

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

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

Издается с 1961 г.

№ 11

Москва 2019

ИНФОРМАЦИОННЫЕ СИСТЕМЫ

УДК 004.4:004.6

Д.Ю. Съедин

Программно-вычислительный комплекс для решения задачи интеграции атрибутивных данных информационных объектов

Представлен программно-вычислительный комплекс, обеспечивающий взаимодействие программных систем для решения задачи интеграции атрибутивных данных информационных объектов, в основе которого лежит разработанное и апробированное математическое и программное обеспечение. Этот комплекс может быть использован для повышения эффективности процессов нечеткого связывания данных при координировании сведений в информационных ресурсах, системах и базах данных.

Ключевые слова: связывание данных, интеграция данных, программно-вычислительный комплекс, программные инструменты взаимодействия программных систем, обработка информации

ВВЕДЕНИЕ

Стремительный рост объемов порождаемой информации требует использования современных технологий для эффективного хранения, контроля и управления данными. Значительное развитие получила концепция создания автоматизированных информационных ресурсов и систем, позволяющих

учитывать и агрегировать большие наборы данных по обширным сферам знания для обеспечения надежного и постоянного доступа к актуальным сведениям по исследуемой проблематике.

Вместе с тем сдерживающим фактором на пути создания таких информационных ресурсов является проблема качества данных. Разнородность агреги-

руемой из различных источников информации и ненормализованное её хранение вызывают большие сложности при создании банков знаний и информационных систем. Решение этой проблемы можно условно разделить на два этапа: первый – предполагает разработку структуры данных и их взаимосвязь для дальнейшего приведения разнородной информации к единому виду, второй – подразумевает непосредственное связывание сведений, относящихся к одним и тем же объектам. В результате обеспечивается согласование и консолидация различных представлений данных, а также сокращение дублирования информации.

ОСОБЕННОСТИ СОЗДАНИЯ ПРОГРАММНО-ВЫЧИСЛИТЕЛЬНОГО КОМПЛЕКСА

Реализация первого этапа – разработки структур данных и их взаимосвязи – предполагает анализ основных информационных объектов, учитываемых в разрабатываемых информационных ресурсах или системах, а также моделирование технологических процессов, отражающих жизненный цикл этих объектов и фиксирующих их состояние во времени. Для этого могут быть использованы получившие широкое распространение методологические инструменты [1], позволяющие отображать и анализировать модели деятельности широкого спектра сложных систем в различных разрезах. Затем, как правило, следует непосредственное проектирование структур данных в реляционном хранилище (базе) данных. При этом важно предусмотреть применение компонентов нормативно-справочной информации, являющихся основой для унификации и нормализации данных, сопровождающих и регламентирующих протекающие в системе бизнес-процессы. Использование последних упрощает задачу проведения мониторинга и аналитики накапливаемых и вновь поступающих в систему данных, снижает необходимость привлечения человеческих ресурсов для построения статистических отчетов по накапливаемому фонду.

Второй этап подразумевает создание и применение инструментов для интеграции и обеспечения структурного единства поступающих и хранящихся сведений и их согласованности с существующими компонентами нормативно-справочной информации. За счет этих инструментов должно обеспечиваться повышение эффективности процесса нечеткого связывания данных для структурирования атрибутивных данных информационных объектов и их дальнейшего координирования при максимальной достоверности получаемых результатов.

Под связыванием данных (*record linkage process*) [2], который является одним из важнейших шагов при решении задачи интеграции, понимается процесс сопоставления пар эквивалентных записей, различающихся по написанию. К математическому и программному обеспечению инструмента, способного повысить эффективность указанного процесса, предъявляются следующие требования:

- отсутствие необходимости ручного определения «важности» атрибутов, характеризующих записи;
- отсутствие требований указания правил сопоставления в явном виде;

- отсутствие принятия предположений о взаимной независимости атрибутов, которые часто не выполняются на практике;
- отсутствие требований по работе со специальными форматами машиночитаемой каталогизации;
- возможность влияния на «самостоятельность» решающего правила при связывании данных;
- поддержка параллельных вычислений для эффективного использования вычислительных ресурсов.

В качестве основы для создания инструмента, удовлетворяющего указанным требованиям, может использоваться, например, алгоритм связывания данных [3], который предназначен для связывания наборов данных, состоящих из текстовых записей, на основе информации, характеризующей атрибуты (поля) этих записей. В основе этого алгоритма лежит эвристическое решающее правило – так называемая адаптивная мера подобия записей:

$$\text{sim}(\text{record}_{i_1}, \text{record}_{i_2}) = \frac{\sum_{j=1}^F \partial_j \text{sim}(\text{field}_{i_1j}, \text{field}_{i_2j})}{\sum_{j=1}^F \partial_j} \quad (1),$$

где ∂_j – весовой коэффициент поля j ;

$$\begin{aligned} \text{sim}(\text{field}_{i_1j}, \text{field}_{i_2j}) = & \frac{1}{2n_{i_1j}} \sum_{k_1=1}^{n_{i_1j}} \max_{k_2} \text{sim}(\text{word}_{i_1jk_1}, \text{word}_{i_2jk_2}) + \\ & \frac{1}{2n_{i_2j}} \sum_{k_2=1}^{n_{i_2j}} \max_{k_1} \text{sim}(\text{word}_{i_1jk_1}, \text{word}_{i_2jk_2}) \\ \text{sim}(\text{word}_{i_1jk_1}, \text{word}_{i_2jk_2}) = & (w_{\text{Lev } j} \text{sim}_{\text{Lev}}(\text{word}_{i_1jk_1}, \text{word}_{i_2jk_2}) + \\ & w_{\text{JW } j} \text{sim}_{\text{JW}}(\text{word}_{i_1jk_1}, \text{word}_{i_2jk_2}) + \\ & w_{\text{Jaccard } j} \text{sim}_{\text{Jaccard}}(\text{word}_{i_1jk_1}, \text{word}_{i_2jk_2})) \\ & \log \left(\alpha + \frac{\beta}{f(\text{word}_{i_1jk_1})} \right) \log \left(\alpha + \frac{\beta}{f(\text{word}_{i_2jk_2})} \right); \end{aligned}$$

где sim_{Lev} – мера, основанная на редакционном расстоянии Левенштейна, $w_{\text{Lev } j}$ – соответствующий ей весовой коэффициент для поля j ;

sim_{JW} – мера, основанная на сходстве Джаро–Винклера, $w_{\text{JW } j}$ – соответствующий ей весовой коэффициент для поля j ;

$\text{sim}_{\text{Jaccard}}$ – мера, основанная на коэффициенте сходства Джаккарда, $w_{\text{Jaccard } j}$ – соответствующий ей весовой коэффициент для поля j ;

$f(\text{word}_{i_1jk_1})$ – частота слова $\text{word}_{i_1jk_1}$ поля j записи i_1 в общем массиве данных;

α, β – масштабирующие коэффициенты, необходимые для коррекции вклада статистической информации при расчете адаптивной меры подобия записей (АМПЗ).

Как видно, формула адаптивной меры подобия записей содержит множество весовых коэффициентов. Задача этих коэффициентов – усилить (ослабить) вклады каждой из мер подобия, входящих в АМПЗ, в зависимости от конкретных атрибутов имеющегося набора данных, а также увеличивать (уменьшать) важность того или иного атрибута набора данных при принятии решения об эквивалентности (различии) пар записей.

Сформировав обучающую выборку, состоящую из текстовых записей с указанием их взаимного соответствия/несоответствия, необходимо осуществить калибровку весовых коэффициентов адаптивной меры подобия записей, а также пороговых величин θ_{low} и θ_{high} . Их смысл заключается в следующем: при сопоставлении пары записей адаптивная мера представляет собой значение, которое необходимо сравнить с этими величинами. Таким образом, для того, чтобы можно было сделать вывод об эквивалентности (различии) пары записей, необходимо вычислить значение АМПЗ и далее сравнить полученное значение со значениями величин θ_{low} и θ_{high} . При значении АМПЗ, превышающем значение величины θ_{high} , делается вывод о том, что записи эквивалентны; при значении АМПЗ, меньшем значения величины θ_{low} , делается вывод о том, что записи различны. В случае, если значение адаптивной меры попадает в промежуток между значениями величин θ_{low} и θ_{high} – никаких выводов не делается. Такой подход обусловлен желанием выделить интервал значений, попадая в который при сопоставлении пары записей, АМПЗ не может адекватно оценить эквивалентность (различие) записей с позиции входящих в её состав алгоритмов нечеткого поиска. В этой ситуации решение предлагается принимать пользователю.

Таким образом:

$$\text{sim}(\text{record}_i, \text{record}_j) = \begin{cases} > \theta_{high} - \text{записи эквивалентны} \\ < \theta_{low} - \text{записи уникальны} \\ \geq \theta_{low} \text{ и } \leq \theta_{high} - \text{решение принимает пользователь} \end{cases} \quad (2)$$

Процесс построения адекватного эвристического решающего правила можно автоматизировать путем применения алгоритма машинного обучения, в качестве которого может быть использован, например, алгоритм из класса стохастических алгоритмов глобальной оптимизации [4, 5], т.е. алгоритмов, дающих приемлемое решение за оптимальное время.

Для обеспечения работы алгоритма оптимизации по обучению адаптивной меры подобия записей необходима функция приспособленности (фитнесс-функция), с помощью которой можно было бы оценивать корректность оптимизации АМПЗ и параметров её оценки. Для обеспечения условия (2) эта функция должна учитывать, с одной стороны, корректность определения пар эквивалентных записей при обучении меры, с другой – корректность определения пар уникальных записей. Кроме того, существенный интерес представляет возможность влиять на

процесс обучения АМПЗ параметрами указанной функции, задаваемыми пользователем до начала обучения. При помощи этих параметров будет определяться, насколько АМПЗ «самостоятельна» в принятии решений и, как следствие, насколько широким следует быть интервалу между пороговыми значениями, оценивающими адекватность АМПЗ. В качестве указанной функции приспособленности может использоваться, например, модифицированная $F_{measure}$:

$$\begin{aligned} \tilde{F}_{measure} &= \\ &= \frac{2(w_{recall} + w_{precision})}{w_{recall} (Recall_{duplicaties}^{-1} + Recall_{uniques}^{-1}) + w_{precision} (Precision_{duplicaties}^{-1} + Precision_{uniques}^{-1})} \quad (3) \end{aligned}$$

где $Recall_{duplicaties}$ – полнота определения пар эквивалентных записей; $Recall_{uniques}$ – полнота определения пар уникальных записей; $Precision_{duplicaties}$ – точность определения пар эквивалентных записей; $Precision_{uniques}$ – точность определения пар уникальных записей; w_{recall} – параметр, отвечающий за «вклад» полнот в общую величину $\tilde{F}_{measure}$; $w_{precision}$ – отвечает за «вклад» точностей.

На основании изложенного в процессе калибровки на каждой итерации алгоритма оптимизации с помощью АМПЗ оцениваются все возможные пары записей из заранее сформированной обучающей выборки. Основываясь на имеющейся информации о том, какие в выборке пары записей эквивалентные, а какие различные и на результатах производимого адаптивной мерой оценивания высчитывается значение $\tilde{F}_{measure}$. Задача алгоритма оптимизации – максимизировать $\tilde{F}_{measure}$ путем варьирования весовых коэффициентов и значений пороговых величин оценки адекватности АМПЗ.

В итоге, откалиброванная АМПЗ, а также значения пороговых величин, оценивающие её адекватность, могут быть использованы для оценки схожести записей согласно условию (2).

Для обеспечения функционирования указанных нами этапов алгоритма связывания данных были разработаны программные средства, совокупность которых реализует математическое и программное обеспечение программно-вычислительного комплекса для решения задачи интеграции атрибутивных данных информационных объектов. Комплекс, используя стандарты информационного взаимодействия (например, *HTTP*, *JDBC*), позволяет получать данные в стандартных (*XML*, *JSON*), либо специализированных форматах из различных информационных систем, *web*-ресурсов или баз данных для последующего связывания. Полученные результаты связывания могут быть сохранены либо в виде массива файлов, либо переданы посредством вышеперечисленных интерфейсов в реляционную базу данных или на *web*-ресурс. Таким образом, программно-вычислительный комплекс, организуя взаимодействие между программными системами, обеспечивает решение задачи интеграции атрибутивных данных информационных объектов (рис. 1).



Рис. 1. Схематическое представление структуры программно-вычислительного комплекса для решения задачи интеграции атрибутивных данных информационных объектов

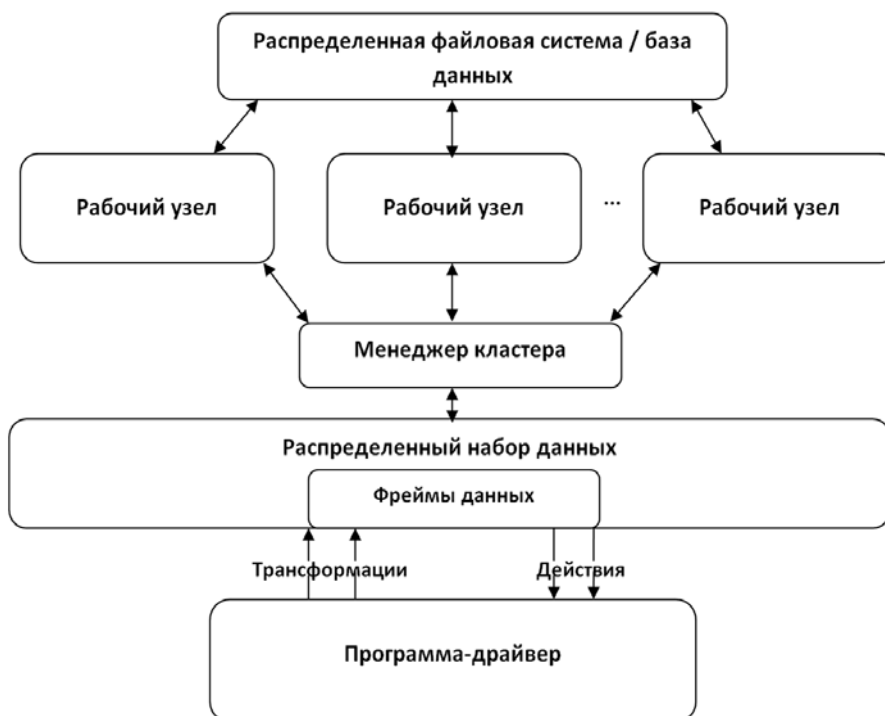


Рис. 2. Схематическое представление функционирования программно-вычислительного комплекса для решения задачи интеграции атрибутивных данных информационных объектов

Схематическое представление функционирования указанного программно-вычислительного комплекса показано на рис. 2. В основе комплекса лежит программа-драйвер, обеспечивающая функционирование программного и математического обеспечения для решения задачи интеграции данных. Эта программа инициирует распределенную обработку данных, хранящихся либо в файловой системе, либо в базе данных; за счет менеджера кластера обеспечивается изоляция распределенных данных при их параллельной обработке на вычислительных узлах и распределение этих данных по сети. Первоначально, в режиме обучения производится калибровка решающего правила на тренировочных наборах данных, после чего в режиме исполнения непосредственно осуществляется процесс связывания. Для обеспечения распределенной обработки данных использовался фреймворк распределенных вычислений *Apache Spark* [6].

ПРИМЕР ПРИМЕНЕНИЯ ПРОГРАММНО-ВЫЧИСЛИТЕЛЬНОГО КОМПЛЕКСА

Апробация представленного программно-вычислительного комплекса была выполнена при решении задачи координации ретроспективных сведений об объектах учета – непубликуемых результатах научно-технической деятельности в государственном информационном фонде непубликуемых научных исследований и их разработок.

На момент проектирования реляционной базы данных Единой государственной информационной системы учета научно-исследовательских, опытно-конструкторских и технологических работ гражданского назначения (ЕГИСУ НИОКТР) [7], аккумулирующей все сведения фонда, предшествующие создаваемой системе информационные ресурсы содержали архивные данные о непубликуемых результатах научно-технической деятельности. При работе над структурой системы с учетом недостатков хранения информации в указанных ресурсах проводился анализ ретроспективных данных для определения возможности соотнесения накопленных сведений, обеспечения их полноты и непротиворечивости при интеграции, а также достижения унификации с существующими установленными компонентами нормативно-справочной информации. При этом для обеспечения достоверного переноса сведений из ретроспективных баз данных и повышения качества структурирования информации в ситуациях, когда связывание данных по четкому сопоставлению оказывалось невозможно, использовался, в том числе, комплекс программного и математического обеспечения, представленный выше.

Из табл. 1 видно, что на момент запуска ЕГИСУ НИОКТР в промышленную эксплуатацию 1 января 2014 г. в информационный фонд системы были введены следующие объекты учета непубликуемых результатов научно-технической деятельности: научно-исследовательские, опытно-конструкторские и технологические работы гражданского назначения (НИОКТР), отчеты по этапам этих работ, диссертации на соискание ученых степеней, а также результаты интеллектуальной деятельности.

Поскольку была цель не только перенести ретроспективные сведения в новые структуры данных, но и обеспечить их достоверное структурирование и адаптацию для возможности эффективного использования, производилось нечеткое связывание данных, хранящихся в текстовых полях объектов учета с соответствующими полями из установленных компонентов нормативно-справочной информации. В табл. 2–4 отражены примеры статистических данных по результатам связывания интегрированных объектов учета с некоторыми справочниками.

Отметим, что все поступающие сведения об объектах учета в ЕГИСУ НИОКТР с 2014 г. унифицированы с требуемыми справочниками и заносятся поставщиками информации через учетные формы системы. В то же время, как показано на примерах в вышеприведенных таблицах, доля привязанных ретроспективных сведений к справочникам составила около 90%, что делает возможным использование этой информации в аналитических сервисах для получения достоверной статистики по накопленному фонду. Исключение составили случаи, когда учетные архивные данные содержали сведения, заведомо не поддающиеся связыванию.

Результатом произведенной интеграции стало обеспечение возможности построения интерактивных сервисов системы, предоставляющих доступ к различным открытым данным об учтенных объектах исследований и разработок. Сервисы системы построены таким образом, чтобы, с одной стороны, можно было проводить анализ конкретных, специфических срезов данных по указанным объектам. С другой стороны, чтобы имела возможность оценивать глобальное состояние ситуации в сфере объектов учета информационного фонда и связанных с ними показателей. Это, в свою очередь, позволяет осуществлять мониторинг ежегодного и ежемесячного изменения количества появляющихся научных работ в нашей стране в зависимости от их типов (рис. 3–6), а также контролировать выделение средств на НИОКТР различными источниками финансирования.

Таблица 1

**Количество объектов учета непубликуемых результатов научно-технической деятельности
с 01.01.1982 по 31.12.2013 гг., интегрированных в фонд системы ЕГИСУ НИОКТР**

НИОКТР	Отчеты	Диссертации	Результаты интеллектуальной деятельности
1 237 647	1 354 361	710 254	65 010

Таблица 2

**Результаты связывания сведений о НИОКТР по годам со справочником контрагентов
(с 2004 по 31.12.2013 гг.)**

год	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Общее кол-во	11005	12394	14444	13318	14568	23108	21136	30999	32797	27205
Привязанные, %	91,1	94,9	94,5	96,0	96,7	98,3	96,3	98,1	97,8	98,6

Таблица 3

**Результаты связывания сведений о диссертациях по годам со справочником контрагентов
(с 2004 по 31.12.2013 гг.)**

год	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Общее кол-во	24918	35288	27862	45756	28231	24723	21136	20391	24199	18109
Привязанные, %	89,1	91,9	89,5	96,2	93,7	98,7	97,1	96,3	95,8	97,4

Таблица 4

**Результаты связывания документов со справочником тематических рубрик
(с 2004 по 31.12.2013 гг.)**

	НИОКТР	Отчеты	Диссертации
Общее кол-во	200974	143428	272960
Привязанные, %	95,6	96,2	94,2



Рис. 3. Визуализация глобальных показателей непубликуемых результатов научно-технической деятельности, выявленных НИОКТР (сайт gosrid.ru).

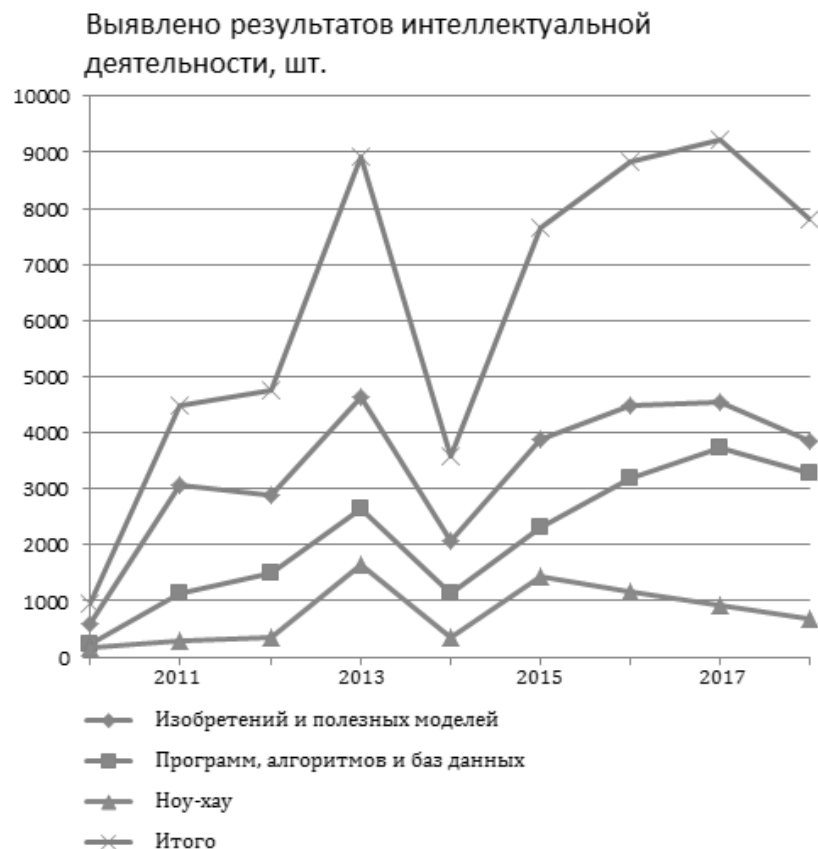


Рис. 4. Визуализация глобальных показателей непубликуемых результатов интеллектуальной научно-технической деятельности (сайт rosrid.ru).

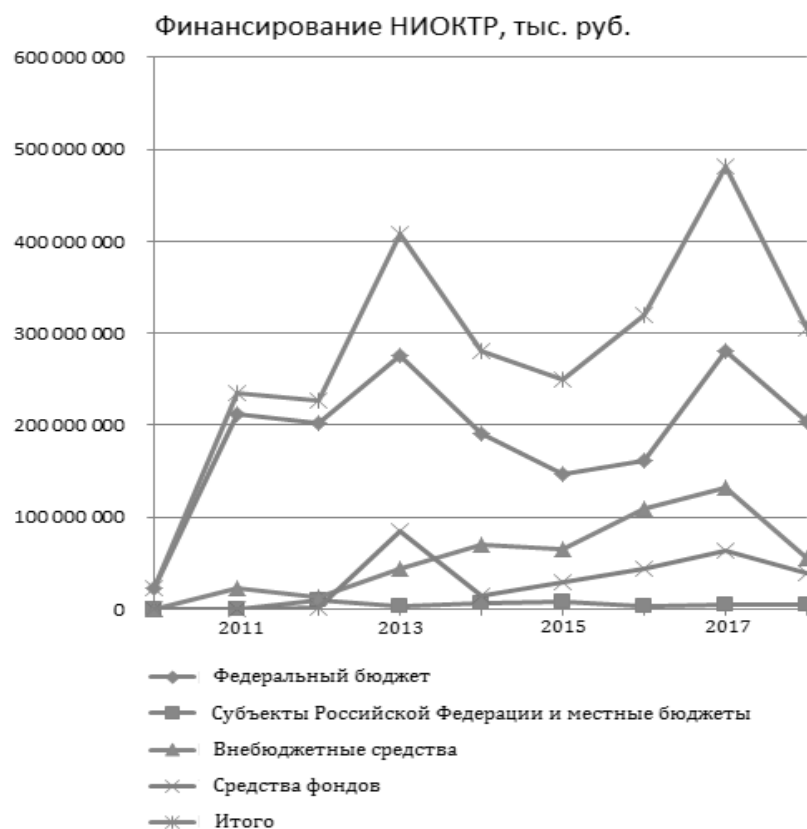


Рис. 5. Визуализация глобальных показателей непубликуемых результатов научно-технической деятельности, профинансированных НИОКТР (сайт rosrid.ru).

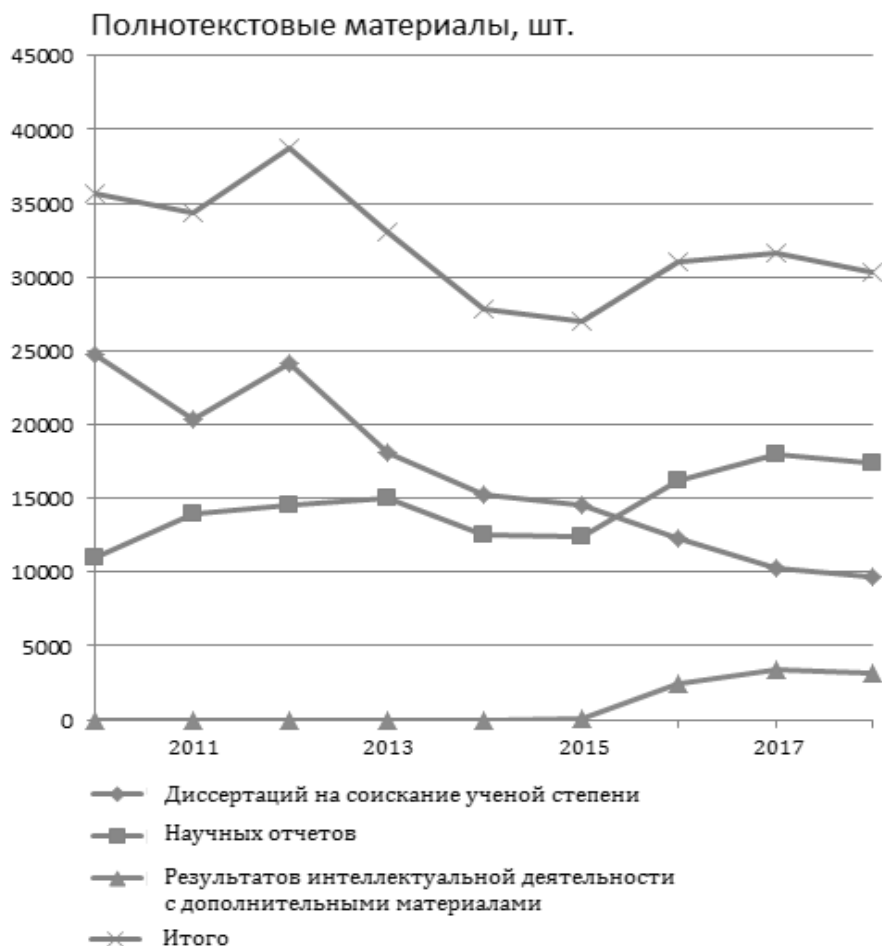


Рис. 6. Визуализация глобальных показателей непубликуемых (полнотекстовые материалы) результатов научно-технической деятельности (сайт rosrid.ru).

Для изучения аналогичных показателей на региональном и отраслевом уровнях запрашиваемые данные могут быть выданы и отображены на интерактивной карте Российской Федерации.

Вместе с тем, на случай возникновения необходимости получения данных по более специфическим параметрам для детального изучения имеется возможность получения такой информации с использованием интегрированного в ЕГИСУ НИОКТР инструментального комплекса. Последний осуществляет графическое отображение структуры и содержимого базы данных, а также обладает функционалом, позволяющим исследовать хранящуюся и вновь поступающую в фонд информацию в различных разрезах без необходимости программирования *SQL*-запросов к базе данных.

ЗАКЛЮЧЕНИЕ

Таким образом, разработанный программно-вычислительный комплекс, организуя взаимодействия информационных ресурсов, систем и баз данных, обеспечивает решение задачи интеграции атрибутивных данных информационных объектов.

Проведенная апробация программно-вычислительного комплекса на примере интеграции ретроспективных объектов учета непубликуемых результа-

тов научно-технической деятельности в государственном информационном фонде показала возможность его применения в качестве программного инструмента интеграции данных. Обеспечение структурирования указанных информационных объектов, их унификация с актуальными компонентами нормативно-справочной информации способствовали эффективному наполнению фонда, что подтверждают интерактивные сервисы информационной системы ЕГИСУ НИОКТР. В тоже время, представленный программно-вычислительный комплекс является инструментом интеграции общего назначения и может быть использован как одно из средств нечеткого связывания данных в других информационных системах, фондах и базах данных.

СПИСОК ЛИТЕРАТУРЫ

1. Атисков А.Ю., Перминов С.В. Гибридная адаптивная технология проектирования бизнес-процессов // Труды СПИИРАН. Вып. 4. – СПб: Наука, 2007.
2. Fellegi I.P., Sunter A.B. A theory for record linkage // Journal of the American Statistical Association. – 1969. – Vol. 64, №328. – P. 1183-1210.
3. Съедин Д.Ю. Разработка и реализация алгоритма связывания данных в государственной ин-

- формационной системе гражданского назначения // Научно-техническая информация. Сер. 2. – 2018. – №7. – С. 32-39; S'edin D. Yu. The Development and Implementation of the Data-Binding Algorithm in the State Civil Information System // Automatic Documentation and Mathematical Linguistics. – 2018. – Vol. 52, № 4. – P. 195-202.
4. Съедин Д.Ю. Новый стохастический гибридный алгоритм глобальной оптимизации на основе алгоритма М-РСА // Информатизация и связь. – 2017. – №1. – С. 143-148.
 5. Съедин Д.Ю. Параллельная реализация гибридного стохастического алгоритма глобальной оптимизации, основанного на алгоритме М-РСА // Информатизация и связь. – 2018. – №1. – С. 150-156.
 6. Reyes-Ortiz J.L., Oneto L., Anguita D. Big Data analytics in the cloud: Spark on Hadoop vs MPI/OpenMP on Beowulf. – 2015. – URL: <https://cyberleninka.org/article/n/1057727.pdf> (дата обращения: 13.07.2019).
 7. Борзых А.Н. Развитие единой государственной информационной системы учета научно-исследовательских, опытно-конструкторских и технологических работ гражданского назначения (ЕГИСУ НИОКТР) // Информатизация и связь. – 2011. – №6. – С. 40-41.

Материал поступил в редакцию 07.08.19

Сведения об авторе

СЪЕДИН Дмитрий Юрьевич – начальник отдела разработки и внедрения Федерального государственного автономного научного учреждения «Центр информационных технологий и систем органов исполнительной власти», Москва
e-mail: syedin@inevm.ru

Особенности онтологии исходных данных и метаданных в области систем виртуальной реальности

Предлагается перечень структурных элементов онтологии для систем виртуальной реальности: классификационные критерии технологий и систем виртуальной реальности; данные функционально-элементного, функционально-параметрического и модельно-параметрического анализа средств, поддерживающих технологии виртуальной реальности; программное обеспечение; рекомендуемые приложения к онтологии.

Ключевые слова: онтология исходных данных, онтология метаданных, технологии виртуальной реальности, системы виртуальной реальности

Под технологиями виртуальной реальности (*Virtual Reality – VR*) понимают совокупность технологических средств, предоставляющих человеку возможность воспринимать информационный контент, созданный аппаратно-программными средствами и действующий на анализаторы (органы чувств) человеческого организма. Объем конструктивно-технологических, программных и исследовательских решений, сложность и многовариантность аппаратно-программного обеспечения в рассматриваемой области настолько высоки, что для дальнейшего результативного совершенствования этого направления, для расширения спектра возможных прикладных применений существующих разработок необходимо формировать единую системную информационную базу, которая может быть представлена в виде онтологии исходных и метаданных в области технологий виртуальной реальности [1].

КРИТЕРИИ КЛАССИФИКАЦИИ ТЕХНОЛОГИЙ ВИРТУАЛЬНОЙ РЕАЛЬНОСТИ

Выбор определенного варианта аппаратно-технологического оснащения, необходимого для реализации технологии *VR*, в значительной степени определяется типом самой технологии.

Основным критерием классификации технологий *VR* является **степень погружения пользователя в виртуальную среду** [2], которая определяется количеством и качеством представления информационных воздействий, т.е. количеством анализаторов человека, воспринимающих сигналы, количеством и качеством представления воздействий, уровнем реализации обратной связи. По данному критерию различают технологии:

- обеспечивающие эффект полного погружения. Характеризуются высокой степенью детализации ви-

зуализируемого контента, мгновенным реагированием на действия пользователя, наличием специальных средств для обеспечения эффекта погружения (*Active Worlds*);

- без эффекта погружения. Предоставляют пользователю аудио-визуальный информационный контент высокого качества, проецируемый на широкоформатный экран;
- обеспечивающие взаимодействие с другими пользователями без эффекта полного погружения;
- на основе интернет-платформ (например, платформ *Second Life*, *Blue Mars*, *Sansar*).

Также технологии *VR* можно классифицировать в зависимости от:

О вида предоставляемого пользователю контента:

- максимально приближенный к реальности;
- иллюзия, в которой происходящее логически связано [3];

О способа представления визуального информационного контента зрителю:

- системы, в которых используются специальные периферийные устройства, относящиеся к классу *HDM* (*Head Mounted Display*, «Дисплей, закрепленный на голове»);
- *CAVE*-системы (*Cave Automatic Virtual Environments*, Комната виртуальной среды), в которых изображение проецируется на стены, пол и потолок; обеспечивается большая величина поля зрения. Положение пользователя, повороты его головы отслеживаются системой трекинга. Происходит глубокое погружение пользователя в виртуальную среду;
- система *MotionParallax3D*, позволяющая проецировать объемное изображение на внешний экран,

на котором пользователь через очки видит объемное изображение реального объекта, находящегося в соответствующей точке реального пространства, причем, изображение на экране меняется при повороте головы, что усиливает эффект присутствия во время просмотра [4];

- технология «виртуальный ретинальный монитор» (*Virtual Retinal Display, VRD; Retinal Scan Display, RSD*), обеспечивающая проекцию изображения на сетчатку глаза. Ее достоинствами являются: большой угол обзора – более 120°, высокие показатели разрешения, качество воспроизведения цвета, яркость, контраст, возможность воспроизведения объемного изображения с большой пространственной глубиной [5].

ФУНКЦИОНАЛЬНО-ЭЛЕМЕНТНЫЙ АНАЛИЗ АППАРАТНЫХ СРЕДСТВ, ПОДДЕРЖИВАЮЩИХ ТЕХНОЛОГИИ ВИРТУАЛЬНОЙ РЕАЛЬНОСТИ

Для выполнения функции «Воспроизведение изображения и звука, трекинга движения головы пользователя» в системах виртуальной реальности, относящихся к классу *HDM*, используются специальные периферийные устройства: очки и шлемы виртуальной реальности. Основными производителями этих устройств являются компании *Google, Samsung, HTC, Facebook и Sony* [6]. Очки виртуальной реальности – это устройство со встроенным модулем *NFC* (*Near Field Communication*, «ближняя бесконтактная связь»), позволяющим использовать смартфон для обработки и отображения 3D-изображения (например, очки *Google Cardboard*) [7]. Шлем виртуальной реальности – наиболее технологичное решение *HDM*. Он выполняет функции:

- воспроизведение (визуализация) изображения.** Элементами, реализующими данную функцию, являются два дисплея (формирующие изображения для каждого глаза) и система линз, предоставляющая возможность корректировки изображения [8]. Такая система визуализации создает полноценный стереоскопический эффект с иллюзией глубины. Одним из важнейших параметров дисплея является угол обзора;

- воспроизведение звука, регулировка его уровня в зависимости от расстояния пользователя до виртуального источника звука.** Многоканальная аудиосистема позволяет определять место нахождения источника звука, т.е. ориентироваться в виртуальном мире с помощью слуха. Для игровых приложений чаще востребованы наушники с технологией *Surround Sound 7.1* (наиболее популярные модели – беспроводные наушники *Astro A50* и *Logitech G930* [9]). Основное требование – объемный звук;

- трекинг** – отслеживание положения головы пользователя в пространстве. К элементам системы трекинга относятся: основные наружные (отслеживающая камера, датчики) и дополнительные внутренние (акселерометр, магнитометр, гироскоп). Акселерометр (микроскопический кристалл) отслеживает изменения скорости перемещения и повороты головы вокруг оси. Магнитометр (датчик Холла) позволяет обнаружить источники магнитного излучения [10]. Гироскоп –

микроэлектромеханическая система, инерциальный датчик, размещенный на печатной плате, по данным акселерометра и модуля *GPS* вычисляет положение устройства в 3D-пространстве [11].

Аппаратные средства, реализующие функцию «**Взаимодействие пользователя с виртуальной средой** (*Virtual Environment – VE*)», предполагают различные варианты управления средой *VR*: при положении сидя или стоя, с возможностью перемещения пользователя в пределах определенного пространства [12]. Реализацию этой функции выполняет чрезвычайно широкий спектр элементов.

В мобильных версиях *VE* чаще используются сенсорные панели и беспроводные джойстики. В стационарных версиях *VE* спектр дополнительных элементов, обеспечивающих взаимодействие пользователя с информационным контентом и усиливающих эффект погружения в *VR*, постоянно расширяется. Используются: джойстики и контроллеры, тактильные нарукавники, тактильные устройства для ног, тактильные накладки для лица, беговые платформы, оптические датчики для трекинга кистей рук (модуль *Leap Motion*), перчатки для реализации обратной связи с виртуальной реальностью, стимуляции тактильных ощущений у пользователя и многое другое.

ФУНКЦИОНАЛЬНО-ПАРАМЕТРИЧЕСКИЙ АНАЛИЗ АППАРАТНЫХ СРЕДСТВ, ПОДДЕРЖИВАЮЩИХ ТЕХНОЛОГИИ ВИРТУАЛЬНОЙ РЕАЛЬНОСТИ

Задачи функционально-параметрического анализа носят прямой и обратный характер. Решение прямой задачи позволяет определить возможные варианты реализации конкретной функции. Обратная задача предполагает обоснование значений параметров конструктивно-технологических и программных средств, обеспечивающих требуемый уровень реализации функции.

Одной из основных функций, реализуемых системами виртуальной реальности, является функция «**Взаимодействие системы и пользователя (интерактивность)**». Варианты ее реализации определяются такими параметрами аппаратно-программных средств и информационного контента, поддерживающих систему, как:

- количество возможных исходов определенного пользовательского действия;

- способность системы в ответ на действия пользователя симулировать ответную реакцию, субъективно воспринимаемую как естественную. Временной интервал между совершением действия пользователем (например, для зрительной системы – поворотом головы, перемещением взгляда) и моментом отображения этого действия виртуальной средой называется задержкой. Как показывает практика использования таких систем *VR*, как авиационные тренажеры, критическая величина задержки, превышение которой разрушает эффект погружения пользователя в виртуальную среду, составляет 20-50 мс;

- способность системы изменять окружающую среду;

- способность системы к ориентации и перемещению в среде.

Таблица 1

**Параметры шлемов виртуальной реальности, поддерживающие функцию «Отображение аудиовизуальной информации»
(фрагмент метаданных)**

Варианты реализации функции							
Типы шлемов VR	Угол поля зрения, формат дисплеев	Разрешение изображения экрана, дополнительные сервисы по сглаживанию изображения	Тип и параметры дисплея	Время отклика дисплея	Внедрение в видеоконтент (в формате 360°) интерактивных элементов	Воспроизведение звука	Совместимость шлема с рабочей платформой
Игровые	По горизонтали 45...120° (угол обзора человека 160°)	Полное разрешение, dpi: 2880×1600 2560×1440 2160×1200	Органический светоизлучающий (Organic Light Emitting Diode, OLED), активная матрица на органических светодиодах (AMOLED) [16]: - идеальный черный цвет - высокий уровень яркости - высокая контрастность изображения - минимальное время отклика - возможность изгиба	IPS LCD: около 5 мс AMOLED: 0,1 мс	Активные метки в виртуальном пространстве, предоставляющие возможность движения по различным траекториям.	Шлемы без наушников в комплекте. Фирменные наушники в комплекте, предложение о замене на модель с эффектом объемного звучания.	«Закрытые» шлемы, которые можно использовать только с ограниченным контентом, воспроизводимым определенными игровыми приставками и смартфонами.
Имеющие свой процессор и подключающиеся к компьютеру	Формат дисплея: 3,5 5,5"	Частота обновления: 80 Гц 90 Гц	- при частоте ниже 250 Гц заметно мерцание		Наличие дополнительного контента (изображения, видео, гиперссылки и т.д.).	Наушники с технологией Surround Sound 7.1.	«Открытые» модели, позволяющие использовать контент из разных источников
Мобильные, в которые вставляется смартфон со специальным приложением		Технология PenTile для внедрения субпикселей и улучшения качества изображения (2 или 3 субпикселя) [14]. Технология Anti-Screen Door Effect (Anti-SDE) (используется сетка, рассеивающая свет от каждого пикселя) [15]	Жидкокристаллический IPS LCD (In-Plane Switching Liquid Crystal Display) -дисплей: -максимальная точность отображения тонов и оттенков -в недорогих синеватый оттенок дисплея - возможность получить идеальный белый цвет -невозможно получить черный цвет, низкая контрастность изображения - отсутствует мерцание -малый запас яркости [17]		Возможность перехода из видеоконтента (в формате 360°) в смоделированную 3D-среду [18].	Бинауральный звук. Встроенные динамики, обеспечивающие простран-ственный звук, без блокировки внешнего звука	

Таблица 2

Фрагмент результатов модельно-параметрического анализа элемента систем VR «Очки виртуальной реальности»

Модель	Функциональные параметры			Технические параметры								Комплектация	Дополнительная информация
	Вывод изображения	Назначение	Регулировка	Частота обновления	Разрешение дисплея	Технология экрана	Угол обзора	Разъемы	Датчики	Системные требования	Масса		
Oculus Rift S	на собственный экран	для ПК	регулировки межзрачкового нет	80 Гц	2560×1440	LCD дисплей с полной RGB-матрицей	110°	«мини-джек» 3,5 мм	тирскоп; акселерометр; магнитометр; инфракрасные датчики	процессор Intel i3-6100/AMD Ryzen 3 1200, FX4350; GPU: NVIDIA GTX 1050Ti или AMD Radeon RX 470; RAM 8 Гб или более; DisplayPort 1.2; USB 3.0; Windows 10	450 г	контроллер движений	пять встроенных камер системы отслеживания движений Oculus Insight; два контроллера Oculus Touch; встроенные динамики
HTC Vive	на собственный экран	для ПК	межзрачкового и фокусировка расстояния	90 Гц	2160×1200 (1200×1080 для каждого глаза)	AMOLED	110°	«мини-джек» 3,5 мм; разъем для подключения зарядного устройства	акселерометр; гироскоп; датчик приближения	GPU: NVIDIA GeForce GTX 970 или AMD Radeon R9 290; CPU: Intel Core i5-4590 или AMD FX 8350; оперативная память: 4 Гб; видео-выход: HDMI 1.4 или DisplayPort 1.2	512 г	наушники; контроллер движений; внешний датчик положения в пространстве	встроенный микрофон; фронтальная камера; порты HDMI и USB 2.0

**Основные факторы, влияющие на выбор программного обеспечения,
которое может быть использовано при разработке и использовании информационного контента
в виртуальной среде**

Особенности программного обеспечения	Аппаратные средства		
	Шлемы и очки VR, подключаемые к ПК	Мобильные устройства	Веб-реализация виртуальной реальности
Функциональные особенности аппаратных средств	<i>HTC SteamVR</i> [19] <i>Oculus Rift</i> <i>OSVR HDK (Open-Source Virtual Reality Hacker Dev Kit)</i> [20]	<i>Samsung Gear VR</i> [21] <i>Google Cardboard</i>	в качестве базового устройства используется смартфон
		Устройства, работающие под управлением <i>Android</i> 4.0 и выше; <i>iOS</i> 7.0 (например, <i>iPhone</i> 4) и выше [22]	
Особенности VR-контента	Использование контроллеров движения Качественная графика	Ограничения по глубине, сложности и детальности сцены	
Программные среды	<i>Unity</i> <i>Unreal</i> Веб-реализации	<i>Unity</i> <i>Google Cardboard SDK (Software Development Kit)</i> Инструменты для создания мобильных приложений <i>VR (Cardboard SDK)</i> [23]	Язык разработки <i>Mozilla A-Frame</i> для создания кроссплатформенных VR-приложений [24] Веб-приложение <i>Vizor</i> [25]

• менее важной функцией системы VR, реализуемой такими аппаратными средствами, как шлемы виртуальной реальности, является «**Отображение аудиовизуальной информации**». Результаты функционально-параметрического анализа систем VR представлены в табл. 1 на примере метаданных по установлению взаимосвязи между вариантами реализации функции «Отображение аудиовизуальной информации» и уровнями параметров, которые достигнуты в наиболее распространенных в настоящее время шлемах виртуальной реальности [13].

Эффективная организация систем VR предполагает формирование аналогичных баз данных, основанных на результатах функционально-параметрического анализа для всего спектра функций, поддерживаемых широким спектром элементов и аппаратных средств, входящих в системы.

Модельно-параметрический анализ элементов систем VR. Цели модельно-параметрического анализа заключаются в формировании хранилищ данных, содержащих достоверную исходную информацию о количественных значениях функциональных и качественных параметров широкого спектра моделей аппаратно-программных средств систем VR, а также позволяющим совершенствовать эти базы данных (упрощение доступа к ним, возможность добавления новых данных и удаление устаревших). Полнота и достоверность исходной информации, возможность ее корректировки определяют степень целесообразности разработки онтологии исходных данных для решения практических задач в области систем VR.

В табл. 2 приведены результаты модельно-параметрического анализа таких элементов систем VR, как очки виртуальной реальности.

Программное обеспечение систем VR. Выбор программного обеспечения, используемого при разработке контента для VE, определяется рядом ряд

факторов. Основные из них, влияющие на выбор программного обеспечения, которое может быть использовано при разработке и использовании информационного контента в виртуальной среде, приведены в табл. 3.

Примерный перечень приложений к онтологии метаданных. Эффективность и актуальность онтологии исходных данных и метаданных в области технологий виртуальной реальности предполагают наличие пополняемой базы приложений, в которую должны входить ссылки на публикации по результатам соответствующих теоретических исследований и практических разработок по конкретным направлениям данной области [26].

ВЫВОДЫ

В настоящей работе сформулированы рекомендации по разработке онтологии исходных и взаимосвязанных данных для развивающейся и востребованной области технологий и систем виртуальной реальности и поддерживающих эти системы аппаратно-программных средств.

Открытый характер части онтологии, относящейся к исходным данным, позволит корректировать и дополнять информацию с учетом совершенствования технологических параметров средств VR и расширения спектра их функциональных возможностей. Открытый характер части онтологии, относящейся к метаданным, является следствием повышения требований пользователя к средствам VR и расширения области их использования.

Наличие онтологии дает возможность профессионалам и пользователям эффективно и результативно применять и совершенствовать средства *Virtual Environment* для решения широкого спектра исследовательских, производственных, культурно-просветительских и творческих задач.

СПИСОК ЛИТЕРАТУРЫ

1. Нестерова Е.И., Смогоржевский А.А. Особенности разработки онтологий метаданных для медиаиндустрии (на примере формирования аппаратно-технологических комплексов конференц-залов) // Научно-техническая информация. Сер. 2. – 2014. – №4. – С. 18-27; Nesterova E.I., Smogorzhevskii A.A. The Properties of Developing Metadata Ontologies for the Media Industry (Based on the Example of Creating Hardware Conference-Hall Complexes) // Automatic Documentation and Mathematical Linguistics. – 2014. – Vol. 48, №2. – P. 86-95. – URL: <http://lamb.viniti.ru/sid2/sid2free?sid2=J12391392> (дата обращения 14.08.18).
2. Chris Woodford Virtual reality. – URL: <https://www.explainthatstuff.com/virtualreality.html> (дата обращения 24.07.19)
3. «Объемная» музыка: как VR-технологии передают звук. – URL: <https://habr.com/ru/company/audiomania/blog/397147/> (дата обращения 24.07.19)
4. Мухамедзянова Д. Мы принесем голографический эффект к пользователям домой. – URL: <https://hightech.fm/2016/08/20/nettle> (дата обращения 24.07.19)
5. VR-очки Avegant: виртуальная реальность на сетчатке глаза. – URL: <https://total3d.ru/trends/117228/> (дата обращения 24.07.19)
6. Таскин Е. «Современное»: Уильям Оккам не одобряет. Что такое технология VR? – URL: <https://concepture.club/post/infopoloz/sovremennoe-uiljam-okkam-ne-odobrjaet-cto-takoe-tehnologija-vr> (дата обращения 18.09.19)
7. Лучшие VR очки и шлемы 2019. – URL: <https://etalongame.com/luchshie-vr-ochki-i-shlemy-2019/> (дата обращения 24.07.19)
8. Виртуальная реальность. – URL: https://spravochnick.ru/informatika/virtualnaya_realnost/ (дата обращения 24.07.19)
9. Обзор беспроводной игровой гарнитуры Logitech G930. – URL: https://www.eldorado.ru/research_center/article/detail.php?ID=442009204 (дата обращения 18.09.19)
10. Гришанков В. Магнитный датчик в смартфоне: что это и как работает? – URL: <https://androidlime.ru/magnetic-sensor-smartphone> (дата обращения 24.07.19)
11. Всё, что нужно знать о датчиках в вашем смартфоне. – URL: <https://ru.ihodl.com/technologies/2017-07-29/vsyo-cto-nuzhno-znat-o-datchikah-v-vashem-smartfone/> (дата обращения 24.07.19)
12. Эффективные устройства управления в VR. – URL: <https://habr.com/ru/post/387935/>
13. ТОП-9 шлемов и очков виртуальной реальности: критерии выбора актуальных моделей. – URL: <https://gadgets-reviews.com/ru/stati/426-obzor-7-glavnykh-shlemov-virtualnoj-realnosti.html#ixzz5tr33SCdj> (дата обращения 24.07.19)
14. Лучшие VR очки и шлемы 2019. – URL: <https://etalongame.com/luchshie-vr-ochki-i-shlemy-2019/> (дата обращения 24.07.19)
15. Samsung Trademarks ‘Anti-Screen Door Effect’ AMOLED Display. – URL: <https://uploadvr.com/samsung-trademarks-anti-screen-door-effect-amoled-display/> (дата обращения 24.07.19)
16. Что такое AMOLED-дисплей? – URL: <https://androidlime.ru/amoled-display> (дата обращения 25.07.19)
17. IPS экран: чем лучше матрица IPS, в чем ее недостатки. – URL: <https://5nch.com/ips-ekran-chem-luchshe-matritsa-ips-v-chem-ee-nedostatki/> (дата обращения 18.09.19)
18. Виртуальная реальность на службе у музея. – URL: <http://blog.maugry.ru/blog/technologies/virtual-reality-in-museums/> (дата обращения 25.07.19)
19. Vive Developer Mini Course. – URL: <https://learn.vrdev.school/p/vive-developer-mini> (дата обращения 25.07.19)
20. OSVR HDK: обзор и первые впечатления. – URL: <https://habr.com/ru/post/368373/> (дата обращения 25.07.19)
21. Gear VR Developer Mini Course. – URL: <https://learn.vrdev.school/p/gear-vr-developer-mini> (дата обращения 25.07.19)
22. Какой смартфон идеально подойдет для VR-очков: гайд по подбору + рейтинг лучших моделей. – URL: <https://vr4you.ru/zhelezo-i-aksessuary/smartfon-dlya-ochkov> (дата обращения 25.07.19)
23. Приложение VR Calibration for Cardboard позволяет настроить смартфон под очки виртуальной реальности. – URL: <https://holographica.space/news/vr-calibration-for-cardboard-4487> (дата обращения 25.07.19)
24. AFRAME. – URL: <https://aframe.io/> (дата обращения 25.07.19)
25. The WebVR Vizards. – URL: <https://site.vizor.io/> (дата обращения 25.07.19)
26. Свежие решения в области IT. Профессиональная группа. – URL: <https://www.professionalgroup.ru/vr-virtualnaya-realnost.html> (дата обращения 25.07.19)

Материал поступил в редакцию 26.07.19.

Сведения об авторах

НЕСТЕРОВА Елена Ивановна – доктор технических наук, доцент, заведующий кафедрой компьютерной графики и дизайна Санкт-Петербургского государственного университета кино и телевидения
e-mail: nesterova.el.i@yandex.ru

Общая и частные задачи многоуровневого синтеза моделей объектов мониторинга

Рассматриваются общая и частные задачи многоуровневого синтеза моделей объектов мониторинга, удовлетворяющих потребностям специалистов предметных областей в построении моделей при решении задач прогнозирования, управления и др. Общая задача формулируется как однокритериальная оптимизационная задача со многими ограничениями. Для оценки синтезируемых моделей предлагается система критериев и показателей эффективности синтеза. В рамках общей постановки задачи и в соответствии с построенной системой показателей формулируются частные задачи многоуровневого синтеза.

Ключевые слова: многоуровневый синтез, моделирование объектов, символические искусственные системы

ВВЕДЕНИЕ

В настоящее время значительное число технических объектов, применяемых при решении практических задач в предметных областях, относятся к сложным техническим объектам. Основным источником данных об объектах являются результаты наблюдения за ними. Наблюдение ведется с применением средств мониторинга.

Наблюдаемые объекты обладают следующими особенностями:

- имеют многоуровневую структуру, изменяющуюся во времени;
- реализуют многие целевые процессы;
- для них характерно наличие сложных внутренних взаимосвязей;
- от них поступают большие объемы данных.

Задачи прогнозирования состояния и управления этими объектами требуют синтеза их моделей.

Приведем примеры задач, возникающих в прикладных предметных областях. Так, в области телекоммуникаций необходимо обеспечить надежную работу сетей кабельного телевидения. Отказы в их работе приводят к потере пользователей. Особенности существующих сетей кабельного телевидения состоят в том, что большинство телевизионных приемников – это старые образцы, выпущенные более 10 лет назад. Кроме того, для передачи данных используются существующие сети, многие из которых имеют низкую пропускную способность. В этих условиях задача обеспечения надежности сетей среднего размера, а также крупных сетей, где число пользователей исчисляется миллионами, становится плохо решаемой.

В области медицины на первый план вышла задача системной обработки диагностических данных. Учеными-медиками обоснована необходимость в системном подходе к лечению пациентов. Это требует оценки целостности состояния организма, его спо-

собности к самовосстановлению, готовности противостоять внешним негативным воздействиям. Современные диагностические средства позволяют получать все необходимые данные с высокой точностью. Однако до настоящего времени отсутствуют средства для их системной обработки.

Рассмотренные и иные схожие примеры определяют новый класс задач – задачи синтеза моделей объектов мониторинга. Синтезируемые модели должны отражать структуру наблюдаемых объектов, учитывать взаимосвязи между их элементами, а также изменения этих взаимосвязей в пространстве и во времени, иметь относительно невысокую сложность.

При решении задач синтеза необходимо учитывать ряд особенностей, в частности, ограниченный доступ к исходным данным, ограниченные ресурсы для сбора и обработки данных, отсутствие полных моделей объектов и среды, ограничения на время решения задач синтеза, точность и достоверность формируемых решений. Перечисленные ограничения рассматриваются как основные. Этот перечень может, при необходимости, дополняться другими ограничениями.

АНАЛИЗ СУЩЕСТВУЮЩИХ МЕТОДОВ ПОСТРОЕНИЯ МОДЕЛЕЙ ОБЪЕКТОВ МОНИТОРИНГА

Для построения моделей наблюдаемых объектов возможно применение моделей и методов, разрабатываемых в рамках трех направлений: символического, эмерджентного и гибридного [1]. Для первого типа характерно использование продукционных и графовых моделей знаний. Для второго типа характерно применение ассоциативных нейросетевых методов, а также других конкурентных моделей и методов. Гибридные архитектуры предусматривают совместное использование моделей знаний и нейросетевых методов. Они ожидаемо обладают наиболее широкими возможностями.

Длительное время большое внимание уделялось символическим искусственным системам. Удалось создать ряд практически действующих, но узкоспециализированных решений. Обзор их можно найти в [1]. Наиболее известными символическими системами являются SOAR [2], MIDAS [3], Disciple [4], ICARUS [5], ARCADIA [6]. К общим недостаткам систем этой группы относятся сложность их обучения, слабая креативность, отсутствие ассоциативного запоминания информации и другие. В [7, 8] при построении искусственных символических систем предложено использовать модели и методы слияния данных. Однако, несмотря на это, создаваемые символические системы остаются ориентированными на решение четко обозначенных частных задач. Ни одна из известных систем не обладает возможностями моделирования реальных объектов в широком смысле.

Эмерджентные модели и методы разрабатываются на принципах биоподобия. Наиболее часто используются нейросетевые модели с ассоциативной обработкой сигналов. Нейросетевые модели и методы хорошо себя зарекомендовали прежде всего в распознавании образов в высокоразмерных данных, в обучении с подкреплением и в реализации ассоциативной памяти. Примеры известных эмерджентных когнитивных нейросетевых решений: иерархическая темпоральная память (*Hierarchical Temporal Memory – HTM*) [9]; система DesTIN [10]; интегрированная биологически основанная архитектура (*Integrated Biologically based Cognitive Architecture – IBCA*) [11]; мобильное адаптивное устройство с нейронной организацией (*Neurally Organized mobile Adaptive Device – NOMAD*) [12]; системы, создаваемые на нейроморфных процессорах TrueNorth [13] и др [14]. Наиболее подходящими для решения задач синтеза объектов мониторинга в рамках эмерджентного подхода являются рекуррентные нейросетевые структуры с ассоциативной обработкой информации. Однако существующие ассоциативные нейросетевые структуры, в частности, Хопфилда, Коско, RMLP, RTRN, Эльмана [1, 9–17], не позволяют достичь в нейросетевых машинах уровня, необходимого для оперативной и глубокой обработки воспринимаемых сигналов. Некоторые новые решения, полученные в последние годы [18–22], позволяют надеяться на прорыв в этой области при дальнейшем их развитии. Однако для создания новых нейросетевых машин на их основе может потребоваться длительное время.

В целом, проведенный анализ показал, что имеющиеся подходы, включая гибридный, не позволяют решать задачи построения моделей объектов с учетом их особенностей и имеющихся ограничений. Кроме того, искусственные системы, по большей части, продолжают оставаться только экспериментальными. Однако, символические искусственные системы, способные к слиянию данных, имеют большие потенциальные возможности. Создание на их основе новых символических систем, способных к моделированию реальных объектов, возможно за счет применения в них аппарата автоматического синтеза.

ОБЩАЯ ЗАДАЧА МНОГОУРОВНЕВОГО СИНТЕЗА МОДЕЛЕЙ ОБЪЕКТОВ МОНИТОРИНГА

Требуется путем синтеза S найти модель B_0 , при которой достигается W_{opt} – экстремум основного показателя эффективности этой модели W при ограничениях на вспомогательные показатели $M(B_v)$:

$$W_{opt}(B_0) = \max_{v \in \Lambda} \{W_v(S, M(B_v))\},$$

$$M(B_v) = F(M_1(B_v), M_2(B_v), \dots, M_n(B_v)) \in E;$$

$$M_1(B_v) \in E_1, M_2(B_v) \in E_2, \dots, M_n(B_v) \in E_n;$$

где Λ – множество вариантов построения вычислительной модели B_0 ; $M_1(B_v), M_2(B_v), \dots, M_n(B_v)$ – группы вспомогательных показателей; $E, E_1, E_2, \dots, E_n$ – области допустимых значений для показателей эффективности.

С учетом требований, предъявляемых к моделям объектов, и имеющихся ограничений предлагается следующий процесс для их синтеза:

$$S: X \rightarrow R \rightarrow O \rightarrow Y,$$

где X – исходные потоки данных, поступающие от наблюдаемых объектов; R – результаты обработки данных; O – синтезированные модели наблюдаемых объектов; Y – новые модели. Синтез моделей осуществляется за счет выявления информационных элементов в исходных потоках данных и их многократного преобразования и связывания между собой.

Для получения данных, необходимых для развития построенных моделей объектов, определяются процессы и программы мониторинга:

$$\tilde{S}: Y \rightarrow P(\text{PRG}) \rightarrow X,$$

где $P(\text{PRG})$ – синтезированные модели процессов (программ) мониторинга. Процессы определяют данные, которые необходимо получить о наблюдаемых объектах, и способы их обработки. Программы содержат машинные инструкции, обеспечивающие реализацию процессов с применением имеющихся технических средств.

Элементы синтезированных по исходным данным моделей O и установленные между ними связи позволяют построить пространство, в котором могут синтезироваться новые модели Y . При синтезе новых моделей задаются исходная и целевые модели, выполняется доказательство существования целевой модели при известной исходной.

Рассмотренный процесс позволяет строить многоуровневые ассоциативные модели наблюдаемых объектов на основе исходных данных. Многоуровневая структура моделей дает возможность рассматривать объект на многих уровнях. За счет этого обеспечивается снижение вычислительной сложности решаемых задач прогнозирования и др. Наличие ассоциативных взаимосвязей в моделях обеспечивает возможность их использования при решении практических задач, которые требовали участия экспертов. Более детально процесс синтеза моделей объектов мониторинга рассмотрен в [23].

СИСТЕМА ПОКАЗАТЕЛЕЙ И КРИТЕРИЕВ ЭФФЕКТИВНОСТИ МНОГОУРОВНЕВОГО СИНТЕЗА МОДЕЛЕЙ ОБЪЕКТОВ МОНИТОРИНГА

В соответствии с рассмотренной постановкой задачи предлагается система показателей и критериев эффективности многоуровневого синтеза, представленная на рисунке.

На нижнем уровне определяются показатели эффективности обработки контента информационных потоков с учетом наблюдаемого контекста. Под контентом понимается содержание потоков данных, поступающих от наблюдаемых объектов. Контекст определяется условиями сбора и обработки данных, ожиданиями конечных потребителей и другими условиями, влияющими на синтез наблюдаемых объектов. Результаты обработки во многом зависят от характеристик исходных данных.

Оценка эффективности синтеза моделей объектов дается исходя из полноты синтезируемых моделей и вычислительной сложности их синтеза. Эффективность синтезируемых процессов мониторинга определяется скоростью достижения поставленной цели. В частности, скоростью получения данных, необходимых для синтеза моделей наблюдаемых объектов, или обработки этих данных. Синтез моделей программ мониторинга оценивается исходя из объема временных и вычислительных ресурсов, необходимых на их выполнение, длин программ. Наряду с перечисленными могут использоваться другие показатели, традиционно применяемые для оценки программ.

При решении прикладных задач с использованием синтезированных моделей объектов в качестве пока-

зателей эффективности могут рассматриваться: число одновременно наблюдаемых объектов, минимальное время получения данных, минимальный объем расходуемых ресурсов и др.

К ключевым показателям многоуровневого синтеза моделей относятся полнота синтезируемых моделей и вычислительная сложность их синтеза. Полнота определяется соотношением известных и неизвестных параметров синтезированной модели. Полнота модели, построенной на момент времени r , определяется как:

$$Q_r = 1 - \frac{1}{c} \sum_{j=1}^c q_{jr},$$

$$q_{1r} = \frac{\| (DA(M_{r-1})) | - Z_1 \|}{Z_1}; \quad q_{2r} = \frac{\| (DB(M_{r-1})) | - Z_2 \|}{Z_2};$$

$$q_{3r} = \frac{\| (DC(M_{r-1})) | - Z_3 \|}{Z_3}; \quad q_{4r} = \frac{\| (FB(M_{r-1})) | - Z_4 \|}{Z_4};$$

$$q_{5r} = \frac{\| (FC(M_{r-1})) | - Z_5 \|}{Z_5},$$

где M_{r-1} – модель на $(r-1)$ момент времени; $|DA|$, $|DB|$, $|DC|$ – допустимые множества входных, внутренних и выходных элементов модели; $|FB|$, $|FC|$ – допустимые множества связей между внутренними и выходными элементами; $Z_1, Z_2, \dots, Z_5$ – мощности соответствующих множеств; c – коэффициент, назначаемый в соответствии с числом параметров модели, $c=5$. При $Q_r = 1$ модель является полной, при $Q_r = 0$ параметры модели не определены.



Система показателей и критериев эффективности многоуровневого синтеза моделей объектов мониторинга

Вычислительная сложность синтеза определяется числом операций или временем, необходимым для построения модели. Верхняя граница времени T_H определяется как

$$T_H \approx c \sum_{i=1}^K m_i^2 \leq c \left(\sum_{i=1}^K m_i \right)^2,$$

где c – постоянный коэффициент; m_i – число условий задачи на i -м уровне. Эта оценка справедлива, когда число условий задач многоуровневого и одноуровневого синтеза одинаково. Принимая во внимание, что на каждом верхнем уровне один шаг синтеза эквивалентен n_i шагам на уровне “0”, оценка нижней границы времени T_L при многоуровневом синтезе равна:

$$T_L \approx c \sum_{i=0}^K \frac{m_i^2}{n_i^2} \leq c \sum_{i=0}^K m_i^2,$$

Тогда среднее время вычисляется как $T = (T_L + T_H) / 2$. В случаях, когда при синтезе моделей объектов известны вероятности перехода p_i на более низкие уровни, для оценки сложности используется следующее выражение:

$$T \approx cn_0^2 + c \sum_{i=0}^K p_i \frac{m_i^2}{n_i^2}.$$

ЧАСТНЫЕ ЗАДАЧИ СИНТЕЗА МОДЕЛЕЙ ОБЪЕКТОВ МОНИТОРИНГА

Частные задачи формулируются в соответствии с показателями, определяемыми рассмотренной системой и критериями эффективности. Выделяются следующие группы частных задач: прикладные, синтез моделей процессов и программ мониторинга, синтез моделей объектов мониторинга, трансформация контента в контексте. Ниже приводятся отдельные частные задачи.

Частные задачи прикладного уровня. К типовым задачам прикладного уровня относится поиск способа синтеза моделей объектов, процессов и программ мониторинга V_o , при котором достигается максимальное число одновременно наблюдаемых объектов с периодичностью P не ниже заданной:

$$\begin{aligned} Q_o(V_o, t) &= \max_{i \in \Omega} Q(V_i(K(t)), K(t), C(t), t), \\ V_i &\in E_v; \quad i = \overline{1, I} \\ P(V_i(K(t)), K(t), C(t), t) &\geq P_{\text{зад}}, \end{aligned}$$

где $V_i(K(t))$ – способ мониторинга; $C(t)$ – контент; $K(t)$ – контекст Ω – множество способов синтеза моделей объектов, процессов и программ мониторинга.

Другой типовой задачей является поиск способа синтеза моделей объектов, процессов и программ мониторинга V_o , обеспечивающего максимальную информативность синтезируемых моделей объектов при ограничениях на точность Z и достоверность R :

$$\begin{aligned} G_o(V_o, t) &= \max_{i \in \Omega} G(Z_i(V_i, R_i), R_i(V_i), C(t), t); \\ Z_i(V_i, R(V_i), t) &\geq Z_{\text{зад}}; \quad R_i(V_i(K(t)), K(t), t) \geq R_{\text{зад}}. \end{aligned}$$

Возможен поиск способа синтеза моделей объектов, процессов и программ V_o , обеспечивающего получение необходимой информации за минимальное время T_o при заданных допустимых значениях точности Z и информативности G синтезируемых моделей:

$$\begin{aligned} T_o(V_o, t) &= \min_{i \in \Omega} T(V_i(K(t)), K(t), C(t), t), \\ G_i(V_i, R(V_i), t) &\geq G_{\text{зад}}; \quad Z(V_i, R_i(V_i), t) \geq Z_{\text{зад}}; \\ R(V_i(K(t)), K(t), C(t), t) &\geq R_{\text{зад}}. \end{aligned}$$

Поиск способа синтеза моделей объектов, процессов и программ V_o может также выполняться по критерию минимального расхода ресурсов, обеспечивающего заданную информативность результирующей модели:

$$\begin{aligned} U_o(V_o, t) &= \min_{i \in \Omega} U(V_i(K(t)), K(t), C(t), t); \\ G(Z(V_i), R(V_i), K(t), t) &\geq G_{\text{зад}}; \\ Z(V_i(K(t)), R(V_i), t) &\geq Z_{\text{зад}}; \\ R(V_i(K(t)), K(t), C(t), t) &\geq R_{\text{зад}}. \end{aligned}$$

Частные задачи синтеза процессов и программ мониторинга. При синтезе процессов и программ следует исходить из достижения максимальной скорости обработки информации D при ограничениях на точность обработки Z , которая определяется объемом информации I и видами обработки L :

$$\begin{aligned} D_o(V_o, \Delta t) &= \max_{i \in \Omega} D(V_i(K(\Delta t)), K(\Delta t), C(\Delta t)); \\ Z(V_i, I(\Delta t), L(\Delta t)) &\geq Z_{\text{зад}}. \end{aligned}$$

Синтез процессов и программ мониторинга V_o можно выполнять, требуя минимизацию затрат на перестройку существующих процессов U с учетом ожидаемого повышения информативности моделей объектов:

$$\begin{aligned} U_o(V_o, t, \Delta t) &= \min_{i \in \Omega} U(\Delta V(t), \Delta G(t)); \\ \Delta V(t) &= |V_i(t) - V'|; \quad \Delta G(t) = |G(V_i(t)) - G(V')|. \end{aligned}$$

Еще один критерий – обеспечение при синтезе максимальной информативности результатов обработки с учетом входного контента и контекста при ограничениях на ресурсы:

$$\begin{aligned} G_o(V_o, \Delta t) &= \max_{i \in \Omega} G_i(V(K(\Delta t)), K(\Delta t)); \\ U(V_i(K(\Delta t))) &\leq U_{\text{зад}}. \end{aligned}$$

Частные задачи синтеза моделей объектов мониторинга. При синтезе моделей объектов возможна постановка задачи поиска способа синтеза V_o , обеспечивающего построение наиболее полной модели по имеющимся контенту и контексту с учетом возможностей получения дополнительной информации при ограничениях на достоверность R и точность Z модели, а также ресурсы U для ее построения:

$$\begin{aligned} G_o(V_o, t) &= \max_{i \in \Omega} G(Z(V_i), R(V_i), K(t), \Delta K) \\ R(V_i, \Delta K) &\geq R_{\text{зад}}; \quad Z(R_i(V_i), \Delta K) \geq Z_{\text{зад}}; \quad U(V_i(\Delta K)) \leq U_{\text{зад}} \end{aligned}$$

Другой возможной задачей синтеза моделей объектов мониторинга является поиск способа V_0 , обеспечивающего максимум достоверности моделей мониторинга с учетом возможностей получения дополнительной информации при ограничениях на полноту G и затрачиваемое время T :

$$R_0(V_0, t) = \max_{i \in \Omega} R_i(V_i(K(t)), K(t), \Delta K_i),$$

$$R(V_i) = (1 - \prod_j (1 - R_j(V_j)))$$

$$G(Z(V_i), R(V_i), \Delta K) \geq G_{\text{зад}}; T(V_i(\Delta K)) \leq T_{\text{зад}}.$$

Частные задачи трансформации данных мониторинга. При решении задач трансформации данных может определяться такой способ V_0 трансформации, который обеспечит достижение минимума времени, необходимого для обработки данных, при условии, что точность Z преобразования и достоверность R результата будет не ниже требуемой:

$$T_0(V_0, t) = \min_{i \in \Omega} T(V_i(K(t)), C(t), t),$$

$$Z(R(V_i), t) \geq Z_{\text{зад}}; R(V_i, K(t), C(t), t) \geq R_{\text{дон}}.$$

В другой постановке задачи необходимо найти способ трансформации V_0 , обеспечивающий максимум точности преобразований при ограничениях на отводимые ресурсы U :

$$Z_0(V_0, t) = \max_{i \in \Omega} Z(R(V_i), K(t), C(t), t);$$

$$R(V_i(K(t)), K(t), C(t), t) \geq R_{\text{дон}};$$

$$U(V_i(K(t)), K(t), C(t), t) \leq U_{\text{зад}}.$$

Также может выполняться поиск способа трансформации V_0 , обеспечивающего максимум информативности контента на единицу объема обрабатываемого потока $K(t)_{\Delta 1}$ при ограничениях на время обработки:

$$G_0(V_0, \Delta t) = \max_{i \in \Omega} G(Z(V_i), R(V_i), K_{\Delta 1}(t), K(t), C(t), t);$$

$$T(V_i(K_{\Delta 1}(t)), K_{\Delta 1}, K(t), C(t), t) \leq T_{\text{зад}}.$$

ПРАКТИЧЕСКИЕ ОЦЕНКИ СЛОЖНОСТИ СИНТЕЗА МОДЕЛЕЙ ОБЪЕКТОВ МОНИТОРИНГА

Сложность синтеза моделей объектов по исходным данным определяется сложностью выявления отдельных элементов в исходных данных, восстановления ассоциативных связей между элементами, а также сложностью формирования на их основе многоуровневых структур и связывания структур между собой.

Для выявления зависимостей между элементами данных, передаваемых в потоках, применяются алгоритмы поиска ассоциативных зависимостей. Среди них наибольшую известность получил алгоритм *A priori*. Оценки сложности алгоритмов приведены в [24].

Для связывания многоуровневых структур применимы модели и методы теории графов. При этом выполняется сопоставление структур, для чего решается задача распознавания изоморфизма графов. Она относится к классу NP , однако, неизвестно, является ли она NP -полной или принадлежит классу P [25]. Время, затрачиваемое на ее решение, экспоненциально возрастает с увеличением числа узлов в графах. Для снижения вычислительной сложности разработано значительное число алгоритмов, включая алгоритм Ульмана (*Ullmann's algorithm*), алгоритм Шмидта – Дрюффеля (*algorithm of Schmidt and Druffel – as SD*), VF алгоритм и его модификацию $VF2$, алгоритм *Nauty*. Оценки сложности для этих алгоритмов приведены в [25]. В частности, алгоритм $VF2$ позволяет сопоставлять атрибутивные графы отношений (*Attributed Relational Graphs – ARG*), содержащие от 2 до 3000 узлов, выполняя при этом не более $O(N!N^3)$ операций, где N – количество узлов в сравниваемых графах [26]. В лучшем случае $VF2$ решает задачу за $O(N!N)$ операций.

При синтезе моделей объектов путем доказательства их существования решается задача прямого и обратного логического вывода [27, 28]. Его сложность зависит от используемой логики. В значительном числе случаев используются дескриптивные логики (*Description Logic – DL*). Их подробное описание и оценки их вычислительной сложности приведены в [29]. Одной из базовых логик семейства DL является ACL логика (*Attributive Language with Complement*). Проблема выполнимости концептов ACL лежит в классе $PSPACE$. К расширениям ACL относится логика $SHIQ$. Она дополнена инверсными ролями, качественными ограничениями кардинальности ролей, транзитивными ролями и иерархией ролей. $SHOIQ$ допускает работу с номиналами. Логика $sROIQ$ включает в себя логику $SHIQ$, а также составные аксиомы вложенности ролей в $TBox$, непересекаемости, рефлексивности, иррефлексивности и асимметричности ролей, универсальную роль и др. $sROIQ$ является $NExpTime$ -трудной.

К наиболее известным машинам вывода можно отнести $KAON2$ и $Pellet$. $KAON2$ поддерживает логику $SHIQ$, $Pellet$ – логику $sROIQ$. В $KAON2$ реализован алгоритм, основанный на резолюции, $Pellet$ реализует табло-алгоритм.

ЗАКЛЮЧЕНИЕ

В статье рассмотрена проблема построения моделей объектов по данным мониторинга, являющаяся актуальной для различных предметных областей, в частности, области телекоммуникаций, области медицины. Определены основные особенности моделируемых объектов и сформулированы предъявляемые к моделям требования. Показано, что для выполнения требований модели должны иметь многоуровневую структуру, относиться к числу ассоциативных моделей.

Анализ существующих подходов, применимых для построения моделей объектов, показал, что в настоящее время они не позволяют моделировать реальные объекты. Предложено развитие символического

ского подхода за счет создания новых систем, основу которых составляет аппарат автоматического синтеза. Новые системы предусматривают многоуровневый синтез моделей объектов за счет выявления информационных элементов в исходных потоках данных и их связывания. Развитие моделей обеспечивается путем синтеза процессов сбора дополнительных данных о наблюдаемых объектах и их обработки. Для реализации процессов формируются исполняемые программы. Результирующие модели наблюдаемых объектов применяются при решении прикладных задач.

В соответствии с рассмотренным процессом синтеза разработана система показателей и критериев эффективности. Структура системы включает: уровень обработки исходных данных, уровень синтеза моделей наблюдаемых объектов, уровень синтеза процессов и программ мониторинга, уровень решения прикладных задач. Для каждого из них сформулированы типовые частные задачи, которые могут на нем решаться. В результате общая задача синтеза сводится к формулировке и решению последовательности частных задач.

В целом, предложенное решение развивает существующий символический подход к построению искусственных систем до уровня, необходимого для синтеза моделей реальных объектов.

СПИСОК ЛИТЕРАТУРЫ

1. Kotseruba I., Tsotsos J.K. 40 years of cognitive architectures: core cognitive abilities and practical applications // *Artificial Intelligence Review* – 2018. – URL: <https://link.springer.com/article/10.1007/s10462-018-9646-y>
2. Laird J. E. The soar cognitive architecture. – Cambridge, MA: MIT Press, 2012.
3. Gore B.F., Hooley B.L., Wickens C.D., Scott-Nash S. A computational implementation of a human attention guiding mechanism in MIDAS. Vol. 5 // *International Conference on Digital Human Modeling*. July 19-24, 2009 – San Diego, CA, 2009.
4. Tecuci G. et al. An experiment in agent teaching by subject matter experts // *Int. J. Hum. Comput. Stud.* – 2000. – Vol. 53, № 4 – P. 583-610.
5. Langley P., Choi D. Learning recursive control programs from problem solving // *Journal of Mach. Learn. Res.* – 2006. – Vol. 7. – P. 493-518.
6. Bridewell W., Bello P.F. Incremental object perception in an attention-driven cognitive architecture // *Proceedings of the 37th Annual Meeting of the Cognition Science Society (CogSci 2015 July 22-25)*. – Pasadena, CA, 2015. – P. 279-284.
7. Glodek M. et al. Fusion paradigms in cognitive technical systems for human-computer interaction // *Neurocomputing*. – 2015. – Vol. 161. – P. 17-37.
8. Perlovsky L.I. Cognitive high level information fusion // *Information Sciences*. – 2007. – P. 2099-2118.
9. Hawkins J., Blakeslee S. On Intelligence. – New York: Times Books, 2004.
10. Arel I., Rose D., Coop R. Destin: a scalable deep learning architecture with application to high-dimensional robust pattern recognition // *Proceedings of the AAAI Workshop on Biologically Inspired Cognitive Architectures*, November. – Washington, D.C.: National Academy of Sciences, 2009.
11. O'Reilly R.C., Braver T.S., Cohen J. // A biologically-based computational model of working memory // *Models of Working Memory* / eds. A. Miyake, P. Shah. – New York: Cambridge University Press, 1999. – P. 375-411.
12. Fleischer J., Gally J., Eldeman G., Krichmar J. Retrospective and prospective responses arising in modeled hippocampus during maze navigation by a brain-based device // *Proceedings of the National Academy of Sciences of America*. Vol. 104. – Washington D.C.: National Academy of Sciences, 2007.
13. Modha S. Introducing a brain-inspired computer TrueNorth's neurons to revolutionize system architecture. – URL: <http://www.research.ibm.com/articles/brain-chip.shtml>
14. Haykin S., Fatemi M., Setoodeh P., Xue Y., Cognitive control // *Proceedings of the IEEE*. – 2012. – Vol. 100. – P. 3156-3169.
15. Haikonen P. The role of associative processing in cognitive computing // *Cognitive Computation*. – 2009. – Vol. 1. – P. 42-49.
16. Palm G. Neural associative memories and sparse coding // *Neural Networks*. – 2013. – № 37. – P. 165-171.
17. Haykin S. Neural networks and learning machines, 3rd ed. – New York: PrenticeHall, 2008.
18. Osipov V., Osipova M. Space-time signal binding in recurrent neural networks with controlled elements // *Neurocomputing*. – 2018. – № 308. – P. 194-204.
19. Osipov V., Nikiforov V. Formal aspects of streaming recurrent neural networks // *Proceedings 15th International Symposium on Neural Networks «Advances in Neural Networks– ISNN 2018»*, LNCS 10878, (Minsk, Belarus, June 25-28, 2018) / eds. T. Huang et al. – Cham, Switzerland: Springer, 2018. – P. 29-36.
20. Osipov V. Space-time structure of recurrent neural networks with controlled synapses // *Proceedings International Symposium on Neural Networks «Advances in Neural Networks – ISNN 2016»*, LNCS 9719 / eds. L. Cheng et al. – Cham, Switzerland: Springer, 2016. – P. 177-184.
21. Osipov V. Structure and basic functions of cognitive neural network machine // *MATEC Web of Conferences*. – 2017. – Vol. 113(20 June).
22. Осипов В.Ю. Ассоциативная интеллектуальная машина с тремя сигнальными системами // *Информационно-управляющие системы*. – 2014. – № 5(72). – С. 12-17.
23. Osipov V. Automatic synthesis of multilevel automata models of biological objects // *International Conference on Computational Science and Its Applications (ICCSA 2019)*. Lecture Notes in Computer Science. – Cham: Springer, 2019.
24. Kusters W.A., Pijls W., Popova V. Complexity analysis of depth first and FP-growth implementations of APRIORI // *Machine learning and data mining in pattern recognition*. MLDM Lecture

- Notes in Computer Science (Lecture Notes in Artificial Intelligence). Vol. 2734. / eds. P. Perner, A. Rosenfeld. – Berlin – Heidelberg: Springer, 2003. – P. 284-292.
25. Foggia P., Sansone C., Vento M.A. Graph-based representations in pattern recognition // Proceedings of the 3rd IAPR TC-15 Workshop, Ischia, Italy, May 23-25, 2001. – Ischia: C.U.E.N., 2001. – P. 188-199.
26. Cordella L.P. Evaluating performance of the VF graph matching algorithm Proc. of the 10th International Conference on Image. Analysis and Processing // IEEE Computer Society Press. – 1999. – P. 1172-1177.
27. Чень Ч., Ли Р. Метод резолюций // Математическая логика и автоматическое доказательство теорем. Гл. 5. – М.: Наука, 1983. – С. 358.
28. Маслов С.Ю. Обратный метод установления выводимости для логических исчислений // Труды математического института им. В.А. Стеклова АН СССР, Т.98, 1968.
29. Donini F. et al. Reasoning in description logics // Principles of knowledge representation. – Stanford, CA: Center for the Study of Language and Information, 1996. – P. 191-236.

Материал поступил в редакцию 25.07.19.

Сведения об авторах

ЖУКОВА Наталия Александровна – кандидат технических наук, доцент, старший научный сотрудник лаборатории информационно-вычислительных систем и технологий программирования Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИ РАН).
e-mail: nazhukova@mail.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81'322.3'36'373.423

К.К. Боярский, Е.А. Каневский

Семантика устойчивых словосочетаний с глаголами*

Рассматриваются вопросы омонимии при компьютерном анализе русскоязычного текста. Анализируется группа устойчивых словосочетаний, состоящих из наречия или местоимения и глагола. Показано, что в зависимости от правого контекста такое словосочетание может быть или наречным/местоименным оборотом, или просто сочетанием двух слов. Сформулированы формальные правила разрешения омонимии, которые могут быть использованы в качестве алгоритмов работы парсеров.

Ключевые слова: компьютерный анализ текста, морфология, словосочетания, наречные обороты, местоименные обороты, омонимия

ВВЕДЕНИЕ

В современном мире ежедневно появляется огромное количество информации, в том числе научной: учебники, монографии, статьи и т. д. Обработка такой информации имеет свою специфику: во-первых, объемы документов в определенной предметной области гораздо меньше, например, объемов новостных сообщений, что вносит ограничения в применение таких статистических методов, как латентное размещение Дирихле, нейронные сети и т. п. [1]; во-вторых, научные документы обладают характерной внутренней структурой [2]; в-третьих, эти документы «низкоконтрастны», т.е. лексически близки друг к другу [3]; в-четвертых, они, с одной стороны, должны достаточно полно охватывать релевантные информационные источники, с другой — удовлетворять требованиям антиплагиата.

Эти особенности вызывают потребность применять для анализа более точные модели, чем просто «мешок слов» (*bag-of-words*), в котором порядок слов в тексте не учитывается вообще [4]. Одни из таких методов — построение дерева подчинения [5]. Однако на этом пути возникают свои сложности, связанные, прежде всего, с высокой степенью омонимичности русскоязычных текстов. В настоящей работе анализируются некоторые возможности снятия омонимии, основанные на анализе устойчивых словосочетаний.

ПОСТАНОВКА ЗАДАЧИ И МЕТОДИКА ИССЛЕДОВАНИЯ

В качестве средства анализа текстов нами применяется парсер *SemSin* [6], осуществляющий глубокий синтаксический анализ русскоязычных текстов. Он использует расширенный вариант словаря В.А. Тузова [7], объемом свыше 175 тыс. слов. Анализ текста осуществляется под управлением продукционных правил [8].

Все устойчивые словосочетания можно разделить на три группы: неизменяемые, с изменяемым первым словом и полностью изменяемые [9]. Большая часть неизменяемых словосочетаний является оборотами, выполняющими в предложении функции одного слова, поэтому при синтаксическом разборе такое словосочетание рассматривается как один узел (токен) и не разбивается на отдельные словоформы. Наиболее полные списки оборотов приведены в Национальном корпусе русского языка (НКРЯ) [10]. Достаточно много оборотов приведено также у С.А. Кузнецова [11]. Однако имеющиеся в свободном доступе системы анализа текстов, как правило, не выявляют устойчивые словосочетания. Исключением является ЭТАП-4 [12] — последняя версия парсера ЭТАП [13], в которой частично определяются обороты.

Ситуация осложняется тем, что одно и то же сочетание нескольких слов в зависимости от контекста может быть оборотом, а может и не быть им. Многие словосочетания такого рода рассмотрены Р.П. Рогожниковой [14], которая отмечает возможность их использования в качестве свободных сочетаний слов, омонимичных оборотам. Сегодня в нашем словаре более 1700 неизменяемых оборотов, из которых около 190 являются такими омонимами.

* Статья подготовлена по результатам исследований по теме НИР: «Исследование согласованного развития городов, регионов и природной среды методами математического моделирования, направленное на устойчивое развитие городской среды, промышленности и транспортной инфраструктуры с использованием методов анализа данных», номер Г.Р. АААА-А19-119021390164-1.

Следует отметить, что перечисленные словари написаны «для людей». Содержащиеся в них примеры интуитивно понятны. Но в них нет четких правил, обеспечивающих работу компьютерных программ. В НКРЯ приведены частоты встречаемости оборотов, но нет информации о частотности появления омонимии, вследствие чего непонятно, нужно ли тратить усилия на борьбу с ней.

Особый интерес представляют случаи, когда одно из слов, входящих в состав устойчивого словосочетания, является глаголом или предикатом. Дело в том, что неверное определение оборота в этой ситуации приводит к полному развалу дерева подчинения и невозможности дальнейшего корректного извлечения информации. Поэтому необходимо сформулировать правила, позволяющие с удовлетворительной точностью решать задачу снятия омонимии для выбранной нами группы словосочетаний. Ограничимся глагольными словоформами *попало*, *придется*, *хотите*, *хочешь* и предикатом *угодно*.

Анализ проводился на материале предложений из НКРЯ и имеющегося в нашем распоряжении массива текстов (объемом около 50 млн слов), составленного из ряда повестей, новостных и спортивных статей. Для каждого словосочетания брались все предложения из основного корпуса НКРЯ, если исследуемое словосочетание встречалось слишком редко, то использовался также газетный корпус. К найденным предложениям добавлялись предложения из нашего массива текстов. Отобранный материал подвергся дополнительной фильтрации, чтобы исключить случаи разрыва словосочетания знаками препинания (тогда это точно не оборот).

Полученные предложения были предварительно проанализированы с целью обнаружения омонимии оборот/не оборот. Предложения, в которых словосочетание не является оборотом, разделялись на группы по синтаксическим признакам (прежде всего, по правому контексту) с помощью специальных команд парсера *SemSin* [15].

Например, для словосочетания *как угодно* было отобрано 1196 предложений. С помощью парсера было выделено 5 групп предложений с возможностью омонимии:

1) 8 предложений с предлогом, стоящим после словосочетания (*Подарки на другой день свадьбы, как угодно с вашей стороны*)¹;

2) 19 предложений со словами *было* или *будет*, стоящими после словосочетания (*Малограмотные комиссары, выпускники Института Красной Профессуры пришли в науку и навели там порядок так, как угодно было коммунистам*);

3) 46 предложений с существительным, стоящим после словосочетания (*Но как угодно Всевышнему!*);

4) 56 предложений с существительным или местоимением в дательном падеже, стоящим справа не-

посредственно после словосочетания или через одно слово (*Впрочем, как угодно моему мужу...*);

5) 147 предложений с инфинитивом в правом контексте (*Конечно, я в университете не был, но живу, как дай бог всякому, и ничего худого за собой не знаю, а впрочем, кому как угодно судить*).

Сформированные группы подвергались окончательному анализу с целью определения частоты встречаемости оборотов и их омонимов.

АНАЛИЗ НАРЕЧНЫХ СЛОВСОЧЕТАНИЙ

Состоящие из двух слов обороты, в которых первое слово является наречием *где*, *как*, *когда*, *куда*, *откуда*, *сколь*, *сколько*, а второе – глаголом или предикатом *угодно*, приведенные в НКРЯ и в книге Рогожниковой [14], сведены в табл. 1. Буквой «О» обозначено наличие у данного словосочетания омонимии (по Рогожниковой). Как будет показано ниже, большинство приведенных оборотов в действительности также являются омонимичными.

Отметим, что списки оборотов совпадают не полностью и не охватывают всех случаев словоупотребления в языке.

Таблица 1

Обороты с наречием и глаголом

НКРЯ	Рогожникова
где попало	где попало
где придется	где придется О
где угодно	где угодно
	где хотите О
	где хочешь О
как попало	как попало О
как придется	как придется
как угодно	как угодно
	как хочешь
	как хотите
когда придется	когда придется
когда угодно	когда угодно
куда попало	куда попало
куда придется	куда придется
куда угодно	куда угодно
откуда попало	
откуда угодно	
сколь угодно	сколь угодно
сколько угодно	сколько угодно
сколько хочешь	сколько хочешь

Из этих наборов слов можно образовать матрицу словосочетаний, представленную в табл. 2, в которой по каждому типу приведено количество предложений, отобранных для анализа, что можно рассматривать как некоторую меру частоты их встречаемости в русском языке. Рассмотрим подробнее все возможные словосочетания такого рода с целью определить или уточнить их омонимию.

¹ Здесь и далее жирным шрифтом выделено собственно словосочетание, слово, по которому предложение отнесено к данной группе, подчеркнуто.

Наречные словосочетания

	попало	придется	удобно	хотите	хочешь
Где	427	258	695	107	202
Как	702	291	1196	1600	1422
Когда	17	178	267	114	179
Куда	343	77	762	175	212
Откуда	14	5	103	8	9
Сколько	0	0	384	0	19
Сколько	4	147	1829	212	742

1. Словосочетания, оканчивающиеся на слово «попало»

Словосочетание этой группы может быть оборотом в функции наречия (за исключением оборота *сколько попало*), а может быть и свободным сочетанием слов, омонимичным обороту, если справа от него расположен предлог или существительное (местоимение) среднего рода в именительном падеже.

Где попало

- **Наречный оборот:** *Фигуры перемешаны: белые и чёрные стоят где попало*².

- **Свободное словосочетание**

- о Контактнo справа стоит предлог (0,5%)³: *Где попало в тебя?*

Как попало

- **Наречный оборот:** *Все жили как попало.*

- **Свободное словосочетание**

- о Контактнo справа от него находится существительное/местоимение (0,7%): *И совершенно неизвестно, как попало оружие к Алтынову-Задонскому.*

- о Контактнo справа от него находится предлог (0,85%): *Как попало на фотографию лицо преступника, отбывающего срок?*

Когда попало

- **Наречный оборот:** *Зато пиццуженец ест что попало и когда попало.*

- **Свободное словосочетание**

- о Контактнo справа от него находится предлог (5,9%): *...чужое имя делается известным тогда только, когда попало в какой-нибудь водоворот происшествий.*

Куда попало

- **Наречный оборот:** *Вырвется она за его границы и уйдет куда попало.*

- **Свободное словосочетание**

- о Контактнo справа от него находится существительное/местоимение (0,6%): *Так, число представляет собой класс эквивалентности, куда попало множество знаков зодиака.*

- о Контактнo справа от него находится предлог (0,3%): *Странным является слово на'божный: сколько мне известно, его в церковнослав. нет, а есть оно только в польском, куда попало из чешского...*

Откуда попало

- **Наречный оборот:** *Артисты были набраны откуда попало...*

- **Свободное словосочетание**

- о Контактнo справа от него находится предлог (7,1%): *... «Сказание о снах Шахаиши», видимо, через персидскую версию перешло при посредстве богومیлов в Византию, затем в Сербию, откуда попало в XIII-XIV вв. на Русь.*

Сколько попало

- **Оборот в роли наречия или числительного:** *Стопорили фургон на сколько попало времени, сами тишком уходили.*

- **Свободное словосочетание**

- о Контактнo справа от него находится предлог (25%): *... Сколько было освобождено? Сколько попало в больницы?*

2. Словосочетания, оканчивающиеся на слово «придется»

Словосочетание этой группы может быть оборотом в функции наречия (за исключением оборота *сколько придется*), а может быть и свободным сочетанием слов, омонимичным обороту, если справа от него расположен инфинитив (в отдельных случаях может быть и предлог).

Где придется.

- **Наречный оборот:** *А то он тут бежит где придется.*

- **Свободное словосочетание**

- о Справа инфинитив (18%): *Никто из солдат и офицеров не знает, где придется действовать.*

Как придется.

- **Наречный оборот:** *Лара вежливо улыбалась, но про себя явно решила делать как придется.*

- **Свободное словосочетание**

- о Справа инфинитив (18%): *Он думает об экспедиции, и о том, как придется теперь действовать.*

Когда придется.

- **Наречный оборот:** *Бегай сколько хочешь, но чужой у соседей, ешь чего придется и когда придется.*

- **Свободное словосочетание**

- о Контактнo справа предлог (3,9%): *Об остальных деятелях Академии художеств особо теперь говорить не буду: когда придется к слову, поману о них.*

- о Справа инфинитив (69%): *Ведь наступит день, когда придется их отдавать!*

² Здесь и далее жирным шрифтом выделены только обороты.

³ Здесь и далее цифры в скобках обозначают частоту встречаемости данной конструкции.

Куда придется.

- **Наречный оборот:** *А Кирилл поднял ноги и со всей силы пнул тварь куда придётся.*
- **Свободное словосочетание**
 - Контактнo справа предлог (5,2%): *Он бил их с яростью крайнего, неутоляемого отчаянья, не глядя, куда придется удар...*
 - Справа инфинитив (9,1%): *Но все же не знаешь, куда придется ложиться, до самого дня, когда ляжешь.*

Откуда придется

- **Наречный оборот:** *И вот писатель, как пчела, берет мед откуда придётся.*
- **Свободное словосочетание**
 - Справа инфинитив (40%): *Приехали доктор с фельдшером, тоже поедут с нами, хотят посмотреть, откуда придется доставлять раненых.*

Сколько придется.

- **Оборот в функции наречия или числительно-го:** *Сколько придется, столько и проживете, — сказал Коминт.*
- **Свободное словосочетание**
 - Справа расположен инфинитив (86%): *Сколько придется ждать, пока закончат все формальности?*
 - Контактнo справа находится предлог, местоимение-существительное, местоимение-прилагательное или существительное (4,8%): *Стоит ли удивляться тому, что солдаты буквально обезумели, подсчитав, сколько придется на долю каждого?*
 - Контактнo справа находится предлог, местоимение-существительное, местоимение-прилагательное или существительное (3,4%): *Один Симон (известный служитель в ресторане) не принимал участия в общем энтузиазме и, по-видимому, рассчитывал, сколько придется ему на водку.*

3. Словосочетания, оканчивающиеся на слово «угодно»

Словосочетание этой группы в большинстве случаев является оборотом в функции наречия (за исключением оборота *сколько угодно*), а может быть и свободным сочетанием слов, омонимичным обороту, если справа стоит существительное или местоимение в дательном падеже, а также глагол *быть*.

Где угодно.

- **Наречный оборот:** *Уехать не уедешь, зато каждый день можно где угодно пропасть.*
- **Свободное словосочетание.**
 - Контактнo справа существительное в дательном падеже (0,1%): *Все увидели, что дар чудотворения, украшавший ее проповедников, не есть какая-либо естественная неведомая сила, которая может быть подчинена человеческому произволу и обращена в предмет купли, но живая сила Божия, которая оказывает свое действие только там, где угодно Промыслу.*

○ Контактнo справа слово *будет* (0,1%): *Со всей охотой, — отвечал Потемкин, — желаю я исполнить волю вашего сиятельства и с радостью останусь, где угодно будет меня определить.*

Как угодно.

- **Наречный оборот:** *Ты можешь одеваться как угодно.*
- **Свободное словосочетание.**
 - Контактнo справа существительное или местоимение в дательном падеже: *Пусть наша любовь увенчается так, как угодно богам.*
 - Контактнo справа слово *было* или *будет* (1,5%): *Понаехали греческие и иные из-за моря патриархи. Ну, вестимое дело, и порешили как угодно было царю да боярам.*

Когда угодно.

- **Наречный оборот:** *Доступ к личным цифровым данным когда угодно и где угодно.*
- **Свободное словосочетание.**
 - Контактнo справа существительное или местоимение в дательном падеже (3%): *Когда угодно тебе, равняем тебя с собою!*
 - Контактнo справа слово *было* или *будет* (3,4%): *Когда мы приблизились к дверям, нас остановили и долгое время мы ждали, когда угодно будет доложить большевистскому начальству о нашем приходе.*
 - Правее находится инфинитив (4,1%): *Так вам когда угодно обедать?*

Куда угодно.

- **Наречный оборот:** *Алла очень захотела уйти из дома — куда угодно, к кому угодно.*
- **Свободное словосочетание.**
 - Контактнo справа существительное или местоимение в дательном падеже (0,5%): *...назначена на выписку из больницы; а потому, куда угодно ей будет отправиться?*
 - Контактнo справа слово *будет* (0,3%): *Если бы я не получил предписания, — отвечал полицеймейстер, — то прежде всего за дерзость наказал бы тебя, но теперь велю отправить, куда угодно будет г. Зембину.*

Откуда угодно (всегда только оборот).

- **Наречный оборот:** *Я могу убежать откуда угодно и куда угодно.*

Сколь угодно (всегда только оборот).

- **Наречный оборот:** *Российский суд может, конечно, принять любое сколь угодно абсурдное решение.*
- **Оборот в роли числительного:** *Здесь в любой момент, любого игрока можно заменить сколь угодно раз.*

Сколько угодно.

- **Наречный оборот:** *Представитель месткома сказал: — Есть можете сколько угодно.*
- **Оборот в роли числительного:** *Я не сомневался, что меня хватит на сколько угодно диссертаций...*
- **Свободное словосочетание.**
 - Контактнo справа существительное или местоимение в дательном падеже (0,3%): *Словом, со свежей рыбой опять трудности, зато рыбных консервов сколько угодно душе.*

4. Словосочетания, оканчивающиеся на слово «хотите»

Словосочетание этой группы может быть оборотом в функции наречия (за исключением оборотов *как хотите* и *сколько хотите*), а может быть и свободным сочетанием слов, омонимичным обороту, если справа от него расположен инфинитив.

Где хотите.

- **Наречный оборот:** *Подождите, пока книгу издадут, а там уж выступайте где хотите.*

- **Свободное словосочетание.**

- Правее находится инфинитив (7,5%): *Где хотите обедать: у себя или со мной?*

- Контактно справа стоит запятая и наречие там (5,6%): *Где хотите, там и доставайте валы.*

Как хотите.

- **Наречный оборот:** *Впрочем теперь некому об этом подумать, и вы можете делать как хотите.*

- **Вводный оборот:** *Нет, как хотите, это я не согласен считать за сад.*

- **Свободное словосочетание.**

- Правее находится инфинитив (0,6%): *И совсем не такие сильные, как хотите казаться.*

- Контактно справа стоит запятая и союз (24%): *Нет, как хотите, а это невозможно.*

Когда хотите.

- **Наречный оборот:** *Делайте снимки когда хотите, чтобы сохранить набор параметров коррекции.*

- **Свободное словосочетание.**

- Правее находится инфинитив (21%): *Когда хотите венчаться, Дмитрий Сергеевич?*

Куда хотите.

- **Наречный оборот:** *Убирайте их куда хотите и как хотите.*

- **Свободное словосочетание.**

- Правее находится инфинитив (6,3%): *Куда хотите вести меня, Нуллус?*

Откуда хотите (всегда только оборот).

- **Наречный оборот:** *Возьмите примеры откуда хотите и сколько хотите.*

Сколько хотите.

- **Наречный оборот:** *Скажите, умоляю, заплачу сколько хотите!*

- **Оборот в роли числительного:** *Я дам Вам сколько хотите времени...*

- **Свободное словосочетание.**

- Правее находится инфинитив (2,4%): *Нужно решить, сколько хотите накопить и за какой срок.*

5. Словосочетания, оканчивающиеся на слово «хочешь»

Словосочетание этой группы может быть оборотом в функции наречия (за исключением оборотов *как хотите*, *сколь хотите* и *сколько хотите*), а может быть и свободным сочетанием слов, омонимичным обороту, если справа от него расположен инфинитив.

Где хочешь.

- **Наречный оборот:** *С такими деньгами где хочешь примут.*

- **Свободное словосочетание.**

- Правее находится инфинитив (9,4%): *Очень подробно представь то место, где хочешь очутиться.*

- Контактно справа стоит запятая и наречие там (4%): *А где хочешь, там и доставай.*

Как хочешь.

- **Наречный оборот:** *Как хочешь с ней разделивайся ты, а я бегу в кусты.*

- **Вводный оборот:** *Короче, как хочешь, хоть умри!*

- **Свободное словосочетание.**

- Правее находится инфинитив (0,6%): *А ты-то как хочешь спастись?*

- Контактно справа стоит запятая и союз (18%): *Как хочешь, но без нее, нарушь ее план!*

Когда хочешь.

- **Наречный оборот:** *Приходи обедать когда хочешь.*

- **Свободное словосочетание.**

- Правее находится инфинитив (41%): *Учись прилежно, когда хочешь быть благополучным.*

- Контактно справа стоит существительное в родительном падеже (1,1%): *...я понимаю ваше огорчение, когда хочешь парня, а тут одни девченки.*

Куда хочешь.

- **Наречный оборот:** *Но не продавай, куда хочешь и за сколько хочешь!*

- **Свободное словосочетание.**

- Правее находится инфинитив (3,8%): *Ты куда хочешь поехать?*

Откуда хочешь (всегда только оборот).

- **Наречный оборот:** *Хорошо, когда сад открытый, входи откуда хочешь.*

Сколь хочешь (всегда только оборот).

- **Наречный оборот:** *Плоти сколь хочешь – их кормить не надо!*

- **Оборот в роли числительного:** *Сколь хочешь годов опы по книгам накинута.*

Сколько хочешь.

- **Наречный оборот:** *Их спокойно можно убрать или добавить еще сколько хочешь.*

- **Оборот в роли числительного:** *Можно заказать сколько хочешь клей, зависит от объемов производства...*

- **Свободное словосочетание.**

- Правее находится инфинитив (один случай): *Говори, сколько хочешь дать!*

- Контактно справа стоит предлог за (0,8%): *Сколько хочешь за пару понятых лимунистых?*

АНАЛИЗ МЕСТОИМЕННЫХ СЛОВСОЧЕТАНИЙ

Определенный интерес представляет группа словосочетаний, состоящая из местоимений (и местоименных прилагательных) и рассмотренных выше глаголов и предиката *удобно*. К сожалению, в НКРЯ [3] обороты такого рода отсутствуют, однако в книге Рогожниковой [12] они приведены. К ним относятся следующие словосочетания: *какой попало*, *какой придется*, *какой угодно*, *какой хотите*, *какой хочешь*, *кто попало*, *кто угодно*, *кто хочешь*, *что попало* (О), *что придется* и *что угодно* (О)⁴.

⁴ Буквой «О» здесь обозначено наличие омонимии по Рогожниковой.

Местоименные словосочетания

	попало	придется	удобно	хотите	Хочешь
Какой	47	37	1448	179	259
Кто	92	230	1767	181	634
Чей	1	0	48	0	4
Что	987	984	2205	1398	1656

Аналогично наречным словосочетаниям, рассмотренным выше, можно построить некоторую матрицу словосочетаний из местоимений *кто, что*, местоименных прилагательных *какой, чей* и имеющегося набора глаголов (см. табл. 3). В этой таблице по каждому типу приведено количество предложений, отобранных для анализа, что можно рассматривать как некоторую меру частоты их встречаемости в русском языке. Следует отметить, что все эти словосочетания характеризуются тем, что в них первое слово является изменяемым в отличие от наречных словосочетаний. Рассмотрим подробнее все возможные словосочетания такого рода с целью определить или уточнить их омонимию.

1. Словосочетания, оканчивающиеся на слово «**попало**»

Словосочетание этой группы может быть оборотом в роли местоимения-прилагательного или местоимения-существительного, а может быть и свободным сочетанием слов, омонимичным обороту, если справа от него расположен определенный предлог.

Какой попало (всегда только оборот).

- **Оборот в роли местоимения-прилагательного:** *И анекдот какой попало здесь не расскажешь.*

Кто попало (всегда только оборот).

- **Оборот в роли местоимения-существительного:** *А зачем сейчас спрашивать советы у кого попало?*

Чей попало (всегда только оборот).

- **Оборот в роли местоимения-прилагательного:** *А, черт, возьми чью попало машину и отвезу.*

Что попало.

- **Оборот в роли местоимения-существительного:** *Вооружайтесь чем попало и выходите на отлов бандитов...*

- **Свободное словосочетание** (в этом случае встречается только словоформа *что*).

- о Контекстно справа стоит один из предлогов *в, к, между, на, под* (4,2%): *Но не так просто найти то, что попало на дно двадцать или двадцать пять веков назад.*

- о Контактно справа стоит местоимение-существительное в дательном падеже (0,5%): *Первое, что попало ему в голову, это было...*

- о Контактно справа стоит словосочетание *под руку* (1,3%): *Я схватил первое, что попало под руку.*

2. Словосочетания, оканчивающиеся на слово «**придется**»

Словосочетание этой группы может быть оборотом в роли местоимения-прилагательного или местоимения-существительного, а может быть и свобод-

ным сочетанием слов, омонимичным обороту, если справа от него расположен инфинитив (в отдельных случаях может быть или предлог, или существительное/местоимение в дательном падеже).

Какой придется.

- **Оборот в роли местоимения-прилагательного:** *Завтра он убежит отсюда вниз по реке на каком придется пароходе.*

- **Свободное словосочетание.**

- о Справа от него находится инфинитив (49%): *Он мечтал о той роли, какую придется ему играть в Европе.*

- о Контактно справа стоит местоимение или существительное в дательном падеже (5,4%): *Какой придется пассажирам по душе?*

Кто придется.

- **Оборот в роли местоимения-существительного:** *Вильнула хвостом и живет с кем придется.*

- **Свободное словосочетание.**

- о Справа от него находится инфинитив (70%): *Тем, кому придется побывать в Вязниках, можно посоветовать следующее...*

- о Контактно справа стоит предлог (2,2%): *Вот дама и предлагает желающим взглянуть на обоих и взять того, кто придется по душе.*

- о Контактно справа стоит местоимение в дательном падеже (0,4%): *– Да, бояр, – продолжала твердым голосом кукона, – я могу отдать себя тому, кто придется мне по сердцу, могу сделаться его женою или невольницей – все равно!*

Что придется.

- **Оборот в роли местоимения-существительного:** *Я решил год готовиться, а пока подрабатывать чем придется на киностудии.*

- **Свободное словосочетание.**

- о Справа от него находится инфинитив (77%): *Второе, с чем придется столкнуться в стенах РБЦ, – пронизывающий до костей холод.*

- о Словосочетание начинается со словоформы *что* и контактно справа стоит один из предлогов *в, до, за, к, на, от, по, с* (1,5%): *А в "Призраках" есть то, что придется по душе любому зрителю.*

- о Словосочетание начинается со словоформы *что* и контактно справа стоит местоимение-существительное (1,3%): *Европеец роется в ветоши, достает, что придется ему впору, обновляет, хозяйничает...*

- о Перед словосочетанием стоит форма глагола и запятая (1,7%): *Знаешь, что придется несладко, и тянешь, сколько возможно.*

3. Словосочетания, оканчивающиеся на слово «**угодно**»

Словосочетание этой группы может быть оборотом местоимения-прилагательного или местоимения-существительного, а может быть и свободным сочетанием слов, омонимичным обороту, если справа стоит существительное или местоимение в дательном падеже

Какой угодно.

- **Оборот в роли местоимения-прилагательного:** *Всего этого в бадье намешано в каких угодно пропорциях.*

- **Свободное словосочетание.**

- Контактнo справа существительное или местоимение в дательном падеже (0,2%): *То есть, Дума должна правительство сделать такое, какое угодно людям, положившим деньги в ее тумбочку.*

- Контактнo справа слово было (0,3%): *Если бы они попали к Колчаку, то они так же жестоко и бессмысленно исполняли бы то дело, какое угодно было Колчаку?*

Кто угодно.

- **Оборот в роли местоимения-существительного:** *С кем угодно могло такое случиться, только не с Пашкой.*

- **Свободное словосочетание.**

- Контактнo справа стоит инфинитив (1 случай): *Господа! – сказал он медленно, освобождая свои руки, – кому угодно заплатить за меня двадцать червонцев?*

Чей угодно (всегда только оборот).

- **Оборот в роли местоимения-прилагательного:** *У кармана нашелся хозяин, а след мог быть чей угодно.*

Что угодно.

- **Оборот в роли местоимения-существительного:** *Если есть наличные деньги, можно заказать всё что угодно.*

- **Свободное словосочетание.**

- Контактнo справа стоит существительное а дательном падеже (0,2%): *Он сохранил старую норму: "Что угодно государю, имеет силу закона".*

- Справа от словосочетания стоит инфинитив, но в предложении отсутствуют словоформы глаголов мочь и уметь (0,2%): *...а доволен был своим жребием и тем, что угодно было учинить с мною Промыслу Господню.*

4. Словосочетания, оканчивающиеся на слово «**хотите**»

Словосочетание этой группы может быть оборотом местоимения-прилагательного или местоимения-существительного, а может быть и свободным сочетанием слов, омонимичным обороту, если справа от него расположен инфинитив или за ним стоит вопросительный знак.

Какой хотите.

- **Оборот в роли местоимения-прилагательного:** *В Петербурге дочери священника Чудакова ка-кой хотите стол заставят вертеться.*

- **Свободное словосочетание.**

- Правее такого словосочетания (но не обязательно контактнo) имеется инфинитив (4,5%): *"Вы какой хотите от меня ответ услышать?" - улыбнулся Владимир Путин.*

- Правее словосочетания находится знак вопроса (5%): *Вы какой хотите суржик?*

Кто хотите.

- **Оборот в роли местоимения-существительного:** *- Вы можете писать о ком хотите, – заметила на это Вера.*

- **Свободное словосочетание.**

- Правее такого словосочетания имеется инфинитив (8,8%): *" Кому хотите вы зло сделать?"*

- Контактнo справа стоит вопросительный знак (2,8%): *Вы кого хотите?*

Что хотите.

- **Оборот в роли местоимения-существительного:** *Ешьте эту кашу и рассказывайте – о чём хотите.*

- **Свободное словосочетание.**

- Правее такого словосочетания имеется инфинитив (9%): *" Вы уверены, что хотите удалить приложение?"*

- Контактнo справа стоит вопросительный знак (4,1%): *– А вы чего хотите? – спросила Афра.*

5. Словосочетания, оканчивающиеся на слово «**хочешь**»

Словосочетание этой группы может быть оборотом местоимения-прилагательного или местоимения-существительного, а может быть и свободным сочетанием слов, омонимичным обороту, если справа от него расположен инфинитив или за ним стоит вопросительный знак.

Какой хочешь.

- **Оборот в роли местоимения-прилагательного:** *Во время гульбы играли песни какие хочешь.*

- **Свободное словосочетание.**

- Правее такого словосочетания (но не обязательно контактнo) имеется инфинитив (4,2%): *Стань таким, каким хочешь видеть других.*

- Правее словосочетания находится знак вопроса (3,5%): *Тот тут же встал: - Какую хочешь фору?*

Кто хочешь.

- **Оборот в роли местоимения-существительного:** *Таким дуэтом кого хочешь можно нагнуть!*

- **Свободное словосочетание.**

- Правее такого словосочетания имеется инфинитив, но отсутствуют словоформы глаголов мочь и уметь (6%): *Затем ты должен написать того, кому хочешь дать команду.*

Чей хочешь (всегда только оборот).

- **Оборот в роли местоимения-прилагательного:** *Пиши меня чьим хочешь рабом, а я, друг, свой стал.*

Что хочешь.

- **Оборот в роли местоимения-существительного:** *И потом, королева может с собой что хочешь сделать.*

- **Свободное словосочетание.**

- Правее такого словосочетания имеется инфинитив, но отсутствуют словоформы глаголов мочь и уметь (8,6%): *Ты не знаешь, во что хочешь вмешаться.*

- Контактнo справа стоит вопросительный знак (4,4%): *– Ну, а ты чего хочешь?*

Омонимия словосочетаний наречий с глаголами

	попало		придется		угодно		хотите		хочешь	
	всего	не оборотов, %	всего	не оборотов, %	всего	не оборотов, %	всего	не оборотов, %	всего	не оборотов, %
Где	427	0,5	258	18	695	0,3	107	13	202	13
Как	702	1,6	291	18	1196	4,3	1600	25	1422	18
Когда	17	5,9	178	63	267	10,5	114	21	179	42
Куда	343	0,9	77	14	762	0,8	175	6,3	212	3,8
Откуда	14	7,1	5	40	103	–	8	–	9	–
Сколь	0	–	0	–	384	–	0	–	19	–
Сколько	4	25	147	94	1829	0,3	212	3,3	742	0,9

Таблица 5

Омонимия словосочетаний местоимений с глаголами

	попало		придется		угодно		хотите		хочешь	
	всего	не оборотов, %	всего	не оборотов, %	всего	не оборотов, %	всего	не оборотов, %	всего	не оборотов, %
Какой	47	–	37	54	1448	0,5	179	9,5	259	7,7
Кто	92	–	230	73	1767	–	181	12	634	6
Чей	1	–	0	–	48	–	0	–	4	–
Что	987	6	984	78	2205	0,4	1398	13	1656	13

ЗАКЛЮЧЕНИЕ

В результате нашего исследования для каждого из 35 наречных и 20 местоименных словосочетаний были определены частоты наличия оборотов и их омонимии, которые сведены в табл. 4 и 5, где приводятся количества словосочетаний данного типа и случаев, когда эти словосочетания не являются оборотами (в процентах).

Кроме того, сформулированы формальные правила разрешения омонимии, которые . могут быть использованы в качестве алгоритмов работы парсеров для улучшения точности разбора предложений со сложной синтаксической структурой.

Предположим, что точность анализа парсера 95%. Тогда если для конкретного словосочетания омонимичность ниже 5%, то его следует всегда считать оборотом и омонимичностью можно пренебречь. Как видно из табл. 4 и 5, из наречных словосочетаний таких насчитывается более 65%, в то время как для местоименных словосочетаний их ровно половина.

СПИСОК ЛИТЕРАТУРЫ

1. Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey // *Proceedings of Frontiers of Computer Science in China*, 2010. – Vol. 4, Issue 2 – P. 280–301.
2. Котюрова М.П. Стилистика научной речи. – М.: Академия, 2010.
3. Авдеева Н.А., Боярский К.К., Гусарова Н.Ф., Добренко Н.В., Каневский Е.А. Исследование специфики применения алгоритмов тематической сегментации для научных текстов. // *Аналитика и управление данными в областях с интенсивным использованием данных. XVII Международная конференция DAMDID/RCDL'2015*. – М.: НИЯУ МИФИ, 2015. – С. 181–189.
4. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. 2-е изд., перераб. и доп. – СПб: БХВ-Петербург, 2007.
5. Тестелец Я.Г. Введение в общий синтаксис. – М.: РГТУ, 2001.
6. Боярский К.К., Каневский Е.А. Семантико-синтаксический парсер SemSin // *Научно-технический вестник информационных технологий, механики и оптики*. – 2015. – Т. 15, № 5. – С. 869–876.
7. Тузов В.А. Компьютерная семантика русского языка. – СПб: Изд-во С.-Петерб. ун-та, 2004.
8. Боярский К.К., Каневский Е.А. Система продукционных правил для построения синтаксического дерева предложения. Прикладна лінгвістика та лінгвістичні технології: MegaLing-2011. – Киев: Довіра, 2012. – С. 73–80.
9. Боярский К.К., Каневский Е.А. Словосочетания, эквивалентные слову // *Компьютерная лингвистика и вычислительная онтология: сборник научных статей. Труды XVIII объеди-*

ненной научной конференции «Интернет и современное общество» (IMS-2015), Санкт-Петербург, 23–25 июня 2015 г. – СПб: Университет ИТМО, 2015. – С. 55–66.

10. Национальный корпус русского языка. – URL: <http://ruscorpora.ru>.
11. Кузнецов С.А. Большой толковый словарь русского языка. – СПб: Норинт, 1998. – URL: <http://proling.iitp.ru/etap4>.
12. Iomdin L., Petrochenkov V., Sizov V., Tsinman L. Etap parser: state of the art // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог», Бекасово 30 мая–3 июня 2012 г. Том 2. Выпуск 11 (18). – М.: Изд-во РГГУ, 2012. – С. 117–131.
13. Рогожникова Р.П. Толковый словарь сочетаний, эквивалентных слову. – М.: Астрель, АСТ, 2003.
14. Боярский К.К., Каневский Е.А. Язык правил для построения синтаксического дерева // Интернет и современное общество: Материалы

XIV Всероссийской объединенной конференции «Интернет и современное общество». – СПб: МультиПроджектСистемСервис, 2011. – С. 233–237.

Материал поступил в редакцию 18.09.19.

Сведения об авторах

БОЯРСКИЙ Кирилл Кириллович – кандидат физико-математических наук, доцент, Университет ИТМО, доцент, Санкт-Петербург
e-mail: boyarin9@yandex.ru

КАНЕВСКИЙ Евгений Александрович – кандидат технических наук, старший научный сотрудник, Институт проблем региональной экономики РАН, ведущий научный сотрудник, Санкт-Петербург
e-mail: eak300@mail.ru

УДК 025.4.06:004.42(470)

Ю. В. Смирнов

О классификаторе программ для электронных вычислительных машин и баз данных

Описывается структура и назначение классификатора программ для электронных вычислительных машин и баз данных, используемого для систематического поиска в Едином реестре российских программ для электронных вычислительных машин и баз данных. При анализе поисковых возможностей этого ресурса выявлены основные достоинства и недостатки в его организации. Предлагается ряд решений для повышения релевантности и пертинентности систематического поиска.

Ключевые слова: импортозамещение, классификатор программ для электронных вычислительных машин и баз данных, единый реестр российских программ для электронных вычислительных машин и баз данных, систематический поиск

Классификатор программ для электронных вычислительных машин и баз данных (далее – Классификатор) был создан на основании приказов № 621 и № 622 Министерства связи и массовых коммуникаций РФ от 31 декабря 2015 г., а в апреле 2016 г. подвергся изменениям согласно приказу Минкомсвязи №134 от 01.04.2016 г.¹

Основная цель Классификатора – обеспечение классификационного поиска программного обеспечения (ПО) и баз данных (БД) в Едином реестре российских программ для электронных вычислительных машин и баз данных (далее Росреестр ПО) [1]. Подробно поисковые возможности этого ресурса рассматриваются в [2]. Росреестр ПО также, как и Классификатор, был создан и запущен в 2016 г. для обеспечения информационной безопасности государ-

ственных и муниципальных учреждений России в условиях санкционного давления стран Запада.

В Федеральном законе от 29 июня 2015 г. № 188-ФЗ² перечислены критерии, которым должны соответствовать программное обеспечение и базы данных, включаемые в Реестр ПО:

1) исключительное право на весь его срок и на территории всего мира принадлежит РФ, субъекту РФ, муниципальному образованию РФ, любой российской организации или гражданину РФ, доля которых в участии превышает 50% и при условии, что иностранные лица не могут влиять на принимаемые решения;

2) программное обеспечение должно быть введено в оборот на территории РФ и доступно в свободной продаже;

3) сумма выплат по лицензионным и иным договорам, предусматривающим предоставление прав иностранным лицам и контролируемым ими российским организациям не должна превышать 30% от годовой выручки за данное ПО;

4) программное обеспечение не должно иметь доступа к государственной тайне и не предоставлять сведения о ней.

¹ Приказ Минкомсвязи России от 31.12.2015 № 621 «Об утверждении классификатора программ для электронных вычислительных машин и баз данных». – URL: <http://base.garant.ru/71338148/> ;
Приказ Минкомсвязи России от 31.12.2015 № 622 «Об утверждении правил применения классификатора программ для электронных вычислительных машин и баз данных». – URL: <https://reestr.minsvyaz.ru/upload/iblock/e2a/622.pdf> ;
Приказ Минкомсвязи России от 01.04.2016 № 134 «О внесении изменений в Классификатор программ для электронных вычислительных машин и баз данных, утвержденный приказом Министерства связи и массовых коммуникаций Российской Федерации от 31.12.2015 № 621». – URL: <https://reestr.minsvyaz.ru/upload/iblock/efa/Классификатор.pdf>.

² Федеральный закон от 29 июня 2015 г. № 188-ФЗ «О внесении изменений в Федеральный закон "Об информации, информационных технологиях и о защите информации" и статью 14 Федерального закона "О контрактной системе в сфере закупок товаров, работ, услуг для обеспечения государственных и муниципальных нужд"». – URL: <http://www.kremlin.ru/acts/bank/39838>.

Отметим, что с 2017 г. разрешены закупки не только российского, но и программного обеспечения из других стран Евразийского экономического союза (ЕАЭС): Армения, Белоруссия, Казахстан, Киргизия.

На конец июня 2019 г. в базе данных Росреестра ПО было зарегистрировано 5460 программных продуктов, отвечающих вышеупомянутым критериям и подлежащих классификации с помощью Классификатора программ для электронных вычислительных машин и баз данных. Индексирование программного обеспечения и баз данных осуществляется Экспертным советом по российскому программному обеспечению при Министерстве связи и массовых коммуникаций Российской Федерации одновременно с принятием решения о включении сведений о программном обеспечении в Росреестр ПО.

Каждая программа для ЭВМ и база данных может быть проиндексирована сразу несколькими индексами Классификатора, которые в Росреестре ПО отображаются в виде наименования раздела или класса. Например, информация о классе ПО «*ABBYU PassportReader SDK*» представлена следующим образом: «Библиотеки подпрограмм (*SDK*), Лингвистическое программное обеспечение, Системы сбора, хранения, обработки, анализа, моделирования и визуализации массивов данных».

Классификатор является иерархическим с использованием последовательного метода кодирования. Глубина классификатора составляет 2 уровня, каждый из которых состоит из двух цифр:

1 уровень – раздел (например: «01 – Встроенное программное обеспечение»);

2 уровень – класс (например: «01.01 – BIOS и иное встроенное программное обеспечение»).

Классификатор состоит из 4 разделов, содержащих 23 класса:

1) 01 – Встроенное программное обеспечение (содержит 2 класса);

2) 02 – Системное программное обеспечение (содержит 7 классов);

3) 03 – Средства разработки программного обеспечения (содержит 5 классов);

4) 04 – Прикладное программное обеспечение (содержит 9 классов).

В структуре классификатора выделяются следующие поля:

- Наименование раздела;
- Наименование класса;
- Описание класса программ для ЭВМ и баз данных;
- Код (числовое обозначение) раздела или класса программ для ЭВМ и БД;
- Код (числовое обозначение) Общероссийского классификатора видов экономической деятельности (ОКВЭД).

Поля классификатора являются стандартными, однако хотелось бы немного подробнее остановиться на поле «Описание класса программ для ЭВМ и БД». Это поле содержит подробное описание класса. Например:

«03 – Средства разработки программного обеспечения

03.03 – Средства версионного контроля исходного кода

Программы, позволяющие хранить несколько версий одного и того же документа и при необходимости возвращаться к более ранним версиям и определять, кто и

когда сделал то или иное изменение» (Приказ Минкомсвязи № 134).

- Во многих случаях описание класса разделено на пункты, по которым удобно ориентироваться. Например, согласно Приказу Минкомсвязи России № 134

«04 – Прикладное программное обеспечение

04.03 – Офисное программное обеспечение

Программы (приложения), предназначенные для обработки электронных документов и другой информации (productivity software).

Включают в себя:

- Файловые менеджеры (*File Manager*) / Программы (приложения), позволяющие создавать, удалять, копировать и перемещать файлы в доступных пользователю хранилищах файлов

- Коммуникационное ПО / Программы, обеспечивающие коммуникации пользователей, в том числе в виде голосовых и видеозвонков, обмена текстовыми сообщениями, файлами, электронными письмами

- Офисные пакеты / Совокупность взаимосвязанных программ (приложений) с унифицированным интерфейсом, предназначенных для создания, просмотра и редактирования электронных документов.

- Почтовые приложения / Программы, обеспечивающие доступ пользователя к его почтовым ящикам с учетом разграничения прав, которые позволяют получать, создавать, просматривать, редактировать, отправлять и хранить сообщения электронной почты».

Исходя из вышеизложенного, можно назвать основные достоинства этого классификатора:

1) подробное описание каждого класса программ и баз данных обеспечивает более точное индексирование и последующий поиск;

2) небольшой объем классификатора обеспечивает наглядность и простоту использования при формировании поискового запроса;

3) наличие соответствий с Общероссийским классификатором видов экономической деятельности (ОКВЭД) связывает рассматриваемый Классификатор с другими российскими классификаторами.

Однако, несмотря на достоинства, результаты поиска с использованием этого Классификатора очень часто являются нерелевантными или непертициентными. Можно выделить три основные причины этого явления:

1) отсутствие возможности объединения поискового запроса с помощью булевой логики «И»;

2) индексирование объектов излишним количеством кодов;

3) несоответствие количества индексируемых объектов количеству классов Классификатора.

Рассмотрим эти причины подробнее.

Отсутствие возможности объединения поискового запроса с помощью булевой логики «И» увеличивает количество результатов поиска, что характерно для логики «ИЛИ». В качестве примера попробуем найти «Систему автоматизации библиотек ИРБИС64 (САБ ИРБИС64)» (№ 1808), которая индексируется двумя классами:

«04.11 – Системы управления процессами организации»;

«04.15 – Информационные системы для решения специфических отраслевых задач».

Для начала произведем раздельный поиск по этим классам и оценим количество результатов:

«04.11» – 1692 результата, расположенных на 85 страницах по 20 позиций на каждой;

«04.15» – 3506 результатов, расположенных на 176 страницах по 20 позиций на каждой.

Произведя объединенный поиск по этим классам («04.11» + «04.15») мы получим 4002 результата на 201 странице по 20 позиций на каждой.

Как видно из результатов, в алгоритме поиска по Классификатору отсутствует возможность объединения запросов с помощью булевой логики «И». Это же подтверждается и найденными позициями. Например: «АРМ сотрудника ДПС» (№5461) находится в результатах поиска, но проиндексирован только одним классом: «Информационные системы для решения специфических отраслевых задач».

Конечно, существует практика упрощения пользовательского интерфейса поиска посредством исключения возможности выбора логики объединения поисковых запросов, однако в алгоритме поиска и последующем ранжировании результатов эта логика учитывается, а самыми первыми показываются результаты наиболее релевантные поисковому запросу. Например, если поисковый запрос состоит из двух понятий, то первыми будут отражаться результаты объединенного поиска с помощью логики «И», как наиболее соответствующие, а затем с помощью логики «ИЛИ». Так устроен интерфейс поисковых систем *Google*, *Yandex* и др., и так же работают их алгоритмы

К сожалению, в поисковом алгоритме Росреестра это не реализовано, что подтверждается анализом результатов. Так при поиске по этим двум классам САБ ИРБИС64 находится на 61 странице по 20 позиций на каждой, причем в большинстве случаев ей предшествуют результаты менее релевантные поисковому запросу, поскольку содержат только один из двух классов, по которым производится поиск. Например: № 54 «БАРС.Портал – Образование» (№ 54) присвоен только один класс – «Информационные системы для решения специфических отраслевых задач», однако в результатах поиска он расположен на самой первой (!) странице.

Индексирование объектов излишним количеством кодов классификатора. Присвоение классов программам и базам данных, как уже было отмечено, осуществляется Экспертным советом по российскому программному обеспечению при Минкомсвязи России одновременно с принятием решения о включении сведений о программном обеспечении в Росреестр ПО на основании документов, предъявляемых заявителем. Обычно каждому продукту присваивается от 1 до 3 классов, однако, поскольку никаких ограничений по количеству присвоенных классов в Правилах применения классификатора (Приказ Минкомсвязи России № 662) не имеется, то некоторые заявители пользуются этой возможностью и подают заявку о включении их программного обеспечения или базы данных сразу в несколько классов, что является излишним.

Например, встраиваемая операционная система *FastSYS* проиндексирована шестью классами (01.02 –

Встроенное программное обеспечение телекоммуникационного оборудования; 02.03 – Утилиты и драйверы; 02.07 – Серверное и связующее программное обеспечение; 02.13 – Средства обеспечения информационной безопасности; 02.15 – Информационные системы для решения специфических отраслевых задач; 02.01 – Операционные системы), несмотря на то, что достаточно было бы указать только два: «02.01 – Операционные системы» и «01.02 – Встроенное программное обеспечение телекоммуникационного оборудования». Остальные присвоенные классы описывают либо часть программного обеспечения этого продукта, (например: 02.03 – Утилиты и драйверы), либо являются слишком общими (например: 02.15 – Информационные системы для решения специфических отраслевых задач).

Однако существует и абсолютный «чемпион» такого индексирования – Общее программное обеспечение «Эльбрус» (№3199), которому присвоено 13 классов из 23, т.е. больше половины всех классов классификатора.

Стоит отметить, что заявители или экспертный совет начинают понимать ошибку излишнего индексирования и корректируют описания некоторых продуктов. Например, Операционной системе общего назначения «*Astra Linux Common Edition*» (№ 4433) до ноября 2018 г. было присвоено 8 кодов классификатора, однако сейчас остался только один: «02.01 – Операционные системы».

Несоразмерность количества индексировемых объектов количеству классов классификатора. В базе данных Росреестра на конец июня 2019 г. зарегистрировано 5460 программ и баз данных, которые распределяются по 23 классам Классификатора. Таким образом, на один класс Классификатора приходится в среднем 237 объектов ($5460/23 = 237,39$).

Для примера, индексирование книг и документов в Государственной публичной научно-технической библиотеке Российской Федерации (ГПНТБ России) производится с помощью Универсальной десятичной классификации (УДК) и Государственного рубрикатора научно-технической информации (ГРНТИ). Согласно статистике полное число записей в машиночитаемом каталоге ГПНТБ России составляет 3829510^3 , количество индексов в УДК – более 68000 [3], в ГРНТИ – более 8000 [3]. Таким образом, на один индекс УДК в среднем приходится 56,31 изданий, а на один индекс ГРНТИ – 478,68. Из этого соотношения видно, что УДК более подробно отражает тематику документов, а ГРНТИ стремится собрать как можно больше документов в одну тематическую рубрику.

Несомненно, усредненные значения полезны, поскольку отражают общую картину, однако для ее полноты необходимо провести более подробное статистическое исследование, а именно подсчитать количество программ и баз данных по каждому классу Классификатора (см. таблицу).

³ Краткий отчет о деятельности ГПНТБ России за 2017 год. – URL: <http://www.gpntb.ru/ofitsialnye-dokumenty/84--12/ofitsialnye-dokumenty/5397-kratkij-otchet-o-deyatelnosti-gpntb-rossii-za-2019-god.html>.

Количество программ и БД в каждом классе Классификатора

Код класса	Описание класса	Количество программ и баз данных
01.01	BIOS и иное встроенное программное обеспечение	48
01.02	Встроенное программное обеспечение телекоммуникационного оборудования	41
02.01	Операционные системы	52
02.03	Утилиты и Драйверы	71
02.05	Средства обеспечения облачных и распределенных вычислений, средства виртуализации и системы хранения данных	91
02.07	Серверное и связующее программное обеспечение	363
02.09	Системы управления базами данных	54
02.11	Системы мониторинга и управления	620
02.13	Средства обеспечения информационной безопасности	429
03.01	Средства подготовки исполнимого кода	20
03.03	Средства версионного контроля исходного кода	3
03.05	Библиотеки подпрограмм (SDK)	106
03.07	Среды разработки, тестирования и отладки	89
03.09	Системы анализа исходного кода на закладки и уязвимости	14
04.01	Прикладное программное обеспечение общего назначения	960
04.03	Офисные приложения	197
04.05	Поисковые системы	153
04.07	Лингвистическое программное обеспечение	75
04.09	Системы управления проектами, исследованиями, разработкой, проектированием и внедрением	378
04.11	Системы управления процессами организации	1692
04.13	Системы сбора, хранения, обработки, анализа, моделирования и визуализации массивов данных	895
04.15	Информационные системы для решения специфических отраслевых задач	3506

Анализируя таблицу, можно выявить наиболее часто применяемые классы, т.е. те, использование которых превышает среднее значение по Классификатору. Таких классов было найдено восемь (в процентном соотношении 34,78%), причем использование пяти из них выше среднего более чем в два раза. Также следует обратить внимание на два класса: «04.11» и «04.15», использование которых превышает среднее значение более чем в 7 и 14 раз соответственно. Этими же двумя классами проиндексировано 30% и 64% всех представленных в БД Росреестра программ и баз данных. Таким образом, можно сделать вывод, что результаты поиска с использованием вышеуказанных 8 классов будут не только нерелевантными, но и непертинентными.

Для увеличения релевантности систематического поиска с помощью Классификатора и улучшения качества поиска предлагаются следующие варианты решения вышеуказанных проблем:

- 1) реорганизация алгоритма поиска;
- 2) установка ограничений на количество присваиваемых классов;
- 3) увеличение глубины классификатора.

Для того, чтобы объединить поисковый запрос с помощью булевой логики «И», необходимо **реорганизовать алгоритм поиска**. Видится два пути решения этой задачи.

Первый. Включить возможность выбора либо с помощью опций, предлагаемых в пользовательском

интерфейсе, либо с помощью синтаксиса языка поискового запроса. В последнем случае это можно сделать, например, с помощью знака «+» (плюс), логик объединения поискового запроса: логики «ИЛИ», которая уже использовалась в поисковой системе Росреестра ПО и логики «И», которую необходимо разработать.

Второй. Реализовать механизм ранжирования результатов поиска от наиболее релевантных (т.е. полностью соответствующих поисковому запросу) к менее релевантным.

Для предотвращения индексирования объектов излишним количеством кодов классификатора следует **ввести ограничения на количество присваиваемых классов в правила** применения классификатора для электронных вычислительных машин и баз данных.

Каждое программное обеспечение и базу данных можно отнести к одному классу, который будет выражать наиболее общее понятие для этого типа программ. Например, рассмотренное выше Общее программное обеспечение «Эльбрус» имеет альтернативное название – Операционная система «Эльбрус», но индексируется не только классом «02.01 – Операционные системы», но и другими 12 классами, которые относятся к входящим в эту систему самостоятельным программам с собственным наименованием. Такое ПО можно называть составным.

Однако, ограничение только одним классом тоже не всегда приемлемо, поскольку существуют программы, которые не содержат в своем составе другое ПО, но выполняют несколько функций, подпадающих под разные классы Классификатора. Например, Платформа «CUBA» (№ 1294) индексируется классами: «03.05 – Библиотеки подпрограмм (SDK)», «03.07 – Среды разработки, тестирования и отладки», «04.15 – Информационные системы для решения специфических отраслевых задач». И если последний класс можно исключить из описания, то два других класса – основные для этого программного обеспечения, поскольку оно является средой разработки, содержащей библиотеки подпрограмм. Небольшое пояснение про библиотеки подпрограмм: они могут быть как самостоятельными продуктами, так и являться частью среды разработки, поэтому в конкретном случае использование этих двух классов вполне оправдано.

Учитывая вышесказанное, наиболее приемлемым представляется ограничение на использование максимум трех классов для каждого ПО и БД, которое следует ввести в правила применения классификатора для электронных вычислительных машин и баз данных.

Увеличение глубины классификатора необходимо, поскольку постоянный и устойчивый рост количества программ и баз данных, проходящих регистрацию в Росреестре, приводит к несоразмерности количества индексируемых объектов и количества классов классификатора.

В первую очередь делению должны подвергнуться наиболее часто используемые классы: «02.07 – Серверное и связующее программное обеспечение», «02.11 – Системы мониторинга и управления», «02.13 – Средства обеспечения информационной безопасности», «04.01 – Прикладное программное обеспечение общего назначения», «04.09 – Системы управления проектами, исследованиями, разработкой, проектированием и внедрением», «04.11 – Системы управления процессами организации», «04.13 – Системы сбора, хранения, обработки, анализа, моделирования и визуализации массивов данных», «04.15 – Информационные системы для решения специфических отраслевых задач».

Увеличение глубины в большинстве случаев не представляется сложным, поскольку все классы снабжены описанием, которое часто имеет перечислительная структура. Например, согласно Приказу Минкомсвязи России № 134 класс «04.01 – Прикладное программное обеспечение общего назначения» содержит описание:

«Прикладные программы, предназначенные для выполнения различных задач». Включают в себя :

- мультимедийное ПО/программы, обеспечивающие представление информации в виде звука, анимированной компьютерной графики, видеоряда;
- плагины (*plug-in*)/независимо компилируемый программный модуль, динамически подключаемый к основной программе и предназначенный для расширения ее возможностей;
- игры и развлечения/программы, организующие игровой процесс (*gameplay*), связь с партнерами по игре, или сами выступающие в качестве партнера».

Классы 02.13, 04.09, 04.11, 04.13 имеют подобное описание. Таким образом, при выделении этих пунктов в отдельные классы программ увеличится глубина Классификатора и впоследствии улучшится релевантность и pertinентность поиска. Приведем примеры такого преобразования:

04.01 – Прикладное программное обеспечение общего назначения

04.01.01 – Мультимедийное ПО

04.01.03 – Плагины (*plug-in*)

04.01.05 – Игры и развлечения

Сложности могут встретиться при углублении классов, в описании которых нет перечислительной структуры – 02.07, 02.11 и 04.15. Например, для класса «02.11 – Системы мониторинга и управления» дано описание:

- «Программно-аппаратные комплексы, предназначенные для изменения, сбора, хранения и анализа рабочих характеристик объектов управляемой системы для оценки ее состояния, выявления неполадок, оповещения, управления настройками и состоянием программных и аппаратных вычислительных средств, сетевых устройств, функционирующих в рамках системы, в том числе средства управления информационной безопасностью» (Приказ Минкомсвязи России № 134), из которого сложно выделить нижестоящие классы.

Необходимо отдельно остановиться на самом используемом в Росреестре классе – «04.15 – Информационные системы для решения специфических отраслевых задач». Описание этого класса содержит перечень различных отраслей:

- «Программные системы специализированного информационного обеспечения государственного управления, жилищно-коммунального хозяйства, социальной защиты и обеспечения норм труда, здравоохранения, обеспечения безопасности, видеонаблюдения и контроля доступа, оборонно-промышленного комплекса, образования, производства и машиностроения, радиочастотной идентификации, промышленности, страхования, строительства, телекоммуникаций (включая IP АТС) и СМИ, торговли, транспорта и перевозок, финансов и банковского сектора, энергетики и нефтегазовой отрасли, юриспруденции, фармацевтики, рекламы, гостиничного и туристического бизнеса» (Приказ Минкомсвязи России № 134).

Исходя из приведенного выше описания, становится понятна причина «популярности» этого класса, которая заключается в том, что почти каждую программу можно включить в этот класс. Например, этим классом проиндексированы и «Система автоматизации библиотек ИРБИС64 (САБ ИРБИС64)» (№ 1808) и «*IndorCAD/Road*: Система проектирования автомобильных дорог» (№ 82). Именно это разнообразие свидетельствует о необходимости углубления этого класса, в результате которого каждое отраслевое ПО будет иметь свой код, а следовательно, увеличится релевантность поиска по классификатору. Это было бы очень полезно при поиске конкретных отраслевых программ.

Углубление этого класса можно произвести двумя способами.

1. Разбить описание класса на нижестоящие классы, например:

04.15.81 – Программные системы информационного обеспечения государственного управления.

04.15.82 – Программные системы информационного обеспечения жилищно-коммунального хозяйства и т.д.

2. Использовать уже созданный и применяемый на практике классификатор отраслей, например, Общероссийский классификатор видов экономической деятельности (ОКВЭД-2), например:

04.15.02 – Лесоводство и лесозаготовки полуприцепов (сформирован из класса ОКВЭД-2 «02»).

04.15.29 – Производство автотранспортных средств, прицепов и полуприцепов (из класса ОКВЭД-2 «29») и т.д.

Также следует отметить, что для сохранения существующих принципов построения Классификатора можно идти не только путем углубления классов, который был предложен выше, но и путем простого увеличения их количества, т.е. выделяя из одного существующего класса несколько новых. Например:

04.01 – Прикладное программное обеспечение общего назначения.

04.21 – Мультимедийное ПО.

04.23 – Плагины (*plug-in*).

04.25 – Игры и развлечения.

В настоящее время Росреестр ПО активно развивается и совершенствуется, что подтверждается количеством зарегистрированных в нем программ и БД. Несмотря на свой «юный возраст» он решает возложенные на него задачи. Однако для повышения релевантности и пертинентности поиска необходимо доработать как сам классификатор программ для электронных вычислительных машин и баз данных, так и алгоритмы поиска, правила применения классификатора

Надеемся, что наша статья поможет разработчикам улучшить реализацию систематического поиска в Росреестре ПО, тем самым облегчив пользователям поиск с помощью Классификатора.

СПИСОК ЛИТЕРАТУРЫ

1. Единый реестр российских программ для электронных вычислительных машин и баз данных. – URL: <https://reestr.minsvyaz.ru/reestr/>.
2. Смирнов Ю.В. Анализ поисковых возможностей Единого реестра российских программ для электронных вычислительных машин и баз данных // Материалы конференции «Информационные технологии, компьютерные системы и издательская продукция для библиотек ("LIBCOM-2018")». – URL: <http://www.gpntb.ru/libcom2018/docl/005.pdf>.
3. Гиляревский Р.С., Шапкин А.В., Белозеров В.Н. Рубрикатор как инструмент информационной навигации. – СПб: Профессия, 2008. – С. 269.

Материал поступил в редакцию 12.08.19.

Сведения об авторе

СМИРНОВ Юрий Викторович – научный сотрудник Государственной публичной научно-технической библиотеки России, Москва
e-mail: yury@gpntb.ru

ВНИМАНИЮ ЧИТАТЕЛЕЙ!

ИЗДАНИЕ УДК

УНИВЕРСАЛЬНАЯ ДЕСЯТИЧНАЯ КЛАССИФИКАЦИЯ
АЛФАВИТНО-ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ
в 2-х томах

Алфавитно-предметный указатель (АПУ) к 4-му полному изданию УДК на русском языке:

Том I содержит АПУ от буквы А до Н;

Том II содержит АПУ от буквы М до Я и указатель латинских наименований к классам УДК 56 Палеонтология, 57 Биологические науки, 58 Ботаника, 49 Зоология, 61 Медицинские науки.

АПУ содержит около 100 000 понятий, представленных в полных таблицах УДК.

При его составлении были учтены изменения, опубликованные в Выпусках № 1 – 6 «Изменения и дополнения к УДК»

Для подписки необходимо направить заявку для оформления счета по адресу:

125190, Россия, Москва, ул. Усиевича, 20, ВИНТИ РАН

Телефоны: 499 155-42-85, 499 151-78-61

E-mail: feo@viniti.ru

<http://www.udcc.ru>

ВНИМАНИЮ ЧИТАТЕЛЕЙ!

ВИНИТИ РАН, как единственный в России владелец лицензии Консорциума УДК, предлагает издания УДК полного четвертого издания на русском языке в печатном и электронном виде:

1. Таблицы УДК

УДК. Том I Общая методика применения УДК. Вспомогательные таблицы. Основные таблицы. Общий отдел. Алфавитно-предметный указатель к Общему отделу

УДК. Том II 1/3 Философия. Психология. Религия. Богословие. Общественные науки (только электронное издание)

УДК. Том III 5/54 Математика. Естественные науки (только электронное издание)

УДК. Том IV 55/59 Геологические и биологические науки (только электронное издание)

УДК. Том V 6/61 Медицинские науки (только электронное издание)

УДК. Том VI (часть 1) 6/621 Прикладные науки. Технология. Инженерное дело (только электронное издание)

УДК. Том VI (часть 2) 622/629 Техника. Инженерное дело (только электронное издание)

УДК. Алфавитно-предметный указатель к т. VI (1 и 2 части) (только электронное издание)

УДК. Том VII 63/65 Сельское хозяйство. Домоводство. Управление предприятием (только электронное издание)

УДК. Том VIII 66 Химическая технология. Химическая промышленность. Пищевая промышленность. Металлургия. Родственные отрасли (только электронное издание)

УДК. Том IX 67/69 Различные отрасли промышленности и ремесел. Строительство (только электронное издание)

УДК. Том X 7/9 Искусство. Спорт. Филология. География. История.

УДК. АПУ (с в о д н ы й) к полному 4-му изданию

УДК. Изменения и дополнения. Выпуск 2 (к т.т. 1–3) (только электронное издание)

УДК. Изменения и дополнения. Выпуск 3 (к т.т. 1–6) (только электронное издание)

УДК. Изменения и дополнения. Выпуск 4 (к т.т. 1–7) (только электронное издание)

УДК. Изменения и дополнения. Выпуск 5 (к т.т. 1–10)

УДК. Изменения и дополнения. Выпуск 6 (к т.т. 1–10)

УДК. Изменения и дополнения. Выпуск 7 (к т.т. 1–10), 2017 г. (только электронное издание)

Для подписки необходимо направить заявку по адресу:

125190, Россия, Москва, ул. Усиевича, 20, ВИНТИ РАН

Телефоны: 499-155-42-85, 499-151-78-61

E-mail: feo@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** выполняют 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>