

# НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ  
ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г.

№ 4

Москва 2023

## ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК [001:311:311]:002

Р.С. Гиляревский, А.Н. Либкинд, И.А. Либкинд

### Классифицирование библиографических ссылок для определения взаимовлияния крупных областей знания\*

*Обсуждается устройство списка тематических категорий Web of Science в связи с некорректностью вычислений по ним количества содержащихся в них профильных публикаций, а также возможный выход из ситуации с использованием библиографических ссылок. Установлено, что сопоставление результатов цитирования между крупными разделами науки подтверждает усиление взаимовлияния естественных, общественных и гуманитарных наук и междисциплинарность исследований.*

**Ключевые слова:** Web of Science, тематические категории, вычисление количества публикаций, классифицирование, библиографические ссылки, междисциплинарность исследований

DOI: 10.36535/0548-0027-2023-04-1

\* Исследование проводилось по государственному заданию ВИНТИ РАН по теме FFFU-0003-2022-0007.

## ВВЕДЕНИЕ

В процессе наших исследований была выявлена некорректность вычислений количества публикаций по классификациям журналов в центрах индексации и цитирования при сопоставлении областей знания из-за рассеивания статей и закона Брэдли-Форда. Осознав это, мы решили понять, как устроены эти классификации на примере платформы *Web of Science* (WoS), поскольку используем в своих исследованиях её данные [1]. Все журналы в этих центрах классифицируются по линейным схемам, в WoS такую классификацию называют списком категорий, которых в 2018 г. было 254. Каждый журнал получает свою тематическую категорию (43,8% журналов даже по несколько), и всем статьям каждого журнала приписываются его категории. Постатейная индексация не проводится из-за отсутствия надежных автоматических систем, а ручная – из-за нехватки ресурсов (21 тыс. названий журналов ежегодно публикует около 2,8 млн статей). По своим классификационным особенностям журналы отличаются от статей и книг непостоянством и разнообразием содержания и тематического профиля. Их трудно распределять по категориям, так как многие охватывают сразу несколько родственных категорий, которые пришлось из-за этого объединять в исследовательские области (*research areas*). Таких исследовательских областей – 153,

часть из них сохраняет свое название в названии категории (табл. 1), эти области входят в пять широких разделов (*big categories*, табл. 2), которые распределены по трем гарфилдовским указателям: по естественным и прикладным наукам (*Science Citation Index Expanded*, *SCIE*), общественным наукам (*Social Sciences Citation Index*, *SSCI*), искусству и гуманитарным наукам (*Arts & Humanities Citation Index*, *A&HCI*). Это несколько напоминает иерархическую классификацию, но ею никоим образом не является.

Между тем, большинство исследователей убеждено в том, что такая классификация хотя и не очень строгая, но все же иерархическая система, по которой индексируются журналы и статьи. Некоторые рассматривают разные уровни этой системы как отдельные классификации, используемые в зависимости от широты профиля классифицируемого журнала. При этом многие полагают, что в каждой категории сосредоточены если не все статьи соответствующей ей тематики, то их основная масса. Но это не так, поскольку из-за рассеивания статей по журналам и по закону Брэдли-Форда по крайней мере две трети статей находится в журналах других тематических категорий. И ничего с этим не поделаешь, поскольку социальные законы так же неумолимы, как и естественнонаучные.

Таблица 1

Фрагмент таблицы категорий, входящих в индекс цитирования *Science Citation Index Expanded* (SCIE) и широкий раздел *Technology*

| Исследовательская область (одна из 153) | Категории (семь из 254)                          |
|-----------------------------------------|--------------------------------------------------|
| Computer Science                        | COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS |
| Computer Science                        | COMPUTER SCIENCE, THEORY & METHODS               |
| Computer Science                        | COMPUTER SCIENCE, INFORMATION SYSTEMS            |
| Computer Science                        | COMPUTER SCIENCE, SOFTWARE ENGINEERING           |
| Computer Science                        | COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE        |
| Computer Science                        | COMPUTER SCIENCE, CYBERNETICS                    |
| Computer Science                        | COMPUTER SCIENCE, HARDWARE & ARCHITECTURE        |

Таблица 2

Перечень широких разделов с количеством входящих в них исследовательских областей

| Широкий раздел              | Количество исследовательских областей |
|-----------------------------|---------------------------------------|
| Arts & Humanities           | 14                                    |
| Social Sciences             | 25                                    |
| Life Sciences & Biomedicine | 76                                    |
| Physical Sciences           | 17                                    |
| Technology                  | 21                                    |
| Всего                       | 153                                   |

## О ТЕМАТИЧЕСКИХ КАТЕГОРИЯХ НА ЦИФРОВОЙ ПЛАТФОРМЕ *WEB OF SCIENCE*

В базах данных на платформе WoS тематическая классификация ведется только по журналам, которые в 2018 г. распределялись по 254 категориям (*categories*). Состав категорий со временем меняется, одни возникают, другие перестают наполняться, но не исчезают из списка (в 1997 г. их насчитывалось 203, в 2021 г. – 256). Не все категории входят в 153 исследовательские области (*research areas*), поскольку для некоторых не нашлось названий и они обозначены как «другие темы» (*other topics*). Пяти широким разделам (*broad categories*) соответствуют три первоначальных указателя цитирования. Но это не трехуровневая система, а простые списки разделов, областей и категорий, которые отчасти пересекаются между собой. Широкие разделы служат для распределения журналов по трем исходным индексам цитирования, а исследовательские области и категории – для классификации журналов во всех базах данных. При этом названия 101 исследовательской области входят в названия их категорий (часть областей вообще имеет только одну категорию), а в 52 – не входят, и никакой последовательности в этом не обнаруживается.

Следует подчеркнуть, что все эти разделы, области и категории предназначены для распределения журналов, и их нельзя сравнивать с такими классификациями для статей, книг, докладов, как Универсальная десятичная классификация или Государственный рубрикатор научно-технической информации. Журнал – это живой, быстро меняющийся организм, который в течение нескольких лет может менять профиль, периодичность и даже название. Журналы с похожими названиями могут иметь более широкий или узкий профиль, что предусмотрено в перечне категорий и составе исследовательских областей. Кроме того, нам известно о существовании такого явления как рассеивание статей, когда в каждом журнале имеется некоторое количество статей не по профилю этого журнала, и закона Брэдфорда, в котором утверждается, что по крайней мере две трети статей находятся в журналах других категорий. Поэтому данные о количестве статей той или иной тематики на основании расчетов по категории WoS не могут служить основанием для корректных заключений о прогрессе или регрессе какой-либо области науки. Не стоит также забывать и о том, что количество ссылок на этой платформе в какой-то мере отражает предпочтения англоязычных ученых и специалистов.

Журналы по категориям располагаются неравномерно – от 363 названий в категории *ECONOMICS* до 7 в категории *ANDROLOGY*, в среднем по медиане – 68 (вычислено по *Journal Citation Reports (JCR)* 2018 г. и указателям *SCIE* и *SSCI*), а всего их было 11878. В это число входят и 5199 мультидисциплинарных журналов (43,8%), из которых 6 имеют по 6 категорий, 36 – по 5, 264 – по 4, 1136 – по 3 и 3757 журналов имеют по 2 категории. Интересно отметить, что наиболее часто упоминаемые как примеры такого типа журналов *Science*, *Nature*, *PLoS ONE* в это число не входят, поскольку они с самого начала имеют одну рубрику *MULTIDISCIPLINARY SCIENCES*, которая насчитывает 69 журналов.

Создается впечатление, что на протяжении шестидесятилетней истории платформы WoS список ка-

тегорий создавался для решения практических задач без заранее определенной методики и не предназначался для использования в качестве важного инструмента науковедческих исследований, каким он стал в последние годы. Мы оказались в сложной ситуации. Научное сообщество нуждается в библиометрических характеристиках своих областей знания, а наиболее продвинутые центры цитирования не располагают соответствующими данными, так как сведения о количестве статей в тематических категориях неверно отражают действительность. В ближайшее время решение этой конкретной проблемы видится в разработке автоматических систем постатейного индексирования или в обращении к реферативным службам, где такая классификация статей по рубрикам осуществляется вручную. Но если смотреть глубже, то современная наука развивается междисциплинарным путем. Уже десятилетия назад Ю. Гарфилд начал создавать карты и атласы научных исследований на основе кластеров социтирования. Среди аналитических инструментов *Web of Science* появился информационный продукт *Citation Topics* (темы цитирования), который показывает, что все крупные и значимые исследования ведутся учеными, получившими образование на разных университетских факультетах и кафедрах и причисляемые научными номенклатурами к разным специальностям. Мы называем эти темы (кластеры) цитирования мульти- или междисциплинарными, но, вероятно, они-то и составляют реальную и перспективную номенклатуру научных специальностей.

### Соотношение категорий *Web of Science* и ядерной зоны закона рассеивания Брэдфорда

Было бы интересно оценить, как тематические категории соотносятся с математической моделью закона Брэдфорда, которую он окончательно сформулировал в своем трактате по документации, опубликованном незадолго до смерти в 1948 г. в 9-й главе «Документальный хаос» в следующей формулировке: «Таким образом, может быть установлен такой закон распределения статей по заданному предмету в научной периодике: если научные журналы расположить в порядке убывания числа помещенных в них статей по какому-либо заданному предмету, то в полученном списке можно выделить ядро журналов, посвященных непосредственно этому предмету, и несколько групп или зон, каждая из которых содержит столько же статей, что и ядро. Тогда количества журналов в ядре и в последующих зонах будут относиться как  $1 : a : a^2 : \dots$ »<sup>1</sup> [2, с. 116].

В начале 9-й главы книги автор уточнил: «В этой и следующей главах термин *научный* подразумевает включение термина *технический*».

<sup>1</sup> “Therefore, the law of distribution of papers on a given subject in scientific periodicals may thus be stated: if scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus, when the numbers of periodicals in the nucleus and succeeding zones will be as  $1 : n : n^2 : \dots$ ”

В начале 9-й главы автор уточнил: “In this and the next chapter, the term *scientific* is intended to include *technical*”.

## Рассеивание журнальных публикаций по этологии (поведению животных)

| Название журнала                                                                            | Категории                                                                                  | Число публикаций           |                               |             | Импакт-фактор | Страна     |
|---------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|----------------------------|-------------------------------|-------------|---------------|------------|
|                                                                                             |                                                                                            | соответствующих категориям | не соответствующих категориям | общее       |               |            |
| Animal Behaviour                                                                            | Zoology; Behavioral Sciences                                                               | 316                        | 14 (Z=5, BS=9)                | 330         | 2,869         | Англия     |
| Behavioral Ecology                                                                          | Zoology; Behavioral Sciences; Ecology                                                      | 233                        | 12 (Z=11, BS=1)               | 245         | 3,311         | США        |
| Behavioral Ecology and Sociobiology                                                         | Zoology; Behavioral Sciences; Ecology                                                      | 197                        | 9 (Z=3, BS=9)                 | 206         | 2,185         | Германия   |
| Behavioural Processes                                                                       | Zoology; Behavioral Sciences; Psychology, Biological                                       | 159                        | 20 (Z=0, BS=20)               | 179         | 1,746         | Нидерланды |
| Animal Cognition                                                                            | Zoology; Behavioral Sciences                                                               | 116                        | 0                             | 116         | 2,209         | Германия   |
| Ethology                                                                                    | Zoology; Behavioral Sciences                                                               | 95                         | 2 (Z=0, BS=2)                 | 97          | 1,398         | Германия   |
| Behaviour                                                                                   | Zoology; Behavioral Sciences                                                               | 79                         | 8 (Z=8, BS=0)                 | 87          | 1,394         | Нидерланды |
| Journal of Comparative Physiology A-Neuroethology Sensory, Neural and Behavioral Physiology | Zoology; Behavioral Sciences; Physiology; Neurosciences                                    | 65                         | 14 (Z=10, BS=4)               | 79          | 2,429         | Германия   |
| Journal of Ethology                                                                         | Zoology; Behavioral Sciences                                                               | 45                         | 0                             | 45          | 1,061         | Япония     |
| Journal of Comparative Psychology                                                           | Zoology; Behavioral Sciences; Psychology; Psychology, Multidisciplinary                    | 44                         | 0                             | 44          | 2,268         | США        |
| Learning & Behavior                                                                         | Zoology; Behavioral Sciences; Psychology Biological; Psychology, Experimental              | 40                         | 9 (Z=9, BS=0)                 | 49          | 1,015         | США        |
| Ethology Ecology & Evolution                                                                | Zoology; Behavioral Sciences                                                               | 38                         | 0                             | 38          | 1,582         | Италия     |
| Journal of Experimental Psychology-Animal Learning and Cognition                            | Zoology; Behavioral Sciences; Psychology; Psychology, Biological; Psychology, Experimental | 28                         | 9 (Z=9, BS=0)                 | 37          | 2,132         | США        |
| Acta Ethologica                                                                             | Zoology; Behavioral Sciences                                                               | 23                         | 0                             | 23          | 1,417         | Германия   |
| Brain Behavior and Evolution                                                                | Zoology; Behavioral Sciences; Neurosciences                                                | 22                         | 25 (Z=25, BS=0)               | 47          | 1 1, 915      | Швейцария  |
| <b>Всего</b>                                                                                |                                                                                            | <b>1500</b>                | <b>122</b>                    | <b>1622</b> |               |            |

\* ПРИМЕЧАНИЕ: Данные обо всех 15 журналах, соответствующих категориям *Web of Science ZOOLOGY* и *BEHAVIORAL SCIENCES*, получены из *Journal Citation Reports* за 2018 г. Просмотрены статьи за 2016 г., которые импортированы в августе 2022 г. В графе «Число публикаций, не соответствующих категориям» приняты следующие обозначения: Z – число публикаций, которые не соответствуют категории *ZOOLOGY*; BS – число публикаций, которые не соответствуют категории *BEHAVIORAL SCIENCES*. Эти числа в скобках не всегда можно складывать, так как при несоответствии статей обеим категориям сумма оказывается большей, чем число статей, не соответствующих категориям.

И в тематической категории *Web of Science* и в ядре модели Брэдфорда статьи каждого журнала сконцентрированы по отношению к его профилю, но по-разному: в первой по импакт-фактору, а во втором – по убыванию продуктивности статей данной тематики. Но как велика эта разница количественно? Ведь научное цитирование предполагает, что большинство библиографических ссылок приходится на близкие по содержанию статьи. Мы решили оценить это на небольшом примере, доступном для ручного индексирования статей. С этой целью был выбран предмет, не совпадающий с какой-либо категорией, но относящийся сразу к двум категориям. Им оказалась этология, изучающая поведение животных. Журналы, содержащие статьи по этой науке, входят в сравнительно крупные категории *ZOOLOGY* (170 журналов) и *BEHAVIOR SCIENCES* (53 журнала). На пересечении этих категорий оказалось 15 журналов (пять из Германии, четыре из США, два из Нидерландов и по одному из Англии, Италии, Швейцарии и Японии) с 1622 статьями на английском языке, частично проиндексированными также в журналах других категорий (некоторым из этих 15 журналов, помимо двух категорий, были присвоены и другие категории: *ECOLOGY*; *PSYCHOLOGY*; *PSYCHOLOGY BIOLOGICAL*; *PSYCHOLOGY EXPERIMENTAL*; *PSYCHOLOGY MULTIDISCIPLINARY*; *NEUROSCIENCES*). При индексировании учитывалось несоответствие содержания статьи каждой из двух выбранных категорий или обоим сразу. Результаты эксперимента представлены в табл. 3.

Выяснилось, что из 1622 статей 122 не соответствуют одной или обоим выбранным категориям, что составляет 7,5%. В какой мере это характеризует среднюю концентрацию статей, профильных для всех тематических категорий, установить трудно. Для распределения статей по модели закона Брэдфорда такая степень концентрации характерна для ядра и свидетельствует о коэффициенте рассеивания выше среднего, который для прикладных наук по данным Брэдфорда близок к 5. В любом случае можно предположить, что если ядро составляет 15 журналов, то во второй зоне их не может быть меньше 75, а в третьей при трехзонной модели – 375. Это означало бы, что, по крайней мере, в *Web of Science* 4,5 тыс. статей по этологии должны содержаться в 465 журналах, поскольку эта модель рассчитана только на журналы, содержащие статьи по заданному предмету, а таких статей в нашем эксперименте было 1,5 тыс.

Интересно также проследить за соответствием и несоответствием статей в наших 15 журналах тематическим категориям. Треть из них содержала только статьи, соответствующие обоим категориям. Три журнала опубликовали статьи, в которых сообщались результаты опытов с животными, но не было тематики поведения этих животных, содержание одной статьи касалось только поведения людей. Разумеется, в реальной жизни такой схематизации не бывает. Чаще всего во вторую и третью зоны модели попадают статьи соответственно из журналов смежных дисциплин и общего содержания, а также из журналов квартальных и вообще с небольшим числом статей.

В сухом остатке от этого небольшого обсервационного исследования осталось впечатление, что те-

матические категории цифровых платформ индексации и цитирования меньше страдают от статей чуждого тематического профиля, нежели брэдфордовское ядро рассеивания статей по журналам, хотя в нем коэффициент зависит от заданного предмета и полноты исследуемого массива журналов.

### Массивы публикаций-источников и библиографических ссылок

Не располагая возможностью корректного вычисления количества публикаций на цифровых платформах цитирования, мы пошли другим путем, используя классифицирование ссылок, метод библиографического сочетания публикаций и, опираясь на самые крупные отрасли науки, сопоставимые с тремя основными указателями цитирования: *SCIE*, *SSCI*, *A&HCI*. Библиографическая ссылка как объект индексирования и классифицирования имеет свои особенности, отличающие её от журнала и от статьи. Так, автор цитирующей статьи ссылается на другие работы, не будучи связан никакими условиями, кроме этических. Он может ссылаться на работы любого типа, принадлежащие любому автору, опубликованные в любом издании в любое время, совершенно не заботясь о том, где индексируется это издание и как эта ссылка будет использоваться в дальнейшем. С одной стороны, ссылка привязана к цитирующей статье, в списке литературы которой она находится и которая имеет тематическую категорию своего журнала (не всегда и необязательно ей соответствуя). С другой стороны, ссылка является библиографическим описанием цитируемой работы, которая может быть статьей или обзором в любом журнале или периодическом сборнике, докладом на конференции, главой монографии любого года опубликования, независимо от факта индексирования в каком-либо информационном издании.

Эти обстоятельства и особенно то, что разделение журналов по тематическим категориям вовсе не означает, что все статьи, получившие категорию журнала, на самом деле относятся к этой тематике, а большинство профильных статей рассеяно по журналам других профилей, вынуждает нас сопоставлять только самые крупные отрасли науки. Несмотря на то, что полученные результаты все равно имеют лишь оценочное значение, к вычислениям во многомиллионном множестве ссылок пришлось подойти со значительными ограничениями, выразившимися в создании большого количества разного рода типов массивов данных.

В настоящей работе исследуются массивы библиографических описаний публикаций двух различных типов: цитирующих публикаций – источников ссылок и самих ссылок на цитируемые публикации. В связи с этим следует учитывать ряд обстоятельств.

Во-первых, каждая публикация-источник, зарегистрированная в WoS, имеет тематическую категорию своего журнала и, следовательно, распределена по основным указателям этой платформы. Цитируемая публикация и библиографическая ссылка на нее каким-либо процедурам классифицирования не подвергается. Это требует по-разному подходить к формированию массивов публикаций-источников и ссылок соответственно. Если для формирования массивов

публикаций-источников можно непосредственно воспользоваться классификацией WoS, то для ссылок такая возможность отсутствует. Поэтому массивы ссылок мы формировали опосредованно через сформированные ранее массивы публикаций-источников. Это значит, что в массив ссылок будут включены те и только те ссылки, которые имеются в публикациях-источниках, соответствующих некоторой тематической категории или набору категорий WoS. В результате мы получим массив ссылок, который лишь очень приблизительно будет соответствовать тематической категории, использованной при формировании массива публикаций-источников. Например, если для формирования массива публикаций-источников будет взята категория «Прикладная физика» (*PHYSICS, APPLIED*), то мы получим массив ссылок, который окажется ограниченным теми цитирующими публикациями-источниками, которые формально соответствуют категории «Прикладная физика». Подчеркнем, именно формально, так как в действительности некоторые из этих цитирующих публикаций-источников необязательно будут соответствовать категории WoS, приписанной им в их журналах, а многие будут опубликованы в журналах других тематических категорий. Задав определенные ограничения на массивы публикаций-источников (годы опубликования, гражданство автора и т. п.), мы получим соответствующие массивы ссылок.

Во-вторых, все данные, которые используются в настоящем исследовании, представляют собой массивы библиографических описаний публикаций и/или массивы элементов (полей платформы) этих библиографических описаний. При этом очень важно выполнение требования, согласно которому массив должен состоять из однотипных элементов, в качестве которых мы будем рассматривать как отдельные элементы библиографического описания, так и библиографическое описание в целом.

Однородный массив. Мы будем называть однородным массив, состоящий из однотипных элементов библиографического описания. Одной из важнейших характеристик этих элементов помимо их структуры является их роль в научной коммуникации. (С этой точки зрения массивы публикаций-источников и массивы ссылок являются *неоднородными* массивами, поскольку их роли в коммуникационном процессе различны). В каждом конкретном случае требования однородности могут уточняться: ужесточаться, либо, напротив, смягчаться. Например, однородными могут быть признаны массивы ссылок и/или массивы публикаций-источников, соответствующих некоторой заданной тематике, и/или содержащие только журнальные публикации, и/или авторами которых являются ученые заданной страны.

Массивы публикаций-источников. В качестве таких массивов мы будем рассматривать множество публикаций-источников, соответствующих одной из трех баз данных основных указателей WoS, а также совокупности публикаций-источников, выделенных из более крупного массива по той или иной тематической категории, иногда с некоторыми дополнительными признаками. Например, массив российских публикаций, соответствующих некоторой тематиче-

ской категории и/или определенному их набору, и/или соавтором каждой из которых является хотя бы один зарубежный ученый, и/или только отечественные ученые.

Массивы ссылок. В качестве таких массивов мы будем рассматривать совокупность ссылок, полученных в результате реструктуризации поля *CR (Cited References* – ссылки на цитируемые публикации) платформы WoS для тех публикаций-источников, которые включены в заданный массив. Это значит, что для того, чтобы сформировать массив ссылок, необходимо вначале сформировать массив цитирующих публикаций-источников. Таким образом, массив ссылок представляет собой совокупность всех тех публикаций, на которые сослались публикации-источники из заданного массива.

Сопоставляемые массивы. Для того чтобы определить силу и характер взаимозависимости и взаимовлияния различных областей знания и отдельных направлений исследований, введем понятие пары сопоставляемых массивов<sup>2</sup>. Два однородных массива, в каждом из которых осуществляется поиск общих (идентичных, совместных) элементов, мы будем называть сопоставляемыми. Для выбора и выделения пары таких массивов можно воспользоваться либо массивами, образованными существующим списком категорий WoS, либо различными объединениями этих массивов (например, совокупностью нескольких тематически близких категорий), а также массивами, на которые наложены определенные ограничения (годы, аффилиации авторов).

Сопоставляемые массивы публикаций-источников. Пара массивов, в каждом из которых осуществляется поиск общих для них (идентичных, совместных) публикаций-источников. Одно из требований к сопоставляемым массивам состоит в том, чтобы при формировании каждой пары в нее включались только однородные массивы.

Совместная публикация-источник (цитирующая). В каждом из двух сопоставляемых массивов публикаций-источников присутствует такая публикация. Факт совпадения (идентичности) двух записей, которые соответствуют одной и той же публикации-источнику, но содержатся в двух различных массивах, устанавливается на основании совпадения значений идентификаторов *UT (unified tag* – идентификационный номер WoS) у этих записей. Важно понимать, что *совместность публикации-источника всегда определяется по отношению к рассматриваемым сопоставляемым массивам публикаций-источников*. Это значит, что одна и та же публикация в одной паре сопоставляемых массивов публикаций-источников может оказаться совместной, а в другой паре таких массивов может и не быть совместной.

Сопоставляемые массивы ссылок. Два массива ссылок, в каждом из которых осуществляется поиск общих (идентичных, совместных) ссылок, – это сопоставляемые массивы. Аналогично тому, как в массивах публикаций-источников в качестве примеров

<sup>2</sup> Можно сформировать также тройки, четверки и т. д. сопоставляемых массивов, однако в настоящей работе такие массивы не рассматриваются.

сопоставляемых массивов ссылок могут рассматриваться пары, образованные из баз данных *SSCI* и *SCIE*, *A&HCI* и *SSCI*, *SSCI* и *A&HCI*, категорий *PHYSICS*, *APPLIED* и *BIOCHEMISTRY & MOLECULAR BIOLOGY*, *MATHEMATICS* и *DEMOGRAPHY*. Важно помнить, что ссылки в WoS в отличие от публикаций-источников непосредственно не подвергаются процедурам классификации – пары сопоставляемых массивов ссылок задаются только соответствующими парами публикаций-источников.

**Совместная ссылка.** В каждом из двух сопоставляемых массивов присутствует совместная ссылка. Аналогично с ситуацией публикаций-источников, совместность ссылки всегда определяется по отношению к паре сопоставляемых массивов, в данном случае – по отношению к паре сопоставляемых массивов ссылок. Важно помнить, что совместность ссылки всегда определяется по отношению к рассматриваемым сопоставляемым массивам. Это значит, что одна и та же ссылка в одной паре сопоставляемых массивов может оказаться совместной, а в другой паре может и не быть совместной. В зависимости от способов выявления совместных ссылок они могут быть двух видов: совместные ссылки 1-го рода и совместные ссылки 2-го рода.

**Совместная ссылка 1-го рода.** Это ссылка, которая включена в состав «Списка литературы» совместной публикации-источника. Подчеркнем, что она считается совместной для определенной пары массивов публикаций-источников. Это значит, что совместными для такой пары сопоставляемых массивов ссылок являются все те публикации, на которые сослались все совместные публикации-источники. Таким образом, принадлежит ли совместная ссылка к 1-му роду или нет, зависит, прежде всего, от сопоставляемых массивов публикаций-источников, и лишь затем – от сопоставляемых массивов ссылок. Естественно, что «Список литературы» в статье после ее опубликования не меняется. Поэтому здесь проверка на идентичность ссылок не требуется, так как в таком списке не может быть двух повторяющихся ссылок. Это что касается признания совместными ссылок 1-го рода, на которые сослалась совместные публикации-источники сопоставляемых массивов. В случае, когда определяется взаимная идентичность двух или более ссылок, каждая из которых была сделана различными публикациями-источниками и при этом каждая из них уже признана совместной ссылкой 1-го рода и, следовательно, включена в их массив, мы прибегнем к результатам выполненной нами идентификации ссылок. Напомним, что эта идентификация основывалась на совпадении их библиографических описаний и/или значений DOI.

**Совместная ссылка 2-го рода.** Две ссылки признаются таковыми, если их библиографические описания и/или DOI в обоих сопоставляемых массивах совпадают. Это означает, что обе ссылки в обоих массивах могут представлять одну и ту же публикацию, т.е. они взаимно идентичны. Для проверки мы обязательно прибегнем к результатам идентификации ссылок. Очевидно, что такая процедура определения взаимной идентичности ссылок 2-го рода может совпасть с процедурой определения взаимной идентичности некоторой ограниченной совокупности

ссылок 1-го рода. Здесь нужно обратить внимание на словосочетание «некоторая ограниченная совокупность». Действительно, утверждение о совпадении процедур определения взаимной идентичности для совместных двух (или более) ссылок 1-го рода справедливо лишь когда идентичность определяется для тех ссылок, которые содержатся в двух (или более) различных массивах совместных публикаций-источников.

**Массив совместных уникальных ссылок.** Это массив таких ссылок, каждая из которых обязательно присутствует в обоих сопоставляемых массивах при условии, что частота встречаемости этой ссылки в этих массивах не учитывается. Здесь нас не интересует, сколько раз была процитирована ссылка (нас не интересует цитируемость ссылки): достаточно того, чтобы эта ссылка содержалась в каждом из сопоставляемых массивов хотя бы один раз. Именно этот факт учитывается при определении числа совместных уникальных ссылок. В ссылках 1-го рода эта характеристика вычисляется подсчетом тех уникальных ссылок, которые содержатся в публикациях-источниках и являются совместными для сопоставляемых массивов (Способ 1), а в ссылках 2-го рода – подсчетом числа ссылок, библиографические описания и/или значения DOI которых идентичны в обоих сопоставляемых массивах (Способ 2).

**Массив совместных неуникальных ссылок.** Это массив таких ссылок, каждая из которых обязательно присутствует в обоих сопоставляемых массивах и при условии, что частота ее встречаемости в этих массивах учитывается.

**Интегральный способ (Способ 3).** Опыт показал, что большая часть массива совместных уникальных ссылок, которые выделяются с помощью Способа 1, совпадает с массивом ссылок, которые выделяются с помощью Способа 2. Однако некоторые части этих массивов все же не совпадают. При этом и совпадения, и несовпадения части этих массивов содержат совместные уникальные ссылки. Это значит, что при сопоставлении двух различных массивов ссылок одновременное использование Способа 1 и Способа 2, т. е. применение интегрального способа, позволяет сформировать максимально возможную по численности совокупность совместных уникальных ссылок.

## Распределение публикаций-источников и/или ссылок

В этом и последующих разделах нам потребуется оперировать с понятием распределения публикаций-источников и/или ссылок. Будем называть распределением по категориям совместных публикаций-источников и/или ссылок, соответствующих некоторой заданной категории WoS, такой список названий категорий, в котором против каждой такой категории указано число совместных публикаций-источников и/или ссылок<sup>3</sup>.

<sup>3</sup> Более общее определение понятия «распределение публикаций» может быть сформулировано следующим образом: таблица, в которой против каждой из содержащихся в ней позиций (тематические категории, журналы, страны, годы опубликования и т. п.) указано число соответствующих публикаций.

Фрагмент распределения публикаций-источников категории *CELL BIOLOGY* по другим категориям\*

| Категории с совместными публикациями-источниками | Число совместных публикаций-источников |
|--------------------------------------------------|----------------------------------------|
| Biochemistry & Molecular Biology                 | 243                                    |
| Biophysics                                       | 130                                    |
| Oncology                                         | 56                                     |
| Geriatrics & Gerontology                         | 32                                     |
| Genetics & Heredity                              | 31                                     |
| Medicine Research & Experimental                 | 27                                     |
| Biotechnology & Applied Microbiology             | 24                                     |
| Immunology                                       | 23                                     |
| Biology                                          | 22                                     |
| Physiology                                       | 22                                     |
| Developmental Biology                            | 18                                     |

\* ПРИМЕЧАНИЕ: В таблицу включены первые 11 (по числу совместных публикаций-источников) из 30 тематических категорий WoS.

В табл. 4 представлен фрагмент распределения публикаций-источников WoS, соответствующих категории *CELL BIOLOGY* по другим категориям. В это распределение включены те российские (совместные) публикации-источники, которые были опубликованы в период 2017–2019 гг. при условии, что автором/соавтором каждой такой публикации был хотя бы один зарубежный ученый.

Как следует из приведенных нами определений совместных ссылок 1-го и 2-го рода, основное их отличие друг от друга состоит в следующем. Для того чтобы публикация была признана в качестве ссылки 1-го рода, необходимым и достаточным условием является требование – чтобы в обоих сопоставляемых массивах на эту публикацию сослалась одна и та же публикация-источник, тогда как в случае ссылки 2-го рода это требование отсутствует, но вступает в силу требование непосредственной идентичности библиографических описаний или совпадений значений DOI.

Способ 1 основывается на вполне понятном утверждении, согласно которому «Список литературы» с момента опубликования некоторой публикации остается неизменным. Это, в частности, значит, что вне зависимости от того, в какой из сопоставляемых массивов включена публикация-источник, она будет содержать один тот же набор ссылок. Следовательно, мы можем формировать массив совместных ссылок, обращаясь к тем публикациям-источникам, которые были признаны совместными и, что сейчас важно подчеркнуть, они также являются и совместными цитируемыми публикациями. Таким образом, совокупность ссылок, которые были процитированы всеми совместными публикациями, составит массив совместных ссылок. Отобрать этот массив мы можем из любого сопоставляемого ему массива ссылок.

При этом способе существует проблема незаполненности поля DOI и различий в элементах библио-

графического описания ссылки. Однако острота этой проблемы существенно меньше, чем при другом способе. В особенности, когда один из сопоставляемых массивов включает ссылки из одной базы данных WoS, а второй – из другой, например, из SCIE и SSCI, соответственно. В частности, хотя бы потому, что мы имеем дело с библиографическими описаниями ссылок из одной и той же базы данных, что сокращает их вариативность. Таким образом, использование Способа 1, опирающегося на совместные публикации-источники, в целом вполне правомерно, за исключением случая, когда в одном и том же массиве библиографические описания одной и той же ссылки или идентификаторы DOI будут различаться. Такая ситуация возможна, когда подобные ссылки появятся в нескольких публикациях-источниках из одного и того же массива и процедурой автоматической (программной) идентификации будут восприняты как разные.

Способ 2 не требует предварительного формирования массива совместных публикаций-источников. Необходимо лишь сформировать массив ссылок на публикации, соответствующие заданной рубрике. Затем следует определить, какие публикации-источники (за исключением тех, которые принадлежат заданной рубрике) также имеют эти ссылки в своем списке. После этого остается лишь определить, к каким категориям принадлежат эти другие («не исключенные») публикации-источники. Подсчитав, сколько публикациями-источниками было сделано ссылок, принадлежащих к каждой из определенных таким способом категорий, мы получим распределение их ссылок по другим категориям. И при Способе 2 существует вероятность, что часть ссылок не попадет в число совместных. Как показал предварительный анализ, отдельные элементы библиографического описания в ссылках на одну и ту же публикацию могут отличаться друг от друга, а поле DOI может оказаться незаполненным. В резуль-

тате, поскольку идентификация выполняется программным образом, такие ссылки не будут признаны совместными и не попадут в число совместных уникальных ссылок<sup>4</sup>. Особенность состоит в том, что несовпадение описаний и незапятнанность поля DOI одной и той же ссылки при Способе 2 имеет место внутри одного массива ссылок, соответствующего только публикациям-источникам некоторой заданной категории, а при Способе 1 – во всем общем массиве ссылок.

Принятое нами деление совместных ссылок на два типа, с одной стороны, вынуждено и вызвано рядом обстоятельств, которые существенно затрудняют процесс идентификации ссылок и ограничивают массивы, которые могут быть подвергнуты его процедурам. Среди этих обстоятельств: различие в требованиях в редакциях разных журналов к составлению «Списка литературы»; изменение со временем этих требований в одном и том же журнале; отсутствие у значительного числа ссылок идентификатора DOI; небрежность авторов при оформлении «Списка литературы». С другой стороны, деление ссылок на два типа позволяет по-новому взглянуть на проблему их использования в качестве инструмента построения соответствующих классификаций и распределений и, в конечном счете, – разработать некоторый новый подход к применению этого инструмента для исследования взаимосвязей и взаимовлияния различных разделов науки.

## Два типа классификации ссылок

Наша задача заключается не только в определении силы взаимосвязи двух сопоставляемых отраслей науки (в частности, WoS-категорий), но и в определении тематического спектра интересующего нас раздела, т. е. распределении ссылок по категориям, которое будет получено в результате этого сопоставления. При этом распределение, полученное по Способу 1, заранее задано распределением тех совместных публикаций-источников, в которых содержатся рассматриваемые ссылки. А если учесть, что распределение публикаций-источников задано распределением журналов, в которых они были опубликованы, то оказывается, что распределение по WoS-категориям ссылок 1-го рода уже задано соответствующим распределением журналов по категориям. При этом тематический спектр конкретного журнала в WoS представлен в виде набора всего из нескольких категорий<sup>5</sup>. Для подтверждения этого можно вернуться к данным, полученным в ходе анализа годового выпуска *Journal Citation Reports* за 2018 г.: каждому из 6679 журналов, т. е. более чем половине (54,1%) зарегистрированных в нем журналов, была приписана только одна категория, 3757 журналам – две, остальным 1442 журналам – от 3 до 6. Такая практика порождает заранее заданную, очень усеченную

и достаточно жесткую классификацию публикаций-источников и совместных ссылок 1-го рода.

Изложенные соображения приводят к необходимости различать два типа принципиально различающихся классификаций ссылок: усеченную (ограниченную) и не усеченную (свободную). *Усеченная (ограниченная) тематическая классификация ссылок* осуществляется через распределение журналов, т. е. публикаций-источников в них. На нее тематические ограничения (например, категориями WoS) накладываются не только при формировании массива цитируемых публикаций, но и, что очень важно, на публикации, в которых эти ссылки содержатся. Действительно, в усеченной классификации ссылок мы для данной категории рассматриваем ссылки, которые содержатся только в тех публикациях-источниках, которые являются совместными для этой категории с ограниченным набором других категорий.

Напротив, *неусеченная (свободная) классификация ссылок* способствует распределению ссылок с минимальным тематическим ограничением. Это классификация (распределение ссылок в публикациях-источниках данной категории по другим категориям), которая удовлетворяет единственному, но также очень важному требованию, а именно: тематические параметры (категории WoS), которыми характеризуются ссылки на цитируемые публикации, не должны накладываться в качестве ограничений на публикации, содержащие эти ссылки. Иначе, в случае свободной классификации ссылок, мы задаем (посредством отбора публикаций-источников, которые соответствуют заданной рубрике, но не обязательно являющиеся совместными) некоторую категорию, а затем смотрим, каким категориям (помимо заданной) соответствуют те публикации-источники, в которых эти ссылки содержатся. При этом для каждой такой рубрики подсчитывается общее число процитировавших публикаций-источников и общее число ссылок.

Проиллюстрируем это различие типов классификации ссылок на примере статьи «*Obstion of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the L*», посвящённой изложению результатов поиска бозона Хигса и опубликованной в 2012 г. в журнале «*Physics Letters B*», которому присвоены три категории: *PHYSICS PARTICLES & FIELDS*, *ASTRONOMY & ASTROPHYSICS* и *PHYSICS NUCLEAR*. Статье, в соответствии с практикой WoS также присвоены эти три категории. Всего по данным WoS, на эту статью сослались 1089 публикаций. Из них 776 была присвоена хотя бы одна из тех трех категорий, которые были присвоены и рассматриваемой статье. Остальным 313 цитирующим публикациям присвоены другие категории, что в случае усеченной классификации ведет к исключению этих публикаций из дальнейшего анализа. Картина кардинально изменится, если мы при поиске ссылок выйдем за пределы тех категорий, которые присвоены цитируемой статье, т.е. используем подход, соответствующий свободной классификации. В этом случае распределение цитирующих публикаций будет состоять уже не из трех, а из 101 категории. В числе 99 «дополнительных» катего-

<sup>4</sup> Задача визуальной идентификации в массиве, превышающем 15 млн ссылок, практически не разрешима, а при программной идентификации различия даже в одном знаке приведут к тому, что публикации не будут признаны идентичными.

<sup>5</sup> С увеличением набора категорий, которые в WoS на данный момент присвоены журналу, автоматически растет и набор категорий, которым характеризуется каждая его публикация.

рий, помимо тех из них, которые тематически относительно близки к указанным трем категориям (*PHYSICS MULTIDISCIPLINARY, PHYSICS MATHEMATICAL, PHYSICS APPLIED, NUCLEAR SCIENCE TECHNOLOGY, PHYSICS ATOMIC MOLECULAR CHEMICAL, PHYSICS CONDENSED MATTER*), присутствуют и тематически не близкие: *BIOLOGY; STATISTICS & PROBABILITY; MATHEMATICS APPLIED, MULTIDISCIPLINARY SCIENCES; COMPUTER SCIENCE, INFORMATION SYSTEMS; INTERDISCIPLINARY APPLICATIONS, INSTRUMENTS & INSTRUMENTATION; RADIOLOGY NUCLEAR MEDICINE, MEDICAL IMAGING*. Особенно интересно, что здесь присутствуют категории, которые соответствуют общественным и гуманитарным наукам: *HISTORY AND PHILOSOPHY OF SCIENCE; EDUCATION, SCIENTIFIC DISCIPLINES; ECONOMICS*. Таким образом, если не ограничивать цитирующие публикации категориями, которые присвоены журналу и цитируемым в нем статьям, то в результате мы получим значительно более богатую картину, соответствующую неусеченной (свободной) классификации ссылок.

Возникает вопрос: зачем тогда вообще рассматривается Способ 1. Дело в том, что при сопоставлении крупных разделов науки (соответствующих трем изначальным указателям цитирования Ю. Гарфилда) нужны некоторые показатели, которые можно получить только этим способом. Мы покажем ряд таких показателей, в частности, долю совместных публикаций-источников и долю совместных ссылок<sup>6</sup>. Доля

<sup>6</sup> Более детальный список показателей, которые позволяет получить Способ 1, используемый совместно со Способом 2, выглядит следующим образом: Массив публикаций-источников (ПИ); Подмассив совместных ПИ; Доля подмассива совместных ПИ: отношение числа совместных ПИ к общему числу ПИ (%); Массив неуникальных (повторяющихся) ссылок; Массив уникальных ссылок; Подмассив неуникальных (повторяющихся) совместных ссылок 1-го рода; Доля подмассива неуникальных (повторяющихся) совместных ссылок 1-го рода: отношение числа неуникальных совместных ссылок 1-го рода к общему числу неуникальных ссылок (%); Подмассив уникальных совместных ссылок 1-го рода; Доля подмассива уникальных совместных ссылок 1-го рода: отношение числа уникальных совместных ссылок 1-го рода к общему числу уникальных ссылок (%); Объединенный подмассив неуникальных (повторяющихся) совместных ссылок - объединение множества совместных ссылок 2-го рода с множеством совместными ссылок 1-го рода (совместное применение способа 1 и способа 2); Доля объединенного подмассива неуникальных совместных ссылок: отношение общего числа неуникальных (повторяющихся) совместных ссылок 2-го рода, дополненное неуникальными совместными ссылок 1-го рода - к общему числу неуникальных ссылок (совместное применение Способа 1 и Способа 2) (%); Объединенный подмассив уникальных совместных ссылок - объединение множества уникальных совместных ссылок 2-го рода с множеством уникальных совместных ссылок 1-го рода (совместное применение Способа 1 и Способа 2); Доля объединенного подмассива уникальных совместных ссылок: отношение общего числа неуникальных (повторяющихся) совместных ссылок 2-го рода, дополненное уникальными совместными ссылками 1-го рода к общему числу неуникальных ссылок (совместное применение Способа 1 и Способа 2).

подмассива совместных публикаций-источников – это отношение числа совместных публикаций-источников к общему числу публикаций-источников в данном массиве, выраженное в процентах. Доля подмассива совместных ссылок – это отношение числа совместных ссылок к общему числу ссылок в данном массиве, выраженное в процентах. Исходя из этих определений, расчет доли подмассива совместных публикаций-источников, а также расчет доли подмассива совместных ссылок можно выполнить по формуле:

$$Sh_a^b = \frac{N_{a\_And\_b}}{N_a} * 100,$$

где:

$Sh_a^b$  – доля тех публикаций в массиве  $a$ , которые присутствуют и в массиве  $b$ ;

$N_{a\_And\_b}$  – число тех публикаций в массиве  $a$ , которые присутствуют и в массиве  $b$ ; фрагмент  $\_And\_$  нижнего индекса указывает на то, что все публикации, включенные в число  $N_{a\_And\_b}$ , в обязательном порядке должны присутствовать и в массиве  $a$  и в массиве  $b$ ;

$N_a$  – общее число публикаций в массиве  $a$ .

Наш подход к определению взаимосвязей и взаимовлияния различных разделов науки заключается в движении от заданной классификации публикаций-источников к свободной классификации ссылок.

## Взаимовлияние крупных отраслей науки

Выберем самые крупные отрасли: естественные, точные и прикладные науки (БД *SCIE*), общественные науки (БД *SSCI*) и искусство и гуманитарные науки (БД *A&HCI*). Определим ряд таких показателей, как совместные публикации-источники и совместные ссылки для пар БД *SSCI* ← БД *SCIE*, БД *A&HCI* ← БД *SCIE*. В табл. 5–9 содержатся значения описанных показателей для различных сопоставляемых пар баз данных на основе массивов публикаций, автором/соавтором каждой из которых является хотя бы один российский ученый. Табл. 5 является сводной, в ней представлены все пары различных баз данных. При этом данные, соответствующие этим парам, в большинстве случаев, охватывают один и тот же период появления публикаций-источников: 1993–2020 гг., т.е. 28 лет. Другой период охвата – это те пары, в которых присутствует хотя бы одна из следующих баз данных: *BKCI-SSH*, *BKCI-SSH* и *CPCI-S*. Дело в том, что данные в первых двух БД (*BKCI-SSH*, *BKCI-SSH*) в *WoS* охватывают период с 2005 г. В случае БД *CPCI-S* мы располагаем данными, начиная только с 2006 г. В итоге было принято решение ограничить период охвата для таких пар пятью годами: 2013–2017 гг. В табл. 5 эти пары помечены символом (\*).

Статистические характеристики основных сопоставляемых баз данных Web of Science

| № строки | Сопоставляемые<br>базы данных WoS | Публикации-<br>источники (ПИ) в SSCI | Совместные<br>публикации-источники | Доля подмассива<br>совместных ПИ (%) | Массив неуникальных<br>ссылок | Массив уникальных<br>ссылок | Подмассив неуникальных<br>совместных ссылок 1-го рода | Доля подмассива неуник.<br>совместных ссылок 1-го рода (%) | Подмассив уникальных<br>совместных ссылок 1-го рода | Доля подмасс. уникальн. совместв.<br>ссылок 1-го рода (%) | Объедин. подмас. неуник.<br>совместных ссылок | Доля объединен. подмас.<br>неуник совм. ссылок (%) | Объединенный подмассив<br>уникальных совместных ссылок | Доля объединен. подмас.<br>уник. совм. ссылок (%) |
|----------|-----------------------------------|--------------------------------------|------------------------------------|--------------------------------------|-------------------------------|-----------------------------|-------------------------------------------------------|------------------------------------------------------------|-----------------------------------------------------|-----------------------------------------------------------|-----------------------------------------------|----------------------------------------------------|--------------------------------------------------------|---------------------------------------------------|
|          |                                   |                                      |                                    |                                      |                               |                             |                                                       |                                                            |                                                     |                                                           |                                               |                                                    |                                                        |                                                   |
| 1        | SSCI ← SCIE                       | 30430                                | 12176                              | 40,0                                 | 1197988                       | 809026                      | 453358                                                | 37,8                                                       | 398844                                              | 49,3                                                      | 806123                                        | 67,3                                               | 404310                                                 | 50,0                                              |
| 2        | SSCI ← A&HCI                      | 30430                                | 1766                               | 5,8                                  | 1197988                       | 809026                      | 105299                                                | 8,8                                                        | 102536                                              | 12,7                                                      | 110269                                        | 9,2                                                | 104220                                                 | 12,9                                              |
| 3        | SSCI ← CPCI-S*                    | 7195                                 | 336                                | 4,5                                  | 316223                        | 224602                      | 5299                                                  | 1,7                                                        | 5182                                                | 2,3                                                       | 21187                                         | 6,7                                                | 11315                                                  | 5,0                                               |
| 4        | SSCI ← CPCI-SSH                   | 30430                                | 378                                | 1,2                                  | 1197988                       | 809026                      | 14929                                                 | 1,2                                                        | 14819                                               | 1,8                                                       | 29624                                         | 2,5                                                | 24520                                                  | 3,0                                               |
| 5        | SSCI ← BKCI-SSH*                  | 7195                                 | 74                                 | 1,0                                  | 316223                        | 224602                      | 2345                                                  | 0,7                                                        | 2345                                                | 1,0                                                       | 5141                                          | 1,3                                                | 4536                                                   | 2,0                                               |
| 6        | SSCI ← BKCI-S*                    | 7195                                 | 11                                 | 0,15                                 | 316223                        | 224602                      | 250                                                   | 0,08                                                       | 250                                                 | 0,11                                                      | 3677                                          | 1,16                                               | 2895                                                   | 1,29                                              |
| 7        | A&HCI ← SSCI                      | 14795                                | 1766                               | 11,9                                 | 409854                        | 369479                      | 105299                                                | 25,7                                                       | 102536                                              | 27,7                                                      | 117427                                        | 28,6                                               | 104363                                                 | 28,2                                              |
| 8        | A&HCI ← SCIE                      | 14795                                | 633                                | 4,3                                  | 409854                        | 369479                      | 52110                                                 | 12,7                                                       | 49401                                               | 13,4                                                      | 119327                                        | 29,1                                               | 52156                                                  | 14,1                                              |
| 9        | A&HCI ← CPCI-S*                   | 4049                                 | 27                                 | 0,7                                  | 143493                        | 126277                      | 2918                                                  | 2,0                                                        | 2870                                                | 2,3                                                       | 6112                                          | 4,3                                                | 4003                                                   | 3,2                                               |
| 10       | A&HCI ← CPCI-SSH                  | 14795                                | 162                                | 1,1                                  | 409854                        | 369479                      | 6463                                                  | 1,6                                                        | 6462                                                | 1,7                                                       | 8026                                          | 2,0                                                | 7470                                                   | 2,0                                               |
| 11       | A&HCI ← BKCI-SSH*                 | 4049                                 | 87                                 | 2,1                                  | 143493                        | 126277                      | 4003                                                  | 2,8                                                        | 3992                                                | 3,16                                                      | 4547                                          | 3,17                                               | 4396                                                   | 3,5                                               |
| 12       | A&HCI ← BKCI-S*                   | 4049                                 | 14                                 | 0,3                                  | 143493                        | 126277                      | 2482                                                  | 1,7                                                        | 2472                                                | 2,0                                                       | 3015                                          | 2,1                                                | 2793                                                   | 2,2                                               |
| 13       | SCIE ← SSCI                       | 892171                               | 12176                              | 1,4                                  | 23791085                      | 13065712                    | 453358                                                | 1,9                                                        | 398844                                              | 3,0                                                       | 491892                                        | 2,1                                                | 404850                                                 | 3,1                                               |
| 14       | SCIE ← A&HCI                      | 892171                               | 633                                | 0,07                                 | 23791085                      | 13065712                    | 52110                                                 | 0,2                                                        | 49401                                               | 0,38                                                      | 57184                                         | 0,2                                                | 51811                                                  | 0,4                                               |
| 15       | SCIE ← CPCI-S*                    | 186290                               | 14208                              | 7,6                                  | 5819456                       | 3661381                     | 174966                                                | 3,0                                                        | 159257                                              | 4,3                                                       | 577357                                        | 9,9                                                | 276998                                                 | 7,6                                               |
| 16       | SCIE ← CPCI-SSH                   | 892171                               | 348                                | 0,04                                 | 23791085                      | 13065712                    | 9578                                                  | 0,04                                                       | 9509                                                | 0,07                                                      | 30645                                         | 0,13                                               | 25664                                                  | 0,20                                              |
| 17       | SCIE ← BKCI-SSH*                  | 186290                               | 1                                  | <0,00                                | 5819456                       | 3661381                     | 192                                                   | <0,00                                                      | 192                                                 | 0,01                                                      | 2324                                          | 0,04                                               | 2069                                                   | 0,06                                              |
| 18       | SCIE ← BKCI-S*                    | 186290                               | 172                                | 0,1                                  | 5819456                       | 3661381                     | 23181                                                 | 0,4                                                        | 22656                                               | 0,62                                                      | 97050                                         | 17                                                 | 74627                                                  | 2,0                                               |
| 19       | CPCI-SSH ← CPCI-S*                | 9520                                 | 2426                               | 25,5                                 | 159495                        | 135985                      | 52566                                                 | 33,0                                                       | 50106                                               | 36,8                                                      | 88137                                         | 55,3                                               | 51422                                                  | 37,8                                              |
| 20       | CPCI-S ← CPCI-SSH*                | 68168                                | 2426                               | 3,6                                  | 998305                        | 716746                      | 52566                                                 | 5,3                                                        | 50106                                               | 6,99                                                      | 59082                                         | 5,9                                                | 52412                                                  | 7,3                                               |
| 21       | BKCI-SSH ← BKCI-S*                | 1382                                 | 147                                | 10,6                                 | 38520                         | 34644                       | 8522                                                  | 22,1                                                       | 8479                                                | 24,5                                                      | 8862                                          | 23,0                                               | 8522                                                   | 24,6                                              |
| 22       | BKCI-S ← BKCI-SSH*                | 3540                                 | 147                                | 4,1                                  | 177083                        | 150239                      | 8522                                                  | 4,8                                                        | 8479                                                | 5,6                                                       | 8705                                          | 4,9                                                | 8584                                                   | 5,7                                               |

\* Базы данных, помеченные символом (\*), охватывают период 2013–2017 гг.; массив ссылок соответствует периоду 1700–2020 гг.; публикации-источники опубликованы в 2013–2017 гг.

Отношение между базами данных в табл. 5 внутри сопоставляемой пары не является полностью симметричным. В графе 1 этой таблицы в каждой строке приводится соответствующая пара БД: слева – БД, которая в данный момент (в данной строке) подвергается исследованию; справа – БД, с которой сопоставляется левая, – ей посвящено большинство граф табл. 5 (2, 4, 5, 6, 8, 10, 12, 14), тогда как остальные графы (3, 7, 9, 13) содержат показатели, общие для данной пары сопоставляемых баз данных. Эти последние показатели представляют собой количества совместных (для данной пары) публикаций-источников и/или ссылок.

Если отношение внутри пары сопоставляемых баз данных в табл. 5 не симметрично, то сами сопоставляемые пары, напротив, являются симметричными. Так, пара «SSCI ← SCIE» (1-я строка) симметрична паре «SCIE ← SSCI» (13-я строка), пара «SSCI ← A&HCI» (2-я строка) симметрична паре «A&HCI ← SSCI» (7-я строка), пара «A&HCI ← SCIE» (8-я строка) симметрична паре «SCIE ← A&HCI» (14-я строка), пара «CPCI-SSH ← CPCI-S» (10-я строка) симметрична паре «CPCI-S ← CPCI-SSH» (20-я строка) и пара «BKCI-SSH ← BKCI-S» (21-я строка) – паре «BKCI-S ← BKCI-SSH» (22-я строка).

При формировании сопоставляемых пар для журнальных БД (SCIE, SSCI и A&HCI) требования к однородности массивов ослаблены и состоят только в том, чтобы они содержали публикации-источники (см. табл. 5). Особое внимание к трем журнальным базам данных здесь объясняется той ролью, которую играют журнальные публикации в научных коммуникациях. Для остальных четырех БД требование к однородности более жесткое: на эти массивы накладывается требование однотипности этих публикаций-источников. Так, для пары «CPCI-SSH ← CPCI-S» (15-я строка) это требование заключается в том, чтобы сопоставляемые массивы содержали только публикации-источники, представляющие собой материалы конференций, а для пары «BKCI-SSH ← BKCI-S» (21-я строка) – публикации-источники, представляющие собой книги и главы из книг.

**Влияние естественных и прикладных наук (SCIE) на общественные науки (SSCI)  
между базами данных основных указателей цитирования Web of Science в 1993–2020 гг.,  
определенное по ссылкам**

| Сопоставляемые базы данных SCIE → SSCI | Период<br>(годы) | Массив публикаций-источников<br>в БД SSCI          | Подмассив совместных с БД SSCI<br>публикаций-источников | Доля подмассива (%) | Массив ссылок | Массив уникальных ссылок | Подмассив неуникальных<br>совместных ссылок 1-го рода | Доля подмассива неуникальных<br>совместных ссылок 1-го рода (%) | Подмассив уникальных совместных<br>ссылок 1-го рода | Доля подмассива уникальных<br>совместных ссылок 1-го рода (%) | Объединенный подмассив<br>неуникальных совместных ссылок | Доля объединенного подмассива<br>неуникальных совместных ссылок (%) | Объединенный подмассив<br>уникальных совместных ссылок | Доля объединенного подмассива<br>уникальных совместных ссылок (%) |
|----------------------------------------|------------------|----------------------------------------------------|---------------------------------------------------------|---------------------|---------------|--------------------------|-------------------------------------------------------|-----------------------------------------------------------------|-----------------------------------------------------|---------------------------------------------------------------|----------------------------------------------------------|---------------------------------------------------------------------|--------------------------------------------------------|-------------------------------------------------------------------|
|                                        |                  | Все публикации российских ученых                   |                                                         |                     |               |                          |                                                       |                                                                 |                                                     |                                                               |                                                          |                                                                     |                                                        |                                                                   |
|                                        | 1993–1997        | 3836                                               | 920                                                     | 24,0                | 57580         | 49379                    | 19486                                                 | 33,8                                                            | 18468                                               | 37,4                                                          | 45711                                                    | 79,4                                                                | 18764                                                  | 38,0                                                              |
|                                        | 1998–2002        | 3842                                               | 1175                                                    | 30,6                | 86503         | 70819                    | 29555                                                 | 34,2                                                            | 28045                                               | 39,6                                                          | 67773                                                    | 78,3                                                                | 28351                                                  | 40,0                                                              |
|                                        | 2003–2007        | 3209                                               | 1105                                                    | 34,4                | 92518         | 72213                    | 35535                                                 | 38,4                                                            | 33346                                               | 46,2                                                          | 86936                                                    | 94,0                                                                | 33760                                                  | 46,8                                                              |
|                                        | 2008–2012        | 4955                                               | 2153                                                    | 43,5                | 136216        | 105431                   | 47353                                                 | 34,8                                                            | 44507                                               | 42,2                                                          | 112386                                                   | 82,5                                                                | 45461                                                  | 43,1                                                              |
|                                        | 2013–2017        | 7195                                               | 3068                                                    | 42,6                | 316223        | 224602                   | 115335                                                | 36,5                                                            | 106705                                              | 47,5                                                          | 272236                                                   | 86,1                                                                | 109382                                                 | 48,7                                                              |
|                                        | 2017–2020        | 7393                                               | 3754                                                    | 50,8                | 508948        | 326067                   | 205952                                                | 40,5                                                            | 187045                                              | 57,4                                                          | 438286                                                   | 86,1                                                                | 189466                                                 | 58,1                                                              |
|                                        | Период<br>(годы) | Публикации только российских ученых                |                                                         |                     |               |                          |                                                       |                                                                 |                                                     |                                                               |                                                          |                                                                     |                                                        |                                                                   |
|                                        |                  | Публикации российских ученых с участием зарубежных |                                                         |                     |               |                          |                                                       |                                                                 |                                                     |                                                               |                                                          |                                                                     |                                                        |                                                                   |
|                                        | 1993–1997        | 3478                                               | 727                                                     | 20,9                | 46938         | 41618                    | 14003                                                 | 29,8                                                            | 13466                                               | 32,4                                                          | 29337                                                    | 62,5                                                                | 13676                                                  | 32,9                                                              |
|                                        | 1998–2002        | 3228                                               | 778                                                     | 24,1                | 56181         | 48434                    | 13968                                                 | 24,9                                                            | 13415                                               | 27,7                                                          | 32091                                                    | 57,1                                                                | 13704                                                  | 28,3                                                              |
|                                        | 2003–2007        | 2489                                               | 646                                                     | 25,9                | 53159         | 44308                    | 15480                                                 | 29,1                                                            | 14837                                               | 33,5                                                          | 39238                                                    | 73,8                                                                | 15177                                                  | 34,2                                                              |
|                                        | 2008–2012        | 3747                                               | 1466                                                    | 39,1                | 70065         | 59159                    | 19629                                                 | 28,0                                                            | 18867                                               | 31,9                                                          | 42426                                                    | 60,5                                                                | 19283                                                  | 32,6                                                              |
|                                        | 2013–2017        | 4445                                               | 1716                                                    | 38,6                | 113902        | 91272                    | 32035                                                 | 28,1                                                            | 30683                                               | 33,6                                                          | 90294                                                    | 79,3                                                                | 31794                                                  | 34,8                                                              |
|                                        | 2017–2020        | 3845                                               | 1771                                                    | 46,1                | 161785        | 117263                   | 56516                                                 | 34,9                                                            | 53514                                               | 45,4                                                          | 154990                                                   | 95,8                                                                | 54765                                                  | 46,7                                                              |
|                                        | 1993–1997        | 259                                                | 42                                                      | 16,2                | 8307          | 5972                     | 820                                                   | 9,9                                                             | 820                                                 | 13,7                                                          | 6072                                                     | 73,1                                                                | 871                                                    | 14,6                                                              |
|                                        | 1998–2002        | 594                                                | 94                                                      | 15,8                | 28575         | 21657                    | 2788                                                  | 9,8                                                             | 2746                                                | 12,7                                                          | 12514                                                    | 43,8                                                                | 3159                                                   | 14,6                                                              |
|                                        | 2003–2007        | 694                                                | 256                                                     | 36,9                | 38010         | 27348                    | 11549                                                 | 30,4                                                            | 11041                                               | 40,4                                                          | 40791                                                    | 107,3                                                               | 11811                                                  | 43,2                                                              |
|                                        | 2008–2012        | 1171                                               | 654                                                     | 55,9                | 64777         | 46405                    | 26565                                                 | 41,0                                                            | 25002                                               | 53,9                                                          | 75256                                                    | 116,2                                                               | 25523                                                  | 55,0                                                              |
|                                        | 2013–2017        | 2608                                               | 1253                                                    | 48,0                | 193470        | 131120                   | 79708                                                 | 41,2                                                            | 74211                                               | 56,6                                                          | 196719                                                   | 101,7                                                               | 75713                                                  | 57,7                                                              |
|                                        | 2017–2020        | 3533                                               | 1977                                                    | 56,0                | 346407        | 215002                   | 149259                                                | 43,1                                                            | 136398                                              | 63,4                                                          | 322640                                                   | 93,1                                                                | 137688                                                 | 64,0                                                              |

Таблица 7

**Влияние искусства и гуманитарных наук (A&HCI) на общественные науки (SSCI) между базами данных  
основных указателей цитирования Web of Science в 1993–2020 гг., определенное по ссылкам**

| Сопоставляемые базы данных A&HCI → SSCI | Период<br>(годы) | Массив публикаций-источников<br>в БД SSCI | Подмассив совместных с SSCI<br>публикаций-источников | Доля подмассива (%) | Массив ссылок | Массив уникальных ссылок | Подмассив неуникальных<br>совместных ссылок 1-го рода | Доля подмассива неуникальных<br>совместных ссылок 1-го рода (%) | Подмассив уникальных<br>совместных ссылок 1-го рода | Доля подмассива уникальных<br>совместных ссылок 1-го рода (%) | Объединенный подмассив<br>неуникальных совместных ссылок | Доля объединенного подмассива<br>неуникальных совместных ссылок (%) | Объединенный подмассив<br>уникальных совместных ссылок | Доля объединенного подмассива<br>уникальных совместных ссылок (%) |
|-----------------------------------------|------------------|-------------------------------------------|------------------------------------------------------|---------------------|---------------|--------------------------|-------------------------------------------------------|-----------------------------------------------------------------|-----------------------------------------------------|---------------------------------------------------------------|----------------------------------------------------------|---------------------------------------------------------------------|--------------------------------------------------------|-------------------------------------------------------------------|
|                                         |                  | Все публикации российских ученых          |                                                      |                     |               |                          |                                                       |                                                                 |                                                     |                                                               |                                                          |                                                                     |                                                        |                                                                   |
|                                         | 1993–1997        | 3836                                      | 224                                                  | 5,8                 | 57580         | 49379                    | 7661                                                  | 13,3                                                            | 7644                                                | 15,5                                                          | 7903                                                     | 13,7                                                                | 7728                                                   | 15,7                                                              |
|                                         | 1998–2002        | 3842                                      | 132                                                  | 3,4                 | 86503         | 70819                    | 5521                                                  | 6,4                                                             | 5516                                                | 7,8                                                           | 6098                                                     | 7,0                                                                 | 5718                                                   | 8,1                                                               |
|                                         | 2003–2007        | 3209                                      | 124                                                  | 3,9                 | 92518         | 72213                    | 6096                                                  | 6,6                                                             | 6008                                                | 8,3                                                           | 7128                                                     | 7,7                                                                 | 6222                                                   | 8,6                                                               |
|                                         | 2008–2012        | 4955                                      | 300                                                  | 6,1                 | 136216        | 105431                   | 16442                                                 | 12,1                                                            | 16052                                               | 15,2                                                          | 19139                                                    | 14,1                                                                | 16439                                                  | 15,6                                                              |
|                                         | 2013–2017        | 7195                                      | 586                                                  | 8,1                 | 316223        | 224602                   | 40619                                                 | 12,8                                                            | 39828                                               | 17,7                                                          | 44977                                                    | 14,2                                                                | 40700                                                  | 18,1                                                              |
|                                         | 2017–2020        | 7393                                      | 400                                                  | 5,4                 | 508948        | 326067                   | 28960                                                 | 5,7                                                             | 28528                                               | 8,7                                                           | 34534                                                    | 6,8                                                                 | 30576                                                  | 9,4                                                               |

| Публикации только российских ученых                |      |     |      |        |        |       |      |       |      |       |      |       |      |
|----------------------------------------------------|------|-----|------|--------|--------|-------|------|-------|------|-------|------|-------|------|
| 1993–1997                                          | 3478 | 213 | 6,1  | 46938  | 41618  | 7317  | 15,6 | 7307  | 17,6 | 7482  | 15,9 | 7345  | 17,6 |
| 1998–2002                                          | 3228 | 116 | 3,6  | 56181  | 48434  | 4056  | 7,2  | 4051  | 8,4  | 4245  | 7,6  | 4113  | 8,5  |
| 2003–2007                                          | 2489 | 94  | 3,8  | 53159  | 44308  | 2700  | 5,1  | 2693  | 6,1  | 3261  | 6,1  | 2764  | 6,2  |
| 2008–2012                                          | 3747 | 214 | 5,7  | 70065  | 59159  | 8463  | 12,1 | 8414  | 14,2 | 9661  | 13,8 | 8548  | 14,4 |
| 2013–2017                                          | 4445 | 343 | 7,7  | 113902 | 91272  | 21237 | 18,6 | 21053 | 23,1 | 23914 | 21,0 | 21448 | 23,5 |
| 2017–2020                                          | 3845 | 243 | 6,3  | 161785 | 117263 | 15451 | 9,6  | 15324 | 13,1 | 17811 | 11,0 | 16078 | 13,7 |
| Публикации российских ученых с участием зарубежных |      |     |      |        |        |       |      |       |      |       |      |       |      |
| 1993–1997                                          | 15   | 6   | 40,0 | 363    | 333    | 141   | 38,8 | 141   | 42,3 | 152   | 41,9 | 143   | 42,9 |
| 1998–2002                                          | 66   | 15  | 22,7 | 3103   | 2848   | 1443  | 46,5 | 1443  | 50,7 | 1612  | 51,9 | 1471  | 51,7 |
| 2003–2007                                          | 65   | 28  | 43,1 | 5650   | 4328   | 2970  | 52,6 | 2911  | 67,3 | 3910  | 69,2 | 2984  | 68,9 |
| 2008–2012                                          | 165  | 82  | 49,7 | 14832  | 10828  | 7957  | 53,6 | 7709  | 71,2 | 11217 | 75,6 | 7914  | 73,1 |
| 2013–2017                                          | 459  | 167 | 36,4 | 36291  | 26980  | 16823 | 46,4 | 16384 | 60,7 | 22624 | 62,3 | 16974 | 62,9 |
| 2017–2020                                          | 478  | 157 | 32,9 | 30832  | 25041  | 13509 | 43,8 | 13300 | 53,1 | 17708 | 57,4 | 13792 | 55,1 |

Таблица 8

**Влияние общественных наук (SSCI) на искусство и гуманитарные науки (A&HCI)  
между базами данных основных указателей цитирования Web of Science в 1993–2020 гг.,  
определенное по ссылкам**

| Сопоставляемые базы данных SSCI → A&HSCI | Период<br>(годы)                                   |                                           |                                                         |                     |               |                          |                                                       |                                                                 |                                                     |                                                               |                                                          |                                                                     |                                                        |                                                                   |
|------------------------------------------|----------------------------------------------------|-------------------------------------------|---------------------------------------------------------|---------------------|---------------|--------------------------|-------------------------------------------------------|-----------------------------------------------------------------|-----------------------------------------------------|---------------------------------------------------------------|----------------------------------------------------------|---------------------------------------------------------------------|--------------------------------------------------------|-------------------------------------------------------------------|
|                                          |                                                    | Массив публикаций-источников<br>в БД SSCI | Подмассив совместных с БД SSCI<br>публикаций-источников | Доля подмассива (%) | Массив ссылок | Массив уникальных ссылок | Подмассив неуникальных<br>совместных ссылок 1-го рода | Доля подмассива неуникальных<br>совместных ссылок 1-го рода (%) | Подмассив уникальных<br>совместных ссылок 1-го рода | Доля подмассива уникальных<br>совместных ссылок 1-го рода (%) | Объединенный подмассив<br>неуникальных совместных ссылок | Доля объединенного подмассива<br>неуникальных совместных ссылок (%) | Объединенный подмассив<br>уникальных совместных ссылок | Доля объединенного подмассива<br>уникальных совместных ссылок (%) |
|                                          | Все публикации российских ученых                   |                                           |                                                         |                     |               |                          |                                                       |                                                                 |                                                     |                                                               |                                                          |                                                                     |                                                        |                                                                   |
|                                          | 1993–1997                                          | 1569                                      | 26                                                      | 1,7                 | 31330         | 30574                    | 888                                                   | 2,8                                                             | 887                                                 | 2,9                                                           | 2768                                                     | 8,8                                                                 | 976                                                    | 38,0                                                              |
|                                          | 1998–2002                                          | 1305                                      | 34                                                      | 2,6                 | 23510         | 22732                    | 2928                                                  | 12,5                                                            | 2924                                                | 12,9                                                          | 6279                                                     | 26,7                                                                | 3062                                                   | 40,0                                                              |
|                                          | 2003–2007                                          | 1058                                      | 72                                                      | 6,8                 | 24555         | 22439                    | 4804                                                  | 19,6                                                            | 4675                                                | 20,8                                                          | 14303                                                    | 58,2                                                                | 4735                                                   | 46,8                                                              |
|                                          | 2008–2012                                          | 2184                                      | 141                                                     | 6,5                 | 61052         | 55015                    | 9139                                                  | 15,0                                                            | 8750                                                | 15,9                                                          | 24252                                                    | 39,7                                                                | 9060                                                   | 43,1                                                              |
|                                          | 2013–2017                                          | 4049                                      | 242                                                     | 6,0                 | 143493        | 126277                   | 22909                                                 | 16,0                                                            | 21802                                               | 17,3                                                          | 60569                                                    | 42,2                                                                | 22918                                                  | 48,7                                                              |
|                                          | 2017–2020                                          | 4630                                      | 118                                                     | 2,6                 | 125914        | 115280                   | 11442                                                 | 9,1                                                             | 11308                                               | 9,8                                                           | 36538                                                    | 29,0                                                                | 12715                                                  | 58,1                                                              |
|                                          | Публикации только российских ученых                |                                           |                                                         |                     |               |                          |                                                       |                                                                 |                                                     |                                                               |                                                          |                                                                     |                                                        |                                                                   |
|                                          | 1993–1997                                          | 1535                                      | 24                                                      | 1,6                 | 30298         | 29622                    | 878                                                   | 2,9                                                             | 878                                                 | 3,0                                                           | 2384                                                     | 7,9                                                                 | 946                                                    | 3,2                                                               |
|                                          | 1998–2002                                          | 1238                                      | 22                                                      | 1,8                 | 20385         | 19892                    | 2023                                                  | 9,9                                                             | 2021                                                | 10,2                                                          | 3429                                                     | 16,8                                                                | 2072                                                   | 10,4                                                              |
|                                          | 2003–2007                                          | 990                                       | 30                                                      | 3,0                 | 18430         | 17729                    | 1196                                                  | 6,5                                                             | 1188                                                | 6,7                                                           | 3247                                                     | 17,6                                                                | 1250                                                   | 7,1                                                               |
|                                          | 2008–2012                                          | 1985                                      | 72                                                      | 3,6                 | 44088         | 42228                    | 3908                                                  | 8,9                                                             | 3872                                                | 9,2                                                           | 9390                                                     | 21,3                                                                | 4002                                                   | 9,5                                                               |
|                                          | 2013–2017                                          | 3480                                      | 110                                                     | 3,2                 | 100551        | 94705                    | 9099                                                  | 9,0                                                             | 8902                                                | 9,4                                                           | 21914                                                    | 21,8                                                                | 9195                                                   | 9,7                                                               |
|                                          | 2017–2020                                          | 4144                                      | 54                                                      | 1,3                 | 94873         | 90334                    | 5465                                                  | 5,8                                                             | 5428                                                | 6,0                                                           | 14368                                                    | 15,1                                                                | 5886                                                   | 6,5                                                               |
|                                          | Публикации российских ученых с участием зарубежных |                                           |                                                         |                     |               |                          |                                                       |                                                                 |                                                     |                                                               |                                                          |                                                                     |                                                        |                                                                   |
|                                          | 1993–1997                                          | 15                                        | 6                                                       | 40,0                | 363           | 333                      | 141                                                   | 38,8                                                            | 141                                                 | 42,3                                                          | 152                                                      | 41,9                                                                | 143                                                    | 42,9                                                              |
|                                          | 1998–2002                                          | 66                                        | 15                                                      | 22,7                | 3103          | 2848                     | 1443                                                  | 46,5                                                            | 1443                                                | 50,7                                                          | 1612                                                     | 51,9                                                                | 1471                                                   | 51,7                                                              |
|                                          | 2003–2007                                          | 65                                        | 28                                                      | 43,1                | 5650          | 4328                     | 2970                                                  | 52,6                                                            | 2911                                                | 67,3                                                          | 3910                                                     | 69,2                                                                | 2984                                                   | 68,9                                                              |
|                                          | 2008–2012                                          | 165                                       | 82                                                      | 49,7                | 14832         | 10828                    | 7957                                                  | 53,6                                                            | 7709                                                | 71,2                                                          | 11217                                                    | 75,6                                                                | 7914                                                   | 73,1                                                              |
|                                          | 2013–2017                                          | 459                                       | 167                                                     | 36,4                | 36291         | 26980                    | 16823                                                 | 46,4                                                            | 16384                                               | 60,7                                                          | 22624                                                    | 62,3                                                                | 16974                                                  | 62,9                                                              |
| 2017–2020                                | 478                                                | 157                                       | 32,9                                                    | 30832               | 25041         | 13509                    | 43,8                                                  | 13300                                                           | 53,1                                                | 17708                                                         | 57,4                                                     | 13792                                                               | 55,1                                                   |                                                                   |

В каждой из табл. 6–9 содержатся данные, соответствующие только одной паре баз данных, а строки в разделах каждой таблицы отличаются друг от друга только периодом охвата, начиная с 1993 г.: 1993–1997 гг., 1998–2002 гг., 2003–2007 гг., 2008–2012 гг., 2013–2017 гг. И только последний период охватывает

2018–2020 гг. Анализируя результаты, представленные в этих таблицах, мы исходим из следующих положений. Базы данных SCIE (журнальные публикации), CPCI-S (материалы конференций) и BKCI-S (книги и отдельные разделы книг) соответствуют области знания «Естественные, точные и прикладные

науки». Базы данных *SSCI* (журнальные публикации), *CPCI-SSH* (материалы конференций) и *BKCI-SSH* (книги и отдельные разделы книг) – области знания «Общественные науки», а две последние (*CPCI-SSH* и *BKCI-SSH*) – также и области знания «Гуманитарные науки и искусство», журнальные публикации по этой области знания отражаются в БД *A&HCI*.

Взаимовлияние, взаимосвязь между областями знания, а также между разделами и направлениями исследований этих областей, может осуществляться по различным каналам. Такими каналами, причем очень важными и различными по своим возможностям и функциям, являются как публикации-источники, так и ссылки. Показатели вида «Доля совместных публикаций» сопоставляемых пар БД характеризуют степень связи (степень взаимовлияния) тех областей знания,

которые представляют БД из данной пары. При этом чем выше значение показателя, тем выше взаимовлияние областей знания, представляемых данной парой БД. Значения этих показателей могут служить в качестве оценки степени влияния (воздействия) каждого из таких каналов на взаимосвязи областей знания и их крупных разделов.

Рассматривая показатели вида «Доля совместных публикаций», можно получить два вида оценок, дополняющих друг друга: с одной стороны, определить степень взаимосвязи различных областей знания, а с другой – степень их влияния друг на друга, т. е. мощность, которой характеризуется тот или иной канал влияния. Поэтому при анализе сведений табл. 6–9 нужно уделять особое внимание показателям вида «Доля совместных публикаций».

Таблица 9

**Влияние естественных и прикладных наук (SCIE) на искусство и гуманитарные науки (A&HCI) между базами данных основных указателей цитирования Web of Science в 1993–2020 гг., определенное по ссылкам**

| Сопоставляемые базы данных SCIE → A&HSCI                        | Период<br>(годы)                                                      |                                                  |                                        |                                   |               |                          |                                                       |                                                                 |                                                     |                                                               |                                                          |                                                                     |                                                        |                                                               |
|-----------------------------------------------------------------|-----------------------------------------------------------------------|--------------------------------------------------|----------------------------------------|-----------------------------------|---------------|--------------------------|-------------------------------------------------------|-----------------------------------------------------------------|-----------------------------------------------------|---------------------------------------------------------------|----------------------------------------------------------|---------------------------------------------------------------------|--------------------------------------------------------|---------------------------------------------------------------|
|                                                                 |                                                                       | Массив публикаций-источников (ПИ)<br>в БД A&HSCI | Подмассив совместных<br>с БД A&HSCI ПИ | Доля подмассива совместных ПИ (%) | Массив ссылок | Массив уникальных ссылок | Подмассив неуникальных совместных<br>ссылок 1-го рода | Доля подмассива неуникальных<br>совместных ссылок 1-го рода (%) | Подмассив уникальных совместных<br>ссылок 1-го рода | Доля подмассива уникальных<br>совместных ссылок 1-го рода (%) | Объединенный подмассив<br>неуникальных совместных ссылок | Доля объединенного подмассива<br>неуникальных совместных ссылок (%) | Объединенный подмассив<br>уникальных совместных ссылок | Доля объединенного подмассива<br>уникальных совместных ссылок |
|                                                                 | Публикации всех российских ученых с участием и без участия зарубежных |                                                  |                                        |                                   |               |                          |                                                       |                                                                 |                                                     |                                                               |                                                          |                                                                     |                                                        |                                                               |
|                                                                 | 1993–1997                                                             | 1569                                             | 26                                     | 1,7                               | 31330         | 30574                    | 888                                                   | 2,8                                                             | 887                                                 | 2,9                                                           | 2768                                                     | 8,8                                                                 | 976                                                    | 3,2                                                           |
|                                                                 | 1998–2002                                                             | 1305                                             | 34                                     | 2,6                               | 23510         | 22732                    | 2928                                                  | 12,5                                                            | 2924                                                | 12,9                                                          | 6279                                                     | 26,7                                                                | 3062                                                   | 13,5                                                          |
|                                                                 | 2003–2007                                                             | 1058                                             | 72                                     | 6,8                               | 24555         | 22439                    | 4804                                                  | 19,6                                                            | 4675                                                | 20,8                                                          | 14303                                                    | 58,2                                                                | 4735                                                   | 21,1                                                          |
|                                                                 | 2008–2012                                                             | 2184                                             | 141                                    | 6,5                               | 61052         | 55015                    | 9139                                                  | 15,0                                                            | 8750                                                | 15,9                                                          | 24252                                                    | 39,7                                                                | 9060                                                   | 16,5                                                          |
|                                                                 | 2013–2017                                                             | 4049                                             | 242                                    | 6,0                               | 143493        | 126277                   | 22909                                                 | 16,0                                                            | 21802                                               | 17,3                                                          | 60569                                                    | 42,2                                                                | 22918                                                  | 18,1                                                          |
|                                                                 | 2017–2020                                                             | 4630                                             | 118                                    | 2,6                               | 125914        | 115280                   | 11442                                                 | 9,1                                                             | 11308                                               | 9,8                                                           | 36538                                                    | 29,0                                                                | 12715                                                  | 11,0                                                          |
|                                                                 | Публикации только российских ученых                                   |                                                  |                                        |                                   |               |                          |                                                       |                                                                 |                                                     |                                                               |                                                          |                                                                     |                                                        |                                                               |
|                                                                 | 1993–1997                                                             | 1535                                             | 24                                     | 1,6                               | 30298         | 29622                    | 878                                                   | 2,9                                                             | 878                                                 | 3,0                                                           | 2384                                                     | 7,9                                                                 | 946                                                    | 3,2                                                           |
|                                                                 | 1998–2002                                                             | 1238                                             | 22                                     | 1,8                               | 20385         | 19892                    | 2023                                                  | 9,9                                                             | 2021                                                | 10,2                                                          | 3429                                                     | 16,8                                                                | 2072                                                   | 10,4                                                          |
|                                                                 | 2003–2007                                                             | 990                                              | 30                                     | 3,0                               | 18430         | 17729                    | 1196                                                  | 6,5                                                             | 1188                                                | 6,7                                                           | 3247                                                     | 17,6                                                                | 1250                                                   | 7,1                                                           |
|                                                                 | 2008–2012                                                             | 1985                                             | 72                                     | 3,6                               | 44088         | 42228                    | 3908                                                  | 8,9                                                             | 3872                                                | 9,2                                                           | 9390                                                     | 21,3                                                                | 4002                                                   | 9,5                                                           |
|                                                                 | 2013–2017                                                             | 3480                                             | 110                                    | 3,2                               | 100551        | 94705                    | 9099                                                  | 9,0                                                             | 8902                                                | 9,4                                                           | 21914                                                    | 21,8                                                                | 9195                                                   | 9,7                                                           |
| 2017–2020                                                       | 4144                                                                  | 54                                               | 1,3                                    | 94873                             | 90334         | 5465                     | 5,8                                                   | 5428                                                            | 6,0                                                 | 14368                                                         | 15,1                                                     | 5886                                                                | 6,5                                                    |                                                               |
| Публикации российских ученых с обязательным участием зарубежных |                                                                       |                                                  |                                        |                                   |               |                          |                                                       |                                                                 |                                                     |                                                               |                                                          |                                                                     |                                                        |                                                               |
| 1993–1997                                                       | 34                                                                    | 2                                                | 5,9                                    | 1032                              | 952           | 10                       | 1,0                                                   | 9                                                               | 0,9                                                 | 384                                                           | 37,2                                                     | 30                                                                  | 3,2                                                    |                                                               |
| 1998–2002                                                       | 47                                                                    | 12                                               | 25,5                                   | 3125                              | 2840          | 905                      | 29,0                                                  | 903                                                             | 31,8                                                | 2850                                                          | 91,2                                                     | 990                                                                 | 34,9                                                   |                                                               |
| 2003–2007                                                       | 65                                                                    | 24                                               | 36,9                                   | 5650                              | 4328          | 1848                     | 32,7                                                  | 1808                                                            | 41,8                                                | 10090                                                         | 178,6                                                    | 2014                                                                | 46,5                                                   |                                                               |
| 2008–2012                                                       | 165                                                                   | 65                                               | 39,4                                   | 14832                             | 10828         | 5040                     | 34,0                                                  | 4789                                                            | 44,2                                                | 16078                                                         | 108,4                                                    | 4996                                                                | 46,1                                                   |                                                               |
| 2013–2017                                                       | 459                                                                   | 130                                              | 28,3                                   | 36291                             | 26980         | 13555                    | 37,4                                                  | 12810                                                           | 47,5                                                | 39620                                                         | 109,2                                                    | 13214                                                               | 49,0                                                   |                                                               |
| 2017–2020                                                       | 478                                                                   | 64                                               | 13,4                                   | 30832                             | 25041         | 5977                     | 19,4                                                  | 5896                                                            | 23,5                                                | 23639                                                         | 76,7                                                     | 6746                                                                | 26,9                                                   |                                                               |

## ЗАКЛЮЧЕНИЕ

Влияние естественных и прикладных наук на общественные, а также взаимосвязь и взаимное влияние общественных и гуманитарных наук и искусства на протяжении 28 лет неизменно и устойчиво возрастали во всех трех категориях (все публикации российских ученых, только российских ученых и с зарубежными учеными). Исключение составляет только «Доля объединенного подмассива не уникальных совместных ссылок (%)», которая росла не очень стабильно по пятилетиям, что может объясняться различиями в методике *Web of Science* при учёте несовместных (совпадающих) ссылок.

Все остальные графы табл. 6–9 свидетельствуют об усилении междисциплинарности научных исследований, поскольку все больше положительных результатов получается на стыке наук. Интересно отметить в табл. 9, что в 1983–1997 гг. доля объединенного подмассива уникальных совместных ссылок (крайняя правая графа) составляла по всем верхним строкам 3,2 %, т. е. что российские ученые естественных и прикладных наук в это время еще не уделяли должного внимания гуманитарным наукам и мало сотрудничали с зарубежными коллегами. В 2017–2020 гг. эта доля увеличилась – стало больше исследований с участием зарубежных ученых, меньше без них. Это означает, что роль гуманитарных наук в творчестве российских ученых-естественников возрастает, особенно в сотрудничестве с зарубежными коллегами. То же самое (только с другими значениями долей) можно сказать и о влиянии общественных наук на искусство и гуманитарные науки, определенное по ссылкам между базами данных основных указателей цитирования *Web of Science* (табл. 8). Можно предположить, что это связано с успехами структурной лингвистики и семиотики языка, которые оказались важны для представителей многих наук.

Совершенно иная картина наблюдается относительно влияния искусства и гуманитарных наук на общественные науки. Из табл. 7 (крайняя правая графа) видно, что доля объединенного подмассива уникальных совместных ссылок неизменно уменьшается от одного пятилетия к другому, причем резче всего в публикациях отечественных авторов вместе с зарубежными. Что касается влияния естественных и прикладных наук на общественные (см. табл. 6), то оно находится в русле общей тенденции постепенного увеличения. При необходимости почти каждая графа табл. 6–9 может быть проанализирована специалистами любой из этих крупных отраслей науки с гипотезами о возможных причинах динамики их взаимовлияния. Таким образом, общее представление о прогрессе науки и панораме ее достижений перемещается из привычной иерархии наук к кластерам библиографических сочетаний различных цитируемых публикаций-источников (т. е. содержащихся в них ссылках на одни и те же более ранние публика-

ции) и соцированию между цитируемыми публикациями (т. е. совпадающих в каждой из них ссылок на более поздние публикации). Эта тенденция идет от карт и атласов науки, составлявшихся Ю. Гарфилдом еще в 1980-е гг., когда эти кластеры размещались в достаточно узких предметных областях и обозначались именами наиболее часто цитируемых авторов. Современные *Citation Topics* нужно называть, и этот способ их обозначения таит в себе угрозу необходимости создавать тематическую классификацию этих названий.

Другой вывод из этого трудоемкого исследования заключается в том, что статистика библиографических ссылок может служить средством для оценки достижений научных работников и их коллективов, по крайней мере, при сопоставлении крупных разделов современной науки. Что касается более мелких подразделений науки, например, на уровне тематических категорий на платформах индексации и цитирования, то это необходимо проверить в дальнейших исследованиях.

## СПИСОК ЛИТЕРАТУРЫ

1. Гиляревский Р.С. О некорректности использования индексов цитирования для вычислений по сопоставлению разделов науки // Научно-техническая информация. Сер. 2. – 2022. – № 2. – С. 21–24; Gilyarevskii R.S. On the incorrect use of citation indexes for assessment by comparison of science divisions // Automatic Documentation and Mathematical Linguistics. – 2022. – Vol. 56, № 1. – P. 26–29.
2. Bradford S.C. Documentation. – London: Lockwood, 1948. – 156 p.

*Материал поступил в редакцию 13.02.23.*

### Сведения об авторах

**ГИЛЯРЕВСКИЙ Руджеро Сергеевич** – доктор филологических наук, профессор, заведующий Отделением Всероссийского института научной и технической информации Российской академии наук (ВИНИТИ РАН); профессор кафедры новых медиа и теории коммуникации факультета журналистики МГУ им. М.В. Ломоносова, Москва  
e-mail: ruggero29@gmail.com

**ЛИБКИНД Александр Наумович** – кандидат технических наук, ведущий научный сотрудник ВИНТИ РАН  
e-mail: anliberty@mail.ru

**ЛИБКИНД Илья Александрович** – ведущий аналитик, ООО Сервисное бюро ВИП, Москва  
e-mail: ilya libkind@hotmail.com

О.П. Шестерникова, В.К. Финн, К.А. Лесько, Л.В. Винокурова

## О применении ДСМ-метода автоматизированной поддержки исследований для предсказания развития сахарного диабета у больных хроническим панкреатитом

*Описано применение алгоритмических средств нахождения эмпирических закономерностей в ДСМ-методе автоматизированной поддержки исследований к массиву пациентов для исследования сахарного диабета третьего типа. Впервые для нахождения эмпирических закономерностей используется обобщенный метод, представляющий тернарный предикат причинности: причина — множество тормозов причины — эффект.*

**Ключевые слова:** ДСМ-метод автоматизированной поддержки исследований (ДСМ-метод АПИ), эмпирические закономерности, сахарный диабет, панкреатит

DOI: 10.36535/0548-0027-2023-04-2

### ВВЕДЕНИЕ

Изучение сахарного диабета третьего типа (сахарного диабета панкреатогенного генеза – СДЗс) с помощью ДСМ-метода автоматизированной поддержки исследований (ДСМ-метода АПИ) проводится Федеральным исследовательским центром «Информатика и управление» (ФИЦ ИУ РАН) совместно с Московским клиническим научным центром имени А.С. Логинова (ГБУЗ МНКЦ им. А.С. Логинова). На базе этого исследования были сформулированы и опробованы многие алгоритмические методы, образующие в настоящий момент современный подход к исследованию в ДСМ-методе АПИ. Первая статья, посвященная изучению СДЗс, была опубликована в 2015 г. [1]; за ней последовала череда публикаций, в которых уточнялись предложенные алгоритмические методы, описывалось и оценивалось их применение [2–5].

Главной особенностью разрабатываемого ДСМ-метода АПИ стало его применение к открытым (постоянно пополняемым) данным, представляемым в виде расширяющихся баз фактов (БФ), получившее название ДСМ-исследования, в то время как ранее используемое понятие ДСМ-рассуждение, являющееся синтезом процедур индукции, аналогии и абдукции, закрепилось за применением процедур ДСМ-метода АПИ к одной базе фактов. В ДСМ-методе автоматизированной поддержки исследований было формализовано понятие эмпирической закономерности. Нахождение эмпирических закономерностей (ЭК) – одно из главных результатов проведения исследова-

ния. Кроме того, в рамках ДСМ-метода АПИ были разработаны шкалы оценки получаемых гипотез о причинах, ЭЗК и ДСМ-исследования в целом [5].

Впоследствии развитие методов ДСМ-исследования на примере онкологических задач в гастроэнтерологии осуществлялось ФИЦ ИУ РАН также совместно с ГБУЗ МНКЦ им. А.С. Логинова [6, 7]. Спустя некоторое время представилось целесообразным вернуться к исходной задаче диагностирования СДЗс для апробации уточнённых методов и сравнения новых результатов с предыдущими.

Второй причиной возвращения к исследованию СДЗс стало развитие инструментальных средств обобщенного метода. Его новый вариант был представлен в [8]. В некоторых публикациях, посвященных интеллектуальной системе для исследования СДЗс, выдвигалось предположение о возможном успешном применении этого метода для исследуемой задачи, подтверждаемое, в частности, положительным исходом теста на обобщенный метод [3, 4]. Таким образом, в настоящем исследовании нами впервые было выполнено применение обобщенного метода к расширяющимся массивам данных.

### ОПИСАНИЕ ЗАДАЧИ ИССЛЕДОВАНИЯ

Поджелудочная железа с функциональной точки зрения состоит из экзокринной и эндокринной частей, деятельность которых взаимосвязана и оказывает существенное влияние на жизнедеятельность организма. При этом экзокринный компонент занимает

около 85% ткани органа [9]. Большая часть неэндокринных заболеваний поджелудочной железы представлена воспалительной патологией – острый и хронический панкреатит (ХП), а также опухолями – в первую очередь раком поджелудочной железы.

Хронический панкреатит приводит к значимому снижению качества жизни и целому спектру тяжелых патологических состояний, среди которых сахарный диабет (СД) и экзокринная недостаточность поджелудочной железы. Необходимость в постоянной поддерживающей терапии и периодической госпитализации пациентов с хроническим панкреатитом создает дополнительную экономическую нагрузку на систему здравоохранения.

Поражение экзокринной части поджелудочной железы при хроническом панкреатите и раке влияет и на эндокринную часть, что проявляется в развитии сахарного диабета третьего типа. Эта форма сахарного диабета, в свою очередь, значительно усиливает тяжесть течения хронического панкреатита и ухудшает прогноз течения этого заболевания [10].

Сахарный диабет третьего типа может быть ошибочно диагностирован как сахарный диабет второго типа (СД 2), что приводит к плохому гликемическому контролю. Помимо этого, следствием поздней диагностики сахарного диабета может быть развитие тяжелых жизнеугрожающих состояний – поражения почек, глаз, опорно-двигательной, сердечно-сосудистой и центральной нервной систем, а также повышение уровня смертности трудоспособного населения от таких заболеваний и патологических состояний как инфаркт миокарда и инсульт [11].

Значимость своевременной и точной диагностики сахарного диабета третьего типа обусловлена тем, что такая форма сахарного диабета может иметь клинические или лабораторные проявления еще до манифестации основного заболевания поджелудочной железы, вызвавшего ее. Это особенно важно для своевременной диагностики рака поджелудочной железы. Однако обилие признаков и сложность их взаимодействий делает крайне затруднительным обобщение их врачом-клиницистом, что может приводить к диагностическим ошибкам и выбору неверной тактики лечения пациентов [10, 12]. Все это обуславливает необходимость автоматизации и применения интеллектуальных систем для анализа медицинских данных.

В связи с вышеизложенным поставленная нами цель является важной с медицинской и социальной точек зрения. База фактов, подготовленная для исследования предсказания развития СДЗс у пациентов с хроническим панкреатитом с использованием ДСМ-метода АПИ, включает важнейшие признаки, позволяющие точно диагностировать сахарный диабет третьего типа и прогнозировать тяжесть его течения.

## БАЗА ФАКТОВ ДЛЯ ИССЛЕДОВАНИЯ

В основе настоящего исследования лежит база фактов, используемая в предыдущих версиях [3, Приложение 1]. Из неё были удалены три признака, отвечающие за наличие мутаций в генах *CFTR*, *PRSS*, *SPINK*. Таким образом, примеры из актуальной базы фактов содержали 36 признаков, представляющие собой клинические и лабораторные показатели.

Общее количество записей в данных – 163, из них 15 пациентов имеют диагноз СД 2, 38 пациентов – диагноз СДЗс, 100 пациентов не имеют диабета. Для предсказания были добавлены 10 пациентов, из которых четыре имели диагноз СДЗс, а шесть не имели диабета. Пациенты с диагнозом СДЗс стали положительными примерами в исследовании (всего 38); пациенты с диагнозом СД 2 и пациенты без диабета – отрицательными (всего 115).

Для моделирования постоянно пополняемой базы фактов и для обнаружения эмпирических закономерностей имеющиеся данные мы разбили на три непересекающихся множества, близкие по размерности:

- 1) множество А (количество положительных примеров 12, отрицательных – 38);
- 2) множество В1 (количество положительных примеров 13, отрицательных – 38);
- 3) множество В2 (количество положительных примеров 13, отрицательных – 39).

На основе перестановок этих множеств были созданы шесть расширяющихся последовательностей – «историй возможных миров» [5]. В предыдущих исследованиях использовались «естественные» расширения (т. е. расширения, образованные в порядке сбора данных экспертами), которые значительно отличались по количеству входящих в них примеров. В настоящем исследовании размерность элементов перестановок была выровнена. Проверка сохранения эмпирических закономерностей во всех историях возможных миров используется для минимизации случайности расширения.

## МЕТОДЫ, ИСПОЛЬЗУЕМЫЕ В ИССЛЕДОВАНИИ

Используемые нами методы можно разделить на две большие группы: 1) те, которые представляют результат в виде бинарного предиката для пар вида  $\langle V, W \rangle$  ( $V$  – переменная для обозначения причины,  $W$  – переменная для обозначения эффекта), и 2) те, которые представляют результат в виде тернарного предиката для троек вида  $\langle V, X, W \rangle$  ( $V$  – переменная для обозначения причины,  $W$  – переменная для обозначения эффекта,  $X$  – переменная для множества тормозов (блокаторов) причины).

Первая группа повторяет методы, использованные ранее в исследовании целесообразности применения компьютерной томографии [7]:

1. Для обнаружения эмпирических закономерностей применялись **16 стратегий**. Стратегия характеризует тип причинности в предметной области (подразумевается возможность существования различных типов причинности в предметной области). Стратегия определяется именами двух предикатов сходства: для положительных (обозначается  $x$ ) и для отрицательных (обозначается  $y$ ) примеров. В исследовании использовались четыре предиката сходства: простое сходство (обозначается  $a$ ), простое сходство с запретом на контрпример (обозначается  $ab$ ), метод различия (обозначается  $ad_0$ ) и метод различия с запретом на контрпример (обозначается  $abd_0$ ). Стратегии обозначаются  $Str_{x,y}$ , где  $x, y \in \{a, ab, ad_0, abd_0\}$  (пример,  $Str_{a,a}$  – стратегия с простым сходством и для положительных и для отрицательных примеров).

Стратегии образуют изоморфные дистрибутивные решетки с частичным порядком, определяемым посредством логического следования, для правил индуктивного вывода (разные для положительных и отрицательных гипотез) [13]. Таким образом, для сравнимых стратегий можно указать более «сильную». Гипотезы, порождаемые более «сильными» стратегиями, представляются более надёжными. Краткое формальное описание стратегий приведено в *Приложении 1*.

Минимальные<sup>1</sup> из полученных гипотез о причине были предоставлены для комментария экспертам.

2. Для оценки полученных эмпирических закономерностей использовалась **шкала  $\langle z, Str, k \rangle$**  [7], где  $z$  – тип закономерности,  $Str$  – стратегия из множества возможных 16 стратегий,  $k$  – количество родителей гипотезы. Тип закономерности (определяется согласно дереву Т-классификаций в [5], обозначается латинскими буквами от **a** до **n**) характеризует устойчивость гипотезы при расширении данных. Количество родителей – важная характеристика гипотезы, отражающая частотность полученной гипотезы как механизма в исходном массиве данных.

3. Ещё одним результатом ДСМ-исследования является **разбиение множества стратегий** относительно отношений эквивалентности и построение классов эквивалентности. Были использованы три типа эквивалентности: совпадение правильных предсказаний (обозначается  $E_1$ ), совпадение правильных предсказаний и ошибок (обозначается  $E_{13}$ ) и совпадение эмпирических закономерностей (обозначается  $E_2$ ). Определяемые отношения эквивалентности образуют решетку, изображенную на рис. 1.

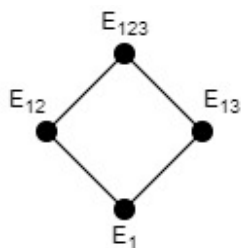


Рис. 1. Решётка отношений эквивалентности стратегий

Для оценки ДСМ-исследования используются два показателя: монотонность объясняемости и предсказания.

Объясняемость – это процент исходного массива данных, объясняемый порождаемыми гипотезами о причинах. Этот показатель рассчитывается для ДСМ-рассуждения и является основой абдуктивного принятия порождаемых гипотез. Объясняемость должна быть не меньше 0,8 (в долях от единицы). В ДСМ-исследовании проверяется монотонное неубывание объясняемости: проверка того, что при пополнении массива новыми примерами показатель объясняемости не уменьшается.

<sup>1</sup> Под минимальными подразумеваются гипотезы, такие, что никакая другая полученная гипотеза не состоит из подмножества признаков этой гипотезы; формальное описание минимальных гипотез представлено, например, в [14].

Предсказания – это те примеры из заданного заранее множества неопределенностей, которые мы предсказываем на основе полученных гипотез о причинах. Результатом предсказания будет либо отнесение к положительным примерам, либо отнесение к отрицательным, либо фактическое противоречие (отнесение сразу к обеим группам), либо невозможность предсказания, отказ от предсказания). Кроме этого, предсказания можно верифицировать. Очевидно, что идеальный случай для предсказания – правильное предсказание положительных и отрицательных случаев без противоречий и отказов от предсказаний. Все предсказания можно классифицировать либо как правильные (обозначаются  $l_0$ ), либо как ошибки, которые, в свою очередь, могут быть трёх типов: неверные предсказания (предсказание положительного примера как отрицательного и наоборот, обозначаются  $a$ ); противоречивые предсказания (предсказание как фактическое противоречие, обозначаются  $b$ ); невозможность предсказания (отказ от предсказания, обозначаются  $c$ ). Предсказания и классификация их правильности (обычно представляемые в виде таблицы) относятся к параметрам ДСМ-рассуждения; в ДСМ-исследовании рассматривается таблица правильных предсказаний и ошибок для последнего расширения, которое является одинаковым во всех историях возможных миров.

**Вторая группа** методов (представление результата в виде тернарного предиката) включает обобщенный метод (обозначается  $ag$ ), описанный в [8]. В этом исследовании были использованы три типа стратегий с обобщенным методом: в первом типе обобщенный метод применялся только для положительных гипотез (для отрицательных использовалось простое сходство), во втором – только для отрицательных (для положительных использовалось простое сходство), в третьем – и для положительных, и для отрицательных. Будем их обозначать как  $\langle ag, a \rangle$ ,  $\langle a, ag \rangle$  и  $\langle ag, ag \rangle$ .

Для нахождения эмпирических закономерностей, представленных тернарным предикатом, использовался тот же алгоритм, что и для бинарного предиката, т.е. рассматривалось только сохранение причины при расширении БФ, без учёта множества тормозов. Представляется интересным рассмотреть, как множество тормозов эмпирических закономерностей ведёт себя при расширении массива данных, однако это выходит за рамки настоящего исследования.

## ОПИСАНИЕ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ

Интеллектуальная система (ИС) является средством проведения исследования и важным компонентом ДСМ-метода АПИ, она имеет определенную архитектуру: 1) база фактов + база знаний (БФ + БЗ), 2) Решатель и 3) интерфейс пользователя. Реализация этой системы для данной задачи как программного обеспечения было основано на ИС для диагностики заболеваний поджелудочной железы, подробно описанной в [6], и доработанной позднее в [7]. Перечислим основные компоненты этой реализации (рис. 2).

1. Основой программной системы является инфраструктурный компонент, который, в свою очередь, включает различные компоненты, необходимые для запуска и корректной работы программного обеспе-

чения. Сюда входят как язык программирования и среда выполнения, так и различные программные модули, обеспечивающие современные подходы (паттерны) к созданию приложения (логирование, DI-контейнеры и др.) [15].

2. Компонент, соответствующий базе фактов, состоит из:

1) непосредственно базы данных (БД) с исходными данными для системы, подготовленной экспертами с использованием стороннего программного обеспечения, например системы управления базами данных (СУБД);

2) провайдера, который читает данные из БД и переводит их во внутренние объекты системы;

3) процедур подготовки БФ, удовлетворяющих условиям применения ДСМ-метода АПИ [16], из этих объектов.

В качестве базы данных в системе используется таблица в формате *Office Open XML (xlsx)*.

3. Компонент, соответствующий Решателю – это модули, реализующие процедуры ДСМ-метода АПИ.

4. Компонент, соответствующий интерфейсу пользователя, содержит как графический интерфейс пользователя (GUI), предоставляемый средой выполнения, так и процедуры для представления, анализа и вывода результатов (например, фильтр только минимальных гипотез или вывод в табличный редактор).

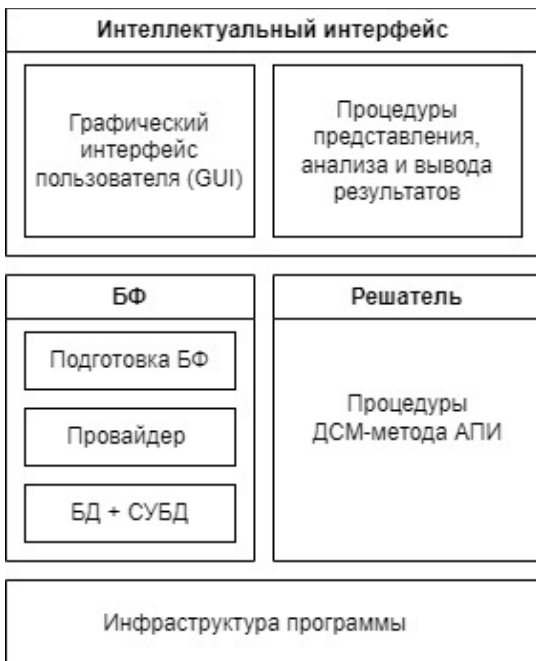


Рис. 2. Структура реализации интеллектуальной системы

Для исследования сахарного диабета третьего типа в исходную систему были внесены два основных изменения: первое касалось представления БФ – ее было необходимо преобразовать из специального формата [3, 4] в табличный; второе изменение предусматривало расширение алгоритмических средств Решателя в части обобщенного метода. Для нахождения гипотез при обобщенном методе был использован алгоритм, соответствующий предикату сходства

обобщенного метода  $M_{ag}^{\sigma} (\sigma \in \{+, -\})$  [8] для фиксированного эффекта, обозначаемого константой Q.

1. Для положительных и отрицательных примеров отдельно находим сходства.

2. Для каждого сходства S находим примеры противоположного знака, которые его включают, обозначим это множество примеров M.

3. Если множество M пусто ( $M = \emptyset$ ), то сходство S не является причиной (соответствует одному из условий, по которому множество тормозов должно быть не пусто).

4. Если множество M не пусто ( $M \neq \emptyset$ ), то ищем множество тормозов Bг для возможной причины S. Для этого формируем множество  $M^*$ , включающее разности элементов множества M и сходства S.

4.1. Находим все возможные сходства элементов из множества  $M^*$ , обозначим это множество Bг\*.

4.2. Для каждого элемента Bг\* из множества Bг\* проверяем, входит ли Bг\* в какой-либо элемент множества «родителей» S. Если нет, то объявляем Bг\* «тормозом» и добавляем в множество Bг.

4.3. Если получившееся множество Bг не пусто ( $Bг \neq \emptyset$ ), то тройка  $\langle S, Bг, Q \rangle$  – выполняет предикат сходства  $M_{ag}^{\sigma}$ .

## РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

Кратко изложим полученные результаты.

1. На последнем расширении БФ было правильно предсказано семь из десяти примеров (правильные предсказания,  $l_0$ ), заданных для предсказания; остальные три были предсказаны как противоречия (ошибки типа b). Из правильно предсказанных семи примеров два пациента относятся к положительным и пять – к отрицательным. Из них оба положительных входят в некоторую эмпирическую закономерность, а из пяти отрицательных в нее входят три.

2. Наилучшие предсказания дают стратегии  $Str_{b, bd0}$  и  $Str_{bd0, bd0}$  (эквивалентны по правильным предсказаниям и ошибкам, эквивалентность  $E_{13}$ ). В таблице приведены правильные предсказания и ошибки предсказаний для всех использованных стратегий.

Таблица ошибок предсказаний (отражает эквивалентность  $E_{13}$ )

|                                                        | $l_0$ |   | $a$ |   | $b$ |   | $c$ |   |
|--------------------------------------------------------|-------|---|-----|---|-----|---|-----|---|
|                                                        | +     | - | +   | - | +   | - | +   | - |
| $Str_{a, a}, Str_{a, d0}, Str_{d0, a}, Str_{d0, d0}$   |       | 1 |     |   | 4   | 5 |     |   |
| $Str_{a, b}, Str_{d0, b}$                              | 1     | 1 |     |   | 3   | 5 |     |   |
| $Str_{a, bd0}, Str_{d0, bd0}$                          | 2     | 1 |     |   | 2   | 5 |     |   |
| $Str_{b, a}, Str_{b, d0}, Str_{bd0, a}, Str_{bd0, d0}$ |       | 5 |     |   | 4   | 1 |     |   |
| $Str_{b, b}, Str_{bd0, b}$                             | 1     | 5 |     |   | 3   | 1 |     |   |
| $Str_{b, bd0}, Str_{bd0, bd0}$                         | 2     | 5 |     |   | 2   | 1 |     |   |

3. Объясняемость для ДСМ-рассуждения во всех возможных мирах имеет значение 1, т. е. присутствует монотонное неубывание во всех историях возможных миров.

4. Отрицательные эмпирические закономерности найдены всеми стратегиями, положительные –

только половиной стратегий, имеющих в отрицательном предикате условие с запретом на контрпримеры, т.е.  $Str_{a,ab}$ ,  $Str_{a,abd0}$ ,  $Str_{ab,ab}$ ,  $Str_{ab,abd0}$ ,  $Str_{ad0,ab}$ ,  $Str_{ad0,abd0}$ ,  $Str_{ad0,ab}$ ,  $Str_{ad0,abd0}$ , в общем виде  $Str_{x,by}$ . Найденные закономерности имеют тип **a**.

5. Все причины в отрицательных минимальных эмпирических закономерностях состоят из одного признака. Всего разными стратегиями было получено восемь минимальных отрицательных эмпирических закономерностей.

6. Разными стратегиями мы получили всего две минимальных положительных эмпирических закономерности. Обе состоят больше чем из одного признака.

7. Применение метода обнаружения эмпирических закономерностей для обобщенного метода дало одну отрицательную эмпирическую закономерность типа **a**. Для причины в этой закономерности был найден один тормоз.

8. На последнем расширении применение обобщенного метода дает два правильных отрицательных предсказания ( $l_0$ ); остальные восемь были предсказаны как фактические противоречия (ошибки типа  $b$ ). Эти два правильных предсказания входят в правильные предсказания, полученные наилучшими стратегиями (см. п. 2).

9. Были построены классы эквивалентности для отношений эквивалентности стратегий (приведены в *Приложении 2*) отдельно для: 1) положительных примеров для предсказания и положительных эмпирических закономерностей (обозначаются  $E_1^+$ ,  $E_2^+$ ,  $E_{13}^+$ ,  $E_{123}^+$ ); 2) отрицательных примеров для предсказания и отрицательных эмпирических закономерностей (обозначаются  $E_1^-$ ,  $E_2^-$ ,  $E_{13}^-$ ,  $E_{123}^-$ ).

Сравнение с результатами, представленными в статье 2015 г. Предсказания на последнем расширении, описанные в [1], имели лучшие показатели правильности и ошибок, чем предсказания в настоящем исследовании ( $l_0=8$ ,  $b=2$  в [1] и  $l_0=7$ ,  $b=3$  в текущем исследовании). Это может быть связано как с ограничением минимального порога родителей для гипотез  $k$ : в предыдущем исследовании пороги подбирались в зависимости от стратегии (15-18 – для положительных гипотез, 22-45 – для отрицательных), в настоящем исследовании минимальный порог везде был задан равным двум, так и с уменьшением числа признаков в БФ.

Сравнение с результатами, представленными в статье 2018 г. В статье [5, Приложение 7] были приведены эмпирические законы (ЭЗ) для двух стратегий  $Str_{a,a}$  и  $Str_{ab,ab}$ . Найденные отрицательные ЭЗ для стратегии  $Str_{ab,ab}$  совпали с эмпирическими закономерностями (ЭК), найденными в настоящем исследовании. Отрицательный ЭЗ, найденный для стратегии  $Str_{a,a}$ , также входит в список полученных ЭЗ в настоящем исследовании. Кроме того изменилась типизация этих ЭЗ: был тип **b**, а стал тип **a** за счет улучшения показателя монотонности роста объясняемости, что также может быть связано с уменьшением порога для родителей гипотезы.

Ниже будут приведены полученные эмпирические закономерности в текстовом виде и представлен комментарий эксперта-медика.

Всего получено две эмпирических закономерности о сахарном диабете третьего типа (СДЗс):

(1) Дефицит массы тела (16,1-18,5) И деформация панкреатического протока И увеличение размеров поджелудочной железы в головке И проведение инсулинотерапии И наличие алкоголя в анамнезе И наличие табакокурения в анамнезе И панкреатическая гипертензия И соматостатин от 0,02 до 0,42 И гастрин от 0,02 до 20 И с-пептид от 0,1 до 2 — все эти признаки вместе приводят к наличию СДЗс; это верно для 8 случаев из 38.

(2) Деформация панкреатического протока И значение с-пептид от 0,1 до 2 является причиной наличия СДЗс; это верно для 22 случаев из 38.

Очевидно, что вторая эмпирическая закономерность является более общим случаем, чем первая, и объединяет большее количество пациентов (22 vs 8). Однако, первую (более частную) можно считать надёжней, так как она порождается стратегиями, в положительном предикате которых есть условие с запретом на контрпримеры.

Эти закономерности предсказывают 2 из 4 положительных примеров, данных на предсказание.

Об отсутствии СДЗс получено восемь эмпирических закономерностей:

(1) Значение с-пептид от 0,1 до 2 является причиной отсутствия СДЗс; это верно для 27 случаев из 115.

(2) Экзокринная недостаточность (значение 1 по шкале) является причиной отсутствия СДЗс; это верно для 26 случаев из 115.

(3) Значение гастрина от 0,02 до 20 является причиной отсутствия СДЗс; это верно для 26 случаев из 115.

(4) Значение соматостатина от 0,02 до 0,42 является причиной отсутствия СДЗс; это верно для 25 случаев из 115.

(5) Возраст пациента от 41 года до 60 лет является причиной отсутствия СДЗс; это верно для 24 случаев из 115.

(6) Нормальный индекс массы тела (18,6-25) является причиной отсутствия СДЗс; это верно для 19 случаев из 115.

(7) Слабая боль (значение 1 по шкале интенсивности боли) является причиной отсутствия СДЗс; это верно для 18 случаев из 115.

(8) Значение гликемического профиля от 4,5 до 6,5 ммоль/л является причиной отсутствия СДЗс; это верно для 16 случаев из 115.

Наилучшей из этих закономерностей можно считать «Гликемический профиль ммоль/л = 4,5-6,5» (8), так как гипотезы с ним, во-первых, порождаются стратегиями с сильным отрицательным предикатом (в частности, предикатом с запретом на контрпример), и поэтому представляются более надёжными, а, во-вторых, он предсказывает три отрицательных примера из шести, представленных на предсказание.

Также интересной гипотезой можно считать «Боль = 1» (7), для которой обобщенным методом был получен тормоз:

(7\*) Слабая боль (значение 1 по шкале интенсивности боли) является причиной отсутствия СДЗс во всех случаях кроме, случаев, когда возраст пациента

от 41 до 60 лет И есть дефицит массы тела И есть хронический кальцифицирующий панкреатит И есть вирусногепатит И проводится инсулинотерапия И есть алкоголь в анамнезе И есть табакокурение в анамнезе И есть кальцинаты И гликемический профиль от 8,1 до 10 И холецистокинин от 0,16 до 3 И секретин от 0,023 до 0,223 И глюкозависимый инсулиотропный пептид от 0,002 до 2,002 И вазоактивный инсулиновый гормон от 0,001 до 4,0 И соматостатин от 0,02 до 0,42 И гастрин меньше 0,02 И с-пептид от 0,1 до 2.

Остальные гипотезы получены стратегиями с более слабым отрицательным предикатом, хотя объединяют большее количество пациентов из исходной базы (соответствует параметру  $k$  – количеству родителей гипотезы). Признаки *соматостатин* = 0,02-0,42, *гастрин* = 0,02-20 и *с-пептид* = 0,1-2 совпадают с признаками, входящими в положительную эмпирическую закономерность.

Комментарий эксперта-медика. В то время как первичная диагностика сахарного диабета (СД) не представляет сложностей для современной медицины и базируется на изучении уровня гликемии и степени толерантности к глюкозе, точное распознавание сахарного диабета третьего типа (СДЗс) затруднено. Поэтому, с одной стороны, каждый случай сахарного диабета второго типа (СД 2) может рассматриваться на ранних этапах обследования как СДЗс, что приводит к повышению сложности, продолжительности и стоимости диагностики, а, с другой стороны, большая часть пациентов с СДЗс могут оставаться длительное время с невыявленным панкреатическим генезом СД. Эта проблема встает особенно остро, если учитывать очень широкую распространенность СД 2 в популяции – 92,5% от всех случаев СД [10, 17].

Использование ДСМ-метода АПИ позволит решить обе проблемы – отсеять из субпопуляции пациентов с СД 2 конкретные случаи СДЗс, не увеличивая при этом нагрузку на диагностические службы системы здравоохранения. И уже сформированная группа пациентов с подозрением на наличие СДЗс должна быть подвергнута прицельному обследованию для выявления заболевания поджелудочной железы.

Формирование диагностической группы может происходить в два этапа: первый – группировка пациентов по наличию стойкого нарушения гликемического профиля на момент обследования или в анамнезе, если речь идет о пациентах с диагностированным сахарным диабетом первого или второго типа; второй – по данным ряда дополнительных лабораторных анализов из пациентов с гипергликемией с помощью ДСМ-метода АПИ отбираются пациенты, удовлетворяющие эмпирическим закономерностям. При этом целесообразно опираться как на закономерности, подтверждающие наличие СДЗс, так и свидетельствующие о его отсутствии. Такой подход позволит избежать ложной диагностики СДЗс среди пациентов с СД 2. Это особенно важно для закономерности об отсутствии СДЗс, поскольку применение только одной закономерности может привести к неверному заключению.

Например, применение лишь самой сильной закономерности «Гликемический профиль ммоль/л = 4,5-6,5»

должно распространяться и на анамнез, так как у пациентов с контролируемым СД 2, уровень глюкозы плазмы крови находится в пределах нормальных значений. Тогда не будут упущены потенциально возможные случаи панкреатического генеза сахарного диабета у пациентов с установленным и контролируемым диагнозом СД 2. Также важна аккуратная трактовка гипотезы «Боль = 1», которая должна рассматриваться как закономерность, исключая СДЗс только при выполнении условий, указанных в описании данной гипотезы. В противном случае при пренебрежении этими условиями может быть упущен ряд пациентов, в том числе с поздними формами хронического панкреатита с формированием СДЗс, у которых вследствие запущенного воспалительного и фиброзного процесса нарушена болевая афферентация от поджелудочной железы.

## ЗАКЛЮЧЕНИЕ

В нашем исследовании было выполнено более корректное нахождение эмпирических закономерностей по сравнению с более ранними исследованиями сахарного диабета панкреатогенного генеза. Хотя наблюдается небольшое ухудшение результата в части предсказаний (предположительно, из-за изменения языка представления), тем не менее, важно отметить совпадение полученных эмпирических закономерностей с эмпирическими закономерностями, найденными с использованием предыдущих версий ДСМ-метода АПИ в исследованиях 2015 и 2018 гг. Более того, все найденные закономерности имеют клиническую интерпретацию, а их оценка, данная системой, соотносится с экспертным мнением о значимости и надежности полученных эмпирических закономерностей.

Впервые был реализован и применен новый вариант обобщенного ДСМ-метода АПИ. Тернарный предикат (причина — множество тормозов причины — эффект), лежащий в его основе, дает возможность моделировать предметную область с более информативным отношением причинности. Пример эмпирической закономерности, найденной этим методом, имеет клиническую интерпретацию: предполагаемая причина может не реализовываться при определенных условиях, которые, в свою очередь, могут формировать некоторую особую группу пациентов.

Таким образом, разработанная нами интеллектуальная система может стать эффективным инструментом в руках врача-клинициста. Использование системы, основанной на ДСМ-методе, позволит улучшить качество диагностики и прогнозирования течения сахарного диабета третьего типа, что, в свою очередь, приведет к повышению эффективности диагностики рака поджелудочной железы и, следовательно, к снижению онкологической смертности.

## СПИСОК ЛИТЕРАТУРЫ

1. Шестерникова О.П., Агафонов М.А., Винокурова Л.В., Панкратова Е.С., Финн В.К. Интеллектуальная система прогнозирования развития сахарного диабета у больных хроническим панкреатитом // Искусственный интеллект и принятие решений. – 2015. – № 4. – С. 12-50.

2. Агафонов М.А., Шестерникова О.П., Винокурова Л.В., Панкратова Е.С., Финн В.К. О принципах и логических средствах, реализуемых в интеллектуальной системе для гастроэнтерологии // Научно-техническая информация. Сер. 2. – 2017. – № 3. – С. 16-39.
3. Шестерникова О.П. О применении интеллектуальной системы прогнозирования развития диабета у больных хроническим панкреатитом // Искусственный интеллект и принятие решений. – 2016. – № 3. – С. 62-71.
4. Шестерникова О.П., Панкратова Е.С. Интеллектуальная система для обнаружения закономерностей в гастроэнтерологических данных // Пятнадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2016 (3-7 октября 2016 г., г. Смоленск, Россия) // Труды конференции. В 3-х томах. Т. 1. – Смоленск: Универсум, 2016. – С. 396-404.
5. Финн В.К., Шестерникова О.П. Эвристика обнаружения эмпирических закономерностей посредством ДСМ-рассуждений // Научно-техническая информация. Сер. 2. – 2018. – № 9. – С. 7-42; Finn V.K., Shesternikova O.P. The Heuristics of Detection of Empirical Regularities by JSM Reasoning // Automatic Documentation and Mathematical Linguistics. – 2018. – Vol. 52, № 5. – P. 215-247.
6. Шестерникова О.П., Финн В.К., Винокурова Л.В., Лесько К.А., Варванина Г.Г., Тюляева Е.Ю. Интеллектуальная система для диагностики заболеваний поджелудочной железы // Научно-техническая информация. Сер. 2. – 2019. – № 10. – С. 41-48; Shesternikova O.P., Finn V.K., Vinokurova L.V., Les'ko K.A., Varvanina G.G., Tyulyaeva E.Yu. An Intelligent System for Diagnostics of Pancreatic Diseases // Automatic Documentation and Mathematical Linguistics. – 2019. – Vol. 53, № 5. – P. 288-294.
7. Шестерникова О.П., Финн В.К., Лесько К.А., Винокурова Л.В. Интеллектуальная система прогнозирования необходимости применения компьютерной томографии // Искусственный интеллект и принятие решений. – 2022. – № 2. – С. 3-16.
8. Финн В.К., Шестерникова О.П. О новом варианте обобщенного ДСМ-метода // Искусственный интеллект и принятие решений. – 2016. – № 1. – С. 57-64.
9. Pour P.M., Standop J., Batra S.K. Are the islet cells the gatekeepers of the pancreas? // Pancreatology. – 2002. – Vol.2, Iss. 5. – P. 440-448.
10. Hart P.A., Bellin M.D., Andersen D.K. et al. Type 3c (pancreatogenic) diabetes mellitus secondary to chronic pancreatitis and pancreatic cancer // Lancet Gastroenterol Hepatol. – 2016. – Vol.1, № 3. – P. 226-237. DOI:10.1016/S2468-1253(16)30106-6.
11. Woodmansey C., McGovern A.P., McCullough K.A. et al. Incidence, Demographics, and Clinical Characteristics of Diabetes of the Exocrine Pancreas (Type 3c): A Retrospective Cohort Study // Diabetes Care – 2017; –Vol. 40, № 11. – P. 1486-1493.
12. Roy A, Sahoo J, Kamalanathan S, Naik D, Mohan P, Kalayarasan R. Diabetes and pancreatic cancer: Exploring the two-way traffic // World Journal of Gastroenterology. –2021 – Vol. 27, № 30. – P. 4939-4962. DOI:10.3748/wjg.v27.i30.4939.
13. Финн В.К. Дистрибутивные решетки индуктивных процедур // Научно-техническая информация. Сер. 2. – 2014. – № 11. – С. 1-30; Finn V.K. Distributive Lattices of Inductive JSM Procedures // Automatic Documentation and Mathematical Linguistics. – 2014. – Vol. 48, № 6. – P. 265-295.
14. Гусакова С.М., Михеенкова М.А. Интеллектуальный анализ данных как инструмент формирования структуры социума // Научно-техническая информация. Сер. 2. – 2016. – № 8. – С. 9-18; Gusakova S.M., Mikheyenkova M.A. Knowledge Discovery as a Tool for the Formation of the Structure of Society // Automatic Documentation and Mathematical Linguistics. – 2016. – Vol. 50, № 5. – P. 179-188.
15. Fowler M., Rice D., Foemmel M., Hieatt E., Mee R., Stafford R. Patterns of Enterprise Application Architecture. – Addison-Wesley Professional, 2002 – 560 p.
16. Финн В.К., Блинова В.Г., Панкратова Е.С., Фабрикантова Е.Ф. Интеллектуальные системы для анализа медицинских данных // ДСМ-метод автоматического порождения гипотез: Логические и эпистемологические основания / сост. О.М. Аншаков, Е.Ф. Фабрикантова; под общ. ред. О.М. Аншакова. – Москва: Книжный дом «ЛИБРОКОМ», 2009. – 432 с.
17. Дедов И., Шестакова М., Викулова О., Железнякова А., Исаков М. Эпидемиологические характеристики сахарного диабета в Российской Федерации: клинико-статистический анализ по данным регистра сахарного диабета на 01.01.2021 // Сахарный диабет. – 2021. – Т. 24, № 3. – С. 204-221. DOI:10.14341/DM12759.

## ПРИЛОЖЕНИЕ 1

### Определения предикатов сходства

$M_{a,n}^{\sigma}(V,W)$  ( $\sigma = +, -, 0, \tau$ ) являются минимальными предикатами, реализующими индуктивные выводы. Они могут быть дополнены более информативными усилениями:

1. Запрет на контрпример (обозначается  $b$ ): положительная причина не может входить в отрицательную и наоборот

$$(b)^+ \forall X \forall Y (((V \subset X) \& (W \subset Y)) \rightarrow (J_{(1,n)}(X \Rightarrow_1 Y) \vee J_{(\tau,n)}(X \Rightarrow_1 Y)))$$

$$(b)^- \forall X \forall Y (((V \subset X) \& (W \subset Y)) \rightarrow (J_{(-1,n)}(X \Rightarrow_1 Y) \vee J_{(\tau,n)}(X \Rightarrow_1 Y)))$$

$$M_{ab,n}^+(V,W) = M_{a,n}^+(V,W) \& (b)^+, M_{ab,n}^-(V,W) = M_{a,n}^-(V,W) \& (b)^-.$$

2. Метод различия (обозначается  $d_0$ ): ищутся такие общие фрагменты отдельно для положительных и отрицательных примеров, что остаток от примера за вычетом фрагмента, который содержится в этом примере, не содержится в других примерах рассматриваемой группы

$$(d_0)^+ \forall X \forall Y \forall Z \forall U ((J_{(1,n)}(X \Rightarrow_1 Y) \& (W \subset Y) \& (V \subset X) \& ((X - V) \subset Z) \& ((X - V) \neq \emptyset) \& \neg(V \subset Z)) \rightarrow \neg J_{(1,n)}(Z \Rightarrow_1 U) \vee \neg(W \subset U)))^2$$

$$(d_0)^- \forall X \forall Y \forall Z \forall U ((J_{(-1,n)}(X \Rightarrow_1 Y) \& (W \subset Y) \& (V \subset X) \& ((X - V) \subset Z) \& ((X - V) \neq \emptyset) \& \neg(V \subset Z)) \rightarrow \neg J_{(-1,n)}(Z \Rightarrow_1 U) \vee \neg(W \subset U)))$$

$$M_{ad_0,n}^+(V,W) \equiv M_{a,n}^+(V,W) \& (d_0)^+, M_{ad_0,n}^-(V,W) \equiv M_{a,n}^-(V,W) \& (d_0)^-.$$

3. Метод различия с запретом на контрпример, который является конъюнкцией этих двух условий:

$$M_{ad_0,n}^+(V,W) \equiv M_{a,n}^+(V,W) \& (d_0)^+ \& (b)^+$$

Предикаты образуют дистрибутивную решётку по логической силе (рис. 3):

Аналогично, существует решётка для отрицаний этих предикатов (рис. 4):

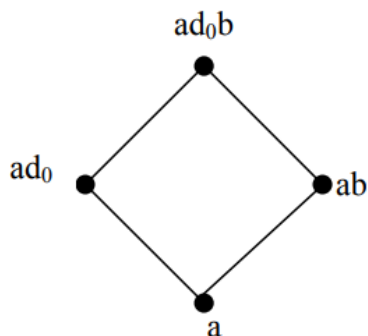


Рис. 3. Дистрибутивная решетка для предикатов схождения, использующихся в исследовании

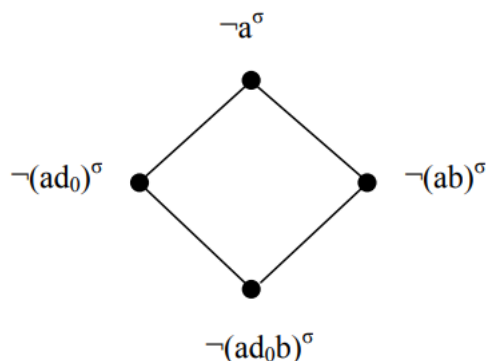


Рис. 4. Дистрибутивная решетка для отрицаний предикатов схождения, использующихся в исследовании

Стратегия  $Str$  определяется парой предикатов схождения: для положительных и для отрицательных примеров. Таким образом, четыре предиката схождения определяют шестнадцать возможных стратегий. Множество стратегий образуют изоморфные решетки по логической силе для каждого из четырёх правил индуктивного вывода: для положительных гипотез, для отрицательных гипотез, для гипотез о фактическом противоречии и для гипотез о неопределенности. Эти решетки являются прямым произведением предыдущих. Например, индуктивному правилу для положительных гипотез соответствует произведение решетки предикатов схождения (см. рис. 3) и решетки отрицаний предикатов схождения (см. рис. 4). Соответствующая результату решетка приведена на рис. 5.

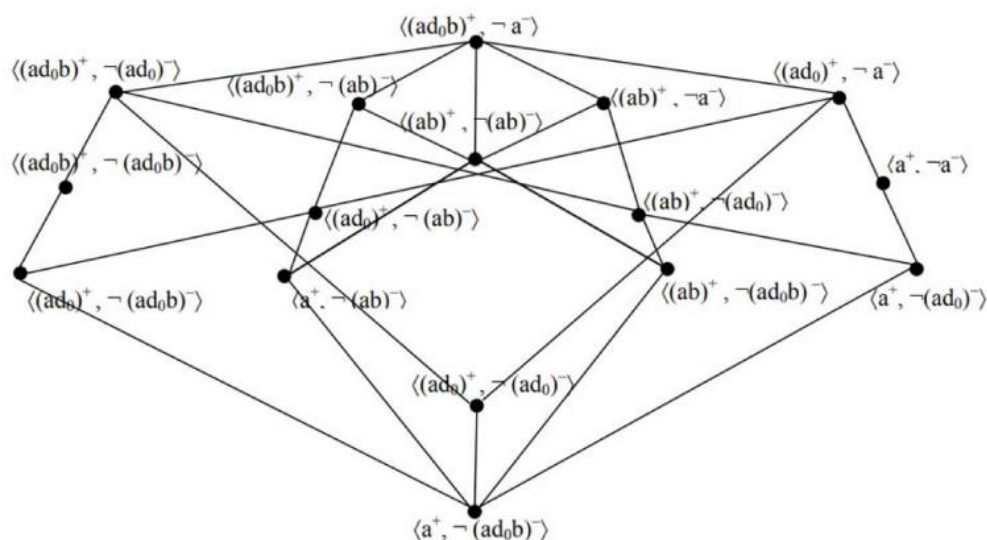


Рис. 5. Диаграмма стратегий для правила индуктивного вывода положительных гипотез

<sup>2</sup> Здесь « $\rightarrow$ » обозначает операцию булевой разности.

## Классы эквивалентности

Стратегии ( $Str_{x,y}$ ) для описания классов эквивалентности представлены в виде таблиц, где по вертикали располагается компонента  $x$ , а по горизонтали компонента  $y$ . Таким образом, стратегии соответствует ячейка в таблице, а классу эквивалентности — объединение нескольких ячеек.

Далее приведено разбиение для всех возможных отношений эквивалентности. Если для стратегии отношение не определено, ячейка закрашена серым. Классы перечислены как множества  $Cl_1$ ,  $Cl_2$  и т. д.

$$E_1^+ = E_{13}^+$$

|                  | y | a                     | ad <sub>0</sub>                      | ab                     | abd <sub>0</sub>                      |
|------------------|---|-----------------------|--------------------------------------|------------------------|---------------------------------------|
| x                |   |                       |                                      |                        |                                       |
| a                |   | <a,a>                 | <a,ad <sub>0</sub> >                 | <a,ab>                 | <a,abd <sub>0</sub> >                 |
| ad <sub>0</sub>  |   | <ad <sub>0</sub> ,a>  | <ad <sub>0</sub> ,ad <sub>0</sub> >  | <ad <sub>0</sub> ,ab>  | <ad <sub>0</sub> ,abd <sub>0</sub> >  |
| ab               |   | <ab,a>                | <ab,ad <sub>0</sub> >                | <ab,ab>                | <ab,abd <sub>0</sub> >                |
| abd <sub>0</sub> |   | <abd <sub>0</sub> ,a> | <abd <sub>0</sub> ,ad <sub>0</sub> > | <abd <sub>0</sub> ,ab> | <abd <sub>0</sub> ,abd <sub>0</sub> > |

$$Cl_1 = \{ Str_{a,ab}, Str_{ad_0,b}, Str_{ab,ab}, Str_{abd_0,ab} \}$$

$$Cl_2 = \{ Str_{a,abd_0}, Str_{ad_0,bd_0}, Str_{ab,abd_0}, Str_{abd_0,abd_0} \}$$

На остальных стратегиях отношение не определено

$$E_1^- = E_{13}^-$$

|                  | y | a                     | ad <sub>0</sub>                      | ab                     | abd <sub>0</sub>                      |
|------------------|---|-----------------------|--------------------------------------|------------------------|---------------------------------------|
| x                |   |                       |                                      |                        |                                       |
| a                |   | <a,a>                 | <a,ad <sub>0</sub> >                 | <a,ab>                 | <a,abd <sub>0</sub> >                 |
| ad <sub>0</sub>  |   | <ad <sub>0</sub> ,a>  | <ad <sub>0</sub> ,ad <sub>0</sub> >  | <ad <sub>0</sub> ,ab>  | <ad <sub>0</sub> ,abd <sub>0</sub> >  |
| ab               |   | <ab,a>                | <ab,ad <sub>0</sub> >                | <ab,ab>                | <ab,abd <sub>0</sub> >                |
| abd <sub>0</sub> |   | <abd <sub>0</sub> ,a> | <abd <sub>0</sub> ,ad <sub>0</sub> > | <abd <sub>0</sub> ,ab> | <abd <sub>0</sub> ,abd <sub>0</sub> > |

$$Cl_1 = \{ Str_{a,a}, Str_{ad_0,a}, Str_{a,ad_0}, Str_{ad_0,ad_0}, Str_{a,ab}, Str_{ad_0,b}, Str_{a,abd_0}, Str_{ad_0,bd_0} \}$$

$$Cl_2 = \{ Str_{ab,a}, Str_{abd_0,a}, Str_{ab,ad_0}, Str_{abd_0,ad_0}, Str_{ab,ab}, Str_{abd_0,b}, Str_{ab,abd_0}, Str_{abd_0,bd_0} \}$$

$$E_1 = E_{13}$$

|                  | y | a                     | ad <sub>0</sub>                      | ab                     | abd <sub>0</sub>                      |
|------------------|---|-----------------------|--------------------------------------|------------------------|---------------------------------------|
| x                |   |                       |                                      |                        |                                       |
| a                |   | <a,a>                 | <a,ad <sub>0</sub> >                 | <a,ab>                 | <a,abd <sub>0</sub> >                 |
| ad <sub>0</sub>  |   | <ad <sub>0</sub> ,a>  | <ad <sub>0</sub> ,ad <sub>0</sub> >  | <ad <sub>0</sub> ,ab>  | <ad <sub>0</sub> ,abd <sub>0</sub> >  |
| ab               |   | <ab,a>                | <ab,ad <sub>0</sub> >                | <ab,ab>                | <ab,abd <sub>0</sub> >                |
| abd <sub>0</sub> |   | <abd <sub>0</sub> ,a> | <abd <sub>0</sub> ,ad <sub>0</sub> > | <abd <sub>0</sub> ,ab> | <abd <sub>0</sub> ,abd <sub>0</sub> > |

$$Cl_1 = \{ Str_{a,a}, Str_{ad_0,a}, Str_{a,ad_0}, Str_{ad_0,ad_0} \}$$

$$Cl_2 = \{ Str_{a,ab}, Str_{ad_0,b} \}$$

$$Cl_3 = \{ Str_{a,abd_0}, Str_{ad_0,bd_0} \}$$

$$Cl_4 = \{ Str_{ab,a}, Str_{abd_0,a}, Str_{ab,ad_0}, Str_{abd_0,ad_0} \}$$

$$Cl_5 = \{ Str_{ab,ab}, Str_{abd_0,b} \}$$

$$Cl_6 = \{ Str_{ab,abd_0}, Str_{abd_0,bd_0} \}$$

$$E_2^+$$

|                  | y | a                     | ad <sub>0</sub>                      | ab                     | abd <sub>0</sub>                      |
|------------------|---|-----------------------|--------------------------------------|------------------------|---------------------------------------|
| x                |   |                       |                                      |                        |                                       |
| a                |   | <a,a>                 | <a,ad <sub>0</sub> >                 | <a,ab>                 | <a,abd <sub>0</sub> >                 |
| ad <sub>0</sub>  |   | <ad <sub>0</sub> ,a>  | <ad <sub>0</sub> ,ad <sub>0</sub> >  | <ad <sub>0</sub> ,ab>  | <ad <sub>0</sub> ,abd <sub>0</sub> >  |
| ab               |   | <ab,a>                | <ab,ad <sub>0</sub> >                | <ab,ab>                | <ab,abd <sub>0</sub> >                |
| abd <sub>0</sub> |   | <abd <sub>0</sub> ,a> | <abd <sub>0</sub> ,ad <sub>0</sub> > | <abd <sub>0</sub> ,ab> | <abd <sub>0</sub> ,abd <sub>0</sub> > |

$$Cl_1 = \{ Str_{ab,a}, Str_{abd_0,a}, Str_{ab,ad_0}, Str_{abd_0,ad_0} \}$$

$$Cl_2 = \{ Str_{ab,ab}, Str_{abd_0,b}, Str_{ab,abd_0}, Str_{abd_0,bd_0} \}$$

На остальных стратегиях отношение не определено

$$E_2^-$$

|                  | y | a                     | ad <sub>0</sub>                      | ab                     | abd <sub>0</sub>                      |
|------------------|---|-----------------------|--------------------------------------|------------------------|---------------------------------------|
| x                |   |                       |                                      |                        |                                       |
| a                |   | <a,a>                 | <a,ad <sub>0</sub> >                 | <a,ab>                 | <a,abd <sub>0</sub> >                 |
| ad <sub>0</sub>  |   | <ad <sub>0</sub> ,a>  | <ad <sub>0</sub> ,ad <sub>0</sub> >  | <ad <sub>0</sub> ,ab>  | <ad <sub>0</sub> ,abd <sub>0</sub> >  |
| ab               |   | <ab,a>                | <ab,ad <sub>0</sub> >                | <ab,ab>                | <ab,abd <sub>0</sub> >                |
| abd <sub>0</sub> |   | <abd <sub>0</sub> ,a> | <abd <sub>0</sub> ,ad <sub>0</sub> > | <abd <sub>0</sub> ,ab> | <abd <sub>0</sub> ,abd <sub>0</sub> > |

$$Cl_1 = \{ Str_{a,a}, Str_{ad_0,a} \}$$

$$Cl_2 = \{ Str_{a,ad_0}, Str_{ad_0,ad_0}, Str_{a,ab}, Str_{ad_0,b}, Str_{a,abd_0}, Str_{ad_0,bd_0} \}$$

$$Cl_3 = \{ Str_{ab,a}, Str_{abd_0,a} \}$$

$$Cl_4 = \{ Str_{ab,ad_0}, Str_{abd_0,ad_0} \}$$

$$Cl_5 = \{ Str_{ab,ab}, Str_{abd_0,b} \}$$

$$Cl_6 = \{ Str_{ab,abd_0}, Str_{abd_0,bd_0} \}$$

$$E_{123}^+$$

|                  | y | a                     | ad <sub>0</sub>                      | ab                     | abd <sub>0</sub>                      |
|------------------|---|-----------------------|--------------------------------------|------------------------|---------------------------------------|
| x                |   |                       |                                      |                        |                                       |
| a                |   | <a,a>                 | <a,ad <sub>0</sub> >                 | <a,ab>                 | <a,abd <sub>0</sub> >                 |
| ad <sub>0</sub>  |   | <ad <sub>0</sub> ,a>  | <ad <sub>0</sub> ,ad <sub>0</sub> >  | <ad <sub>0</sub> ,ab>  | <ad <sub>0</sub> ,abd <sub>0</sub> >  |
| ab               |   | <ab,a>                | <ab,ad <sub>0</sub> >                | <ab,ab>                | <ab,abd <sub>0</sub> >                |
| abd <sub>0</sub> |   | <abd <sub>0</sub> ,a> | <abd <sub>0</sub> ,ad <sub>0</sub> > | <abd <sub>0</sub> ,ab> | <abd <sub>0</sub> ,abd <sub>0</sub> > |

$$Cl_1 = \{ Str_{a,ab}, Str_{ad_0,b} \}$$

$$Cl_2 = \{ Str_{a,abd_0}, Str_{ad_0,bd_0} \}$$

$$Cl_3 = \{ Str_{ab,ab}, Str_{abd_0,b} \}$$

$$Cl_4 = \{ Str_{ab,abd_0}, Str_{abd_0,bd_0} \}$$

|             |         |          |            |           |             |                                              |
|-------------|---------|----------|------------|-----------|-------------|----------------------------------------------|
| $E_{123}^-$ |         |          |            |           |             | $Cl_1 = \{ Str_{a,a}, Str_{ad0,a} \}$        |
|             | y       | a        | $ad_0$     | ab        | $abd_0$     | $Cl_2 = \{ Str_{a,ad0}, Str_{ad0,ad0} \}$    |
| x           |         |          |            |           |             | $Cl_3 = \{ Str_{a,ab}, Str_{ad0,b} \}$       |
|             | a       | <a,a>    | <a,ad0>    | <a,ab>    | <a,abd0>    | $Cl_4 = \{ Str_{a,abd0}, Str_{ad0,bd0} \}$   |
|             | $ad_0$  | <ad0,a>  | <ad0,ad0>  | <ad0,ab>  | <ad0,abd0>  | $Cl_5 = \{ Str_{ab,a}, Str_{abd0,a} \}$      |
|             | ab      | <ab,a>   | <ab,ad0>   | <ab,ab>   | <ab,abd0>   | $Cl_6 = \{ Str_{ab,ad0}, Str_{abd0,ad0} \}$  |
|             | $abd_0$ | <abd0,a> | <abd0,ad0> | <abd0,ab> | <abd0,abd0> | $Cl_7 = \{ Str_{ab,ab}, Str_{abd0,b} \}$     |
|             |         |          |            |           |             | $Cl_8 = \{ Str_{ab,abd0}, Str_{abd0,bd0} \}$ |

Материал поступил в редакцию 03.02.23.

## Сведения об авторах

**ШЕСТЕРНИКОВА Ольга Павловна** – инженер-программист отдела разработки программного обеспечения ФГБУ «Центральный научно-исследовательский институт организации и информатизации здравоохранения» Минздрава РФ (ЦНИИОИЗ МЗ РФ), Москва  
e-mail: oshesternikova@gmail.com

**ФИНН Виктор Константинович** – доктор технических наук, профессор, заслуженный деятель науки РФ, главный научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН, научный руководитель Отделения интеллектуальных систем в гуманитарной сфере РГГУ, Москва  
e-mail: ira.finn@gmail.com

**ЛЕСЬКО Константин Александрович** – кандидат медицинских наук, врач-рентгенолог рентгеновского отделения отдела лучевых методов диагностики и лечения Московского Клинического Научного Центра имени А.С. Логинова Департамента здравоохранения г. Москвы (ГБУЗ МКНЦ им. А.С. Логинова ДЗМ)  
e-mail: k\_lesko@mail.ru

**ВИНОКУРОВА Людмила Васильевна** – доктор медицинских наук, ведущий научный сотрудник отделения патологии поджелудочной железы и желчных путей Московского Клинического Научного Центра имени А.С. Логинова Департамента здравоохранения г. Москвы (ГБУЗ МКНЦ им. А.С. Логинова ДЗМ)  
e-mail: vinokurova52@mail.ru

# Существование больших подрешеток, изоморфных булевой алгебре, в решетке кандидатов

*Вероятностный подход к извлечению знаний с помощью операции сходства возник как способ преодоления необходимости порождения (потенциально экспоненциальной) решетки всех кандидатов (например, для случая булевой алгебры). В настоящей работе доказано, что для случайной плотной обучающей выборки, порождаемой последовательностью Бернулли, вероятность того, что в решетке кандидатов возникнет большая подрешетка, изоморфная булевой алгебре, будет стремиться к единице при увеличении размера выборки.*

**Ключевые слова:** обучающая выборка, решетка кандидатов, булева алгебра

**DOI:** 10.36535/0548-0027-2023-04-3

## ВВЕДЕНИЕ

Идеи В.К. Финна о синтезе познавательных процедур привели к созданию ДСМ-метода автоматического порождения гипотез [1]. На основании обучающей выборки с помощью бинарной операции сходства он порождает кандидаты в гипотезы о причинах целевого свойства  $\langle A, B \rangle$ , где  $A \subseteq O$  — список обучающих примеров, а  $B \subseteq F$  — совокупность их общих признаков, называемая *фрагментом* этого кандидата.

Однако при компьютерной реализации ДСМ-метода возникли вычислительные трудности: экспоненциальный взрыв числа кандидатов, существование [2] «фантомных» кандидатов, приводящих к переобучению, и сложность вычислительных процедур. Самым критичным оказался эффект экспоненциально большого числа кандидатов относительно размера обучающей выборки. Наиболее наглядно это проявляется в случае  $n$ -мерной булевой алгебры, где обучающая выборка содержит все коатомы. Пусть  $O = \{o_1, o_2, \dots, o_n\}$  будет множеством обучающих примеров (коатомов), каждый из которых описывается признаками из списка  $F = \{f_1, f_2, \dots, f_n\}$ , и  $o_i I f_j \leftrightarrow j \neq i$ .

Ясно, что  $o_{j1} \cap o_{j2} \cap \dots \cap o_{js} = F \setminus \{f_{j1}, f_{j2}, \dots, f_{js}\}$ , так как добавление в сходство примера  $o_k$  с номером  $k$  удаляет из фрагмента признак  $f_k$  с тем же самым номером  $k$ . Очевидно, что минимальная обучающая выборка из табл. 1 занимает  $n^2$  бит, а порождаемая из нее булева алгебра кандидатов требует  $2^n \cdot n$  бит памяти.

Для преодоления этих трудностей мы предложили использовать вероятностные алгоритмы [3] порождения случайного подмножества решетки кандидатов и основанный на них вероятностно-комбинаторный формальный метод извлечения знаний (ВКФ-метод) [4].

Используя вероятностную модель последовательности Бернулли [5] для порождения обучающих примеров, мы докажем, что при  $m \rightarrow \infty$  вероятность порождения  $m$ -мерной булевой алгебры как подрешетки кандидатов для случайной обучающей выборки из  $k \geq m$  примеров, описываемых  $n \geq e \cdot m \cdot \ln m$  бинарными признаками, стремится к единице. Этот результат показывает, что случай неполиномиального размера решетки кандидатов встречается не так уж и редко.

Таблица 1

**Минимальная обучающая выборка для булевой алгебры**

| $O \setminus F$ | $f_1$    | $f_2$    | $\dots$  | $f_n$    |
|-----------------|----------|----------|----------|----------|
| $o_1$           | 0        | 1        | $\dots$  | 1        |
| $o_2$           | 1        | 0        | $\dots$  | 1        |
| $\vdots$        | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $o_n$           | 1        | 1        | $\dots$  | 0        |

Далее представим формальные определения используемых понятий, а также формулировку и доказательство основной теоремы. Это доказательство аналогично оценке хвоста вероятности в задаче о собирателе купонов, хорошо изученной в теории вероятностей (см., например, [5]).

## НЕОБХОДИМЫЕ ОПРЕДЕЛЕНИЯ И ФАКТЫ

Обучающей выборкой обозначим бинарное отношение  $I \subseteq O \times F$  между элементами множества  $O$ , которые мы называем *обучающими объектами*, и элементами множества  $F$ , которые мы называем (бинарными) *признаками*. Если в строчке, соответствующей объекту  $o \in O$ , и столбце, соответствующем

признаку  $f \in F$ , стоит единица, то мы утверждаем, что объект  $o$  обладает признаком  $f$ , и обозначаем это через  $oIf$ , если не обладает – то объект  $o$  не имеет признака  $f$ .

Для подмножества  $A \subseteq O$  объектов его сходством называется подмножество  $A' = \{f \in F: \forall o \in A [oIf]\} \subseteq F$ . По определению  $\emptyset' = F$  для пустого множества объектов. С вычислительной точки зрения эта операция совпадает с последовательным вычислением побитового умножения строк, соответствующих отобранным во множество  $A$  объектам.

Для подмножества  $B \subseteq F$  признаков его сходством называется подмножество  $B' = \{o \in O: \forall f \in B [oIf]\} \subseteq O$ . Полагаем  $\emptyset' = O$  для пустого множества признаков. Эта операция совпадает с последовательным вычислением побитового умножения столбцов, соответствующих отобранным во множество  $B$  признакам.

**Определение 1.** Пару  $\langle A, B \rangle$  назовем **кандидатом**, если  $A = B' \subseteq O$  и  $B = A' \subseteq F$ .

Легко проверить, что множество всех кандидатов (для фиксированной обучающей выборки) образует решетку с операциями:

$$\langle A, B \rangle \wedge \langle C, D \rangle = \langle (B \cap D)', B \cap D \rangle, \quad (1)$$

$$\langle A, B \rangle \vee \langle C, D \rangle = \langle A \cap C, (A \cap C)' \rangle. \quad (2)$$

Порядок на кандидатах задается правилом:

$$\langle A, B \rangle \leq \langle C, D \rangle \Leftrightarrow B \subseteq D. \quad (3)$$

Для того, чтобы получить  $m$ -мерную булеву алгебру как подрешетку кандидатов для обучающей выборки из  $k \geq m$  примеров, описываемых  $n$  бинарными признаками, нам потребуется найти  $m$  строк (перестановкой можно считать их первыми строками) и  $m$  столбцов, которые содержат единственный ноль и  $m-1$  единиц, так, чтобы для каждой строки в отобранных столбцах встречался ровно один ноль. Например, в табл. 2 в столбце  $f_j$  единственный ноль встречается в строке, соответствующей обучающему примеру  $o_i$ .

Таблица 2

Случайная обучающая выборка для булевой алгебры

| $O \backslash F$ | $f_1$            | ...      | $f_j$    | ...      | $f_n$            |
|------------------|------------------|----------|----------|----------|------------------|
| $o_1$            | $\delta_{1,1}$   | ...      | 1        | ...      | $\delta_{1,n}$   |
| $\vdots$         | $\vdots$         | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$         |
| $o_{i-1}$        | $\delta_{i-1,1}$ | $\vdots$ | 1        | $\vdots$ | $\delta_{i-1,n}$ |
| $o_i$            | $\delta_{i-1,1}$ | $\vdots$ | 0        | $\vdots$ | $\delta_{i,n}$   |
| $o_{i+1}$        | $\delta_{i+1,1}$ | $\vdots$ | 1        | $\vdots$ | $\delta_{i+1,n}$ |
| $\vdots$         | $\vdots$         | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$         |
| $o_m$            | $\delta_{m,1}$   | ...      | 1        | ...      | $\delta_{m,n}$   |

Обозначим через  $c_i$  столбец высоты  $m$ , где  $m-1$  позиция заполнены единицами, а единственный ноль встречается на  $i$ -ой позиции. Тогда ситуацию с  $j$ -м столбцом табл. 2 можно описать равенством  $f_j = c_i$ .

Рассмотрим теперь серию испытаний Бернулли  $\delta_{1,1}, \dots, \delta_{m,n}$ , заполняющую табл. 2 с вероятностью успеха  $p = (1-1/m)$ . Значение для параметра  $p$  определяется как оценка максимального правдоподобия, чтобы обеспечить в среднем ровно один ноль в столбце высоты  $m$ . Тогда вероятность  $P[f_j = c_i]$  получить столбец  $e_i$  равна  $s = p^{m-1}(1-p)$  для любых  $i$  и  $j$ , причем события  $[f_j = c_i]$  и  $[f_{j'} = c_{i'}]$  независимы для разных  $i, i', j$  и  $j'$ .

## ОСНОВНАЯ ТЕОРЕМА

Предложенные нами рассуждения показывают, что мы имеем ситуацию, полностью аналогичную задаче о собирателе купонов (см., например, [5]).

**Теорема 1.** При  $m \rightarrow \infty$  и  $n \geq e \cdot m \cdot \ln m$  для вероятности  $P[\forall j \exists i (f_j = c_i)]$  порождения  $m$ -мерной булевой алгебры как подрешетки кандидатов имеем:

$$\lim_{m \rightarrow \infty} P[\forall j \exists i (f_j = c_i)] \geq 1 - m^{(1-e \cdot m \cdot s)} \rightarrow 1.$$

**Доказательство.** Рассмотрим вероятность дополнительного события  $P = P[\exists j \forall i (f_j \neq c_i)]$ .

Из неравенства Буля получаем оценку:

$$P \leq m \cdot \prod_{i=1}^{e \cdot m \cdot \ln m} (1 - P[f_j = c_i])$$

из-за независимости событий  $[f_j = c_i]$  и  $[f_{j'} = c_{i'}]$  для разных  $i, i', j$  и  $j'$ . Используя неравенство  $(1-s) \leq e^{-s}$ , имеем оценку  $P \leq e^{(\ln m \cdot s \cdot e \cdot m \cdot \ln m)} = m^{(1-s \cdot e \cdot m)}$ , где  $s$  – общая вероятность событий  $[f_j = c_i]$  для разных  $i$  и  $j$ .

Оценим теперь:

$$e \cdot m \cdot s = e \cdot m \cdot (1-1/m)^{m-1} \cdot (1/m) = e \cdot (m-1)^{m-1} / m^{m-1}.$$

Из известного неравенства  $(1+1/(m-1))^{m-1} < e$  следует, что  $e \cdot m \cdot s > 1$ , откуда получаем  $m^{(1-e \cdot m \cdot s)} \rightarrow 0$  при  $m \rightarrow \infty$ , что завершает доказательство теоремы.

## ЗАКЛЮЧЕНИЕ

Из доказанной теоремы следует, что в плотной обучающей выборке с использованием  $n$  признаков, описывающих  $m$  обучающих примеров, вероятность порождения булевой алгебры размерности  $m = \sqrt{n}$  в качестве подрешетки кандидатов становится близкой к единице, так как для достаточно больших  $m$  выполняется  $e \cdot m \cdot \ln m < m^2 = n$ . Из-за экспоненциальности числа элементов  $m$ -мерной булевой алгебры мы получаем неполиномиальность числа кандидатов, что делает еще более востребованным применение ВКФ-метода [4] для извлечения знаний из выборок достаточно большого размера.

\* \* \*

Автор благодарит своих коллег по ВЦ им. А.А. Дородницына ФИЦ ИУ РАН за поддержку и полезные дискуссии. Особую благодарность автор выражает своей аспирантке Л.А. Якимовой за совместную работу, обсуждения и поддержку.

## СПИСОК ЛИТЕРАТУРЫ

1. ДСМ-метод автоматического порождения гипотез: логические и эпистемологические основания / под ред. О.М. Аншаков, В.К. Финн. – Москва: URSS, 2009. – 430 с.
2. Виноградов Д.В. Скорость сходимости к пределу вероятности порождения случайного сходства при наличии контр-примеров // Научно-техническая информация. Сер. 2. – 2018. – № 2. – С. 21–24; Vinogradov D.V. The Rate of Convergence to the Limit of the Probability of Encountering an Accidental Similarity in the Presence of Counter Examples // Automatic Documentation and Mathematical Linguistics. – 2018. – Vol. 52, № 1. – P. 35-37.
3. Виноградов Д.В. Вероятностное порождения гипотез в ДСМ-методе с помощью простейших цепей Маркова // Научно-техническая информация. Сер. 2. – 2012. – № 9. – С. 20-27; Vinogradov D.V. Random Generation of Hypotheses in the JSM Method using Simple Markov Chains // Automatic Documentation and Mathematical Linguistics. – 2012. – Vol. 46, № 5. – P. 221-228.
4. Vinogradov D.V. Machine Learning Based on Similarity Operation // Communication in Computer and Information Science. – 2018. – Vol. 934. – P. 46–59.
5. Феллер В. Введение в теорию вероятностей и ее приложения: в 2-х т. Т. 1 / пер. с англ. – Москва: Мир, 1984. – 528 с.

*Материал поступил в редакцию 11.02.23*

### Сведения об авторе

**ВИНОГРАДОВ Дмитрий Вячеславович** – доктор физико-математических наук, ведущий научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН, Москва  
e-mail: vinogradov.d.w@gmail.com

## Концепция и языковые средства представления и обработки знаний на смысловом уровне

*Обосновываются требования, которым должны удовлетворять языки представления знаний, и делается вывод, что причины неудач предыдущих проектов в рамках парадигмы систем, основанных на знаниях, состоят в отсутствии языков представления знаний, ориентированных на работу с выраженными в понятийной форме знаниями. Основными элементами системы являются база знаний, база данных и машина вывода. Факты в БД хранятся в виде сети экземпляров понятий, общие знания представляются в виде онтологии предметной области, состоящей из совокупности определений понятий, условий целостности и правил. Средства определения онтологии позволяют решить задачу описания семантики понятий, используя модель понятия типа представление/содержание, что дает возможность представлять и обрабатывать информацию на смысловом уровне. Достоинством предлагаемого подхода является возможность использования одного языка для ввода не только фактов и запросов, но и для определения понятий и правил.*

**Ключевые слова:** языки представления знаний, глубинная структура, онтология, база знаний, факт, машина вывода, концептуальное логическое выражение, унификация

DOI: 10.36535/0548-0027-2023-04-4

### ВВЕДЕНИЕ

Обработка знаний – главный предмет исследований в области искусственного интеллекта – в настоящее время осуществляется в основном в рамках парадигмы нейронных сетей [1, 2]. Такая ситуация сложилось не более пятнадцати лет назад по двум причинам. С одной стороны, основная парадигма – системы, основанные на знаниях [3, 4], – которая была в центре внимания в предыдущие годы, не принесла никаких практически значимых результатов, а с другой стороны, нейронные сети продемонстрировали высокую эффективность для ряда приложений. Однако нельзя упускать из виду тот факт, что парадигма нейронных сетей реализует только одну индуктивную компоненту интеллектуальной деятельности человека – формирование общих правил по примерам их применения. С помощью нейронных сетей в принципе нельзя описать в виде алгоритма сформированную в процессе обучения процедуру. Известно, что интеллектуальная деятельность человека включает механизм дедукции – применение общих правил в конкретной ситуации, и что основной объем знаний современный человек получает в языковой форме, прежде всего на естественном языке (ЕЯ). Поэтому и дедуктивные правила чаще всего формулируются на ЕЯ.

Обработка знаний включает два этапа: преобразование знаний во внутренний формат системы с учетом ранее полученных знаний и их обработку в соответствии с запросами. Обработка знаний, задаваемых на ЕЯ, означает, что в отличие от поверхностного анализа должна использоваться технология, предполагающая семантический анализ исходного текста, завершающийся построением его глубинной структуры. Создание семантического парсера, способного выполнить такой анализ, представляет собой чрезвычайно сложную задачу, решить которую в полной мере так и не удалось. В связи с этим усилия исследователей сконцентрировались на поиске вариантов представления знаний независимо от способа их формирования. Основная задача этих исследований заключалась в попытках создания так называемых языков представления знаний (ЯПЗ) [3–9], текст на которых в базе знаний (БЗ) системы был бы формальным представлением извлеченных знаний. Очевидно, что для **обработки знаний на смысловом уровне (ОЗСУ)** языки представления знаний, помимо очевидных требований о достаточной выразительной мощности и эффективности реализации, должны удовлетворять еще одному важнейшему требованию, вытекающему из свойств понятий, которые используются в ЕЯ. В подавляющем числе случаев такие понятия обладают

смысловым содержанием (семантикой), которое подразумевается при их употреблении. Отсюда следует, что ЯПЗ, используемые в таких системах, должны поддерживать возможность определения семантики понятий. Проведем с этих позиций краткий анализ возможности использования существующих ЯПЗ в системах обработки знаний на смысловом уровне.

Основные типы языков представления знаний можно разбить на группы, исходя из моделей представления знаний, на которых они основаны: логически ориентированные ЯПЗ, фрейм-ориентированные ЯПЗ, ЯПЗ на основе продукций, а также их комбинации [3].

К первой группе относится семейство языков с общим названием семантические сети. Данные языки – это различные подмножества языка исчисления предикатов первого порядка, наиболее известного формального аппарата представления знаний. К недостаткам такого подхода можно отнести отсутствие для них готовых реализаций машин вывода и принципиальную невозможность обеспечить вычислительную эффективность из-за декларативного способа представления знаний. Кроме того, использование предикатов для представления понятий значительно обедняет возможности ЯПЗ, так как язык первого порядка напрямую не поддерживает манипулирование отдельными кортежами (как будет показано далее, без этого невозможно решить поставленные задачи). Наконец, еще более важно, что семантические сети не содержат адекватных механизмов для определения семантики понятий.

Языки, относящиеся ко второй группе, классическими представителями которых являются *FRL*, *KRL* [3, 6], поддерживают возможность работы с экземплярами понятий, имея преимущество в этом отношении перед семантическими сетями. Главный недостаток

этих языков – высокая сложность и трудоемкость создания баз знаний на их основе, исходя из описания знаний на ЕЯ. Эти языки также не содержат механизмов для определения семантики понятий.

Языки, относящиеся к третьей группе [3, 7], не являются «понятийно ориентированными», что сразу исключает возможность их применения для реализации ОЗСУ-систем.

Более современные версии языков представления знаний, такие как *KIF*, *CycL*, *Loom* [10], не выходят за рамки парадигмы языков первого порядка или фреймов. *KIF* подобен таким фреймовым языкам, как *KL-ONE*, но в отличие от *KL-ONE* его основное применение связано с обменом знаниями между системами; *CycL* – это язык онтологии; *Loom* – это язык, служащий для построения интеллектуальных приложений, сочетающий фреймовый и логический подходы.

Из приведенного краткого обзора понятно, что ни один из существующих ЯПЗ не позволяет определять семантику понятий. Соответственно становится очевидной задача, которую необходимо решить для создания систем обработки знаний на смысловом уровне – разработать понятийно ориентированный язык представления знаний, позволяющий определять и использовать смысловое содержание понятий.

## СТРУКТУРА СИСТЕМ ОБРАБОТКИ ЗНАНИЙ НА СМЫСЛОВОМ УРОВНЕ

Для удобства дальнейшего изложения введем следующее определение: **информационные системы, способные на смысловом уровне обрабатывать знания, сформированные и вводимые пользователем в формате глубинной структуры, будем называть ОЗСУ-системами.**

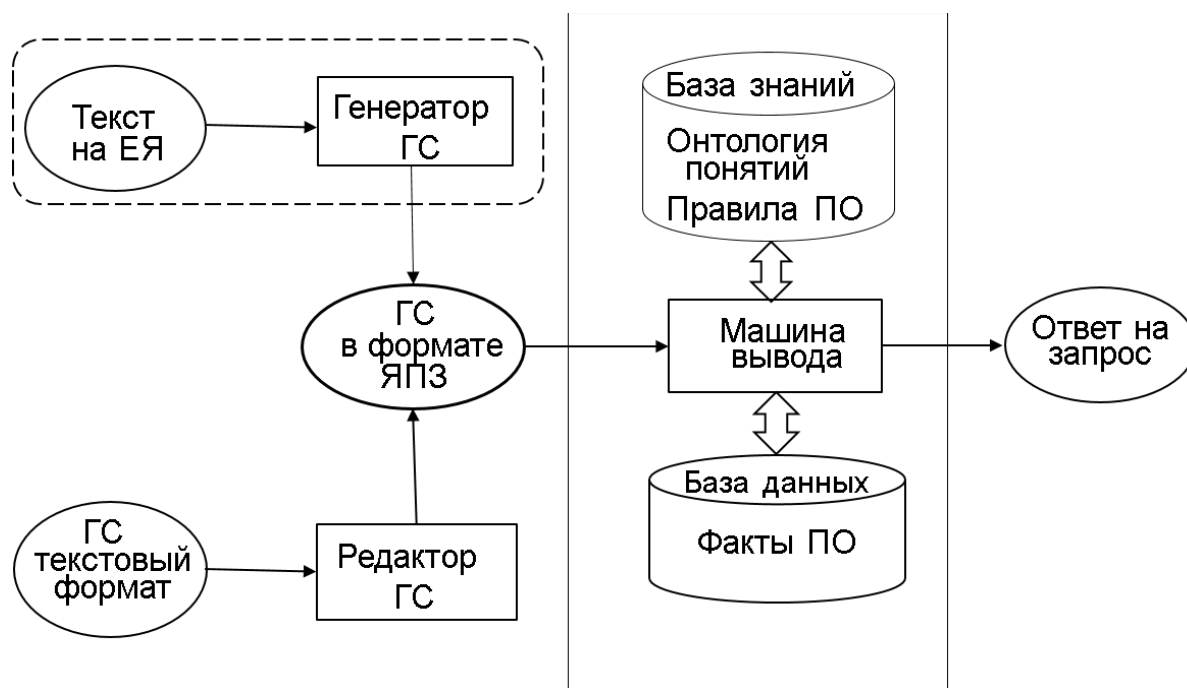


Рис. 1. Структура систем обработки знаний на смысловом уровне (ОЗСУ-систем)

Без потери общности, можно считать, что ОЗСУ-системы состоят из двух компонентов, которые условно можно назвать интерфейсом и ядром. Интерфейс обеспечивает формирование глубинной структуры (ГС) информации, вводимой пользователем, и передачу ее на вход ядра. Этой информации должно быть достаточно для формирования ответов на запросы и реализации операций создания и/или редактирования фактов и понятий, хранящихся в ядре в формате ЯПЗ. Поэтому ядро должно содержать отдельный компонент, машину вывода, назначением которой является интерпретация действий, закодированных в понятиях, представленных в формате глубинной структуры. Кроме машины вывода в ядро системы входят база знаний и база данных. В базе знаний хранятся определения понятий и правила предметной области (ПО), а в базе данных – конкретные факты ПО, выраженные с помощью этих понятий. Машина вывода, обрабатывая входные сообщения, манипулирует сведениями из базы данных под управлением базы знаний (рис. 1).

ОЗСУ-системы позволяют не только обращаться с запросами к имеющимся данным, но и генерировать новые факты предметной области. Реализация такой системы для конкретной предметной области осуществляется посредством определения структурных и семантических свойств понятий, включая условия целостности.

Отметим, что по своей функциональности ОЗСУ-системы принципиально отличаются от БД и не могут быть разработаны с помощью традиционных технологий для баз данных по нескольким причинам. Во-первых, понятия, в терминах которых были введены факты, могут не совпадать с понятиями, использованными в запросе, хотя по смыслу они могут быть тождественны запросу. Для решения этой проблемы система должна поддерживать логический вывод, с помощью которого сформированный в результате запроса ответ будет включать также факты, логически следующие из запроса. Во-вторых, реляционная модель не поддерживает отношений род-вид и часть-целое, которые широко используются при определении понятий. И, наконец, в-третьих, ОЗСУ-системы позволяют определять не только онтологию, но и правила предметной области, включая условия целостности. В частности – это «правила поведенческой активности», позволяющие предсказывать возможные события, которые могут произойти при текущем состоянии множества хранящихся фактов и правил.

Приведем пример работы системы, иллюстрирующий требуемую функциональность. Предположим, что в онтологии предметной области для учета успеваемости студентов определено понятие отличник, которое на ЕЯ может быть сформулировано следующим образом: Отличником называется студент, имеющий по всем предметам, по которым он обязан сдавать экзамены, оценку 5. Предположим также, что поступила информация: «Студент Иванов из группы Б22-104 является отличником», хотя оценки Иванова могут быть и неизвестны. Рассмотрим, какие последствия влечет за собой ввод в систему данного факта. Предварительно отметим, что если определены усло-

вия целостности, то ввести данный факт можно только в том случае, если в БД среди оценок Иванова (если они есть) не имеется оценок по предметам, которые он должен сдавать, отличных от 5. Далее: если обратиться с запросом об оценке Иванова по предмету, который он обязан сдавать, то система, за счет определения понятия отличник, выдаст ответ 5 даже в том случае, когда данные о результатах экзаменов отсутствуют. Если попытаться ввести отличную от 5 оценку Иванова по некоторому предмету, то будет зафиксировано противоречие. В ответ на запрос об имеющихся отличниках будут выданы, как явно определенные отличники (в данном случае Иванов), так и те, которые логически, исходя из определения, являются отличниками по результатам экзаменов. Если в учебном заведении существует правило, согласно которому каждый отличник получает повышенную стипендию, то соответствующая операция по изменению значения стипендии должна быть выполнена.

Принципиально важно, что все перечисленные функции реализуются автоматически, исходя только из формального определения понятия отличник. Не только определения, но и факты могут содержать информацию, которая на языке первого порядка требует использования кванторов всеобщности (например, Студент Иванов из группы Б22-104 сдал все экзамены на 4), а также кванторов существования.

Даже рассмотренный пример дает представление о сложности реализации указанного функционала ОЗСУ-систем.

## ОНТОЛОГИЯ ПРЕДМЕТНОЙ ОБЛАСТИ

Основной элемент онтологии – понятие. В настоящей работе предлагаются методы описания семантики понятий, дающие возможность поддержать основные способы определения понятий, которые применяются в человеческой практике. По способу определения понятия делятся на первичные и производные, смысловое содержание которых определяется через другие понятия. Производные понятия делятся на задаваемые либо с помощью прямого определения семантики, либо с помощью определения видовых отличий.

Приложения работают с отображениями реальных объектов и событий предметной области в виде фактов, хранящихся в БД. Структура фактов базы данных определяется с помощью понятий. Это означает, что определения понятий должны содержать информацию, с помощью которой может быть сконструирован запрос, позволяющий выбрать факты из БД и/или сформировать процедуру создания фактов в базе данных по их описаниям в формате глубинной структуры.

**Первичные понятия** – это простейший тип понятий. Смысловое содержание таких понятий определяется только их интерпретацией во внешней среде. Именно такой тип понятий используются в большинстве онтологий, представляя собой, по существу, обобщение понятия сущности в реляционной модели данных. Оно состоит в определении **атрибутов** с указанием доменов их значений и **операндов** – возможных связей с другими понятиями. Например,

понятие *студент* можно определить, как имеющее атрибуты: *Фамилия*, *Номер\_зачетной\_книжки* т.п., экземпляры которого должны быть связаны с экземпляром понятия *группа*. По умолчанию все связи в первичных понятиях имеют тип «многие ко многим», но при наличии информации о кардинальности связей она может быть указана в определении понятия. Определение понятия *Студент* можно записать следующим образом:

```
def_concept #Студент (Фамилия(Строка),  
Номер_зачетной_книжки(Число), учится.#Группа)
```

**Понятия-действия** могут описывать изменения значений атрибутов и/или связей некоторых объектов предметной области, представленных экземплярами соответствующих понятий. Например, понятие *Переезд* по смыслу означает изменение текущего адреса субъекта на новый адрес. Информация, содержащаяся в определении подобных понятий в принципе должна также включать описание действий, которые должны быть выполнены при создании экземпляра данного понятия в БД.

Но чтобы не нарушать принципа формирования онтологии, можно декларативную и функциональную части понятия размещать отдельно – декларативную в онтологии, а функциональную – в разделе правил (см. далее).

Для указанного примера это будет выглядеть следующим образом:

сначала идет декларативная часть –

```
def_concept #Переезд (Адрес_новый (Строка),  
переезжает#Человек),
```

а затем функциональная –

```
#Переезд (Адрес_новый, переезжает #Человек) -->  
#Человек (Адрес = Адрес_новый/).
```

В данном случае действие заключается в изменении значения атрибута *Адрес* экземпляра понятия *Человек*.

#### **Понятия с прямым определением семантики.**

Рассмотрим гипотетический пример понятия *Коррупционная\_сделка*, смысловое содержание которого можно описать следующим образом: «Предприниматель покупает некоторый товар по определенной цене, а потом продает его через небольшой промежуток времени по гораздо более низкой цене чиновнику». Нетрудно видеть, что это определение описывает в обобщенной форме факт, состоящий в следующем утверждении:

имеются экземпляры понятий *Торговая операция*, с одним и тем же *Товаром* и одним и тем же *Предпринимателем*, дополненные условиями, накладываемыми на значения их атрибутов *Цена* и *Дата*.

Обобщение состоит в том, что в описании условий и экземпляров этих понятий основные переменные не означены, что позволяет называть это описание фактом-прототипом. Если определить экземпляр понятия *Коррупционная\_сделка* как первичное, т.е. указать в качестве атрибутов *Цена*, *Цена/1* и *Дата*, *Дата/1*, а в качестве операндов – *#Предприниматель*, *#Чиновник*, *#Товар*, *#Торговая операция* и *#Торговая*

*операция/1* с именами связей покупка и продажа, то при этом будет потеряна информация о семантике понятия. Действительно, в соответствии с определением семантики операнды *#Предприниматель* и *#Товар* являются общими для обоих экземпляров понятия *#Торговая операция*, а значения атрибутов *Цена*, *Цена/1* и *Дата*, *Дата/1* должны совпадать со значениями соответствующих атрибутов операндов.

При поиске экземпляров понятий с прямым определением семантики, как в данном случае, система позволяет найти все экземпляры, которые являются виртуальными, но выводятся из имеющихся данных. Например, таким способом можно найти экземпляры понятия *Коррупционная\_сделка* в БД, содержащей факты о покупках и продажах. Если предположить, что определение понятия включает также атрибуты и операнды, которые относятся непосредственно к экземпляру понятия, например, это могут быть даты поступления информации, то указанные параметры должны быть включены в запрос.

Представление данного определения, в котором используется встроенный предикат *больше*, с лингвистической шкалой оценки степени, может иметь следующий вид:

```
def_concept #Коррупционная_сделка  
(#Торговая_операция (покупатель#Предприниматель,  
#Товар, Дата, Цена),  
#Торговая_операция/1 (продавец#Предприниматель,  
покупатель#Чиновник, #Товар, Дата/1, Цена/1),  
больше_незначительно (Дата/1, Дата),  
больше_существенно (Цена, Цена/1))
```

**Понятия, определяемые через видовые отличия**, – это наиболее распространенный способ определения понятия с помощью указания его видовых отличий по отношению к взятому за основу родительскому понятию. Отношение специализации (синоним отношения род-вид) образует дерево специализации. Корнем дерева является родовое понятие, все специализированные понятия называются его **категориями**, наиболее специализированные категории образуют листья дерева. В качестве родового понятия потенциально может быть любое понятие с прямым определением семантики. При манипулировании экземплярами категорий в зависимости от контекста имя категории может пониматься и как обозначение конкретного экземпляра, и как обозначение множества всех экземпляров данной категории. В такой интерпретации все экземпляры дочерней категории принадлежат родительской категории. Поэтому при описании категорий традиционно используется принцип наследования, распространяемый на атрибуты, связи и категориальные ограничения. Это означает, что если у родительского понятия определяются атрибуты (ограничения), то они принадлежат (справедливы) всем дочерним категориям этого понятия. То же самое – если у родительского понятия есть ограничения, то они справедливы для всех дочерних категорий.

Дерево специализации допускает определение не только альтернативных (например, таких категорий студентов, как отличники и неуспевающие, которые образуют таксономию на множестве всех студентов), но и пересекающихся категорий. Например, человек

может быть студентом, композитором и художником одновременно. Причем, у каждой из категорий может быть свой специфический набор атрибутов, связей и ограничений.

Кроме категорий, определенных через видовые отличия, можно создавать производные категории с помощью операций объединения, пересечения и дополнения первичных и/или ранее созданных производных категорий.

Абстрагируясь от способа реализации дерева специализаций, можно утверждать, что при создании экземпляра любой категории автоматически создаются экземпляры во всех категориях, которые являются предками данной категории в дереве специализаций. Для доступа к этому экземпляру можно использовать имя любой родительской категории. При этом возможна ситуация, когда созданный экземпляр принадлежит категории, которая может обладать большим числом атрибутов и операндов, чем определено в выбранной для доступа родительской категории. Поэтому для удобства в перечень параметров шаблона любой родительской категории допускается включать любые параметры, которые определены в дочерних категориях.

Например, можно задать вопрос: «Найти фамилию человека, который написал картину “Бурлаки на Волге”», хотя в формальном определении понятия *Человек* такого операнда как *написание\_картины* не определено. Аналогично, пусть вводится факт: «Картину “Бурлаки на Волге” написал человек по фамилии Репин», тогда будет создан экземпляр понятия-категории *художник* по правилу принадлежности операнда *написание\_картины* именно этому понятию.

Описание категорий, как и раньше, определяется указанием общего вида экземпляра категории, в котором представлены его видовые отличия. Кроме того, в описание добавляется и имя родительской категории.

Отметим, что, если определена ассоциативная связь между некоторым понятием *A* и каким-либо понятием, принадлежащим дереву специализации, то в сети экземпляров соответствующая связь может быть между экземпляром этого понятия *A* и экземпляром любого понятия этого дерева специализации.

Приведем примеры определений категорий. Пусть требуется определить понятие *Перевод*, понимаемое как текст, получаемый в результате выполнения перевода. Определение категории *Перевод*, как частного случая категории *Текст*, при всей очевидной условности примера может выглядеть следующим образом:

```
def_concept #Перевод (Father = Текст)
(результат#Текст, исполнитель#Переводчик,
  Качество(Число))
```

**Отношение между понятиями типа часть-целое** определяется достаточно просто и логично. Когда объявляется понятие-целое, то в его состав включается как атрибут понятие-часть. Причем имя связи не указывается. Например, для понятия *Книга*, состоящего из обычных атрибутов *Название*, *Число\_страниц* и имеющего связь с другим понятием *Автор*, может быть указано в качестве части понятие *Иллюстрация*.

Понятие *Иллюстрация*, естественно, имеет свои собственные атрибуты.

Вот как это будет выглядеть на описываемом языке представления знаний:

```
def_concept # Книга
(Название (Строка), Число_страниц (Число),
  автор#Человек, #Иллюстрация)
```

По умолчанию кардинальность отношения часть-целое определяется как один ко многим. При наличии более точной информации она может быть указана в определении понятия.

**Правила функционирования предметной области** описываются в форме продукций; каждая продукция традиционно имеет левую и правую части. Если истинна левая часть продукции, то должна быть выполнена правая часть. Чаще всего выполнение правой части – это либо генерация новых фактов, либо изменение атрибутов экземпляров понятий. Рассмотрим пример:

```
def rule #Экзамен (сдает#Отличник, Предмет) -->
  #Экзамен (Оценка = 5),
```

где утверждается, что если отличник сдавал какие-то предметы, то получал по ним исключительно оценки отлично, т.е. выполнение правой части продукции приводит к назначению или изменению значения атрибута *Оценка* у тех нескольких экземпляров понятия *Экзамен*, которые связаны отношением *сдает* с конкретным экземпляром понятия *Отличник*.

## БАЗА ДАННЫХ

Модель данных, ссылки на которую мы вынуждены были давать в предыдущем разделе, представляет собой концептуальную сеть. Элементами этой модели являются экземпляры понятий. Для поддержки хронологии экземпляры понятий могут быть определены как имеющие версии. Кроме того, экземпляры понятий могут быть связаны друг с другом с помощью неориентированных бинарных связей. **Факты** предметной области, на работу с которыми ориентирована модель, с содержательной точки зрения представляют собой утверждения о том, что некоторые экземпляры понятий, представленные своими характеристиками (атрибутами), связаны между собой определенными отношениями. Факты хранятся в БД в виде совокупности взаимосвязанных экземпляров понятий, описывающих факт.

Факты могут быть хранимыми и виртуальными. В первом случае идентификация тела факта реализуется посредством использования идентифицирующего экземпляра, который связан со всеми экземплярами понятий, составляющих факт. Хранимые факты необходимы для более быстрого доступа к экземплярам понятий, относящихся к конкретному факту. Это уже скорее элемент реализации.

Виртуальный факт – это чаще всего экземпляр производного понятия, обладающего семантикой. В качестве виртуальных можно рассматривать факты, которые являются результатами исполнения запросов. Для виртуальных фактов идентифицирующие их экземпляры не создаются.

Набор фактов, которые можно рассматривать как индикаторы событий, происходящих в системе, в определенном смысле описывает состояние этой системы. Совокупность знаний и фактов определяет поведение системы.

Кроме конкретных данных, например: «Студент Иванов получил оценку хорошо на экзамене по физике», факт может содержать информацию, для представления которой на языке исчисления предикатов требуется использовать кванторы. Так, в факте «Некоторые студенты группы Б22-104 сдали экзамен по физике досрочно» используется квантор существования. Хотя невозможно однозначно идентифицировать в БД экземпляры студентов, о которых идет речь, данное предложение несет определенную информацию, обеспечивая положительный ответ на вопрос о наличии студентов, сдавших досрочно экзамены в группе Б22-104. Еще один случай представляют факты, описание которых на языке исчисления предикатов первого порядка требует использования кванторов всеобщности, как, например, в факте «Студент Иванов из группы Б22-104 сдал все экзамены на отлично».

Для доступа к данным используется техника, основанная на процедуре сопоставления с образцом. Безусловно, выбор нотации для представления образцов имеет важнейшее значение. Образцы описываются с помощью, так называемых **концептуальных логических выражений**. Предлагаемая здесь нотация основана на нотации пролога. Логические операции изображаются символами «,» (и), «;» (или) и «not» (не). Используются также круглые скобки и

символ /, в основном как ограничители. Основным элементом выражения является **терм**, который состоит из идентификатора экземпляра понятия с префиксом «#» и **шаблона**, записываемого справа от идентификатора в скобках. Шаблон содержит атрибуты и идентификаторы других экземпляров понятий, с которыми связан описываемый экземпляр, называемые далее переменными. Если значения атрибутов известны, то они записываются после символа «=», следующего за соответствующим атрибутом. Значения идентификаторов экземпляров не устанавливаются пользователями, а присваиваются системой при создании экземпляра. Данные в БД описываются выражениями, в которых все атрибуты означены (рис. 2).

Приведем пример текстового описания структуры данных, представленной на рис. 2:

```
#Студент (Фамилия = 'Иванов', учится.  
#Группа (Номер = 'Б22-104'), сдает.#Экзамен  
  (Предмет = Физика, Оценка = 4),  
сдает.#Экзамен/1 (Предмет = Матем, Оценка = 5))
```

Если необходимо использовать два разных экземпляра одного понятия, то к переменным, представляющим идентификаторы разных экземпляров, должны добавляться символ «/» и любые дополнительные символы. При этом данное правило распространяется и на атрибуты.

Потенциально переменные в системе могут иметь три состояния: быть неозначены, иметь неопределенное значение, и конкретное значение, соответствующие типу.

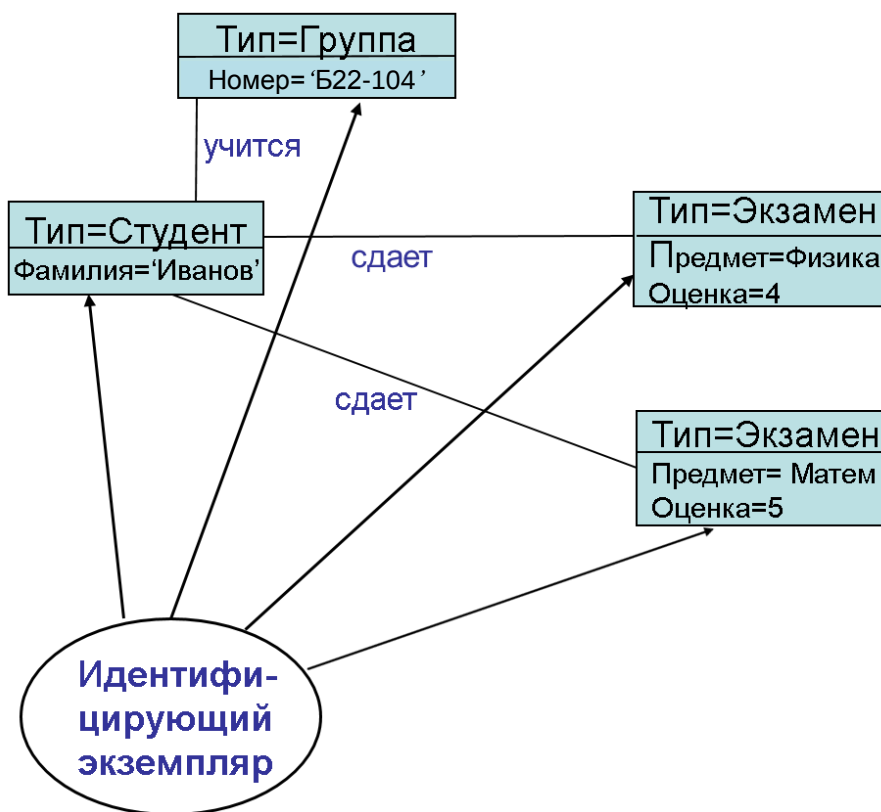


Рис. 2. Факт и идентифицирующий его экземпляр

Важнейшее свойство выражений – это возможность их интерпретации в качестве логических выражений, в которых предикатами выступают термы, а символы «,», «;», «not» обозначают соответствующие логические операции. Терм представляется как предикат, определенный на множестве экземпляров понятия в БД, имя которого задано в идентификаторе экземпляра. По определению терм выражения, называемый образцом, принимает значение *T* или *F* на выбранном экземпляре понятия, в зависимости от того «унифицируется» или нет шаблон терма с этим экземпляром.

**Унификация** между образцом и соответствующим термом базы данных определяется следующим образом:

1) множество переменных в образце должно быть подмножеством множества соответствующих переменных в экземпляре БД, с которым выполняется унификация;

2) унификация терма образца состоит в последовательной унификации переменных образца с соответствующими переменными экземпляра БД, начиная с переменной в идентификаторе экземпляра;

3) если у экземпляра понятия образца имеются операнды, то унификация рекурсивно выполняется для каждого операнда при условии, что повторные вхождения переменных в них имеют значения, которые они получили ранее;

4) унификация завершается успехом при любом значении переменной экземпляра БД, если соответствующая переменная образца не означена. При этом она получает значение этой переменной экземпляра. Если переменная образца означена, то соответствующая переменная экземпляра БД должна иметь такое же значение;

5) во всех остальных случаях унификация завершается неудачей.

Доступ по образцу применяется также для доступа к левым и правым частям правил базы знаний в ходе логического вывода.

Интерпретация термов, как логических выражений, а также отсутствие ориентированности связей дают возможность проведения эквивалентных преобразований концептуальных логических выражений. Так, порядок перечисления термов и/или конструкций внутри шаблонов произволен. Установить тождественность двух выражений можно путем приведения их к единственной с точностью до порядка записи термов канонической форме.

Например, если сформулировать утверждения о студенте следующим образом: «Студент Иванов сдавал экзамены по математике и физике. По математике он получил 4, а по физике 5. Он учится в группе Б22-104», то глубинная структура может иметь следующий вид:

```
#create_fact
(#Экзамен(сдает.#Студент(Фамилия = 'Иванов'),
Предмет = Физика),
#Экзамен/1(сдает.#Студент(Фамилия = 'Иванов'),
Предмет = Матем.)),
#create_fact(#Экзамен(Оценка = 4), #Экзамен/1
(Оценка = 5)).
```

```
#create_fact(#Студент(Фамилия = 'Иванов',
#Группа(Номер = 'Б22-104'))).
```

Для того чтобы узнать оценку по физике студента Иванова, достаточно сформировать концептуальное выражение:

```
#Экзамен (сдает.#Студент (Фамилия = 'Иванов'),
Предмет = Физика, Оценка),
```

либо:

```
#Студент (Фамилия = 'Иванов', сдает.#Экзамен
(Предмет = Физика, Оценка)).
```

Отметим, что концептуальные логические выражения могут использоваться не только в логической интерпретации, но и как специальная команда доступа к данным. В этом случае, при передаче ей управления, делается попытка означить переменные термов с помощью выборки первого по порядку экземпляра из базы данных, который унифицируется с ним. При возвратах переменным – атрибутам или операндам – ставятся в соответствие следующие по порядку подходящие значения или экземпляры.

## ЗАКЛЮЧЕНИЕ

В настоящей работе описывается структура системы обработки знаний и обосновываются требования, которым должны удовлетворять языки представления знаний. В состав системы входят база знаний, база данных и машина логического вывода. БД представляет собой совокупность фактов, хранящихся в системе в виде сети экземпляров понятий. Универсальные знания представляются в виде онтологии предметной области, состоящей из совокупности определений понятий, условий целостности и правил функционирования. Концептуальное описание понятий дает возможность представлять и обрабатывать информацию на смысловом уровне. Хотя настройка на предметную область предполагает явное построение всех определений, она принципиально может выполняться и в режиме обучения по примерам. Предлагаемый подход позволяет использовать один язык для ввода не только фактов и запросов, но и для определения понятий и правил, что в принципе помогает описать глубинную семантическую структуру предложений, сформулированных на естественном языке. Перспективы настоящей работы во многом зависят от решения задачи построения глубинной структуры текста на ограниченном естественном языке.

## СПИСОК ЛИТЕРАТУРЫ

1. Курвилль А., Гудфеллоу Я., Бенджио И. Глубокое обучение. – Москва: ДМК Пресс, 2018.
2. Николенко С., Кадури А., Архангельская Е. Глубокое обучение. Погружение в мир нейронных сетей. – Санкт-Петербург: Питер, 2018.
3. Nilsson N. The Quest for Artificial Intelligence. A History of Ideas and Achievements. – Cambridge: Cambridge University Press, 2021.

4. Гаврилова Т., Хорошевский В. Базы знаний интеллектуальных систем. – Санкт-Петербург: Изд-во Питер, 2001. – 384 с.
5. Рассел С., Норвиг П. Искусственный интеллект: современный подход. – 4-е изд. – Москва: Диалектика, 2006.
6. Осипов Г. Методы искусственного интеллекта. – Москва: Физматлит, 2011. – 296 с.
7. Осипов Г. Лекции по искусственному интеллекту. – Москва: Ленанд, 2018. – 272 с.
8. Sergievsky G. Concept of Inductive Programming Supporting Anthropomorphic Information Technology // Journal of Computer and Systems Sciences. – 2011. – Т. 50, № 1. – Р. 38-50.
9. Сергиевский М.В. Шаблоны унифицированного языка моделирования для проектирования информационных систем // Научно-техническая информация. Сер. 2. – 2020. – № 1. – 19-27; Sergievsky M. Modeling Unified Language Template for Designing Information Systems // Automatic Documentation and Mathematical Linguistics. – 2020. – Т. 54, № 1. – Р. 26-31.
10. Bordini R., Hubner J., Wooldridge M. Programming Multi-Agent Systems in AgentSpeak Using Jason. – WILEY, 2007. – 294 p.

*Материал поступил в редакцию 31.01.23.*

#### **Сведения об авторах**

**СЕРГИЕВСКИЙ Георгий Максимович** – кандидат технических наук, доцент, независимый исследователь

**СЕРГИЕВСКИЙ Максим Владимирович** – кандидат технических наук, доцент, доцент Национальный Исследовательский Ядерный Университет МИФИ, Москва.  
e-mail: sermax@yandex.ru