МОДЕЛИ И МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА НЕПОЛНЫХ ДАННЫХ ДЛЯ ПОСТРОЕНИЯ ФОРМАЛЬНЫХ ОНТОЛОГИЙ

В.А. Семёнова[1], С.В. Смирнов[2]

[1]Самарский государственный аэрокосмический университет им. академика С.П. Королёва (национальный исследовательский университет),
[2]Институт проблем управления сложными системами РАН

В статье представлены модели и методы онтологического анализа данных, направленного на выявление понятийной структуры, или формальной онтологии, исследуемой предметной области. Реалии накопления эмпирической информации отражены в модели обобщенной таблицы «объекты-свойства», а неполнота этой информации влечет необходимость использования моделей многозначной логики. Впервые подробно описаны модели и метод учета «ограничений существования», которые могут быть априори известны исследователю предметной области применительно к измеряемым свойствам объектов обучающей выборки.

Сложные информационные системы (в самом широком охвате, включая интернет, Big Data и.т.п.) результативны лишь при надежном и согласованном представлении их предмета. Систематизация, разработка и использование таких информационных моделей составляют современное содержание онтологического подхода в вычислительных науках.

В отличие от философии онтология в информатике описывает некоторую ограниченную сферу знания, предметную область (ПрО). Поэтому в силу множественности наук и ПрО, когда каждая из них имеет свою собственную или даже несколько конкурирующих терминологий, здесь в противоположность философии приобретает смысл употребление множественного числа для термина, т.е. говорят об «онтологиях». Более того, различают лингвистические и формальные онтологии, где последние наследуют парадигму моделей и методов представления знаний, разработанных в искусственном интеллекте [1].

Формальная онтология описывает ПрО с помощью стандартизирующей терминологии – словаря, а также неоднородных связей между определенными в нем понятиями. Моделирующими примитивами онтологической спецификации служат классы, отношения, функции и аксиомы, что в некотором смысле сближает эти структуры представления знаний и соответствующие компьютерные ресурсы с алгебраическими системами А.И. Мальцева [2].

Сегодня можно указать три основных пути разработки онтологий [3]:

- наиболее используемый связан с прямой формализацией опыта и знаний экспертов, которые с помощью компьютерных языковых средств либо автоформализуют свои представления о ПрО, либо фиксируют их с помощью инженера по знаниям;
- при наличии развитой инфраструктуры работы со знаниями актуальные онтологии могут синтезироваться в результате человеко-машинных процедур композиции/декомпозиции апробированных формальных онтологий разного уровня и направленности [4];
- третий путь связан с автоматическим «выводом» формальной онтологии из доступных данных. Эти данные рассматриваются как результат измерений объектов исследуемой ПрО и сводятся в стандартизованные таблицы «объекты-свойства» [5], анализ которых приводит к выявлению понятийной структуры ПрО. Наиболее результативные методы этого направления опираются на ветвь теории решёток – анализ формальных понятий (АФП) [6].

Рассматривая третий путь построения формальных онтологий - интеллектуальный онтологический анализ данных (ОАД), - важно обратить внимание на иную, чем в других случаях роль исследователя. Фактически его основной задачей становится выдвижение

гипотез о свойствах объектов исследуемой ПрО и затем априорное комплектование арсенала измерительных процедур (органов чувств, вербальных возможностей экспертов, искусственных сенсоров, приборов, систем и т.д.), с помощью которых интересующая его ПрО будет зондироваться.

В предлагаемом сообщении внимание сосредоточено на двух аспектах ОАД. Во-первых, это отражение реалий сбора данных о ПрО, приводящее к необходимости использования для представления эмпирической информации моделей многозначной логики, а, во-вторых (и в связи с первым), на моделировании и методе учета в работе априори известных исследователю зависимостей между измеряемыми свойствами.

В контраст с общепринятой точкой зрения ОАД исходит из положения, что всякое измерение свойства объекта может дать специальный результат «**None**», который может свидетельствовать о семантическом несоответствии исследуемого объекта и измерительной процедуры, о нахождении значения измеряемого свойства за порогами чувствительности, вне динамического диапазона средства измерений [7]. В фундаментальном АФП подобные эффекты достигаются в результате выполнения когнитивной процедуры, именуемой концептуальным шкалированием [6]. Здесь исследователь может априори субъективно «расщепить» действительный диапазон процедуры измерения свойства, образуя набор новых свойств объектов ПрО, фактически измеряемых после этого в бинарной шкале наименований {**X**, **None**}, где **X** - лингвистическая константа, собирательно обозначающая любой символ шкалы динамического диапазона измерительной процедуры.

Так или иначе, «None-концепция» позволяет изменить парадигму анализа экспериментального материала и для начала естественным образом преобразовать таблицу «объекты-свойства» в совокупность оценок истинности базовых семантических суждений (БСС) о ПрО вида b_{xy} = «объект x обладает свойством y»:

$$\left\| b_{xy} \right\| = \begin{Bmatrix} \textbf{Истина}, & \text{если результат измерения свойства } y \text{ объекта } x \text{ есть } \mathbf{X}; \\ \textbf{Ложь}, & \text{в противоположном случае.} \end{Bmatrix}$$

Именно на обработку таких данных ориентирован АФП, в котором используются следующие обозначения и модели [6]:

- $\boldsymbol{K} = (G^*, M, I)$ – формальный контекст, где $G^* = \{g_i\}_{i = 1,\ldots, r}$, $r = |G^*| \geq 1$ - набор объектов исследуемой ПрО, попавших в поле зрения исследователя (т.е. множество объектов обучающей выборки: $G^* \subseteq G$, где G – всё мыслимое множество объектов ПрО), $M = \{m_j\}_{j = 1,\ldots, s}$, $s = |M| \geq 1$ - множество измеряемых у объектов свойств, I – бинарное соответствие «объекты-свойства», т.е. совокупность оценок $\|b_{ij}\| \in \{\textbf{Истина}, \textbf{Ложь}\}$;
- операторы Галуа φ, ω (общая нотация «'») для контекста $\boldsymbol{K}$:
  $\varphi(X) = X' = \{m_j \mid m_j \in M, \forall g_i \in X: g_i I m_j\}$ - общие свойства объектов, составляющих $X \subseteq G^*$, или Галуа-проекция X на M;
  $\omega(Y) = Y' = \{g_i \mid g_i \in G^*, \forall m_j \in Y: g_i I m_j\}$ - объекты, которые обладают всеми свойствами из $Y \subseteq M$, или Галуа-проекция Y на G^*;
- (X, Y) – формальное понятие, у которого $X \subseteq G^*$ - объем, $Y \subseteq M$ - содержание, причем $X = Y'$, $Y = X'$;
- $\boldsymbol{B}(\boldsymbol{K})$ - множество формальных понятий контекста $\boldsymbol{K}$;
- $(\boldsymbol{B}(\boldsymbol{K}), \leq)$ –замкнутая решетка понятий, где $(X_1, Y_1) \leq (X_2, Y_2)$, если $X_1 \subseteq X_2$, или эквивалентно $Y_1 \supseteq Y_2$.

На практике качество исходного эмпирического материала таково, что, по крайней мере, для части БСС оценка истинности расплывчата (например, формируется экспертом на основе опыта или интуиции), и для оценивания таких суждений естественнее употреблять истинностные значения, вводимые многозначными логиками. Однако тогда возникает вопрос о модели, объясняющей происхождение и способ вычисления подобных оценок.

В общем плане ясно, что многозначность оценок истинности БСС вызывает неполнота данных о ПрО (неточность, противоречивость, неопределенность и т.п.), которая, однако, не находит отражения в стандартной структуре таблицы «объекты-свойства». Причины неполноты вызываются реалиями накопления эмпирической информации: выполнением, как правило, многократных независимых измерений свойства $m_j \in M$ у объекта $g_i \in G^*$; использованием для измерения одного и того же свойства m_j нескольких различных процедур (конгруэнтных источников информации); дифференциацией доверия к различным процедурам измерения. Поэтому в качестве адекватной модели исходных данных предлагается обобщенная таблица «объекты-свойства», описываемая кортежем

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

где:

- $Se = \bigcup_{i=1}^{r} Se_{(i)}$ - множество всех выполненных при зондировании ПрО серий измерений, $|Se| = \sum_{i=1}^{r} |Se_{(i)}| = m$, и $Se_{(i)} = \{se_{(i)k}\}_{k=1,\dots q_{(i)}}$, $q_{(i)} \geq 1$, $i = 1,\dots, r$ - множество серий измерений, которым подвергнут объект $g_i \in G^*$;
- $Pr = \bigcup_{j=1}^{s} Pr_{(i)}$ - арсенал всех используемых при зондировании ПрО процедур измерения, $|Pr| = \sum_{j=1}^{s} |Pr_{(j)}| = n$, и $Pr_{(j)} = \{pr_{(j)k}\}_{k=1,\dots p_{(j)}}$, $p_{(j)} \geq 1$, $j = 1,\dots, s$ - множество конгруэнтных процедур измерения свойства $m_j \in M$, причем всякая процедура $pr_{(j)k}$ характеризуется степенью доверия к ее результатам $t_{(j)k}$;
- $A = (a_{ij})_{i=1,\dots, m;\ j=1,\dots, n}$ – матрица результатов серий измерений Se свойств M у объектов из выборки G^*, выполненных с помощью процедур измерения Pr. Элементами этой матрицы могут быть константы **X** и **None**, а также еще две лингвистические константы. Константа **Failure** фиксирует отказ, сбой измерительного средства, воздержание при голосовании и т.п., т.е. тот нередко наблюдаемый на практике «результат» работы измерительной процедуры, который собирательно можно квалифицировать как «отказ от выполнения измерения» [8]. Константа **NM** (*not measured*) указывает, что в действительности в рассматриваемой серии измерений отдельная процедура измерения не использовалась (введение этого формального результата необходимо для сохранения двумерного характера обобщенной таблицы «объекты-свойства»).

Модель (1) позволяет вычислять «мягкие» оценки истинности БСС о ПрО. В [9] показано как на основе подобных данных формируется нечеткий формальный контекст $\boldsymbol{K}$, в котором соответствие «объекты-свойства» I – нечеткое отношение. В [8] на основе (1) и избрания для моделирования более адекватной многозначной векторной логики V^{TF} [10] строится нестрогий формальный контекст $\boldsymbol{K}$ с нестрогим соответствием I, которое образуют векторные оценки истинности $\|b_{ij}\| \in \langle b^+_{ij}, b^-_{ij}\rangle$, $b^+_{ij}, b^-_{ij} \in [0, 1]$, где компонент (аспект истинности) b^+_{ij} - *Истина* - формируется свидетельствами, подтверждающими БСС, а компонент (аспект) b^-_{ij} - *Ложь* - отрицающими БСС.

К сожалению на сегодня не существует эффективных методов вывода понятийной структуры ПрО непосредственно из «мягких» формальных контекстов[1]. Результативные методы основаны на предварительной α-аппроксимации входящих в такие контексты «мягких» соответствий «объекты-свойства» при задании исследователем порога доверия к исходным данным [12, 13]. Затем к полученным бинарным аппроксимациям применяются (с различными дополнениями) апробированные АФП-методы вывода понятий. Тем не менее и этот подход в общем случае оказывается некорректным, поскольку прием

[1] Например, сложный в теоретическом и вычислительном плане метод, использующий оператор замыкания нечеткого множества [11], представляет лишь академический интерес, поскольку генерирует гигантское количество нечетких понятий даже для малоразмерных «разреженных» нечетких контекстов.

α-сечения не учитывает зависимостей между измеряемыми свойствами, которые априори известны исследователю.

Интеллектуализации метода α-сечения в случае концептуальной сопряженности групп измеряемых свойств – зависимости между свойствами, возникающей в результате уже упоминавшегося когнитивного концептуального шкалирования, - посвящена работа [12]. Общие модели характерных типов связей между свойствами предложены в [14] в форме бинарных отношений «ограничений существования». В частности, пара свойств $m_j, m_k \in M, j \neq k$ для любого объекта ПрО (и, следовательно, для $\forall g_i \in G^*$) может быть:

- несовместимой, если, обладая свойством m_j, объект g_i заведомо не обладает свойством m_k, и наоборот, т.е. $E(m_j, m_k) \leftrightarrow \forall g_i \in G^*: m_j \in \{g_i\}' \rightarrow m_k \notin \{g_i\}'$;
- обусловленной, если, обладая свойством m_j, объект g_i непреложно обладает свойством m_k (хотя обратное может быть неверно), т.е. $C(m_j, m_k) \leftrightarrow \forall g_i \in G^*: m_j \in \{g_i\}' \rightarrow m_k \in \{g_i\}'$.

Для совместного описания таких «простых» ограничений и парной несовместимости измеряемых свойств в группах концептуально сопряженных свойств (ГКСС) приходится строить иерархическую модель ограничений существования. В ней простые ограничения существования задаются отдельно на уровне свойств, которые были известны до концептуального шкалирования (или «архисвойств»), и отдельно, ниже - на уровне ГКСС - фиксируется парная несовместимость измеряемых свойств в каждой ГКСС.

Предложенная модель ограничений существования свойств укрупнено будет определена кортежем (M_A, E_A, C_A), где:

- M_A - множество актуальных для исследователя архисвойств объектов ПрО, $1 \leq |M_A| \leq |M|$, $M_A = M_{A1} \cup M_{A2}$, $M_{A1} \cap M_{A2} = \varnothing$; M_{A1} – подмножество одиночных архисвойств (т.е. «нерасщеплённые» архисвойства); M_{A2} – подмножество архисвойств, подвергнутых концептуальному шкалированию (иначе говоря, архисвойства, «расщеплённые» не менее чем на два измеряемых свойства), или множество всех ГКСС: $M_{A2} = \{ Gr_1, ..., Gr_{|M_{A2}|} \}$, - причем в каждой ГКСС составляющие ее измеряемые свойства несовместимы, т.е. $(\forall m_j, m_k \in Gr_i, j \neq k) \rightarrow E(m_j, m_k) =$ **Истина**;
- E_A – пары несовместимых архисвойств, $E_A \subseteq M_A \times M_A$, $|E_A| \leq C^2_{|M_A|}$ (число сочетаний);
- C_A – пары обусловленных архисвойств, $C_A \subseteq M_A \times M_A$, $|C_A| \leq A^2_{|M_A|}$ (число размещений).

На уровне архисвойств простые ограничения существования определяют совокупность так называемых нормальных подмножеств [14]. Подмножество архисвойств $Z \subseteq M_A$ нормально тогда и только тогда, когда оно замкнуто и совместимо: Z замкнуто, если оно содержит все архисвойства, обусловленные любым элементом Z, т.е. $\forall m_j \in Z$ $(\exists m_k \in M_A: C(m_j, m_k) \rightarrow m_k \in Z)$; Z совместимо, если любые два элемента Z не связаны отношением несовместимости, т.е. $\forall m_j \in Z$ $(\exists m_k \in M_A: E(m_j, m_k) \rightarrow m_k \notin Z)$.

Очевидно, что объект обучающей выборки на уровне архисвойств может обладать лишь нормальным их подмножеством. А любое нормальное подмножество архисвойств для отдельно взятой ГКСС (т.е. «расщеплённого» архисвойства) объекта обучающей выборки устанавливает одну из двух возможностей: либо все измеряемые свойства, образующие группу, у объекта должны отсутствовать, либо объект должен обладать каким-либо одним и только одним измеряемым свойством из данной группы.

Таким образом метод рационального α-сечения «мягкого» формального контекста включает следующее:

- на основе априори формируемого исследователем ПрО состава архисвойств, «простых» ограничений их существования, а также сведений о предпринятых операциях

концептуального шкалирования строятся нормальные подмножества архисвойств и группы концептуально сопряженных измеряемых свойств;

- по указываемому исследователем порогу доверия к исходным данным определяются бинарные оценки истинности одиночных архисвойств;
- для каждого объекта обучающей выборки выявляется совокупность тех нормальных подмножеств архисвойств, которые санкционируют произведенные бинарные оценки истинности одиночных архисвойств. Если для какого-то объекта указанная совокупность окажется пустой, то констатируется недоопределенность данных для выбранного порога доверия, т.е. невозможность бинарной аппроксимации «мягкого» формального контекста;
- для каждого объекта обучающей выборки и каждой его ГКСС производится выбор порога доверия к соответствующим данным, обеспечивающего наиболее правдоподобную их бинарную аппроксимацию в рамках подхода, предложенного в [12]. Эти действия производятся для каждого санкционирующего нормального подмножества анализируемого объекта обучающей выборки с сохранением наиболее правдоподобной аппроксимации и определяющего ее порога.

Очерченные в статье модели и методы реализованы в прототипе программной лаборатории для ОАД на платформе Excel и VBA. Возможности разработанных методов и средств экспериментально исследованы при анализе данных в области экспертной идентификации внешности людей, играющей важную роль в криминалистике.

Литература

1. Искусственный интеллект. – В 3-х кн. Кн. 2. Модели и методы: Справочник / Под ред. Д.А. Поспелова. – М.: Радио и связь, 1990. – 304 с.
2. Мальцев А.И. Алгебраические системы. – М.: Наука, 1970. – 392 с.
3. Смирнов С.В. Онтологическое моделирование в ситуационном управлении // Онтология проектирования. – 2012. - №2(4). - С. 16-24.
4. Онтологии в системах искусственного интеллекта: способы построения и организации / А.В. Смирнов, М.П. Пашкин, Н.Г. Шилов и др. // Новости искусственного интеллекта. - 2002. - № 1. - С. 3-13 (Часть 1); № 2. - С. 3-9 (Часть 2).
5. Анализ данных и процессов / А.А. Барсегян, М.С. Куприянов, И.И. Холод и др. – СПб.: БХВ-Петербург, 2009. – 512 с.
6. Ganter B., Wille R. Formal Concept Analysis. Mathematical foundations. - Berlin-Heidelberg: Springer-Verlag, 1999. - 290 p.
7. Смирнов С.В. Онтологический анализ предметных областей моделирования // Известия Самарского научного центра РАН. - 2001. - Т. 3. № 1. - С. 62-70.
8. Офицеров В.П., Смирнов С.В. Использование V^{TF}-логики для определения формальных контекстов и построения онтологий предметных областей // Проблемы управления и моделирования в сложных системах: Труды XV междунар. конф. (25-28 июня 2013 г., Самара, Россия). – Самара: СамНЦ РАН, 2013. - С. 291-297.
9. Смирнов С.В. Нечеткий анализ формальных понятий и задача вывода онтологий // Труды второй международной конференции «Информационные технологии интеллектуальной поддержки принятия решений» ITIDS+RRS'2014 (18-21 мая 2014 г., Уфа, Россия). Т. 1 – Уфа: Уфимский гос. авиационный тех. ун-т, 2014. – С. 5-10.
10. Аршинский Л.В. Векторные логики: основания, концепции, модели. - Иркутск: Иркутский гос. ун-т, 2007. – 228 с.
11. Computing the lattice of all fixpoints of a fuzzy closure operator / R. Belohlavek, B. De Baets, B. Outrata, J. Vychodil // IEEE Trans. on Fuzzy systems. - 2010. - Issue 3. - Vol.18. - P. 546-557.
12. Офицеров В.П., Смирнов В.С., Смирнов С.В. Метод альфа-сечения нестрогих формальных контекстов в анализе формальных понятий // Проблемы управления и моделирования в сложных системах: Труды XVI междунар. конф. (30 июня - 03 июля 2014 г., Самара, Россия). – Самара: СамНЦ РАН, 2014. - С. 228-244.
13. De Maio C., Fenza L.V., Senatore S. Towards Automatic Fuzzy Ontology Generation // In: Proceedings of the 2009 IEEE International Conference on Fuzzy Systems (Jeju Island, Korea, 2009, August 20-24). P. 1044–1049.
14. Lammari N., Metais E. Building and maintaining ontologies: a set of algorithms // Data & Knowledge Engineering. – 2004. - Vol. 48(2). - P. 155-176.