ДСМ-метод: теоретико-множественное объяснение

О.М. Аншаков

В работе дается оригинальное изложение основных понятий и правил ДСМ-метода на языке обычной теории множеств (без использования аппарата неклассических логик). Рассматривается архитектура и принципы работы ДСМ-системы. Обсуждается связь ДСМ-метода и анализа формальных понятий. Даются практические рекомендации для разработчиков нестандартных версий ДСМ-систем.

Ключевые слова: ДСМ-метод, интеллектуальный анализ данных, правдоподобные рассуждения, анализ формальных понятий, соответствие Галуа.

Введение

ДСМ-метод автоматического порождения гипотез был предложен В.К.Финном в работах [1–3]. Основные работы по ДСМ-методу, вышедшие до 2009 года, можно найти в сборниках [4] и [5]. В настоящее время ДСМ-метод рассматривается как оригинальная совокупность логико-комбинаторных технологий интеллектуального анализа данных, использующая формализованные (с помощью неклассических логик) правила правдоподобных рассуждений. Если говорить более подробно, то ДСМ-метод обнаруживает закономерности в данных с помощью *индукции* в стиле Д.С.Милля [6], порождает правила устранения неопределенности в данных, используя *фальсификацию* в стиле К.Поппера [7], формирует предсказания с помощью оригинальной версии рассуждений по *аналогии* и оценивает правдоподобие порожденных гипотез с помощью некоторого варианта *абдукции* Ч.С.Пирса [8]. Аббревиатура «ДСМ» — это инициалы Джона Стюарта Милля, формализованные правила индуктивной логики которого являются идеологическим фундаментом ДСМ-метода.

Современный взгляд на ДСМ-метод, представленный его основателем, содержится в статьях [9–11] и книге [12]. В классическом ДСМ-методе для обнаружения закономерностей использовался формализованный аналог метода сходства Д.С.Милля. В статьях [9, 10] В.К.Финн дает формализованные описания всех правил индуктивной логики Д.С.Милля. В настоящее время идет интенсивная работа над созданием систем анализа данных, поддерживающих разные методы индуктивной логики Д.С.Милля. В этой связи, прежде всего, необходимо упомянуть статьи А.Ю.Волковой, например [13].

В большинстве работ по ДСМ-методу правила правдоподобных рассуждений описывались с помощью многозначных логик с J-операторами в стиле Россера и Тьюркетта [14]. Это обстоятельство делает изучение основ ДСМ-метода непростой задачей для разработчиков программного обеспечения, не имеющих специальной подготовки в области неклассических логик. Возможно, это является одной из причин того факта, что ДСМ-метод является не настолько широко распространенным методом анализа данных, насколько он этого заслуживает.

Попытки объяснить ДСМ-метод с помощью другой математической техники делались неоднократно, но нельзя сказать, что этих попыток было много. В этой связи следует упомянуть статью В.Б.Борщева [15], которая, по-видимому, была первой работой, содержащей нестандартное изложение ДСМ-метода.

В статье О.М.Аншакова [16] ДСМ-метод был изложен на языке обычной теории множеств. Используя упрощенную версию ДСМ-метода, описанную в [16], исследователи из Института механики МГУ им. М.В.Ломоносова создали систему автоматической классификации структур двухфазных сплавов. Подробное описание этой системы имеется в кандидатской диссертации А.С.Шундеева [17]. Работа [16] имеет также педагогическое значение, так как упрощенное изложение ДСМ-метода облегчает его пони-

мание и использование для разработки интеллектуальных систем. В учебном пособии Г.С.Осипова [18] ДСМ-метод излагается по материалам статьи [16].

В дальнейшем теоретико-множественный подход к ДСМ-методу использовался в работах А.А.Липкина, в частности, в его кандидатской диссертации [19]. Теоретико-множественный подход к правилам ДСМ-метода позволил автору этих работ сравнительно просто определить правила обобщенного ДСМ-метода для случая так называемого ранжированного ДСМ-метода и разработать прототип соответствующей ДСМ-системы.

Резюмируя сказанное выше, можно сделать вывод о том, что наличие упрощенного (даже вульгаризированного) описания ДСМ-метода является полезным обстоятельством. Оно способствует снижению барьера вхождения в сообщество разработчиков ДСМ-систем (или ДСМ-подобных систем) для независимых исследователей и инженеров. Кроме того, упрощенное описание ДСМ-метода дает возможность достаточно быстро (хотя и поверхностно) ознакомиться с ним начинающим исследователям и разработчикам, которые в дальнейшем могут продолжить изучение ДСМ-метода со всеми его тонкостями и подробностями.

Данная статья посвящена упрощенному описанию элементов ДСМ-метода на языке наивной теории множеств. Содержание этой статьи не дублирует содержание работы [16] (и вообще имеет с ним мало общего), несмотря на то, что статья [16] также посвящена изложению ДСМ-метода на теоретико-множественном языке.

В статье [16] автор делает попытку рассмотреть ДСМ-метод с некоторой общей точки зрения, например, обобщить понятие *фрагмента* исследуемого объекта, используя понятие *характеристики* и формализованного *описания*. Опыт показал, что такое обобщение оказалось невостребованным при программной реализации ДСМ-систем. Кроме того, в [16] предполагалось, что структура исследуемых объектов может быть

представлена различными способами: в виде вектора, строки, графа, множества, мультимножества и т.п. (конкретный способ представления объектов в [16] не назывался). Однако случаи использования в ДСМ-системах представления объектов, отличного от представления в виде множества, очень редки, поэтому вряд ли в популярном изложении имеет смысл о них упоминать. И, наконец, в [16] все-таки говорится о нестандартных истинностных значениях, вводятся соответствующие обозначения, описываются (хотя и неформально) правила правдоподобных рассуждений, т.е., дается представление о логических основаниях ДСМ-метода. В настоящее время автор считает, что в популярном изложении ДСМ-метода можно обойтись и без этого.

В предлагаемой читателю статье, в отличие от [16], акцент делается не на логических, а на архитектурных и алгоритмических проблемах.

1. Назначение, компоненты и основной алгоритм ДСМ-метода

В работе [9] В.К.Финн выделяет пять компонентов ДСМ-метода:

(1) условия применимости,

(2) ДСМ-рассуждения,

(3) представление знаний в виде открытых квазиаксиоматических теорий (КАТ),

(4) метатеоретические принципы и средства исследования рассуждений и предметных областей (в том числе дедуктивная имитация рассуждений, процедурная семантика и препроцессинг, результатом которого является выбор стратегий рассуждения и соответствующей им процедурной семантики),

(5) интеллектуальные системы типа ДСМ (ИС-ДСМ).

В статье [11] к ним добавляется компонент:

(6) обнаружение эмпирических закономерностей завершающих процесс knowledge discovery в базе фактов (БФ) посредством ИС-ДСМ.

В данной работе, нас будет интересовать только компонент (5) (ДСМ-системы), остальных компонентов мы будем касаться лишь в том случае, если это будет необходимо для объяснения работы ДСМ-системы. В данной статье под *ДСМ-системой* будем понимать систему интеллектуального анализа данных, т.е. извлечения знаний из данных (обнаружения закономерностей в данных) с помощью формализованных правдоподобных рассуждений, основанных на (довольно сильно модифицированных) «канонах Д.С.Милля» [6].

ДСМ-система предназначена для обнаружения связи *между структурой объекта и его поведением*. Эта связь интерпретируется как причинно-следственная связь. Как правило, *характерные черты структуры объекта считаются причинами особенностей его поведения*, но в некоторых случаях, принята обратная интерпретация: совокупность элементов поведения считается причиной множества характеристик его структуры. Это принято в обратном ДСМ-методе, введенном в статье С.М.Гусаковой, М.А.Михеенковой и В.К.Финна [20], см. также [4, ч. III, гл. 3].

Элементы структуры объекта будем называть *атомами*. Элементы поведения будем называть (целевыми) *свойствами*. Исследуемый *объект* представляется в виде конечного множества атомов. Он может обладать (или не обладать) некоторым множеством целевых свойств. Предполагается, что как у наличия, так и у отсутствия набора целевых свойств может быть причина (не обязательно единственная), эта причина является *фрагментом* структуры объекта. Так как объекты представлены множествами, фрагменты представляются подмножествами объектов, т.е. опять же конечными множествами атомов. Модель причинно-следственных связей, используемая ДСМ-методом, может быть описана графом, изображенном на рис. 1.

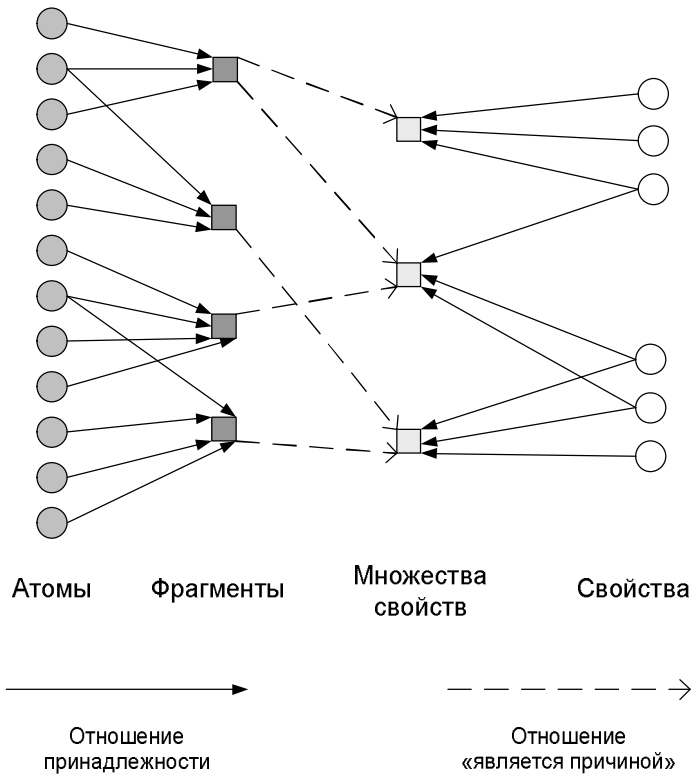


Рис. 1. Сущности ДСМ-метода и связи между ними

ДСМ-система решает две основные задачи:

(1) формирование гипотез о возможных причинах наличия (отсутствия) свойств у объектов (возможные причины являются фрагментами — подмножествами — объектов),

(2) формирование гипотез о наличии (отсутствии) свойств у тех объектов, для которых это было неизвестно.

Формирование гипотез о возможных причинах происходит с помощью *формализованных индуктивных рассуждений*. Формирование гипотез о наличии (отсутствии) свойств можно интерпретировать как применение *формализованных рассуждений по аналогии*.

Классический алгоритм ДСМ-метода, включающий работу ДСМ-системы, представляет собой итеративную процедуру, которую можно изобразить диаграммой деятельности UML[1] (см. рис. 2). Рассмотрим основные блоки этой процедуры.

Подготовка данных в контексте работы ДСМ-системы означает преобразование данных во внутренний формат ДСМ-системы. Исходные данные могут быть представлены в некотором стандартном формате, например в виде CSV-файла, где каждому объекту соответствует строка таблицы. Для работы ДСМ-системы объекты необходимо представить в виде множеств. В результате процедуры подготовки данных формируется *база фактов*, представленных во внутреннем формате ДСМ-системы.

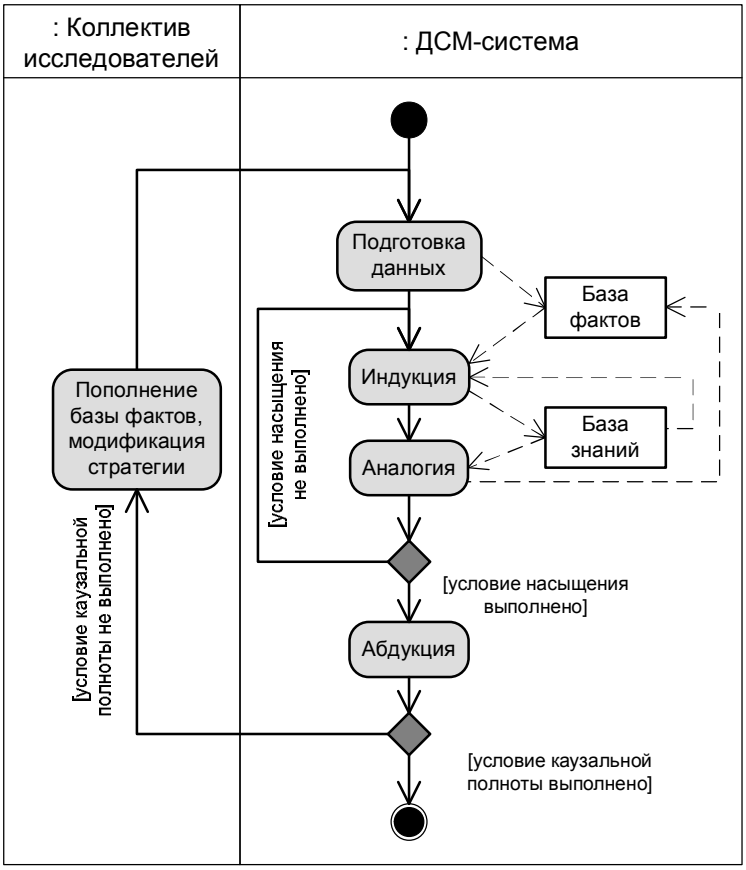


Рис. 2. Алгоритм ДСМ-метода

[1] По языку UML существует обширная литература, сошлемся, например, на наиболее научное и непредвзятое, по мнению автора, изложение из книги [21].

Индукция, т.е., применение формализованных правил индуктивных рассуждений, предназначена для порождения гипотез о возможных причинах наличия (отсутствия) целевых свойств. Гипотезы о возможных причинах представляют некоторые общие закономерности, которые включаются в *базу знаний* ДСМ-системы.

Аналогия, т.е., применение формализованных правил рассуждений по аналогии, предназначена для порождения гипотез о наличии (отсутствии) целевых свойств у тех объектов, для которых это неизвестно. Гипотезы о наличии (отсутствии) целевых свойств имеют такой же формат, как исходные факты, и включаются в *базу фактов*.

Проверка условия насыщения: условие насыщения считается выполненным, если невозможно получить новые гипотезы о наличии (отсутствии) свойств у тех объектов, для которых это было неизвестно. Поскольку количество таких объектов конечно, насыщение обязательно (рано или поздно) наступает. Если насыщение наступило, происходит выход из цикла *Индукция–Аналогия* и переход к блоку абдукции.

Абдукция — это проверка условия (критерия) каузальной полноты. Условие каузальной полноты считается выполненным, если для каждого объекта, обладающего (не обладающего) требуемым набором целевых свойств найдена причина наличия (отсутствия) этого набора свойств. Если условие каузальной полноты выполнено, то работа ДСМ-системы завершается, в противном случае происходит расширение исходного набора данных и, возможно, происходит модификация стратегии работы ДСМ-системы.

Пополнение базы фактов и модификация стратегии производится пользователями и администраторами (настройщиками) ДСМ-системы. Пополняется набор исходных данных, который преобразуется в базу фактов уже самой ДСМ-системой. Модификация стратегии включает различные операции, в частности, может быть выбрана альтернативная разновидность ДСМ-метода, изменены параметры правил индукции и аналогии и т.п.

ДСМ-рассуждением будем называть часть алгоритма ДСМ-метода, полученную удалением процедур *подготовки данных* и *пополнения базы фактов и модификации стратегии*. Эта часть относится только к ДСМ-системе и играет в ДСМ-системе очень важную роль. Введем некоторые понятия, относящиеся к работе ДСМ-рассуждению, следуя, в основном, формулировкам из статьи В.К.Финна [9]:

- *Шагом* ДСМ-рассуждения будем называть однократное выполнение процедуры индукции или процедуры аналогии.
- *Тактом* ДСМ-рассуждения будем называть однократное последовательное выполнение процедур индукции и аналогии.
- *Этапом I* ДСМ-рассуждения будем называть внутренний цикл работы ДСМ-метода, состоящий из последовательных тактов, выполняемых до тех пор, пока не будет выполнено условие насыщения.
- *Этапом II* ДСМ-рассуждения будем называть выполнение процедуры абдукции и проверку условия каузальной полноты.

В данной статье нас будет интересовать, главным образом, Этап I ДСМ-рассуждений. Мы будем называть этот этап *ядром ДСМ-системы*. Также в этой статье будут изложены некоторые общие соображения относительно процедуры подготовки данных.

2. Архитектура ДСМ-системы

Архитектура ДСМ-системы представлена на рис. 3. Она включает: ДСМ-решатель, базу фактов (данных), базу знаний и интерфейс, разделенный на два компонента: интерфейс конечного пользователя и интерфейс для настройки системы. Конечный пользователь формулирует конкретные задачи, а настройщик системы определяет стратегию и методы подготовки данных, которые будут применяться для различных

задач конечного пользователя. Роль настройщика для ДСМ-системы аналогична роли инженера по знаниям для продукционных экспертных систем.

База знаний ДСМ-системы имеет сложную структуру. Она включает:

- *предметные знания* (они описывают известные закономерности предметной области и используются, в основном, для подготовки данных), предметные знания включают правила или процедуры перевода данных во внутренний формат и правила или процедуры для определения структуры объектов,
- *метазнания* — правила или процедуры, представляющие формализованные правдоподобные рассуждения, — индукцию и аналогию,

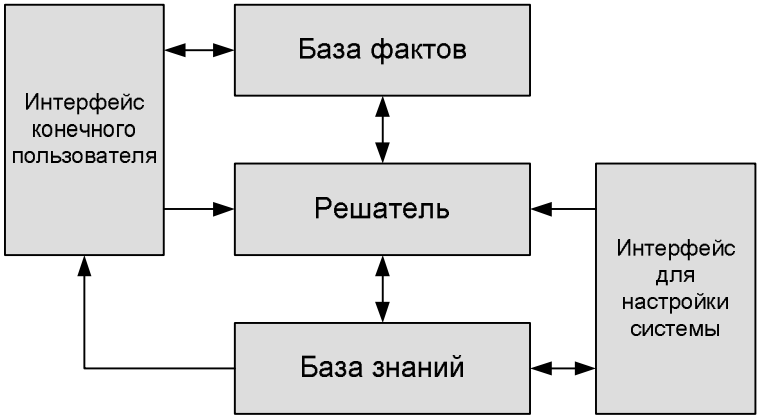


Рис. 3. Архитектура ДСМ-системы

- *гипотезы о причинах наличия (отсутствия) целевых свойств* — они имеют формат аналогичный формату данных, но описывают общие закономерности, поэтому должны интерпретироваться как знания.

База фактов содержит:

- сведения о *наличии или отсутствии целевых свойств*,
- сведения о *структуре объектов*.

Интерфейс конечного пользователя позволяет организовать ввод исходных данных и сформулировать задание для ДСМ-системы.

Интерфейс для настройки системы позволяет модифицировать правила и/или процедуры индукции и аналогии, выбрать разновидность ДСМ-метода и задать его параметры.

Решатель исполняет ту часть алгоритма ДСМ-метода, которая относится к работе ДСМ-системы (см. рис. 2). Решатель содержит *вычислитель* и *рассуждатель*. Вычислитель исполняет процедуры подготовки данных и обнаружения сходства объектов, рассуждатель реализует процедуры правдоподобных рассуждений: индукции, аналогии и абдукции.

По отношению к внутреннему циклу из алгоритма, изображенного на рис. 2 (ядру ДСМ-системы) база фактов и база знаний делятся на *постоянную* и *модифицируемую* части. Постоянная часть базы фактов — это сведения о структуре объектов. Модифицируемая часть базы фактов — это сведения о наличии (отсутствии) у объектов целевых свойств. Постоянная часть базы знаний — это предметные знания и метазнания. Модифицируемая часть базы знаний — это гипотезы о возможных причинах наличия (отсутствия) целевых свойств объектов. Структура базы фактов и базы знаний изображена на рис. 4.

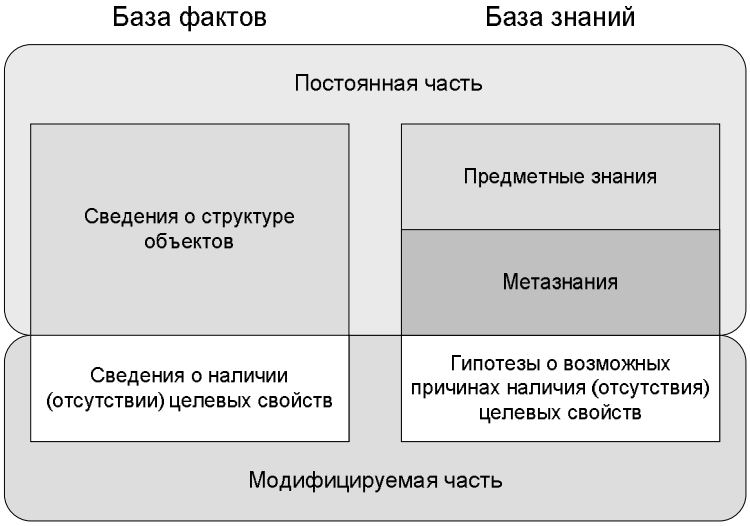


Рис. 4. Структура базы фактов и базы знаний

Модифицируемая часть базы знаний изменяется с помощью процедуры индукции, модифицируемая часть базы фактов изменяется с помощью процедуры аналогии. Постоянные части базы фактов и базы знаний не изменяются ядром ДСМ-системы.

Подчеркнем, что относительно знаний нигде в явной форме не говорилось, имеем ли мы в виду *процедурные знания* (алгоритмы) или *декларативные знания* (системы правил). Это связано с тем обстоятельством, что каждый компонент базы знаний (см. рис. 4) может быть реализован и как система процедурных, и как система декларативных знаний. Исключением являются гипотезы о возможных причинах, которые по форме напоминают факты и являются декларативными знаниями. Предметные знания, как правило, реализуются как процедурные. Метазнания могут быть реализованы и как процедурные, и как декларативные знания, хотя с точки зрения теории они являются декларативными знаниями (системой правил формализованных правдоподобных рассуждений).

3. Атомы, объекты, свойства, базовые логические функции

Рассмотрим теоретико-множественную версию математической модели предметной области, лежащую в основе ДСМ-системы.

Как уже было сказано выше, подлежащие рассмотрению *объекты* представляются в виде конечных множеств. Все возможные элементы таких множеств образуют *универсум атомов*, который будем обозначать через $\mathbf{A}$. Множество подлежащих рассмотрению объектов — *универсум объектов* — будем обозначать через $\mathbf{O}$. Очевидно, $\mathbf{O} \subseteq 2^{\mathbf{A}}$, где через $2^{\mathbf{A}}$ обозначено множество всех подмножеств множества $\mathbf{A}$. Каждое подмножество множества $\mathbf{A}$ будем называть *фрагментом*, а множество $2^{\mathbf{A}}$ — *универсумом фрагментов*. Универсум фрагментов будем обозначать через $\mathbf{F}$.

Разумеется, для реальной ДСМ-системы универсум фрагментов — это просто абстракция. Никто никогда не пытался целиком построить этот универсум, да это и невозможно, поскольку в реальных задачах универсум атомов $\mathbf{A}$ может содержать несколько сот, а то и тысяч, элементов.

Множество всех (целевых) свойств, которыми (по нашим представлениям) могут обладать объекты, будем называть *универсумом целевых свойств* и обозначать через $\mathbf{T}$. Совокупность подлежащих рассмотрению множеств целевых свойств будем называть *универсумом значимых множеств целевых свойств* и обозначать через $\mathbf{P}$. Относительно множеств целевых свойств необходимо учитывать следующие обстоятельства:

- как правило, множество целевых свойств воспринимается как единое целое,
- обычно, множество $\mathbf{T}$ невелико по объему или просто одноэлементно,
- в большинстве случаев, элементы множества $\mathbf{P}$ являются одноэлементными подмножествами множества $\mathbf{T}$.

Через $\mathbf{B}$ будем обозначать множество классических истинностных значений: *false* и *true*. Обозначать эти значения будем через 0 и 1, соответственно, т.е., $\mathbf{B} = \{0,1\}$.

Будем предполагать, что у нас есть возможность определить следующие вычислимы функции со значениями из $\mathbf{B}$:

- $\mathrm{P}^+ : \mathbf{O} \times \mathbf{P} \to \mathbf{B}$ — функция «обладает множеством свойств», для любых $o \in \mathbf{O},\ p \in \mathbf{P}$

$$\mathrm{P}^+(o, p) = \begin{cases} 1, & \text{если объект } o \text{ обладает множеством свойств } p, \\ 0, & \text{в противном случае.} \end{cases}$$

- $\mathrm{P}^- : \mathbf{O} \times \mathbf{P} \to \mathbf{B}$ — функция «не обладает множеством свойств», для любых $o \in \mathbf{O},\ p \in \mathbf{P}$

$$\mathrm{P}^{-}(o,p)=\begin{cases}1, & \text{если объект } o \text{ не обладает множеством свойств } p,\\ 0, & \text{в противном случае.}\end{cases}$$

– $\mathrm{C}^{+}:\mathbf{F}\times\mathbf{P}\to\mathbf{B}$ — функция «является возможной причиной наличия множества свойств», для любых $s\in\mathbf{F},\ \ p\in\mathbf{P}$

$$\mathrm{C}^{+}(s,p)=\begin{cases}1, & \text{фрагмент } s \text{ является причиной наличия множества свойств } p,\\ 0, & \text{в противном случае.}\end{cases}$$

– $\mathrm{C}^{-}:\mathbf{F}\times\mathbf{P}\to\mathbf{B}$ — функция «является возможной причиной отсутствия множества свойств», для любых $s\in\mathbf{F},\ \ p\in\mathbf{P}$

$$\mathrm{C}^{-}(s,p)=\begin{cases}1, & \text{фрагмент } s \text{ является причиной отсутствия множества свойств } p,\\ 0, & \text{в противном случае.}\end{cases}$$

Необходимо отметить, что значения логических функций P^{+}, P^{-}, C^{+} и C^{-} зависят от результатов наблюдений и/или экспериментов или от результатов *недостоверных* рассуждений. Сведения, на основании которых определяются значения функций P^{+}, P^{-}, C^{+} и C^{-}, могут быть неполными, неточными и даже противоречивыми. Поэтому мы не можем определенно говорить о наличии или отсутствии свойств и, тем более, о возможных причинах этого наличия или отсутствия.

Более правильно говорить о том, что имеются аргументы за (или против) наличия или отсутствия свойств и т.п. Например, функцию $\mathrm{P}^{+}:\mathbf{O}\times\mathbf{P}\to\mathbf{B}$ правильнее было бы описать так: для любых $o\in\mathbf{O},\ \ p\in\mathbf{P}$

$$\mathrm{P}^{+}(o,p)=\begin{cases}1, & \text{если имеется достаточно убедительный набор аргументов ЗА то,}\\ & \text{что объект } o \text{ обладает множеством свойств } p,\\ 0, & \text{в противном случае.}\end{cases}$$

А функцию $C^+ : \mathbf{F} \times \mathbf{P} \to \mathbf{B}$ правильнее описать следующим образом: для любых $s \in \mathbf{F}$, $p \in \mathbf{P}$

$$C^+(s,p) = \begin{cases} 1, & \text{если имеется достаточно убедительный набор аргументов ЗА то,} \\ & \text{что фрагмент } s \text{ является причиной наличия множества свойств } p, \\ 0, & \text{в противном случае.} \end{cases}$$

Аналогично, могут быть неформально описаны функции P^- и C^-. Что понимается под «достаточно убедительным набором аргументов ЗА», зависит от конкретной разновидности ДСМ-метода, а, иногда, и от конкретной задачи, решаемой ДСМ-системой.

4. Положительные и отрицательные примеры и гипотезы

Будем использовать введенные в предыдущем разделе обозначения для универсумов атомов, объектов, фрагментов и т.п.

Объект $o \in \mathbf{O}$ будем называть *положительным примером* для множества свойств $p \in \mathbf{P}$, если $P^+(o,p) = 1$, т.е., если существует достаточно убедительный набор аргументов *за* то, что o *обладает* множеством свойств p.

Множество всех положительных примеров для набора свойств p будем обозначать через $\mathbf{O}^+(p)$.

Объект $o \in \mathbf{O}$ будем называть *отрицательным примером* для множества свойств $p \in \mathbf{P}$, если $P^-(o,p) = 1$, т.е., если существует достаточно убедительный набор аргументов *за* то, что o *не обладает* множеством свойств p.

Множество всех отрицательных примеров для набора свойств p будем обозначать через $\mathbf{O}^-(p)$.

Объект $o \in \mathbf{O}$ будем называть *противоречивым примером* для множества свойств $p \in \mathbf{P}$, если он одновременно является и положительным и отрицательным примером для этого множества свойств, т.е., если $\mathrm{P}^{+}(o, p) = 1$ и $\mathrm{P}^{-}(o, p) = 1$. Это возможно лишь в случае достаточно убедительного набора аргументов как *за* наличие, так и *за* отсутствие у объекта o множества свойств p.

Множество всех противоречивых примеров для набора свойств p будем обозначать через $\mathbf{O}^{0}(p)$.

Объект $o \in \mathbf{O}$ будем называть *неопределенным примером* для множества свойств $p \in \mathbf{P}$, если он не является ни положительным, ни отрицательным примером для этого множества свойств, т.е., если $\mathrm{P}^{+}(o, p) = 0$ и $\mathrm{P}^{-}(o, p) = 0$. Это возможно в случае отсутствия достаточно убедительного набора аргументов и *за* наличие, и *за* отсутствие у объекта o множества свойств p.

Множество всех неопределенных примеров для набора свойств p будем обозначать через $\mathbf{O}^{\tau}(p)$.

Объект $o \in \mathbf{O}$ будем называть *чисто положительным примером* для множества свойств $p \in \mathbf{P}$, если $\mathrm{P}^{+}(o, p) = 1$ и $\mathrm{P}^{-}(o, p) = 0$. В этом случае существует достаточно убедительный набор аргументов *за* то, что o *обладает* множеством свойств p и не существует достаточно убедительного набора аргументов *за* то, что o *не обладает* множеством свойств p.

Множество всех чисто положительных примеров для набора свойств p будем обозначать через $\mathbf{O}^{(+)}(p)$.

Объект $o \in \mathbf{O}$ будем называть *чисто отрицательным примером* для множества свойств $p \in \mathbf{P}$, если $\mathrm{P}^{-}(o, p) = 1$ и $\mathrm{P}^{+}(o, p) = 0$. В этом случае существует достаточно

убедительный набор аргументов *за* то, что *о не обладает* множеством свойств p и не существует достаточно убедительного набора аргументов *за* то, что *о обладает* множеством свойств p.

Множество всех чисто отрицательных примеров для набора свойств p будем обозначать через $\mathbf{O}^{(-)}(p)$.

Очевидно, что между множествами $\mathbf{O}^{+}(p)$, $\mathbf{O}^{-}(p)$, $\mathbf{O}^{0}(p)$, $\mathbf{O}^{\tau}(p)$, $\mathbf{O}^{(+)}(p)$ и $\mathbf{O}^{(-)}(p)$ имеют место следующие соотношения:

$$\mathbf{O}^{0}(p)=\mathbf{O}^{+}(p)\cap\mathbf{O}^{-}(p),$$

$$\mathbf{O}^{(+)}(p)=\mathbf{O}^{+}(p)\setminus\mathbf{O}^{0}(p),$$

$$\mathbf{O}^{(-)}(p)=\mathbf{O}^{-}(p)\setminus\mathbf{O}^{0}(p),$$

$$\mathbf{O}^{\tau}(p)=\mathbf{O}\setminus\left(\mathbf{O}^{+}(p)\cup\mathbf{O}^{-}(p)\right),$$

$$\mathbf{O}^{\tau}(p)=\mathbf{O}\setminus\left(\mathbf{O}^{(+)}(p)\cup\mathbf{O}^{(-)}(p)\cup\mathbf{O}^{0}(p)\right),$$

$$\mathbf{O}=\mathbf{O}^{+}(p)\cup\mathbf{O}^{-}(p)\cup\mathbf{O}^{\tau}(p),$$

$$\mathbf{O}=\mathbf{O}^{(+)}(p)\cup\mathbf{O}^{(-)}(p)\cup\mathbf{O}^{0}(p)\cup\mathbf{O}^{\tau}(p).$$

Термин «пример» мы используем по отношению к объектам в контексте обладания набором целевых свойств. Термин «гипотеза» мы будем использовать по отношению к фрагментам, которые могут быть возможными причинами наличия или отсутствия наборов целевых свойств.

Фрагмент $s\in\mathbf{F}$ будем называть *положительной гипотезой* для множества свойств $p\in\mathbf{P}$, если $\mathrm{C}^{+}(s,p)=1$, т.е., если существует достаточно убедительный набор аргументов *за* то, что *s является возможной причиной наличия* множества свойств p.

Множество всех положительных гипотез для набора свойств p будем обозначать через $\mathbf{F}^{+}(p)$.

Фрагмент $s \in \mathbf{F}$ будем называть *отрицательной гипотезой* для множества свойств $p \in \mathbf{P}$, если $\mathrm{C}^{-}(s, p) = 1$, т.е., если существует достаточно убедительный набор аргументов *за* то, что s *является возможной причиной отсутствия* множества свойств p.

Множество всех отрицательных гипотез для набора свойств p будем обозначать через $\mathbf{F}^{-}(p)$.

Фрагмент $s \in \mathbf{F}$ будем называть *противоречивой гипотезой* для множества свойств $p \in \mathbf{P}$, если он одновременно является и положительной и отрицательной гипотезой для этого множества свойств, т.е., если $\mathrm{C}^{+}(s, p) = 1$ и $\mathrm{C}^{-}(s, p) = 1$. Это возможно лишь в случае достаточно убедительного набора аргументов как *за* то, что s является возможной причиной наличия, так и *за* то, что s является возможной причиной отсутствия множества свойств p.

Множество всех противоречивых гипотез для набора свойств p будем обозначать через $\mathbf{F}^{0}(p)$.

Фрагмент $s \in \mathbf{F}$ будем называть *неопределенной гипотезой* для множества свойств $p \in \mathbf{P}$, если он не является ни положительной, ни отрицательной гипотезой для этого множества свойств, т.е., если $\mathrm{C}^{+}(s, p) = 0$ и $\mathrm{C}^{-}(s, p) = 0$. Это возможно в случае отсутствия достаточно убедительного набора аргументов и *за* то, что s является возможной причиной наличия, и *за* то, что s является возможной причиной отсутствия множества свойств p.

Множество всех неопределенных гипотез для набора свойств p будем обозначать через $\mathbf{F}^{\tau}(p)$.

Фрагмент $s \in \mathbf{F}$ будем называть *чисто положительной гипотезой* для множества свойств $p \in \mathbf{P}$, если $\mathrm{C}^{+}(s, p) = 1$ и $\mathrm{C}^{-}(s, p) = 0$. В этом случае существует достаточно убедительный набор аргументов *за* то, что s является возможной причиной *наличия* множества свойств p, и не существует достаточно убедительного набора аргументов *за* то, что s является возможной причиной *отсутствия* множества свойств p.

Множество всех чисто положительных гипотез для набора свойств p будем обозначать через $\mathbf{F}^{(+)}(p)$.

Фрагмент $s \in \mathbf{F}$ будем называть *чисто отрицательной гипотезой* для множества свойств $p \in \mathbf{P}$, если $\mathrm{C}^{-}(s, p) = 1$ и $\mathrm{C}^{+}(s, p) = 0$. В этом случае существует достаточно убедительный набор аргументов *за* то, что s является возможной причиной *отсутствия* множества свойств p, и не существует достаточно убедительного набора аргументов *за* то, что s является возможной причиной *наличия* множества свойств p.

Множество всех чисто отрицательных гипотез для набора свойств p будем обозначать через $\mathbf{F}^{(-)}(p)$.

Между множествами гипотез $\mathbf{F}^{+}(p)$, $\mathbf{F}^{-}(p)$, $\mathbf{F}^{0}(p)$, $\mathbf{F}^{\tau}(p)$, $\mathbf{F}^{(+)}(p)$ и $\mathbf{F}^{(-)}(p)$ имеют место соотношения, аналогичные соотношениям между соответствующими множествами примеров, а именно:

$$\mathbf{F}^{0}(p) = \mathbf{F}^{+}(p) \cap \mathbf{F}^{-}(p),$$

$$\mathbf{F}^{(+)}(p) = \mathbf{F}^{+}(p) \setminus \mathbf{F}^{0}(p),$$

$$\mathbf{F}^{(-)}(p) = \mathbf{F}^{-}(p) \setminus \mathbf{F}^{0}(p),$$

$$\mathbf{F}^{\tau}(p) = \mathbf{F} \setminus \left(\mathbf{F}^{+}(p) \cup \mathbf{F}^{-}(p) \right),$$

$$\mathbf{F}^{\tau}(p) = \mathbf{F} \setminus \left(\mathbf{F}^{(+)}(p) \cup \mathbf{F}^{(-)}(p) \cup \mathbf{F}^{0}(p) \right),$$

$$\mathbf{F} = \mathbf{F}^{+}(p) \cup \mathbf{F}^{-}(p) \cup \mathbf{F}^{\tau}(p),$$

$$\mathbf{F} = \mathbf{F}^{(+)}(p) \cup \mathbf{F}^{(-)}(p) \cup \mathbf{F}^{0}(p) \cup \mathbf{F}^{\tau}(p).$$

5. Сходства и кластеры. Соответствия Галуа

Будем находить возможную причину набора целевых свойств как общую часть (пересечение) некоторого семейства объектов, этим набором свойств обладающих. В работах по ДСМ-методу общую часть объектов называют их *сходством*, т.е. термин «сходство» употребляется в смысле *операции* (пересечения), а не отношения, обладающего свойствами рефлексивности и симметричности (такие отношения называются отношениями *толерантности* [22]).

Здесь нужно отметить следующее обстоятельство: мы допускаем, что *у одного и того же* набора целевых свойств могут быть *различные* причины. Если мы рассмотрим множество объектов, обладающих некоторым набором свойств p, то каждая возможная причина наличия этого свойства порождает свой *кластер* (в один кластер собираются объекты, имеющие общую причину наличия у них свойства p). Заметим, что возможная причина (фрагмент), которая определяется как пересечение некоторого семейства объектов, должна включаться в каждый объект, принадлежащий этому семейству.

Теперь дадим формальные определения двум двойственным понятиям: *сходства* семейства объектов и *кластера* фрагмента.

Пусть $O \subseteq \mathbf{O}$ — некоторое семейство объектов, представленных в виде множеств. *Сходством* этого семейства будем называть фрагмент равный пересечению мно-

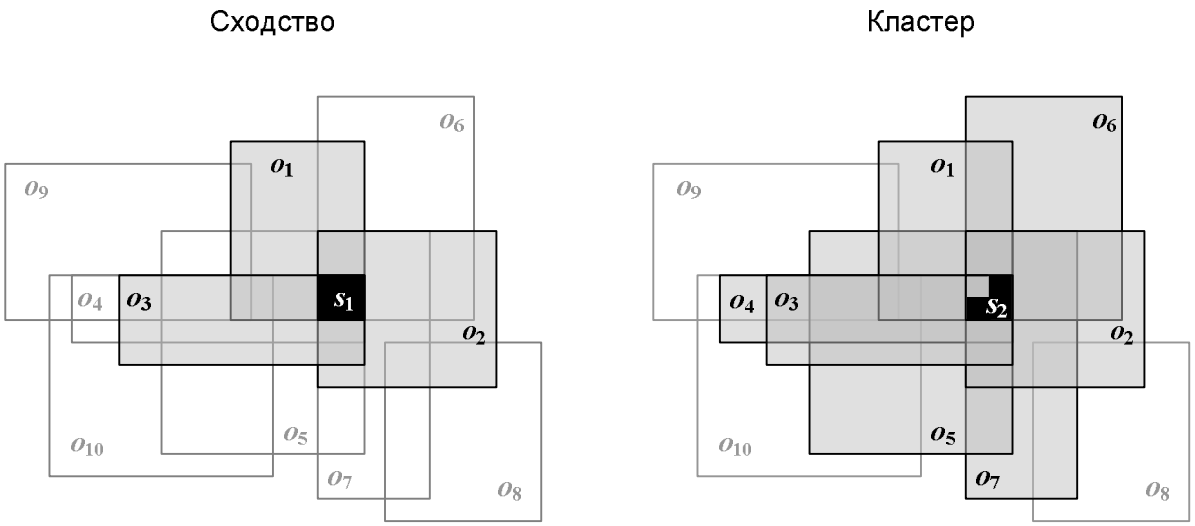
жеств, представляющих объекты. Обозначать сходство семейства объектов O будем через O^{Si}. Таким образом, по определению,

$$O^{\mathrm{Si}} = \bigcap_{o \in O} o = \{a \in \mathbf{A} \mid (\forall o \in O)(a \in o)\}. \qquad (1)$$

Пусть $s \in \mathbf{F}$ — некоторый фрагмент, т.е. подмножество универсума **A**. *Кластером* фрагмента s назовем семейство всех объектов из **O**, включающих s. Обозначать кластер фрагмента s будем через s^{Cl}. Таким образом, по определению,

$$s^{\mathrm{Cl}} = \{o \in \mathbf{O} \mid s \subseteq o\} \quad \{o \in \mathbf{O} \mid (\forall a \in s)(a \in o)\}. \qquad (2)$$

Цепочки равенств (1) и (2) мы в дальнейшем рассмотрим более подробно, когда будем говорить о связи ДСМ-метода и анализа формальных понятий.



$O = \{o_1, o_2, o_3\}, \qquad s_1 = O^{\mathrm{Si}} = o_1 \cap o_2 \cap o_3 \qquad s_2^{\mathrm{Cl}} = \{o \in \mathrm{O} \mid s_2 \subseteq o\} = \{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$

Рис. 5. Сходство и кластер

На рис. 5 дана наглядная иллюстрация определениям сходства семейства объектов и кластера фрагмента. Объекты на этом рисунке изображены в виде прямоугольников. Фрагмент может быть любой фигурой (не обязательно прямоугольником). На ле-

вой части рисунка хорошо видно, что семейство объектов, порождающих сходство, не обязано включать *все* объекты, которые порождают соответствующее пересечение. На правой части рисунка видно, что фрагмент, порождающий кластер, не обязательно является *пересечением* всех элементов кластера.

Определив сходство семейства объектов и кластер фрагмента, мы, фактически, определили два отображения: $^{\mathrm{Si}}: 2^{\mathbf{O}} \to 2^{\mathbf{A}}$ и $^{\mathrm{Cl}}: 2^{\mathbf{A}} \to 2^{\mathbf{O}}$. Заметим, что на семействах множеств $2^{\mathbf{O}}$ и $2^{\mathbf{A}}$ теоретико-множественное включение $\subseteq$ является отношение частичного порядка. Пара отображений $\left\langle {}^{\mathrm{Si}}, {}^{\mathrm{Cl}} \right\rangle$ обладает некоторым набором свойств, описывающим связь этих отображений с частичным порядком $\subseteq$ на семействах множеств $2^{\mathbf{O}}$ и $2^{\mathbf{A}}$.

Для любых $O, O_1, O_2 \in 2^{\mathbf{O}}$ и любых $s, s_1, s_2 \in 2^{\mathbf{A}}$ верно:

(i) $O_1 \subseteq O_2$ влечет $O_2{}^{\mathrm{Si}} \subseteq O_1{}^{\mathrm{Si}}$ (антитонность отображения $^{\mathrm{Si}}$),

(ii) $s_1 \subseteq s_2$ влечет $s_2{}^{\mathrm{Cl}} \subseteq s_1{}^{\mathrm{Cl}}$ (антитонность отображения $^{\mathrm{Cl}}$),

(iii) $O \subseteq O^{\mathrm{Si}\circ\mathrm{Cl}}$ (экстенсивность композиции $^{\mathrm{Si}\circ\mathrm{Cl}}$),

(iv) $s \subseteq s^{\mathrm{Cl}\circ\mathrm{Si}}$ (экстенсивность композиции $^{\mathrm{Cl}\circ\mathrm{Si}}$).

Справедливость утверждений (i)–(iv) означает, что пара отображений $\left\langle {}^{\mathrm{Si}}, {}^{\mathrm{Cl}} \right\rangle$ является соответствием Галуа в смысле определения из [23].

Понятие *соответствия Галуа* (Galois connection) относится к частично упорядоченным множествам. В [23] это понятие определяется следующим образом:

Пусть $\langle M, \leq \rangle$ и $\langle M', \leq' \rangle$ — два частично упорядоченных множества. Упорядоченная пара отображений $\langle \varphi, \psi \rangle$, где $\varphi : M \to M'$, $\psi : M' \to M$, называется *соответствием Галуа* (между этими частично упорядоченными множествами), если для любых $a, b \in M$ и любых $a', b' \in M'$ верно:

(a) $a \le b$ влечет $b^{\varphi} \le' a^{\varphi}$ (антитонность отображения φ),

(b) $a' \le' b'$ влечет $b'^{\psi} \le a'^{\psi}$ (антитонность отображения ψ),

(c) $a \le a^{\varphi\circ\psi}$ (экстенсивность композиции $\varphi\circ\psi$),

(d) $a' \le' a'^{\psi\circ\varphi}$ (экстенсивность композиции $\psi\circ\varphi$).

Здесь мы используем постфиксную форму представления значения функции. Т.е., вместо $\varphi(a)$ пишем a^{φ} (аналогично, вместо $\psi(a')$ пишем a'^{ψ}).

Можно показать, что пара антитонных отображений $\langle\varphi,\psi\rangle$, где $\varphi: M \to M'$, $\psi: M' \to M$, является соответствием Галуа (между частично упорядоченными множествами $\langle M,\le\rangle$ и $\langle M',\le'\rangle$) в том и только в том случае, если для любых $a \in M$ и $a' \in M'$:

$$a' \le' a^{\varphi} \text{ тогда и только тогда, когда } a \le a'^{\psi}.$$

Этот факт позволяет дать соответствию Галуа эквивалентное определение, несколько более короткое, чем определение из [23].

Применяя приведенное выше утверждение к соответствию Галуа $\left\langle {}^{\mathrm{Si}}, {}^{\mathrm{Cl}} \right\rangle$ получаем, что для любых $O \in 2^{\mathbf{O}}$ и $s \in 2^{\mathbf{A}}$:

$$O \subseteq s^{\mathrm{Cl}} \text{ тогда и только тогда, когда } s \subseteq O^{\mathrm{Si}}. \qquad (3)$$

Рис. 5 с $O = \{o_1, o_2, o_3\}$ и $s = s_2$ является наглядной иллюстрацией соотношения (3). Однако, нас прежде всего будет интересовать тот случай, когда в соотношении (3) включения будут заменены на равенства. Для того, чтобы рассматривать этот случай более содержательно, введем следующее определение:

Пусть $O \in 2^{\mathbf{O}}$, $s \in 2^{\mathbf{A}}$. Будем говорить, что пара $\langle O, s \rangle$ удовлетворяет *условию исчерпываемости* относительно множества $\mathbf{O}$, если справедливы равенства:

$$O = s^{\mathrm{Cl}}, \tag{4}$$

$$s = O^{\mathrm{Si}}. \tag{5}$$

Равенства (4) и (5), а также (1) и (2), потребуются нам при обсуждении связи ДСМ-метода с анализом формальных понятий. Условие исчерпываемости будет необходимо при определении так называемых решающих предикатов ДСМ-метода.

Связь конструкций ДСМ-метода с соответствиями Галуа обнаружил С.О.Кузнецов. См., например, [24].

6. Кандидаты в возможные причины

Пусть s — фрагмент, т.е., $s \in \mathbf{F} = 2^{\mathbf{A}}$. Будем говорить, что s является *кандидатом в возможные причины наличия набора свойств* $p \in \mathbf{P}$, если существует $O \subseteq \mathbf{O}^{(+)}(p)$ такое, что выполняются следующие условия:

(i) пара $\langle O, s \rangle$ удовлетворяет условию исчерпываемости относительно $\mathbf{O}^{(+)}(p)$;

(ii) $s \neq \varnothing$;

(iii) $|O| \geq 2$.

Здесь через $|O|$ обозначена мощность множества O.

Приведем некоторые аргументы в пользу именно такого определения кандидата в возможные причины. Рассмотрим сначала условие (i). Напомним, что условие исчер-

пываемости предполагает справедливость равенства (5). Значит, $s = O^{\mathrm{Si}}$, т.е., s является сходством (пересечением) множества объектов O.

ДСМ-метод основан на предположении о том, что возможная причина *наличия* набора свойств должна определяться как нечто общее (сходство) некоторого множества *положительных* примеров для этого набора свойств. В данном случае мы поступаем более жестко и отбираем только *чисто положительные* примеры.

Условие исчерпываемости кроме равенства (5) включает также равенство (4). Значит, $O = s^{\mathrm{Cl}}$. Требование справедливости равенства (4) обусловлено менее очевидными соображениями, связанными с эффективностью алгоритмов порождения всех различных пересечений.

Заметим, что один и тот же фрагмент s может быть получен как пересечение различных (иногда даже дизъюнктных) семейств объектов, но существует *единственное* семейство O, для которого пара $\langle O, s \rangle$ удовлетворяет условию исчерпываемости относительно некоторого заранее выбранного универсума объектов. Этот факт позволяет определять такие алгоритмы построения всех (разных) пересечений, во время работы которых каждое такое пересечение порождается ровно один раз. Т.е., эти алгоритмы будут иметь сложность *линейную относительно результата* (количества порожденных пересечений). Разумеется, в наихудшем случае, даже такие алгоритмы имеют сложность *экспоненциальную относительно размера исходных данных* (количества объектов). Однако, данные реальных задач таковы, что наихудший случай для них практически невозможен.

Рассмотрим теперь условие (ii). Оно требует, чтобы кандидат в возможные причины был непуст. Это условие можно объяснить исходя из элементарного здравого смысла. Отсутствие чего бы то ни было не может быть возможной причиной наличия непустого набора свойств.

Условие (iii) связано с классической формулировкой метода сходства (method of agreement) Д.С.Милля [6], в котором присутствуют «… два или более случаев рассматриваемого явления…». Каждому такому случаю соответствует объект из семейства O.

Рассмотрим теперь вопрос о том, как можно изменить определение кандидата в возможные причины с целью его адаптации для решения собственных прикладных задач. Сначала необходимо описать общую структуру этого определения. *Оно содержит преамбулу и три условия.* В преамбуле говорится о том, что существует семейство объектов O, каким-то образом связанное с фрагментом s, который рассматривается в качестве кандидата в возможные причины. Это семейство объектов O фактически порождает фрагмент s. В работах по ДСМ-методу такое семейство объектов может быть названо по-разному: орбита, множество образующих, множество порождающих, множество родителей (фрагмента s). В данной работе мы для краткости будем называть семейство O *родителем* фрагмента s.

В условии (i) раскрывается характер связи между фрагментом s и его родителем O. В условии (ii) накладываются ограничения на сам фрагмент s. В условии (iii) ограничения накладываются на родителя O.

Предположим, что мы собираемся сохранить общую структуру этого определения. Вопрос о том, как можно изменить преамбулу и условие (i) рассмотрим позже. А сейчас рассмотрим возможные модификации условий (ii) и (iii).

Модификация условия (ii) может потребоваться для того, чтобы более точно отразить характер предметной области. Предположим, что экспертам предметной области известен некоторый набор фрагментов, которые ни при каких обстоятельствах не могут быть возможными причинами множества целевых свойств p. Обозначим этот набор «незначащих» фрагментов через N. Будем предполагать, что $\varnothing \in N$. Тогда условие (ii) будет записано в виде

$$s \notin N.$$

Теперь предположим, что множество атомов **A** содержит подмножество S «незначащих элементов». Примерами таких элементов могут служить стоп-слова, от которых принято избавляться при анализе текстов. Логично предположить, что фрагмент, полностью состоящий из «незначащих элементов» сам будет «незначащим» и не может рассматриваться в качестве кандидата в возможные причины. В этом случае условие (ii) представляется в виде

$$s \not\subseteq S.$$

Возможна также ситуация, когда специалистов в предметной области, прежде всего, будут интересовать такие возможные причины, размер которых находится в определенном диапазоне. Тогда условие (ii) будет записано в виде

$$a \le |s| \le b,$$

где $0 < a < b.$

Теперь рассмотрим возможные варианты модификации условия (iii). Во-первых, может быть изменено пороговое значение мощности родителя. Например, условие (iii) может быть переписано в виде

$$|O| \ge 10.$$

Тогда от фрагмента s будет требоваться, чтобы он являлся пересечением по крайней мере 10 различных чисто положительных примеров.

Во-вторых, пороговое значение мощности может зависеть от количества чисто положительных примеров или от общего количества примеров. В этом случае, мы бу-

дем учитывать частоту встречаемости фрагмента в рассматриваемых примерах. Например, можно записать условие (iii) в виде

$$|O| \geq 0{,}1 \cdot \left|\mathbf{O}^{(+)}(p)\right|.$$

Тогда фрагмент s должен быть общей частью по крайней мере 10% чисто положительных примеров.

В некоторых случаях (например, при анализе текстов) наиболее полезными могут оказаться фрагменты, которые встречаются не слишком часто, но и не слишком редко. Для таких случаев условие (iii) может быть переписано в виде

$$a \cdot \left|\mathbf{O}^{(+)}(p)\right| \leq |O| \leq b \cdot \left|\mathbf{O}^{(+)}(p)\right|,$$

где $0 < a < b < 1$.

Что касается преамбулы и условия (i), то наиболее естественной их модификацией была бы замена множества *чисто* положительных примеров $\mathbf{O}^{(+)}(p)$ на множество положительных примеров $\mathbf{O}^{+}(p)$. Эта замена (вместе с некоторыми изменениями стратегии ДСМ-рассуждений) позволила бы рассматривать аргументы «за» и аргументы «против» совершенно независимо друг от друга, оставляя последнее решающее слово эксперту.

Возможно и более радикальное изменение условия (i), включающее замену условия исчерпываемости на более простое условие. Автор данной статьи неоднократно высказывал мысль о необходимости «отделить в ДСМ-методе логику от комбинаторики» и отдельно рассматривать вопросы о *порождении сходств* (пересечений множеств, представляющих объекты) и *проверке этих сходств на соответствие условиям*, выраженным средствами логики предикатов первого порядка. Это идея не была поддержана

членами сообщества исследователей, занимающихся ДСМ-методом, и впоследствии автор от нее отказался.

Следует отметить, что с условием исчерпываемости связаны весьма непростые и тонкие вопросы математического, концептуального и технологического характера. Их подробное обсуждение в относительно популярной работе автору кажется неуместным. Однако некоторые замечания все-таки необходимо сделать.

В условие исчерпываемости входит равенство

$$s = O^{\mathrm{Si}} = \bigcap_{o \in O} o.$$

Если попытаться выразить его непосредственно через бинарную операцию сходства (пересечения), то в результете будет получена формула переменной длины, содержащая квантор по натуральным числам

$$\exists n \, \exists o_1 \ldots \exists o_n \left(s = o_1 \cap \ldots \cap o_n \right). \qquad (6)$$

Эта формула включается в формулу, описывающую так называемый решающий предикат ДСМ-метода (фактически, в определение кандидата в возможные причины), встречающуюся во многих работах по ДСМ-методу. В статье В.К.Финна [9] определение решающего предиката сопровождается подробными комментариями.

Однако, формула (6) не является формулой логики предикатов первого порядка. Это осложняет представление алгоритмов ДСМ-метода средствами логического программирования, хотя для ДСМ-метода, использующего рассуждения по формальным правилам, аппарат логического программирования является вполне естественным.

Теоретическая основа для применения логического программирования в ДСМ-методе была заложена в работах Д.В.Виноградова [25], [26]; наиболее известная прак-

тическая реализация ДСМ-метода на языке логического программирования описана в работе М.А.Михеенковой и Т.Л.Феофановой [27].

Д.В.Виноградовым в [28] было доказано, что формулу (6) (а значит и условие исчерпываемости) в контексте некоторой теории ДСМ-метода можно представить с помощью формулы логики предикатов первого порядка. Это весьма интересный теоретический результат, но, к сожалению, первопорядковое представление для формулы (6) не очень удобно для практического использования.

Заканчивая отступление, посвященное условию исчерпываемости, и возвращаясь к основной теме данного раздела, необходимо отметить, что мы рассмотрели лишь малую часть вариантов определения кандидата в возможные причины. Подводя итог обсуждению вопроса о модификациях этого определения, можно сказать, что различные его варианты позволяют достаточно тонко настраивать ДСМ-подобную систему, приспосабливая ее к решению конкретных задач.

Определение *кандидата в возможные причины отсутствия набора свойств* $p \in \mathbf{P}$ является *двойственным* определению кандидата в возможные причины наличия этого набора свойств, данному в начале текущего раздела. Требуется просто заменить $\mathbf{O}^{(+)}(p)$ на $\mathbf{O}^{(-)}(p)$. Возможные модификации нового определения будут аналогичны возможным модификациям исходного определения.

7. Решающие предикаты. Правила индукции

Пусть $s \in \mathbf{F}$, $p \in \mathbf{P}$. Тот факт, что s является кандидатом в возможные причины наличия множества свойств p будем обозначать через $\mathrm{M}^{+}(s, p)$. Тот факт, что s является кандидатом в возможные причины отсутствия множества свойств p будем обозначать через $\mathrm{M}^{-}(s, p)$.

Условия $M^+(s,p)$ и $M^-(s,p)$ будем называть *положительным и отрицательным решающими предикатами индукции*, соответственно. Решающие предикаты позволяют дать краткую формулировку правилам индукции, которые мы будем интерпретировать как правила вычисления значений булевых функций C^+ и C^-. Заметим, что формулировка правил в данной статье будет существенно проще традиционной формулировки правил ДСМ-метода. Это связано с тем обстоятельством, что в этой статье используется расширенный набор базовых функций (их четыре: P^+, P^-, C^+ и C^-), а не две, как в традиционном ДСМ-методе.

Оператор присваивания будем обозначать через «$:=$». Это достаточно традиционное обозначение для оператора присваивания, оно должно восприниматься однозначно, кроме того это обозначение (хотя и немного) отличается от знака равенства.

Итак, мы имеем два правила индукции:

$$\frac{C^+(s,p)=0, M^+(s,p)}{C^+(s,p):=1} \qquad (I+)$$

$$\frac{C^-(s,p)=0, M^-(s,p)}{C^-(s,p):=1} \qquad (I–)$$

Этим правилам будут соответствовать условные операторы:

(I+) **if** $C^+(s,p)=0$ **and** $M^+(s,p)$ **then** $C^+(s,p):=1$;

(I–) **if** $C^-(s,p)=0$ **and** $M^-(s,p)$ **then** $C^-(s,p):=1$.

Подчеркнем, что эти правила и условные операторы имеют отношение к математической модели, а не к конкретной программной реализации.

Фрагменты (элементы множества $\mathbf{F}$) заранее неизвестны, они порождаются в процессе работы ДСМ-системы, и мы полагаем, что, по умолчанию, значение функций

C^+ и C^- для любого нового (только что построенного) фрагмента равно 0. Так что условия $C^+(s,p)=0$ и $C^-(s,p)=0$ можно не проверять, они всегда будут выполнены после того, как был построен фрагмент s.

Проверка условия $M^+(s,p)$ сводится к проверке условий (i)–(iii) из определения кандидата в возможные причины наличия набора свойств p. Самое сложное из них — условие (i) не нужно проверять, так алгоритмы порождения фрагментов, используемые в ДСМ-методе, строят только такие фрагменты, которые заведомо удовлетворяют условию (i). Остается только проверка сравнительно простых условий (ii) и (iii). То же самое можно сказать о проверке условия $M^-(s,p)$.

Из правил (I+) и (I–) можно получить следствия, связывающие решающие предикаты M^+ и M^- с принадлежностью фрагмента одному из множеств гипотез: $\mathbf{F}^{(+)}(p)$, $\mathbf{F}^{(-)}(p)$, $\mathbf{F}^0(p)$ и $\mathbf{F}^\tau(p)$. Эти следствия можно интерпретировать как правила «перетаскивания» гипотезы из множества неопределенных гипотез $\mathbf{F}^\tau(p)$ в одно из перечисленных выше множеств гипотез. В частности, гипотеза может остаться в множестве $\mathbf{F}^\tau(p)$. Приведем список этих правил.

(Ind +) $$\frac{s\in\mathbf{F}^\tau(p), M^+(s,p), \neg M^-(s,p)}{s\in\mathbf{F}^{(+)}(p)}$$

(Ind –) $$\frac{s\in\mathbf{F}^\tau(p), M^-(s,p), \neg M^+(s,p)}{s\in\mathbf{F}^{(-)}(p)}$$

(Ind 0) $$\frac{s\in\mathbf{F}^\tau(p), M^+(s,p), M^-(s,p)}{s\in\mathbf{F}^0(p)}$$

$$\text{(Ind } \tau\text{)} \qquad \frac{s \in \mathbf{F}^{\tau}(p), \neg \mathrm{M}^{+}(s,p), \neg \mathrm{M}^{-}(s,p)}{s \in \mathbf{F}^{\tau}(p)}$$

Те читатели, которые знакомы с ДСМ-методом, легко поймут, что приведенные выше правила «перетаскивания» суть сформулированные несколько иначе классические *правила простого ДСМ-метода без запрета на контрпример*.

Автор не ставил себе задачу описать на теоретико-множественном языке все используемые в настоящее время разновидности ДСМ-метода. Задача была более скромная: показать, что теоретико-множественное описание основных понятий ДСМ-метода, не использующее аппарат неклассических логик, в принципе возможно. Поэтому никакие другие версии ДСМ-метода в этой статье рассматриваться не будут.

8. Решающие предикаты. Правила аналогии

Правила аналогии позволяют делать предсказания о наличии или отсутствии у объектов набора целевых свойств. Общая идея этих правил такова:

- если объект содержит причину наличия набора свойств p, то он *должен* обладать этим набором свойств,
- если же объект содержит причину отсутствия набора свойств p, то он *не должен* обладать этим набором свойств.

Разумеется, противоречия в этом случае возможны, но они корректно обрабатываются.

Дадим сначала некоторые определения. Будем говорить, что объект $o \in \mathbf{O}$ является *кандидатом на наличие набора свойств* $p \in \mathbf{P}$, если существует $s \in \mathbf{F}^{(+)}(p)$ такое, что $s \subseteq o$. Аналогично, будем говорить, что объект $o \in \mathbf{O}$ является *кандидатом на отсутствие набора свойств* $p \in \mathbf{P}$, если существует $s \in \mathbf{F}^{(-)}(p)$ такое, что $s \subseteq o$.

Тот факт, что o является кандидатом на *наличие* набора свойств p, будем обозначать через $\Pi^+(o,p)$. Тот факт, что o является кандидатом на *отсутствие* набора свойств p, будем обозначать через $\Pi^-(o,p)$.

Условия $\Pi^+(o,p)$ и $\Pi^-(o,p)$ будем называть *положительным и отрицательным решающим предикатом аналогии, соответственно*. Решающие предикаты позволяют дать краткую формулировку правилам аналогии, которые мы будем интерпретировать как правила вычисления значений булевых функций P^+ и P^-. Использование решающих предикатов делает формулировку правил аналогии очень похожей на формулировку правил индукции. Приведем эти правила.

$$\frac{\mathrm{P}^+(o,p)=0,\Pi^+(o,p)}{\mathrm{P}^+(o,p):=1} \qquad (\mathrm{A}+)$$

$$\frac{\mathrm{P}^-(o,p)=0,\Pi^-(o,p)}{\mathrm{P}^-(o,p):=1} \qquad (\mathrm{A}-)$$

Этим правилам будут соответствовать условные операторы:

(A+) **if** $\mathrm{P}^+(o,p)=0$ **and** $\Pi^+(o,p)$ **then** $\mathrm{P}^+(o,p):=1$;

(A–) **if** $\mathrm{P}^-(o,p)=0$ **and** $\Pi^-(o,p)$ **then** $\mathrm{P}^-(o,p):=1$.

Как и в случае правил индукции, можно получить следствия из правил (A+) и (A–), которые можно интерпретировать как правила «перетаскивания» объекта из множества неопределенных примеров $\mathbf{O}^{\tau}(p)$ в одно из множеств: $\mathbf{O}^{(+)}(p)$, $\mathbf{O}^{(-)}(p)$, $\mathbf{O}^{0}(p)$ или сохранения его в множестве $\mathbf{O}^{\tau}(p)$. Приведем список этих правил.

$$(\mathrm{An}+) \qquad \frac{o\in\mathbf{O}^{\tau}(p),\Pi^+(o,p),\neg\Pi^-(o,p)}{s\in\mathbf{O}^{(+)}(p)}$$

(An –) $$\frac{o \in \mathbf{O}^{\tau}(p), \Pi^{-}(o,p), \neg\Pi^{+}(o,p)}{s \in \mathbf{O}^{(-)}(p)}$$

(An 0) $$\frac{o \in \mathbf{O}^{\tau}(p), \Pi^{+}(o,p), \Pi^{-}(o,p)}{o \in \mathbf{O}^{0}(p)}$$

(An τ) $$\frac{o \in \mathbf{O}^{\tau}(p), \neg\Pi^{+}(o,p), \neg\Pi^{-}(o,p)}{s \in \mathbf{O}^{\tau}(p)}$$

9. Работа ДСМ-системы. Итерация применения правил

Ядро ДСМ-метода составляют две фазы:

- **Фаза индукции.** С помощью правил индукции формируются гипотезы о возможных причинах набора целевых свойств. С точки зрения математической модели, фаза индукции состоит в вычислении значений функций C^{+} и C^{-}. Но не нужно забывать, что в этой же фазе один из аргументов этих функций — фрагмент s — только формируется. Алгоритм построения множества фрагментов, используемых в качестве аргументов функций C^{+} и C^{-}, вносит наибольший вклад в вычислительную сложность процедур ДСМ-метода. Построенные фрагменты образуют множества положительных и отрицательных гипотез (возможных причин наличия и возможных причин отсутствия набора целевых свойств, соответственно). Напомним, что положительная гипотеза находится как сходство (пересечение) положительных примеров (объектов, обладающих набором целевых свойств), а отрицательная гипотеза находится как сходство (пересечение) отрицательных примеров (объектов, не обладающих набором целевых свойств).

- **Фаза аналогии.** С помощью правил аналогии формируются гипотезы о наличии или отсутствии набора целевых свойств. С точки зрения математической модели происходит вычисление значений функций P^+ и P^-. Никаких новых объектов при этом не строится но в соответствии с определениями из раздела 4 изменяются множества примеров. Множество неопределенных примеров сужается. Множества положительных и отрицательных примеров расширяются.

После формирования гипотез о наличии или отсутствии набора целевых свойств Этап I ДСМ-метода (в смысле статьи [9]) может быть закончен. Это происходит в случае так называемого *одношагового* ДСМ-метода. Но, поскольку у нас изменилось соотношение между положительными, отрицательными и неопределенными примерами, мы можем попробовать еще раз применить процедуры индукции и аналогии. В этом случае ядро представляет собой циклическую процедуру, а соответствующий вариант ДСМ-метода называется *итеративным*.

Выход из цикла осуществляется при выполнении *условия насыщения*, когда множества примеров перестают изменяться. Поскольку множество неопределенных примеров может только уменьшаться, а множества положительных и отрицательных примеров — только расти, и все эти множества — конечны, рано или поздно условие насыщения должно стать выполненным.

В случае итеративного ДСМ-метода завершение Этапа I происходит после выполнения условия насыщения и наступает Этап II — применение абдукции (проверка условия каузальной полноты). В этой статье Этап II не рассматривается.

Среди исследователей, занимающихся ДСМ-методом нет единства в вопросе о необходимости итераций в ядре ДСМ-метода. Существуют как сторонники, так и убежденные противники итераций.

Противники итераций утверждают, что надежность новых примеров, полученных с помощью правил аналогии, достаточно низка; повторяя применение правил, мы с каждым шагом уменьшаем правдоподобие получаемых гипотез, и, в конце концов, получаем гипотезы, не имеющие никакого отношения к исследуемому набору целевых свойств. Для обоснования этой позиции иногда приводится (искусственный) пример из работы [29] (см. также [5, гл. 5]), в которой доказывается существование сколь угодно длинных последовательностей непересекающихся гипотез.

Позиция автора данной статьи по этому вопросу такова:

- математического доказательства того факта, что итерация *всегда* бесполезна, — нет;
- случаи, когда итерация оказывается бесполезной, — возможны;
- для экспериментального доказательства бесполезности (как и полезности) итерации — недостаточно данных;
- можно привести (искусственный) пример, демонстрирующий возможность получения с помощью итерации процедур индукции и аналогии новых содержательных гипотез.

Завершим этот раздел описанием упомянутого выше примера. Идея этого примера такова:

- в качестве объектов берем множества ключевых слов,
- единственное целевое свойство — иметь отношение к оружию;
- в качестве отрицательных примеров берутся множества ключевых слов, имеющих отношение к сельскохозяйственным орудиям.

Порождение гипотез и примеров демонстрирует эволюцию вооружений и сельскохозяйственной техники. Работа процедур ДСМ-метода представлена в таб. 1.

Таб. 1. Демонстрация возможностей итерации процедур индукции и аналогии

Номер шага	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Фаза	С	И	А	И	А	И	А	И	А	И	А	И	А	И	А
Объект															
{лук, меч},	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
{лук, копье},	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
{меч, копье},	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
{лук, арбалет},			+	+	+	+	+	+	+	+	+	+	+	+	+
{меч, арбалет},			+	+	+	+	+	+	+	+	+	+	+	+	+
{арбалет, шпага},					+	+	+	+	+	+	+	+	+	+	+
{меч, шпага},			+	+	+	+	+	+	+	+	+	+	+	+	+
{шпага, мушкет},							+	+	+	+	+	+	+	+	+
{шпага, пистолет},							+	+	+	+	+	+	+	+	+
{арбалет, мушкет},					+	+	+	+	+	+	+	+	+	+	+
{арбалет, пистолет},					+	+	+	+	+	+	+	+	+	+	+
{мушкет, ружье},									+	+	+	+	+	+	+
{пистолет, ружье},									+	+	+	+	+	+	+
{ружье, винтовка},											+	+	+	+	+
{пистолет, винтовка},									+	+	+	+	+	+	+
{винтовка, пулемет},													+	+	+
{пистолет, пулемет},									+	+	+	+	+	+	+
{пистолет, револьвер},									+	+	+	+	+	+	+
{винтовка, револьвер},													+	+	+
{мотыга, серп},	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
{серп, соха},	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
{соха, мотыга},	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
{серп, коса},			–	–	–	–	–	–	–	–	–	–	–	–	–
{мотыга, коса},			–	–	–	–	–	–	–	–	–	–	–	–	–
{мотыга, плуг},			–	–	–	–	–	–	–	–	–	–	–	–	–
{соха, плуг},			–	–	–	–	–	–	–	–	–	–	–	–	–
{коса, косилка},					–	–	–	–	–	–	–	–	–	–	–
{плуг, косилка},					–	–	–	–	–	–	–	–	–	–	–
{плуг, трактор},					–	–	–	–	–	–	–	–	–	–	–
{косилка, трактор}.							–	–	–	–	–	–	–	–	–
Фрагмент															
{лук},		+	+	+	+	+	+	+	+	+	+	+	+	+	+
{копье},		+	+	+	+	+	+	+	+	+	+	+	+	+	+
{меч},		+	+	+	+	+	+	+	+	+	+	+	+	+	+
{арбалет},				+	+	+	+	+	+	+	+	+	+	+	+
{шпага},						+	+	+	+	+	+	+	+	+	+
{мушкет},								+	+	+	+	+	+	+	+
{пистолет},								+	+	+	+	+	+	+	+
{ружье},										+	+	+	+	+	+
{винтовка},												+	+	+	+
{пулемет},														+	+
{револьвер},														+	+
{мотыга},		–	–	–	–	–	–	–	–	–	–	–	–	–	–
{серп},		–	–	–	–	–	–	–	–	–	–	–	–	–	–
{соха},		–	–	–	–	–	–	–	–	–	–	–	–	–	–
{коса},				–	–	–	–	–	–	–	–	–	–	–	–
{плуг}				–	–	–	–	–	–	–	–	–	–	–	–
{косилка},						–	–	–	–	–	–	–	–	–	–
{трактор},								–	–	–	–	–	–	–	–

Условные обозначения:

- И — индукция, А — аналогия, С — «старт» (начальное состояние);
- «+» — положительный(ая) пример (гипотеза);
- «–» — отрицательный(ая) пример (гипотеза);
- пустая клетка — неопределенный(ая) пример (гипотеза).

В этом примере неопределенными являются те гипотезы, которые еще не построены, т.е. не получено соответствующее пересечение.

10. ДСМ-метод и анализ формальных понятий

Основополагающей работой по анализу формальных понятий (formal concept analysis) считается работа Р.Вилле [30]. Наиболее основательное изложение результатов этого направления исследований содержится в книге [31]. ДСМ-метод и анализ формальных понятий появились примерно в одно время и около десяти лет развивались независимо друг от друга.

Анализ формальных понятий позиционировался как ветвь прикладной алгебры (более конкретно — теории решеток). Исследователей в этой области интересовали, прежде всего, алгебраические и алгоритмические вопросы. Проблемы практического применения теоретических результатов не были для данного направления центральными, хотя иногда основоположники анализа формальных понятий занимались и прикладными проблемами (см., например, [32]).

ДСМ-метод позиционировался, прежде всего, как система технологий интеллектуального анализа данных, предназначенная для решения практических задач в разных предметных областях: фармакологии, медицине, социологии и т.п. В ранних работах по ДСМ-методу можно выделить два направления:

- исследования неклассических логик, являющихся фундаментом для описания процедур ДСМ-метода (наиболее типичной работой такого рода является статья [33]),
- прикладные исследования, посвященные применению ДСМ-метода в различных предметных областях (например, работы [34], [35] и [36]).

Работ, посвященных проектированию эффективных алгоритмов, в первое десятилетие ДСМ-метода было сравнительно немного. Среди немногих исключений следует отметить, например, работу [37]. В это же время в рамках анализа формальных понятий велась достаточно интенсивная работа по описанию разнообразных алгоритмов порождения всех формальных понятий для заданного формального контекста.

Связь между ДСМ-методом и анализом формальных понятий обнаружил С.О.Кузнецов — автор одного из наиболее эффективных алгоритмов ДСМ-метода — алгоритма «Замыкай-по-Одному» [38]. После обнаружения этой связи в ДСМ-системах стали широко использоваться алгоритмы, разработанные для порождения формальных понятий (подробный анализ таких алгоритмов имеется в работе С.О.Кузнецова и С.А.Объедкова [39]). Особенно популярным стал алгоритм Норриса [40], который в настоящее время является стандартом де факто для ДСМ-систем.

Прежде чем объяснять связь между ДСМ-методом и анализом формальных понятий (АФП) введем основные термины, относящиеся к АФП.

Формальным контекстом будем называть упорядоченную тройку $\langle G, M, I \rangle$, где G — непустое *множество объектов*, M — непустое *множество признаков*, $I \subseteq G \times M$, т.е. I задает некоторое соответствие между элементами множеств G и M.

Тот факт, что $\langle g, m \rangle \in I$ будем, как обычно, обозначать через $g\, I\, m$. Если $g\, I\, m$, то будем говорить, что *объект* g *обладает признаком* m.

Пусть A — множество объектов, т.е. $A \subseteq G$. *Содержанием* (*интенсионалом*) множества объектов A будем называть множество признаков, которыми обладают все объекты из A. Интенсионал множества A будем обозначать через A^{In}. По определению,

$$A^{\mathrm{In}} = \{m \in M \mid (\forall g \in A)(g\, I\, m)\}. \tag{7}$$

Пусть B — множество признаков, т.е. $B \subseteq M$. *Объемом* (*экстенсионалом*) множества признаков B будем называть множество объектов, которые обладают всеми признаками из B. Экстенсионал множества B будем обозначать через B^{Ex}. По определению,

$$B^{\mathrm{Ex}} = \{g \in G \mid (\forall m \in B)(g\, I\, m)\}. \tag{8}$$

Введя понятия интенсионала и экстенсионала мы, фактически, определили два отображения: $^{\mathrm{In}}: 2^G \to 2^M$ и $^{\mathrm{Ex}}: 2^M \to 2^G$. Можно показать, что пара отображений $\left\langle {}^{\mathrm{In}}, {}^{\mathrm{Ex}} \right\rangle$ является соответствием Галуа (между частично упорядоченными множествами $\langle 2^G, \subseteq \rangle$ и $\langle 2^M, \subseteq \rangle$). Т.е., мы получили ситуацию, достаточно похожую на рассмотренную в разделе 5, где были определены отображения $^{\mathrm{Si}}: 2^{\mathbf{O}} \to 2^{\mathbf{A}}$ и $^{\mathrm{Cl}}: 2^{\mathbf{A}} \to 2^{\mathbf{O}}$, такие, что пара $\left\langle {}^{\mathrm{Si}}, {}^{\mathrm{Cl}} \right\rangle$ является соответствием Галуа (между частично упорядоченными множествами $\langle 2^{\mathbf{O}}, \subseteq \rangle$ и $\langle 2^{\mathbf{A}}, \subseteq \rangle$).

Для простоты будем отождествлять объект g с множеством признаков, которыми этот объект обладает. Тогда вместо $g\, I\, m$ (g обладает признаком m) мы можем писать $m \in g$.

Интенсионал множества объектов A в этом случае будет состоять из всех признаков, общих для объектов из A (т.е., будет пересечением объектов из A). Тогда равенство (7) может быть переписано следующим образом:

$$A^{\mathrm{In}} = \bigcap_{g \in A} g = \{ m \in M \mid (\forall g \in A)(m \in g) \}. \qquad (9)$$

Экстенсионал множества признаков B будет состоять из всех объектов, содержащих все признаки из B. В этом случае равенство (8) может быть переписано следующим образом:

$$B^{\mathrm{Ex}} = \{ g \in G \mid B \subseteq g \} = \{ g \in G \mid (\forall m \in B)(m \in g) \}. \qquad (10)$$

Теперь сходство между парами равенств (1), (2) и (9), (10) стало совершенно очевидным.

Итак, непосредственной проверкой можно убедиться в том, что упорядоченная тройка $\langle \mathbf{O}, \mathbf{A}, э \rangle$, где $\mathbf{O}$ — множество объектов ДСМ-метода, $\mathbf{A}$ — множество атомов (каждый объект является подмножеством множества атомов), э — отношение «содержит» (o э a тогда и только тогда, когда $a \in o$), является формальным контекстом, для которого интенсионал совпадает со сходством семейства объектов ($O^{\mathrm{In}} = O^{\mathrm{Si}}$), а экстенсионал — с кластером фрагмента ($s^{\mathrm{Ex}} = s^{\mathrm{Cl}}$).

Подчеркнем, что при погружении ДСМ-метода в анализ формальных понятий роль признаков будут играть атомы, из которых состоят объекты.

Введем теперь еще один термин из анализа формальных понятий.

Пусть $\langle G, M, I \rangle$ — формальный контекст, $A \subseteq G$, $B \subseteq M$. Упорядоченную пару $\langle A, B \rangle$ назовем формальным понятием, если верны следующие равенства:

$$A = B^{\mathrm{Ex}}, \tag{11}$$

$$B = A^{\mathrm{In}}. \tag{12}$$

Множество A будем называть объемом (экстенсионалом) формального понятия $\langle A, B \rangle$, а множество B — его содержанием (интенсионалом).

Для рассмотренного выше формального контекста $\langle \mathbf{O}, \mathbf{A}, э \rangle$ равенства (11) и (12) будут равносильны равенствам (4) и (5), соответственно. В этом случае, пара $\langle O, s \rangle$ (где $O \subseteq \mathbf{O}$, $s \subseteq \mathbf{A}$) является формальным понятием тогда и только тогда, когда она удовлетворяет условию исчерпываемости относительно множества $\mathbf{O}$ (см. раздел 5).

Процедуры ДСМ-метода должны находить все пары, удовлетворяющие условию исчерпываемости. Для этого достаточно найти (породить) все формальные понятия в некотором формальном контексте. Как уже было сказано выше, в таком направлении исследований как анализ формальных понятий было разработано немало алгоритмов, эффективно решающих эту задачу.

11. Подготовка данных

Данные для ДСМ-системы могут быть разными. Их внешнее представление может существенно отличаться от внутреннего, с которым работают процедуры ДСМ-метода. Например, в анализе данных по фармакологии объектами считаются химические соединения. Исходные данные в этом случае содержат представления структурных формул (или трехмерной структуры молекул) химических соединений в каком-либо стандартном формате.

Для работы ДСМ-системы необходимо представить соединения в виде множеств. Для этого используется особая система кодирования, разработанная специали-

стами в предметной области — так называемый фрагментарный код суперпозиций подструктур (ФКСП) [41]. Задача автоматизации перевода внешнего представления во внутреннее для этого случая весьма нетривиальна. Решение этой задачи включает разработку оригинальных алгоритмов обработки химических графов и создание эффективной компьютерной программы, выполняющей такой перевод. Решению этой задачи была посвящена диссертация Д.А.Добрынина [42].

Существуют и другие предметные области, для которых формирование внутреннего представления данных для ДСМ-системы является нетривиальной задачей и требует участия эксперта. Однако в большинстве случаев данные для анализа записаны в прямоугольную таблицу, в которой каждому объекту соответствует единственная строка таблицы. Столбцы таблицы содержат значения параметров объекта, и существует единственный выделенный столбец, соответствующий целевому свойству. Для определенности, будем считать, что в этом столбце могут содержаться только три возможных значения: «1» (объект обладает свойством), «0» (объект не обладает свойством) и «?» или пустая клетка (информация о наличии свойства у объекта отсутствует).

В случае такого табличного представления совокупности объектов часто говорят, что каждый объект представлен в виде вектора. Это не совсем правильно, так как термин «вектор» принято употреблять по отношению к структуре, все элементы которой имеют один и тот же тип. Однако на практике в разных столбцах могут содержаться значения разных типов (целые, вещественные, строковые и т.п.). Т.е., в общем случае объект представлен в виде записи (кортежа) или строки таблицы реляционной базы данных. Мы все-таки будем использовать по отношению к такой смешанной структуре термин «вектор», подразумевая, что наш вектор является гибридным или обобщенным.

Предположим, что таблица, представляющая данные, содержит n столбцов. Один столбец (обычно первый) содержит идентификаторы объектов, и еще один стол-

бец (обычно последний) содержит значения целевого свойства. Оставшиеся $n-2$ столбца образуют *представления объектов*. У таблицы имеется *заголовочная строка*, содержащая имена (идентификаторы) атрибутов, другими словами — *заголовки столбцов*. Разумеется, у разных столбцов должны быть разные заголовки.

Наша задача — заменить векторы, представляющие объекты, на подмножества некоторого универсума, сохранив (насколько это возможно) смысл информации, содержавшейся в исходном векторе, и не допустив чрезмерного увеличения размеров универсума.

Прежде всего, нам необходимо разбить атрибуты на два класса:

- индивидуальные (не требующие группировки) — атрибуты, для которых исходная таблица содержит *мало* различных значений,
- групповые (требующие обязательной группировки) — атрибуты, для которых исходная таблица содержит *много* различных значений.

Типичным примером индивидуального атрибута является целевое свойство. Соответствующий столбец может содержать всего три возможных значения. Типичный пример группового атрибута — идентификатор объекта. Все значения в соответствующем столбце разные. Но, ни целевое свойство, ни идентификатор в представление объекта не входят.

Пороговые значения, отличающие «малое количество возможных значений» от «большого количества возможных значений», зависят от конкретной задачи. В любом случае, «мало» означает «намного меньше количества строк в таблице», а много — «равно или не намного меньше количества строк в таблице».

Чаще всего, групповые атрибуты соответствуют данным, представляющим результаты измерений, например: рост, масса тела, температура, возраст и т.п. Обычно такие атрибуты имеют вещественный тип. Для группировки результатов измерений

наиболее естественно использовать интервалы значений. О том, на сколько интервалов следует разбить множество возможных значений атрибута, и какие нужно установить пороговые значения, лучше спросить у эксперта предметной области.

Однако может получиться и так, что групповой атрибут является *номинальным*, т.е. для него неизвестно естественное отношение линейного порядка. Рассмотрим, например, атрибут «цвет». Предположим, что в столбце «цвет» содержится 256 вариантов различных значений. Это довольно много. Кроме того, предположим, что (в контексте рассматриваемой задачи) на множестве значений параметра «цвет» нельзя определить естественное отношение порядка. Тогда мы должны выбрать какой-либо осмысленный способ классификации, не использующий интервалы. Например, поместить в один класс «оттенки серого», в другой — «оттенки красного» и т.д.

Если же никакого естественного способа классификации для группового атрибута найти не удается, то такой атрибут следует исключить из представления объекта, так как все равно никакая закономерность, в которой было бы существенным участие этого атрибута, обнаружена не будет.

Для единообразия будем предполагать, что множество возможных значений любого (группового или индивидуального) атрибута разбивается на конечное число классов. Каждый класс индивидуального атрибута содержит в точности один элемент.

Через $\mathrm{Partition}(A)$ обозначим множество (имен) классов, на которые разбито множество возможных значений атрибута A. Через R обозначим множество атрибутов, участвующих в векторном представлении объектов. Через U обозначим универсум для представления объектов в виде множеств. Положим по определению,

$$U = \{\langle A, C\rangle \mid A \in R, C \in \mathrm{Partition}(A)\}.$$

Пусть v — векторное представление объекта. Через $v(A)$ будем обозначать значение атрибута A в строке v. Через $\mathrm{Set}(v)$ будем обозначать представление объекта в виде множества, соответствующее векторному представлению v. Положим по определению,

$$\mathrm{Set}(v) = \{\langle A, C\rangle \mid A \in R, C \in \mathrm{Partition}(A), v(A) \in C\}.$$

Предположим, что наши данные записаны в таб. 2.

Таб. 2. Исходные данные (векторное представление)

Идентификатор	Глаза	Нос	Губы	Целевое свойство
Маша	Большие	Средний	Средние	1
Даша	Средние	Средний	Пухлые	1
Саша	Маленькие	Длинный	Тонкие	0

Для простоты будем считать, что все атрибуты индивидуальны. Тогда для универсума получим следующее представление.

$$\begin{aligned} U = \{&\langle \text{Глаза, Большие}\rangle, \langle \text{Глаза, Средние}\rangle, \langle \text{Глаза, Маленькие}\rangle, \\ &\langle \text{Нос, Средний}\rangle, \langle \text{Нос, Длинный}\rangle, \\ &\langle \text{Губы, Средние}\rangle, \langle \text{Губы, Пухлые}\rangle, \langle \text{Губы, Тонкие}\rangle\}. \end{aligned}$$

В качестве имени векторного представления будем использовать идентификатор строки. Тогда объекты будут представлены в виде множеств следующим образом:

$$\mathrm{Set}(\text{Маша}) = \{\langle \text{Глаза, Большие}\rangle, \langle \text{Нос, Средний}\rangle, \langle \text{Губы, Средние}\rangle\},$$

$$\mathrm{Set}(\text{Даша}) = \{\langle \text{Глаза, Средние}\rangle, \langle \text{Нос, Средний}\rangle, \langle \text{Губы, Пухлые}\rangle\},$$

$$\mathrm{Set}(\text{Саша}) = \{\langle \text{Глаза, Маленькие}\rangle, \langle \text{Нос, Длинный}\rangle, \langle \text{Губы, Тонкие}\rangle\}.$$

Теперь обратим внимание на то, что в ДСМ-методе причина *отсутствия* набора целевых свойств всегда описывается через *наличие* каких-либо атомов во фрагменте. Атомы могут интерпретироваться как признаки или компоненты структуры. Но ведь вполне возможна такая ситуация, когда причиной отсутствия набора свойств является не наличие каких-либо компонентов, а их отсутствие. Например, *отсутствие* качественного профессионального образования является возможной причиной *отсутствия* престижной работы.

Более того, отсутствие каких-либо компонентов может быть причиной наличия целевых свойств. Например, *отсутствие* в организме витамина B1 является причиной *наличия* у пациента болезни бери-бери.

В связи с этими обстоятельствами, следует отнестись к подготовке данных более внимательно и попытаться представить «отсутствие» через «наличие». Для этого мы должны в два раза увеличить наш универсум, добавив «отрицательные» классы. Для нашего примера это означает, что универсум должен включать не только пары

$$\langle \text{Глаза, Большие}\rangle,\ \langle \text{Глаза, Средние}\rangle,\ \langle \text{Глаза, Маленькие}\rangle,$$

но и пары

$$\langle \text{Глаза, не Большие}\rangle,\ \langle \text{Глаза, не Средние}\rangle,\ \langle \text{Глаза, не Маленькие}\rangle.$$

Теперь опишем эту ситуацию более формально. Через $\mathrm{Dom}(A)$ будем обозначать *домен* атрибута A, т.е. множество его возможных значений. Для любого класса $C \in \mathrm{Partition}(A)$ положим

$$\overline{C} = \mathrm{Dom}(A) \setminus C.$$

Через $\mathrm{Cover}(A)$ будем обозначать *расширение разбиения* множества значений атрибута A на классы или *покрытие* домена атрибута A. Положим по определению,

$$\mathrm{Cover}(A) = \mathrm{Partition}(A) \cup \left\{ \overline{C} \mid C \in \mathrm{Partition}(A) \right\}.$$

Определим *расширенный универсум* EU следующим образом:

$$EU = \left\{ \langle A, C \rangle \mid A \in R, C \in \mathrm{Cover}(A) \right\}.$$

Для каждого векторного представления v *определим расширенное представление в виде множества* следующим образом:

$$\mathrm{ESet}(v) = \left\{ \langle A, C \rangle \mid A \in R, C \in \mathrm{Cover}(A), v(A) \in C \right\}.$$

Теперь посмотрим, какую интерпретацию получат эти определения в контексте нашего примера. Сначала договоримся вместо $\overline{C}$ писать просто «не C». Для расширенного универсума получим следующее представление:

$$\begin{aligned} EU = \{ & \langle \text{Глаза, Большие} \rangle, \langle \text{Глаза, Средние} \rangle, \langle \text{Глаза, Маленькие} \rangle, \\ & \langle \text{Нос, Средний} \rangle, \langle \text{Нос, Длинный} \rangle, \\ & \langle \text{Губы, Средние} \rangle, \langle \text{Губы, Пухлые} \rangle, \langle \text{Губы, Тонкие} \rangle, \\ & \langle \text{Глаза, не Большие} \rangle, \langle \text{Глаза, не Средние} \rangle, \langle \text{Глаза, не Маленькие} \rangle, \\ & \langle \text{Нос, не Средний} \rangle, \langle \text{Нос, не Длинный} \rangle, \\ & \langle \text{Губы, не Средние} \rangle, \langle \text{Губы, не Пухлые} \rangle, \langle \text{Губы, не Тонкие} \rangle \}. \end{aligned}$$

Теперь рассмотрим расширенные представления объектов в виде множеств:

$$\mathrm{ESet}(\text{Маша}) = \{\langle\text{Глаза, Большие}\rangle, \langle\text{Нос, Средний}\rangle, \langle\text{Губы, Средние}\rangle,$$
$$\langle\text{Глаза, не Средние}\rangle, \langle\text{Глаза, не Маленькие}\rangle, \langle\text{Нос, не Длинный}\rangle,$$
$$\langle\text{Губы, не Пухлые}\rangle, \langle\text{Губы, не Тонкие}\rangle\},$$

$$\mathrm{ESet}(\text{Даша}) = \{\langle\text{Глаза, Средние}\rangle, \langle\text{Нос, Средний}\rangle, \langle\text{Губы, Пухлые}\rangle,$$
$$\langle\text{Глаза, не Большие}\rangle, \langle\text{Глаза, не Маленькие}\rangle, \langle\text{Нос, не Длинный}\rangle,$$
$$\langle\text{Губы, не Средние}\rangle, \langle\text{Губы, не Тонкие}\rangle\},$$

$$\mathrm{ESet}(\text{Саша}) = \{\langle\text{Глаза, Маленькие}\rangle, \langle\text{Нос, Длинный}\rangle, \langle\text{Губы, Тонкие}\rangle,$$
$$\langle\text{Глаза, не Большие}\rangle, \langle\text{Глаза, не Средние}\rangle, \langle\text{Нос, не Средний}\rangle,$$
$$\langle\text{Губы, не Средние}\rangle, \langle\text{Губы, не Пухлые}\rangle\}.$$

Несмотря на кажущуюся избыточность, расширенные представления в виде множеств несут много полезной информации. Из таб. 2 видно, что объекты «Маша» и «Даша» являются положительными примерами для целевого свойства. Если взять «узкое» представление этих объектов в виде множеств, т.е., $\mathrm{Set}(\text{Маша})$ и $\mathrm{Set}(\text{Даша})$, то их сходство (пересечение) содержит лишь один элемент — $\langle\text{Нос, Средний}\rangle$. Это довольно бедное содержание, чтобы его всерьез рассматривать в качестве кандидата в возможные причины наличия целевого свойства.

Однако, если мы возьмем пересечение множеств $\mathrm{ESet}(\text{Маша})$ и $\mathrm{ESet}(\text{Даша})$, то оно, в частности, содержит такие элементы как

$$\langle\text{Глаза, не Маленькие}\rangle,\ \langle\text{Нос, не Длинный}\rangle,\ \langle\text{Губы, не Тонкие}\rangle,$$

что выглядит гораздо более содержательным. Т.е. в этом случае мы действительно смогли обнаружить кандидата в возможные причины целевого свойства.

В заключение этого раздела подчеркнем, что представление объектов в виде множеств — это всего лишь абстракция. Чтобы использовать его в программе, необхо-

димо выбрать одно из возможных машинных представлений множеств. Заметим, что в ДСМ-системах множества обычно представляют в виде битовых строк.

Заключение

В работе был предложен нестандартный подход к описанию конструкций ДСМ-метода. Этот подход был использован для определения решающих предикатов и правил ДСМ-метода. В работе также обсуждалась связь между ДСМ-методом и анализом формальных понятий, обосновывалась возможность использования процедурами ДСМ-метода алгоритмов порождения формальных понятий.

Разумеется, эта статья не охватывает всех (и даже всех основных) проблем, связанных с ДСМ-методом. Автор планирует написать продолжение этой статьи, в котором собирается рассмотреть следующие вопросы:

- обобщенный и несимметричный ДСМ-метод,
- алгоритмы ДСМ-метода,
- проектирование ДСМ-системы,
- нестандартные варианты ДСМ-метода.

Автор желает читателям успехов в самостоятельном изучении ДСМ-метода и применении его для решения прикладных задач.

Литература

1. *Финн В.К.* О возможностях формализации правдоподобных рассуждений средствами многозначных логик // Всесоюз. симпозиум по логике и методологии науки.— Киев: Наукова думка, 1976.— С. 82–83.

2. *Финн В.К.* Базы данных с неполной информацией и новый метод автоматического порождения гипотез // Диалоговые и фактографические системы информационного обеспечения.— М., 1981.— С. 153–156.
3. *Финн В.К.* О машинно-ориентированной формализации правдоподобных рассуждений в стиле Ф.Бэкона — Д.С.Милля // Семиотика и информатика. — 1983. — Вып. 20. — С. 35–101.
4. Автоматическое порождение гипотез в интеллектуальных системах / Сост. Е.С.Панкратова, В.К.Финн; Под. общ. ред. В.К.Финна. — М.: ЛИБРОКОМ, 2009. — 528 с.
5. ДСМ-метод автоматического порождения гипотез: Логические и эпистемологические основания / Сост. О.М.Аншаков, Е.Ф.Фабрикантова; под. общ. ред. О.М.Аншакова. — М.: ЛИБРОКОМ, 2009. — 433 с.
6. *Милль Д.С.* Система логики силлогистической и индуктивной. М.: ЛЕНАНД, 2011, 832 с.
7. *Поппер К.Р.* Логика и рост научного знания. М.: Прогресс, 1983, 605 с.
8. *Peirce C.S.* Abduction and induction. In: Buchler J (ed) Philosophical Writings of Peirce, 1995, Dover, NY, pp 150–156
9. *Финн В.К.* Индуктивные методы Д.С. Милля в системах искусственного интеллекта. Часть I // Искусственный интеллект и принятие решений. — 2010. — № 3. — С. 3–21.
10. *Финн В.К.* Индуктивные методы Д.С. Милля в системах искусственного интеллекта. Часть II // Искусственный интеллект и принятие решений. — 2010. — № 4. — С. 14–40.

11. *Финн В.К.* Об определении эмпирических закономерностей посредством ДСМ - метода автоматического порождения гипотез // Искусственный интеллект и принятие решений. — 2010. — № 4. — С. 41–48.
12. *Финн В.К.* Искусственный интеллект: Методология, применения, философия, М.: КРАСАНД, 2011. — 448 с.
13. *Волкова А.Ю.* Алгоритмизация процедур ДСМ-метода автоматического порождения гипотез // НТИ. Сер 2. — 2011. — № 5.— С. 6–12.
14. *Rosser J.B., Turquette A.R.* Many-valued logics. Amsterdam: North-Holland. 1951.
15. *Борщев В.Б.* О постулатах ДСМ-метода // Новости искусственного интеллекта. Спец. вып. К 60-летию В.К.Финна.— М.: 1993.— С. 16–26.
16. *Аншаков О.М.* Об одной интерпретации ДСМ-метода автоматического порождения гипотез // НТИ. Сер 2. — 1999. — № 1–2.— С. 45–53.
17. *Шундеев А.С.* Логико-языковые средства автоматизации производственных процессов: диссертация на соискание ученой степени кандидата физико-математических наук: 05.13.11: М., 2005, 168 с.
18. *Осипов Г.С.* Лекции по искусственному интеллекту. М.: УРСС, 2009, 272 с.
19. *Липкин А.А.* ДСМ-метод порождения гипотез для объектов, описываемых атрибутами с весами: диссертация на соискание ученой степени кандидата технических наук: 05.13.17: М.:, 2008, 117 с: ил.
20. *Гусакова С.М., Михеенкова М.А., Финн В.К.* О логических средствах автоматизированного анализа мнений // НТИ. Сер 2. — 2001. — № 5.— С. 4–24.
21. *Новиков Ф.А., Иванов Д.Ю.* Моделирование на UML. Теория, практика, видеокурс. — СПб.: Профессиональная литература, Наука и Техника, 2010. — 640 с.: ил.
22. *Шрейдер Ю.А.* Равенство, сходство, порядок. — М.: Наука, 1971. — 257 с.: ил.

23. Математическая Энциклопедия. Т. 1 (А–Г). Ред. коллегия: И.М.Виноградов (глав. ред.) [и др.] — М., «Советская Энциклопедия», 1977, 1152 с.: ил.

24. *Кузнецов С.О.* ДСМ-метод на языке соответствий Галуа // НТИ. Сер 2. — 2006. — № 12. — С. 1–7.

25. *Виноградов Д.В.* Логические программы для квазиаксиоматических теорий // НТИ. Сер 2. — № 1–2. — 1999. — С. 61–64

26. *Виноградов Д.В.* Корректные логические программы для правдоподобных рассуждений // НТИ. Сер. 2. — 2001. — № 5. — С. 25–28.

27. *Михеенкова М.А., Феофанова Т.Л.* Обучающая ДСМ-система для анализа социологических данных // Вестник Российского государственного гуманитарного университета. Серия «Информатика. Информационная безопасность. Математика», 2009 г., вып. 10, с. 152 –169.

28. *Виноградов Д.В.* Формализация правдоподобных рассуждений в логике предикатов первого порядка // НТИ. Сер 2. — 2000. — № 11. — С. 17–20.

29. *Аншаков О.М., Скворцов Д.П., Финн В.К.* О дедуктивной имитации некоторых вариантов ДСМ-метода // Семиотика и информатика.— 1993.— Вып. 33.— С. 164–233.

30. *Wille R.* Restructuring Lattice Theory: an Approach Based on Hierarchies of Concepts // Ordered Sets / Ed. by I. Rival. — Dordrecht; Boston: Reidel, 1982. — P. 445–470.

31. *Ganter B., Wille R.* Formal Concept Analysis: Mathematical Foundations, Berlin: Springer-Verlag, 1999.

32. *Ganter B., Wille R.* “Conceptual scaling,” in Applications of combinatorics and graph theory to the biological and social sciences / Ed. by F. Roberts. New York: Springer-Verlag, 1989, pp. 139–167.

33. *Anshakov O. M., Finn V. K., Skvortsov D. P.* On axiomatization of many-valued logics associated with formalization of plausible reasonings // Studia Logica. — 1989. — Vol. 48, N 4. — P. 423–447.

34. *Панкратова Е.С., Ивашко В.Г., Авидон В.В., Блинова В.Г., Бодягин Д.А.* Экспериментальная проверка новой версии ДСМ-метода автоматического порождения гипотез // НТИ. Сер. 2.— 1988.— № 2.— С. 18–21.

35. *Панкратова Е.С., Блинова В.Г., Финн В.К.* О возможности применения ДСМ-метода в задаче распознавания химического канцерогенеза // Экпертные системы: состояние и перспективы / Под ред. Д.А.Поспелова.— М.: Наука, 1989.— С. 131–138.

36. *Панкратова Е.С., Ивашко В.Г., Блинова В.Г., Попов Д.В.* Применение ДСМ-метода порождения гипотез для прогноза противоопухолевой активности и токсичности соединений, принадлежащих к различным классам химических соединений // Экпертные системы: состояние и перспективы / Под ред. Д.А.Поспелова.— М.: Наука, 1989.— С. 139–146.

37. *Забежайло М.И., Ивашко В.Г., Кузнецов С.О., Михеенкова М.А., Хазановский К.П., Аншаков О.М.* Алгоритмические и программные средства ДСМ-метода автоматического порождения гипотез // НТИ. Сер. 2. — 1987. — № 10. — С. 1–14.

38. *Кузнецов С.О.* Быстрый алгоритм построения всех пересечений объектов из конечной полурешетки // НТИ, Сер.2. — 1993. — № 1. — С. 17–20.

39. *Kuznetsov S.O., Obiedkov S.A.* Comparing Performance of Algorithms for Generating Concept Lattices // Journal of Experimental and Theoretical Artificial Intelligence. — 2002. — Vol. 14. — N. 2–3, pp. 189–216.

40. *Norris E.M.* An Algorithm for Computing the Maximal Rectangles in a Binary Relation // Revue Roumaine de Mathématiques Pures et Appliquées. — 1978. — N 23(2). — pp. 243–250.

41. *Avidon V.V., Pomerantsev I.A., Golender V.E., Rozenblit A.B.* Structure-activity relationship oriented languages for chemical structure representation // Journal of Chemical Information and Computer Sciences. — 1982. — Vol. 22. — N 4. — pp. 207–214.

42. *Добрынин Д.А.* Инструментальные средства для представления информации о структуре химических соединений и их сходстве в интеллектуальных системах: диссертация на соискание ученой степени кандидата технических наук: 05.25.05.- М.:, 2003. — 118 с.: ил.