ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 4** **Москва 2017**

ОБЩИЙ РАЗДЕЛ

УДК 14 : 001.102

Е.А. Плешкевич

Информация как реальность в неживой природе или атрибутивная концепция информации 2.0: проблемы и перспективы развития

Проанализирована атрибутивная концепция «четырех миров реальности» предложенная К.К. Колиным, доказывающая существование информации в физическом мире неживой природы в контексте «мира идеальной реальности первого рода». Показана, ошибочность данного суждения.

Ключевые слова: *информация, информационный подход, идеальное и материальное, атрибутивная концепция информации, концепция «четырех миров реальности»*

ВВЕДЕНИЕ

Информатизация сегодня сопровождается усилением интереса к природе информации. Ярким примером этого стала дискуссия, развернувшаяся на страницах сборника «Научно-техническая информация» и других изданий. В этой дискуссии принимали участие известные советские и российские ученые А.Д. Урсул и К.К. Колин [1-3], предложившие обновленную версию атрибутивной концепции информации. Напомним, что в контексте атрибутивной концепции, сложившейся еще в 1960-1970-х гг., информация рассматривается как всеобщее свойство материи, ее атрибут и что она существует не только в живой природе и социуме, но и в неживой природе. При этом информация связывается с отражением, разнообразием, неоднородностью. Альтернативная атрибутивной, функциональная концепция также связывает природу информации с отражением и разнообразием, однако предполагает, что информация или, вернее, информационные процессы могут иметь место только в самоорганизующихся и самоуправляемых системах в живой природе, социуме и кибернетических системах, созданных человеком.

АТРИБУТИВНАЯ КОНЦЕПЦИЯ ИНФОРМАЦИИ

Определенный успех в развитии именно атрибутивных представлений об информации, А.Д. Урсул связывает, во-первых, с результативностью использования отдельных положений атрибутивной концепции в исследованиях в области космологии, химии и других естественных наук (при этом часто эта результативность рассматривается им как подтверждение обоснованности атрибутивного видения информации), а во-вторых, с появлением концепции «четырех миров реальности», разработанной К.К. Колиным и подтверждающей обоснованность факта всеобщности информации ее существованием в неживой природе[1]. Рассмотрим, насколько справедливы эти аргументы.

Представляется, что в первом случае мы имеем дело с определенной подменой теории информации информационным подходом. Теория информации как совокупность базовых представлений о ней, описывающих ее уникальные характеристики и свойства, ее генезис всегда содержит определенные противоречия по тем или иным пунктам. В отличие от нее информационный подход – это совокупность некоторых методологических положений, обеспечивающих изучение информационных аспектов явлений действительности. Таким образом, если во главе угла концепции стоит постижение сущности явления, то информационный подход ориентирован на методологическую эффективность и универсальность. Ради достижения этих целей можно оставить за скобками отдельные спорные положения этой концепции. Например, в средние века методика астрономического измерения широты в мореплавании с использованием астролябии и секстанта базировалась на геоцентрической концепции Птолемея. Однако доказанная эффективность применения некоторых положений концепции Птолемея не оказалась решающим аргументом в споре с гелиоцентрической концепцией Коперника. В принципе, это касается и информационного подхода. В свое время В.С. Тюхтин отметил, что информационного процесса в неживой природе нет, но применение информационных методов при ее изучении вполне допустимо и может быть полезно, если эти методы не абсолютизируются [5, с. 88–91]. Таким образом, обновление атрибутивной концепции мы связываем с разрешением ее внутренних противоречий, а не с расширением области использования ее отдельных положений. И в этом аспекте нас в первую очередь привлекает анализ концепции «четырех миров реальности».

А.Д. Урсул отметил, что в основе концепции К.К. Колина лежит его же идея «о том, если одной из общих закономерностей природы является ее стремление к равновесию, симметрии, равномерному распределению материи и энергии, то такая тенденция реализуется с помощью движения. Это означает, что первопричиной этого движения выступает информация, поскольку именно информация (выступающая в форме асимметрии, неравномерности распределения материи и энергии) ведет к появлению движения от ее высокой концентрации, к тем областям пространства, где эта концентрация является более низкой» [1, с. 3]. Для нас не вполне понятны эти «стремления» природы к равновесию и равномерности. Безусловно, эти явления существуют, например, при распределении тепловой энергии, когда при взаимодействии более нагретое тело остывает, а менее нагретое – нагревается до тех пор, пока их температуры не выровняются. Однако существуют и другие физические явления, описываемые законом всемирного тяготения, когда более крупные тела притягивают менее крупные, что ведет к концентрации вещества и, соответственно, энергии. Более того, заполнение пустоты имеет случайный и хаотичный характер. В условиях земного тяготения вода заполняет пустоты, однако не потому что она стремится их заполнить, а потому, что ее притягивает Земля, когда же воду притягивает Луна, она выходит из этих пустот. Более того, согласно общей теории относительности, пространство вследствие гравитационного искривления деформировано, т.е. неоднородно и в нем присутствуют так называемые «черные дыры». Иными словами стремление к равномерности неочевидно и требует доказательства. Это, во-первых. Во-вторых, допустим, что информация ведет движение материи из областей с ее высокой концентрацией в области низкой концентрации, тогда кто или что ведет движение в обратную сторону? Или обратного движения не существует даже на локальном уровне? И, наконец, в третьих, если информация имеет форму асимметрии и неравномерности распределения материи и энергии, то, в конце концов, достигнув равномерности и симметрии распределения материи, информация должна исчезнуть, хотя бы локально.

Сегодня, пишет И.С. Пригожин, мы знаем, что увеличение энтропии отнюдь не сводится к увеличению беспорядка, ибо порядок и беспорядок возникают и существуют одновременно [6, с. 48]. Так, может быть равномерность и неравномерность, симметрия и асимметрия, равновесие и отсутствие такового возникают и существуют одновременно и никто никуда не стремится, а движение идет случайным образом и в обоих направлениях? Может, мы имеем дело с явлением рассеивания материи и энергии, может, движение – это естественное состояние материи, вектор движения которой задается импульсами большого и других взрывов во вселенной, гравитационными силами, и сопровождающие эти процессы различия являются следствием, а не первопричиной? Может различия сами по себе могут породить лишь большие или меньшие различия и не стоит умножать сущности без надобности? Определенно сказать сложно, в любом случае поиск ответов за автором концепции.

КОНЦЕПЦИЯ «ЧЕТЫРЕХ МИРОВ РЕАЛЬНОСТИ» К.К. КОЛИНА

Первая особенность концепции «четырех миров реальности»связана с тем, что сама идея выделения особых миров была заимствована К.К. Колиным у К. Поппера. Напомним читателям, что К. Поппер, исследуя мир науки, выделил три его формы: 1) мир физических объектов и состояний; 2) мир психиче-

[1] Частично мы уже касались концепции К.К. Колина [4], и в настоящей статье мы продолжаем ее рассмотрение.

ских и ментальных состояний; 3) мир объективного знания, куда входят содержание научных гипотез, литературные произведения и т.д. Мир физических объектов взаимодействует с миром психических состояний, а тот порождает мир объективного знания, который не зависим от своих создателей. Таким образом, знание по К. Попперу не зависит от познающего субъекта. Вторая особенность концепции К.К. Колина заключается в связи информации с идеальными (нематериальными) структурами реальности. «Важно отметить, пишет он, что информация не является материальным объектом или процессом. Она является феноменом идеальной реальности» [3, с.56].

Построение этой концепции начинается с разделения парадигм реальности на материально-энергетическую и материально-информационную. В рамках первой парадигмы реальность состоит из движущейся материи и энергии, помимо которых больше ничего не существует. Вторая парадигма включает физическую (материальную) и идеальную (нематериальную) реальности. В структуру идеальной реальности К.К. Колин включает три компонента:

1) объективная идеальная реальность первого рода (ИР–1) неразрывно связана с физической реальностью и возникает в результате взаимодействия отдельных компонентов физической реальности и их взаимного отражения, что, собственно, и обусловливает существование всего многообразия окружающего нас мира. Необходимо специально отметить, что этот вид идеальной реальности называется объективным, потому что он напрямую не связан с деятельностью сознания человека и не является продуктом этой деятельности, а порождается физической реальностью в результате действия всеобщего закона отражения;

2) субъективная идеальная реальность (ИР–2) включает феномен сознания человека, а также продукты деятельности сознания, существующие внутри него;

3) объективная идеальная реальность второго рода (ИР–3) включает всю совокупность нематериальных продуктов деятельности сознания, находящихся вне него. Сюда, в частности, относятся нематериальные продукты культуры и искусства, а также наука, религия и т. п. [2].

Сопоставляя схемы миров К. Поппера и К.К. Колина, можно увидеть, что к имеющимся трем мирам К. Поппера добавлен четвертый – мир объективной идеальной реальности первого рода, который, по мнению К.К. Колина, служит своеобразным каналом информационного взаимодействия физического мира и мира сознания. Как полагает К.К. Колин, этот новый мир, во-первых, возникает в результате взаимодействия материальных объектов и представляет собой совокупность их взаимных отражений в процессе этого взаимодействия; во-вторых, он не только обеспечивает взаимодействие между физическим миром и миром сознания, но и служит ареной взаимодействия объектов физической реальности; в-третьих, если бы этого мира не существовало, то никакие взаимодействия материальных объектов в природе вообще не были бы возможными.

Опираясь на эти умозаключения, К.К. Колин делает следующие выводы: поскольку информация является всеобщим свойством реальности, то она существует не только в сознании человека, но и в физическом мире в виде так называемой связанной информации, которую иногда называют также «физической» или «структурной» информацией. Концепция «четырех миров реальности» принципиально позволяет с единых концептуальных позиций (всеобщности информации как атрибута всех компонентов реальности) изучать не только общие закономерности, но и специфику проявления феномена информации в различных компонентах реальности. В итоге он приходит к выводу, что информация не является материальным объектом или процессом, а представляет собой идеальный феномен реальности. При этом идеальные процессы возникают в результате взаимодействия физических процессов и представляют собой отражения последствий этого взаимодействия [2].

СУЩЕСТВУЕТ ЛИ ИДЕАЛЬНЫЙ МИР РЕАЛЬНОСТИ ПЕРВОГО РОДА?

Аргумент в пользу атрибутивной природы информации строится на утверждении существования мира «идеальной реальности первого рода» и порождаемой им информации. Насколько обоснован этот аргумент? В качестве доказательства существования этого мира К.К. Колин предлагает провести мыслительный эксперимент с физической реальностью.

«*Рассмотрим фрагмент физической реальности,* пишет он*, в котором содержатся два материальных объекта: А и В. Предположим, что объект А представляет собой шар из пластичного материала, а объект В — это шар для игры в бильярд, более твердый по своей консистенции по сравнению с объектом А. Предположим далее, что оба рассматриваемые нами объекта приведены в соприкосновение некоторым усилием, а затем вновь разъединены.*

В результате этого взаимодействия пишет К.К. Колин*, на поверхности объекта А образовалась вмятина С, которая представляет собой след, оставленный объектом В на поверхности объекта А. Что можно теперь сказать о результатах данного взаимодействия двух материальных объектов с точки зрения изменения структуры рассматриваемого нами фрагмента реальности? Оказывается, что эта структура изменилась весьма существенным образом. И дело не только в том, что изменилась форма поверхности объекта А, которая теперь имеет вмятину. Принципиально важным является другое, а именно то, что эта вмятина представляет собой отображение (зеркальную копию) той части объекта В, которая входила в соприкосновение с объектом А. Таким образом, можно утверждать, что в наблюдаемом нами фрагменте реальности, где ранее находились лишь два материальных объекта, в результате их взаимодействия возник третий объект С, который представляет собой след (вмятину), являющийся отображением некоторой части поверхности объекта В на поверхности объекта А. Этот новый объект С не является материальным, и поэтому он должен рассматриваться как объект идеальной реальности. Но ведь этот объект реально существует, он не является плодом нашего вообра-*

жения. Это не ментальный продукт деятельности нашего сознания, а вполне реальный и объективно существующий результат взаимодействия материальных объектов физической реальности. Следовательно, продолжает К.К. Колин, *объективная идеальная реальность первого рода существует»* [2, с.142].

Однако мы усомнились в качестве проведенного мыслительного эксперимента, в его результатах и решили его повторить. Начнем с того, что источником, первопричиной этого взаимодействия выступает внешнее физическое воздействие на объекты А и В (на оба или на один из них), в результате которого им был придан импульс (сообщена некоторая скорость), который определил силу последующего взаимодействия и повлиял на результат столкновения. Взаимодействие случилось в определенном пространстве и имело форму механического столкновения объектов А и В, в результате чего произошли: 1) пластическая деформация поверхности соприкосновения, форма которой отражает геометрическую форму объекта В и величину энергии взаимодействия; 2) деформация общей геометрической формы объектов, либо увеличение плотности одного из них; 3) нагревание объектов и выделение части тепла в атмосферу и т.д. Суммируя, можно сказать, что в результате столкновения объекты А и В в большей или меньшей степени изменили свою геометрическую форму (полностью или частично), вес, плотность и т.д. Но, учитывая, что опыт все же мыслительный, остановимся на том, что изменения коснулись только объекта А, который трансформировался в объект А'. В итоге можно сказать что в результате взаимодействия возникла новая физическая реальность включающая А' и В.

Сопоставим описания эксперимента. Во-первых, из всего набора физических изменений, вызванных механическим столкновением, в отличие от нас, К.К. Колиным выбрано только одно изменение в виде деформации. Как мы знаем, мыслительный характер опыта не исключает комплексность рассмотрения физики явления, поэтому ограничение выбора нуждается в пояснении. Во-вторых, так называемая зеркальная копия отражает не только геометрию шара, но и величину импульса. В-третьих, что такое «вмятина С», введенная К.К. Колиным в описание объекта А, и как благодаря ней возникла новая«идеальная» реальность?

Согласно нашему описанию мыслительного эксперимента произошло механическое взаимодействие фрагментов материи (объекты А и В) и энергии, приведшее к трансформации фрагмента физической реальности, включающей объекты А' и В. Согласно К.К. Колину, реальность также изменилась, но содержание этих изменений иное: они коснулись появления на объекте А деформации в виде вмятины, геометрическая форма которой симметрично отображает геометрическую форму части В. Ссылаясь на симметрию К.К. Колин идентифицирует вмятину как новый объект С, имеющий идеально реалистическую природу.

Итак можно увидеть, что в обоих описаниях мыслительного эксперимента отмечена деформация части объекта А, геометрическая форма которой отчасти определена геометрической формой объекта В, т.е. в этой части мы едины. Различия появляются в описании картины реальности, в интерпретации места деформации. В нашей картине реальности, назовем ее физической, место деформации в форме вмятины является частью объекта А', а в описании К.К. Колина – самостоятельным объектом С.

Чем же обусловлена эта разница? Мы полагаем, что она обусловлена мыслительной фрагментацией К.К. Колиным картины физической реальности. Она касается фрагментации реального объекта А' посредством мыслительного вычленении из него фрагмента деформации (вмятины) как отдельного объекта С. Поясним нашу мысль подробнее. Мы представляем объекты А и В, поскольку ранее видели их прототипы, т.е. реально существующие самостоятельные материальные объекты. Как мы представляем себе вмятину или след? Как вмятину в чем-то, как след от чего-то. Иными словами, вмятина – это часть пространства, отграниченного физическими границами материального объекта от остального пространства. Мы, безусловно, можем так же фрагментировать и пространство, обозначив его в качестве самостоятельного объекта. Таким образом, объект С – это результат мыслительной фрагментации объекта А' и пространства вокруг него. Проиллюстрируем эту мысль. Допустим в начале эксперимента, расположив объекты А и В друг против друга, мы сфотографировали их. На полученной фотографии мы можем увидеть сферические объекты А и В и пространство вокруг них. Вторая фотография сделана после столкновения. На ней объекты А' и В и все то же пространство, но ввиду вмятины изменилась геометрия одного из них, т.е. в той части, где была материя, изображено пространство. И вот здесь К.К. Колин совершает определенные действия, а именно – берет маркер и выделяет участок с изменениями, т.е. фрагментирует изображенный на фотографии объект А' и пространство вокруг него. Таким образом, реально существует объект А' и пространство вокруг него, а вот объект С – это результат мыслительной фрагментации, проведенной после столкновения. При этом можно было и до столкновения фрагментировать любой участок пространства или объектов на фотографии и присвоить им статус объектов.

Итак, объект С – это идеализация фрагмента реальности в виде части материального пространства ограниченного контурами вмятины в объекте А'. Иными словами, **сам фрагмент пространства, занимаемого вмятиной, материален, идеально его выделение и описание**.

Что касается соотнесения геометрической формы вмятины и геометрической формы части объекта В, то К.К. Колин заостряет внимание на принципиальной зеркальной геометрической симметрии. Как мы уже отметили, произошло воздействие одного объекта на другой, результатом которого стала деформация одного из них. Несложно предположить, что характер деформации может быть обусловлен материальными свойствами объектов (геометрической формой и плотностью), а также силой их столкновения. К.К. Колин отмечает, что геометрическая форма вмятины отображает геометрию объекта В. Иными

словами, сравнение частей объекты B и A' свидетельствует, что их взаимодействие носило материально-энергетический характер. Однако можно сказать, что геометрическая форма вмятины идеальна, т.е. максимально точно и близко отражает геометрическую форму части объекта B, участвовавшей в столкновении. Но тогда мы просто используем понятие идеальное в ином значении, при этом идеальность сферического соответствия является следствием пластичности (мягкости) вещества объекта A.

ЗАКЛЮЧЕНИЕ

Резюмируя можно сказать, что эксперимент мог протекать в природе исключительно в материально-энергетической форме. Природа деформации и ее геометрическая форма обусловлены физическими свойствами материи. Фрагментация деформации и ее обозначение в качестве самостоятельного объекта C имеют внешний по отношению к эксперименту, характер. Иными словами, мы не увидели и не можем увидеть его самостоятельно, о нем нам могут только сообщить. Таким образом, феномен **идеальной реальности в данном эксперименте нами не выявлен**. Соответственно, открытый К.К. Колиным так называемый «мир идеальной реальности первого рода», который должен служить доказательством атрибутивной природы информации, в природе не существует.

Нам представляется, что дальнейшее развитие атрибутивной и функциональной концепции информации может идти в следующих направлениях. Одно из них связано с дальнейшим развитием этих концепций путем устранения выявленных ранее противоречий. Однако в этом плане концепция К.К. Колина не делает попыток к их разрешению[2]. Второй путь связан с осознанием бесперспективности «войны» атрибутистов и функционалистов. Может быть, хватит, образно выражаясь, «катать шары» в пользу той или иной концепции, может быть лучше «расширить» горизонт рассмотрения проблемы? Предпосылки для этого мы наблюдаем в современных исследованиях в области космологии, цифровой и квантовой физики, развивающих антропный принцип, согласно которому «мы видим Вселенную такой, потому что только в такой Вселенной мог возникнуть наблюдатель, человек». Возможно, выход на новый, более высокий уровень обобщения, высветит новые аспекты природы информации, и теоретический конфликт будет разрешен.

СПИСОК ЛИТЕРАТУРЫ

1. Урсул А.Д. Информация и информационный подход: от информатики к глобалистике // Научно-техническая информация. Сер.1. – 2012. – №2. – С. 1–11.
2. Колин К.К. Философия информации и структура реальности: концепция «четырех миров» // Знание, понимание, умение. – 2013. – №2. – С. 136–147.
3. Колин К.К. Философские тезисы о природе информации // Вестник Международной академии наук (Русская секция). – 2015. – №1 (17). – С. 52–58.
4. Плешкевич Е.А. Дискуссия о природе информации и путях построения ее философской концепции (обзор) // Научно-техническая информация. Сер.1. – 2015. – №4. – С.14–18.
5. Тюхтин В.С. Отражение, системы, кибернетика. Теория отражения в свете кибернетики и системного подхода. – М. : Наука, 1972. – 256 с.
6. Пригожин И.С. Философия нестабильности // Вопросы философии. – 1991. – №6. – С. 46-52.
7. Дубровский Д.И., Вержбицкий В.В. Категории информации: философский обзор // Философские науки. – 1976. – №1. – С. 148-157.

Материал поступил в редакцию 31.10.16.

Сведения об авторе

ПЛЕШКЕВИЧ Евгений Александрович – доктор педагогических наук, главный научный сотрудник Государственной публичной научно-технической библиотеки СО РАН, г. Новосибирск
e-mail: eap1966eap@mail.ru

[2] Основные аргументы против атрибутивной концепции информации был предложен еще в 1970-х гг. [7].

УДК 16 : 001.12

Д.П. Скворцов

О ‘квантово-механической’ природе недоучёта модально-временных соображений при анализе некоторых логических парадоксов*

Каждое точное определение мира просто обязано быть парадоксом.

Станислав Ежи Лец.
Непричёсанные мысли

Предложено объяснение некоторых логических парадоксов[1], исходящее из идеи изменения ситуаций (возможных миров).

Под подразумеваемым здесь ‘квантово-механическим’ эффектом имеется в виду одна известная популярная интерпретация (хотя, утверждается, и ошибочная[2]) ‘принципа неопределённости’ Гейзенберга: невозможность одновременного измерения двух параметров (в канонических изложениях, например: координаты и импульса) частицы (при возможности замерить любой из них по отдельности) мотивируется тем, что само измерение одного влияет, меняет состояние всей системы в целом, в результате чего второе измерение (другого параметра) определяет его уже в новом состоянии, где значение первого вообще говоря уже вовсе не обязано совпадать с ранее замеренным (в состоянии предыдущем: что минуло, канув в небытие). Такой эффект зачастую не замечаемого и упускаемого из виду изменения состояний, перехода из одного в другое, вовсе отличное от первоначально исходного, но с которым оно ненароком по недосмотру смешивается и перепутывается – именно это и есть тот модальный аспект: идея восприятия, рассмотрения ‘возможных миров’ (восходящая к Лейбницу – или с ним связываемая?), о котором пойдёт здесь речь. Иногда этот эффект проявляется, выражается, оформляется в виде времени, то есть какого-то протекающего, развёртывающегося процесса – но отнюдь не всегда и не обязательно, и не в форме же суть. Ну а если этот хитрый и тонкий аспект не учитывать, то наложение и слияние различных несовместимых между собой состояний в как бы одно единое (единственное!) – оно и способно породить видимость, ‘кажимость’ противоречия: которое мы, наивно обескураженные, и зовём тревожно пугающим словом ‘парадокс’.

Это как если бы пёструю, разноцветную спираль, постепенно старательно карабкающуюся, взбирающуюся ввысь по поверхности цилиндра, вдруг всю разом взять, да и скинуть, обрушить, спроектировать вниз на плоскость: принудив, заставив её, словно в беличьем колесе, безнадёжно бегать по кругу (ну точнее, по одной и той же окружности) – а тогда слияние вместе, в одной точке многообразия несовместимых цветов или звуков со всех разных уровней, витков спирали и породило бы неизбежное ощущение неразберихи и какофонии, то есть опять же: парадокс.

Ну а теперь давайте попытаемся это увидеть и проследить на паре отдельных примеров.

Начнём с ‘судебного парадокса’ (в англоязычных источниках: Paradox of the Court), называемого иначе ‘парадоксом (или софизмом) Протагора и Эватла’[3] .

---

* Этот текст представляется своего рода стилистическим продолжением, подобным предыдущей публикации автора: «Два курьёзных замечания на околоматематические темы» (НТИ. Сер.2. – 2009. – №4. – С. 12-15) – при том, что никакой зависимости или связи по содержанию между ними и нет. Есть ещё одно обстоятельство, то ли сближающее, то ли различающее эти два текста: если тот был написан в новогоднюю ночь на 1-е января 2009 г., то этот задуман и начат в «чёрную пятницу» 13 ноября 2015 г., что могло коварным образом наложить на него некий зловещий свой отпечаток (за что и приношу читателям свои извинения).

[1] быть может, ‘кажущихся парадоксов’? – что и хотелось здесь попытаться уяснить, пояснить или прояснить...

[2] Физики могут возразить, что понимание (или объяснение), описываемое в нескольких ближайших строках (вслед за этой сноской!) на деле не имеет реального отношения ни к настоящей квантовой механике, ни к её принципу Гейзенберга; но поскольку предмет нашего обсуждения отнюдь не квантовая теория, а анализ нескольких логических парадоксов, и для нас полезна именно такая интерпретация, то для краткости и выразительности именно её и будем здесь называть (пусть неточно и неадекватно) ‘принципом Гейзенберга’.

[3] хотя иногда участников ситуации разделяют, именуя: кто ‘парадоксом Эватла’, а кто-то – Протагора

Сюжет вкратце: второй получил у первого ряд уроков (кажется, по адвокатскому красноречию, или как это у них там называлось?) и договорился расплатиться за учёбу после первого им выигранного судебного процесса, но прослушав курс, юридической практикой заниматься не стал, надеясь компенсировать отсутствие заработка возможностью не платить учителю (вроде как бы и не за что?). Однако тот, исчерпав свой запас терпения (или может, обидевшись?), решил всё же ‘стряхнуть’ свои денежки с нерадивого ученика, и предъявил тому иск об оплате. Тем самым поставив перед судом – согласно традиционно принимаемой интерпретации парадокса – неразрешимую коллизию (или дилемму): отказ в иске, то есть позволение ученику не платить, сам становится выигранным им делом, и влечёт обязанность расплатиться, а напротив, удовлетворение иска о платеже станет для того проигрышем, и значит, платить опять-таки не за что. Разумеется, ясное дело, каждый из оппонентов пытается трактовать коллизию в свою пользу. Протагор небескорыстно заявляет: ты в любом случае должен мне деньги – или, в первом случае, по нашему уговору (дело ты ж выиграл), ну а во втором: постановленьем суда. На что уклонист, разумеется, тотчас парирует: а вот накося выкуси, хоть бы при всяком исходе, но ни черта ты с меня не получишь [4] – в первом суд разрешает, во втором же наш договор (выигранных дел у меня как не было, так и нет). Вот такая выходит петрушка, говорят комментаторы. Парадокс, дескать, братцы – ничего не попишешь…

Но попробуем вникнуть и взглянуть повнимательней: столь ли уж ахово безнадёжно положенье судьи? Вынося решение отказать Протагору в удовлетворении иска, он в своём полном безоговорочном праве: никаких оснований что-либо требовать договор с учеником тому пока ведь не предоставляет. Хотя, ключевое тут слово *пока*: ведь и правда, как толковали нам прежние комментаторы, сам такой приговор станет делом, выигранным Эватлом, ну и всё как по-писаному. Только это ж ведь не в момент принятия, вынесения – а когда он будет объявлен, да ещё и после вступления в силу. Даже если пренебречь обычной отсрочкой для апелляции, пусть немедленно – всё равно: ситуация, состояние дел уже переменилось. Вот он где и зарыт, милый наш ‘Гейзенберг’: выскакивает, как чёртик из табакерки, и с довольным видом потирает ручками, снимая любые коллизии. Сам судебный вердикт изменил картину мира и обстоятельств (подобно тому самому первому измерению в популярном иллюстративном объяснении ‘неопределённости’, с которого всё начиналось). Ну а раз уж мы попали в новый блаженный мир (к слову, счастливый для Протагора, но вовсе не для угодившего впросак неплательщика), то и рассуждать теперь надо заново, оставив позади прошлые муки сомнения [5].

Ну и где же тут парадокс?! Растаял, смылся, без следа провалился в прогал между разных ‘возможных миров’. А иллюзия мнимого парадокса возникала у иных комментаторов как раз из слияния вместе, наложения и соединения несовместимых суждений и рассуждений, протекающих в различных мирах и при разных обстоятельствах (как раз кстати вспоминаем нашу бедненькую спираль, безжалостно до противоречия сплющиваемую в окружность). И ведь сколько бывает в жизни высказываний, чьё значение чудесно зависит хоть от места, от времени, обстоятельств, погоды – да мало ль чего ещё?[6] Ну положим, скажу я, что нахожусь в этой вот комнате, и ведь будет чистейшая правда, ан и запросто может сделаться, оказаться ложным спустя долю минуты, достаточную, чтобы мне выйти за дверь – мы ж не видим тут ни малейшего парадокса. А и чем же Эватл с Протагором тут отличаются: за что им страдать? Разве что с ними чуток накрутили, наворотили, да ещё нипочём ни про что парадоксом их обозвали, чтобы нас, наивных, запутать и голову заморочить – а мы и рады, готовы попасться, клюнуть на удочку, да наживку и заглотать. Вот и вся недолга... Не так ли?

А теперь припомним и взглянем на ещё один знаменитый парадокс, с куда как более трагичным финалом: ‘парадокс неожиданной казни’ (или иначе: ‘парадокс узника’)[7]. Как известно, в воскресенье под вечер заключённого в камере смертника посещает, положим, лично начальник тюрьмы (по слухам, скажем, прославленный своей честностью и правдивостью) и с лёгкой дрожью в голосе (возможно, слегка играя: как кошка с мышкой?) сообщает ему, что на грядущей неделе тебя непременно казнят, но когда, я тебе не скажу, и никто не ответит, и ещё накануне дня казни ты его знать не будешь. Вот такая печаль: подоспело. Но к несчастью, у узника есть мозги, и порою он их вдруг включает. И тут как раз он задумывается и вдруг замечает: а ведь надумай они меня казнить в воскресенье, ровно неделю спустя, так ведь

---

[4] он ведь как Буратино (в ответ на задачку Мальвины): “Ну не отдам же я этому некту яблоко: хоть он дерись!”

[5] Между прочим, по одному из вариантов изложенья сюжета, именно так оно и случилось. Когда ученик на угрозу судом небрежно отмахнулся, словно от назойливой мухи: мол, не стану платить, хоть судись не судись, всё равно я ведь прав – но тогда искушённый учитель парировал: значит, будем дважды судиться. Что и сделал, исполнив угрозу и доведя предприятие до конца: проиграв первый процесс, предъявил иск повторно, и теперь уже выиграл ‘за изменением обстоятельств дела’ – после чего уклониста уже по суду принудили расплатиться. Будь так, древнегреческий суд выходит и выглядит куда мудрее иных комментаторов, приверженных и подверженных гнёту воображаемых парадоксов: и отлично ведь интуицией, чутьём понимал и принимал ‘принцип возможных миров’ в стиле нашего ‘Гейзенберга’ – задолго до появленья последнего.

[6] а порою и вовсе: просто от способа понимания (то ли от индивидуально-личного, или от общераспространённого) терминов, слов, в нём употреблённых: ну, скажем, холодно сегодня или тепло? – это уж кому как…

[7] А хотя и тот был для Эватла малость печален, но уж никак не сравнить! Правда, ведь и этот, новый, иногда представляют ‘парадоксом неожиданного экзамена’ (или, ещё мягче: ‘внезапной контрольной’) – зато менее кровожадно, но логическая суть не меняется (хоть положим, иной студент и заявит сдуру, в запале, что экзамен похуже, страшнее смерти – но уж это, конечно же, безмерная глупость, ну или на худой конец, эдакая ‘фигура речи’).

выйдет, начальничек наш соврамши – я ж в субботу под вечер, не дурак, догадаюсь, что деваться мне дальше некуда. Но в голове и дальше шарики крутятся, и его осеняет: раз я уже знаю, что до воскресенья мне ну никак не дожить, так выходит, и в субботу нельзя: я же в пятницу это как-нибудь, да небось, осознáю. Разумеется, день за днём (чем дальше и дольше думать, тем получается, всё ближе развязка) добирается узник в мыслях до понедельника – и вдруг разом впадает в полный непревзойдённый восторг: а ведь не сумеет начальник меня казнить, лжецом не прослыв (и выходит, сидеть мне тут до естественной смерти: хоть не радость, конечно, ан ведь жить-то охота…). Иногда представляющие парадокс тем и кончают: дескать, тоже мне из начальника безупречный правдивец: изволил наобещать, а исполнить не в состоянии. Только ж это ещё отнюдь не финал, говорят иные (малость более сообразительные), итог здесь куда печальнее. Мол, блаженствует узник наш пару дней в эйфории, поплёвывая себе в потолок с опрометчиво беззаботным ощущением безопасности – а, положим, в среду после обеда входит палач и с порога: а ну ка, пожалуй, милок, к ответу. И казнят его всенепременно, и ведь прав оказался начальник: накануне ж тому ну никак невдомёк, насколько он в мыслях своих прокололся и в какой просак угодил. Вот такой опять же парадокс получается…

Только это всё-таки всего лишь одно из возможных изложение и трактовка – а попробуем глянуть хоть маленько поглубже (или на шаг подальше?).

Очень многие комментаторы понимают, что понятие ‘знания’ (на котором тут вся игра) запредельно неоднозначно, нечётко и неточно, и зачастую воспринимается очень по-разному (сравн. сноску 6). И правда же: ну как оно описать и определить, что́ кто-то знает или не знает (или только полагает, что знает[8])? Зачастую формально-логические определения, описания понятия ‘знания’ включают, предполагают, предусматривают извлечение (или способность к извлечению?) всех возможных логических следствий из массива собственных знаний субъекта. Ну разумеется, подобное допущение категорически, абсолютно нереалистично и далеко уходит за пределы правдоподобия. Никто из живущих ни на что подобное заведомо не способен: не говоря уж о полной невозможности объять необъятное или об ограничениях времени, но и просто ни один нормальный вменяемый человек и не станет даже пытаться выводить заключения напропалую. Предпочитается как правило то (из доступного), что хоть чем-нибудь интересно: то ли приятно, или же что неприятно, скажем, опасно (хотя с этим сложнее: о дурном порой и не всегда сто́ит задумываться, а от опасности по-страусиному укрыться в песок – хотя и это зачастую чревато). Но уж точно вряд ли кто станет извлекать следствия безразличные и нерелевантные (разве что ненароком, случайно, мимоходом или по недосмотру). Однако в любом случае ясно, что анализ подобных нюансов и обстоятельств не имеет ни малого отношения к логике и ей неподвластен (уж это скорее из области психологии или мало ль чего ещё?). Поэтому оставим это всё в стороне, а вернёмся к нашему злосчастному злополучному узнику.

Естественно, состояние его знаний зависит не столько от реального состояния и обстояния всяческих дел в мире, прошлых или будущих (на которые он всё равно никак теперь не влияет, а только от них зависит – ведь он не всезнайка и не пророк, чтобы предсказать день собственной казни), сколько от его убеждений или предположений: от того, во что или чему он готов поверить или довериться. Первоначально он, похоже, готов поверить словам начальника, полагая того правдивым – но потом начинает сам рассуждать. А вот тут-то и скрыт подвох. Когда, достигнув до понедельника (правда, не доживя, а до него, так сказать, ‘домыслив’), он радостно останавливается и с успокоенным облегчением отключает мозговые извилины – здесь он и совершает критическую ошибку. Конечно, человеку свойственно расслабляться от приятной радостной новости и ею восторженно упиваться, забывая о печальной реальности – однако подобное могло бы скорее быть естественным для беззаботного студента из ‘внезапной контрольной’ (сравн. сноску 7), готового утешиться чем ни попадя, лишь бы уклониться от тоскливой необходимости подготовки к неизбежному экзамену. Ну а уж наш-то персонаж, коль ходит под жизнью, висящей на волоске, мог бы быть посерьёзнее и поосмотрительнее[9]. Ну а если бы он не перестал шевелить своими извилинами, то возможно, и докумекал бы, и заметил, что ведь это всё у него получилось и вывелось потому только и из того, что ещё в начале сам же принял за истину слово начальника о назначенной казни на предстоящей неделе – а теперь, как ни в чём не бывало, пришёл к заключению прямо противоположному, и на том успокоился, нимало почему-то не озаботившись между собой сопоставить. Ну и как же это возможно в здравом уме такому вместе поверить?! И выходит, узнику теперь придётся пересмотреть и отказаться от своего доверия к высказанному заявленью начальника. Хотя вроде бы правдивость его, это сказавшего, и предполагается общеизвестной, но и что же с того? (ну положим, если вдруг кто-то, кому все мы до сих пор доверяли, вдруг заявит, что дважды два пять, так и в этом что ли неужели с ним соглашаться? – или даже если не так уж он грубо в лоб, а завернёт что-нибудь эдакое чуть похитрее, но из чего мы скоренько всё равно именно это самое $2 \times 2 = 5$ вдруг и получим? что тогда?).

[8] К слову, в этой связи припоминается одно суждение Асан Дабсовича Тайманова, произнесённое им однажды по ходу доклада на семинаре кафедры логики в МГУ (и показавшееся, на вкус автора, удачным – оттого так и запавшее в память): Что значит, когда я говорю, что я нечто понимаю? Это попросту означает, я как-то так сумел себя убедить, что вроде как я это понимаю. Ну а если мои все предпринятые попытки в том себя убедить ни к какому успешному результату не привели и окончились крахом, неудачей, ничем – вот тогда я и говорю: нет, этого я не понимаю (или, возможно: понять не могу).

[9] Правда, это всё равно ему бы нимало не помогло, ну а так он хотя бы нашёл себе в том какое-то облегчение – подобно неизлечимо больному, что тешит себя надеждой, будто все у него путём, а диагноз безделка, и авось оно как-нибудь само рассосётся…

А теперь, коль поверить не удалось, то и узник наш переходит в совсем иное состояние своих рассуждений (быть может, это чуть адекватней, точнее, чем: состояние знаний?): из прежнего ‘мира’ непомерно неограниченного доверия, оказавшегося (или показавшегося?) чрезмерным (так сказать, из мира ‘пере-доверия’, ‘пере-информированности’, на поверку представившегося несбыточным и неосуществимым) в состояние категорично полного недоверия и неверия, уже не включающее обсуждаемого обещанья начальника. Ну а раз оно, это обещание, теперь вовсе не принято, то и следствий из него никаких не получится. Вот он, очередной наш ‘Гейзенберг’, возвращающий огорчённого, расстроенного, дезинформированного узника (после некоторых его рассуждений и умозаключений, явившихся фактически бесполезными, потому как не приведших ни к чему продуктивному) в исходное состояние неопределённости, неизвестности, безнадёжности: изгнанного из привидевшегося, примерещившегося ему обманчиво блаженного рая, которому для него не суждено состояться и сбыться. Облом… И когда он будет в какой-то из неопределённо предсказанных дней неизбежно казнён, то слова начальника в полной мере и в полном объёме окажутся правдой – только для узника это ровным счётом ничего не меняет: он же всё равно из-за некоторой малость замаскированной казуистичности обещания не оказался способен ему ни поверить, ни его опровергнуть (да и какая ему, несчастному, до того, собственно, разница?)[10] …

Ну а парадокс-то, похоже, исчез? – просто ни ему до узника, ни узнику до него никакого нет дела… А возникало у нас ощущение кажущегося парадокса, опять же, как раз из смешения разных, но ошибочно неразличаемых ‘миров’: реального положения вещей (наступающего после своевременной казни) и ‘виртуального’[11] умственного состояния узника, да ещё поначалу не доведённого до логической точки, а оборванного на полуслове – точнее, на полу-мысли (хотя впрочем, если и довести рассуждение до упора – как мы видим, оно мало бы что дало и мало чему и кому способно помочь).

В заключение, к слову, отметим пару попутных обстоятельств и замечаний – впрочем, достаточно праздных. Одно: отклоняя высказывание начальника, приведшее его к противоречию, субъект не обязан исключать его полностью (подобно тому, как, скажем, несогласие с заявлением $2 \times 2 = 5$ от заслужившего доверие собеседника вовсе не требует непременно забыть и отвергнуть разом всё, что он до или после того говорил: ну кому из нас не могло бы случиться от заскока или минутного затменья, чтобы вдруг сгоряча или впопыхах ляпнуть какую-то чушь?). То есть, учитывая прежнее доверие, можно бы попытаться не всё отвергать, а сохранить и учесть максимально допустимое из сообщённой им информации, что ещё не ведёт нас к противоречию. Трудность в том, что подобный максимальный объём содержания навряд окажется однозначным – и неясно, кто лучше и кого из них выбрать. Например, в нашем случае: есть хотя бы один вариант с обещаньем за неделю казнить, исключив гарантию, что сам узник о дне не узнает, а другая возможность: наоборот, её и оставить – мол, всё равно заранее не предскажешь, но уж тогда без ограничения срока (по формуле: не скажу, когда, но в какой-то из дней это непременно случится). Впрочем, это всё предмет особой науки, теории, именуемой ‘пересмотр убеждений’ (belief revision), со своими специальными подходами, методами и результатами (удачными или не слишком?), вдаваться в которую (и в которые) мы здесь уж точно не намерены и не станем.

И другой праздный вопрос: а чего ради, собственно, начальник счёл уведомить узника, объявив ему своё странное обещание (или может, предупреждение?): неужели из чистого правдолюбия? Ну и что же он, умник, хотел (или имел в виду) тому сообщить, или чего пытался достичь? Всё равно ж его слова ни в реальности ничего не затронули и не повлияли, да и собеседнику ни знаний, ни проку дать ничуть не смогли, лишь запутали голову. А быть может, к тому и стремился? (то ли надеясь подарить блеск надежды? или напротив, коварно сбить с толку и смутить, испортив тому остающиеся считаные дни? и способен ли он был это спрогнозировать, предвидеть, предугадать?). Или, дескать, охота проявить и продемонстрировать, насладиться, поупиваться всемогуществом вседержителя? Или право же, и вовсе ничего такого отнюдь не предполагал и ничего не старался добиться: ну просто так, сказал и сказал – язык же ведь без костей, и никакого ему за это не будет. А впрочем, это всё исключительно чистые домыслы. И какая в том разница, и стоит ли вовсе доискиваться, что́ могло взбрендить этому ‘правдорубу’, коли дадено право играть и распоряжаться предоставленными ему во власть чужими судьбами или жизнями?..

А теперь: не взглянуть ли, что сходного скрыто между этих двух наших историй? А вот в каждой есть персонаж. Если в первой – малость легкомысленный шалопай, попытавшийся (или может быть, понадеявшийся?) задарма невзначай с кондачка объегорить учителя, но в итоге сам попавшийся тому на крючок, заготовленно притаившийся за искусно раскинутой сетью. Во второй – всё куда суровей, беспощадней и безнадёжней: так жестоко и круто, и непомерно бо́льшее на кону. Но и там персонаж вдруг взял да и очутился опрометчивой жертвой своего легковерия к собственным рассуждениям (вымыслам? или домыслам?), понадеявшись на хромой кобыле оппонента объехать, уличив того в якобы лжи, а себе приписав мнимую неуязвимость – но опять же сам угодил на удочку коварной ловушки, где куда ни

[10] Ну или это как если бы людям, привыкшим и знающим, что земля плоская и на трёх китах, сказануть вдруг что-нибудь этакое про круженье всяких круглых планет вокруг солнца – они ж просто неспособны будут в это поверить, как в абсолютно не совместимое со всем их жизненным опытом, во всех неисчислимых их поколениях: для них это будет уж явная ложь, или даже вовсе бред сивой кобылы. Между тем как в нашей с вами картине, системе восприятия мира оно в точности так и есть. Но только: а в чём же тут парадокс?..

[11] И не напомнит ли это нынешний в век Интернета ставший уже расхожим до заезженности мотив разрыва между виртуальностью и реальностью, и о несчастных, которые, от неспособности здраво их разделить, в него проваливаются, попадают и пропадают в этот безнадёжный прогал?.. Но это ж отнюдь и вовсе не парадокс, а обиход нашей жизни… Разве не так?

кинь, а всё клином, и всенепременно против тебя, ну а тот, кого ты так мечтал обойти, ускользнув из-под неизбежности: как ни глянь, вопреки всем твоим доводам, ан куда ни деваться, всё равно представится правым. Только если первый наш недотёпистый неудачник, под непреложным гласом судьи, платит всего-навсего лишь деньгами, то второй ведь – всей бесценною жизнью. Да и зависит ли уже она от него (и от всех и всяких его соображений, размышлений или же измышлений?), коль судебный палач (или это сам вестник-исполнитель неотвратимо непреклонной судьбы?) как раз успел навострить свой неумолимый топор?..

И не выше ли этот трагический драматизм сюжета, чем любая логика якобы парадокса и подобные изощрённые изыски, лишь способные заморачивать головы и участникам ситуации, да и тех, кто пытается о ней размышлять?..

На сём и позвольте закончить[12].

* * *

В заключение автор хотел бы искренне выразить благодарность за полезные обсуждения, соображения и предложения Д. Виноградову, Р. Гиляревскому, Е. Забинковой, Е. Золину, Д. Савельеву.

Материал поступил в редакцию 21.02.17.

Сведения об авторе

СКВОРЦОВ Дмитрий Павлович – кандидат физико-математических наук, старший научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН, Москва
e-mail: skvortsovd@yandex.ru.

[12] Навряд ли хотя бы одно из высказанных, рассмотренных здесь соображений по парадоксам можно было бы назвать действительно новым: наверняка любое из них в той или иной форме и вариации кем-нибудь уже предлагалось, выдвигалось и обсуждалось. Тем не менее, автор позволяет себе оставить надежду, что всё представленное в целом способно хоть для кого-нибудь что-нибудь да и прояснить. На том и спасибо...

УДК 004.94 : 004.82

С.Л. Гольдштейн, С.С. Печеркин, А.Н. Аверьянова

О моделях экстрактора знаний для разрешения проблемных ситуаций со сложным объектом

Представлены системно-структурные, кортежные и алгоритмическая модели экстрактора знаний, необходимые для разрешения проблемных ситуаций со сложным объектом. Модели дают графическое описание морфогенеза и функционирования экстрактора, развитие которого предлагается, в основном, за счет введения структур системной интеграции.

Ключевые слова: *экстрактор знаний, системная интеграция, моделирование, сложный объект, проблемная ситуация*

ВВЕДЕНИЕ

Современный технологический уклад предполагает приоритет экономики знаний [1]. При этом знания выступают как невидимые активы фирмы, способные на порядок повышать или понижать ее рыночную оценку по сравнению с бухгалтерской [2].

В теории и технологии систем, основанных на знаниях [3], при создании баз знаний используют методы извлечения (*elicitation*) или экстракции (*extraction*) знаний из документов и от экспертов (как из информационной «тары» [4, 5]) коммуникативными (пассивными или активными) или текстологическими методами, соответственно [6–8]. При этом имеют место три процедуры: предварительной (суб) и основной экстракции знаний, а также их интеграции и структуризации. Цель первой – низкоресурсный (прежде всего, по времени, т.е. быстрый) отбор «перспективных» по ценности и релевантности источников информации, второй – полноресурсное извлечение смысла из отобранных источников, третьей – интеграция и структуризация приобретаемых знаний, например, до авторских, а лучше инвариантных, тезаурусных онтологий, как основы для ответа пользователю, разрешающему проблемную ситуацию. Следует особо отметить, что экстракции знаний посвящены многочисленные публикации отечественных и зарубежных исследований на уровне монографий, статей, патентов. Эта тематика напрямую связана, в частности, с идеологией развития Интернета (проекты типа Web 2.0).

Процедура 1 представлена в [9], для процедуры 2 используют разные методы извлечения знаний: нейронные сети, ближайшего соседа, дискриминантного и кластерного анализов, линейного программирования, генетических алгоритмов [3, 5] и др., для процедуры 3 применяют семантические веб-технологии [10–13]. Так решается проблема создания предметных баз знаний, используемых для подсказок по ситуациям малой и средней сложности.

Однако, система, основанная на знаниях, призвана обеспечить не только предметную, но и системную (мета) подсказку, наиболее полезную для разрешения проблемных ситуаций со сложным объектом, специфична, так как должна предоставлять пользователю знания не только о сложном объекте, но и о разноаспектной системно-интегрированной [14, 15] деятельности по разрешению связанной с этим объектом проблемной ситуации. Модели таких систем предлагаются, но проработаны все еще недостаточно [16–26]. В настоящей статье поставлена и решена задача моделирования одной из структур такой системы, основанной на знаниях – экстрактора знаний, необходимых для разрешения проблемных ситуаций со сложным объектом.

ПРЕДПОСЫЛКИ РАЗРАБОТКИ ЭКСТРАКТОРА ЗНАНИЙ

Для уточнения авторской позиции процесса принятия решений приведены две предпосылки об информации и механизме приобретения знаний.

В качестве предпосылки 1 из общеизвестной подборки (коллекции) определений информации выделим и будем ориентироваться на следующее: «Информация – это запомненный выбор одного варианта из несколько возможных и равноправных» [27–30]. В качестве предпосылки 2 взята схема приобретения знаний, приведенная на рис. 1.

При этом инфосырье «А» характеризуется неустойчивым состоянием (хаосом), и представляет собой (в соответствии с [29]) «перемешивающий слой», не-

обходимый для организации выбора информации и последующего ее запоминания. А смыслы «С» можно представить с помощью основных известных моделей: семантики предпочтения, концептуальной зависимости, «смысл-текст», универсального сетевого языка UNL [31–35] и др.

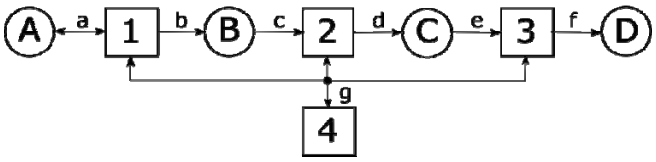


Рис.1. Технологическая схема приобретения (*acquisition*) знаний

Обозначения устройств: 1 – отбора (разведки) релевантных источников информации, 2 – экстракции смыслов из релевантных источников информации, 3 – структуризации знаний и их последующей трансформации для предоставления пользователю, 4 – управления; инфосырья и инфопродуктов: A – потенциальные источники знаний (тексты от экспертов и/или из документов – бумажных и электронных), B – релевантные источники знаний, C – смыслы релевантных источников, D – интегрированные, структурированные и позиционированные знания; каналы связи: a– запросно-ответный, b – f – логистические, g – управленческие.

Системно-структурную и алгоритмическую модели экстрактора знаний представим (рис. 2) исходя из компилятивного научного прототипа, выбранного по методике [36]. Основными источниками этого прототипа являются модели, представленные различными авторами в аспекте приобретения знаний и интеллектуальном анализе данных [3, 5, 7, 8, 12, 37, 38].

Недостаток прототипа, выявленный в соответствии с требованиями системноструктурной полноты, – отсутствие средств интеллектуальной и системной поддержек. Парированием этого недостатка может быть введение в структуру экстрактора механизма 10 системной интеграции [39] и адаптация к этой процедуре механизмов 1÷8.

Функционирование развитого и ориентированного на компьютерную реализацию экстрактора отражает рис. 3.

Видно, что блоки 7, 9 и 11 выполняют по сути роль супервизора – помогающего лицу, принимающему решения, – ЛПР (блок 12) и реализуются в параллель (блок 2) с линейкой информационно-технологических блоков 4, 5, 6, 8, 10, 13.

Построение тезаурусной онтологии (блок 5), достаточно хорошо описанное в литературе, можно представить в виде кортежной модели [40]:

$$\mathrm{ThOn} = \langle \mathrm{Cn}, \mathrm{Rn}, \mathrm{Ct}; \mathrm{R1} \rangle, \qquad (1)$$

где: Cn – понятия, Rn – отношения между понятиями, Ct – контент, релевантный Cn и Rn, а когнитивную карту (CM) – в рамках блока 6 – как в [41, 42]:

$$\mathrm{CM} = \langle \mathrm{ThOn}, \mathrm{Lim}; \mathrm{R2} \rangle, \qquad (2)$$

где Lim – контурные ограничения ThOn; R1 и R2 – матрицы связи – смежности или инцидентности (здесь и далее представленные в виде R_i).

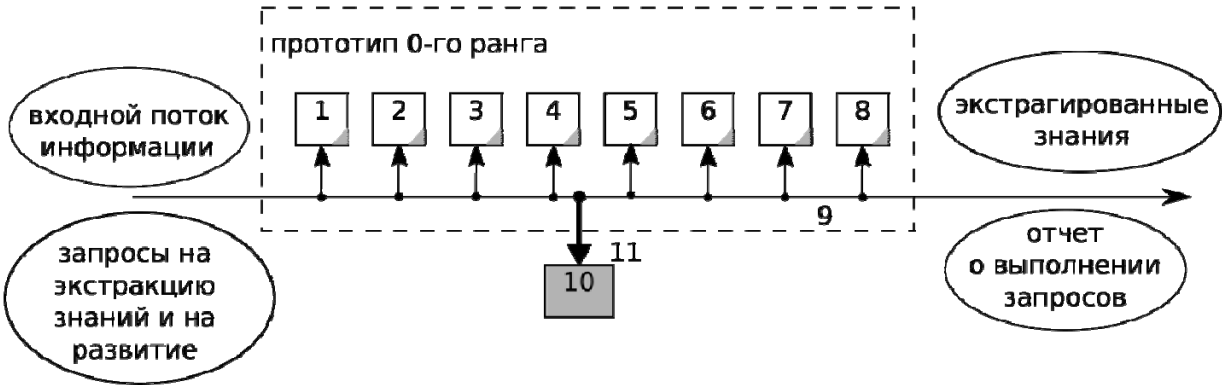


Рис.2. Системно-структурная модель экстрактора знаний о сложном объекте по компилятивному прототипу и предлагаемому решению, обозначенному штриховкой здесь и далее

Обозначения механизмов: 1 – отбора ценной информации, 2 – построения тезаурусной онтологии понятий о сложном объекте и о системно-интегрированной деятельности по разрешению проблемной ситуации с этим объектом, 3 – вычленения из тезаурусной онтологии когнитивных карт, релевантных проблемной ситуации и запросу, 4 – подбора и анализа моделей объекта и моделей разрешения проблемной ситуации, 5 – визуализации объекта и ситуации, 6 – формулирования экстрагированных научных знаний в пертинентной форме, 7 – оценки итогов экстрагирования, 8 – управления экстрагированием знаний; 9, 11 – интерфейсов, 10 – системной интеграции на уровне экстрактора.

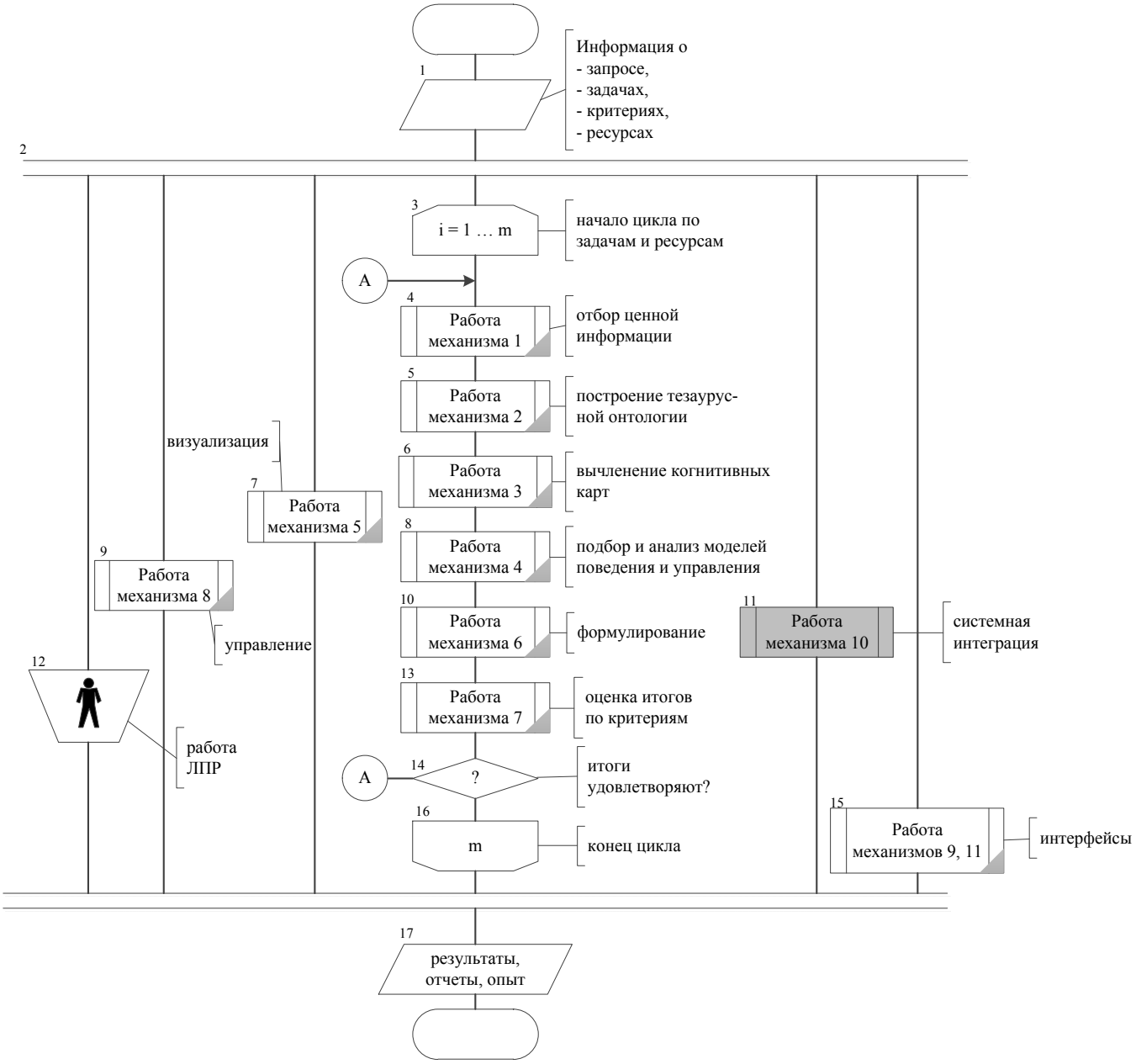


Рис.3. Алгоритм функционирования экстрактора знаний по прототипу и предлагаемому решению (ЛПР – лицо, принимающее решения)

Для обобщенного количественного описания деятельности, представленной на рис. 3, целесообразно исходить из естественно-научного представления о действии как функции координат объекта и скорости их изменения [43]. Применительно к функционированию экстрактора знаний уместны следующие уточнения. База знаний может быть позиционирована в системе координат онтологического пространства [57] и перемещаема в нем (в зависимости от наполнения) из исходного состояния в желаемое с требуемой скоростью под воздействием управленческого ресурса, который, как обычно, целесообразно минимизировать.

СИСТЕМНО-СТРУКТУРНЫЕ МОДЕЛИ МЕХАНИЗМОВ ЭКСТРАКТОРА ЗНАНИЙ

Каждый из механизмов на рис.2 затем представили с учетом прототипных описаний и предлагаемых решений, направленных на улучшение качества функционирования, в кортежном формализме. При декомпозиции всех структур удовлетворено правило Ингве–Миллера (7±2).

Модель механизма 1:

$$M1 = \langle C1.1 \div C1.6;\ C1.7,\ C1.8;\ R3 \rangle, \qquad (3)$$

где системы по компилятивному прототипу [6–13, 16, 29]: C1.1 – позиционирования ЛПР в информационно-когнитивном пространстве, C1.2 – задания критериев и определения ценности информации, C1.3 – предоставления информации ЛПР, C1.4 – отбора ценной информации, C1.5 – оценки качества процесса отбора и его результатов; система, предлагаемая в качестве развития M1: C1.7 – системной интеграции на уровне механизма 1; системы C1.6, C1.8 – интерфейсы.

Система С1.1 может быть развита, исходя из модели:

$$PS = \langle EC1, EC2; R4 \rangle, \quad (4)$$

где PS – позиция ЛПР; ЕС1 и ЕС2 – оценки компетентности лица, принимающего решение в тезаурусах о сложном объекте и о деятельности по разрешению проблемных ситуаций с этим объектом, соответственно.

Относительно системы С1.2 следует отметить, что ценность информации (VI), отбираемой для экстракции знаний, может быть определена в соответствии с моделью:

$$VI = \langle VI1, VI2, VI3; R5 \rangle, \quad (5)$$

где имеются частные составляющие ценности:
VI1 – по целеполаганию для практических действий, VI2 – по предварительной осведомленности ЛПР, например, по тезаурусу, VI3 – по своевременности.

В системе С1.3 предоставление информации может быть обеспечено «поисковиками» Интернета, рубрикаторами бумажных библиотек, рейтингами экспертов. А для системы С1.4 важен учет всех трех известных задач (выбора, оптимального выбора и общей задачи принятия решений) в зависимости от степени формализации множества альтернатив и принципов выбора.

Интегральная оценка качества в системе С1.5 должна включить частные оценки как по процессу (например, своевременность, технологичность, затратность), так и по результату.

Таким образом прототипные решения по системам С1.1 – С1.5следует развить с учетом введения в структуру Ml дополнительной системы 1.7. На входе механизма Ml – список источников информации, а также запросы: на отбор ценной информации и на развитие, а на выходе – ценная информация и отчеты о работе.

Представление о механизме 2 дает кортеж:

$$M2 = \langle C2.1 \div C2.10; C2.11, C2.12; R6 \rangle, \quad (6)$$

где системы по компилятивному прототипу [3, 40, 44, 45]: С2,1– формирования списка / ключевых слов из запроса, С2.2 – привязки определений /дефиниций из авторитетных словарно-энциклопедических источников к ключевым словам запроса, С2.3 – построения иерархической онтологии, С2.4 – выявления «боковых» / функциональных связей и построения сетевой тезаурусной онтологии, С2.5 – наполнения вершин и связей тезаурусной онтологии контентом, С2.6 – устранения противоречий между знаниями из бумажноэлектронных сведений и от экспертов, С2.7 – позиционирования наполненной тезаурусной онтологии в инфопространстве, С2.8 – управления механизмом 2, С2.9 – оценки результатов; система, предлагаемая в качестве развития М2: С 2.11 – системной интеграции на уровне механизма 2; системы С2,10 и С2,12– интерфейсы.

Системы С2.1 и С2.2 в комментариях не нуждаются. Для реализации же системы С2.3 необходим учет типологии знаний (ТК):

$$TK = \langle TK1 \div TK5; R7 \rangle, \quad (7)$$

где структуризация знаний: ТК1 – по типу (процедурное, фактологическое, декларативное), ТК2 – по содержанию (бытовое, предметное, мета), ТКЗ – по форме (вербальная, графическая, аналитическая), ТК4 – по латентности (эксплицитное, имплицитное), ТК5 – по генезису (от фактов – безусловные, от теорий – условные).

При этом будем полагать, что справедлив кортеж:

$$T = \langle M, L, S; R8 \rangle, \quad (8)$$

где Т – устный и/или письменный текст, М – смысл, как содержание (семантика) текста и как субъективный образ, возникающий при понимании текста в нотациях герменевтики, L – языковое оформление смысла, S – стилевое оформление текста.

Кроме того, для системы С2.3, ориентированной на создание иерархической онтологии, важно правильное начало декомпозиции:

$$C2.3 = \langle SS2.3.1, SS2.3.2, SS2.3.3; R9 \rangle, \quad (9)$$

где подсистемы – это вершины 1-го уровня тезаурусной онтологии; т.е. термины: SS2.3.1 – сложного объекта, SS2.3.2 – системно-интегрированной деятельности, SS2.3.3 – оценок качества деятельности.

Для этого требуются, прежде всего, представления о генезисе знаний:

$$K = \langle Inf, Y; R10 \rangle, \quad (10)$$

где К – знания, Inf – информация, Y – интегральный критерий качества информации,

$$Y = \prod_{i=1}^{6} y_i^{a_i}, \quad (11)$$

где y_i – частные критерии: y_1 – селективности, y_2 – упорядоченности, y_3 – способа получения, y_4 – оформленности, y_5 – социальной значимости, y_6 – признаваемости, α_i – вес, причем $\sum_{i=1}^{6} a_i$, с учетом правила:

$$\text{If } Y > Y^{TP} \text{ then Inf (y)} \in K \text{ else Inf (y)} \notin K, \quad (12)$$

где Y^{TP} – требуемое значение качества информации.

О системах С2.4÷С2.6 в литературе имеется достаточно информации. А для создания системы С2.7 необходимы представления о позиции знаний:

$$P = \langle P1, P2, P3; R11 \rangle, \quad (13)$$

где позиции: Р1– фрагмента тезаурусной онтологии, релевантного задаче, в онтологическом пространстве системы знаний, как хранилища, Р2 – знаний о задаче

по отношению к фокусу внимания и мотивации ЛПР [16], Р3 – когнитивного потенциала ЛПР, как активного ресурса для решения задачи [46].

Недостаток (системно-структурная неполнота) 2-го прототипа 1-го ранга предлагается парировать введением системы С2.11 и адаптацией систем С2.1÷С2.9 к специфике разрешаемой проблемной ситуации. На входе М2 – поток ценной информации и запросы: на тезаурусную онтологию и на развитие, на выходе – согласованная тезаурусная онтология и отчеты о работе.

Механизм 3 представлен кортежем:

$$M3=\langle C3.1 \div C3.7;\ C3.8,\ C3.9;\ R12\rangle, \qquad (14)$$

где системы по компилятивному прототипу [41, 42]: С3.1 – вычленения структур, релевантных запросу, С3.2 – учета влияний, представленных связями, С3.3– анализа устойчивости когнитивной карты, С3.4 – анализа сценария развития ситуации, С3.5 – управления механизмом 3, С3.6 – оценки итогов, система, предлагаемая в качестве развития М3: С3.8 – системной интеграции на уровне механизма 3; С3.7, С3.9 – интерфейсы).

Когнитивная карта – это семейство моделей, представляющих структуру причинно-следственных влияний слабоструктурированной ситуации. Обязательной частью любой модели семейства является ориентированный граф, в котором вершины соответствуют факторам ситуации, а дуги – прямым причинно-следственным влияниям факторов друг на друга. Когнитивный подход [46] поддерживает разрешение проблемных ситуаций методами, учитывающими когнитивные аспекты познавательных процессов человека.

Недостаток этого 3-го прототипа 1-го ранга предлагается компенсировать добавлением системы С3.8 и адаптацией систем С3.1÷С3.6 к проблематике. На входе М3 – тезаурусная онтология и запросы: на когнитивные карты и на развитие, на выходе – когнитивные карты и отчеты о работе.

Механизм 4 представлен кортежем:

$$M4 = \langle C4.1 \div C4.8;\ C4.9,\ C4.10;\ R13\rangle, \qquad (15)$$

где системы по компилятивному прототипу [42, 47, 48]: С4.1 – построения алгебраических уравнений взаимодействия элементов карты, С4.2 – построения модели поведения в виде дифференциального уравнения с суммой частных вкладов о темпе нарастания качества, С4.3 – построения модели управления в виде дифференциального уравнения с обобщенными вкладами, С4.4 – анализа устойчивости, С4.5 – анализа динамики, С4.6 – управления механизмом 4, С4.7 – оценки результатов; система, предлагаемая в качестве развития М4: С4.9 – системной интеграции на уровне механизма 4; системы С4.8, С4.10 – интерфейсы.

Недостаток этого 4-го прототипа 1-го ранга может быть парирован введением системы С4.9 и доводкой систем С4.1÷ С4.7. На входе механизма М4 – когнитивные карты и запросы: на модели и на развитие, на выходе – модели и отчет о работе.

Механизм 5 также обладает системно-структурной неполнотой. Он представлен кортежем:

$$M5 = \langle C5.1 \div C5.8;\ C5.9,\ C5.10;\ R14\rangle, \qquad (16)$$

где системы визуализации по прототипу [49]: С5.1 – инфосырья, С5.2 – инфополупродуктов, С5.3 – инфопродуктов, С5.4 – ситуаций и процессов, С5.5 – ресурсов для разрешения ситуаций, С5.6 – управления механизмом 5, С5.7 – оценок результатов; система, предлагаемая в качестве развития М5: С5.9 – системной интеграции на уровне механизма 5; системы С5.8, С5.10 – интерфейсы.

Решение, предлагаемое в работах [50–53], структурно связано с введением системы С5.9 и адаптацией систем С5.1÷С5.7, а по сути – с 3D-визуализацией сложного объекта и пятиплоскостной визуализацией ситуации с объектом, что требует супервычислительных мощностей. На входе механизма М5 – данные для визуализации и запросы: на визуализацию и развитие, на выходе – визуализация и отчеты о работе.

Механизм 6 отражает модель:

$$M6 = \langle C6.1 \div C6.7;\ C6.8,\ C6.9;\ R15\rangle, \qquad (17)$$

где системы по компилятивному прототипу [18, 38]: С6.1 – поддержки сепарирования новых предметных и системных знаний, Сб.2 – поддержки формулирования нового предметного знания, С6.3 – поддержки формулирования нового системного знания, С6.4 – оценки качества решения задачи, Сб.5 – поддержки воплощения формализованных знаний в еще неформализованные, С6.6 – управления механизмом 6, С6.7 – оценки результатов; система, предлагаемая в качестве развития Мб: С 6.8– системной интеграции на уровне механизма 6; системы Сб.7, С6.9 – интерфейсы.

Система 6.5 необходима для предоставления пользователю знаний в привычной, понятной ему форме. Наши предложения связаны с созданием инструментальной среды системотехнического облуживания сложных объектов [54, 55] для развития систем 6.1÷6.6 и введением, как и в механизмах 1÷5, системы 6.8. На входе механизма 6 – когнитивный полуфабрикат и запросы: на формулировку знаний и на развитие, на выходе – сформулированные знания и отчеты о работе.

Механизм 7 представлен кортежем:

$$M7 = \langle C7.1 \div C7.5;\ C7.6,\ C7.7;\ R16\rangle, \qquad (18)$$

где системы по прототипу [56]: С7.1 – критериев, С7.2 – оценки по дифференциальным критериям, С7.3 – оценки по интегральным критериям, С7.4 – управления механизмом 7; система, направленная на развитие М7: С7.6 – оценки уровня системной интеграции, системы С7.5 и С7.7 – интерфейсы.

На входе механизма М7 – информация о работе экстрактора знаний и запросы: на оценку итогов и на развитие, на выходе – оценки и отчеты о работе.

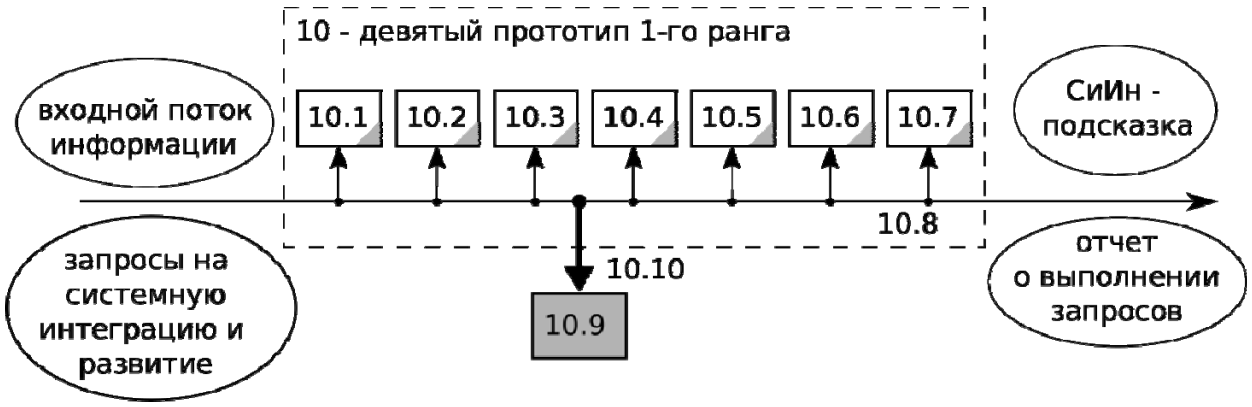


Рис. 4. Системно-структурная модель механизма системной интеграции по нашему же прототипу [39] и новому предлагаемому решению

Обозначения системы: 10.1 – интегрированного корпоративного бизнеса заказчика запроса, 10.2 – системно-интегрированной логистики заказчика, 10.3 – интегрированных информационных технологий заказчика, 10.4 – интегрированной полимедиавизуализации информации, 10.5 – управления механизмом системной интеграции, 10.6 – системно-научной поддержки, 10.7 – человеко-машинной интеллектуальной поддержки, 10.9 – позиционирования знаний о деятельности и объекте в онтологическом пространстве; 10.8, 10.10 – интерфейсов.

Механизм 8 отражает классический взгляд на управление:

$$M8 = \langle C8.1 \div C8.6;\ C8.7,\ C8.8;\ R17 \rangle, \qquad (19)$$

где системы по прототипу [48]: С8.1 – фиксации реального и желаемого состояний, С8.2 – задания критериев качества управления, С8.3 – ресурсов управления, С8.4 – мониторинга результатов управления, С8.5 – парирования помех; система, направленная на развитие М8: С8.7 – системной интеграции на уровне механизма 8; системы С8.6 и С8.8 – интерфейсы.

На входе механизма М8 – информация для выработки управления и запросы: на управление и на развитие, на выходе – управленческие решения и отчеты о работе. Механизм 10 системной интеграции (рис. 4) имеет структуру, инвариантную и для систем: 1.7, 2.11, 3.8, 4.9, 5.9, 6.8, 7.6 и 8.7.

В основу работы подсистемы 10.9 заложены модели, полученные ранее в работе [16].

ВЫВОДЫ

Поставлена и решена задача моделирования структуры экстрактора знаний для разрешения проблемных ситуаций со сложным объектом, развитие которого предложено за счет введения структур системной интеграции и модификации прототипных механизмов.

Приведены системно-структурная и алгоритмическая модели экстрактора, ориентированного на решение поставленной задачи.

Представлен пакет системно-структурных моделей восьми механизмов экстрактора, призванных обеспечить наполнение системы знаний контентом, достаточным для системно-интеграционной поддержки деятельности лица, принимающего решение.

Уместен вывод о переходе к следующим уровням моделирования – информационному и математическому.

СПИСОК ЛИТЕРАТУРЫ

1. Глазьев С.Ю. Стратегия опережающего развития России в условиях глобального кризиса. – М.: Экономика, 2010. – 255 с.
2. Макаров В.Л. Экономика знаний: уроки для России // Вестн. Рос. акад. наук. – 2003. – Т.73, №5. – С. 450 – 456.
3. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2000. – 384 с.
4. Extracting Clasificationknoledge of Internet Documents with Mining Term Associations: A Semantic Approach. / Shian-Hua Lin [и др.]. – URL: https://scholar.google.ru/citations?view_op=view_citation&hl=vi&user=ohtkptoAAAAJ&citation_for_view=ohtkptoAAAAJ:u-x6o8ySG0sC (дата обращения 01.10.2016)
5. Витяев Е.Е. Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов. – Новосибирск: НГУ, 2006. – 293 с.
6. Гаврилова Т.А. Извлечение знаний: «пассивные» методы. – URL: http://pandia.ru/text/77/214/98428.php. (дата обращения 01.10.2016)
7. Гаврилова Т.А. Извлечение знаний: лингвистический аспект // Корпоративные системы. – 2001. – №10 (25). – С. 24-28.
8. Гаврилова Т.А. Активные индивидуальные методы извлечения знаний из данных. – URL: http://kmsoft.ru/publications/km/big/active_method_isvl_knowlgdata.html. (дата обращения 15.10.2014)
9. Ландэ Д.В. Поиск знаний в Интернет. – М.: Вильямс, 2005. – 272 с.

10. Браславский П., Колычев И. Автоматическое реферирование веб-документов с учетом запроса // Интернет-математика 2005. Автоматическая обработка веб-данных. – М., 2005. – С. 485-501.
11. Ильин Н., Киселев С., Танков С., Рябышкин В. Технологии извлечения знаний из текстов // Открытые системы. СУБД. – 2006. – №6.
12. Дюк В.А. DataMining – интеллектуальный анализ данных // BYTE (Россия). 1999. – № 9. – С. 18–24.
13. О'Лири Д.Е. Управление корпоративными знаниями. – URL: http://hr-portal.ru/article/ upravlenie-korporativnymi-znaniyami (дата обращения 12.07.2016).
14. Печеркин С.С. Теоретическое описание и развития системной интеграции для научно-практических структур: дисс. канд. физ.-мат. наук. – Екатеринбург, 2002. – 204 с.
15. Гольдштейн С.Л. Системная интеграция бизнеса, интеллекта, компьютера. – Екатеринбург: ИД «Пироговъ», 2006. – 392 с.
16. Гольдштейн С.Л., Инюшкина О.Г., Кормышев В.М. Развитие системы управления знаниями для разрешения ситуаций в бизнесе. – Екатеринбург: ИД «Пироговъ», 2006. – 220 с.
17. Гольдштейн С.Л., Кудрявцев А.Г. Разрешение проблемных ситуаций при поддержке систем, основанных на знаниях. – Екатеринбург: ИД «Пироговъ», 2006. – 218 с.
18. Бельков С.А., Гольдштейн С.Л., Ткаченко Т.Я. Гипертекстовый тезаурус системных знаний // Научно-техническая информация. Сер. 2. – 1996. – №3. – С. 1-11.
19. Гольдштейн С.Л. Кудрявцев А.Г. Идея рациональной структуризации знаний (на примере математики) // Научно-техническая информация. Сер. 2. – 1997. – №2. – С. 13-17.
20. Гольдштейн С.Л., Ткаченко Т.Я., Устьянцев Д.А. Об одном способе построения системы знаний // Научно-техническая информация. Сер. 2. – 1997. – №10. – С. 8-26.
21. Браславский П.И., Гольдштейн С.Л., Ткаченко Т.Я. Тезаурус как средство описания систем знаний // Научно-техническая информация. Сер. 2. – 1997. – №11. – С. 16-22.
22. Гольдштейн С.Л., Кудрявцев А.Г., Ткаченко Т.Я. Моделирование систем знаний: системно-информационный и физико-технический аспекты // Научно-техническая информация. Сер. 2. – 1998. – №8. – С. 17-32.
23. Гольдштейн С.Л., Кудрявцев А.Г., Ткаченко Т.Я. Моделирование систем знаний: вычислительный эксперимент // Научно-техническая информация. Сер. 2. – 2000. – №2. – С. 15-21.
24. Гольдштейн С.Л., Кудрявцев А.Г. Моделирование систем знаний: динамический подход // Научно-техническая информация. Сер. 2. – 2000. – №4. – С. 11-16.
25. Гольдштейн С.Л., Кудрявцев А.Г. Системно-информациологические предпосылки для математического моделирования запросно-ответных процессов и механизмов // Научно-техническая информация. Сер. 2. – 2004. – №4. – С. 1-9.
26. Гольдштейн С.Л., Кудрявцев А.Г. Абстрактно-когнитивный тезаурус по диалогу // Научно-техническая информация. Сер. 2. – 2005. – №7. – С. 23-35.
27. Новейший философский словарь. 2-е изд. – М.: Интерпрессервис, 2001. – 1280 с.
28. Карагодин В.И. Информация и феномен информации. – М.: Пущино, АН СССР, 1991. – 201с.
29. Чернавский Д.С. Синергетика и информация (динамическая теория информации). – М.: Едиториал УРСС, 2004. – 288 с.
30. Стратонович Р.Л. Теория информации. – М.: Соврадио, 1975. – 424 с.
31. Лопатина М.Ю. Способы языковой актуализации семантики предпочтения: дисс. канд. филолог, наук. – Барнаул, 2006. – 170 с.
32. Шапкин П.А. Модели и методы аппликативного моделирования концептуальных зависимостей: дисс. канд. технич. наук. – М., 2010. – 178 с.
33. Большаков И.А., Гельбух А.Ф. Модель «Смысл и Текст»: тридцать лет спустя. – URL: http://www.gelbukh.com/CV/publication/2000/Forum-MTM-rus.htm (дата обращения 27.07.2003)
34. Мельчук И.А. Русский язык в модели «Смысл о Текст». – М.: Вена, 1995. – 370 с.
35. Скребцова Т.Г. UNL как способ представления семантики предложения // Структурная и прикладная лингвистика. – 2008. – № 7. – URL: http://ojs.lib.pu.ru/index.php/SPL/issue/view/1/showToc (дата обращения 18.09.2014).
36. Гольдштейн С.Л., Печеркин С.С. Системный метод прототипирования // Вестник РАЕН. – 2010. – Т. 10, № 1.– С. 45-50.
37. Приобретение знаний / под ред. С. Осуги, Ю. Саэки. – М: Мир, 1990. – 804 с.
38. Нонака И., Такеучи Х. Компания – создатель знания. – М.: Олимп-Бизнес, 2003. – 384 с.
39. Гольдштейн С.Л., Печеркин С.С., Гольдштейн М.Л. О механизме системной интеграции // Системы управления и информационные технологии. – 2011. – № 3 (45). – С. 127-131.
40. Нариньяни А.С. Кентавр по имени ТЕОН: Тезаурус + Онтология // Труды международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. Т. 1 Аксаково, 2001. – URL: http://www.artint.ru/articles/narin/teon.htm (дата обращения 15.08.2010)
41. Люгер Д.Ф. Искусственный интеллект: стратегии и методы решения сложных проблем. – М.: Вильямс, 2003. – 864 с.
42. Хасанова Н.В. Оценка и управление научно-образовательным потенциалом на основе структурных, когнитивных и динамических моделей: диссер. канд. техн. наук. – Уфа: УГАТУ, 2006. – 188 с.
43. Самарский А.А.,Михайлов А.П. Математическое моделирование. – М.: Наука, Физмат, 1997. – 320 с.
44. Добров Б.В., Иванов В.В., Лукашевич Н.В. Онтологии и тезаурусы: модели, инструменты,

примечания. – М.: Интернет-Университет, 2009. – 173 с.
45. Гаврилова Т.А. Онтологический инжиниринг: от истории к формированию прикладных онтологий // В сб. «Когнитивные исследования» / под ред. В.Д. Соловьева, Т.В. Черниговской. Вып.2. – М.: Изд-во «Институт психологии РАН», 2008. – С. 293 –308.
46. Величевский Б.М. Когнитивная наука: основы психологии познания, т.2. – М.: Смысл, 2006. – 432 с.
47. Краснощеков П.С., Петров А.А. Принципы построения моделей. М.: МГУ, 1983. 264 с.
48. Красовский Н.Н. Управление динамической системой. – М.: Наука, 1985. – 520 с.
49. Визуализация информации. Каталог 2010-2011, – М.. Полимедиа, 2010. 116 с. – URL: http://rtk22.ru/download/%CA%E0%F2%E0%EB%EE%E3_POLYMEDIA.pdf (дата обращения 01.09.2016)
50. Гольдштейн С.Л., Печеркин С.С., Гольдштейн М.Л. О системах виртуальной дополненной реальности и их применении в медицине // Системная интеграция в здравоохранении, е-журнал. – 2011. – №1. – С. 5-16.
51. Гольдштейн С.Л., Тюлюкин А.В. Инструментальная оболочка для визуализации работ с каркасом тезаурусной системы знаний // Алгоритмы и программные средства параллельных вычислений. Вып.7. – Екатеринбург: ИММ УрО РАН, 2003. – С. 50-70.
52. Гольдштейн С.Л., Тюлюкин А.В. Параллелизм при графической визуализации работ по системам знаний // Алгоритмы и программные средства параллельных вычислений. №8. – Екатеринбург: ИММ УрО РАН, 2004. – С. 109-129.
53. Гольдштейн С.Л., Печеркин С.С., Гольдштейн М.Л. Ядерно-медицинская установка. Патент 2464658 от 10.07.2012.
54. Гольдштейн С.Л., Ткаченко Т.Я. Концептуально-системная модель инструментальной среды системотехнического обслуживания сложных объектов. Деп. ВИНИТИ №3707-В91, 1991. – 30 с.
55. Ткаченко Т.Я. Инструментальная среда системотехнического обслуживания сложных объектов. – Екатеринбург: УГТУ-УПИ, 2002. – 203 с.
56. Блохина С.И. и др. Моделирование деятельности логопеда: критерии оценки // ИНФОР «БОНУМ», спец. Выпуск, Вестник РЦО РА1S1-1. – Челябинск, 2000. – С. 45 – 53.
57. Гольдштейн С.Л., Кудрявцев А.Г., Печеркин С.С. Об онтологическом пространстве системной интеграции // Вестник РАЕН. – 2014. – №1. – С. 133 – 139.

Материал поступил в редакцию 18.10.16.

Сведения об авторах

ГОЛЬДШТЕЙН Сергей Людвигович – доктор технических наук, профессор, заведующий кафедрой вычислительной техники, Физико-технологический институт Уральского федерального университета им. первого Президента России Б.Н. Ельцина, г. Екатеринбург
e-mail: s.l.goldshtein@urfu.ru

ПЕЧЕРКИН Сергей Сергеевич – кандидат физико-математических наук, научный сотрудник кафедры вычислительной техники,Физико-технологический институт Уральского федерального университета им. первого Президента России Б.Н. Ельцина, г. Екатеринбург
e-mail: specherkin@gmail.com

АВЕРЬЯНОВА Анна Николаевна – старший преподаватель кафедры вычислительной техники, Физико-технологический институт Уральского федерального университета им. первого Президента России Б.Н. Ельцина, г. Екатеринбург
e-mail: a.n.averianova@urfu.ru

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 519.6 : 004

Д.В. Виноградов

Эффективность ленивых вычислений для поиска сходств в ВКФ-системе

Исследуется вопрос ускорения вычислений сходства в ВКФ-методе с помощью ленивых вычислений. Показано, что ускорение достигается тем больше, чем больше отличаются числа обучающих примеров и признаков, их описывающих. Но даже в случае совпадения этих чисел ускорение происходит в два раза.

Ключевые слова: *формальный контекст, ВКФ-кандидат, замыкай-по-одному, ленивые вычисления, производящая функция*

ВВЕДЕНИЕ

Начиная с 1981 г., профессор В.К. Финн и его ученики развивают так называемый ДСМ-метод автоматического порождения гипотез [1], названный так в честь известного английского философа, экономиста и логика Джона Стьюарта Милля. Используя технику многозначных логик, В.К. Финну с коллегами [2, 3] удалось поставить систему индуктивной логики Милля [4] на четкие логические основания. Ключевой компонентой этого подхода является бинарная операция сходства [5]. Примерно в это же самое время аналогичный подход (но основанный не на логике, а на теории решеток) был разработан группой зарубежных исследователей под руководством профессора Рудольфа Вилле под названием теория формальных понятий (ТФП) [6]. Однако отечественный подход включил в рассмотрение контр-примеры, чего не имеется у зарубежных авторов.

Российские ученые под руководством проф. В.К. Финна смогли создать ряд компьютерных систем, получивших общее название ДСМ-систем интеллектуального анализа данных [7].

Однако обнаружились некоторые сложности в применении ДСМ-метода к анализу данных: во-первых, множество порождаемых ДСМ-гипотез может оказаться экспоненциально велико по сравнению с размером обучающей выборки; во-вторых, С.О. Кузнецовым [8], М.И. Забежайло [9, 10] и др. были доказаны пессимистические оценки сложности для многих ДСМ-процедур (так называемые **NP**-полнота и **#P**-полнота); в-третьих, попытки присоединить дедуктивный вывод оказались неудачными: Д.П. Скворцовым [11] установлена неарифметичность стандартного подхода через кванторы по конечным множествам, а автором работы [12] – невыразимость этой теории средствами логики предикатов первого порядка.

Наконец, автор [13] сумел обнаружить еще одну трудность: существование так называемых случайных ДСМ-гипотез, возникающих при вычислении сходства двух (или более) обучающих примеров, каждый из которых имеет свой механизм порождения целевого свойства. Это сходство оказывается фрагментом (набором общих признаков), случайно имеющемся в каждом из этих объектов. Возникновение таких сходств следует рассматривать как аналог «переобучения» для многих методов машинного обучения, когда максимальный учет информации из обучающей выборки приводит к модели, демонстрирующей плохую предсказательную способность.

Для преодоления указанных трудностей автором работы [14] был предложен вероятностно-комбинаторный подход. Так как некоторые ингредиенты заимствованы из теории формальных понятий, автор настоящей статьи назвал его **вероятностно-комбинаторный формальный метод**, сокращенно ВКФ-метод.

Для апробации ВКФ-метода автором была создана программная система, названная ВКФ-системой, которая была с успехом применена к нескольким массивам из репозитория данных для тестирования алгоритмов машинного обучения.

Ключевой процедурой ВКФ-системы является вероятностный алгоритм нахождения случайного ВКФ-кандидата с помощью спаривающей цепи Маркова. В нем используются операции «Замыкай-по-одному-вверх» и «Замыкай-по-одному-вниз».

Внимательное изучение этих операций позволяет выделить в них два шага: побитовое умножение и перебор всех объектов или признаков. Ясно, что второй шаг имеет сложность значительно превосходящую такую же для побитового умножения. Но если идет последовательность однотипных операций (или только вниз, или только вверх), то второй шаг можно вычислить один раз – в самом конце. Эта идея соответствует ленивым вычислениям.

Вопросу изучения эффективности внедрения описанных ленивых вычислений и посвящена настоящая работа.

ВСПОМОГАТЕЛЬНЫЕ ОПРЕДЕЛЕНИЯ И ФАКТЫ

Сходство является бинарной операцией на множестве X, объемлющем множество объектов, т.е. представляет собой отображение $\cap: X \times X \to X$. Элементы множества X мы будем называть *фрагментами*.

Для независимости результата нахождения сходства нескольких объектов, операция сходства должна удовлетворять аксиомам *нижней полурешетки*.

Для выражения тривиальности сходства имеется специальный *пустой фрагмент* $\varnothing$ со свойством наименьшего элемента.

Важнейшим примером для нас будет нижняя полурешетка, состоящая из битовых строк фиксированной длины с побитовым умножением в качестве операции сходства. Пустым фрагментом будет являться строка, состоящая из одних нулей. Каждый бит может быть отождествлен с бинарным признаком. Тогда битовая строка соответствует множеству признаков, в которых встречаются единицы. При этом операция сходства соответствует пересечению множеств признаков, а пустой фрагмент – пустому множеству признаков. В этом примере строка из одних единиц будет соответствовать наибольшему элементу.

Легко доказать известный результат теории решеток [15] о том, что **любую конечную (более того, полную) нижнюю полурешетку с наибольшим элементом можно превратить в решетку**.

Операция $x \cup y$ задается как последовательное сходство (в произвольном порядке) множества $\{z_1,...,z_k\}$ всех *общих верхних граней* для x и y – элементов полурешетки со свойствами $z_j \cap x = x$ и $z_j \cap y = y$. Сходством одноэлементного множества является фрагмент того элемента, который в нем содержится.

Важность этого примера объясняется тем, что операция побитового умножения допускает эффективную реализацию на современных ЭВМ. Существуют специальные классы объектов (например, *dynamic_bitset* в C++), реализующие удобное оперирование битовыми строками.

С другой стороны, теорема Рудольфа Вилле [6] утверждает, что эта конструкция позволяет построить любую конечную решетку из нижней полурешетки битовых строк с операцией побитового умножения, если к ней добавить наибольший элемент.

Собирая вместе битовые строки, представляющие объекты, мы получаем прямоугольную таблицу I, которую мы будем называть *формальным контекстом* [6]. Формальный контекст можно понимать как бинарное отношение между элементами множества O, которые мы называем *именами объектов* (или даже объектами), и элементами множества F, которые мы называем *признаками*. Если в строке, соответствующей объекту $o \in O$, и столбце, соответствующим фрагменту $f \in F$, стоит единица, то мы говорим, что *объект* ***o*** *обладает признаком* ***f*,** и обозначаем это через *oIf*. В противном случае, говорим, что *объект* ***o*** *не имеет признака* ***f***.

Для подмножества $A \subseteq O$ объектов его *сходством* называется подмножество $A' = \{f \in F : \forall o \in A[oIf]\} \subseteq F$. Полагаем $\varnothing' = F$.

На самом деле, это определение совпадает с последовательным вычислением побитового умножения строк, соответствующих отобранным во множество A объектов.

Для подмножества $B \subseteq F$ признаков его *сходством* называется подмножество $B' = \{o \in O : \forall f \in B[oIf]\} \subseteq O$. Полагаем $\varnothing' = o$.

Легко проверить следующие свойства [6]:

$$\forall A \subseteq O\left[A \subseteq A''\right] \qquad \forall B \subseteq F\left[F \subseteq F''\right] \tag{1}$$

$$\forall A_1 \forall A_2\left[A_1 \subseteq A_2 \Rightarrow A_1' \supseteq A_2'\right] \quad \forall B_1 \forall B_2\left[B_1 \subseteq B_2 \Rightarrow B_1' \supseteq B_2'\right] \tag{2}$$

$$\forall A \subseteq O\left[A' = A'''\right] \qquad \forall B \subseteq F\left[B' = B'''\right] \tag{3}$$

Определение 1.

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

Обычно такие пары называют *формальными понятиями*, но мы предпочитаем сменить название, так как оригинальное название подвергается обоснованной критике со стороны специалистов по философии и искусственному интеллекту.

Легко проверить, что **множество всех ВКФ-кандидатов** (для фиксированного формального контекста) **образует решетку**.

Теорема Р. Вилле [6, 15] гласит: **Любая конечная решетка изоморфна решетке всех ВКФ-кандидатов для подходяще выбранного формального контекста**.

В работе [16] конструктивное доказательство теоремы Вилле применяется для формализации структур признаков при описании медицинских данных.

Введем ключевое понятие, определяемое в дальнейшем изложении.

Определение 2.

Операция ***замыкай-по-одному-вниз*** на ВКФ-кандидате $\langle A, B\rangle$ и объекте $o \in O$ порождает пару $CbODown(\langle A, B\rangle, o) = \left\langle \left(A \cup \{o\}\right)'', \left(A \cup \{o\}\right)' \right\rangle$. Операция ***замыкай-по-одному-вверх*** на ВКФ-кандидате $\langle A, B\rangle$ и признаке $f \in F$ порождает пару $CbOUp(\langle A, B\rangle, f) = \left\langle \left(B \cup \{f\}\right)', \left(B \cup \{f\}\right)'' \right\rangle$.

Операция *CbODown* соответствует шагу алгоритма «Замыкай-по-одному», который был предложен С.О. Кузнецовым [17] для вычисления всех ВКФ-кандидатов перебором сверху-вниз.

Лемма 1. *Для любого ВКФ-кандидата* $\langle A,B\rangle$ *и любого объекта* $o\in O$ *пара* $ChODown(\langle A,B\rangle,o)$ *является ВКФ-кандидатом.*

Аналогично, для любого ВКФ-кандидата $\langle A,B\rangle$ *и любого признака* $f\in F$ *пара* $ChOUp(\langle A,B\rangle,f)$ *является ВКФ-кандидатом.*

Это утверждение легко проверяется с использованием формул (1), (3) и Определения 2.

Определение 3.

Порядок на ВКФ-кандидатах зададим правилом $\langle A_1,B_1\rangle\le\langle A_2,B_2\rangle$, если $B_1\subseteq B_2$.

Это двойственное (с точки зрения теории формальных понятий) определение приводится в настоящем виде для согласованности с традицией отечественной школы.

Лемма 2. *Для всякой упорядоченной пары ВКФ-кандидатов* $\langle A_1,B_1\rangle\le\langle A_2,B_2\rangle$ *и любого объекта* $o\in O$ *имеем* $ChODown(\langle A_1,B_1\rangle,o)\le ChODown(\langle A_2,B_2\rangle,o)$.

Аналогично, для всякой упорядоченной пары ВКФ-кандидатов $\langle A_1,B_1\rangle\le\langle A_2,B_2\rangle$ *и любого признака* $f\in F$ *имеем* $ChOUp(\langle A_1,B_1\rangle,f)\le ChOUp(\langle A_2,B_2\rangle,f)$.

Это утверждение легко проверяется с использованием формул (2) и Определения 3.

Data: множество обучающих (+)-примеров; внешние функции *ChOUp*(,) и *ChODown*(,) операций «замыкай-по-одному»
Result: ВКФ-кандидат $\langle A,B\rangle$
$O :=$ (+)-примеры, $F :=$ признаки; $I\subseteq O\times F$ – формальный контекст для (+)-примеров;
$R := O\cup F$; $Min := \langle O,O'\rangle$; $Max := \langle F',F\rangle$;
while ($Min \neq Max$) **do**
    Выбираем случайный элемент $r\in R$;
    if ($r\in O$) **then**
        $Min := ChODown(Min,r)$;;
    end
    else
        $Min := ChOUp(Min,r)$;;
    end
end
$\langle A,B\rangle := Min$:

Алгоритм 1. Спаривающая цепь Маркова

Заметим, что состоянием изменяемых переменных в цикле (= состоянием спаривающей цепи Маркова) является упорядоченная пара ВКФ-объектов $\langle A_1,B_1\rangle\le\langle A_2,B_2\rangle$.

Первоначально меньший ВКФ-кандидат совпадает с наименьшим ВКФ-кандидатом $Min := \langle O,O'\rangle$, а больший – с наибольшим $Max := \langle F',F\rangle$.

В цикле к обоим ВКФ-кандидатам применяется одна и та же операция *ChODown* с выбранным объектом, или *ChOUp* с выбранным признаком.

Процесс останавливается, когда меньший ВКФ-кандидат совпадет в большим. Тогда этот общий ВКФ-кандидат и выдается Алгоритмом 1.

Следующая теорема из статьи [14] объясняет, откуда здесь возникает цепь Маркова.

Теорема 1. *Алгоритм 3 соответствует цепи Маркова.*

Определение 4.

Состояние вида $\langle A,B\rangle=\langle A,B\rangle$ спаривающей цепи Маркова для совпадающей пары ВКФ-кандидатов называется ***эргодическим***. Состояние вида $\langle A_1,B_1\rangle<\langle A_2,B_2\rangle$ называется ***невозвратным***.

Теорема 2. *Вероятность того, что состояние* $\langle A_1,B_1\rangle\le\langle A_2,B_2\rangle$ *спаривающей цепи Маркова, окажется невозвратным, стремится к нулю, когда* $t\to\infty$.

В [18] приводится доказательство этого результата, но это – классический результат в теории цепей Маркова [19], поэтому здесь мы не будем приводить его доказательство.

ОСНОВНОЙ РЕЗУЛЬТАТ

В настоящее время ВКФ-кандидаты вычисляются согласно Алгоритму 1 с использованием операций *ChODown* и *ChOUp*.

Согласно Определению 2

$$ChODown(\langle A,B\rangle,o)=\left\langle\left(A\cup\{o\}\right)'',\left(A\cup\{o\}\right)'\right\rangle.$$

Если вычисление пересечения $\left(A\cup\{o\}\right)'=B\cap o'$ фрагмента текущего ВКФ-кандидата с фрагментом выбранного объекта ***o*** соответствует побитовому умножению соответствующих строк, то операция $\left(A\cup\{o\}\right)''=\left(B\cap o'\right)'$ формирования нового списка родителей может потребовать побитово перемножить с полученным ранее пересечением почти все объекты, чтобы проверить, обладает ли еще какой-нибудь объект полученным пересечением.

Для улучшения ситуации предлагается (лениво) откладывать вычисления второй производной, пока последовательный выбор нескольких объектов для *ChODown* не сменится выбором признака с переходом к операции *ChOUp*.

Аналогично, операция *ChOUp* имеет в своем составе потребляющую много времени компоненту $\left(B\cup\{f\}\right)''=\left(A\cap f'\right)'$. Здесь тоже можно лениво откладывать вычисления этой части до тех пор, пока выбор нескольких признаков для *ChOUp* не сменится выбором объекта с переходом к операции *ChODown*.

Возникает вопрос о степени экономии, достигаемой такой процедурой. Мы предложим анализ этой проблемы с помощью техники рекуррентных событий [20].

Рассмотрим последовательность пар типов (объект или признак) объектов, выбираемых в ходе работы Алгоритма 1. Ясно, что это – последовательность (вообще говоря, бесконечная) испытаний Бернулли $\langle\sigma_1,...,\sigma_j,...\rangle$ с вероятностью успеха (например, выбо-

ра признака), равной $p=\frac{n}{n+k}$, где n – число признаков, а k– число обучающих примеров.

Определим два рекуррентных события, где $E_\sigma=\left[\sigma_j=\sigma\neq(1-\sigma)=\sigma_{j+1}\right]$, где $\sigma\in\{0,1\}$

Нас интересуют вероятности

$v_j^{(\sigma)}=P\left[E_\sigma \text{ впервые происходит при } j-\text{ом испытании}\right]$

и соответствующая производящая функция моментов

$$\psi_\sigma(z)=\sum_{j=1}^{\infty}v_j^{(\sigma)}\cdot z^j=$$

$$=\sum_{j=1}^{\infty}P\left[\sigma_j=\ldots=\sigma_j=\sigma\neq1-\sigma=\sigma_{j+1}\right]\cdot z^j$$

для распределения вероятностей $v_j^{(\sigma)}$ ($v_0^{(\sigma)}=0$ по соглашению).

Легко видеть, что

$$v_j^{(0)}=P\left[E_0 \text{ впервые происходит при } j-\text{ом испытании}\right]=p\cdot(1-p)^j. \quad (4)$$

$$v_j^{(1)}=P\left[E_1 \text{ впервые происходит при } j-\text{ом испытании}\right]=(1-p)\cdot p^j. \quad (5)$$

Суммирование геометрических прогрессий дает:

Лемма 3.

Производящие функции моментов равны

$$\psi_0(z)=\frac{p\cdot(1-p)\cdot z}{1-(1-p)\cdot z}=p\cdot\left(\frac{1}{1-(1-p)\cdot z}-1\right)$$

и

$$\psi_1(z)=\frac{p\cdot(1-p)\cdot z}{1-p\cdot z}=(1-p)\cdot\left(\frac{1}{1-p\cdot z}-1\right).$$

Найдем теперь среднее время d между переходами:

$$d:=E\begin{bmatrix}\text{первый номер испытания},\\ \text{когда происходит событие } E_0 \text{ или } E_1\end{bmatrix}.$$

Теорема 3.

Выполняется $d=\frac{p}{(1-p)}+\frac{(1-p)}{p}$.

Доказательство. Вычисляем производную

$$\psi'(z)=\psi_0'(z)+\psi_1'(z)=\frac{p\cdot(1-p)}{\left(1-(1-p)\cdot z\right)^2}+\frac{(1-p)\cdot p}{\left(1-p\cdot z\right)^2}$$

и подставляем туда 1, так как $d=\psi'(1)$.

Итак, выигрыш от введения ленивых вычислений может составить $d=\frac{p}{(1-p)}+\frac{(1-p)}{p}=\frac{n}{k}+\frac{k}{n}$ раз.

Ясно, что это число тем больше, чем больше разница между числом k – обучающих примеров и числом n – признаков, используемых для описания объектов. Но даже в худшем случае $k=n$ – это сокращение вызовов трудоемкой операции не меньше двух раз.

Следует отметить, что при реализации ленивых вычислений слегка изменится распределение выдачи Алгоритма 1. Например, исчезнет возможность спариться на наибольшем и наименьшем ВКФ-кандидатах. Но с содержательной точки зрения, это несущественно, так как такие ВКФ-кандидаты определенно не пройдут дополнительные тесты на перевод их в статус ВКФ-гипотез: максимальный ВКФ-кандидат не будет иметь достаточного числа родителей, а минимальный будет иметь контр-примеры.

В настоящей работе получена оценка эффективности ленивых вычислений для нахождения сходства в ВКФ-методе. Показано, что ускорение достигается тем больше, чем больше отличаются числа обучающих примеров и признаков, их описывающих. Но даже в случае совпадения этих чисел ускорение вычислений происходит в два раза.

* * *

Автор благодарит проф. В.К. Финна за внимание к работе, проф. Е.М. Бениаминова за полезные обсуждения и своих коллег по лаборатории 35 ФИЦ ИУ РАН за поддержку и полезные дискуссии.

СПИСОК ЛИТЕРАТУРЫ

1. ДСМ-метод автоматического порождения гипотез: Логические и эпистемологические основания / ред.: В.К. Финн, О.М. Аншаков. – М.: URSS, 2009. – 432 с.
2. Аншаков О.М., Скворцов Д.П., Финн В.К. Логические средства экспертных систем типа ДСМ // Семиотика и информатика. – 1986. – Вып. 28. – С. 65–102
3. Аншаков О.М., Скворцов Д.П., Финн В.К. О дедуктивной имитации некоторых вариантов ДСМ-метода автоматического порождения гипотез // Семиотика и информатика. – 1993. – Вып. 33. – С. 164–233
4. Милль Дж.Ст. Система логики силлогистической и индуктивной: Изложение принципов доказательства в связи с методами научного исследования: пер. с англ. Изд. 5. – М.: URSS, 2011. – 832 с.
5. Гусакова С.М., Финн В.К. Сходства и правдоподобный вывод // Известия АН СССР. Сер. «Техническая кибернетика». – 1987. – № 5. – С. 42–63
6. Ganter Bernard, Wille Rudolf. Formal Concept Analysis. Transl. from German. – Berlin: Springer–Verlag, 1999. – 284 p.
7. Финн В.К. Об интеллектуальном анализе данных // Новости искусственного интеллекта. – 2004. – № 3. – С. 3–18.
8. Кузнецов С.О. Интерпретация на графах и сложностные характеристики задач поиска закономерностей определенного вида // Научно-техническая информация. Сер. 2. – 1989. – № 1. – С. 23–28.
9. Забежайло М.И. О некоторых оценках сложности вычислений в ДСМ-рассуждениях. Часть I. // Искусственный интеллект и принятие решений. – 2014. – № 2. – С. 3-18; Zabezhailo M.I. Some Capabilities of Enumeration Control in the DSM Method. Part One // Scientific and Technical Information Processing. – 2014. – № 6. – P. 335-347.

10. Забежайло М.И. О некоторых оценках сложности вычислений в ДСМ-рассуждениях. Часть II // Там же. – 2014. – № 3. – С. 3-21; Zabezhailo M.I. Some Capabilities of Enumeration Control in the DSM Method. Part Two // Scientific and Technical Information Processing. – 2014. – № 6. – P. 348-361.
11. Скворцов Д.П. О некоторых способах построения логических языков с кванторами по кортежам // Семиотика и информатика. – 1983. – Вып. 20. – С. 102–126
12. Виноградов Д.В. Формализация правдоподобных рассуждений в логике предикатов // Научно-техническая информация. Сер. 2. – 2000. – № 11. – С. 17–20; Vinogradov D.V. Formalizing plausible arguments in predicate logic // Automatic Documentation and Mathematical Linguistics. – 2000. – Vol. 34, №6. – P. 6-10.
13. Виноградов Д.В. Вероятность порождения случайного ДСМ-сходства при наличии контрпримеров // Научная и техническая информация, Сер. 2. – 2015. – № 3. – С. 1–5; Vinogradov D.V. The Probability of Encountering an Accidental DSM Similarity in the Presence of Counter Examples // Automatic Documentation and Mathematical Linguistics. – 2015. – Vol. 49, №2. – P. 43-46.
14. Vinogradov D.V. VKF-method of hypotheses generation // Communications in Computer and Information Science. – 2014. – Vol. 436. – P. 237–248
15. Davey B.A., Priestley H.A. Introduction to Lattices and Order. 2nd eds. – Cambridge: Cambridge University Press, 2002. – 298 p.
16. Панкратова Е.С., Виноградов Д.В. Формальное описание настройки интеллектуальных ДСМ-систем на область клинической и лабораторной диагностики // Научно-техническая информация. Сер. 2. – 2011. – № 9. – С. 1–5; Pankratova E.S., Vinogradov D.V. Formal Description of Adaptation of Intelligent JSM-Systems for Clinical and Laboratory Data Analysis // Automatic Documentation and Mathematical Linguistics. – 2011. – Vol. 45, № 5. – P. 213-217.
17. Кузнецов С.О. Быстрый алгоритм построения всех пересечений объектов из нижней полурешетки // Научно-техническая информация. Сер. 2. – 1993. – № 1. – С. 17–20.
18. 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.
19. Кемени Дж., Снелл Дж., Кнепп А. Счетные цепи Маркова / пер. с англ. – М.: Наука, Гл. ред. физ.-мат. лит.,1987. – 416 с.
20. Феллер В. Введение в теорию вероятностей и ее приложения. Т. 1 / пер. с англ. – М.: Мир, 1984. – 528 с.

Материал поступил в редакцию 26.01.17.

Сведения об авторе

ВИНОГРАДОВ Дмитрий Вячеславович – кандидат физико-математических наук, старший научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН и Российского государственного гуманитарного университета, Москва
e-mail: vinogradov.d.w@qmail.com

УДК 81 : 001.101

В.И. Хайруллин

Об одном из базовых принципов структурирования информации

Рассматривается проблема структурирования информации, основывающаяся на принципе языковой избирательности, который признается одним из базовых принципов. Отстаивается позиция, в соответствие с которой понятие языковой избирательности формировалось в науке о языке как осознание факта неодинакового членения отражаемой реальности разными языками. Значительное место отводится вопросу смысловой структуры.

Ключевые слова: *информация, информационный, принцип, ситуация, термин, языковая избирательность, действительность, смысловая структура.*

Проблема организации и структурирования информации в последнее время привлекает внимание специалистов [1–6]. В настоящей статье рассматривается один из базовых принципов структурирования информации, а именно – принцип языковой избирательности.

Прежде всего, представляется необходимым очертить границы исследования, т.е. дать определение понятия.

Итак, языковая избирательность заключается в наличии у каждого языка предпочтительных способов как отражения предметов, обозначения процессов, так и описания ситуаций действительности. Например, один и тот же предмет или явление могут быть представлены в языке с самых различных сторон: по длине, высоте, ширине, плотности, гибкости, что фиксируется в семантике термина, представляющего данный предмет или явление. При номинации предмета или явления происходит «поворачивание» его разными сторонами. Человек видит эти разные стороны, однако не все в одинаковой степени отчетливо. То же самое происходит при описании одной и той же ситуации как в одном языке, так и в разных языках: в каждом из них ситуация «поворачивается» разными сторонами, причем у каждого языка могут быть предпочтительные стороны, или признаки, для представления этой ситуации. Такое «разностороннее» описание ситуации различными языками представляется наиболее полным, поскольку те или иные признаки ситуации, оставшиеся неописанными в одном языке, могут называться при описании этой ситуации в другом языке.

Тот факт, что языки по-разному расчленяют отражаемую действительность, не свидетельствует о существовании какого-либо «магического круга», который язык якобы ограничивает в сознании человека, и за пределы которого человек не может выйти. Человек способен осознавать условность языковой картины мира и сопоставлять ее с фактическим знанием и опытом. Вместе с тем, важность изучения явления языковой избирательности при построении особых картин мира как разных способов организации информации об отражаемой реальности не вызывает сомнения. Без этого трудно правильно осознать устройство и функционирование языковой системы, самой сущности человеческого языка, выполняющего функцию орудия мысли и коммуникации, а также невозможно понять те языковые способы, по которым информация структурируется в языке.

Понятие языковой избирательности формировалось в науке о языке как осознание факта неодинакового членения отражаемой реальности разными языками и изучалось путем сопоставления значений единиц языковой системы. Весомо меньше внимания уделялось проявлениям языковой избирательности в процессе функционирования языков при построении речевых произведений – высказываний. Анализ лингвистического материала приводит к выводу, что избирательность способа описания ситуации прослеживается и в выборе признаков описываемой ситуации, т. е. при оформлении смысловой структуры высказывания, которое рассматривается в современном языкознании как важнейшая коммуникативная единица, актуализирующая единицы языковой системы разного уровня (слово-термин, терминологическое словосочетание, предложение-специальный текст), соотнесенная с конкретной коммуникативной ситуацией и способная выступать в качестве единицы более крупного коммуникативно-прагматического образования – дискурса. В высказывании проявляются все свойства структуры и функционирования языка, в том числе и принцип языковой избирательности.

Наиболее значительное проявление принципа языковой избирательности – это расхождения в семантике сопоставляемых терминологических единиц разных языков, раскрывающие своеобразие членения отражаемой реальности. Исследование языковой избирательности в отношении речевых высказываний,

описывающих процессы и явления, имеет целью раскрыть особенности семантики высказываний, свойственные данному языку и данному функциональному стилю. При этом речь должна идти не о конкретном содержании отдельных высказываний, а о способах структурирования информации в высказывании вообще, о смысловой структуре высказывания.

В современной науке понятие смысловой структуры не имеет однозначного определения, однако большинство специалистов в области информационной семантики признают, что смысловая, или семантическая, структура состоит из элементов смысла, которыми выступают либо глубинные падежи, или актанты, семантические компоненты, семантические функции, или даже аргументы и элементарные смыслы – у разных исследователей они именуются по-разному.

При всем различии имеющихся концепций, что отражает сложность и многогранность структуры информационного содержания высказывания, большинство из них ставит целью выявить глубинные аспекты смысла, в большей или меньшей степени не зависящие от собственно языковых особенностей организации информации в высказывании. Это та часть информационного содержания, которая непосредственно определяется описываемой ситуацией, и ее структура отражает структуру мыслительных образов, воспроизводящих в идеальной форме эту ситуацию. Изучение глубинных связей содержания информации с его когнитивными и реальными коррелятами — важнейшая задача современной науки, решение которой должно пролить свет на сложную динамику взаимодействия языка и мышления. Вместе с тем, несомненный интерес представляют и собственно когнитивно-лингвистические аспекты смысловой структуры, различия в организации передаваемой информации, которые «навязываются» спецификой каждого отдельного языка.

Понятие «смысловая структура» должно охватывать особенности упорядоченного представления всего смысла высказывания, который выступает как весьма сложное и многоплановое образование, формирующееся под воздействием как глубинных, так и поверхностных факторов — семантики языковых единиц, составляющих высказывание. При описании смысловой структуры следует учитывать и семантику лексических единиц, организующих эти единицы и устанавливающих определенные связи между ними. При актуализации высказывания, т. е. при включении его в текст, смысловая структура дополняется способом соотнесения с действительностью, который включает коммуникативную функциональность, ситуативную ориентированность и избирательность способа описания ситуации. Изменение любого элемента смысловой структуры приводит к изменению содержания высказывания.

Таким образом, смысловая структура создается набором языковых единиц, их синтаксической организацией и способом соотнесения с передаваемой информацией и отражаемой действительностью. Иными словами, компонентами (элементами) смысловой структуры являются семантика синтаксической модели; семантика лексических единиц, входящих в высказывание; ситуативно-коммуникативные элементы смысла, отражающие способ соотнесения высказывания с действительностью. Ситуативно-коммуникативные элементы, в свою очередь, включают: коммуникативную функциональность, ситуативную ориентированность и избирательность способа описания ситуации.

Элементы языковой избирательности следует искать не только в различии грамматики и лексики, но и в способах представления ситуации. В связи с этим основное внимание мы обращаем не на различия грамматики и лексики, а на различия в самом построении высказывания.

СПИСОК ЛИТЕРАТУРЫ

1. Арутюнов В.В. О некоторых результатах приоритетных исследований в области информационной безопасности // Научно-техническая информация. Сер. 1. – 2016. – № 2. – С. 1-7; Arutyunov V.V. The Results of Priority Research in the Field of Information Security // Scientific and Technical Information Processing. – 2016. – Vol. 43, № 1. – P. 42-46.
2. Белоногов Г.Г., Гиляревский Р.С. Хорошилов А.А. О природе информации // Научно-техническая информация. Сер. 2. – 2009. – № 1. – С. 1-7.
3. Гиляревский Р.С., Черный А.И. Доктор Юджин Гарфилд: научно-информационная деятельность // Научно-техническая информация. Сер. 1. – 2009. – № 5. – С. 32-35.
4. Лаврик О.Л., Шевченко Л.Б. Информационное сопровождение как новый этап развития информационной деятельности // Научно-техническая информация. Сер. 1. – 2006. – № 9. – С. 19-22.
5. Терещенко С.С. Структуры данных в автоматизированных информационных системах // Научно-техническая информация. Сер. 2. – 1997. – № 9. – С. 8-17; Tereshchenko S.S. Data structures in automatic information systems // Automatic Documentation and Mathematical Linguistics. – 1997. – Vol. 31, № 5. – P. 11-22.
6. Хайруллин В.И. Информативность терминологии международно-правовых документов с позиций перевода // Научно-техническая информация. Сер. 1. – 2016. – № 2. – С. 38-40.

Материал поступил в редакцию 01.02.17.

Сведения об авторе

ХАЙРУЛЛИН Владимир Ихсанович – доктор филологических наук, профессор, заведующий кафедрой «Иностранный язык» Уфимского государственного нефтяного технического университета, профессор кафедры международного права и международных отношений Башкирского государственного университета, г. Уфа.
e-mail: vladimir-blt@mail.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322 : 002

А.Н. Либкинд, В.А. Маркусова, И.А. Либкинд, Н.М. Камень, В.Ю. Фадеев

Наукометрические аспекты идентификации авторов российских публикаций*

Изложены результаты наукометрического исследования массива российских публикаций, отраженных в Web of Science, для выявления реального количества отечественных ученых, принимавших участие в их написании. Предложены алгоритмы и методы идентификации авторов, а также оценки эффективности этих методов.

Ключевые слова: *идентификация, написание фамилий, имен и отчеств авторов, соотношение автороуказаний и реальных ученых, публикационная активность, российский вклад в мировую науку, Web of Science*

ВВЕДЕНИЕ

Происходящий в мире рост научной литературы, особенно в странах тихоокеанского континента, и высокие темпы развития цифровых депозитариев связаны с потребностью устранения неоднозначности в написании фамилии авторов. Проблема идентификации является одной из широко обсуждаемых в современной библиометрии и трудно разрешимой, поскольку по ряду причин имя одного и того же автора может быть написано по-разному. Атрибуция публикации автора имеет большое значение для понимания возникновения новых идей, исследования цитируемости, изучения модели соавторства, предсказания возникновения новых областей [1] исследования мобильности и при оценке научной деятельности индивидуальных исследователей. Анализ литературы показал, что проблема идентификации авторов и установления авторства выходит далеко за рамки задач оценки состояния и тенденций в науке и охватывает самые различные области: от филологии до судебно-медицинской экспертизы [2]. В последние годы в Интернете появились различные системы Research gate, Mendeley, Goodle Scholar, требующие индивидуальной регистрации. Такие системы регистрации как *ResercherID – Web of Science, Author ID – Scopus, ORCID* также являются попыткой научного сообщества решить проблему идентификации авторов с его публикациями. Условно направление исследований по устранению неоднозначности можно разделить на работы с учетом национальных особенностей языков: хинди, майя, – и исследования, связанные с использованием теории графов или алгоритмов кластеризации [3, 4].

Интересный подход к решению этой проблемы был предложен группой специалистов в уже упомянутой работе [1]. Для устранения неоднозначности в написании фамилии одного и того же автора и его идентификации со списком публикаций предложена новая модель для идентификации публикаций индивидуальных исследователей Нидерландов. Изначальной точкой исследования был список, состоящий из 8378 фамилий профессоров за 1980–2011 гг., зарегистрированных в Национальной БД Нидерландов, имеющих публикации в *Web of Science*. В процессе анализа список публикаций был расширен с использованием идентификаторов Scopus, адресов электронной почты, ручной верификация и связи автор-адрес. В качестве «золотого стандарта» использовалась БД, содержащая верифицированные фамилии 1400 авторов в БД Центра по науке и технике Нидерландов *CWTS*. Трудности с решением проблемы атрибуции публикаций с конкретным автором приводят к ошибкам в такой известной аналитической БД как *Essential Science Indicators*, принадлежавшей фирме *Thomson Reuters* (ныне Clarivate Analytics). Как сообщается в работе [5] китайские и корейские авторы являются значительно более продуктивными в БД *ESI* не потому, что они являются «звездами», а потому что отсутствует однозначная атрибуция авторов с одинаковыми фамилиями или с одинаковыми написаниями разных фамилий.

В настоящей работе предпринята попытка комплексного анализа особенностей идентификации авторов. Обработаны данные о 6089 конкретных ученых – авторов российских публикаций, в том числе

* Работа выполнена по гранту РФФИ № 17-07-00256 и грантам РГНФ № 17-02-00078 и № 17-02-00157

о 4487 отечественных ученых, Исследуемый массив насчитывал 37942 автороуказаний[1]. Для формирования исходных данных используется международная информационная система *Web of Science* (*WoS*). Подавляющее количество библиографических описаний публикаций, а, следовательно, и авторов (как российских, так и зарубежных), в этой базе данных приводится на английском языке. Например, за 2009 – 2013 гг. доля описаний российских публикаций на английском языке в БД *Science Citation Index – Expanded* (*SCI-E*), входящей в *WoS*, составляет 94,9%.

ЦЕЛИ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Основная цель идентификации, состоит в том, чтобы записи имен авторов российских публикаций были «трансформированы» в сведения о конкретных отечественных ученых и об их зарубежных коллегах-соавторах. Таким образом, в результате идентификации должен стать возможным переход от понятия «автор» к понятию «ученый» или, говоря обобщенно, к понятию «персона». Эту цель можно «разложить» на несколько составляющих.

Во-первых, разделить одинаковые (совпадающие) исходные написания фамилии, имени и отчества (ФИО), которые относятся к разным исследователям, что является первым шагом на пути «превращения» автора в персону. Забегая вперед, скажем, что этого разделения можно достичь привлечением данных об аффилиации, которые автор указал в соответствующей публикации, а также данных о тематике этой публикации (тематические категории *WoS*).

Во-вторых, стандартизовать (унифицировать) различающиеся варианты написаний ФИО, которые относятся к одному и тому же исследователю, и тем самым выявить все публикации анализируемого массива, автором которых является данный ученый.

В-третьих, по возможности, установить пол автора, что самым непосредственным образом будет способствовать решению задач идентификации. Кроме того, гендерная статистика является важной демографической характеристикой кадрового потенциала как науки в целом, так и отдельных ее разделов.

В-четвертых, определить национальную принадлежность автора; мы будем учитывать не только то, в какой стране работает данный ученый, но и его, условно говоря, «этническую» принадлежность. При этом авторов российских публикаций мы будем рассматривать не так, как это принято в этнографии, а очень обобщенно: с той степенью обобщенности (неточности, огрубления), которая будет достаточной для решения задач идентификации. Судить об этнической принадлежности мы, чаще всего, будем по ФИО автора. Сочетание этих двух характеристик (страна, где работает автор российской публикации и его этническая принадлежность) позволит сформировать несколько «социально-этнических» групп авторов (соавторов) российских публикаций: российские ученые, экспаты – граждане России, экспаты – выходцы из России/из бывшего СССР, иностранные ученые. В результате, мы сможем определить роль и вклад авторов каждой из этих групп в отечественную науку.

Кроме того, идентификация и «превращение» автора в «персону» позволит получить данные о вкладе каждого рассматриваемого ученого в конкретный раздел исследований, установить тематический спектр и научный уровень его изысканий. Обобщая эти сведения, можно выявить ведущих специалистов в той или иной области и разделе науки, определить количество ученых, занятых в данном разделе исследований, а также число активных исследователей в конкретной организации, городе, регионе и в стране в целом. В конечном счете – выявить динамику всех этих характеристик.

Для достижения указанных целей следует решить три взаимно дополняющие задачи. Первая состоит в разделении *одинаковых исходных написаний ФИО* авторов, которые относятся к *разным исследователям*. Этого можно достичь с помощью программной идентификации[2], когда одинаковым написаниям ФИО, при условии, что им соответствуют одинаковые написания организаций – мест работы авторов и других атрибутов аффилиации, будет приписан один и тот же идентификатор. Тем же автороуказаниям (записям), которым при полностью идентично написанных ФИО соответствуют разные организации, будут присвоены разные идентификаторы.

Вторая задача, в отличие от первой, состоит не в разделении одинаковых (по написанию) записей ФИО, а, напротив – в объединении тех записей ФИО, которые несколько *различаются по написанию*, однако принадлежат *одному и тому же ученому*. Эта задача, в основном, может решаться с помощью визуальной идентификации, хотя при этом необходимо использовать и программные методы.

И наконец, третья задача, которая, по сути, представляет собой уточнение результатов, полученных при решении первой и второй задач. Дело в том, что в ряде случаев автороуказания с одинаковыми ФИО и различной аффилиацией, могут принадлежать одному и тому же ученому, который просто указал различные места работы. Решение этой задачи состоит в выявлении таких случаев и в последующем объединении соответствующих автороуказаний.

[1] Автороуказание – запись в массиве (в списке, в таблице), содержащая, как минимум, два элемента:
1) фамилию, имя и отчество (ФИО), которые указал автор (соавтор) публикации; 2) указание на эту публикацию, оформленное в виде ссылки или идентификатора. При этом число автороуказаний, соответствующих некоторой публикации, равно количеству авторов этой публикации. В рассматриваемом массиве автороуказание включает также данные аффилиации автора (страна, город и организация – место работы автора), а публикация представлена ее идентификатором, который, при необходимости позволяет перейти непосредственно к полному ее библиографическому описанию. Таким образом, далее упоминая автороуказание, мы будем иметь в виду триаду «ФИО-публикация-аффилиация».

[2] Здесь следует отметить, что программной идентификации автора должна предшествовать идентификация организаций – мест его работы, которая основывается на данных об аффилиации (страна, город, организация), указанной автором в соответствующей публикации.

ФОРМИРОВАНИЕ МАССИВА ДЛЯ ИССЛЕДОВАНИЯ

Уточним некоторые понятия, которые использованы нами при формировании массива авторов, их идентификации и рассмотрении результатов этой идентификации.

Российская публикация – это такая статья, автором/соавтором которой является хотя бы один российский ученый. Она признается российской и в том случае, когда в число ее авторов входят зарубежные ученые. Следовательно, в качестве авторов российских публикаций мы рассматриваем как российских ученых, так и их зарубежных соавторов.

Автороуказание в исследуемом массиве, как уже упомянуто, включает, помимо ФИО в авторском написании и ссылки (идентификатора) на публикацию, данные аффилиации автора (страна, город и организация – место работы автора). Идентификатор при необходимости позволяет перейти к полному библиографическому описанию статьи.

Полное и неполное написание ФИО – предварительный анализ исследуемого массива показал, что, по степени полноты, встречаются следующие написания ФИО авторов:

1) фамилия (surname, last name) и два инициала (initials) – имени (first name) и отчества (patronymic);
2) фамилия, имя и второй инициал;
3) фамилия и имя;
4) фамилия и первый инициал;
5) фамилия, имя и отчество;
6) фамилия и три инициала.

Варианты написаний ФИО конкретного ученого (персоны) – совокупность различающихся и, при этом, неповторяющихся (уникальных) написаний ФИО данного ученого, указанных в его публикациях и включенных в исследуемый массив.

Массив автороуказаний – совокупность автороуказаний, соответствующих исследуемому массиву публикаций. Другими словами, это совокупность всех без исключения триад «публикация – ФИО – аффилиация». Пользуясь терминологией теории множеств можно сказать, что совокупность ФИО в нашем массиве автороуказаний представляет собой мультимножество [6].

Для того чтобы получить надежные и достоверные данные об авторах российских публикаций, отследить динамику количества и другие характеристики этих публикаций, временной охват должен быть достаточно большим – не менее пяти лет. Исходный массив автороуказаний, соответствующий этим публикациям, составил бы, оценочно, более одного миллиона записей. Понятно, что обработка такого массива, включая идентификацию содержащихся в нем элементов библиографических данных, заняла бы слишком много времени, что для целей настоящего исследования было неприемлемо.

Пришлось сформировать относительно небольшой, но репрезентативный массив. В нашем случае репрезентативность означает, что, во-первых, авторы и публикации, попавшие в массив, представляют как можно большее количество тематических разделов отечественной науки; во-вторых, необходимо, чтобы набор написаний ФИО авторов в этом массиве охватывал бы как можно больше различных их вариантов и, конечно, максимально возможное число ученых – авторов этих публикаций.

Видимо, наилучшим способом удовлетворить требования репрезентативности было бы формирование совокупности автороуказаний путем случайной выборки. Однако последовательная (строгая) реализация такого подхода наталкивается на ряд трудностей. В связи с этим было принято паллиативное решение, которое позволило сформировать массив в форме квазислучайной выборки.

На первом этапе были отобраны российские публикации 2009 – 2013 гг., которые содержались в крупнейшей БД *WoS* – *SCI-E* при условии, что эти публикации соответствовали бы одной тематической категории *WoS* «*Nanoscience & Nanotechnology*» (*N&N*). Выбор тематики объясняется тем, что это направление исследований достаточно динамично, а число российских публикаций по этой тематике за указанный период относительно невелико – 2319 документов.

Второй этап заключался в выделении уникальных (неповторяющихся) написаний ФИО авторов отобранных публикаций по тематике *N&N* (их число оказалось равным 6726) и в формировании списка этих ФИО.

Третий и заключительный этап формирования исходных данных для исследования состоял в следующем: для каждого написания ФИО из списка, полученного на втором этапе, помимо автороуказаний, соответствующих *N&N*, были установлены и все прочие российские публикации 2006 – 2013 гг., которые были отражены в БД *WoS*. Необходимо учитывать, что один и тот же вариант написания ФИО может соответствовать нескольким ученым, область научных интересов которых будет находиться вне тематики *N&N*.

С помощью такого приема мы смогли охватить не только ученых – авторов публикаций по тематике *N&N*, но и значительно более широкий круг исследователей, которые были авторами публикаций по другим тематикам и, следовательно, охватить эти публикации. Кроме того, в исследуемый массив попали публикации, авторами которых были ученые – авторы статей по тематике *N&N*, которые работали и в других разделах исследований.

В результате был сформирован массив, который насчитывает 37942 автороуказаний и включает 15456 публикаций[3]. Наш метод достижения репрезентативности можно признать удовлетворительным, поскольку публикации, представленные в рассматриваемом массиве автороуказаний, распределены по 156 тематическим категориям *WoS* и охватывают все области естественнонаучных исследований. Набор

[3] Заметим, что определение среднего количества авторов в публикации, исходя из приведенных здесь численных значений автороуказаний и публикаций, было бы некорректным. Дело в том, что массив автороуказаний был сформирован случайным (вернее, квазислучайным) образом. В этом массиве в больших количествах могут встречаться случаи, когда публикация, написанная коллективом авторов, будет представлена неполным набором этих авторов или, более того, иногда одним единственным автором.

написаний ФИО авторов в этом массиве охватывает несколько тысяч вариантов написаний ФИО авторов, а общее число уникальных авторов насчитывает 6079 персон, из которых 4487 – отечественные ученые.

ПРИЧИНЫ, ВЫЗЫВАЮЩИЕ РАЗЛИЧИЯ В НАПИСАНИЯХ ФИО ОДНОГО И ТОГО ЖЕ АВТОРА

В анализируемом массиве авторов российских публикаций могут присутствовать ФИО и российских, и зарубежных авторов. Предварительный анализ написаний ФИО авторов позволил высказать два предположения.

Предположение 1. Различие в написаниях ФИО российских авторов прежде всего вызывается разницей в транслитерации латинскими символами одного и того же кириллического символа[4].

Для того чтобы проверить это предположение, исследуемый массив был просмотрен на предмет выделения тех латинских символов и их сочетаний, которые используются в публикациях из рассматриваемого массива при транслитерации ФИО российских авторов. Затем полученные данные были сопоставлены с транслитерацией букв русского языка, предусмотренной в ГОСТ 7.79-2000 [7].

Из этого сопоставления следует, что только для 15 букв русского алфавита транслитерация, использованная в массиве ФИО авторов, полностью совпадает с транслитерацией, предусмотренной в ГОСТ 7.79-2000**;** для шести букв русского алфавита при транслитерации в массиве ФИО авторов используются три варианта, тогда как ГОСТ предусматривает только один вариант; еще для пяти букв в исследуемом массиве ФИО при транслитерации используются два варианта; по одному случаю для кириллического символа при транслитерации в массиве ФИО – шесть вариантов и пять вариантов соответственно; еще один случай, когда ГОСТ предусматривает два варианта, а в массиве ФИО встречается три варианта; и один случай, когда и в ГОСТ, и в массиве ФИО по одному варианту, однако эти варианты не совпадают; два случая, когда в массиве ФИО используются двухбуквенные сочетания, тогда как ГОСТ этого не предусматривает вовсе, причем, для одного из этих двухбуквенных сочетаний в массиве ФИО для транслитерации применяется два варианта, а для другого еще больше – четыре.

Понятно, чем больше вариантов, которые используются для транслитерации одного и того же символа, тем больше вероятность того, что ФИО одного и того же ученого в различных публикациях, автором которых является этот ученый, будет написано поразному. Это, естественно, внесет искажения в значения самых различных библиометрических и наукометрических показателей. При этом, чем больше таких ученых представлено в массиве авторов и чем больше публикаций у этих ученых, тем большее абсолютное и относительное искажение будет внесено в значения этих показателей, в частности: в число публикаций конкретного ученого; в число активных ученых (в данном направлении исследований, данной области науки, в данном городе, субъекте РФ, в стране в целом); в гендерную статистику науки; в определение тенденций этих показателей во времени. Только тщательная идентификация написаний ФИО авторов дала возможность избежать этих искажений, по крайней мере, существенно уменьшить их степень и влияние. При идентификации зарубежных ученых – авторов российских публикаций полезными оказались справочник [8] и статья [9].

Обработка исследуемого массива показала, что различия в написаниях одного и того же ФИО вызывают также орфографические ошибки, когда в статье ФИО автора указано неверно, что, правда, встречается очень редко. Примеры обнаруженных в исследуемом массиве вариантов написаний имен авторов приведены в табл. 1.

Таким образом, полученные нами данные подтверждают правильность предположения о существенном влиянии различных вариантов транслитерации, используемых в исследуемом массиве.

Предположение 2. Написание ФИО автора зависит и от требований к этому написанию, принятых в редакции того или иного журнала. Если это предположение справедливо, то ФИО одного и того же автора в случае, если его статьи опубликованы в разных журналах, может быть написано по-разному.

Для того чтобы проверить справедливость этого предположения, мы провели мини-исследование, которое состояло в следующем. Из массива российских авторов было выбрано несколько ФИО, соответствующих известным отечественным ученым. Затем были определены журналы, из представленных в *WoS*, в которых неоднократно публиковались эти ученые. После этого были найдены библиографические описания статей этих ученых, которые, естественно, содержат и ФИО авторов. Приведем результаты этого исследования.

В ряде журналов (*Applied Organometallic Chemistry, Chemistry-A Eropean Journal, Coordination Chemistry Reviews, Dalton Transactions, European Journal of Inorganic Chemistry, Heteroatom Chemistry, Inorganic Chemistry, Inorganic Chemistry Communications, Inorganica Chimica Acta, Mendeleev Communications*) известный химик академик РАН Глеб Арсентьевич Абакумов как автор статей, опубликованных в этих журналах, представлен следующим образом: *Abakumov, Gleb A.* В другой группе журналов (*Doklady Chemistry, High Energy Chemistry, Russian Chemical Bulletin, Russian Journal of Coordination Chemistry, Russian Journal of General Chemistry, Russian Journal of Physical Chemistry A, Russian Journal of Physical Chemistry B*) этот автор представлен как *Abakumov, G. A.*

[4] Важно учитывать, что здесь речь идет, прежде всего, о российских авторах, поскольку их ФИО в «естественных» условиях написаны на кириллице. Но это касается также и тех зарубежных авторов, родной язык которых использует тот или иной вариант кириллицы. Их проблемы в настоящей работе не обсуждаются, так как, с одной стороны, они требуют специального рассмотрения, а с другой – доля таких авторов в массиве авторов российских публикаций невелика.

Таблица 1

Примеры транслитерации имен, встречающиеся в исследуемом массиве авторов*

Имя на русском языке	Транслитерированные варианты имени
Алексей	Alexey (10), Alexei (129), Aleksey (32), Aleksei (3)
Андрей	Andrey (240), Andrei (208), Andre (18)
Дмитрий	Dmitriy (4), Dmitrii (5), Dmitri (124), Dmitry (186)
Григорий	Grigory *(28)*, Gregory *(17)*, Grigorii *(1)*
Людмила	Lyudmila (22), Ludmila (2), Liudmila (2)
Михаил	Michael (66), Mikhail (184), Michail (6)
Наталья	Natalia *(83)*, Natalya *(5)*, Nataliya *(6)*
Сергей	Sergey *(418)*, Sergei *(114)*
Юрий	Yuriy *(33)*, Yurii *(106)*, Yuri *(286)* Yury *(36)*, *Juri (17)*

* В скобках указано число употреблений данного варианта

Другой пример: российский математик Александр Васильевич Абанин в ряде журналов (*Izvestiya Mathematics, Journal of Approximation Theory, Proceedings of the American Mathematical Society, Siberian Mathematical Journal*) представлен как *Abanin, A. V.*, тогда как в других журналах (*Journal of Mathematical Analysis and Applications, Studia Mathematica*) – как *Abanin, Alexander V.*

Еще два примера. Известный специалист в области фотохимии академик РАН Михаил Владимирович Алфимов в качестве автора публикаций в ряде журналов (*Heteroatom Chemistry, New Journal of Chemistry, Organometallic Chemistry, Physical Chemistry Chemical Physics*) дается как *Alfimov Michael V.*, тогда как в других журналах (*High Energy Chemistry, Doklady Biochemistry and Biophysics, Journal of Organic Chemistry, Russian Chemical Bulletin, Russian Journal of Applied Chemistry, Theoretical and Experimental Chemistry, Uspekhi Khimii*) – как *Alfimov M. V.* Специалист в области оптической физики доктор физ.-мат. наук Александр Васильевич Баранов как автор в одном случае представлен следующим образом*: Baranov, A. V. (Journal of Applied Physics, Journal of Raman Spectroscopy, Journal of Optical Technology, Optics and Spectroscopy*), а в другом – как *Baranov, Alexander V.* (*Beilstein Journal of Nanotechnology, Journal of Physical Chemistry-C, Nanomedicine, Optics Letters*).

Иногда один и тот же автор в одном и том же журнале может быть представлен и первым, и вторым вариантами: А) с ФИО, содержащим помимо фамилии, только инициалы имени и отчества; Б) с ФИО, содержащим помимо фамилии также и полное имя автора. В частности, это оказалось справедливо для двух приведенных выше авторов: Александра Васильевича Абанина – в журнале *Comptes Rendus Mathématique*, и Александра Васильевича Баранова - в журнале *Nanotechnology*.

Таким образом, подтверждается предположение о том, что различия в написании ФИО одного и того же автора могут быть вызваны различающимися требованиями к рукописям в редакциях различных журналов. Помимо чистой констатации, это эмпирически установленное обстоятельство было использовано при идентификации авторов. Например, встречаются два варианта написания ФИО, и есть основания предполагать, что речь идет об одном и том же человеке (совпадающие организации – места работы, очень близкая тематика публикаций). В этом случае особенности «журнальных» сведений об авторе могут представлять собой важную дополнительную информацию при отнесении этих сведений к одному и тому же или, напротив, к различным ученым.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

Выделение групп авторов

В основу идентификации авторов были положены данные триады «ФИО – публикация – аффилиация». При этом в подавляющем большинстве случаев (более 90%) данные об аффилиации были ранее подвергнуты идентификации, что позволило их уточнить, четко выделить страну, город и организацию, в которой работает автор. В ряде случаев (порядка 20%) при идентификации авторов пришлось обращаться за дополнительными данными в Интернет. При идентификации использовались особенности структуры написания ФИО. Из табл. 2, в которой эти особенности представлены, следует, что доля написаний ФИО, содержащих имя (first name) автора (строки 2, 3 и 5), составляет 23,61%. Это обстоятельство активно использовалось при непосредственной идентификации, когда приходилось объединять написания ФИО автора, которые содержали фамилию и *инициалы* (один, два, и иногда и три), с написаниями ФИО, которые содержали фамилию и *полное написание имени*.

Большинство выявленных особенностей написаний ФИО и влияние на эти написания рассмотренных выше факторов были использованы при идентификации авторов, при отнесении автора к той или иной группе ученых, при определении гендерной структуры публикаций (табл. 3) и «превращении» авторов в персон.

Таблица 2

Распределение различных написаний фамилии, имени и отчества авторов

№ п/п	Написание ФИО авторов	Число написаний	Доля написаний, %
1	Фамилия и два инициала	23986	63,22
2	Фамилия, имя и второй инициал	8055	21,23
3	Фамилия и имя	3214	8,47
4	Фамилия и первый инициал	2497	6,58
5	Фамилия, имя и отчество	173	0,46
6	Фамилия и три инициала	17	0,04
7	Итого	37942	100,00

Не менее активно использовались особенности фамилии и/или имени при определении «этнической принадлежности» автора для последующего отнесения автора к одной из социально-этнических групп, формирование которых основывается на одновременном учете «этнической» и «государственной» (страновой) принадлежности автора:

a) российские авторы, т.е. ученые, которые в качестве аффилиации указали только российские организации;

b) экспаты – граждане России, т.е. ученые, которые в качестве аффилиации указали и российские, и зарубежные организации;

c) экспаты – выходцы из России или СССР (ученые российской диаспоры): авторы с типично российской (советской) фамилией, которые в качестве своей аффилиации указали только зарубежные организации;

d) иностранные авторы из стран – бывших республик СССР (ученые ближнего зарубежья);

e) иностранные авторы из остальных стран мира (ученые дальнего зарубежья).

Принципы формирования этих групп авторов детально описаны в работе [10]. Здесь же отметим только следующее. К группе «экспаты – выходцы из России или СССР» относились, по преимуществу, ученые, которые не работают в России, но носят «типично российские» фамилию и имя. В группу «ученые ближнего зарубежья» входят авторы, работающие в одной из стран – бывших республик СССР и носящие, по преимуществу, типичные для данной страны фамилии и имена. Однако если такой автор с такими ФИО в качестве аффилиации указал только российские организации, то он включался в группу «российские авторы». Это, конечно, при условии если не обнаружены некоторые дополнительные данные, указывающие на то, что этого ученого следует все же отнести в группу «иностранные авторы из ближнего зарубежья».

Очень важными оказались особенности фамилии и имени автора и при принятии решения о том, к какой гендерной группе (мужчины или женщины) следует отнести данного автора. С этой целью учитывалось окончание русских фамилий (например, окончание «ov» – мужчина, окончание «ova» – женщина). Кроме того, принималось во ванимание, носит ли данный автор мужское или женское имя. Для зарубежных ФИО в большинстве случаев для отнесения автора к той или иной гендерной группе, приходилось обращаться к дополнительным данным в Интернете.

Оценка эффективности идентификации

Под оценкой эффективности идентификации авторов будем понимать отношение числа уникальных (неповторяющихся) ФИО авторов анализируемого массива публикаций, выявленного непосредственно с помощью интерфейса *WoS* и его опции *Results Analysis*, к аналогичной характеристике этого же массива, которая была получена в результате идентификации. Это отношение назовем коэффициентом идентификации k_{ident}.

$$k_{ident} = \frac{n_{WoS}}{n_{ident}} \quad ,$$

где:

n_{WoS} – число уникальных ФИО авторов, выявленное с помощью интерфейса *WoS*;

n_{ident} – число уникальных авторов (персон), полученное в результате идентификации.

Следует отметить, что непосредственно осуществить оценку всего анализируемого массива написаний ФИО авторов очень сложно из-за способа его формирования, поскольку этот массив был создан случайным (точнее, квазислучайным) образом. Анализ этого массива показал, что ему соответствуют 15456 публикаций, т.е. помимо 2319 публикаций по *N&N* он включал также публикации, которые соответствовали самым различным категориям *WoS*. Учитывая способ организации этого массива, далеко не все попавшие в него публикации, могут быть представлены полным списком авторов. Действительно, если среди авторов некоторой публикации из этого массива есть автор, ФИО которого отсутствует в исходном массиве уникальных написаний ФИО по категории *N&N*, то такой автор в общий массив авторо-указаний не попадет. Следовательно, «лобовое», непосредственное сопоставление результатов идентификации авторов указанных 15456 публикаций с аналогичным массивом авторов, который может быть получен с помощью интерфейса *WoS* и его опции *Results Analysis*, было бы некорректным. Более того, такое сопоставление представляется практически нереализуемым, поскольку потребовалось бы каждую из указанных публикаций искать в *WoS*, т.е. произвести 15456 поисков, а затем из результата каждого такого поиска еще и отбирать те автороуказания, которые соответствуют автороуказаниям из анализируемого массива.

Таблица 3

Распределение авторов, публикаций и автороуказаний по социально-этническим и гендерным группам

№ п/п	Социально-этнические группы авторов (персон)	Уникальные авторы (персоны)		Публикации		Автороуказания		Гендерные группы					
								Мужчины		Женщины		Группа не установлена	
		Число уникальных авторов (персон) в группе	Доля уникальных авторов в группе от общего числа уникальных авторов, %	Число публикаций в группе	Доля публикаций в группе от общего числа публикаций, %	Число автороуказаний группы	Доля автороуказаний группы от общего числа автороуказаний, %	Доля уникальных авторов (персон) в группе (%)	Доля автороуказаний в группе, %	Доля уникальных авторов (персон) в группе, %	Доля автороуказаний в группе, %	Доля уникальных авторов (персон) в группе, %	Доля автороуказаний в группе, %
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1	Российские авторы	3989	65,6	13624	88,1	27398	72,2	73,0	81,0	23,8	18,0	3,1	1,0
2	Экспаты – граждане России	196	3,2	2291	14,8	4241	11,2	82,7	92,1	16,8	7,9	0,5	0,1
3	Экспаты – выходцы из России / СССР	302	5,0	915	5,9	1277	3,4	76,5	81,5	19,9	17,2	3,6	1,3
4	Иностранные авторы (ближнее зарубежье)	133	2,2	190	1,2	359	0,9	66,2	79,4	14,3	11,7	19,5	8,9
5	Иностранные авторы (дальнее зарубежье)	1407	23,2	2025	13,1	4425	11,6	61,0	76,9	10,6	8,3	28,4	14,8
6	Иностр. авторы из ближнего зарубежья, работающие в РФ	14	0,2	41	0,3	70	0,2	28,6	55,7	7,1	25,7	64,3	18,6
7	Иностр. авторы из дальнего зарубежья, работающие в РФ	38	0,6	146	0,9	172	0,5	63,2	86,6	10,5	4,1	26,3	9,3
8	Всего в исследуемом массиве	6079	100,0	15456		37942	100,0						

Учитывая эти трудности, оценка эффективности идентификации была выполнена только для той части рассматриваемого массива автороуказаний, которая соответствует 2319 российским публикациям тематической категории *N&N* за период 2009-2013 гг. Выбор этой части массива автороуказаний объясняется следующим. Во-первых, все автороуказания, которые соответствуют публикациям по *N&N*, прошли процедуру идентификации. Во-вторых, в отличие от ситуации с публикациями, соответствующими общему массиву автороуказаний, каждая *N&N*-публикация представлена полным списком авторов. В-третьих, для выделения аналогичного массива с помощью интерфейса *WoS* потребовалось осуществить не 15456 поисков, а только один, задав в режиме *Advanced Search* запрос, в качестве одного из условий которого была указана тематика *Nanoscience & Nanotechnology*. В результате этого поиска для 2319 российских публикаций по *N&N* за 2009-2013 гг. было получено 7480 уникальных (согласно *WoS*) авторов, которым соответствовало 11586 автороуказаний. Если же воспользоваться результатами идентификации, то для этого же массива публикаций и примерно такого же числа автороуказаний мы получим 5946 уникальных авторов (персон)[5]. Таким образом, коэффициент эффективности идентификации k_{ident}:

$$k_{ident} = \frac{n_{WoS}}{n_{ident}} = \frac{7480}{5946} = 1{,}36 \ .$$

Это значит, что эффективность выполненной идентификации авторов для массива российских публикаций по *N&N* достаточно высока: удалось сократить список авторов на 1534 фамилии (которые, по сути являлись «виртуальными»), т.е. на 25% от списка, полученного с помощью интерфейса *WoS*. В действительности, в результате идентификации произошло следующее: 1) выполнено объединение *различающихся* написаний ФИО одного и того же ученого; 2) осуществлено объединение *одинаковых* написаний ФИО одного и того же ученого, который указал несколько *различных аффилиаций*. В результате, все те публикации, которые в *WoS* «привязаны» к этим виртуальным авторам, были приписаны к соответствующим реальным ученым, которые действительно являются авторами этих публикаций. И конечно, в ходе идентификации были разделены совпадающие написания ФИО, если они соответствовали различным ученым.

[5] На первый взгляд может показаться, что имеет смысл сопоставить числа 6726 и 5946 и именно таким образом определить эффективность идентификации. Однако такое сопоставление было бы некорректным, поскольку первое число соответствует числу вариантов написания ФИО из массива авторов по тематике *N&N*, тогда как второе – числу реальных ученых, являющихся авторами публикаций по этой тематике. Еще раз напомним, что среди множества уникальных написаний ФИО есть написания, каждое из которых может принадлежать разным авторам (разным персонам). С другой стороны, в этом множестве присутствуют различающиеся написания, которые принадлежат одному и тому же автору. Собственно задача идентификации и состояла в совместном решении этих двух проблем.

Анализ результатов идентификации авторов всего массива автороуказаний

Данные анализа 37942 автороуказаний позволили получить ряд важных результатов (см. табл. 3). Так, в ходе идентификации удалось отнести того или иного автора к соответствующей *социально-этнической группе*. К сожалению, для отнесения ряда авторов к соответствующей *гендерной группе* наших данных оказалось недостаточно: для 582 ученых (9,6%) гендерную группу установить не удалось (правда, вклад таких ученых в количество автороуказаний невелик – 1,9%); в табл. 3 доля таких «неустановленных» авторов зависит от социально-этнической группы. Для российских авторов, а также для экспатов – граждан России и экспатов – выходцев из России / СССР эта доля (по отношению к объему группы) невелика: 1,0%, 0,1% и 1,3% соответственно, тогда как для двух групп иностранных авторов эта доля значительна (8,9% и 14,8% для авторов из ближнего и дальнего зарубежья соответственно).

Трем группам авторов, которые составляют российские ученые, работающие только в нашей стране (65,6%), экспаты – граждане России (3,2%) и экспаты – выходцы из нашей страны (5,0%), соответствует 86,8% автороуказаний: 72,2%, 11,2% и 3,4% соответственно. Обнаруживается слабый «приток мозгов» в Россию: 48 иностранных ученых указали в качестве одной из своих аффилиаций российскую организацию. Этот процесс выявляет тенденцию к некоторому росту (в 2009 г. было 18 автороуказаний этих ученых, а в 2013 г. – 78), хотя, учитывая столь малые численные показатели, делать на их основании какие-либо выводы преждевременно.

Гендерное распределение в рассматриваемых социально-этнических группах выглядит следующим образом: максимальная доля женщин по отношению к общему числу персон соответствует группе «российские авторы» (23,8%), затем следуют экспаты – выходцы из России и/или бывшего СССР) (19,9%), экспаты – граждане России (16,8%), иностранцы из ближнего зарубежья (14,3%), иностранцы из дальнего зарубежья (10,6%). В целом по всему массиву этот показатель составляет 20,0%. Доля женщин в автороуказаниях ниже и составляет 15,6%, в социально-этнических группах: российские авторы – 18,0%; экспаты – выходцы из России и/или бывшего СССР – 17,2%; иностранцы (ближнее зарубежье) – 11,7%; иностранцы (дальнее зарубежье) – 7,9%; экспаты – граждане России – 9,6%.

Публикационная активность авторов конкретной социально-этнической группы – это отношение доли автороуказаний этой группы к доле уникальных авторов (персон) в группе. Если обратиться к данным табл. 3 (отношение значений в графе 8 к значениям в графе 4), то максимальная публикационная активность соответствует группе «экспаты – граждане России» (3,5), затем следуют российские авторы (1,1), экспаты – выходцы из России / СССР (0,68), иностранные авторы из дальнего зарубежья (0,5) и иностранные аиторы из ближнего зарубежья (0,41).

Таким образом, самую высокую публикационную активность обнаруживают экспаты – граждане России: доля их автороуказаний в 3,5 раза больше, чем

их доля в общем числе персон (11,2% против 3,2%). Для экспатов – выходцев из России / СССР ситуация обратная: их доля в общем числе персон составляет 5,0%, тогда как в автороуказаниях их доля – 3,4%. Доля зарубежных авторов (из ближнего и дальнего зарубежья) в общем числе персон составляет 25,3%. Однако их доля в автороуказаниях в 1,7 раза меньше, чем их доля в общем числе персон.

Если же обратиться к данным табл. 4, которые отражают цитируемость публикаций, то порядок следования этих групп будет несколько иным (цифры в скобках соответствуют средней цитируемости за 2009-2013 гг.): экспаты – выходцы из России/СССР (20,6), экспаты – граждане России (15,3%), иностранные авторы из дальнего зарубежья (13,8%), российские авторы (5,8%), иностранные авторы из ближнего зарубежья (5,3%). Таким образом, экспаты – выходцы из России/СССР, несмотря на относительно низкую публикационную активность и небольшую долю в общем числе ученых – авторов российских публикаций (5%), являются важным фактором, способствующим повышению качества отечественных исследований, увеличению востребованности и «видимости» (visibility) российской науки. Табл. 4, как и табл. 3 свидетельствует о важной роли группы «экспаты – граждане России» и группы ученых из дальнего зарубежья. К сожалению, вклад ученых из стран – бывших республик СССР в российские исследования очень невелик, а научный уровень исследований с их участием – невысок.

Данные, полученные в результате идентификации авторов, позволяют получить еще ряд важных и интересных библиометрических характеристик в зависимости от группы (социально-этнической, гендерной) авторов. К ним относятся: доля публикаций, поддержанных научными фондами; средняя цитируемость и средневзвешенный импакт-фактор поддержанных этими фондами публикаций; среднее число авторов в расчете на одну поддержанную публикацию. Возможности анализа связей, которые проявляются в результате идентификации авторов и других элементов библиографического описания можно проследить на рисунке.

В верхнем левом углу публикации; в верхнем правом углу – ученые – авторы этих публикаций; ниже, правее авторов – организации, в которых работают авторы; в нижнем правом углу – города, в которых расположены организации; внизу посередине – тематические категории, которые соответствуют показанным публикациям; в левом нижнем углу – научные фонды, поддержавшие эти публикации. Линии указывают на связи между соответствующими объектами, а толщина этих линий – на силу этих связей. Чтобы не перегружать рисунок, на нем не отражены журналы, а также поддержанные фондами исследовательские проекты (гранты).

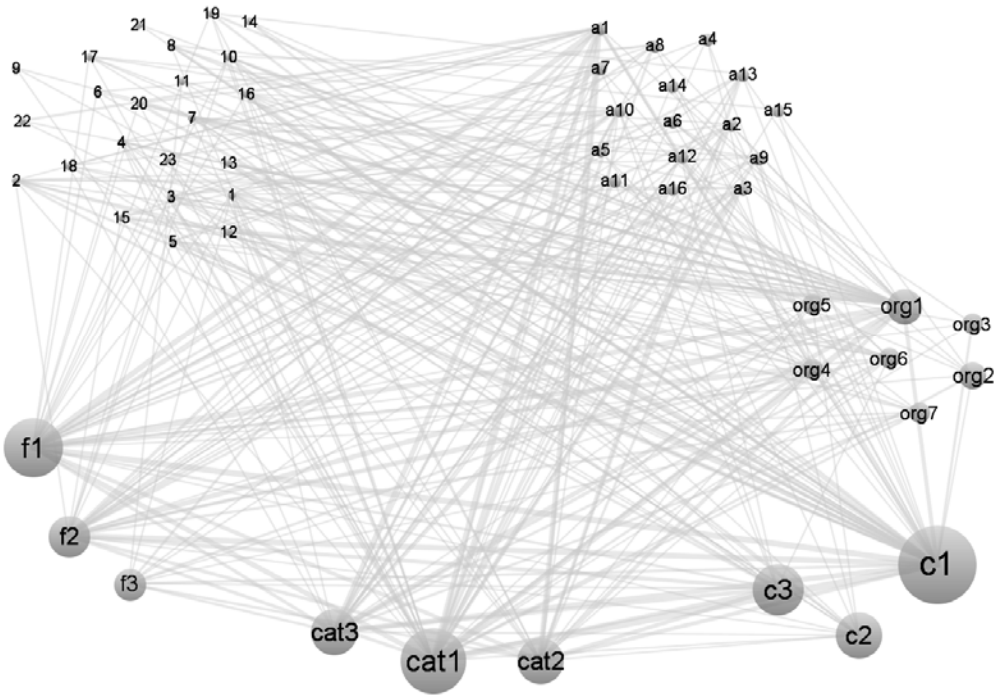
Учитывая особенности анализируемого массива, этот анализ в настоящей работе представляется нецелесообразным. Логика библиометрического исследования основывается на явном или неявном предположении, согласно которому объектом такого исследования должна выступать некая целостная, логически замкнутая тематика [11, 12], например, определенный научный раздел. В нашем случае, напротив, распределение публикаций из анализируемого массива по 156 тематическим категориям WoS оказалось очень неравномерным: каждой из первых 10 групп объектов соответствовало более одной тысячи публикаций, тогда как каждой из 20 последних групп – по одной публикации. Значительная часть публикаций, попавших в массив автороуказаний, представлена в нем не полным списком авторов. Эти обстоятельства не только затруднили бы интерпретацию показателей, полученных на основе такого массива, но и могли бы внести в значения этих показателей существенные искажения.

Таблица 4

Распределение среднего числа ссылок 2009–2013 гг. по социально-этническим группам (на начало января 2016 г.)[6]

№ п/п	Социально этнические группы авторов (персон)	Среднее число ссылок в расчете на публикацию					
		Год опубликования					В среднем в публикациях 2009-2013 гг.
		2009	2010	2011	2012	2013	
1	Российские авторы, работающие только в России	7,5	7,1	5,9	4,9	3,6	5,8
2	Экспаты – граждане России	24,0	16,3	15,3	12,4	8,9	15,3
3	Экспаты – выходцы из России / СССР	31,2	20,7	18,8	14,6	11,2	20,6
4	Иностранные авторы (ближнее зарубежье)	5,7	8,8	3,7	5,3	3,3	5,5
5	Иностранные авторы (дальнее зарубежье)	18,1	16,6	13,6	10,8	7,8	13,8
6	Итого	9,8	8,3	6,9	5,9	4,3	7,0

[6] В качестве публикаций, соответствующих социально-этнической группе, признаются российские публикации, автором каждой из которых является хотя бы один ученый данной группы. Это значит, что авторами такой публикации могут быть и ученые из других групп. Как следствие, одна и та же публикация может быть отнесена к нескольким группам.



Связи, проявившиеся при идентификации авторов и других элементов библиографического описания

Тем не менее, идентификация авторов публикаций, соответствующих тому или иному направлению исследований, открывает широкие возможности для углубленного наукометрического анализа.

ВЫВОДЫ

Проведенное наукометрическое исследование позволило на основании данных *Web of Science* уточнить реальный вклад в мировую науку российских публикаций, которые признаются таковыми, если хотя бы один из их авторов является гражданином России, выходцем из России или СССР, проживающим за рубежом, даже если все остальные авторы – иностранцы. Было показано, что, несмотря на то, что *WoS* одна из лучших мировых информационных систем, ее данные неточны, поскольку ее методы идентификации авторов несовершенны.

В ходе исследования разработаны более совершенные методы идентификации авторов научных публикаций, позволяющие вычленить из указанных системой авторов реальных ученых, принимавших участие в написании этих публикаций и отнести к ним эти публикации. Выяснилось, что отношение указываемых системой авторов к реальным ученым (эффективность идентификации по применявшейся методике) составляет 1,36.

Определены социально-этнические группы ученых, которые формируются на основе совместного учета данных об «этническом» происхождении ученого и сведений о его аффилиации, которые он указывает в своих публикациях. Установлен вклад в российские исследования каждой из этих групп, что позволяет осмыслить правильность политики в отношении публикационной активности авторов российских публикаций. Абсолютное большинство (более 90%) авторов удалось отнести к соответствующей гендерной группе. Определена доля женщин в каждой из указанных социально-этнических групп и их доля в публикациях этих групп.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Reijnhoudt L., Costas R., Noyons E., Borner K., Scharnhorst A. ‘Seed + expand’: a general methodology for detecting publication oeuvres of individual researchers // Scientometrics. – 2014. – Vol. 101, № 2. – P. 1403-1417. – DOI: 10.1007/s11192-015-1699-y
2. Ferilli S. A sentence structure-based approach to unsupervised author identification // Journal of Intelligent Information Systems. – 2016. – Vol. 46, № 1. – P. 1-19.
3. Wu H., Li B., Pei YJ; H.J. Unsupervised author disambiguation using Dempster-Shafer theory // Scientometrics. – 2014. – Vol. 101, № 3. – P. 1955-1972. – DOI: 10.1007/s11192-014-1283-x
4. Shin D., Kim T., Choi J., Kim J. Author name disambiguation using a graph model with node splitting and merging based on bibliographic information Scientometrics. – 2014. – Vol. 100, № 1. – P. 15-50. – DOI: 10.1007/s11192-014-1289-4

5. Harzing AW. Health warning: might contain multiple personalities-the problem of homonyms in Thomson Reuters Essential Science Indicators// Scientometrics. – 2015. – Vol. 105, № 3. – P. 2259-2270. – DOI: 10.1007/s11192-015-1699-y
6. Петровский А.Б. Пространства множеств и мультимножеств. – М.: УРСС, 2003. – 248 с.
7. ГОСТ 7.79 – 2000 (ИСО 9 - 95). Система стандартов по информации, библиотечному и издательскому делу. Правила транслитерации кирилловского письма латинским алфавитом. – Минск: Межгосударственный совет по стандартизации, метрологии и сертификации, 2002. – 19 с.
8. Гиляревский Р.С., Старостин Б.А. Иностранные имена и названия в русском тексте: Справочник. – М.: Высшая школа, 1985. – 303 с.
9. Хачко О.А., Глобачева Э.Я. Индексирование имен авторов испано- и португалоязычных стран в зарубежных журналах, а также в РЖ и БД ВИНИТИ РАН // Научно-техническая информация. Сер. 2. – 2016. – № 9. – С. 24-31.
10. Либкинд А.Н., Маркусова В.А., Терехов А.И., Рубвальтер Д.А., Либкинд И.А. Результаты выполнения конкурсных исследовательских проектов: библиометрия вклада различных групп ученых, организаций, городов, регионов и стран // Научно-техническая информация. Сер. 1. – 2015. – № 11 – С. 16-28.
11. Арапов М.В., Либкинд А.Н. Научные документы в зеркале информатики // Вопросы информационной теории и практики. – 1982. – № 47. – С. 47-81.
12. Libkind A.N. One approach to study communication in science // Scientometrics. – 1985. – № 3-4. – P. 217-234.

Материал поступил в редакцию 16.02.17.

Сведения об авторах

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

МАРКУСОВА Валентина Александровна – доктор педагогических наук, зав. Отделением ВИНИТИ РАН
e-mail: valentina.markusova@gmail.com

ЛИБКИНД Илья Александрович – системный аналитик, Сервисное бюро «ВИП»
e-mail: libkind_ilya@hotmail.com

КАМЕНЬ Наталья Марковна – научный сотрудник ВИНИТИ РАН
e-mail: natakhak@mail.ru

ФАДЕЕВ Владислав Юрьевич – научный сотрудник ВИНИТИ РАН
e-mail: favlad@hotmail.com

ИНФОРМАЦИОННОЕ ПИСЬМО И ПРИГЛАШЕНИЕ

МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ К 65-ЛЕТИЮ ВИНИТИ РАН

«ИНФОРМАЦИЯ В СОВРЕМЕННОМ МИРЕ»

Москва, 25-26 октября 2017 г.

подробная информация на сайте: http://www.viniti.ru

Главный организатор:

Всероссийский институт научной и технической информации Российской академии наук (ВИНИТИ РАН)

Соорганизаторы:

Российская академия наук
Федеральное агентство научных организаций
Российский фонд фундаментальных исследований
Министерство образования и науки РФ

Проблемно-тематическое направление конференции: современный издательский процесс, интеллектуальная собственность, научные библиотеки, информационное обеспечение научной и инновационной деятельности, информационные технологии для научной и библиотечной отрасли, информационная безопасность, международное сотрудничество и информационный обмен, инфометрия, классификации, стандартизация, образование для отрасли, экономика информации

Основные вопросы, предлагаемые к обсуждению:

● Популяризация научных знаний: Новые модели распространения научной информации
● Редакционно-издательская деятельность в цифровой среде: продукты и сервисы
● Издательские стандарты и технологии
● Перспективы развития книжного дела. Проекты и программы
● Взаимодействие цифровых и печатных ресурсов в научно-технической библиотеке
● Информационно-библиотечное обслуживание: сервисный подход
● Управление данными и навигация в современной научной библиотеке
● Научные библиотечные консорциумы – основные подписчики на научную литературу
● Перспективы развития национальных систем научно-технической информации
● Государственные проекты и программы поддержки информационного обеспечения научно-образовательной деятельности
● Тенденции развития региональных аналитических центров
● Информационное обеспечение экспертной деятельности. Использование информационно-аналитических систем для управления наукой и образованием
● Формальные и неформальные каналы развития современных научных коммуникаций

● Современные агрегаторы научной литературы открытого доступа как источник научной информации
● Машинная обработка данных и аналитические исследования: Приоритеты и сотрудничество
● Использование специальных сервисов компании CrossRef для идентификации научных публикаций
● Роль поисковых систем в современном издательском процессе
● Защита данных от несанкционированного использования. Маркеры безопасности. Политика безопасности открытых систем
● Вопросы достоверности и доверенности при обработке информационного потока
● Межгосударственный обмен научно-технической информацией на евразийском пространстве
● Информационное взаимодействие в рамках СНГ
● Международное партнерство при хранении и обработке больших массивов данных
● Современное состояние систем классификации знаний как инструмента индексирования и поиска данных по перспективным направлениям науки и критическим технологиям
● Современные библиометрические методы определения научных лидеров: Новые математические модели
● Анализ читательской аудитории научной литературы путем вебометрического анализа
● Подготовка специалистов в сфере научно-информационной деятельности
● Мастер-класс по работе с классификационными системами (УДК, ГРНТИ)
● Информация как источник цифрового капитала и фактор социальных изменений
● Информационная деятельность как фактор национальной экономики
● Новейшие бизнес-модели для публикаций открытого и закрытого доступа

На конференции планируются доклады представителей ведущих информационных центров и научно-технических библиотек России, СНГ и дальнего зарубежья.

В рамках юбилейной конференции состоится научно-практический семинар по классификационным системам «Перспективные направления научных исследований и критические технологии в классификационных системах». Предполагается проведение специализированных обучающих мероприятий по УДК индексированию. Запланировано заседание методического совета пользователей ГРНТИ и УДК. Участники конференции получат свидетельства о повышении квалификации.

Материалы конференции будут опубликованы в сборнике Трудов и на CD-ROM, основные – в сборнике **«Научно-техническая информация»**.

Доклады

Принимаются оригинальные работы, имеющие научное и прикладное значение, соответствующие тематическим направлениям конференции и НЕ ОПУБЛИКОВАННЫЕ ГДЕ-ЛИБО РАНЕЕ.

Предлагаемый доклад должен отвечать следующим требованиям:

1. Необходимо указать название доклада, фамилию, имя, отчество (полностью) авторов/соавторов, название организации, город, страну, выделить автора, который будет представлять доклад.

2. Необходимо наличие аннотации, раскрывающей содержание доклада. Размер аннотации - не более 850 знаков (включая пробелы).

3. Доклады принимаются только в электронной форме; тексты – в формате MS Word; схемы, диаграммы, фотографии, сканированные виды экранов и т. п. - в формате JPG. Объем доклада вместе с аннотацией, рисунками, приложениями и т.п. не более 10 страниц формата А4.

4. Доклад необходимо выслать по электронной почте до 11 сентября 2017 г. в адрес оргкомитета: conf@viniti.ru

Доклады, не соответствующие вышеуказанным требованиям,
НЕ РАССМАТРИВАЮТСЯ.

Программный комитет оставляет за собой право определять статус доклада (пленарный доклад, доклад, стендовый доклад), включать принятые доклады в те или иные секции.

Время для выступления: пленарные доклады – 15–20 мин., доклады на отдельных мероприятиях – до 10 мин. Доклады включаются в Труды на основании решения экспертов оргкомитета.

Контакты: 125190, Москва, ул. Усиевича, 20, ВИНИТИ РАН
Телефоны: 8 (499) 152 61 13, 8 (499) 155 42 52, 8 (499) 151 02 61. Факс 8 (499) 943 00 60
Интернет-сайт: http://www.viniti.ru Эл. почта: conf@viniti.ru

База данных (БД) ВИНИТИ РАН

Федеральная база отечественных и зарубежных публикаций по естественным, точным и техническим наукам, генерируется с 1981 г., обновляется ежемесячно, пополнение составляет около 1 млн документов в год. Тематическое наполнение соответствует реферативному журналу ВИНИТИ. Для поиска одновременно по всем или нескольким тематическим фрагментам генерируется единая Политематическая БД.

БД ВИНИТИ РАН в сети INTERNET

Сервер ВИНИТИ - http://www.viniti.ru – обеспечивает on-line доступ к Базе данных ВИНИТИ РАН круглосуточно без выходных.

На основе БД ВИНИТИ РАН предоставляются следующие услуги:

- Диалоговый поиск научно-технической информации **в режиме on-line**;
- **Демо-версия**, позволяющая ознакомиться с основными функциями поисковой системы, составом данных, формами представления документов и получить навыки работы с системой;
- **Поисковые эксперты** ВИНИТИ выполнят тематический поиск по разовым или постоянным запросам, а также окажут **консультационные услуги**.

БД ВИНИТИ РАН на CD-ROM

Любые наборы тематических фрагментов БД ВИНИТИ или их разделов за любой период с 1981 г., а также **проблемно-ориентированные выборки** из БД ВИНИТИ по актуальным направлениям научных исследований могут быть предоставлены на договорной основе **в поисковой системе (ИПС) "Сокол"**, работающей под управлением Microsoft Windows и обеспечивающей следующие возможности:

- **Чтение** документов в режиме последовательного просмотра или выборочно по оглавлению за весь период заказанной ретроспективы
- **Поиск** документов по автору, заглавию, источнику, ключевым словам или словосочетаниям, реферату, рубрикам, году издания, стране, языку и т.д. (всего более 20 признаков)
- **Словарь** системы поможет правильно подобрать термины для поиска и выбрать глубину их усечения.
- **Для уточнения поиска** можно дополнительно использовать год издания документа, язык текста документа, рубрики, шифры тематических разделов БД.
- Выполненные **запросы можно сохранять** для их последующего использования и/или редактирования.

125190, г. Москва, ул. Усиевича, 20, БД ВИНИТИ РАН.
Отдел взаимодействия с потребителями – (499) 155-45-25, (499) 152-58-81
E-mail: csbd@viniti.ru, sales@viniti.ru

WWW: http://www.viniti.ru