencies between the measured properties of KD objects.

Finally the methodology includes a series of pragmatic principle determined the method of transformation of the formal concept lattice outputted from empirical data into formal ontology of KD.

The fully presented methodical complex is expected to introduce in developed by the Institute for Control of Complex Systems RAS system of semantic modeling and design on mass software platform.

Finally, it should be noted that from the position of traditional data analysis techniques presented methodical complex should be attributed to varieties of data clustering methods. When clustering incomplete data from congruent sources (e.g. hyperspectral information in Earth remote sensing), this complex can compete with traditional methods such as *k*-means [20].

References

1. Mal'cev AI. Algebraic Systems. Berlin-Heidelberg, New York: Springer Verlag, 1973.
2. Smirnov SV. Ontological modeling in situational management. Ontology of Designing, 2012; 2(4): 16-24. [in Russian]
3. Suárez-Figueroa MC, Gómez-Pérez A, Fernández-López M. The NeOn Methodology for Ontology Engineering. In: Ontology Engineering in a Networked World. Springer Berlin Heidelberg, 2012: 9-34.
4. Lomov PA. Automation of synthesis of composite content ontology design pattern. Ontology of Designing, 2016; 6(2): 162-172 [in Russian]. DOI 10.18287/2223-9537-2016-6-2-162-172.
5. Ganter B, Wille R. Formal Concept Analysis. Mathematical foundations. Springer Berlin-Heidelberg, 1999.
6. Formal Concept Analysis Homepage. URL: http://www.upriss.org.uk/fca/fca.html.
7. Godin R, Mili H, Mineau GW, Missaoui R, Arfi A, Chau TT. Ontology Design with Formal Concept Analysis. Theory and Application of Object Systems (TAPOS), 1998; 4(2): 117-134.
8. Obitko M, Snasel V, Smid J. Ontology Design with Formal Concept Analysis. In: V. Snasel, R. Belohlavek (Eds.): Proc. of the CLA 2004 International Workshop on Concept Lattices and their Applications (Ostrava, Czech Republic, September 23-24, 2004). TU of Ostrava, Dept. of Computer Science, 2004: 111-119.
9. Ganter B, Wille R. Conceptual scaling. In: Applications of Combinatorics and Graph Theory to the Biological and Social Sciences. Springer, New York, 1989: 139-167.
10. Smirnov SV. Non-strictFormal Concept Analysis. Proc. 5th All-Russian Conf. "Knowledge – Ontologies – Theories" (Novosibirsk, Russia, October 6-8, 2015). Novosibirsk: Sobolev Institute of Mathematics, SB of RAS, 2015; 2: 142-150. [In Russian]
11. Arshinskii LV. Substantial and formal deductions in logics with vector semantics. Automation and Remote Control, 2007; 68(1): 139-148.
12. Belohlavek R, De Baets B, Outrata B, Vychodil J. Computing the lattice of all fixpoints of a fuzzy closure operator. IEEE Trans. on Fuzzy systems, 2010; 18(3): 546-557.

13. Tho QT, Hui SC, Fong ACM, Cao TH. Automatic Fuzzy Ontology Generation for the Semantic Web. IEEE Trans. on Knowledge and Data Engineering, 2006; 18(6): 842-856.
14. Yang KM, Kim EH, Hwang SH, Choi SH. Fuzzy Concept Mining based on Formal Concept Analysis. Int. J. of Computers, 2008; 2(3): 279-290.
15. Lammari N, Metais E. Building and maintaining ontologies: a set of algorithms. Data & Knowledge Engineering, 2004; 48(2): 155-176.
16. Pronina VA, Shipilina LB. Using the relationships between attributes to build domain ontology. Control Science, 2009; 1: 27-32. [In Russian]
17. Samoilov DE, Smirnov SV. Data formation and processing in Formal Concept Analysis: subjective aspects. CEUR Workshop Proceedings, 2016; 1638: 806-812. DOI: 10.18287/1613-0073-2016-1638-806-812.
18. Smirnov SV. Building knowledge domain ontologies with structural relationships based on Formal Concept Analysis. Proc. 3rd All-Russian Conf. "Knowledge –Ontologies – Theories" (Novosibirsk, Russia, October 3-5, 2011). Novosibirsk: Sobolev Institute of Mathematics, SB of RAS, 2011; 2: 103-112. [In Russian]
19. Kovartsev AN, Smirnov VS, Smirnov SV. Intelligent Design of Class Structure Model based on Ontological Data Analysis. Proceedings of the Institute for System Programming of the RAS, 2015; 27(3): 73-86.
20. Zimichev EA, Kazanskiy NL, Serafimovich PG. Spectral-spatial classification with k-means++ particional clustering. Computer Optics, 2014; 38(2): 281-286.