

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

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

Издается с 1961 г.

№ 12

Москва 2024

ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ

УДК 004.4'242:005.4/.5

Л.А. Мыльников, А.Д. Салтыкова

Структурно-функциональное моделирование в нотации *Event-driven Process Methodology* как инструмент для генерации кода программ, имитирующих процессную деятельность*

Рассматривается необходимость получения объективных оценок эффективности организации процессной деятельности на основе цифровых двойников. Описана методика установления соответствия между структурно-функциональными моделями в нотации EPM (Event-driven Process Methodology) и знаково-символьным описанием, которое может быть переведено на язык программирования на основе группы правил. Приведены примеры получения знаково-символьного описания и программ на языке R по заданной структурно-функциональной модели, а также результаты статистического исследования процессов, которые они описывают.

Ключевые слова: EPM, Event-driven Process Methodology, имитационное моделирование, статистическое исследование, оценка эффективности процесса, функциональное программирование

DOI: 10.36535/0548-0027-2024-12-1

* Исследование выполнено при финансовой поддержке Правительства Пермского края в рамках научного проекта № С-26/692.

ВВЕДЕНИЕ

Эффективность применения информационных технологий различна в разных сферах деятельности в связи с применением специализированного программного обеспечения, а также с некоторыми функциональными возможностями (уровнем автоматизации), методикой и трудоемкостью использования, требованиями к квалификации пользователей, сложившимися традициями применения ИТ решений и формой представления данных в свойственных каждой предметной области управленческих/производственных/технологических процессах.

Для описания процессной деятельности используют структурно-функциональные модели в таких нотациях как BPMN [1], eEPC [2], IDEF [3], представляющие собой схемы с выделенными элементами (действия, ресурсы) и связями между ними. Более современные нотации ArchiMate (TOGAF) [4], OPM [5], SysML [6] позволяют на единой схеме показать связь между процессной деятельностью и средствами ее реализации и поддержки, обозначить целевые показатели эффективности.

Изучение эффективности процессной деятельности с использованием выделенных KPI (Key Point Indicators – ключевых показателей эффективности) и PPI (Process Performance Indicators – показателей эффективности процессной деятельности) происходит, как правило, путем сбора статистики уже реализованных процессов, что открывает путь к их реконструкции [7].

Исследование, проведенное McKinsey [8], показывает, что компании, эффективно применяющие методы оптимизации протекающих процессов, достигают до 30% сокращения операционных затрат, увеличивают производительность на 35% и повышают удовлетворенности клиентов в среднем на 30%.

Анализировать эффективность процессов до их реализации (рис. 1) возможно только с помощью имитации их работы на структурно-функциональных моделях. Необходимым функционалом для этого обладают сети Петри [9], однако они имеют ограничение,

связанное с проведением обращений/заявок/запросов по одному. Имитационное моделирование с использованием специализированных средств и языков (Simulink, AnyLogic, SciLab, Simula, Algol) не имеет установившегося стандарта для графических обозначений и универсального языка программирования для описания моделей, что затрудняет их чтение и создание.

Такие подходы к исследованию структурно-функциональных моделей, как AS-IS ->TO-BE или методологии BPM [10] и CMMN [11] носят экспертный характер и не позволяют получать объективные данные об эффективности тех или иных процессов.

В отличие от упомянутых описаний и подходов, в нотации Event-driven Process Methodology (EPM) в явном виде выделены блоки, задающие управляющее воздействие (S), блоки окружения (NV), результата (R) и действия (V), позволяющие задавать начальные условия, блоки условия, функциональные описания для представления модели в знаково-символьной форме, что соответствует требованиям к функциональным языкам программирования.

Наличие способа формализации модели на псевдоязыке позволяет переходить на различные языки программирования – как универсальные, так и специализированные – для моделирования, проводить любые статистические исследования и получать объективные оценки эффективности на основе имитационного моделирования с опорой на центральную предельную теорему [12].

Получение объективных оценок открывает, в свою очередь, путь к использованию методов выбора решений, основанных как на привлечении экспертов, так и на использовании критериального выбора.

МЕТОДИКА

Для исследования структурно-функциональных моделей путем имитации их работы необходимо установить соответствия между схематическим и знаково-символьными представлениями. Для нотации EPM [13] такое соответствие может быть установлено, как приведено в таблице 1.

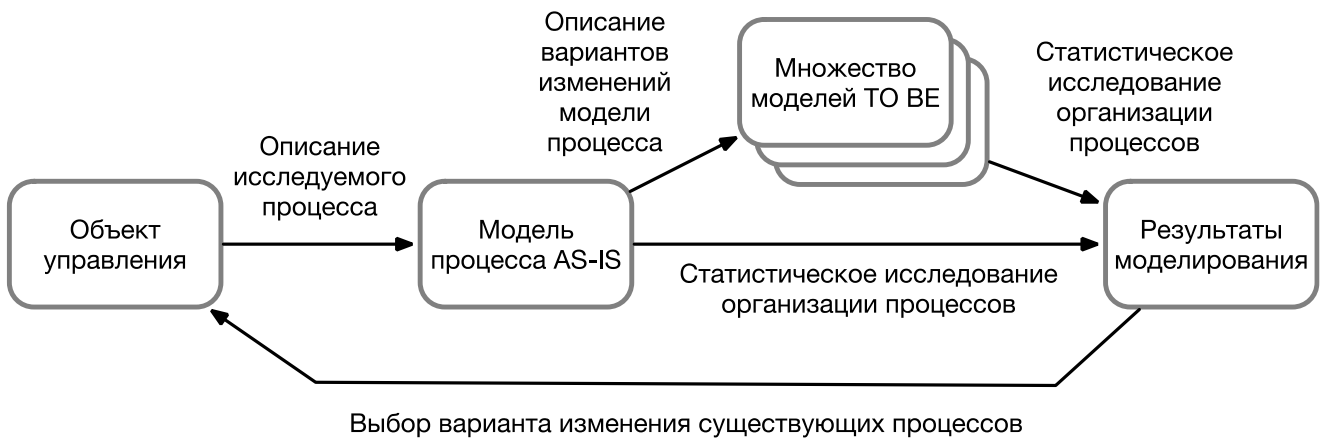


Рис. 1. Схема выбора способа организации процессной деятельности на основе объективных данных

Артефакты нотации EPM, их функционал, назначение, знаково-символьное представление

№	Схематическое обозначение	Функционал	Назначение	Знаково-символьное описание
1	 Блок инициализации (S)	Генерирует простое натуральное число, используя поток Пуассона с заданной вероятностью или периодически повторяющееся значение	Активация работы блока действия, генерация числа запросов/заявок/обращений	$S = \text{Prob}(\langle \text{число значений} \rangle, \langle \text{вероятность} \rangle)$ $S = \text{Periodic}(\langle \text{число значений} \rangle, \langle \text{период} \rangle)$
2	 Блок действия (V)	Реализует операцию, результатом которой будут данные, или задержку, связанную с вычислительной сложностью, в таком случае результатом будут запросы, которые появятся на выходе через некоторое время	Обработка запросов	$R = V(\langle \text{сложность} \rangle, S, NV)$ $R = F(S, NV)$ – для случая самостоятельно реализуемой функции
3	 Блок результата (R)	Хранение и передача на следующем такте результата в следующий блок по имеющейся связи	Хранение результата работы блока действия (V)	
4	 Блок окружения (NV)	Хранит данные, необходимые для работы блока действия (V)	Хранение данных, необходимых для работы модели	$NV = R$ $NV \$ \langle \text{имя переменной} \rangle = \langle \text{значение} \rangle$
5	 Блок условия	Используется для ожидания выполнения действий (синхронизации и управления последовательностью процессов)	Реализация циклов и условных переходов	do $\langle \text{условие} \rangle$ then $\langle \text{действия} \rangle$ else $\langle \text{действия} \rangle$
6		Присвоение	Передача данных	= означает присвоение значения блоку, в который входит стрелка + означает объединение данных при вхождении множества стрелок; ∩ означает выделение данных при объединении результатов нескольких проходов в одном блоке при организации циклов

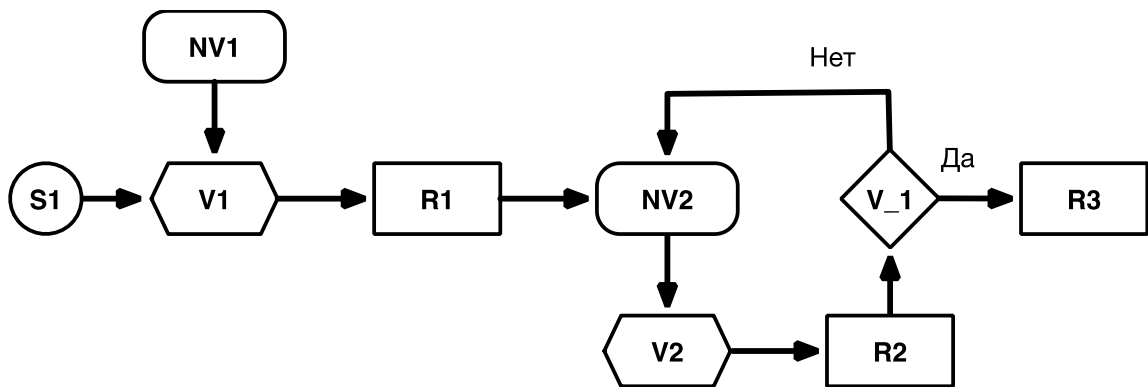


Рис. 2. Процесс приготовления тоста: S1-поток запросов на приготовление тостов, V1 – подготовительные операции (определимся со степенью прожарки, достать хлеб), V2 – приготовление тоста, V_1 – проверка готовности тоста, R2- тост, R3- готовый тост

Использование знаково-символьного описания позволяет вводить начальные условия (пп. 1 и 4 в таблице 1), основные функции (пп. 2 и 3 в таблице 1), условия и циклы (п. 5 в таблице 1). Для обмена данными и через связи необходимо использовать операторы (см. п. 6 в таблице 1).

Правила построения моделей описаны в [13]: 1) использование блоков действия (V) в связке с блоком S, NV (не обязательно) и R, 2) в блок условия может входить только блок R и выходы этого блока могут быть соединены только с блоками R или NV.

С учетом правил нотации введенные соответствия между структурно-функциональным и знаково-символьным представлениями позволяют формализовать написание программ для исследования эффективности процессов. Например, если рассмотреть подготовку тостов в [14] и на рис. 2, то путем сопоставления графических элементов знаково-символьного представления мы получаем знаково-символьное описание, приведенное в листинге 1.

Листинг 1. Знаково-символьное описание, соответствующее приведенной на рис. 2 структурно-функциональной модели:

```
S1= Prob(<число значений>, <вероятность>);
R1=V(<вычислительная сложность>, S1, NV1);
NV2=R1;
R2= V(<вычислительная сложность>, 1, NV2);
do v_1(R2) then
    R2=V(<вычислительная сложность>, 1,
NV2+R2);
else R3=R2.
```

В полученном представлении используются векторы значений, что, с одной стороны, является преимуществом, так как позволяет обрабатывать данные за один проход, а с другой – для получения статистически значимых результатов при повторных вызовах и циклах требует пересмотра данных, задаваемых блоками S, если для их определения использовались случайно сгенерированные числа (функция *Prob*).

Объясним это на примере, в котором требуется заданное число повторов или изменение логики работы в зависимости от номера прохода. Для учета числа проходов необходимо использовать глобальный счетчик повторов (*i*): если некоторый операционный блок должен быть выполнен при некоторых условиях или неоднократно (например, два раза последовательно), а данные, поступающие в блок действия, рассматриваются как заявки, требующие обработки, то такая ситуация может быть описана приведенной на рис. 3 структурно-функциональной моделью [15].

В знаково-символьной форме такую структурно-функциональную модель (см. рис. 3) можно представить листингом 2.

Листинг 2. Знаково-символьное описание, соответствующее приведенной на рис. 3 структурно-функциональной модели:

```
S1= Prob(<число значений>, <вероятность>);
i=1;
R1(i)=V(<вычислительная сложность>, S1,
NULL);
```

```
do V2(i) then
    i=i+1;
    S1= Prob(<число значений>, <вероятность>);
    R1(i)=V(<вычислительная сложность>,
S1+R1, NULL);
else R2=R1(i)∩R1(i-1).
```

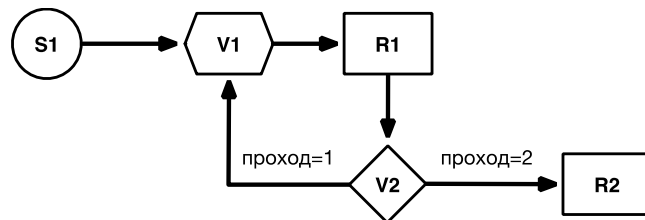


Рис. 3. Пример структурно-функциональной модели с повторяющимся блоком в процессе обработки заявки:

S1 – блок генерации потока заявок (начального вектора запросов), V1 – блок обработки заявок, R1 – результаты обработки заявок в блоке V1, V2 – условный блок используется для проверки того, сколько раз заявки прошли через блок V1, R2 – результаты обработки заявок

Особенность, связанная с работой модели, приведенной на рис. 3, заключается в том, что при каждом новом использовании значения S1 должны обновляться.

В случае, если повторы влияют на окружение, то мы получаем ситуацию, показанную на рис. 4.

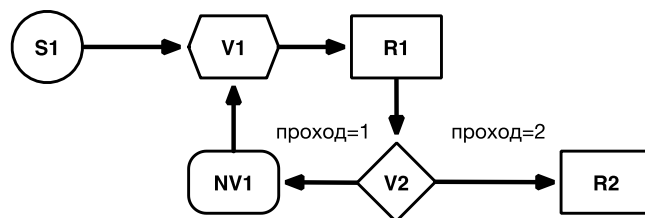


Рис. 4. Пример структурно-функциональной модели с повторяющимся блоком в процессе обработки заявки:

S1 – блок генерации потока заявок (начального вектора запросов), V1 – блок обработки заявок, R1 – результаты обработки заявок в блоке V1, V2 – условный блок используется для проверки того, сколько раз заявки прошли через блок V1, R2 – результаты обработки заявок

В знаково-символьной форме работу приведенной на рис. 4 структурно-функциональной модели можно представить листингом 3:

Листинг 3. Знаково-символьное описание, соответствующее приведенной на рис. 4 структурно-функциональной модели:

```
S1=Prob(<число значений>, <вероятность>);
i=1;
R1(i)=V(<вычислительная сложность>, S1,
NV1);
```

```

do V2(i) then
  i=i+1;
  S1= Prob(<число значений>, <вероят-
ность>);
  R1(i)=V(<вычислительная сложность>,
S1,R1);
else R2=R1(i).

```

Получаемые знаково-символьные представления удовлетворяют требованиям, предъявляемым к функциональным языкам программирования [16], а это значит, что между получаемым знаково-символьным представлением и другими языками программирования (включая императивные) может быть установлено соответствие (необходимые функции и операции могут быть реализованы на другом языке, а части кода в знаково-символьной нотации поставлен в соответствие код на другом языке программирования) и при добавлении команд по сбору статистики, её выводу на экран или в файл по построению графиков может быть выполнен анализ работы процессов, описанных структурно-функциональными моделями.

ИССЛЕДОВАНИЕ ЗНАКОВО-СИМВОЛЬНЫХ МОДЕЛЕЙ

Для получения программы на одном из существующих языков программирования необходима реализация функций *Prob*, *Periodic*, *V*, *F* (дополнительные функции, реализующие работу алгоритма для элемента *V*). Пример реализации функции *Prob* на языке R приведен в листинге 4. Все значения потока заявок/запросов/обращений обрабатываются за один проход, так как такие переменные как *S*, *R*, *NV* являются векторами.

Листинг 4. Реализация на языке R функции, выдающая множество случайных значений, в которой параметр *P* – вероятность, *N* – число элементов для заполнения:

```

S_prob<-function(N, P, ID){
  RES<-vector(mode = "integer", length = N)
  RES[1:N]<-rpois(N,P)

  S1<-data.frame(S=RES, ID=vector(mode =
"numeric", length = N))
  ID_S<-ID
  for (i in 1:N){
    if (RES[i]>0){
      Vect<-as.vector(ID_S+1:RES[i], mode =
"numeric")
      ID_S<-ID_S+RES[i]
      S1$ID[i]<-list(Value=Vect)
    }
  }
  S_prob<-S1
}

```

Кроме этого, для хранения данных в переменных, построения графиков, их выдачи на экран и сохранения во внешний файл могут потребоваться дополнительные типовые части кода, которые также могут быть реализованы с применением механизма функций на используемом для реализации языке программирования (см. например листинги 5 и 6).

Листинг 5. Часть кода, выполняющая преобразование структуры данных после получения результата R и сбор статистики в переменную *X* на языке R:

```

NV <- subset(R, select=c(R, ID_Out))
colnames(NV) <- c('S', 'ID')
X2 <- XES(R)
X <- rbind(X, X2)

```

Листинг 6. Часть кода, позволяющая на основе собранной статистики (переменная *X*) определять время прохождения заявки/запроса/обращения через процесс (переменная *s*) и представлять полученные данные в виде скрипичной диаграммы на языке R:

```

l<-unique(X$ID)
s<-NA
for (i in 1:length(l)){
  s[i]<-sum(X$W[X$ID==l[i]])
}
vioplot(s, col = "lightgray", panel.first=grid())

```

Для реализации условного блока могут использоваться конструкции «if ... then ... else ...» или «while ... do ...», которые присутствуют в большинстве языков программирования.

Используя правило замены, основанное на соответствиях, можно записать код программы, который будет приближен по виду к знаково-символьному представлению (листинг 7).

Листинг 7. Программа для имитации работы модели, представленной на рис. 2:

```

N<-1000 # число рассчитываемых шагов работы модели
# Генерируем поток запросов на приготовление тостов (ID)
S<-S_prob(N, 0.9, 1000)
#Моделируем прохождение событий через блок V1
R1<-V(1, S, "V1")
#Собираем статистику по работе блока и прохождению через него проектов
X<-XES(R1)
#NV2=R1 (в том числе убираем лишние данные)
NV2<-subset(R1, select=c(R, ID_Out))
colnames(NV2) <- c('S', 'ID')
R2<-V(1, NV2, "V2")
NV3<-subset(R2, select=c(R, ID_Out))
colnames(NV3) <- c('S', 'ID')
X2<-XES(R2)
X<-rbind(X,X2)
# Генерируем случайное число (0 или 1)
if (sample(0:1, 1)) { # Если выпало 1, выполняем этот блок кода
  R2 <- V(1, NV2, "V2")
  NV3 <- subset(R2, select=c(R, ID_Out))
  colnames(NV3) <- c('S', 'ID')
  X2 <- XES(R2)
  X <- rbind(X, X2)
}
R3<-R2
# Общее время прохождения тоста
l<-unique(X$ID)
s<-NA
for (i in 1:length(l)){
  s[i]<-sum(X$W[X$ID==l[i]])
}

```

В результате мы получаем доступ к собираемой при имитации статистике, которая может использоваться для анализа эффективности структурно-функциональной модели (рис. 5 и 6).

Получаемое знаково-символьное представление позволяет исследовать и сложные процессы, которые не могут быть реализованы с использованием других нотаций, например, процессы с разными типами заявок и процессы, использующие общую базу (функционал, оборудование) для их реализации (например, структурно-функциональная модель, приведенная на рис. 7).

Рассматриваемый подход позволяет строить знаково-символьное соответствие для таких случаев и получать результаты моделирования с учетом взаимного влияния разных типов заявок и при совместном использовании блоков действия (V) (например, графики загрузки каждого блока операций – рис. 8, время обслуживания заявок по типам – рис. 9).

Как видно из приведенных рассуждений методика перехода от структурно-функциональных моделей к знаково-символьному описанию и затем к программному коду не накладывает ограничений на тип и структуру рассматриваемых процессов, что позволяет исследовать методами имитационного моделирования информационные процессы, производственные

процессы, процессы, связанные с жизнедеятельностью и т.д.

При этом, построив структурно-функциональную модель, мы в автоматическом режиме можем получить знаково-символьное описание, программный код для моделирования и данные моделирования (данные и графики).

Ограничением используемого подхода является внутреннее время, которое представляется в шагах/тактах расчета (i), что позволяет работать только с дискретными системами и свидетельствует о необходимости установления соответствия шагов расчета реальному времени (это можно сделать на основе информации о производительности используемого оборудования и алгоритмов).

Для структурно-функциональных моделей в нотации Event-driven Process Methodology (EPM) может быть установлено соответствие со знаково-символьным описанием и осуществлён его перевод на существующий язык программирования. Знаково-символьное представление позволяет изучать эффективность процессной деятельности методами статистического анализа, сравнивать варианты реализации этих процессов на данных, получаемых в результате имитационного моделирования.

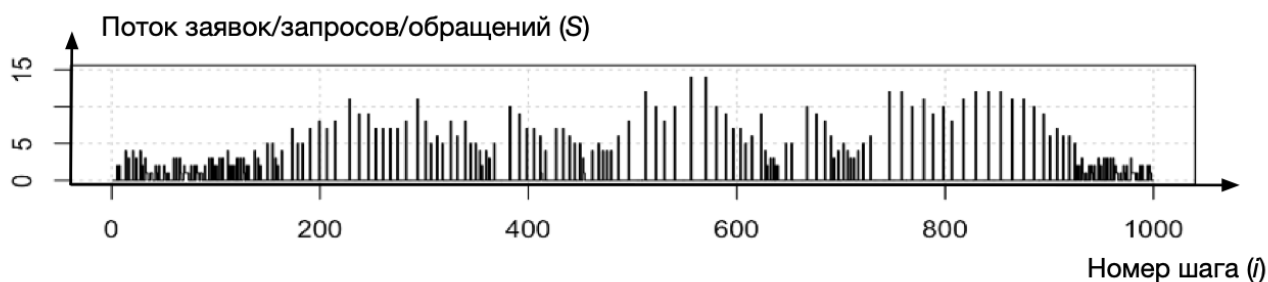


Рис. 5. Поток запросов при выполнении программы, приведенной в листинге 7, имитирующей работу структурно-функциональной модели, показанной на рис. 2

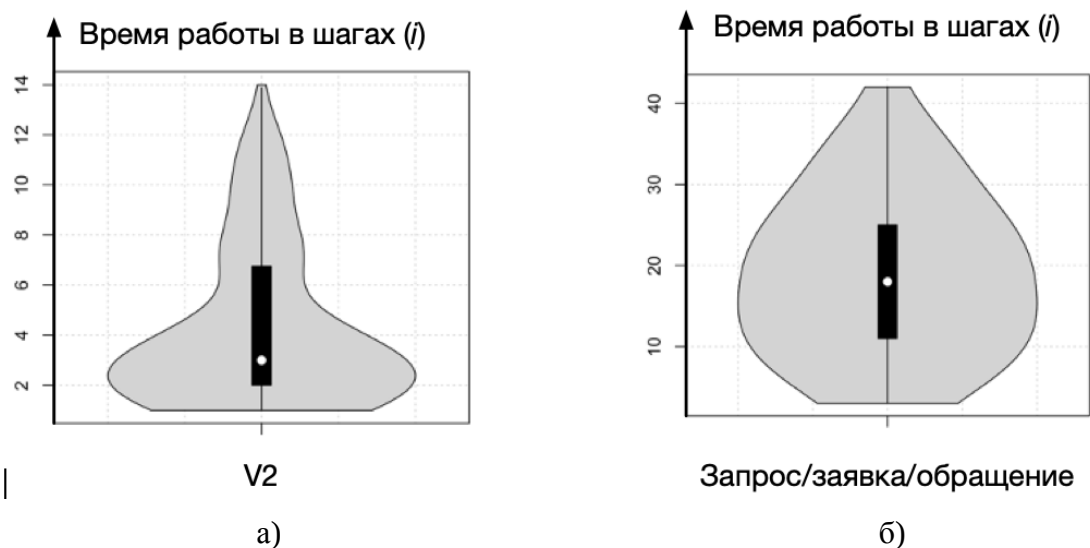


Рис. 6. Статистические данные о работе структурно-функциональной модели, приведенной на рис. 2: а) время загрузки блока V2 в шагах работы (i), б) время обработки запроса

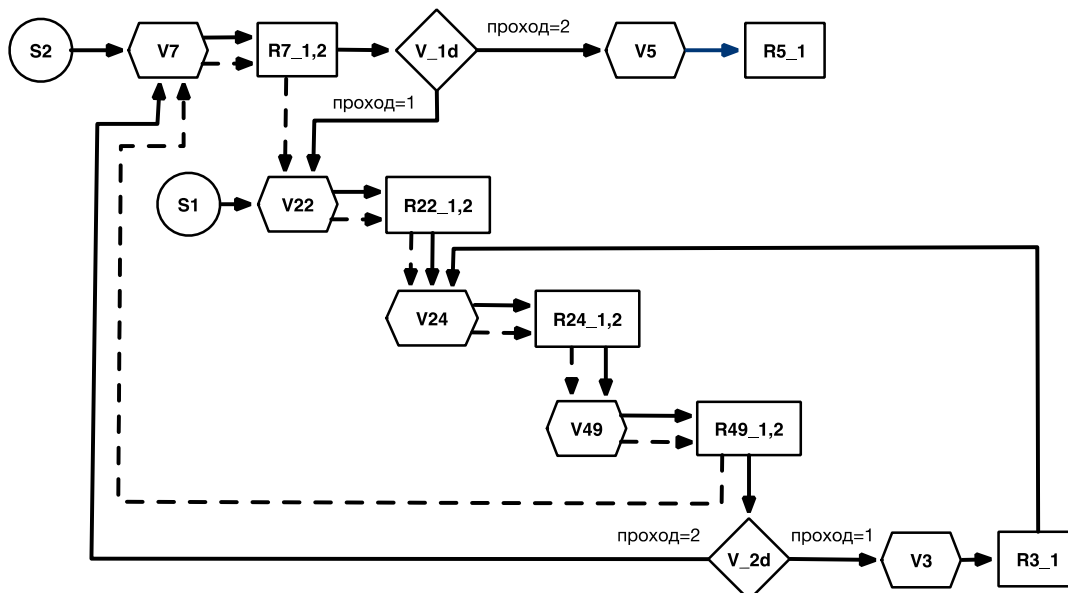


Рис. 7. Структурно-функциональная модель процесса покраски ротора (чёрный цвет, один тип заявок) и сердечника (пунктир, другой тип заявок): S2 – ротор до обработки, S1 – сердечник до обработки, V7 – контроль, V22 – обезжиривание, V24 – окраска, V49 – сушка, V3 – зачистка, V5 – исправление дефекта покрытия, R22_1,2 – сердечник/ротор после обезжиривания, R24_1,2 – сердечник/ротор после окрашивания, R49_1,2 – сердечник/ротор, прошедший процедуру сушки, R3_1 – ротор после зачистки, R7_1,2 – сердечник/ротор после проведения контроля, R5_1 – ротор после исправления дефекта покрытия, V_1d – блок условия (проход 1 = после операции контроля ротор направляется на процедуру обезжиривания, проход 2 = после операции контроля ротор направляется на операцию исправления дефекта покрытия), V_2d – блок условия (проход 1 = после операции сушки ротор направляется на процедуру зачистки, проход 2 = после операции сушки ротор направляется на процедуру контроля)

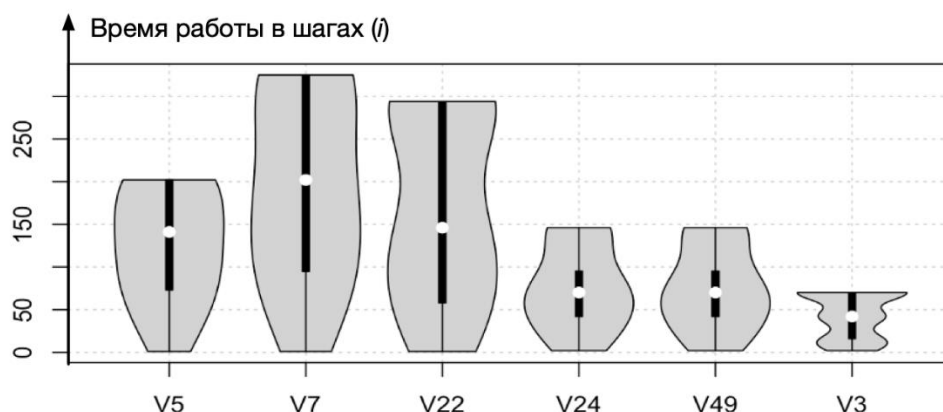


Рис. 8. Скрипичные диаграммы времени загрузки блоков операций при вероятностях появления заявок типа $P(S1) = 0,9$ и $P(S2) = 0,9$



Рис. 9. Скрипичные диаграммы времени обработки заявок типов S1 и S2 при вероятностях появления заявок типа $P(S1) = 0,9$ и $P(S2) = 0,9$

СПИСОК ЛИТЕРАТУРЫ

1. Business Process Model and Notation (BPMN). Object Management Group. – 2013. – URL: <https://www.omg.org/bpmn/>.
2. Scheer A.-W. ARIS. Modellierungsmethoden, Metamodelle, Anwendungen. 4th ed. – Berlin: Springer, 2001. – 244 p.
3. Marca D., McGowan C.L. SADT: structured analysis and design technique. – New York: McGraw Hill, 1988.
4. ArchiMate® 3.1 Specification // The Open Group. – 2019. – URL: <https://pubs.opengroup.org/architecture/archimate31-doc/copyright.html>.
5. Dori D. Object-Process Methodology. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. DOI: 10.1007/978-3-642-56209-9.
6. Delligatti L. SysML Distilled: A Brief Guide to the Systems Modeling Language. – Addison-Wesley, 2013. – 304 p.
7. Van der Aalst W. Process Mining. – Berlin, Heidelberg: Springer Berlin Heidelberg, 2016. DOI: 10.1007/978-3-662-49851-4.
8. Morgan J. The Future Leader: 9 Skills and Mindsets to Succeed in the Next Decade. – Wiley, 2020. – 300 p.
9. Питерсон Д. Теория сетей Петри и моделирование систем. – Москва: Мир, 1984. – 264 с.
10. Estruch A., Heredia Á. J. A. Event-driven manufacturing process management approach // Business Process Management: 10th International Conference – BPM 2012 (Tallinn, Estonia, September 3-6, 2012). Proceedings 10. – Berlin Heidelberg: Springer, 2012. – P. 120–133.
11. Routis I., Bardaki C., Dede G., Nikolaidou M., Kamalakis T., Anagnostopoulos D. CMMN evaluation: the modelers' perceptions of the main notation elements // Software and Systems Modeling. – Dec. 2021. – Vol. 20, № 6. – P. 2089–2109. DOI: 10.1007/s10270-021-00880-3.
12. Doyle P.G. Grinstead and Snell's Introduction to Probability. – Providence: American Mathematical Society, 2006. – 518 p.
13. Мыльников Л.А., Салтыкова А.Д. Нотация Event-driven Process Methodology для исследования информационных процессов // Научно-техническая информация. Сер. 2. – 2024. – № 7. – С. 19–30. DOI: 10.36535/0548-0027-2024-07-2.
14. Wujec T. Wicked Problem Solving (Practitioner Workbook). – Toronto, Canada: Wujec Group Inc., 2021. – 57 p.
15. Мыльников Л.А. О структурно-функциональном моделировании процессов с выделенным субъектом управления // Научно-техническая информация. Сер. 2. – 2022. – № 2. – С. 9–20; Mylnikov L.A. Structural and Functional Modeling of Processes with a Dedicated Control Subject // Automatic Documentation and Mathematical Linguistics. – 2022. – Vol. 56, № 1. – P. 42–54. DOI: 10.3103/S0005105522010083.
16. Филд А., Харрисон П. Функциональное программирование. – Москва: Мир, 1993. – 638 с.

Материал поступил в редакцию 06.09.24.

Сведения об авторах

МЫЛЬНИКОВ Леонид Александрович – кандидат технических наук, доцент, заведующий научно-учебной лабораторией междисциплинарных эмпирических исследований Национального исследовательского университета – «Высшая школа экономики»; доцент кафедры Информационные технологии и автоматизированные системы Пермского национального исследовательского политехнического университета
e-mail: LAMylnikov@hse.ru

САЛТЫКОВА Анастасия Дмитриевна – младший научный сотрудник научно-учебной лаборатории междисциплинарных эмпирических исследований Национального исследовательского университета – «Высшая школа экономики»
e-mail: ADSaltykova@hse.ru

Об отражении в ДСМ-методе автоматизированной поддержки исследований причинности в медицине (на примере установления связи негативной шизофрении с генетическими параметрами)

Рассматриваются некоторые теории причинности и связь с ними ДСМ-метода автоматизированной поддержки исследований (ДСМ-метода АПИ), который способен обнаруживать в данных различные типы причинно-следственных зависимостей с помощью соответствующих стратегий. Приводятся некоторые подходы к причинности в медицине (единственность причины, множественность причин, независимые причины) и их отражение в ДСМ-методе АПИ.

На примере исследования установления связи негативной шизофрении с генетическими параметрами показан выбор подходящего типа причинности, определяемого в ДСМ-методе АПИ с помощью стратегий. В рассматриваемом исследовании использовались две стратегии ДСМ-метода АПИ: с использованием метода простого сходства с запретом на контрпример и обобщенного метода. Применение первого метода продемонстрировало низкий показатель объясняемости порождаемыми гипотезами исходного массива данных, что не удовлетворяет критерию абдуктивного принятия гипотез. Применение обобщенного метода улучшило показатели объясняемости, что может свидетельствовать о более удачном представлении связи генетических параметров с заболеванием, изучаемым в исследовании. Это представление связано с тернарным отношением сходства, лежащим в основе обобщенного метода: «причина — множество тормозов (блокаторов) причины — эффект».

Ключевые слова: ДСМ-метод автоматизированной поддержки исследований, причинность, исследование, гены, ассоциированные с шизофренией

DOI: 10.36535/0548-0027-2024-12-2

ВВЕДЕНИЕ

ДСМ-метод автоматизированной поддержки исследований (ДСМ-метод АПИ) – это метод интеллектуального анализа (*knowledge discovery*), направленный на извлечение знания из данных. Его особенность заключается в обнаружении зависимостей причинно-следственного типа.

В основе ДСМ-метода АПИ лежат идеи Д.С. Милля [1] об индуктивном обобщении как причине изучаемого эффекта. Причинно-следственная связь представляется в виде бинарного отношения «причина — эффект» или тернарного отношения «причина — множество тормозов (блокаторов) причины — эффект». Тернарное отношение использует идею Д.С. Милля о том, что «условием применимости методов индуктивного вывода являются не только установление сходства рассматриваемых фактов, но и отсутствие препятствий,

противодействующих гипотезе о причине; такими препятствиями могут быть другие причины» [2].

ДСМ-метод АПИ содержит разные типы причинности, соответствующие различным стратегиям. С помощью оценок он позволяет подбирать тип причинности, который лучше всего отражает отношения в исследуемых данных.

В ДСМ-методе АПИ возможно существование множественности причин по Д.С. Миллю: «неверно, будто каждое единичное следствие должно быть связано с одной только причиной, с одним рядом условий, будто всякое явление может быть произведено лишь одним путём. Часто существует несколько независимых друг от друга способов, при помощи которых можно вызвать одно и то же явление» [1, с. 343]. Эта идея соответствует тому, что для исследуемого с помощью метода эффекта может быть множество гипотез о причинах.

Цель настоящей статьи – обсуждение процесса выбора стратегии ДСМ-метода АПИ, соответствующей типу причинности, на примере проведенного исследования [3] обнаружения генетических механизмов формирования эмоционального и волевого дефицитов при шизофрении, протекающей с преобладанием негативных расстройств. Для этого мы хотим кратко рассмотреть: причинности в науке, в целом и в медицине в частности и причинность в ДСМ-методе АПИ и её представление с помощью стратегий и методов, а также связать их на примере выбора стратегии для приведенного исследования.

О ПРИЧИННОСТИ В ЭПИСТЕМОЛОГИИ, ПРИКЛАДНЫХ ОБЛАСТЯХ И ДСМ-МЕТОДЕ АПИ

Обнаружение причинно-следственных связей – это одна из актуальных задач искусственного интеллекта (ИИ) и машинного обучения (МО). Необходимость интеграции МО и анализа причинно-следственных отношений отмечается в отчете State of AI [4] за 2020 год. В другом отчете — One Hundred Year Study on AI (AI100) за 2021 г. [5, SQ3] — одной из важных открытых проблем ИИ называется выявление и использование причинности. Джуда Перл, создатель байесовских сетей и один из основателей направления *Causal Inference*, считает причинно-следственные рассуждения (*causal reasoning*) необходимым компонентом общего ИИ (*artificial general intelligence, AGI*) [6]. Анализ причинно-следственных связей актуален в социальных науках, медицине, экономике.

В эпистемологии существуют различные подходы к причинности: как к регулярности [7], вероятностный подход [8], контрфактуальный подход [9] и многие другие. Понимание причинности в ДСМ-методе АПИ связано с регулярностью. В статье Стэнфордской энциклопедии философии «Regularity and Inferential Theories of Causation» [7] отмечаются особенности таких подходов:

1) следствие всегда регулярно следует за причиной; регулярная связь понимается противоположно отношению причинной силы («This regular association is to be understood by contrast to a relation of causal power»);

2) теория восходит к идеям о регулярности Д. Юма и Д.С. Милля;

3) закономерность (*regularities*) связана с выводом: следствие выводится из некоторых причин в некоторой теории с использованием подходящей логики.

Причинность является важной концепцией в различных прикладных областях, в частности, в медицине. Выявление причин заболеваний направлено на эффективное лечение и профилактику. Основная медицинская наука, занимающаяся проблемами причинности в медицине – это эпидемиология.

В [12] выделяются две концепции причинности:

(1) «единственная причина» – болезнь имеет одну причину, а отдельная причина вызывает одну болезнь;

(2) «сочетанного действия комплекса причин» – болезни имеют несколько причин, а отдельная причина может вызывать разные болезни.

Первая концепция приобрела популярность во второй половине XIX в. в связи с исследованием ин-

фекционных болезней: концепция Р. Коха о «единственной и специфической причине» при оценке роли микроорганизмов (цит. по [13]). Вторая концепция появилась, когда оказалось невозможным определить «единственную причину» для некоторых инфекционных и, особенно, неинфекционных заболеваний.

Философ Д. Гиллис в книге «Causality, Probability, and Medicine» [13] отмечает, что в XVII-XIX вв. в медицине (и в науке в целом) доминировало представление о детерминистской причинности (*deterministic causality*): при всех случаях наступления события А за ним следует событие В («A causal claim of the form A causes B is deterministic, only if, *ceteris paribus*, whenever A occurs, it is followed by B»). Однако с середины XX в. под влиянием активно развивающейся эпидемиологии наибольшую популярность в медицине получила концепция индетерминированной причинности (*indeterministic causality*), в которой за событием А не всегда следует событие В («if a casual claim of the form A causes B is indeterministic, then A can occur without always being followed by B»). Примером может служить утверждение: «Курение вызывает рак легких» – что, очевидно, не всегда верно.

Для обозначения индетерминированных причин часто используется понятие «факторы риска» [12]. К ним относят элементы социальной и природной среды, особенности поведения людей, индивидуальные особенности организма. Кроме факторов риска могут существовать «факторы антириска», связанные со снижением количества новых случаев заболевания или его предотвращением.

Причинность в ДСМ-методе АПИ определяется как «операциональная обусловленность эффекта, сохраняющаяся при расширении данных» – *качественная эмпирическая причинность*). «Качественная» означает, что «причинность определяется не метрическими условиями, а зависимостями отношений», а «эмпирическая» подразумевает «зависимость результатов ДСМ-исследований от множества фактов» [10].

Операциональность в ДСМ-методе АПИ включает [11]:

- операцию сходства, интерпретируемую как выделение общего фрагмента у двух и более примеров, а также отношение сходства, определяемое на основе этой операции;

- М-предикаты сходства, использующие отношение сходства;

- Стратегии, определяемые двумя М-предикатами сходства (для группы с наличием исследуемого эффекта и для группы с отсутствием рассматриваемого эффекта).

ОБ ИССЛЕДОВАНИИ

Рассмотрим в качестве примера медицинское исследование [3], – имея задачу установить, одинаковы или различны генетические механизмы формирования эмоционального и волевого дефицитов при шизофрении, протекающей с преобладанием негативных расстройств. Кратко перечислим основные особенности этой задачи:

1) шизофрения – это мультифакторное заболевание, включающее различные генетические и средовые влияния [15];

2) единственного гена шизофрении не существует [16] (установлено около трехсот ассоциированных с шизофренией локусов в генах), предполагаемым фактором заболевания является сложная система взаимодействия генов;

3) набор генетических и средовых факторов частично не идентичен у отдельных индивидуумов;

4) всё, отмеченное в пп. 1-3, касается не только шизофрении в целом, но и отдельных дименсий заболевания, среди которых, в том числе, негативные расстройства: волевой и эмоциональный дефицит (о генетической предикции их возникновения см. в [17]).

Цель исследования – порождение знаний о причинно-следственных зависимостях между генетическими параметрами и двумя дименсиями негативных расстройств при шизофрении. Для этого исследования использовался ДСМ-метод АПИ.

Одно из условий применения ДСМ-метода АПИ – возможность выделения в исходных данных примеров с наличием исследуемого эффекта и примеров с отсутствием такого эффекта, называемых, соответственно, положительными и отрицательными примерами ((+)- и (-)-примерами).

Мы использовали метод простого сходства с запретом на контрпример и обобщенный метод. Далее приведем неформальное описание этих методов.

В методе простого сходства с запретом на контрпример (обозначается как *ab*) заложено представление порождаемого причинно-следственного отношения в виде пары «причина — эффект». Причина является фрагментом, общим для нескольких исходных примеров. Исходные примеры, породившие причину, называются «родителями гипотезы». Количество родителей (обозначается как *k*) – важная характеристика, которая отражает представленность полученной гипотезы о причине в исходном массиве данных.

При использовании метода с запретом на контрпример общий фрагмент в примерах-родителях сравнивается с примерами из противоположной группы (для положительных – в отрицательных и наоборот). Этот общий фрагмент становится гипотезой, только если не содержится ни в одном примере противоположной группы (такой пример из противоположной группы называется контрпримером).

Отметим, что гипотезы (отобранные фрагменты) могут быть «неатомарными». Например, в нашем исследовании гипотезы могут содержать несколько генов и, таким образом, формировать некоторый неделимый с точки зрения задачи кластер, взаимодействия внутри которого требуют интерпретации на биологическом и химическом уровнях.

В обобщенном методе причинно-следственные отношения представляются в виде троек «причина — множество тормозов (блокаторов) причины — эффект». Тормоз также является фрагментом, общим для нескольких исходных примеров.

При использовании обобщенного метода (обозначается как *ag*) полученный общий фрагмент также сравнивается с объектами из противоположной группы. Но, если этот фрагмент содержится в каких-либо объектах противоположной группы, то предполагается, что эта гипотеза не приводит к исследуемому

эффекту в силу наличия какого-то другого, блокирующего, фрагмента, называемого тормозом. Если впоследствии процедура (алгоритм для обобщенного метода приведен в [18]) может найти тормоз (или несколько тормозов, *множество тормозов*), то фрагмент входит в список окончательных гипотез вместе с множеством тормозов. Множество тормозов можно интерпретировать как «внутренний контекст» причины [19] («внешний контекст» – ситуация – рассматривается в ситуационном ДСМ-методе АПИ [20]). Таким образом, обобщенный метод вводит отношения между формируемыми кластерами.

В нашей работе была использована база данных (БД) больных шизофренией, создаваемая в лаборатории клинической генетики Научного центра психического здоровья (ФГБНУ НЦПЗ), из которой формировались базы фактов (положительных и отрицательных примеров) для проведения экспериментов.

Каждая запись из БД содержит четыре гена: BDNF (ген, кодирующий мозговой нейротрофический фактор; полиморфизм Val66Met), 5HTR2A (ген серотонинового рецептора 2a, полиморфизм rs6313 или T102C или 102C>T), HTTLPR (ген переносчика серотонина, полиморфизм HTTLPR-L/S), ZNF804A (ген цинк-пальцевого белка, полиморфизм rs1344706). Каждый ген представлен двухбуквенным сочетанием, соответствующим двум аллелям гена.

В табл. 1. приведем примеры представления генетической информации.

Таблица 1

Пример записей с генетической информацией о людях

№ п/п	BDNF	5HTR2A	HTTLPR	ZNF804
1	AA	TC	LL	AC
2	GG	CC	LL	AC

Каждая запись о пациенте в исходном массиве содержит клинические данные, представляющие собой формализованные описания симптоматики шизофрении согласно методике PANSS (*Positive and Negative Syndrome Scale*, Шкала позитивных и негативных синдромов). На основе этих данных были сформированы группы пациентов со снижением эмоций и воли, которые в экспериментах представляют собой эффекты.

В итоговую БД для исследования вошли 64 записи о пациентах по типу эмоционального дефицита и 172 записи о пациентах по типу волевого снижения. Эта БД содержала также 126 записей о здоровых добровольцах, не больных шизофренией. В этих записях нет клинических данных о заболевании.

В рамках исследования [3] было проведено четыре эксперимента – ДСМ-рассуждения.

Эксперимент 1: сравнение группы пациентов с эмоциональным снижением (положительные примеры) со здоровыми людьми (отрицательные примеры);

Эксперимент 2: сравнение группы пациентов с волевым снижением (положительные примеры) со здоровыми людьми (отрицательные примеры);

Эксперимент 3: сравнение группы пациентов с эмоциональным снижением (положительные примеры) с группой, которая объединяет здоровых людей и пациентов с волевым снижением (отрицательные примеры);

Эксперимент 4: сравнение группы пациентов с волевым снижением (положительные примеры) с группой, которая объединяет здоровых людей и пациентов с эмоциональным снижением (отрицательные примеры);

Каждое из четырех ДСМ-рассуждений было проведено сначала с использованием метода с запретом на контрпример, затем с применением обобщенного метода. Переход к другому методу был обусловлен тем, что одна из значимых оценок ДСМ-рассуждения — показатель объясняемости (ρ^σ , $\sigma \in \{+, -\}$) полученными гипотезами исходной базы фактов — имел низкое значение при использовании метода с запретом на контрпример. Показатель объясняемости лежит в основе абдуктивного принятия полученных гипотез. Он рассчитывается как доля объясняемых гипотезами примеров в общем количестве примеров в базе фактов и должен быть не меньше выбранного порога (обычно 0,8) для принятия порождаемых гипотез.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

В эксперименте 1 при использовании метода простого сходства с запретом на контрпример объясняемость составила 0,15 ($\rho^+=0,15$), при использовании обобщенного метода — 0,89 ($\rho^+=0,89$).

В эксперименте 2 при использовании метода простого сходства с запретом на контрпример объясняемость составила 0,16 ($\rho^+=0,16$), при использовании обобщенного метода — 0,8 ($\rho^+=0,8$).

В эксперименте 3 при использовании метода простого сходства с запретом на контрпример объясняемость составила 0,03 ($\rho^+=0,03$), при использовании обобщенного метода — 0,24 ($\rho^+=0,24$).

В эксперименте 4 при использовании метода простого сходства с запретом на контрпример объясняемость составила 0,11 ($\rho^+=0,11$), при использовании обобщенного метода — 0,58 ($\rho^+=0,58$).

Все полученные гипотезы о причинно-следственных зависимостях были представлены для анализа экспертам в виде высказываний, соответствующих бинарным и тернарным отношениям причинности. Примеры таких высказываний приведены на рис. 1 (для бинарного предиката) и рис. 2 (для тернарного предиката).

Улучшение показателей объясняемости (хотя и недостаточное в эксперименте 3 и 4 для абдуктивного принятия) при смене используемого метода во всех четырех экспериментах может свидетельствовать о том, что для исследуемой задачи лучше подходит отношение причинности, учитывающее внутренних контекст (обобщенный метод): оно может более точно отображать взаимодействие генов (ген-ген интеракции [21]).

С точки зрения машинного обучения (МО), изменение бинарного отношения причинности на тернарное может служить примером решения проблемы недообучения (*underfitting*) — увеличения сложности используемой модели.

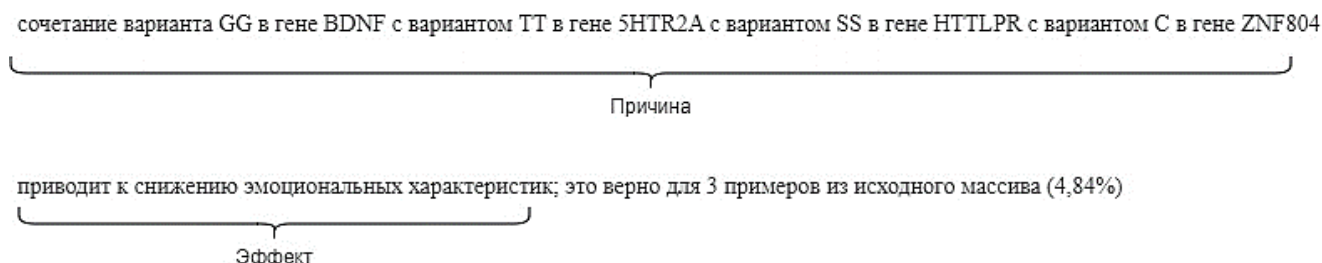


Рис. 1. Пример высказывания для представления гипотезы о причине (бинарное отношение)

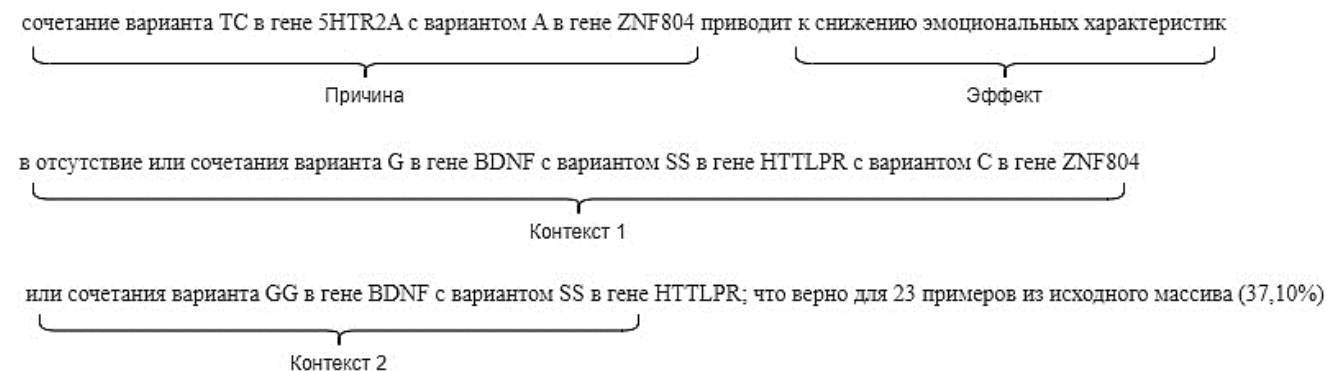


Рис. 2. Пример высказывания для представления гипотезы о причине (тернарное отношение)

Вывод о выборе тернарного отношения причинности подтверждается заключением экспертов о том, что порожденные гипотезы, возможно, отражают явление эпистаза. Термин «эпистаз» был впервые введен Уильямом Бейтсоном в 1908 г. для описания маскирующего эффекта, при котором специфическому варианту или аллелю в одном локусе не позволяет проявить свои эффекты вариант из другого локуса [22]. Эпистаз, также часто используемый для общего обозначения взаимодействия между различными генами, в последние годы стал ключевым элементом в молекулярно-генетических исследованиях психических расстройств [23]. Как возможный механизм, играющий роль при шизофрении, эпистаз имеет особое значение, потому что, если эффект одного локуса изменен или замаскирован эффектами в другом локусе, то способность обнаружить первый локус, вероятно, будет снижена [24]. Следовательно, исследование комбинированных эффектов этих двух локусов также будет затруднено таким взаимодействием. Если задействовано более двух локусов, то добавляется еще один уровень сложности во взаимосвязи генотипа и риска заболевания из-за возможности «многоходового» взаимодействия.

Представление о взаимодействии генов лежит в основе концепции генных сетей — «групп координированно функционирующих генов, взаимодействующих друг с другом как через свои первичные продукты (РНК и белки), так и через разнообразные метаболиты и другие вторичные продукты функционирования генных сетей» [25]. Реконструкция генных сетей для различных заболеваний — актуальная задача биоинформатики. Для этого используются различные методы статистики, машинного обучения, анализа текстов (пример реконструкции генной сети для шизофрении см. в [26]).

Таким образом, обнаружение взаимодействующих кластеров генов с помощью ДСМ-метода АПИ может стать инструментом исследования генных сетей.

ЗАКЛЮЧЕНИЕ

В настоящей статье на примере задачи исследования генетических механизмов формирования эмоционального и волевого дефицитов при шизофрении, протекающей с преобладанием негативных расстройств, показано, как происходит подбор типа причинности в ДСМ-методе АПИ. При этом рассматривались только два типа причинности, определяемые методом простого сходства с запретом на контрпример и обобщенным методом (тернарная причинность). Однако количество типов причинности в ДСМ-методе АПИ значительно шире. В качестве оценки для выбора причинности выступал показатель объясняемости порожденными гипотезами причинно-следственных отношений исходного массива данных. Эта оценка может быть расширена другими показателями (например, верификацией предсказаний, выполняемых порождаемыми причинами, устойчивостью результатов при расширении массива данных — обнаружением эмпирических закономерностей [27]). По заключению экспертов показавший лучшую объясняемость обобщенный метод, предлагающий тернарное отношение причинности («причи-

на — множество тормозов (блокаторов) причины — эффект»), может более точно отражать существующие биологические механизмы изучаемого заболевания — взаимодействие генов.

СПИСОК ЛИТЕРАТУРЫ

1. Милль Дж. Ст. Система логики силлогистической и индуктивной: Изложение принципов доказательства в связи с методами научного исследования: пер. с англ. / предисл. и прилож. В. К. Финна. Изд. 5-е, испр. и доп. — Москва: ЛЕНАНД, 2011. — 832 с.
2. Финн В.К., Шестерникова О.П. О новом варианте обобщенного ДСМ-метода // Искусственный интеллект и принятие решений. — 2016. — № 1. — С. 57-64.
3. Шестерникова О.П., Фабрикантова Е.Ф., Голимбет В.Е., Лежейко Т.В., Романов Д.В. Эмпирическая теория формирования эмоционального и волевого дефицита при шизофрении с преобладанием негативных расстройств, полученная на основе анализа связи клинических и генетических данных с применением ДСМ-метода автоматизированной поддержки исследований // Врач и информационные технологии. — 2024. — № 2. — С. 82-90. DOI: 10.25881/18110193_2024_2_82.
4. State of AI Report. — URL: <https://www.stateof.ai>.
5. Littman M.L., Ajunwa I., Berger G. et al. Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report. — Stanford, CA: Stanford University, September 2021. — URL: <http://ai100.stanford.edu/2021-report>.
6. Pearl J. AI, and Causality: What Role Do Statisticians Play? // Amstatnews. — 2023. — 1 sept. — URL: <https://magazine.amstat.org/blog/2023/09/01/judeapearl>.
7. Holger A., Guenther M. Regularity and Inferential Theories of Causation // The Stanford Encyclopedia of Philosophy (Fall 2021 Edition) / ed. Edward N. Zalta. — URL: <https://plato.stanford.edu/archives/fall2021/entries/causation-regularity/>
8. Hitchcock Ch. Probabilistic Causation // The Stanford Encyclopedia of Philosophy (Spring 2021 Edition) / ed. E.N. Zalta. — URL: <https://plato.stanford.edu/archives/spr2021/entries/causation-probabilistic>.
9. Menzies P., Beebe H. Counterfactual Theories of Causation // The Stanford Encyclopedia of Philosophy (Summer 2024 Edition) / eds. Edward N. Zalta, Uri Nodelman. — URL: <https://plato.stanford.edu/archives/sum2024/entries/causation-counterfactual/>
10. Финн В.К. ДСМ-метод автоматизированной поддержки исследований и аппарат понятий для искусственного интеллекта // Искусственные общества. — 2024. — Т. 19, Вып. 2. — URL: <https://artsoc.jes.su/s207751800030907-6-1/>. DOI: 10.18254/S207751800030907-6
11. Финн В.К. Дистрибутивные решетки индуктивных процедур // Научно-техническая информация. Сер. 2. — 2014. — № 11. — С. 1-30.
12. Савилов Е.Д., Астафьев В.А., Анганова Е.В., Кичигина Е.Л., Жданова С.Н.

- Причинность в эпидемиологии: пособие для врачей. – Иркутск: РИО ГБОУ ДПО ИГМАПО, 2015 – 36 с.
13. Gillies D. Causality, Probability, and Medicine. – London: Routledge, 2018.
14. Stilo S.A., Murray R.M. Non-Genetic Factors in Schizophrenia // Current psychiatry reports – 2019 Sep 14. – № 21(10):100. DOI: 10.1007/s11920-019-1091-3.
15. Незнанов Н.Г., Рукавишников Г.В., Касьянов Е.Д., Филиппов Д.С., Кибитов А.О., Мазо Г.Э. Биопсихосоциальная модель в психиатрии как оптимальная парадигма для современных биомедицинских исследований. Обзорные психиатрии и медицинской психологии. – 2020. – № 2. – С. 3-15. DOI: 10.31363/2313-7053-2020-2-3-15.
16. Резник А.М., Костюк Г.П., Морозова А.Ю., Захарова Н.В. Проблемы предпосылок шизофрении по данным молекулярно-генетических исследований // Health, Food & Biotechnology. – 2019. – №1. – URL: <https://cyberleninka.ru/article/n/problemy-predposylok-shizofrenii-po-dannym-molekulyarno-geneticheskikh-issledovaniy-1> (дата обращения: 01.01.2023).
17. Golimbet V., Lezheiko T., Korovaitseva G., Kolesina N., Plakunova V., Mikhailova V. Immune-related genes are differentially associated with negative symptoms subdomains in patients with schizophrenia // European Psychiatry. – 2022. – № 65(S1). – P. S286-S286. DOI: 10.1192/j.eurpsy.2022.732.
18. Шестерникова О.П., Финн В.К., Леско К.А., Винокурова Л.В. О применении ДСМ-метода автоматизированной поддержки исследований для предсказания развития сахарного диабета у больных хроническим панкреатитом // Научно-техническая информация. Сер. 2. – 2023. – № 4. – С. 16 – 25; Shesternikova O.P., Finn V.K., Lesko K.A., Vinokurova L.V. Application of the JSM Method of Automated Research Support to Predict Diabetes Development in Patients with Chronic Pancreatitis // Automatic Documentation and Mathematical Linguistics. – 2023. – Vol. 57, № 2. – P. 91–100. DOI: <https://doi.org/10.3103/S0005105523020085>.
19. Гусакова С.М., Михеенкова М.А. О формировании эмпирических теорий в плохо формализованных областях // Труды Двадцать первой Национальной конференции по искусственному интеллекту с международным участием КИИ-2023. – Москва: РАИИ, 2023. – Т. 1. – С. 169-179.
20. Финн В.К., Михеенкова М.А. О ситуационном расширении ДСМ-метода автоматического порождения гипотез // Автоматическое порождение гипотез в интеллектуальных системах – Москва: URSS, 2009. – 528 с.
21. Von Cheong E., O'Tuathaigh C.M.P. Modeling Gene–Gene Interactions in Schizophrenia // Handbook of Behavioral Neuroscience. – 2016. – Vol. 23. – P. 327–343. Elsevier. DOI: <https://doi.org/10.1016/B978-0-12-800981-9.00019-5>.
22. Bateson W. Mendel's principles of heredity. – Cambridge: Cambridge University Press, 1909.
23. Gelernter J. Genetics of complex traits in psychiatry // Biological Psychiatry. – 2015. – № 77. – P. 36–42.
24. Cordell H.J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans // Human Molecular Genetics. – 2002. – № 11. – P. 2463–2468.
25. Колчанов Н.А., Игнатъева Е.В., Подколотная О.А., Лихошвай В.А., Матушкин Ю.Г. Генные сети // Вавиловский журнал генетики и селекции. – 2013. – Т. 17, № 4/2. – С. 833-850. – URL: <https://vavilov.elpub.ru/jour/article/viewFile/206/208>
26. Дохоян А.Ю., Глущенко М.В., Орлов Ю.Л. Реконструкция генной сети шизофрении для поиска генов-мишеней // Ульяновский медико-биологический журнал. – 2022. – № 3. – С. 6–22. DOI: 10.34014/2227-1848-2022- 3-6-22.
27. Финн В.К. Об эмпирических закономерностях и интеллектуальном анализе данных в ДСМ-методе автоматизированной поддержки исследований // Научно-техническая информация. Сер. 2. – 2023. – № 12. – С. 14-33; Finn V.K. On Empirical Regularities in the JSM Method of Automated Research Support // Automatic Documentation and Mathematical Linguistics. – 2023. – Vol. 57, № 6. – P. 362– 381.

Материал поступил в редакцию 09.10.24.

Сведения об авторах

ШЕСТЕРНИКОВА Ольга Павловна – ведущий программист, Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва
e-mail: oshesternikova@gmail.com

ФАБРИКАНТОВА Елена Федоровна – кандидат технических наук, главный специалист, Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва
e-mail: el.fabrikantova@ya.ru

РОМАНОВ Дмитрий Владимирович – доктор медицинских наук, профессор, ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» (Сеченовский Университет), ведущий научный сотрудник, ФГБНУ «Научный центр психического здоровья», Москва
e-mail: newt777@mail.ru

Исследование тематической и терминологической трансформации коллекций научных статей по междисциплинарным областям знания*

Представлена методика количественного исследования динамики изменения тематической структуры коллекций и набора употребляемых терминов, направленная на отбор документов в тематические коллекции, создание тематических (классификационных и предметных) профилей этих коллекций в различные хронологические периоды и сравнительный анализ этих профилей. Профили рассматриваются как дескриптивные множества, в качестве веса ключевого слова используется TF-IDF. Методика апробирована на двух коллекциях: «Робототехника и роботизированные системы» и «Интеллектуальные системы, искусственный интеллект и машинное обучение».

Ключевые слова: наукометрия, информационный поиск, статьи в научных журналах, тематические коллекции статей, тематические профили коллекций, ГРНТИ, терминологический анализ, TF-IDF, дескриптивные множества, коэффициент сходства, БД ВИНТИ РАН

DOI: 10.36535/0548-0027-2024-12-3

ВВЕДЕНИЕ

Наукометрические исследования различных источников научно-технической информации (журнальных статей, докладов на конференциях, монографий, патентных документов) с целью выявления инновационных и перспективных направлений развития науки и техники выполняются схожими методами, но результаты этих исследований используются по-разному. Государственные органы, научные фонды, университеты, корпорации определяют наиболее приоритетные направления науки и технологии для коррекции государственных программ, грантов, образовательных программ, инвестиционных стратегий. Центры научно-технической информации и научно-технические библиотеки используют результаты наукометрических исследований для повышения качества информационного обеспечения науки, решая следующие задачи:

- организация поступления во входной поток информационного центра (или в фонд библиотеки)

новых документальных источников информации, отражающих приоритетные области научных исследований и опытно-конструкторских работ;

- коррекция номенклатуры издаваемых информационных продуктов (реферативных журналов, библиографических указателей, баз данных, информационно-аналитических обзоров;

- совершенствование классификационных схем, используемых для индексирования научно-технической литературы;

- дополнение новой научной лексикой терминологических словарей и информационно-поисковых тезаурусов, используемых для индексирования литературы ключевыми словами.

Новые технологии определяются следующими атрибутами: радикальная новизна, быстрый рост, согласованность, заметное влияние на процессы производства знаний, а также неопределенность и неоднозначность будущего влияния на социально-экономическую сферу. Показатели радикальной новизны и согласованности новых технологий могут быть формализованы наукометрическими методами. Согласованность и её сохранение с течением времени отличают технологии, которые приобрели определенную устойчивость, от тех, которые всё еще находятся в состоянии постоянного изменения. Первоначально может ис-

* Работа выполнена в рамках государственного задания ВИНТИ РАН по теме № FFFU-2022-000 «Аналитические исследования потока научно-технической информации и формирование тематических, классификационных и предметных профилей по приоритетным направлениям науки, техники и технологий».

пользоваться множество терминов, и уменьшение их количества, а также появление сокращений или аббревиатур может сигнализировать о постоянстве и указывать на общие интерпретации и, таким образом, на согласованность. Создание специализированных конференций, специальных выпусков и новых специализированных журналов может являться дополнительным признаком согласованности, также как и появление новых категорий в устоявшихся системах классификации [1].

Наличие множества терминов, обозначающих наукометрические методы выявления и прогнозирования новых и растущих направлений научных исследований и технологий, свидетельствует о новизне данной области исследования и пока ещё слабой согласованности, как например, в следующих англоязычных публикациях: *tech mining (TM)* [2, 3], *technology emergence (TE)*, *technology opportunities analysis (TOA)*, *technological forecasting (TF)* [4], *ERT (emerging research topics)* [5].

Отсутствие стабильной терминологии затрудняет поиск релевантных публикаций, так как одинаковые или очень близкие по смыслу понятия часто бывают скрыты под разными терминами и аббревиатурами. Наиболее простая классификация методов *ERT* предложена в [5]. Методы делятся по следующим категориям:

- 1) основанные на цитировании (*citation-based methods*);
- 2) основанные на лексике (*lexical-based methods*);
- 3) гибридные подходы, основанные на цитировании и лексике (*citation and lexical hybrid approaches*).

Классификационные схемы относятся к информационно-поисковым языкам (ИПЯ), а коды классификаций (УДК, ГРНТИ, МПК и т.д.) являются лексическими единицами этих языков. Поэтому наукометрические методы, основанные на статистике классификации источников информации, правомерно отнести ко второй категории.

В предыдущих исследованиях [6, 7] мы изучали динамику трансформации тематических профилей серийных изданий, используя тематические коллекции документов, извлечённые из базы данных ВИНТИ РАН, которые удовлетворяли двум условиям:

- тематика коллекции достаточно точно описывается набором рубрик Государственного рубрикатора научно-технической информации (ГРНТИ) [8];
- документы достаточно полно классифицируются этими рубриками ГРНТИ в реферативной базе данных ВИНТИ РАН (БД ВИНТИ).

Выполнение этих условий позволяет отбирать документы в тематическую коллекцию, составляя поисковый запрос из кодов классификационной схемы – кодов ГРНТИ.

В настоящей работе исследуются тематические коллекции, которые этим условиям удовлетворяют не полностью. Хотя в ГРНТИ присутствуют рубрики «28.23 Искусственный интеллект» (ИИ) и «55.30 Робототехника», этими рубриками индексируются далеко не все документы, релевантные этим рубрикам. Обычно документу присваивается некоторое количество рубрик ГРНТИ, которое отражает не все его те-

матические аспекты. Поскольку «искусственный интеллект» и «робототехника», являясь междисциплинарными научными дисциплинами, проникли во многие области науки и техники, и этот процесс продолжается, то документы в БД ВИНТИ часто индексируются только теми рубриками ГРНТИ, которые отражают область применения ИИ или робототехники (Машиностроение, Транспорт, Медицина, Космические исследования и т.д.). Можно провести аналогию с математическими методами и программным обеспечением, которые давно применяются практически во всех отраслях науки, но это не даёт основания индексировать большинство статей рубриками из разделов «27. Математика» или «50.41 Программное обеспечение вычислительных машин, комплексов и сетей».

Ключевые слова в совокупности с названием и аннотацией статьи более полно отражают тематические аспекты документа, что позволяет найти релевантные публикации, даже если они в явном виде не были классифицированы рубриками ГРНТИ 28.23 и 55.30. Кроме того, при помощи ключевых слов можно найти релевантные документы, которые вообще не классифицировались рубриками ГРНТИ, в том числе, в других российских и зарубежных базах данных. На сегодняшний день ВИНТИ РАН является единственной организацией в России, в которой предметные специалисты выполняют постатейную классификацию опубликованной научно-технической литературы в области естественных, технических и точных наук. В других базах данных используются предметные рубрикаторы с меньшим количеством рубрик и иерархических уровней, которые предназначены лишь для сужения области поиска. При этом основным инструментом поиска являются запросы, состоящие из ключевых слов. Поэтому отбор ключевых слов для создания таких запросов является первым и необходимым шагом для поиска и отбора документов в тематические коллекции.

Для сравнительного анализа тематических коллекций журнальных статей и изменения их тематической структуры во времени выбраны два хронологических периода (2017-2019 и 2021-2023) и два тематических направления, названия которых сформулированы следующим образом:

- Робототехника и роботизированные системы (РРС);
- Интеллектуальные системы, искусственный интеллект и машинное обучение (ИСИМО).

МЕТОДИКА ИССЛЕДОВАНИЯ

Тематику коллекции статей или отдельного серийного издания можно определить статистическим методом и представить в виде тематического профиля. Тематический профиль отдельного серийного издания, коллекции статей или множества серийных изданий, включённых в тематическое направление, может быть двух видов: классификационный и предметный. Основу классификационного профиля составляет перечень классификационных кодов (в нашем случае кодов рубрик ГРНТИ), в основе предметного профиля – перечень ключевых слов, которые используются в качестве предметных рубрик.

Тематические профили обоих видов (классификационный и предметный) можно рассматривать как дескриптивные множества. Дескриптивное множество – конечное множество, каждому элементу которого поставлено в соответствие неотрицательное число («вес») [9].

Классификационный профиль можно представить в виде конечного множества пар значений $(c_i, \text{Frequency}(c_i))$, где c_i – код i -го класса, $i = 1, 2, \dots, n$, n – количество классов в классификационной схеме, $\text{Frequency}(c_i)$ – частота i -го класса. В зависимости от задачи исследования частота может быть абсолютной или относительной.

Таким образом, классификационный профиль тематической коллекции позволяет увидеть распределение абсолютных и относительных частот классов (в нашем случае рубрик ГРНТИ), присвоенных статьям.

Предметный профиль можно представить в виде конечного множества пар значений $(w_j, \text{Weight}(w_j))$, где w_j – j -е ключевое слово, $j = 1, 2, \dots, m$, m – количество ключевых слов в словаре.

В качестве веса ключевого слова в настоящей работе используется $TF-IDF$ – широко распространённая статистическая мера оценки важности лексических единиц в тексте, которая может иметь различные варианты названия в публикациях: $TF-IDF$, $tf-idf$, $TF*IDF$, $TFIDF$, $Tf-idf$ [10].

Перечислим основные концептуальные принципы применяемой методики исследования тематической и терминологической трансформации тематических коллекций.

1. Количественное сравнение тематических профилей за разные хронологические периоды позволяет исследовать тематическую трансформацию предметной области, которая соответствует тематике коллекции. Тематическая коллекция при этом является статистической выборкой, а тематические профили (классификационные и предметные) – инструментами её исследования. Количественное сравнение предметных профилей позволяет также исследовать терминологическую трансформацию предметной области.

2. Поскольку оба вида тематических профилей (классификационный и предметный) можно представить в виде дескриптивных множеств, то к ним применимы математические методы (операции над множествами, меры сходства множеств).

3. Мера $TF-IDF$ вычисляется не в контексте отдельного документа, а в контексте всей тематической коллекции, которая рассматривается как один большой документ, а другие тематические коллекции БД ВИНТИ – как другие большие документы. Поскольку база данных ВИНТИ РАН является политематической, её тематические коллекции относятся к различным областям науки и техники. Поэтому статистический показатель IDF (*inverse document frequency*) позволяет достаточно хорошо отделять специальную научную лексику от общепотребительных слов и общенаучной лексики.

4. Скалярная величина $TF-IDF = TF \times IDF$ полезна для отбора релевантных слов, ранжирования ключевых слов и сравнения предметных профилей, но не позволяет объяснить причину существенного измене-

ния $TF-IDF$ во времени. Иным словами, $TF-IDF$ позволяет выявить изменение, но не позволяет ответить на вопрос: «Почему изменился статистический вес ключевого слова?» Для ответа на этот вопрос необходимо анализировать динамику изменения сомножителей, т.е. значений TF и IDF .

СТРАТЕГИЯ ПОИСКА

Поиск статей для отбора в тематические коллекции выполнялся в следующих метаданных:

- название статьи на языке оригинала;
- перевод названия статьи на русский язык (для статей на иностранном языке);
- ключевые и неключевые (пояснительные) слова, присвоенные статье в тематических выпусках Реферативного журнала ВИНТИ РАН.

Перед составлением текста запроса все поисковые термины были разделены на три множества:

А (безусловно релевантные): если термин присутствует в метаданных статьи, то с высокой вероятностью статья соответствует тематической коллекции;

В (условно релевантные): если термин присутствует в метаданных статьи, то со средней вероятностью статья соответствует тематической коллекции;

С (нерелевантные): если термин присутствует в метаданных статьи, то с высокой вероятностью статья не соответствует тематической коллекции.

Нерелевантные термины из множества С используются в запросах для отсеивания нерелевантных значений терминов из множества В, поскольку термины из множества В, как правило, обладают лексической неоднозначностью.

Далее приведём фрагмент запроса по тематике ИИ в упрощённом виде на примере нескольких словосочетаний со словами «интеллектуальный» и «распознавание», которые обладают лексической неоднозначностью. В упрощённом примере запроса термины даны не в нормальной (словарной), а в наиболее часто встречающейся форме и указаны логические связи между ними. В приведённом примере справа от термина в фигурных скобках указано множество, которому оно принадлежит:

'интеллектуальная система'{A} OR 'интеллектуальный анализ'{A} OR 'интеллектуальное управление'{A} OR 'интеллектуальный помощник'{A} OR ('интеллектуальный'{B} AND NOT 'интеллектуальная собственность'{C} AND NOT 'интеллектуальный капитал'{C}) AND NOT 'интеллектуальная деятельность'{C} AND NOT 'интеллектуальные права'{C}) OR 'распознавание изображений'{A} OR 'распознавание образов'{A} OR 'распознавание лиц'{A} OR 'распознавание объектов'{A} OR ('распознавание'{B} AND NOT 'молекулярное распознавание'{C} AND NOT 'кинетическое распознавание'{C} AND NOT 'распознавание клеток'{C} AND NOT 'распознавание энантиомеров'{C})

Этот пример демонстрирует, как условно релевантные слова «интеллектуальный» и «распознавание», выделенные подчёркиванием, уточняются в запросе благодаря использованию релевантных и нерелевантных устойчивых словосочетаний, включающих в себя эти слова.

Отметим, что в русском языке используется единственное прилагательное «интеллектуальный» в обо-

их контекстах: (система, анализ, ...) и (собственность, капитал, ...). В английском языке такая неоднозначность отсутствует, поскольку в разных контекстах употребляются разные прилагательные: «*intelligent*» (*system, analysis, ...*) и «*intellectual*» (*property, capital, ...*). Это упрощает запросы на английском языке. Однако существительное «*recognition*» в английском языке обладает такой же неоднозначностью, как существительное «распознавание» в русском языке.

Разделить термины на 3 множества и уточнить первоначальные поисковые запросы позволил визуальный просмотр частотных словарей, созданных по тематической коллекции документов. В частотные словари были включены слова и биграммы, извлечённые из названий и переводов названий, а также ключевые и неключевые слова и словосочетания, которые были присвоены статьям в результате координатного индексирования в ВИНТИ. Благодаря просмотру частотных словарей и уточнению запросов в 2-3 итерации удалось добиться высокой точности поиска статей для отбора в тематические коллекции.

Текст запроса для отбора в коллекцию ИСИМО состоит более чем из 300 поисковых терминов, а текст запроса для отбора в коллекцию РРС содержит около 40 поисковых терминов на русском и английском языках, поэтому мы можем их привести:

'робот' OR 'роботизированный' OR 'беспилотный' OR 'безэкипажный' OR 'необитаемый аппарат' OR 'БПЛА' OR 'АНПА' OR 'автономный аппарат' AND NOT 'автономный дыхательный аппарат' OR 'дрон' OR 'квадрокоптер' OR 'мультикоптер' OR 'роевые системы' OR 'управление роем' OR 'автономный движитель' OR 'автономная навигация' OR 'автономная навигационная' OR 'автономный транспорт' OR 'манипулятор' OR 'robot' OR 'robotic' OR 'unmanned' OR 'UAV' OR 'AUV' OR 'UGV' OR 'drone' OR 'metedrone' OR 'quadcopter' OR 'quadrotor' OR 'multirotor' OR 'multicopter' OR 'swarm system' OR 'swarm operation' OR 'autonomous vehicle' OR 'autonomous navigation' OR 'autonomous vehicle' OR 'manipulator'

При поиске необходимо учитывать не только возможные словоизменения и инверсию слов в словосочетаниях, но и присутствие других слов внутри словосочетания. Поэтому мы использовали в запросах результаты стемминга слов и шаблоны слов, заданные посредством регулярных выражений. Приведём примеры регулярных выражений со словом «распознавание» для поиска нерелевантных словосочетаний с учётом вклинивания других слов. Символ % означает усечение, в квадратных скобках перечислены символы, допустимые в данной позиции строки.

Примеры:

- шаблон '%распознавани[ея]% молекул%' найдёт подстроки 'распознавание молекул' и 'распознавание органических молекул';
- шаблон '%распознавани[ея]% клеток%' найдёт подстроки 'алгоритм распознавания клеток' и 'распознавание нормальных и раковых клеток'.

В результате уточнения поисковых запросов и проведения поиска в тематические коллекции включены релевантные статьи из сериальных изданий (СИ), отражённые в реферативной базе данных ВИНТИ РАН

в течение 2017-2023 гг. Объёмы тематических коллекций следующие:

- 24275 статей в коллекции РРС (из них 8503 статей в российских СИ);
- 53821 статей в коллекции ИСИМО (из них 14176 статей в российских СИ).

Тематические коллекции пересекаются: 3648 статей были включены в обе коллекции.

ИССЛЕДОВАНИЕ КЛАССИФИКАЦИОННЫХ ПРОФИЛЕЙ ТЕМАТИЧЕСКИХ КОЛЛЕКЦИЙ

В настоящей работе мы ограничили область исследования статьями из российских сериальных изданий. К российским статьям были отнесены все статьи из сериальных изданий, публикуемых в России, т.е. страна проживания или работы автора статьи не учитывалась. Хронологический период 2017-2023 гг. был разделён на два периода: 2017-2019 (старый) и 2021-2023 (новый), которые сравнивались между собой.

Робототехника и роботизированные системы (РРС)

Для создания классификационных профилей тематических коллекций использовались частотные распределения статей по рубрикам ГРНТИ 2-го уровня иерархии, которые далее кратко именуется как ГРНТИ-2.

Тематическая коллекция российских статей по РРС в БД ВИНТИ РАН составляла в 2017-2019 гг. 3005 статей, в 2021-2023 гг. – 4375 статей. Массив статей из российских СИ по РРС 2017-2019 гг. распределён по 212 рубрикам ГРНТИ-2, 2021-2023 гг. – по 228 рубрикам, при общем увеличении на 45,5% количества статей в более поздний период.

В табл. 1 приведены относительные частоты первых 13 рубрик из классификационного профиля тематической коллекции по РРС в 2021-2023 гг. Для сравнения также указаны доли статей, относящихся к конкретным рубрикам.

Сумма долей статей, относящихся к различным рубрикам, составляет >100% и является неаддитивной величиной, так как одна статья может относиться к неограниченному количеству рубрик ГРНТИ.

Приведём наиболее интересные результаты сравнения классификационных профилей новой (2021-2023 гг.) и старой (2017-2019 гг.) коллекций.

В тематической коллекции РРС наблюдается небольшое увеличение (на 13,5%) доли статей в рубрике 28.23 Искусственный интеллект и уменьшение на 5,5% доли статей в рубрике 55.47 Авиастроение. Вместе с тем, следует отметить, что при увеличении общего объёма тематической коллекции РРС в 1,46 раза (с 3005 до 4375 статей) в 2021-2023 гг. – абсолютное значение количества статей в рубрике «Искусственный интеллект» выросло в 1,65 раз (с 490 до 810 статей), а в рубрике «Авиастроение» – в 1,37 раза (с 498 до 684 статей).

Рубриками ГРНТИ-2, в которых наблюдался значительный рост доли российских статей в 2021-2023 гг. по сравнению с 2017-2019 гг., являются:

- 89.57 Исследования Земли из космоса (рост доли статей в 6,7 раз)
- 73.37 Воздушный транспорт (рост – в 3,8 раз)
- 30.15 Общая механика (рост – в 5,2 раз)

73.31 Автомобильный транспорт (рост – в 2,7 раз)
47.49 Радиотехнические системы зондирования, локации и навигации (рост – в 2,6 раз).

Массив российских журналов в БД ВИНТИ РАН, отражавших тематику по РРС в 2017-2019 гг., включал 753 СИ, с числом отраженных релевантных публикаций ≥ 1 . Общее количество профильных документов составило 3005 статей. В 2021-2023 гг. профильный массив документов в БД ВИНТИ РАН составил 4375 публикаций из 871 российских сериальных изданий. Таким образом, количество наименований российских СИ в тематической коллекции увеличилось на 15,7% при увеличении количества статей на 45,6%.

Интеллектуальные системы, искусственный интеллект и машинное обучение (ИСИИМО)

Тематическая коллекция российских статей по ИСИИМО в БД ВИНТИ РАН составляла в 2017-2019 гг. – 4530, в 2021-2023 гг. – 7757 документов.

Массив публикаций 2017-2019 гг. распределился по 300 рубрикам ГРНТИ-2, в 2021-2023 гг. – по 330 рубрикам, при общем увеличении на 71,2% количества статей в более поздний период.

В табл. 2 приведены относительные частоты и доли статей первых 15 рубрик из классификационного профиля тематической коллекции по ИСИИМО.

Таблица 1

ТОП-13 рубрик классификационного профиля по коллекции «Робототехника и роботизированные системы» в 2021-2023 гг.

Рубрика		Относительная частота рубрики ГРНТИ-2, %	Доля статей рубрики ГРНТИ, %
код	название		
55.30	Робототехника	12,1	18,6
28.23	Искусственный интеллект	12,0	18,5
89.57	Исследования Земли из космоса	11,2	17,3
55.47	Авиастроение	10,1	15,6
73.37	Воздушный транспорт	5,4	8,3
30.15	Общая механика	3,0	4,7
47.49	Радиотехнические системы зондирования, локации и навигации	2,8	4,4
73.31	Автомобильный транспорт	2,8	4,3
50.49	Автоматизированные системы организационного управления	1,9	3,0
55.45	Судостроение	1,9	3,0
55.57	Тракторное и сельскохозяйственное машиностроение	1,6	2,5
73.34	Водный транспорт	1,3	2,0
73.41	Промышленный транспорт	1,2	1,9
Сумма		67,3%	

Таблица 2

ТОП-15 рубрик классификационного профиля коллекции статей по «Интеллектуальным системам, искусственному интеллекту и машинному обучению» в 2021-2023 гг.

Рубрика		Относительная частота рубрики ГРНТИ-2, %	Доля статей рубрики ГРНТИ, %
код	название		
28.23	Искусственный интеллект	9,4	12,1
50.41	Программное обеспечение вычислительных машин, комплексов и сетей	3,9	5,1
73.31	Автомобильный транспорт	3,8	4,9
44.29	Электроэнергетика	3,2	4,1
50.49	Автоматизированные системы организационного управления	3,0	3,9
82.15	Организация управления экономикой	2,6	3,3
89.57	Исследования Земли из космоса	2,4	3,1
34.39	Физиология человека и животных	2,3	3,0
20.23	Информационный поиск	2,1	2,7
50.09	Элементы, узлы и устройства автоматики и вычислительной техники	2,0	2,6
73.29	Железнодорожный транспорт	1,8	2,3
52.47	Разработка нефтяных и газовых месторождений	1,8	2,3
47.49	Радиотехнические системы зондирования, локации и навигации	1,3	1,7
20.01	Общие вопросы информатики	1,3	1,7
90.01	Общие вопросы метрологии	1,2	1,6
Сумма		42,1%	

Рубриками ГРНТИ-2, в которых наблюдается наиболее значительный рост доли российских статей в тематической коллекции ИСИИМО в 2021-2023 по сравнению с 2017-2019 гг. являются:

73.31 Автомобильный транспорт (рост доли статей в 1,9 раз)

50.49 Автоматизированные системы организационного управления (рост – в 1,6 раз)

89.57 Исследования Земли из космоса (рост – в 1,6 раз).

Массив российских журналов в БД ВИНТИ РАН, отражавших тематику по ИСИИМО в 2017-2019 гг., включал 1073 СИ, с числом релевантных публикаций ≥ 1 . Общее количество профильных документов составило 4530 статей. В 2021-2023 гг. профильный массив документов в БД ВИНТИ РАН составил 7757 публикаций из 1484 российских серийных изданий. Таким образом, количество наименований российских СИ в тематической коллекции увеличилось на 38,3% при увеличении количества статей на 71,2%.

ИССЛЕДОВАНИЕ ПРЕДМЕТНЫХ ПРОФИЛЕЙ ТЕМАТИЧЕСКИХ КОЛЛЕКЦИЙ

Современные наукометрические методы исследования, основанные на лексике, обычно используют векторные представления (эмбединги) и тематическое моделирование. Векторные модели представления лексических единиц и текстов выражают их дистрибутивную семантику и применяются для решения различных задач интеллектуального анализа текстов, например кластеризации, классификации, поиска по сходству, подбора синонимов. Вероятностные тематические модели позволяют выполнять мягкую кластеризацию документов и ключевых слов, при которой один документ или одно ключевое слово могут с разной вероятностью принадлежать разным кластерам – выявленным темам.

В отличие от классов классификационной схемы и ключевых слов, которые могут использоваться годами и даже десятилетиями (что позволяет исследовать динамику роста количества документов, индексированных классификационным кодом или ключевым словом), выявленные темы документов и ключевых слов значительно более изменчивы во времени. Это затрудняет исследование динамики соответствующих им кластеров в течение достаточно продолжительного периода, например 5–10 лет. Приведём примеры двух подходов к решению этой проблемы.

В работе [4] выполнен поиск скрытых связей между патентными документами и публикациями, посвящёнными технологиям искусственного интеллекта. Используется тематическое моделирование методом *LDA* (*Latent Dirichlet Allocation*). Поскольку в разные хронологические периоды (1990-1999, 2000-2009, 2010-2017) были получены разные кластеры, авторы вычисляют Евклидово расстояние между любыми двумя темами в разные периоды и строят матрицы расстояний между темами. Это позволяет найти похожие кластеры и исследовать их динамику во времени.

Авторы работы [11] предлагают новый метод, который объединяет *MPA* (*Main Path Analysis*), метод

классификации текстов и тематические модели (используется метод *LDA*). Решается похожая задача – разделить статьи и патентные документы на семантически-близкие кластеры и исследовать связи между ними. Тематика статей и патентных документов – интеллектуальное производство, индустрия 4.0, применение *RFID*-технологии и цепочки поставок продовольствия. Метод *MPA* используется в сетях цитирования для выявления основных путей научных исследований. Полученные методом *MPA* кластеры статей, привязанные к основным путям, используются для обучения модели автоматической классификации (применяется эмбединг *BERT*) и последующей автоматической классификации патентных документов. Это позволяет привязать патентные документы к основным путям научных исследований, выявленным в сети цитирования. На заключительной стадии выполняется тематическое моделирование коллекций статей и патентных документов, соответствующих основным путям научных исследований, и получение списков наиболее частотных ключевых слов, которые характеризуют выявленные скрытые темы.

В настоящей работе мы исследуем динамику частоты встречаемости терминов напрямую: минуя скрытые темы и кластеры, к которым могут принадлежать документы и ключевые слова в разные годы. Для сравнения выбраны два хронологических периода: 2017-2019 и 2021-2023 гг. Вместо кластеризации мы используем готовые результаты в БД ВИНТИ: коды рубрик тематической классификации статей и разделение политематической базы данных на тематические фрагменты.

В качестве терминов для создания предметных профилей использовались ключевые слова, которыми документ был индексирован в ВИНТИ РАН, а также слова и биграмы, извлечённые из названий статей. Согласно действующему стандарту, при индексировании свободными ключевыми словами, взятыми из текста, длину словосочетаний рекомендуется ограничивать двумя-тремя словоформами [12]. Статистика по БД ВИНТИ показывает, что массив ключевых слов в целом соответствует этой рекомендации. Ключевые слова при индексировании в БД ВИНТИ обозначаются меткой *K* и, в отличие от пояснительных слов (обозначенных меткой *H*), используются для заголовков предметного указателя. В 2021-2023 гг. из общего количества ключевых слов 49,5% являлись словосочетаниями различной длины (от 2 до 13 слов). Из них 88,9% имели длину 2 или 3 слова. В этой статистике предлоги и союзы учитывались как слова.

Для взвешивания терминов в предметных профилях мы использовали *TF-IDF*. Если для вычисления *TF-IDF* слов и словосочетаний в контексте документа необходимо иметь массив других документов, различающихся между собой по тематике, то для вычисления *TF-IDF* в контексте тематической коллекции необходимо иметь другие тематические коллекции, причём в достаточном количестве и слабо пересекающиеся по тематике. Соблюдение последнего условия особенно важно для вычисления показателя *IDF*, который позволяет отделять общенаучную лексику,

распределённую по большинству тематических коллекций, от специальных научных терминов, используемых в меньшинстве тематических коллекций. В данной работе в качестве «других тематических коллекций» использовались тематические фрагменты БД ВИНТИ в количестве более 190 (соответствуют выпускам Реферативного журнала ВИНТИ), которые обеспечивали широкий охват естественных, точных и технических наук.

Поскольку мы вычисляли *TF-IDF* в контексте тематической коллекции, а не документа, каждая тематическая коллекция документов рассматривалась как один большой мультидокумент. В результате получилось от 191 до 198 мультидокументов для каждого года с 2017 по 2023 г. Количество описаний статей (название статьи + ключевые слова статьи) в этих мультидокументах варьировалось от 668 до 16030, в среднем 3142 статьи в одном годовом мультидокументе.

Годовое значение *TF-IDF* вычислялось для каждого термина в пределах годового мультидокумента, затем для периода 2017-2019 и 2021-2023 были вычислены средние арифметические *TF-IDF* за 3 года. Мы использовали традиционный метод вычисления *TF-IDF*, т.е. без каких-либо модификаций.

TF (*term frequency* — частота термина) — отношение числа вхождений некоторого термина в документ к общему числу терминов в документе. Таким образом, оценивается важность термина *t* в пределах документа *d*.

$$tf(t, d) = \frac{f(d, t)}{\sum_{k \in d} f(k, d)}, \quad (1)$$

где $f(t, d)$ есть число вхождений термина *t* в документ *d*, а в знаменателе — общее число терминов в данном документе.

IDF (*inverse document frequency* — обратная частота документа) — инверсия частоты, с которой некоторый термин встречается в документах политематического множества документов. Для каждого уникального термина в пределах конкретного политематического множества документов существует единственное значение *IDF*:

$$idf(t, D) = \log \frac{|D|}{|\{d : d \in D \& t \in d\}|}, \quad (2)$$

где: $|D|$ — число документов в политематическом множестве *D*;

$|\{d : d \in D \& t \in d\}|$ — число документов в множестве *D*, в которых встречается термин *t*, т.е. $tf(t, d) \neq 0$. В формуле (2) мы использовали натуральный логарифм.

Мера *TF-IDF* является произведением *TF* и *IDF*: $tfidf(t, d, D) = tf(t, d) \cdot idf(t, D)$.

Робототехника и роботизированные системы (PPC)

Тематическая коллекция статей по PPC была разделена на два хронологических массива по годам отражения документов в БД ВИНТИ: новый (2021-2023) и старый (2017-2019). Сравнение нового и старого массивов позволяет разделить термины на различные категории, из которых наиболее интересны следующие:

- 1) термин присутствует в новом массиве, но отсутствует в старом массиве;
- 2) термин присутствует в обоих массивах, и его *TF-IDF* существенно увеличился;
- 3) термин присутствует в обоих массивах, и его *TF-IDF* существенно уменьшился;
- 4) термин отсутствует в новом массиве, но присутствует в старом массиве.

В качестве примера в табл. 3 приведены *TOP-15* русскоязычных терминов в коллекции PPC, относящихся к категориям 1 и 2. Всего в списке было 3711 терминов (слов и словосочетаний), присутствовавших хотя бы в одном из хронологических массивов. Индексы *new* и *old* в табл. 3 обозначают новый и старый хронологический периоды. В табл. 3 и 4 применяется *Z*-нормализация для центрирования и масштабирования случайных величин по формуле:

$$Z(X) = \frac{X - M(X)}{\sigma(X)},$$

где $M(X)$ — среднее значение, $\sigma(X)$ — стандартное отклонение.

Неестественный порядок слов в некоторых словосочетаниях объясняется тем, что эти словосочетания вводятся в БД ВИНТИ для использования при создании предметного указателя к выпускам Реферативного журнала. Для сортировки и группировки записей указателя главное слово выносится вперёд, например: «системы управления позиционные», «роботы подъемно-транспортные», «роботы медицинские». При статистической обработке словосочетаний возможная инверсия слов учитывалась — в нормальной форме слова внутри словосочетания были упорядочены в алфавитном порядке.

Соотношения TF_{new} / TF_{old} и IDF_{new} / IDF_{old} в двух последних столбцах табл. 3 позволяют объяснить причину существенного изменения *TF-IDF*. Мы видим, что причиной существенного роста во всех примерах в табл. 3 была величина *TF*, т.е. частота термина. При этом *IDF* уменьшился почти во всех примерах, но не так существенно, как увеличился *TF*.

Интеллектуальные системы, искусственный интеллект и машинное обучение (ИСИИМО)

Аналогично предыдущей тематической коллекции, коллекция ИСИИМО была разделена на два хронологических массива. В качестве примера в табл. 4 приведены *TOP-15* русскоязычных терминов в коллекции ИСИИМО, относящихся к категориям 1 и 2. Всего в списке было 5825 терминов (слов и словосочетаний), присутствовавших хотя бы одним из хронологических массивов.

**Робототехника и роботизированные системы:
TOP-15 ранее отсутствовавших или увеличивших вес терминов**

1. Термины присутствуют в 2021-2023 и отсутствуют в 2017-2019			
Слово или словосочетание в наиболее часто употребляемой форме	$Z(TF-IDF_{new})$	$Z(TF_{new})$	$Z(IDF_{new})$
автономные транспортные средства	3,78	1,03	0,87
функциональность	2,87	2,66	-0,55
система распознавания образов	2,08	0,23	1,87
человека робота	1,76	0,35	0,99
системы управления позиционные	1,70	0,15	1,83
предотвращение столкновений	1,34	0,34	0,50
автономное вождение	1,16	0,21	0,77
радиосвязь с самолетами	1,06	0,01	1,96
роботы подъемно-транспортные	0,96	-0,02	1,83
вождение	0,93	0,30	0,14
траектории перемещения	0,93	0,09	0,89
беспилотный транспорт	0,88	0,10	0,74
подкреплением	0,86	0,47	-0,22
тросового робота	0,83	-0,03	1,58
взаимодействия человека	0,80	0,14	0,47
2. Термины присутствуют в 2017-2019 и 2021-2023; $TF-IDF$ существенно увеличился			
Слово или словосочетания в наиболее часто употребляемой форме	$Z(TF-IDF_{new}) - Z(TF-IDF_{old})$	TF_{new} / TF_{old}	IDF_{new} / IDF_{old}
роботы манипуляционные	5,41	1,67	0,94
системы управления адаптивные	4,48	6,17	0,86
автомобили беспилотные	4,46	3,82	0,82
обучение роботов	4,26	1,98	0,97
робототехнические системы	4,17	1,20	0,96
БПЛА	4,04	1,15	0,73
роботы медицинские	2,80	1,38	0,94
роботы коллаборативные	2,75	1,58	0,92
робототехнические устройства	2,68	4,94	0,80
экзоскелеты	2,63	3,82	0,92
автономных транспортных	2,62	3,93	0,81
схваты	2,16	1,88	1,04
мобильные роботы	1,90	0,75	0,96
роботизированной	1,90	0,98	0,94
транспортных средств	1,80	1,80	0,88

Таблица 4

**Интеллектуальные системы, искусственный интеллект и машинное обучение:
TOP-15 ранее отсутствовавших или увеличивших вес терминов**

1. Термины присутствуют в 2021-2023 и отсутствуют в 2017-2019			
Слово или словосочетание в наиболее часто употребляемой форме	$Z(TF-IDF_{new})$	$Z(TF_{new})$	$Z(IDF_{new})$
изображения гиперспектральные	5,02	0,56	0,28
автомобили интеллектуальные	3,89	0,17	1,29
распознавание действий	3,75	0,14	1,42
системы распознавания образов	3,52	0,06	2,05
глубокие нейронные сети	3,07	0,30	0,21
беспроводные технологии	2,89	0,22	0,41
обработка оптической информации	2,71	0,00	1,94
помощью машинного	2,57	0,43	-0,32
электромобили	2,43	0,47	-0,44
граничные вычисления	1,99	0,04	0,85
вождения	1,97	0,16	0,12
метод искусственных нейронных сетей	1,93	-0,05	1,82
биомаркеры	1,87	0,39	-0,55
семантическая сегментация	1,74	0,07	0,41
обучения прогнозирования	1,73	0,28	-0,37

2. Термины присутствуют в 2017-2019 и 2021-2023; <i>TF-IDF</i> существенно увеличился			
Слово или словосочетания в наиболее часто употребляемой форме	$Z(TF-IDF_{new}) - Z(TF-IDF_{old})$	TF_{new} / TF_{old}	IDF_{new} / IDF_{old}
нейронных сетей метод	11,32	2,34	0,87
глубокого обучения	8,06	3,50	0,28
сверточные нейронные сети	7,80	3,43	0,49
обучения с подкреплением	6,94	3,57	0,52
подкреплением	6,75	5,55	0,58
ИСЗ	6,51	2,08	1,05
классификация с обучением	6,46	4,21	0,71
интеллектуальные транспортные системы	6,27	2,32	0,88
наблюдений обработка	5,68	1,97	1,02
географические объекты	5,03	2,20	0,97
умный город	4,79	1,53	0,81
дистанционное зондирование	4,77	2,00	0,93
спутниковые наблюдения	3,88	2,19	0,92
автомобильный транспорт	3,69	2,59	0,99
обучение роботов	3,51	2,54	0,95

Как и в первой коллекции, в этих примерах мы видим рост *TF-IDF* исключительно за счёт роста частоты термина (*TF*), при этом обратная частота документа (*IDF*) либо изменилась несущественно, либо значительно уменьшилась для некоторых терминов: глубокое обучение, сверточные нейронный сети, обучение с подкреплением. Это говорит о распространении этих терминов по таким областям науки и техники, в которых они раньше не употреблялись.

СХОДСТВО ТЕМАТИЧЕСКИХ ПРОФИЛЕЙ

Коэффициент сходства двух тематических профилей (классификационных или предметных) позволяет дать количественную оценку их сходства. Коэффициент сходства принимает значения от 0 до 1. Если коэффициент сходства двух тематических профилей журнала (за новый и старый периоды) стремится к нулю, то тематика очень сильно изменилась. Если коэффициент сходства стремится к единице, то тематика практически не изменилась.

В качестве коэффициента сходства двух тематических профилей мы использовали меру сходства Сьеренсена¹, которая для дескриптивных множеств *A* и *B* рассчитывается как [13]:

$$K_{0;-1} = \frac{2 \sum_1^r \min(a_i, b_i)}{\sum_1^r (a_i) + \sum_1^r (b_i)},$$

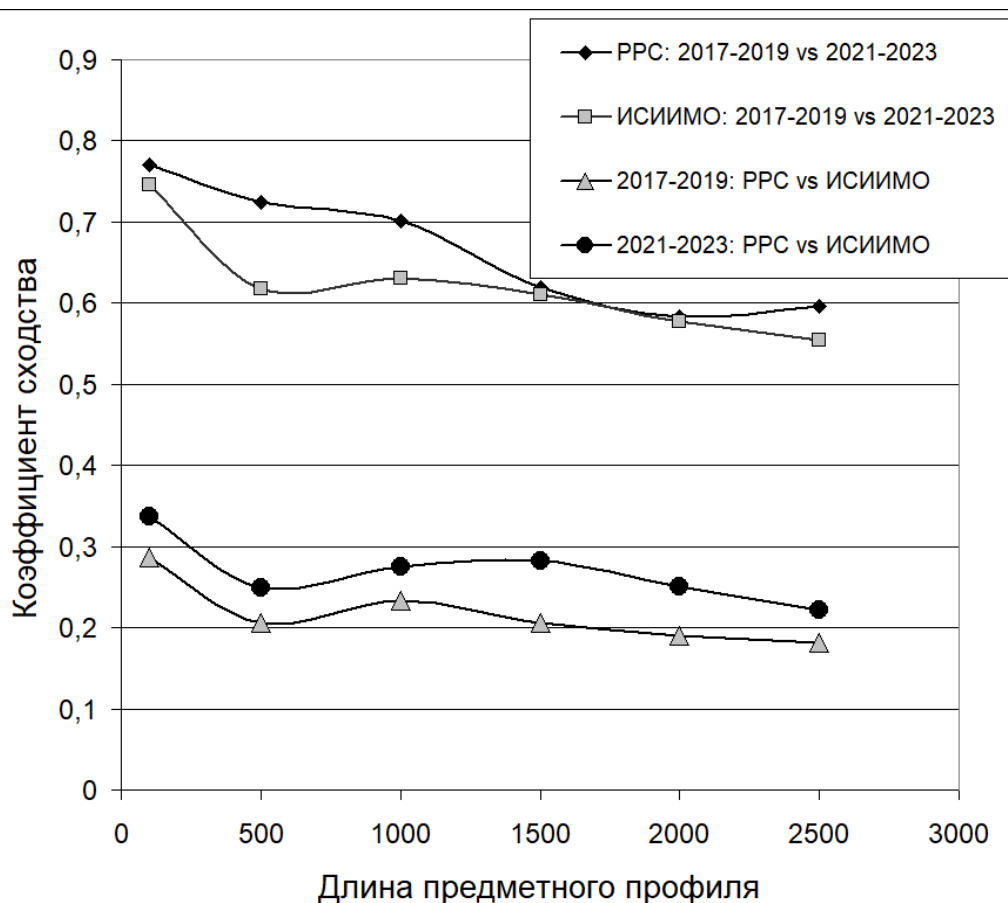
где a_i , b_i – соответственно веса i -го элемента в множествах *A* и *B*. Величина r зависит от вида профиля. В случае сравнения классификационных профилей количество используемых рубрик ГРНТИ-2 не превышало 330, поэтому $r = 330$. Сравнение предметных профилей выполнялась по первым 2000 ключевым

словам, выбранным из списков, ранжированных по убыванию *TFIDF*. Поэтому размер словаря используемых при сравнении ключевых слов не мог превысить $r = 4000$. Если в тематическом профиле менее r элементов, то веса отсутствующих элементов считаются равными нулю.

Вычисление коэффициентов сходства классификационных профилей (период 2017-2019 гг. сравнивался с периодом 2021-2023 гг.) показало, что коллекция РРС претерпела за 4 года более значительную тематическую трансформацию, чем коллекция ИСИМО. В случае РРС имеем $K_{0;-1} = 0,69$, в случае ИСИМО $K_{0;-1} = 0,77$.

Коэффициенты сходства, вычисленные по предметным профилям, показали более существенную терминологическую трансформацию коллекции ИСИМО во времени (в течение 4 лет) по сравнению с коллекцией РРС в тот же период. График на рисунке демонстрирует зависимость коэффициента сходства от длины предметного профиля, который использовался при вычислении этого показателя. Длина предметного профиля x означает, что при вычислении использовались первые x (*TOP-x*) ключевых слов из списка, упорядоченного по убыванию *TF-IDF*. Переменная x принимала следующие значения: 100, 500, 1000, 1500, 2000, 2500. На рисунке кривая линия, обозначенная как ИСИМО: 2017 vs 2021, расположена ниже кривой, обозначенной как РРС: 2017 vs 2021, т.е. коэффициент сходства для коллекции ИСИМО оказался меньше при всех значениях x . Наибольшая разница коэффициентов сходства (0,617 и 0,724) наблюдается при $x = 500$, наименьшая разница (0,577 и 0,584) при $x = 2000$. При таком большом количестве ключевых слов в сравнение предметных профилей вовлекается общенаучная, общеупотребительная и случайная лексика, что делает результаты сравнения менее достоверными. Если необходимо исследовать трансформацию специальной научной терминологии, которая при ранжировании по *TF-IDF* как правило оказывается вверху списка, то лучше сравнивать верхние фрагменты ранжированных списков. Поэтому при сравнении мы ограничились первыми 500 – 1000 ключевыми словами предметных профилей.

¹Мера сходства названа по фамилии Sørensen датского биолога, в русскоязычных публикациях она имеет различные варианты транслитерации – Соренсен, Серенсен, Сьренсен, Сьерсенсен.



Сравнение предметных профилей тематических коллекций

ЗАКЛЮЧЕНИЕ

Разработанная методика исследования тематической и терминологической трансформации коллекций научных статей состоит из нескольких этапов:

- составление и уточнение поисковых запросов на основе получения и просмотра частотных словарей ключевых слов;
- поиск и отбор документов в тематические коллекции;
- формирование классификационных профилей тематических коллекций;
- формирование предметных профилей тематических коллекций;
- сравнительный анализ классификационных профилей двух хронологических периодов для исследования тематической трансформации;
- сравнительный анализ предметных профилей двух хронологических периодов для исследования тематической трансформации;
- сравнительный анализ предметных профилей двух хронологических периодов для исследования терминологической трансформации.

В процессе работы выявлены некоторые проблемы, которые нуждаются в дальнейшем исследовании и решении.

1. Мера $TF-IDF$, изобретённая для статистического взвешивания слов и словосочетаний в документах, нуждается в модификации для использования её в мультидокументах – массивах документов, связан-

ных общей тематической принадлежностью. Особенно это касается величины обратной частоты документа (IDF), которая учитывает единичное присутствие термина в документе. Для мультидокументов единичного присутствия термина недостаточно, так как в большом тематическом массиве документов термин мог оказаться один раз совершенно случайно. Для вычисления IDF термина на множестве мультидокументов необходимо использовать пороговое значение минимального количества употребления термина в мультидокументе.

2. Эффективность любого статистического метода взвешивания терминов может оцениваться по его способности достаточно точно разделять термины на два множества: общенаучные и специальные термины. К сожалению, сравнить результаты разделения терминов не с чем – эталонные словари отсутствуют. В терминологических словарях по различным областям науки и техники специальные и общенаучные термины собраны вместе, и граница между ними не проведена. В качестве наглядного примера приведём «Толковый словарь по искусственному интеллекту» [14], который содержит около 550 профильных терминов. В этом словаре можно увидеть следующие термины: *алгоритм, база данных, высказывание, гипотеза, граф, действие, доказательство, теоремы* и многие другие, широко распространённые в других областях науки. Вполне понятно желание авторов собрать в одном словаре большинство терминов, которые могут понадобиться при изуче-

нии ИИ. Однако очевидно, что подобные словари не могут служить эталонными выборками для оценки и сравнения различных статистических мер важности терминов.

3. В настоящей работе использовался ограниченный набор метаданных: названия статей и ключевые слова, присвоенные статьям специалистами ВИНТИ РАН. Для повышения качества статистических исследований необходимо расширять этот набор за счёт авторских ключевых слов и аннотаций статей. В целях повышения достоверности статистического исследования необходимо также расцеплять ключевые слова, присвоенные человеком, на униграммы (слова) и биграммы, так как понятие «устойчивого словосочетания» субъективно. Одни авторы и редакторы индексируют статьи короткими ключевыми словами, другие предпочитают длинные лексические конструкции, иногда в качестве ключевых слов используются целые фразы.

СПИСОК ЛИТЕРАТУРЫ

1. Rotolo D., Hicks D., Martin B.R. What is an emerging technology? // Research Policy. – 2015. – Vol. 44, Issue 10. – P. 1827–1843. – URL: <https://doi.org/10.1016/j.respol.2015.06.006> (дата обращения 31.10.2024).
2. Huang Y., Zhu D., Qian Y., Porter A. L., Liu Y., Guo Y. A hybrid method to trace technology evolution pathways: a case study of 3D printing // Scientometrics. – 2017. – Vol. 111. – P. 185–204. – URL: <https://doi.org/10.1007/s11192-017-2271-8> (дата обращения 31.10.2024).
3. Михеев А.В. Возможности анализа текстовой информации научно-технологической направленности для обоснования инновационного развития энергетики // Информационные и математические технологии в науке и управлении. – 2017. – № 4(8). – С. 166–176.
4. Li M., Wang W., Zhou K. Exploring the technology emergence related to artificial intelligence: A perspective of coupling analyses // Technological Forecasting & Social Change. – 2021. – Vol. 172. – P. 1–13. – URL: <https://doi.org/10.1016/j.techfore.2021.121064> (дата обращения 31.10.2024).
5. Xu H., Winnink J., Yue Z., Zhang H., Pang H. Multidimensional Scientometric indicators for the detection of emerging research topics // Technological Forecasting & Social Change. – 2021. – Vol. 163. – P. 1–24. – URL: <https://doi.org/10.1016/j.techfore.2020.120490> (дата обращения 31.10.2024).
6. Солошенко Н.С., Федорец О.В., Домнина Т.Н. Исследование трансформации тематических профилей сериальных изданий во входном потоке документов информационного центра (на примере БД ВИНТИ РАН) // Научно-техническая информация. Сер. 1. – 2022. – № 10. – С. 15–24. – URL: <https://doi.org/10.36535/0548-0019-2022-10-3> (дата обращения 31.10.2024).
7. Солошенко Н.С., Федорец О.В., Домнина Т.Н. Количественная и качественная трансформация потока российских научно-технических и производственных сериальных изданий в БД ВИНТИ РАН (2010–2021) // Научно-техническая информация. Сер. 1. – 2022. – № 12. – С. 2–7. – URL: <https://doi.org/10.36535/0548-0019-2022-12-1> (дата обращения 31.10.2024).
8. Государственный рубрикатор научно-технической информации (ГРНТИ). – 2023. – URL: <https://www.gpntb.ru/grnti.html> (дата обращения 31.10.2024).
9. Сёмкин Б.И. Deskriptivnye mnozhestva i ikh prilozheniya // Issledovanie sistem. Tom. 1. Analiz slozhnykh sistem. – Vladivostok: DVNTS AN SSSR, 1973. – С. 83–94.
10. Wikipedia contributors. (2024, July 26). Tf-idf. In Wikipedia, The Free Encyclopedia. – 2024. – URL: <https://en.wikipedia.org/w/index.php?title=Tf%E2%80%93idf&oldid=1236851603> (дата обращения 31.10.2024).
11. Yu D., Yan Z. Combining machine learning and main path analysis to identify research front: from the perspective of science-technology linkage // Scientometrics. – 2022. – Vol. 127. – P. 4251–4274. DOI: <https://doi.org/10.1007/s11192-022-04443-1> (дата обращения 31.10.2024).
12. ГОСТ Р 7.0.66-2010. Индексирование документов. Общие требования к координатному индексированию / Федеральное агентство по техническому регулированию и метрологии. – Москва, 2011. – 16 с. (Система стандартов по информации, библиотечному и издательскому делу).
13. Юрцев Б.А., Семкин Б.И. Изучение конкретных и парциальных флор с помощью математических методов // Ботанический журнал. – 1980. – Т. 65, № 12. – С. 1706–1718.
14. Толковый словарь по искусственному интеллекту / авторы-сост.: А.Н. Аверкин, М.Г. Гаазе-Рапопорт, Д.А. Поспелов. – 2021. – URL: <https://www.raai.org/pages/UGFnZVR5cGU6MTAwMw==> (дата обращения 31.10.2024).

Материал поступил в редакцию 11.11.24.

Сведения об авторах

ФЕДОРЕЦ Олег Владимирович – кандидат технических наук, начальник Отдела программных систем Всероссийского института научной и технической информации Российской академии наук (ВИНТИ РАН), Москва
e-mail: ovf@viniti.ru

СОЛОШЕНКО Наталия Сергеевна – кандидат педагогических наук, заведующий отделом комплектования ВИНТИ РАН
e-mail: solns@viniti.ru

Использование цифровых нитей и систем цифровых двойников в социо-кибер-физических системах*

Рассматриваются вопросы построения социо-кибер-физических систем и использования в них цифровых двойников. Предлагается механизм непрерывного динамического перестроения цифровых двойников для различных этапов жизненного цикла систем. Механизм предусматривает трансформацию моделей, лежащих в основе цифровых двойников, которая может выполняться не только при переходе между различными этапами жизненного цикла системы, но и в рамках одного этапа. В качестве примера показана трансформация компетентностной модели человека.

Ключевые слова: социо-кибер-физические системы, цифровые двойники, жизненный цикл системы, цифровой двойник человека

DOI: 10.36535/0548-0027-2024-12-4

ВВЕДЕНИЕ

Впечатляющие успехи в области микроэлектроники, технологии разработки программного обеспечения и механизмов работы со знаниями открывают перед разработчиками прикладных информационно-ориентированных систем, в частности, кибер-физических и социо-кибер-физических, самые широкие перспективы. Появляется возможность создавать социо-кибер-физические системы принципиально нового уровня сложности, состоящие из большого числа разнородных элементов, включая природные объекты, живые существа, в частности, люди и группы людей.

Такие системы, в большинстве своем, могут быть отнесены к классу систем (*System of Systems, SoS*) [1], они характеризуются наличием большого числа заинтересованных сторон, которые имеют разные интересы и играют разные роли относительно рассматриваемой социо-кибер-физической системы.

Применительно к кибер-физическим системам цифровой двойник обычно определяют как цифровую модель или множество моделей кибер-физической (социо-кибер-физической) системы или ее отдельных подсистем (элементов), которые используются при проектировании социо-кибер-физических систем на одном или нескольких этапах их жизненного цикла. Цифровой двойник – это достаточно сложное и многоаспектное понятие.

Применительно к распределенным социо-кибер-физическим системам с точки зрения их архитектуры можно выделить следующие основные описания цифрового двойника [2–11] – это:

- программный компонент, инкапсулирующий модель или комплекс моделей некоторой существу-

ющей, существовавшей или еще, или уже несуществующей социо-кибер-физической системы;

- программный компонент, инкапсулирующий частную модель некоторой наблюдаемой системы;
- контроллер (прокси, заглушка) какой-либо сущности;
- программный компонент, инкапсулирующий модель или комплекс моделей некоторого класса сущностей;
- программный компонент, обеспечивающий формирование некоторого представления (видения) наблюдаемой и управляемой социо-кибер-физической системы, которое формируется в интересах определенной заинтересованной стороны;
- цифровой двойник – это архитектурный стиль.

Таким образом цифровой двойник можно рассматривать как одну или несколько архитектурных точек зрения.

На пути дальнейшего развития социо-кибер-физических систем можно выделить две ключевые проблемы, с которыми сталкиваются разработчики:

- 1) высокий уровень гетерогенности структуры современных социо-кибер-физических систем;
- 2) построение существенной части современных социо-кибер-физических систем по принципу системы систем, что приводит к тому, что информация об их структуре и поведении может быть ограничена или просто недоступна.

В настоящей статье рассматривается один из аспектов построения социо-кибер-физических систем, построение и перестроение цифровых двойников и входящих в их состав моделей на различных этапах жизненного цикла больших и сложных социо-кибер-физических систем [1].

* Работа выполнена при финансовой поддержке гранта, бюджетная тема №FFZF-2022-0006.

Предмет нашего исследования – попытка ответить на следующие вопросы.

1. Что такое цифровой двойник, зачем он нужен и в чем состоит отличие цифрового двойника от модели?

2. Какие функции выполняет цифровой двойник на разных этапах жизненного цикла социо-кибер-физической системы?

3. Каким образом можно использовать цифровой двойник при построении социо-кибер-физической системы на разных этапах их жизненного цикла?

4. Каким образом можно перестраивать модели, входящие в состав цифрового двойника при переходе между различными этапами жизненного цикла?

5. Какими должны быть модели, используемые при построении цифрового двойника, в частности, какими должны быть модели человека как элемента социо-кибер-физической системы?

Мы представляем один из возможных подходов построению и перестройке цифровых двойников, используемых в качестве элемента социо-кибер-физической системы.

Основные понятия и определения

Кибер-физическая система (cyber-physical system) – информационно-технологическая концепция, подразумевающая интеграцию вычислительных ресурсов в физические сущности любого вида, включая биологические и рукотворные объекты.

Социо-кибер-физическая система – это понятие можно определить, как расширение понятия кибер-физическая система за счет включения в состав человека и коллектива людей, а также их моделей в качестве элемента элементов системы.

Цифровой двойник – обычно определяют как цифровую модель некоторой сущности произвольной природы, которая связана некоторым отношением. При этом типы отношений могут быть самыми разными [12]. Возникает закономерный вопрос: Чем цифровой двойник отличается от цифровой модели и как соотносятся эти понятия?

СОВРЕМЕННОЕ СОСТОЯНИЕ ТЕХНИКИ ПОСТРОЕНИЯ СОЦИО-КИБЕР-ФИЗИЧЕСКИХ СИСТЕМ И ЦИФРОВЫХ ДВОЙНИКОВ

Можно выделить следующие типовые черты современных социо-кибер-физических систем:

1) высокая сложность и гетерогенность;
2) высокий уровень динамики структуры и поведения этих систем и, как следствие, проблемы отслеживания текущего состояния и предсказания их будущих состояний;

3) отдельные элементы систем могут существенно эволюционировать как во времени, так и под влиянием внешних факторов;

4) по разным причинам полная информация о структуре и поведении отдельных элементов таких систем недоступна;

5) отдельные элементы систем могут скрывать, препятствовать или искажать информацию о себе или о других элементах.

Анализ доступных публикаций показывает, что вопросам построения собственно социо-кибер-физических систем уделено не так много внимания, хотя имеется большое число работ, посвященных теории си-

стем [13–15]. Это можно объяснить тем, что отдельные элементы и модели социо-кибер-физических систем существенно различаются в зависимости от их назначения, свойств и входящих в их состав моделируемых сущностей.

В подавляющем большинстве публикаций предметом рассмотрения являются относительно узкие классы систем такие, как, например, социо-кибернетические системы [8], когда в качестве элементов выступают только люди и коллективы людей. При этом многие реальные системы включают в себя именно разнородные элементы.

Имеется очень большое число публикаций, посвященных производственным системам [16–18], в частности индустрии 4.0 [19], которые безусловно представляют для нас определенный интерес. С точки зрения своего жизненного цикла современные социо-кибер-физические системы приближаются к корпоративным системам, и для управления их жизненным циклом могут применяться механизмы похожие, но не аналогичные используемым в фреймворке TOGAF [20].

Основное внимание в имеющихся публикациях уделяется применению цифровых двойников на этапах проектирования и производства и значительно меньше – использованию их на этапах эксплуатации и модернизации. Сегодня вопросы построения цифровых двойников времени эксплуатации рассматриваются как независимые задачи, что приводит к дополнительным издержкам. Практически не рассматривается использование цифровых двойников в распределенных федеративных системах [21], построенных, в частности, по принципу системы систем [1]. Следует заметить, что, с одной стороны, в настоящее время большое число реальных социо-кибер-физических систем строятся как распределенные системы систем, а с другой стороны, чаще всего они рассматриваются как некая монолитная сущность.

Таким образом, анализ доступных публикаций показывает, что большая их часть посвящена использованию цифровых двойников в процессе производства. Цифровые двойники, относящиеся к разным этапам жизненного цикла системы, образуют цифровые нити [22], а множество их архитектур на разных этапах и фазах жизненного цикла – архитектурный континуум.

ПОДХОД К ПОСТРОЕНИЮ СОЦИО-КИБЕР-ФИЗИЧЕСКИХ СИСТЕМ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ АРХИТЕКТУРНОГО КОНТИНИУМА И ЦИФРОВЫХ ДВОЙНИКОВ

Основная идея предлагаемого подхода состоит в использовании цифровых нитей и заключается в следующем:

1) использование систем цифровых двойников на всех этапах жизненного цикла социо-кибер-физических систем;

2) системы цифровых двойников, построенные на ранних этапах жизненного цикла, образуют цифровую нить;

3) цифровые двойники, построенные на ранних этапах жизненного цикла, могут быть полезны при

построении цифровых двойников, используемых на более поздних этапах жизненного цикла.

Предмет рассмотрения – сложные социо-кибер-физические системы, построенные по принципу системы систем, которые в основе своей являются антропогенными системами.

Еще один предмет рассмотрения в настоящей статье – это цифровые двойники [21, 23], которые понимаются как один из возможных подходов к реализации парадигмы социо-кибер-физических систем. В рамках развиваемого подхода обсуждаются организация и проектирование сложных интеллектуальных (когнитивных [24]) распределенных социо-кибер-физических систем с высоким уровнем структурной динамики. Такие системы, способные реализовывать адаптивное поведение и построенные на основе цифровых двойников, могут быть отнесены к классу больших и сложных систем.

Предлагаемый подход обеспечивает сокращение совокупной стоимости владения социо-кибер-физической системой за счет уменьшения стоимости создания и поддержания в актуальном состоянии системы моделей.

В рамках развиваемого подхода цифровой двойник рассматривается как элемент социо-кибер-физических систем, включающий в себя 3 части:

$$\text{SYS} = \langle \text{OBS}, \text{DT}, \text{STHC} \rangle,$$

где: **SYS** – некоторая социо-кибер-физическая система;

OBS – наблюдаемая система произвольной природы;

DT – множество ее цифровых двойников;

STHC (*Stakeholders Concerns*) – интересы заинтересованных сторон, в качестве которых могут выступать люди (операторы, владельцы, продавцы, покупатели и др.), а также другие подсистемы любой природы [25].

Одной и той же наблюдаемой системе может соответствовать произвольное число цифровых двойников, которые строятся исходя из интересов заинтересованных сторон.

Социо-кибер-физическую систему можно представить себе состоящей из двух связанных между собой миров: мира взаимодействующих между собой разнотипных сущностей (физических элементов, ФЭ), природных сущностей (ПЭ), виртуальных сущностей (ВЭ), живых существ, в частности, людей и мира виртуальных сущностей (цифровых двойников (ЦД), агентов (А)) и множества заинтересованных сторон (рис. 1). При этом каждая заинтересованная сторона имеет разные интересы относительно системы.

Цифровой двойник сложной социо-кибер-физической системы, как правило, представляет собой распределенную систему, структура которой необязательно совпадает с физической структурой данной системы.

В общем случае, система цифровых двойников может быть построена как система систем. В зависимости от способов взаимодействия отдельных систем построенная система может быть управляемая, федеративная или коалиционная [1].

Социо-кибер-физические системы представляют собой очень широкий класс систем, разного назначения и разного масштаба, к которым предъявляются самые разнообразные требования.

Являясь гетерогенными, социо-кибер-физические системы включают в свой состав как антропогенные подсистемы, так и людей, коллективы (группы) людей, природные объекты и т.п. Все эти элементы могут иметь разные жизненные циклы, которые могут очень существенно различаться. Другими словами, жизненные циклы отдельных элементов и жизненный цикл всей социо-кибер-физической системы или ее экземпляров часто не совпадают. Например, система создаётся в нескольких экземплярах по заказу или на продажу. После того, как система создана в некотором количестве экземпляров, она передается владельцам (покупателям, заказчикам) (рис. 2).

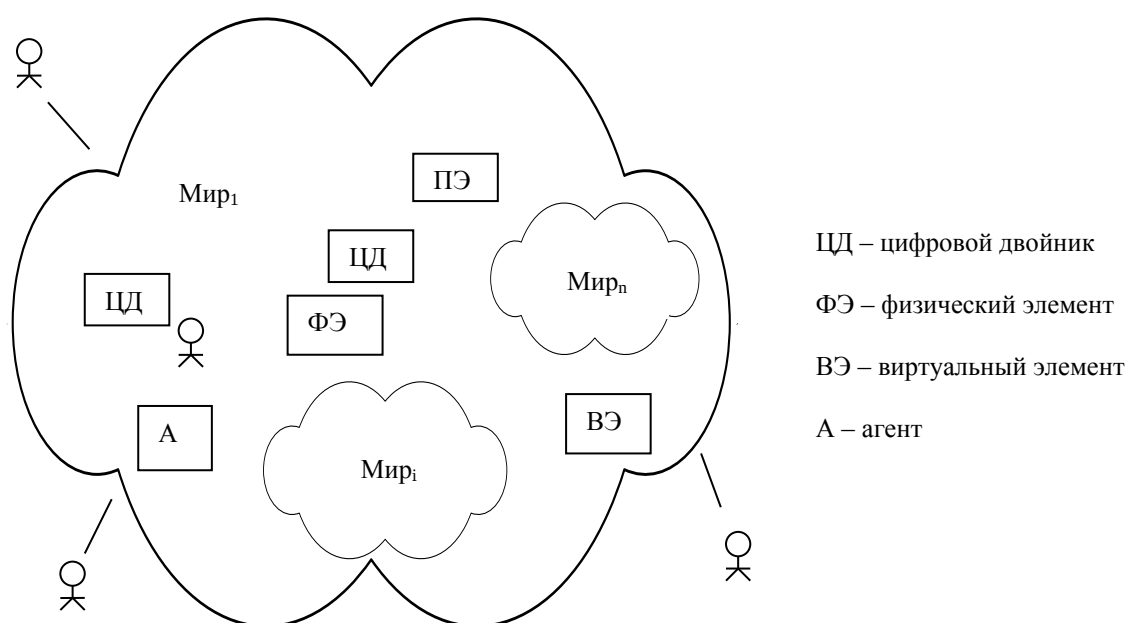


Рис. 1. Цифровой двойник как элемент социо-кибер-физической системы (идеализированная модель)

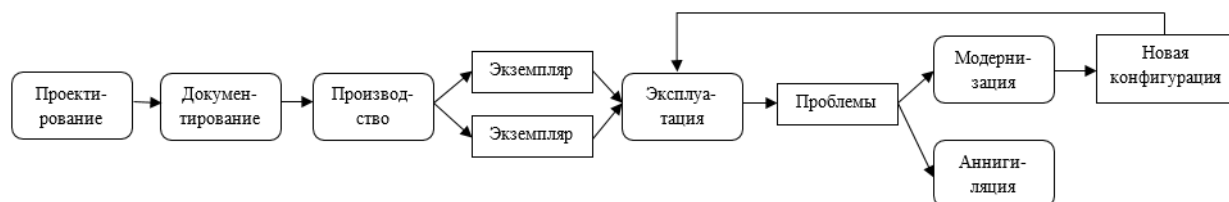


Рис. 2. Простейший жизненный цикл (Система на продажу)

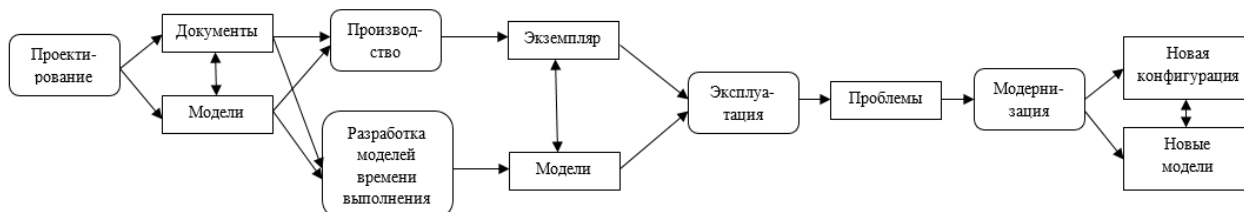


Рис. 3. Непрерывная архитектура на базе цифровых двойников

В предлагаемом подходе модели строятся на базе моделей, применяемых на более ранних этапах жизненного цикла системы (рис. 3). При этом на всех или на отдельных этапах жизненного цикла можно использовать цифровые двойники, т. е. модели. Бывают также некоторые промежуточные варианты.

Чтобы объединить в единую систему разнородные элементы, связанные между собой с помощью разнотипных механизмов, требуется выбрать некоторый вариант описания сущностей разной природы. Цифровая модель выступает в качестве унифицированного представления сущности произвольной природы. При этом цифровой двойник рассматривается как одна из разновидностей модели. Отличие цифрового двойника от модели достаточно тонкое. Цифровой двойник безусловно является моделью, а точнее цифровой моделью некоторой сущности произвольной природы. Попытки дать четкое определение понятия цифрового двойника нельзя признать успешными, хотя известно много его определений [2–7]. Возникают сомнения в принципиальной возможности определить это понятие. Наиболее перспективным представляется подход, в соответствии с которым цифровой двойник определяется как некоторое отношение между сущностями, когда одна из сущностей является цифровой моделью другой сущности. При этом как модель, так и моделируемая сущность могут представлять собой систему, состоящую из множества элементов. В общем случае в качестве моделируемой сущности выступает гетерогенная структура произвольной сложности, а в качестве модели – гетерогенная иерархическая цифровая модель. Структура моделируемой социо-киберфизической системы и структура модели чаще всего не совпадают.

Отношения между наблюдаемой сущностью и цифровым двойником могут быть разными:

1) на этапе проектирования – цифровой двойник описывает еще не существующую сущность;

2) на этапе создания – цифровой двойник управляет процессом создания сущности;

3) на этапе функционирования – цифровой двойник с одной стороны, обеспечивает возможности интеграции в единую систему элементов разной физической природы; с другой стороны, может использоваться для поддержания механизмов архитектурной гибкости [26];

4) на этапе модернизации целевой системы – наличие цифрового двойника может облегчить процесс миграции на другие платформы.

Цифровой двойник практически любой достаточно сложной социо-кибер-физической системы, как правило, описывается с помощью многих моделей. Множество этих моделей можно определить термином полимодель. Полимодель – это многоуровневая модель, элементами которой являются другие модели [27].

В комплекс моделей включаются модели, относящиеся к разным этапам жизненного цикла. Таким образом, применительно к сложным социо-киберфизическим системам полимодели могут быть представлены как множество моделей:

- относящихся к разным фазам жизненного цикла;
- относящихся к разным подсистемам для каждого этапа жизненного цикла;
- отражающих интересы разных заинтересованных сторон;
- описывающих разные состояния социо-киберфизической системы.

Очевидно, что число подобных моделей оказывается очень велико.

Одной из ключевых проблем построения моделей социо-кибер-физической системы является её гетерогенность. В качестве эталонной модели (*reference model*) такой системы мы предлагаем использовать систему взаимосвязанных моделей, которые могут иметь динамическую структуру и реализовывать разное поведение [28, 29].

Применительно к антропогенным подсистемам социо-кибер-физической системы требуется иметь модели, относящиеся к разным этапам жизненного цикла этой системы, при этом она описывается как множество множеств моделей, относящихся к разным этапам её жизненного цикла. Такую систему моделей можно представить следующим образом:

$$M = \{M_i, Tr_{ij}\},$$

где: M_i – множество множеств моделей;
 Tr_{ij} – множество правил трансформации моделей.

Как указывалось ранее, множество моделей, используемых в составе цифрового двойника на разных этапах жизненного цикла, образуют некоторую связанную последовательность. Эти модели описывают одну и ту же наблюдаемую сущность на разных этапах жизненного цикла. Для полного описания наблюдаемой системы на всех этапах жизненного цикла множество моделей, относящихся к разным этапам жизненного цикла, можно представить следующим образом:

$$M = \{M_{ij}, STH, TR\},$$

где: M_{ij} – множество моделей, относящихся к разным этапам жизненного цикла;

i – этап жизненного цикла (проектирование, производство, эксплуатация, модернизация, аннигиляция);

j – модель, соответствующая определенной точке зрения;

STH – заинтересованные стороны, представленные своими интересами;

TR – правила перехода между моделями, относящимися к разным этапам.

Следует заметить, что для разных этапов жизненного цикла могут применяться не только разные модели, но и разные заинтересованные стороны.

Использование модельного континуума не является обязательным, однако создание с нуля моделей, относящихся к разным этапам жизненного цикла наблюдаемой системы, может быть источником ошибок, а также быть дороже по сравнению с подходом, основанным на трансформации моделей.

Для модельного континуума ключевая проблема – это трансформация моделей.

Трансформация моделей – одна из центральных задач проектирования цифровых нитей. Трансформацию моделей TR можно определить следующим образом:

$$M_i = TR(M_{i-1}, CK, R),$$

где: M_i – модель i -го уровня;
 TR – функция трансформации;
 M_{i-1} – известные модели;
 CK – контекстное знание;
 R – требования модели.

Применительно к задачам трансформации моделей, используемых при построении цифровых нитей,

можно различать по пяти основным параметрам – фазам жизненного цикла социо-кибер-физической системы: проектирование, производство, эксплуатация, модернизация и аннигиляция.

Можно выделить два основных типа моделей: отдельно стоящая модель (*stand alone*) и модель, как часть цифрового двойника. В первом случае модель предназначена для проведения экспериментов с целью оценки некоторых параметров системы. Во втором случае модель существует во взаимосвязи с наблюдаемой системой.

Модель может строиться на терминах сырых данных, информации или знаний.

Модель может быть моделью класса систем, моделью конкретной системы и моделью конкретной системы в конкретный момент времени.

Можно выделить два основных типа трансформации моделей: горизонтальную и вертикальную. Вертикальную трансформацию можно определить как трансформацию моделей, относящихся к одному этапу жизненного цикла социо-кибер-физической системы, горизонтальную – как трансформацию моделей, когда исходная и конечная модели относятся к разным этапам жизненного цикла.

На рис. 4 показан фрагмент укрупненной классификации моделей, представленной в виде диаграммы классов, используемых в цифровых двойниках.

Модели, входящие в состав цифровых двойников, могут иметь следующие виды: модели антропогенных физических сущностей, модели абстрактных сущностей, модели природных сущностей, модели живых сущностей, модели человека, модели коллективов людей и т.д. Модели могут быть: структурными моделями или моделями поведения; полными моделями наблюдаемых систем либо частными моделями; статическими или динамическими; монолитными или представлять собой систему моделей, в частности, быть набором разнотипных моделей. Следует отметить, что это далеко не полная классификация.

Число типов моделей, используемых при построении социо-кибер-физических систем, и моделей, ориентированных на использование в процессе функционирования этих систем, на практике может быть достаточно большим.

Общее число моделей, используемых в социо-кибер-физической системе можно оценить следующим образом:

$$N = (\text{число подсистем}) \times ((\text{число точек зрения}) \times (\text{число моделей, описывающих точку зрения})).$$

Типы моделей определяются как интересами пользователей, так и спецификой социо-кибер-физической системы и ее элементов. Например, в качестве моделей аналоговых систем обычно используются дифференциальные уравнения; в качестве моделей дискретных систем – автоматные [30, 31] и графовые модели (графы свойств, графы знаний [32]). В настоящей статье детально рассматриваются только дискретные модели дискретных систем с учетом особенностей их применения в качестве моделей времени выполнения.

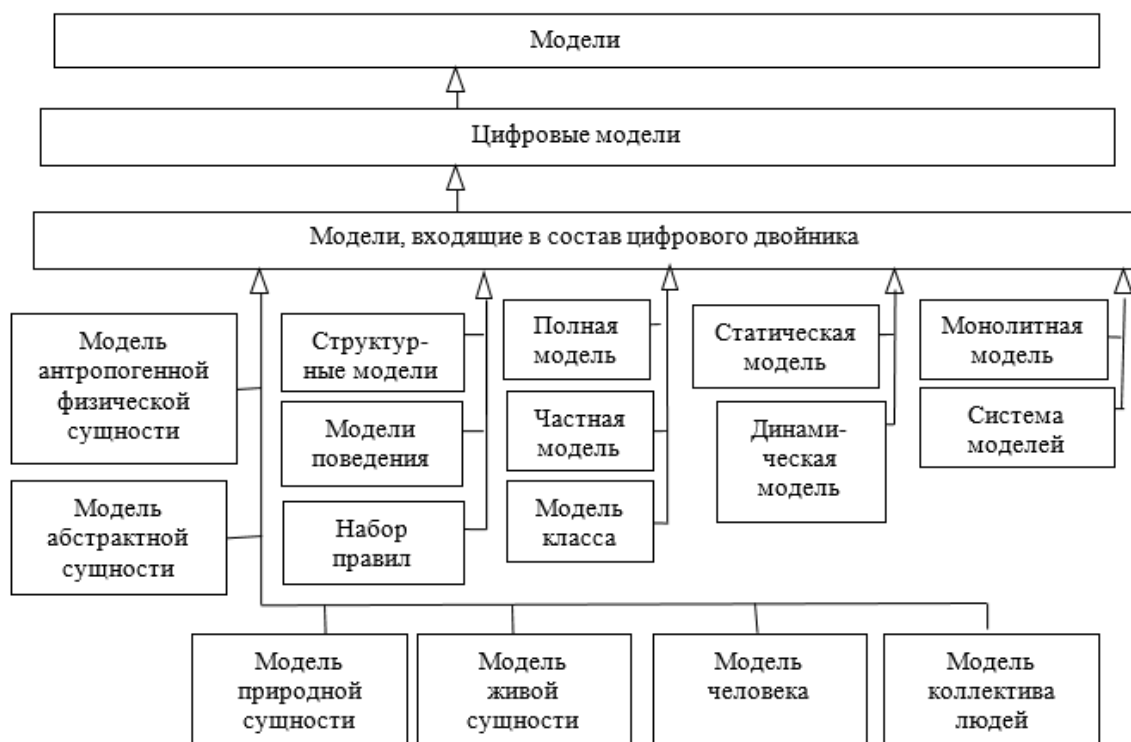


Рис. 4. Типы моделей, используемых в цифровых двойниках

Подробнее рассмотрим структурно-функциональные (СФ) модели, как наиболее широко используемые.

Структурно-функциональные модели – это модели, описывающие системы в терминах их структур и поведения. В качестве таких моделей предлагаем рассматривать дискретные модели дискретных систем, но можно использовать и другие модели.

Для описания структуры основной интерес представляют графовые модели, для описания поведения – разные модели. В настоящей статье в качестве функциональных рассматриваются такие модели, как сети Петри, BPMN, YAWL (*Yet Another Workflow Language* [33]), позволяющие описывать параллельные асинхронные процессы.

С одной стороны, крайне привлекательно выглядит идея использования готового языка для построения подобной модели, такой как UML или SysML или их профиля. Но здесь появляются две проблемы:

- XML нотация, используемая в UML (SysML), часто оказывается слишком сложной для обработки применительно к цифровому двойнику времени выполнения;
- структурные и поведенческие модели в UML (SysML) синхронизируются в ручном режиме.

С другой стороны, крайне желательно, чтобы структурно-функциональные модели были достаточно близки к UML (SysML) моделям, используемым на этапе проектирования. Это должно упростить процедуры генерации структурно-функциональных моделей времени выполнения.

Перечисленные модели представляют собой модели, ориентированные на этап проектирования, но

они не ориентированы на использование в качестве моделей времени выполнения. Проблемы вызывают два момента: 1) в UML (SysML) отсутствуют механизмы автоматической синхронизации моделей; 2) с помощью UML (SysML) достаточно сложно описать системы с изменяющейся структурой и поведением. Однако представляется целесообразным сохранить возможно более тесную связь с традиционными UML (SysML) моделями.

В качестве обобщенной структурно-функциональной модели (метамодели) мы предлагаем использовать модель акторов, представленную в [30], где акторы рассматриваются как агенты, которые сопоставляют каждое входящее сообщение с кортежами, включающими конечный набор сообщений, отправляемых другим акторам. Новое поведение формируется в терминах нового набора акторов и зависит от предыстории.

Формально структурно-функциональную модель можно определить, как:

MSFM = $\langle \text{Inmsg}, \text{Outmsg}, \text{Act}, \text{BM}, \text{BMT}, \text{newA} \rangle$,

где: **Inmsg** – множество допустимых входных сообщений;

Outmsg – множество выходных сообщений;

BM – множество моделей поведения;

BMT – таблица, определяющая правила перехода между моделями поведения;

Act – множество акторов;

newA – множество правил создания новых акторов.

Акторная модель – это одноуровневая модель, однако ее можно достаточно просто обобщить для описания многоуровневой системы, если определить актора, как:

$Act ::= \langle Eact \rangle | \langle MMM \rangle$,

где **Eact** – элементарный актор.

Предлагаемая структурно-функциональная модель представляет собой одну из возможных реализаций и адаптацию акторной модели применительно к цифровым двойникам. Эту модель можно определить, как:

$MM = \langle Act, Afr, Lnk, Tk, Res, BP \rangle$,

где: **Act** – множество акторов;

Afr – множество правил запуска акторов;

Lnk – множество типов связей между акторами;

Tk – множество типов сообщений (токены, фишки);

Res – множество типов ресурсов;

BP – множество бизнес-процессов, описывающих функциональность наблюдаемой социо-кибер-физической системы.

Эти компоненты, в свою очередь, можно определить следующим образом:

$\langle Act \rangle ::= \langle \text{Преобразование} \rangle | \langle \text{Управление} \rangle | \langle \text{Порт} \rangle$,

$\langle Afr \rangle ::= \langle \text{Правила запуска} \rangle | \langle \text{Правила формирования результата} \rangle$

$\langle Lnk \rangle ::= \langle \text{Связи по ресурсам} \rangle | \langle \text{Связи по данным} \rangle | \langle \text{Связи по управлению} \rangle$

$\langle Tk \rangle ::= \langle \text{Токен данных} \rangle | \langle \text{Токен управления} \rangle | \langle \text{Ресурсный токен} \rangle$,

$\langle Res \rangle ::= \langle \text{Физический ресурс} \rangle | \langle \text{Виртуальный ресурс} \rangle$,

$\langle BP \rangle ::= \langle \text{Статический} \rangle | \langle \text{Динамический} \rangle$.

В простейшем случае используются статические бизнес-процессы (БП), когда во время функционирования все операторы и связи между ними известны заранее. Для статических бизнес-процессов модель может быть представлена в виде размеченного (labeled) графа [32], содержащего вершины двух типов: акторы и ресурсы. Акторы соединены между собой дугами, определяющими информационные зависимости. Ресурсные вершины соединены дугами, которые соответствуют физическим или логическим связям между отдельными ресурсами. Акторы и ресурсные вершины соединены дугами, имеющими смысл "выполняется на". Эти два графа можно рассматривать как отдельные подграфы:

$MM = \{Gd, Gr, Ldr\}$,

где: **Gd** – размеченный граф данных;

Gr – ресурсный граф;

Ldr – связи между вершинами **Gd** и **Gr**.

Принципиально **Gr** можно определить через атрибуты вершин **Gd**, но граф **Gr** представляет интерес как предмет синтеза.

Каждый актор имеет **n** входящих и **m** исходящих дуг. Входящие дуги: произвольное число дуг данных, одна дуга **C** для директивного запуска актора, одна дуга **Q** для получения запросной фишки и одна дуга **R** для получения фишки с информацией о готовности ресурса. Актор в общем случае имеет следующие выходные дуги: произвольное количество дуг данных и одну **R**-дугу, по которой выдается фишка с информацией об освобождении ресурса.

Таким образом, в системе акторов циркулируют следующие типы токенов:

- данных (D-токены);
- принудительного запуска актора (C-токены);
- запросов (Q-токены);
- ресурсные (R-токены).

Каждый из токенов может принадлежать к классу либо стационарных, либо подвижных токенов. Стационарные токены не меняют своей позиции в процессе функционирования, а подвижные могут появляться и исчезать. По своей сути подвижные токены идентичны токенам в сетях Петри.

В предлагаемой нами модели реализуется следующая логика работы.

Если на всех входных дугах появились токены, то актор срабатывает. С каждого входа исчезает по одному подвижному токену. На выходах появляются новые токены в соответствии с логикой работы актора. Все стационарные токены сохраняют свои позиции.

Ресурсный граф **Gr** отражает текущую структуру, которая в основе своей является виртуальной, поскольку элементы различной природы могут быть представлены цифровыми двойниками. Ресурсный граф **Gr** можно определить, как:

$Gr = \langle R, Lr, Ldr \rangle$,

где: **R** – ресурсы;

Lr – связи между ресурсами;

Ldr – связи с **Ad**.

Сам ресурсный элемент может находиться в состояниях свободен-занят, исправен-неисправен. В основе статической модели лежит граф потока данных.

Рассмотренные структурно-функциональные модели времени выполнения могут быть расширены для случаев многоуровневых систем и систем с изменяемыми в процессе функционирования структурой и поведением.

ПРИМЕР ПОСТРОЕНИЯ ЦИФРОВОГО ДВОЙНИКА

В качестве примера рассмотрим цифровой двойник студента. Этот пример интересен тем, что в качестве одного из аспектов применения цифрового двойника человека выступает процесс обучения, который в значительной степени базируются на компетентностных моделях. Следует иметь в виду, что это одна из многих точек зрения, описывающих человека как элемент социо-кибер-физической системы.

В основу предлагаемого подхода к построению цифрового двойника человека могут быть положены следующие принципы:

- подсистема обучения – одна из подсистем системы более высокого уровня, например, университета;
- образовательный процесс – один из бизнес-процессов, протекающих в образовательной организации;
- образовательный процесс – как динамический бизнес-процесс.

Модель образовательного процесса может быть построена по типу модели производственной системы [34], при этом выделяются потоки: компетенций (преобразование исходных компетенций обучаемых в требуемые конечные компетенции), информационные (планирование и управление образовательным процессом) и стоимости (в частности, временной стоимости).

В качестве основных частных можно рассматривать следующие модели: обучаемого, обучающего (преподавателя), предметного домена, потребителей образовательных ресурсов, подразделения, отвечающего за подготовку (переподготовку), организации, в интересах которой выполняется подготовка.

Можно выделить следующие основные группы заинтересованных сторон: обучаемый, например, студент; потребители – специалисты (промышленность), преподаватели (университет), общество.

Каждая из перечисленных групп заинтересованных сторон имеет собственные интересы, которые могут определяться в разных терминах и различаться.

Компетентностная модель описывает процесс формирования и накопления компетенций, который можно описать в терминах компетентностных состояний, и строится в терминах компетенций, компетентностных состояний и переходов между компетентностными состояниями.

Для работы с машинно-ориентированными компетентностными моделями могут быть использованы идеи, заложенные в таксономии Блума [35], где выделяется несколько уровней учебных целей: знание, понимание, применение, анализ, синтез, оценка.

Таксономия Блума определяет следующие уровни компетенции, что позволяет описывать итерационные модели обучения: применение (умение использовать изученный материал в конкретных условиях и новых ситуациях), анализ (умение разбить материал на составляющие так, чтобы ясно выступала структура), синтез (умение комбинировать элементы, чтобы получить целое, обладающее новизной), оценка (умение оценивать значение того или иного материала с учетом целей обучения). Следует заметить, что число уровней и их наполнение не представляется чем-то абсолютным, т. е. их число и содержание определяют спецификой конкретной предметной области.

Принципиально можно ограничиться использованием классической модели компетенций, которая базируется на триаде «знать-уметь-иметь представление».

Компетентностное состояние можно определить, как множество компетенций, которыми владеет обучаемый. Переход между компетентностными состояниями является результатом изучения некоторого

курса (модуля) или участия в выполнении некоторого проекта. Подобный процесс может быть описан в терминах автомата (машины состояний).

При изменении компетентностного состояния каждая компетенция может преобразоваться одним из четырех способов: не изменять свое состояние, тогда появляется новая компетенция; изменять свое состояние и переходить на более высокий уровень, когда на базе имеющихся и добавляемых состояний формируется компетенция более высокого уровня. Состояние компетенций инкапсулируется в атрибутах. Компетентностная модель по своей сути является автоматной моделью.

Термин "компетенция" определяется как "знания, умения, навыки, модели поведения и личностные характеристики, при помощи которых достигаются желаемые результаты"; понятие "компетентность" – как множество компетенций, которые требуются для выполнения профессиональных задач. Обычно принято различать поведенческие, технические и лидерские компетенции. Следует заметить, что компетентностный подход все более активно используется при построении интеллектуальных систем и, в первую очередь, систем гибридного интеллекта [24].

Применительно к системам образования компетенции используются для описания процесса получения знаний, навыков и умений обучаемыми. Можно считать, что профессиональное образование представляет собой процесс формирования у человека определённого набора компетенций, необходимых для решения практических задач.

Другими словами, речь идет о построении бизнес-процесса, который переводит обучаемого в требуемое компетентностное состояние.

На практике задача работы с компетентностными моделями может встречаться в различных постановках: создание (синтез) компетентностных моделей, выполнение запросов к моделям, в ряде случаев может также возникнуть необходимость в постановке и решении задачи актуализации компетентностных моделей.

В основе своей компетентностные модели — это модели, описывающие знания о знаниях и навыках, которыми располагают люди (студенты, разработчики, менеджеры). Поэтому при создании компетентностных моделей представляется целесообразным ориентироваться на механизмы работы со знанием, в частности, на онтологии, графы знаний и графы свойств [32, 33].

Компетентностную модель можно определить как множество компетентностных состояний и множество правил перехода между компетентностными состояниями. Каждое компетентностное состояние включает в себя множество компетенций более низкого уровня. Все компетенции должны быть определены в онтологии. Компетентностную модель можно определить, как:

$$CM = \langle CS, TR, O \rangle,$$

где: **CM** – компетентностная модель;
CS – компетентностное состояние;
TR – функция переходов между состояниями;
O – онтология предметной области.

CS можно считать множеством элементарных компетенций, определенных в онтологии **O**; **TR** – как множество правил перехода между **CS**. Чтобы перейти из некоторого исходного **CS_i** в требуемое состояние **CS_{i+1}**, можно использовать образовательный ресурс **R_j**. Динамика функционирования представленной компетентностной модели может быть описана в терминах многоуровневого, относительно конечно-го, автомата [30, 31].

Компетентностное состояние можно описать с помощью графа знаний, т.е. выделить следующие типичные задачи, решение которых основывается на использовании цифрового двойника человека: построение образовательных траекторий, проверка актуальности образовательной программы, построение компетентностной модели специалиста, подбор исполнителей для выполнения проекта.

Построение образовательной траектории – это задача организации оптимального в некотором смысле бизнес-процесса с учетом ограничений.

Задача построения компетентностной модели специалиста может быть сформулирована следующим образом.

Дано: 1) исходная компетентностная модель; 2) информация об освоённой образовательной программе (курсе); 3) информация об участии в практической деятельности.

Требуется: 1) построить результирующую компетентную модель; 2) определить степень схожести модели требований и полученной компетентностной модели.

Аналогично может быть определена задача подбора исполнителей для выполнения проекта.

Дано: 1) компетентностные модели специалистов (команды специалистов); 2) требуемая компетентностная модель.

Требуется определить: модель, которая соответствует требуемой компетентностной модели.

Если компетентностная модель представлена в виде графа знаний или графа свойств [32, 33], то решение этих задач не должно вызывать значительных трудностей. Основная проблема заключается в разработке онтологии, описывающей соответствующую предметную область.

Таким образом, компетентностные модели могут быть полезны при построении цифровых двойников студентов, преподавателей, специалистов.

Следует заметить, что построение компетентностных моделей в форме онтологий или графов знаний требует наличия эталонного словаря, содержащего перечень компетенций, который может быть представлен, в частности, в форме онтологии.

ЗАКЛЮЧЕНИЕ

В настоящей статье рассматриваются вопросы использования цифровых двойников при построении социо-кибер-физических систем на разных этапах жизненного цикла социо-кибер-физической системы. Современный этап развития информационных технологий характеризуется, в первую очередь, постоянным увеличением уровня сложности создаваемых информационно-ориентированных систем и попытками включить человека в состав кибер-физической

системы, т.е. перейти от кибер-физической системы к социо-кибер-физической системе. Проблема построения модели человека состоит в том, что, с одной стороны, отсутствует полная информация о человеке как объекте моделирования, а, с другой стороны, человек может обучаться и его поведение как элемента социо-кибер-физической системы может зависеть от очень многих факторов.

Использование цифровых нитей представляет собой действенный инструмент сокращения совокупной стоимости владения социо-кибер-физической системой. Цифровые двойники, построенные на ранних этапах жизненного цикла такой системы, могут быть использованы при построении цифровых двойников для более поздних этапов её жизненного цикла.

Предлагаемый подход был сформулирован в результате реализации ряда проектов в различных предметных областях, таких как системы мониторинга сетей кабельного телевидения, производственные системы, транспортные системы и системы образования [36, 37].

Дальнейшее развитие предлагаемого подхода планируется осуществлять по следующим четырем направлениям: 1) переход к практической реализации цифрового двойника человека (создание моделей, онтологий и фреймворков); 2) расширение области применения цифрового двойника человека, способного работать с ментальными категориями; 3) анализ возможных подходов к обучению моделей, в частности, распределенных; 4) исследование возможности использования предлагаемого подхода при построении гибридных интеллектуальных систем.

СПИСОК ЛИТЕРАТУРЫ

1. Jamshidi M. System of systems engineering: innovations for the 21st century. – New York: Deakin University, 2008. – 617 p.
2. Haße H. et al. A Taxonomy of Digital Twins. – Salt Lake City, USA: 26th Americas Conference on Information Systems (AMCIS). – 2020. – 10 p.
3. Ricci A. et al. Web of digital twins // ACM Transactions on Internet Technology. – 2022. – Vol. 22, № 4. – P. 1-30.
4. Digital Twins: Basics and Applications / eds. Z. Lv, E. Fersman. – Springer International Publishing, 2022. – 101 p.
5. Cyber-physical Systems and Digital Twins: Proceedings of the 16th International Conference on Remote Engineering and Virtual Instrumentation / eds. M.E. Auer et al. – Springer, 2019. – Vol. 80.
6. Digital twin technology: fundamentals and applications / ed. M. Vohra. – New York: John Wiley & Sons, 2023. – 271 p.
7. Thelen A. et al. A comprehensive review of digital twin—part 1: modeling and twinning enabling technologies // Structural and Multidisciplinary Optimization. – 2022. – Vol. 65, № 12. – 354 p.
8. Gräßler I. et al. The Digital Twin of Humans. – Cham: Springer International Publishing. – 2023. – 270 p.
9. Shah F. Modeling Human Group Behavior In Virtual Worlds. – 2011.

10. Tao F. et al. Digital twin driven smart design. – New York: Academic Press, 2020. – 300 p.
11. Crespi N., Drobot A.T., Minerva R. The Digital Twin: What and Why? – Cham: Springer International Publishing, 2023. – P. 3-20.
12. Охтилев М.Ю., Соколов Б.В., Юсупов Р.М. Интеллектуальные технологии мониторинга и управления структурной динамикой сложных технических объектов. – Наука, 2006. – 408 с.
13. Olcott S., Mullen C. Digital twin consortium defines digital twin // Digital Twin Consortium. – 2020. – Vol. 3.
14. Von Bertalanffy L. The history and status of general systems theory // Academy of management journal. – 1972. – Vol. 15, № 4. – P. 407-426.
15. Черняк Ю.И. Системный анализ в управлении экономикой. – Москва: Экономика, 1975. – 191 с.
16. Nath S.V., Schalkwyk P.V. Building Industrial Digital Twins: Design, Develop, and Deploy Digital Twin Solutions for Real-World Industries Using Azure Digital Twins. – Birmingham, UK: Packt Publishing, 2021. – 286 p.
17. Blasch E., Bosse E., Lambert D. High-Level Information Fusion Management and System Design. – Norwood, MA: Artech House Publishers, 2012. – 369 p.
18. Liggins M., Hall D., Llinas J. Handbook of multisensor data fusion: theory and practice. – Boca Raton, FL: Taylor & Francis Group, 2009. – 850 p.
19. Patnaik S. New Paradigm of Industry 4.0 Internet of Things // Big Data & Cyber Physical Systems. – Cham, Switzerland: Springer Nature, 2020 – 187 p.
20. The TOGAF Standard, 10th Edition. – URL: <https://www.opengroup.org/togaf> (дата обращения: 10.10.2024).
21. Cyber-physical systems of systems: foundations—a conceptual model and some derivations: the AMADEOS legacy / eds. A. Bondavalli, S. Bouchenak, H. Kopetz. – Springer, 2016. – Vol. 10099.
22. Taber M., Zhang J., Strom T., Immerman D., Duncan D. Digital Thread // Building Continuity Across Products, Processes, and People. – 2022. – 15 p.
23. Michael J. et al. Integration challenges for digital twin systems-of-systems. – Pittsburgh, PA, USA: Proceedings of the 10th IEEE/ACM International Workshop on Software Engineering for Systems-of-Systems and Software Ecosystems. – 2022. – P. 9-12.
24. Cyber-physical Systems and Digital Twins: Proceedings of the 16th International Conference on Remote Engineering and Virtual Instrumentation / eds. M.E. Auer et al. – Springer, 2019. – Vol. 80. – 862 p.
25. Водяхо А.И., Жукова Н.А., Ананьева В.Я. Использование систем цифровых двойников при построении социо-кибер-физических систем // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. – 2024. – Т. 20, Вып. 4. – С. 50–64.
26. Bloomberg J. The agile architecture revolution: how cloud computing, rest-based SOA, and mobile computing are changing enterprise IT. – New York: John Wiley & Sons, 2013. – 287 p.
27. Barricelli B.R., Casiraghi E., Fogli D. A survey on digital twin: Definitions, characteristics, applications, and design implications // IEEE Access. – 2019. – Vol. 7. – P. 167653-167671.
28. Vodyaho A., Zhukova N., Delhibabu R., Subbotin A. Continuous agile cyber-physical systems architectures based on digital twins // Future Generation Computer Systems. – 2023. – Vol. 153. – P. 350-359. DOI: 10.1016/j.future.2023.11.024.
29. Ананьева В.Я., Водяхо А.И., Жукова Н.А., Червонцев М.А. Использование динамических цифровых двойников при построении киберфизических систем // Известия СПбГЭТУ «ЛЭТИ». – 2024. – Т. 17, № 3. – С. 44-55. DOI: 10.32603/2071-8985-2024-17-3-44-55.
30. Осипов В.Ю. Автоматический синтез программ действий интеллектуальных роботов // Программирование. – 2016. – № 3. – С. 47-54.
31. Osipov V.Y. Automatic synthesis of action programs for intelligent robots // Programming and Computer Software. – 2016. – Vol. 42. – P. 155-160.
32. Blumauer A., Nagy H. The knowledge graph cookbook: Recipes that work. – edition mono/monochrome, 2020. – 255 p.
33. Aalst W. Process mining: data science in action. – Berlin Heidelberg: Springer-Verlag, 2016. – 477 p.
34. Patnaik S. New paradigm of Industry 4.0 // Internet of Things, Big Data & Cyber Physical Systems. – Cham: Springer, 2020. – 187 p.
35. Bloom B.S. et al. Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain. – New York: Longman, 1956. – P. 1103-1133.
36. Aung T.T. et al. Data processing model for mobile IoT systems // Procedia Computer Science. – 2021. – Vol. 186. – P. 235-241.
37. Subbotin A.N., Zhukova N.A., Man T. Video processing algorithm in foggy environment for intelligent video surveillance // Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Vol. 2. – Springer International Publishing, 2022. – P. 702-715.

Материал поступил в редакцию 24.09.24.

Сведения об авторах

ВОДЯХО Александр Иванович – доктор технических наук, профессор, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
e-mail: aivodyaho@mail.ru

ЖУКОВА Наталия Александровна – доктор технических наук, ведущий научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук
e-mail: nazhukova@mail.ru

АНАНЬЕВА Варвара Яновна – аспирант, ассистент, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
e-mail: varvara.spb99@mail.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81'322.4

А.Ю. Егорова, И.М. Зацман, В.О. Романенко

Метод оценки воспроизводимости результатов машинного перевода с помощью ChatGPT*

Рассматривается возможность воспроизводимости на интервалах времени результатов применения чат-бота ChatGPT для машинного перевода с русского языка на английский с помощью метода количественной оценки «веса» каждой из категорий, характеризующих траекторию изменений во времени переводов одних и тех же фраз. Описан эксперимент на корпусе из 50 фраз на русском языке, которые переводились на английский язык еженедельно в течение 12 недель. Таким образом был сформирован массив из 600 пар фраз, каждая из которых содержит фразу на русском языке и ее английский перевод. Для каждой из 50 оригинальных фраз и соответствующих ей 12 переводов была организована серия из 12 аннотаций, включающих рубрики классификации ошибок перевода или пометку об их отсутствии. Все серии переводов были разделены на шесть категорий в зависимости от вида траектории изменений в серии: снижение качества перевода на интервале, повышение качества, колебание качества, изменение набора ошибок без динамики качества перевода, изменение перевода без динамики его качества и серии переводов без изменения.

Ключевые слова: ChatGPT; мониторинг; машинный перевод; воспроизводимость результатов; интервальное оценивание

DOI: 10.36535/0548-0027-2024-12-5

ВВЕДЕНИЕ

Чат-бот с генеративным искусственным интеллектом ChatGPT был выпущен американской компанией OpenAI в ноябре 2022 г. Через три месяца число его пользователей превысило 120 млн. Он стал использоваться в научных экспериментах по оценке его результативности и степени воспроизводимости результатов его работы.

В эксперименте, описанию которого посвящена настоящая статья, степень воспроизводимости определялась с использованием на интервале времени метода повторяющейся экспертной оценки результатов машинного перевода (МП) с помощью ChatGPT. Оценка повторно применялась одними и теми же операторами и экспертами на одних и тех же текстах. Предмет оценивания, выполняемого при таких условиях, в англоязычной литературе обозначается термином *repeatability* [1].

Мониторинг функционирования этого чат-бота на интервалах времени позволил выявить следующую характерную черту: полученные с помощью ChatGPT результаты актуальны только для конкретного момента времени, но не для интервала мониторинга. Воспроизводимость результатов работы ChatGPT была исследована при решении и других задач, например, в эксперименте по категоризации с его помощью пациентов, поступающих в отделение неотложной медицинской помощи. Цель этого исследования заключалась в решении проблемы: «Могут ли врачи скорой помощи использовать ChatGPT для “сортировки” поступающих пациентов?» Проведенный эксперимент дал отрицательный эффект из-за низкого уровня воспроизводимости результатов работы ChatGPT [2].

В работе [3] была определена воспроизводимость результатов применения ChatGPT 3.5 и ChatGPT 4 для ответов на вопросы экзамена по медицине, которые были заданы каждой версии чат-бота трижды. По итогам эксперимента ChatGPT 4 продемонстрировал значительное повышение уровня стабильных

* Исследование выполнено с использованием ЦКП «Информатика» ФИЦ ИУ РАН.

правильных ответов по сравнению с ChatGPT 3.5: 77,8% и 44,9% соответственно.

Схожий эксперимент описан в [4] на материале ответов на вопросы по проверке слуха (*test of hearing*) на интервале времени: в этом случае ChatGPT 3.5 продемонстрировал 75-79% неизменно правильных ответов, а ChatGPT 4 – 85-88%.

Мониторинг функционирования чат-бота выполнялся и на примере тех задач, которые Дэвид Арел (Научный институт Вейцмана) называл «псевдоалгоритмическими» и относил к проблемам когнитивной сложности [5, с. 402]. Например, была проведена оценка результатов использования ChatGPT при обработке текстов на естественном языке, включая понимание прочитанного и машинный перевод [6, 7]. В работе [8] качество перевода, выполненного с английского языка на японский с помощью ChatGPT, было оценено по четырем параметрам по шкале от 1 до 10: коммуникативная точность (3,7), лингвистическая корректность (5,3), достоверность (5,4), культурная адаптивность (9,6). При этом не производилась фиксация допущенных ошибок и их категоризация.

Отметим, что в большинстве перечисленных работ при тестировании и оценке результатов работы ChatGPT при решении задач когнитивной сложности не выполнялся мониторинг воспроизводимости этих результатов *на интервалах времени*. Однако во время тестирования было замечено, что ответы на один и тот же запрос могут различаться в разные моменты времени. Цель настоящей статьи – на основе экспериментального массива фраз (текстовых фрагментов) оценить количественно воспроизводимость результатов работы ChatGPT как машинного переводчика с русского языка на английский.

Подобное исследование воспроизводимости результатов машинного перевода с русского языка на французский было опубликовано ранее в работе [7]. Описываемые эксперименты по оценке воспроизводимости результатов работы ChatGPT были выполнены с использованием надкорпусной базы данных (НБД), разработанной М.Г. Кружковым [9], и типологии ошибок, разработанной под руководством Н.В. Бунтман [10].

МЕТОД ОЦЕНКИ ВОСПРОИЗВОДИМОСТИ РЕЗУЛЬТАТОВ МАШИННОГО ПЕРЕВОДА

Для мониторинга воспроизводимости результатов машинного перевода с помощью ChatGPT были взяты 50 текстовых фрагментов из русско-английского подкорпуса Национального корпуса русского языка (НКРЯ) [11]. Каждый фрагмент содержит двухкомпонентный коннектор. Под коннектором понимается языковая единица, функция которой состоит в выражении логико-семантического отношения, существующего между соединенными с ее помощью частями текстового фрагмента [12, с. 17]. Например, в табл. 1 коннектор *так как (расстояние) то* устанавливает отношение причины между фрагментами «при этом он употреблял и души» и «результаты этого лечения подвергаются, конечно, сомнению».

Выбор текстовых фрагментов основывался на данных ранее выполненного эксперимента [13], в ходе которого был произведен мониторинг качества

французских переводов, полученных с помощью Google Translate для 250 русскоязычных фрагментов (см. два примера перевода в табл. 1). По итогам эксперимента [13] был разработан метод оценки воспроизводимости результатов машинного перевода (МП). В соответствии с этим методом выполнялось многократное повторение перевода через определенный промежуток времени и результаты фиксировались в надкорпусной БД с их последующим аннотированием.

Применение этого метода в НБД позволило зафиксировать 250 серий переводов, каждой из которых была присвоена одна из шести категорий, характеризующих траекторию на заданном интервале времени [14]:

- 1) повышение качества МП в серии;
- 2) снижение качества МП;
- 3) колебание качества МП;
- 4) изменение набора ошибок в МП без динамики его качества;
- 5) изменение МП без динамики его качества;
- 6) серия МП без изменений.

Пример текстового фрагмента, машинные переводы которого на французский язык продемонстрировали снижение качества на протяжении 12 месяцев проведения первого эксперимента, представлен в табл. 1.

В ходе другого эксперимента этот метод мы применили для оценки воспроизводимости переводов на английский язык 250 текстовых фрагментов, выполненных с помощью Google Translate. Для английских серий переводов была зафиксирована высокая степень стабильности на протяжении 10 месяцев: около 95% серий переводов остались без изменений. При этом качество переводов также не изменялось, т.е. если сначала ошибка была зафиксирована, то, как правило, она сохранялась в последующих версиях перевода. Например, представленный в табл. 2 перевод фрагмента на английский язык оставался стабильным на протяжении 10 месяцев мониторинга, при этом во всех версиях перевода сохранялась лексическая ошибка: из-за омонимии лексем «души» (мн. ч. от существительного женского рода «душа») и «души» (мн. ч. от существительного мужского рода «душ») произошел ошибочный выбор английского эквивалента (т.е. «souls» вместо «showers»).

Два описанных выше эксперимента по мониторингу качества МП иллюстрируют предлагаемый нами подход к оценке воспроизводимости результатов МП, который был использован в эксперименте с ChatGPT, представленном в настоящей статье.

Метод оценки воспроизводимости результатов МП, схема которого представлена на рис. 1, включает пять этапов. Прежде всего, формируется массив экспериментальных данных с применением поиска в НКРЯ, в котором учитываются: жанр текстов, период их создания, авторство текстов, то количество текстовых фрагментов, которое необходимо для проведения эксперимента. Затем выбирается машинный переводчик, выполняется перевод отобранных фрагментов в заданные моменты времени и производится их запись в НБД. Таким образом, для каждого отобранного текстового фрагмента формируется серия его машинных переводов.

Таблица 1

Переводы фрагмента с русского языка на французский с помощью Google Translate, характеризующиеся снижением качества на временном интервале в один год

Исходный текстовый фрагмент	Коннектор в исходном тексте	Перевод исходного фрагмента	Коннектор в переводе и типы ошибок
Но так как при этом он употреблял и души, то результаты этого лечения подвергаются, конечно, сомнению... («Преступление и наказание», Ф.М. Достоевский)	так как то	Mais comme il utilisait en même temps des âmes, les résultats de ce traitement sont bien sûr remis en question ... (МП с Google Translate зафиксирован в 1-й месяц мониторинга (02.03.2019))	comme ErrorSemant
		Mais comme il a également consommé des âmes, les résultats de ce traitement sont bien sûr remis en question ... (МП с Google Translate зафиксирован в 12-й месяц мониторинга (05.02.2020))	comme 2 ErrorSemant

Таблица 2

Перевод фрагмента текста на английский язык с помощью Google Translate

Исходный текстовый фрагмент	Коннектор в исходном тексте	Перевод исходного фрагмента	Коннектор в переводе и типы ошибок
Но так как при этом он употреблял и души, то результаты этого лечения подвергаются, конечно, сомнению...	так как то	But since at the same time he also used souls, the results of this treatment are, of course, doubtful ...	since ErrorSemant

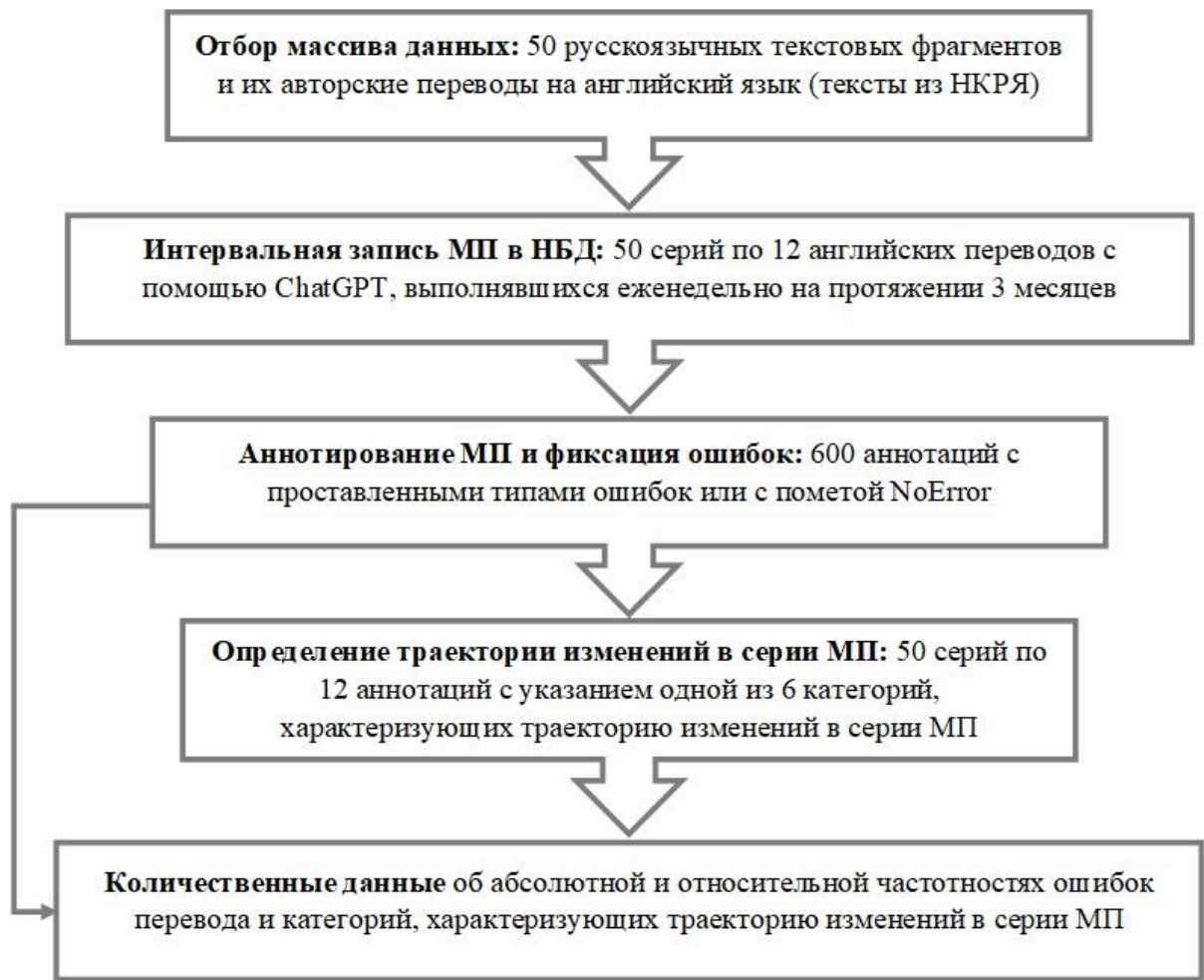


Рис. 1. Схема метода оценки воспроизводимости результатов машинного перевода

Затем эксперты составляют аннотации полученных серий переводов, фиксируют ошибки МП при их наличии и по результатам аннотирования определяют для каждой серии переводов категорию, характеризующую траекторию изменений МП в серии. В итоге экспертный метод оценки воспроизводимости результатов МП позволяет получать количественные данные об абсолютной и относительной частотностях ошибок перевода, а также о процентном соотношении разных категорий.

ПРИМЕНЕНИЕ МЕТОДА ОЦЕНКИ ВОСПРОИЗВОДИМОСТИ РЕЗУЛЬТАТОВ МАШИННОГО ПЕРЕВОДА

Для исследования воспроизводимости английских МП с помощью ChatGPT было отобрано 50 русскоязычных фрагментов, которые ранее продемонстрировали те или иные изменения при переводе на французский язык с помощью Google Translate (по 10 фрагментов для каждой из пяти категорий изменчи-

вых траекторий, т. е., кроме категории 6). При отборе принималось во внимание разнообразие коннекторов во фрагментах.

Эти 50 фрагментов переводились на английский язык посредством ChatGPT версии 3.5 [15] с периодичностью один раз в неделю на протяжении 3-х месяцев. Ежедневный мониторинг позволил зафиксировать высокочастотные колебания в результатах перевода. Всего было проведено 12 сессий машинного перевода для 50 фрагментов, т.е. для каждого фрагмента на русском языке было получено 12 переводов на английский язык. Их аннотирование с использованием классификации ошибок (табл. 3) позволило сформировать 600 аннотаций. В табл. 4 представлен пример серии из 12 МП одного из 50 фрагментов.

Результаты эксперимента обрабатываются и хранятся в надкорпусной базе данных коннекторов, где и были сформированы аннотации¹ для каждого из 600 переводов.

Таблица 3

Трехуровневая классификация ошибок в машинных переводах (без рубрики NoError)

Классификация ошибок в машинных переводах		
1-й уровень	2-й уровень	3-й уровень
Общезыковые ошибки	1. ErrorGram (грамматическая ошибка)	1.1 Морфологическая ошибка во фрагменте текста, не вводимом коннектором
		1.2 Морфологическая ошибка во фрагменте текста, вводимом коннектором
		1.3 Синтаксическая ошибка во фрагменте текста, не вводимом коннектором
		1.4 Синтаксическая ошибка во фрагменте текста, вводимом коннектором
	2. ErrorSemant (лексическая ошибка, искажение смысла в переводимом фрагменте)	2. Лексическая ошибка
	3. ErrorOrth (орфографическая ошибка)	3.1 Орфографическая ошибка (кроме формы коннектора)
		3.2 Орфографическая ошибка в форме коннектора
	4. ErrorPunct (пунктуационная ошибка)	4. Пунктуационная ошибка
Ошибки в употреблении коннектора	5. Cyrillic (слова кириллицей)	5. Слова кириллицей
	6. Latin (русское слово латинским шрифтом)	6. Русское слово латинским шрифтом
	7. Lacuna (пропуск фрагмента текста)	7. Пропуск фрагмента текста
	8. Pleonasm (избыточный перевод, повтор)	8. Избыточный перевод, повтор
	9. ErrorPartCNT (часть неоднословного коннектора переведена ошибочно/не переведена)	9.1 Переведена только первая часть неоднословного коннектора
		9.2 Переведена только вторая часть неоднословного коннектора
		9.3 Первая часть неоднословного коннектора переведена ошибочно
		9.4 Вторая часть неоднословного коннектора переведена ошибочно
	10. ErrorTotalCNT (коннектор переведен несуществующей единицей)	10. Коннектор переведен несуществующей единицей
	11. ErrorMorphCNT (коннектор ошибочно заменен языковой единицей, не являющейся коннектором)	11. Коннектор ошибочно заменен языковой единицей, не являющейся коннектором
	12. ErrorCNT (семантическая ошибка в выборе коннектора)	12. Семантическая ошибка в выборе коннектора

¹ Ввод результатов машинного перевода в НБД был выполнен С.Т. Мачавариани и В.О. Романенко.

**Серия из 12 переводов фрагмента текста на английский язык с помощью ChatGPT
(источник исходного фрагмента – «Преступление и наказание», Ф.М. Достоевский)**

№ п/п	Исходный текстовый фрагмент	Коннектор в исходном тексте	Перевод исходного фрагмента	Коннектор в переводе и типы ошибок	Дата перевода
1	Но так как при этом он употреблял и души, то результаты этого лечения подвергаются, конечно, сомнению...	так как то	But since he used a person's <u>soul</u> in the process, the results of this treatment are certainly questionable...	since <ErrorSemant>	24.10.2023
2			But since he used <u>brandy</u> along with it, the effectiveness of this treatment is, of course, open to doubt...	since <ErrorSemant>	31.10.2023
3			But since he was also <u>consuming</u> <u>spirits</u> during this treatment, the results of this therapy are, of course, questionable...	since <2ErrorSemant>	07.11.2023
4			But as he also <u>consumed</u> <u>spirits</u> , the results of this treatment are, of course, questionable...	as <2ErrorSemant>	14.11.2023
5			But since he used ' <u>dukha</u> ' as a remedy, the results of this treatment are certainly doubtful...	since <Latin>	21.11.2023
6			But since he used (consumed) <u>spirits</u> during this treatment, the results of this therapy are, of course, open to doubt...	since <ErrorSemant>	28.11.2023
7			But since he also used ' <u>dukh</u> ' (soul/spirit) during this treatment, the results of this therapy are, of course, questionable...	since <Latin>	05.12.2023
8			But since he used <u>souls</u> in this process, the results of this treatment are certainly subject to doubt...	since <ErrorSemant>	12.12.2023
9			But since he also consumed <u>alcohol</u> during this treatment, the results of this therapy are, of course, questionable...	since <ErrorSemant>	19.12.2023
10			But since , in doing so, he also <u>consumed</u> <u>souls</u> , the results of this treatment are, of course, subject to doubt...	since <2ErrorSemant>	26.12.2023
11			But as he used ' <u>souls</u> ' during this treatment, the results of this therapy are, of course, subject to doubt...	as <ErrorSemant>	02.01.2024
12			But as he was <u>consuming</u> the <u>souls</u> at the same time, the results of this treatment are, of course, subject to doubt...	as <2ErrorSemant>	09.01.2024

Аннотация состоит из следующих частей (табл. 4): исходный текстовый фрагмент (предложение, несколько предложений или часть предложения), содержащий коннектор, его перевод на английский язык, выполненный с помощью ChatGPT, и описание² ошибок при их наличии.

Для формирования аннотаций был использован второй уровень ранее разработанной классификации ошибок МП [10], которая применялась и в предыдущих экспериментах [16]. Второй уровень, помимо рубрики «Ошибок нет», включает 12 рубрик, характеризующих как ошибки в употреблении коннектора (4 рубрики), так и общеязыковые ошибки в переведенном фрагменте (8 рубрик). Связи между вторым и третьим уровнями классификации ошибок показаны

в табл. 3. Выбор второго уровня классификации ошибок обусловлен его достаточностью для решения задачи мониторинга воспроизводимости результатов МП с помощью ChatGPT.

Затем все зафиксированные в аннотациях ошибки были проанализированы экспертом в пределах каждой серии переводов в сопоставлении друг с другом. Таким образом, было выявлено, какие типы ошибок появляются/исчезают с течением времени в серии переводов одного и того же фрагмента текста на русском языке.

Как уже было отмечено, характер траектории изменений в серии машинного перевода на заданном интервале времени определяет одну из шести категорий [14], к которой относится эта серия. В табл. 4 приведена серия из 12 аннотаций для переводов фрагмента текста на английский язык, полученных с помощью ChatGPT в рамках эксперимента.

² Аннотирование русскоязычных предложений и их переводов было выполнено А.Ю. Егоровой.

В первых двух переводах (см. табл. 4) зафиксировано по одной лексической ошибке (*soul* и *brandy*); в переводах 3 и 4 – по две лексических ошибки (*consuming* и *spirits*); в 5 – 9 версиях перевода лексические ошибки чередуются с ошибками типа *Latin* (написание русского слова «души» латинским шрифтом, а именно *dukha* и *dukh*); в 10 и 12 переводах зафиксировано по две лексических ошибки, схожих по своему содержанию с ошибками из переводов 3 и 4. Такой характер траектории изменения результатов МП обуславливает отнесение представленной серии аннотаций к категории, характеризующей колебание качества машинного перевода.

АНАЛИЗ РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТА

В ходе эксперимента по мониторингу результатов МП, полученных с помощью ChatGPT, все 600 переводных соответствий были проаннотированы, т.е. для каждого переводного соответствия был проставлен тип коннектора в исходном тексте и в переводе, а также зафиксированы ошибки перевода при их наличии.

В табл. 5 представлены количественные данные абсолютной и относительной частотности для типов ошибок, зафиксированных в переводах с помощью ChatGPT.

В 59% полученных аннотаций не было зафиксировано ошибок, что может свидетельствовать об относительно высоком качестве перевода, который чат-бот ChatGPT способен генерировать в языковой паре русский-английский (по нашим данным, в паре русский-французский этот показатель находится на уровне 68%).

На лексические ошибки (*ErrorSemant*) приходится 27% от общего числа зафиксированных ошибок (включая рубрику *NoError*). Пример такой ошибки представлен в первой строке табл. 4.

На грамматические ошибки (*ErrorGram*) приходится 7% ошибок. Этот показатель может указывать на то, что ChatGPT, как правило, успешно справляется с формальной грамматической структурой предложения при переводе на английский язык.

Пунктуационные ошибки (*ErrorPunct*) были зафиксированы в 2% случаев. Это может быть связано с особенностями пунктуационной системы английского языка (меньшая зависимость от синтаксиса, большая гибкость).

Что касается ошибок, связанных с переводом коннекторов, то на них приходится суммарно около 4%. Это может указывать на то, что система МП весьма успешно справляется с подбором адекватных переводных эквивалентов для русских коннекторов. 1% составляет ошибки типа *Latin* (русское слово латинским шрифтом). Такие ошибки возникают, когда системе МП не удается подобрать адекватный переводной эквивалент и слово транслитерируется (см. строка 7 в табл. 4).

По итогам применения метода оценки воспроизводимости результатов МП на интервале времени все серии переводов были распределены по категориям и была дана количественная оценка «веса» каждой категории с указанием их абсолютной и относительной частотности (табл. 6). Больше всего серий (50%) было определено в категорию, характеризующую колебание качества МП. Для этой категории, как было продемонстрировано в табл. 4, характерно попеременное увеличение/уменьшение числа ошибок.

Второй по числу серий переводов (44%) стала категория, характеризующая изменение машинного перевода без динамики его качества. При этом качество МП на интервале времени может оставаться на одном уровне при сохранении ошибок или при их отсутствии.

Категории, характеризующие повышение и снижение качества МП, получили 4% и 2% от общего числа серий переводов, соответственно. Нулевое значение получила рубрика, характеризующая изменение набора ошибок в МП без динамики его качества. Нулевым является и значение последней рубрики в табл. 6, что указывает на то, что переводы 50 отобранных русскоязычных фрагментов, выполненные с помощью ChatGPT, проявляют нестабильность, т.е. со временем меняются.

Таблица 5

Количественные данные по типам ошибок в машинном переводе, включая рубрику *NoError*

№ п/п	Код ошибки	Число ошибок	Доля от общего числа ошибок (включая случаи <i>NoError</i>), %
1	NoError	385	59
2	ErrorSemant	178	27
3	ErrorGram	43	7
4	ErrorPunct	14	2
5	ErrorCNT	10	2
6	ErrorPartCNT	7	1
7	ErrorTotalCNT	7	1
8	Latin	4	1
9	Cyrillic	0	0
10	Lacuna	0	0
11	Pleonasm	0	0
12	ErrorOrth	0	0
13	ErrorMorphCNT	0	0
	Всего:	648	100,0

**Распределение серий переводов по категориям,
характеризующим изменение результатов машинного перевода**

№ п/п	Название рубрики, характеризующей траекторию серии МП	Число серий переводов	Доли от общего числа серий переводов, %
1	Повышение качества МП	2	4
2	Снижение качества МП	1	2
3	Колебание качества МП	25	50
4	Изменение набора ошибок в МП без динамики его качества	0	0
5	Изменение МП без динамики его качества	22	44
6	МП без изменений	0	0
Всего:		50	100,0

ЗАКЛЮЧЕНИЕ

Эксперимент с применением метода оценки воспроизводимости результатов машинного перевода позволяет сформулировать выводы о работе чат-бота ChatGPT как русско-английского машинного переводчика.

Во-первых, в процессе повторных генераций переводов иногда может наблюдаться как улучшение, так и ухудшение качества МП, но, как правило, результат повторного применения ChatGPT характеризуется не однонаправленным изменением качества, а скорее его колебанием на интервале мониторинга.

Во-вторых, разработанный метод оценки воспроизводимости результатов МП на интервалах времени дает возможность:

- определять качество перевода экспериментального текстового массива (как долю элементов массива, которые были переведены без ошибок);

- вычислять абсолютные и относительные частотности типов ошибок, включенных в классификацию из работы [10];

- распределять серии переводов по категориям, характеризующим траекторию изменений МП с указанием абсолютной и относительной частотностей каждой категории.

В-третьих, в процессе работы ChatGPT как русско-английского машинного переводчика наиболее часто встречаются лексические ошибки, на которые приходится 27% от общего числа зафиксированных ошибок.

В заключение отметим, что проведенный эксперимент продемонстрировал на примере чат-бота ChatGPT новизну представленного метода оценки воспроизводимости машинного перевода, не имеющего отечественных и зарубежных аналогов. Этот метод впервые предоставил возможность определять траектории изменений в сериях повторных машинных переводов на временном интервале.

СПИСОК ЛИТЕРАТУРЫ

- Belz A.A. Metrological Perspective on Reproducibility in NLP // *Computational Linguistics*. – 2022. – Vol. 48, Iss. 4. – P. 1125–1135. DOI: 10.1162/coli_a_00448.
- Franc J.M., Cheng L., Hart A., Hata R., Hertelendy A. Repeatability, reproducibility, and diagnostic accuracy of a commercial large language model (ChatGPT) to perform emergency department triage using the Canadian triage and acuity scale // *Canadian Journal of Emergency Medicine*. – 2024. – Vol. 26, Iss. 1. – P. 40-46. DOI:10.1007/s43678-023-00616-w.
- Funk P.F. et al. ChatGPT's Response Consistency: A Study on Repeated Queries of Medical Examination Questions // *European Journal of Investigation in Health, Psychology and Education*. – 2024. – Vol. 14. – P. 657–668. DOI:10.3390/ejihpe14030043.
- Kochanek K., Skarzynski H., Jedrzejczak W.W. Accuracy and Repeatability of ChatGPT Based on a Set of Multiple-Choice Questions on Objective Tests of Hearing // *Cureus*. – 2024. – 16(5): e59857. – URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11157293/>
- Harel D. *Algorithmics* // *The Spirit of Computing*. – Reading, MA, USA: Addison-Wesley, 1987. – 514 p.
- De Winter J.C. Can ChatGPT Pass High School Exams on English Language Comprehension? // *International Journal of Artificial Intelligence in Education*. – 2023. DOI:10.1007/s40593-023-00372-z.
- Егорова А.Ю., Зацман И.М., Романенко В.О. Машинный перевод с помощью ChatGPT: мониторинг воспроизводимости результатов // *Системы и средства информатики*. – 2023. – Т. 33, № 3. – С. 117-128.
- Zhang M. A Study on the Translation Quality of ChatGPT // *International Journal of Educational Curriculum Management and Research*. – 2024. – Vol 5, Iss. 1. – P. 153-163. DOI:10.38007/IJECMR.2024.050121.
- Кружков М.Г. Концепция построения надкорпусных баз данных // *Системы и средства информатики*. – 2021. – Т. 31, № 3. – С. 101-112.
- Бунтман Н.В., Гончаров А.А., Зацман И.М., Нуриев В.А. Количественный анализ результатов машинного перевода с использованием надкорпусных баз данных // *Информатика и ее применения*. – 2018. – Т. 12, № 4. – С. 96-105.
- Национальный корпус русского языка. – URL: <https://ruscorpora.ru/>

12. Инькова-Манзотти О.Ю. Коннекторы противопоставления во французском и русском языках. Сопоставительное исследование. – Москва: Информ-электро, 2001. – 429 с.
13. Егорова А.Ю., Зацман И.М., Кружков М.Г., Нуриев В.А. Нестабильность нейронного машинного перевода // Системы и средства информатики. – 2020. – Т. 30, № 2. – С. 124-135.
14. Егорова А.Ю., Зацман И.М., Кружков М.Г., Нуриев В.А. Индикаторная оценка неустойчивости нейронного машинного перевода // Системы и средства информатики. – 2021. – Т. 31, № 2. – С. 139-151.
15. ChatGPT. – URL: <https://chat.openai.com/>
16. Егорова А.Ю., Зацман И.М., Кружков М.Г., Нуриев В.А. Методика темпоральной оценки неустойчивости машинного перевода // Системы и средства информатики. – 2020. – Т. 30, № 3. – С. 67-80.

Материал поступил в редакцию 25.09.24.

Сведения об авторах

ЕГОРОВА Анна Юрьевна – научный сотрудник Федерального исследовательского центра "Информатика и управление" Российской академии наук; старший преподаватель Московского государственного лингвистического университета, Москва
e-mail: anna.yu.egorova@yandex.ru

ЗАЦМАН Игорь Моисеевич – доктор технических наук, заведующий отделом Федерального исследовательского центра "Информатика и управление" Российской академии наук, Москва
e-mail: izatsman@yandex.ru

РОМАНЕНКО Варвара Олеговна – инженер 1-й категории Федерального исследовательского центра "Информатика и управление" Российской академии наук, Москва
e-mail: varvarorama@gmail.com

УКАЗАТЕЛЬ СТАТЕЙ, опубликованных в сборнике «Научно-техническая информация», и АВТОРСКИЙ УКАЗАТЕЛЬ за 2024 год

Указатель статей

ОБЩИЙ РАЗДЕЛ					
Максимов Н.В. Знания и информация в контексте естественного и искусственного интеллекта	1 (1)	1*	Гоннова С.М., Пивинская И.И., Разуваева Е.Ю. Проблемы национального технологического суверенитета и потенциал государств – участников СНГ	11 (1)	1
Сюнтюрено О.В., Червинская Н.В. Информационное обеспечение науки и промышленности Китая	1 (1)	15	Данилина Я.В., Гербина Т.В. К проблеме систематизации предметной области: устойчивое развитие	12 (1)	1
Косяков Д.В., Блинов П.Ю. Научная результативность федеральных исследовательских центров: есть ли эффект от реструктуризации?	2 (1)	1	ОРГАНИЗАЦИЯ ИНФОРМАЦИОННОЙ РАБОТЫ		
Тодосийчук А.В. Наукометрические показатели в системе оценки результативности науки и научного труда	3 (1)	1	Шорин О.Н. Коммуникационные процессы в научной среде	1 (1)	21
Кохно П.А. Теории информации оценки вероятности событий	4 (1)	1	Васильева В.С. О роли национальной киноиндустрии в экономике страны: наукометрические и статистические методы оценки	1 (1)	30
Шрайберг Я.Л., Дмитриева Е.Ю., Смирнова О.В., Червинская Н.В., Метлова А.В. Построение взаимных отражений рубрик ГРНТИ и индексов УДК по тематическим направлениям естественных и точных наук	5 (1)	1	Мельникова Е.В. Актуальность применения инструментария искусственного интеллекта в современных наукометрических исследованиях	2 (1)	19
Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А. Наука с интенсивным использованием данных. Проблемы и развитие четвертой парадигмы	5 (2)	1	Шрайберг Я.Л., Мосеева Д.С. О создании национальной системы научно-технической информации в условиях нарастающей цифровизации и в новых социально-политических реалиях	3 (1)	8
Редькина Н.С. Влияние инфраструктуры открытой науки на развитие мирового рынка информационных ресурсов	6 (1)	1	Ударцева О.М. Вебометрическая оценка зарубежных информационных систем текущих исследований	4 (1)	6
Ибрагимов Р.С., Яшалова Н.Н. О перспективах применения робототехники в промышленном секторе экономики	7 (1)	1	Ермолаев В.А., Яшалова Н.Н., Щербина А.В., Рубан Д.А. Библиографическая база данных как инструмент научного подхода к решению задачи социально-экономического развития (на примере речного туризма)	4 (1)	15
Максимов Н.В. Информация и информационные взаимодействия	7 (2)	1	Миронов В.В. Наукометрический анализ субъективного интереса участников научного мероприятия к представленным докладам и мероприятию в целом	4 (1)	23
Мельникова Е.В. Искусственный интеллект как природоподобная технология и его применение в информационном и библиотечном деле	8 (1)	1	Дворовенко О.В. Методические службы библиотек в социальных сетях: назначение, виды, формы представления	5 (1)	7
Калачихин П.А. Об экономической эффективности классификаторов научных знаний	9 (1)	3	Шведенко В.В. Анализ инструментальных средств разработки информационного обеспечения умных пространств	6 (1)	13
Михеенкова М.А. Формализованная эвристика построения объясняющей типологии	9 (2)	3	Алексеев Д.С. Методология определения косвенных информационных признаков сложных систем	7 (1)	7
Тодосийчук А.В. Национальные цели в системе управления научно-технологическим развитием	10 (1)	1			

* 1 – означает номер сборника, (1) – серию, 1 – страницу

Васильева И.Н., Корнеева Н.Д., Богатова Р.С. Расширение научно-технического сотрудничества России со странами Ближнего Востока и Северной Африки на основе использования потенциала ученых-соотечественников и данных публикационной активности	7 (1) 13	Бескаравайная Е.В., Харыбина Т.Н. Характеристики информационных потоков для научных исследований	7 (1) 24
Кохно П.А. Инновационное производство с учётом информации о неопределённостях и рисках	8 (1) 7	Гуреев В.Н., Курмышева Л.К., Мазов Н.А. Обзор исследований по экспертной оценке российских научных журналов	8 (1) 14
Баканов А.С. Анализ информационных ресурсов организации с использованием ключевых слов	9 (1) 11	Редькина Н.С. Веб-архивы: классификация и способы организации	9 (1) 18
Бескаравайная Е.В., Харыбина Т.Н. Выбор и использование автоматизированных библиотечно-информационных систем	10 (1) 10	Анисимов И.В., Макеенко М.И., Трищенко Н.Д. Подходы к отбору источников для цитирования в статьях российских исследователей медиакоммуникации	9 (1) 25
Пителинский К.В., Маковой С.О., Цапин Д.М. Дихотомический вопрос роботизации и ее каузальный характер: создание интеллектуальной мобильной системы	11 (1) 12	Мазов Н.А., Гуреев В.Н., Синявина С.В. Актуальность создания полнофункционального электронного архива печатных версий российских научных журналов	10 (1) 17
Шевченко Л.Б. Обучение использованию инструментов для поддержки открытых научных исследований	11 (1) 25	Гребенщикова Е.Г. Научные публикации в эпоху искусственного интеллекта	11 (1) 39
Слащева Н.А., Цветкова В.А. Возможности интеллектуального анализа данных в научной библиотеке	11 (1) 33	Мустафаев Ф.Ф., Джабаров С.Г. Влияние нововведенных требований к присуждению ученых степеней на научные показатели страны	12 (1) 23
Калюжная Т.А., Плешакова М.А. Библиометрический анализ темы «Репозиторий исследовательских данных»	12 (1) 12	Якубов А.В. Об оцифрованности авторефератов диссертаций чеченских ученых	12 (1) 29

ДОКУМЕНТАЛЬНЫЕ ИСТОЧНИКИ ИНФОРМАЦИИ

Мыльникова А.В., Чернышева П.А. Повышение качества перевода при использовании методов изменения порядка слов при предобработке текстов для систем машинного перевода	2 (1) 26
Фролова Т.И., Кажберова В.В. Отечественный медиаконтент: динамика основных повесток (2017–2023 гг.)	3 (1) 14
Крымская А.С. Информационные ресурсы профессиональных ассоциаций в сфере мировой политики и международных отношений	4 (1) 31
Дьяченко Е.Л., Губа К.С., Потопов И.В., Мироненко А.Ю. Сравнение подходов к стратификации российских журналов: наукометрические индикаторы, международные базы данных и национальные списки	5 (1) 13
Гоннова С.М., Чуйкова Н.А., Пивинская И.И., Разуваева Е.Ю. Национальные системы научной и научно-технической экспертизы стран СНГ – информационный ресурс для межгосударственного обмена	6 (1) 22

ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ

Съедин Д.Ю. Разработка модификации стохастического алгоритма глобальной оптимизации РСА, обеспечивающей приближенное решение задачи коммивояжера	1 (2) 1
Мильман Б.Л., Журкович И.К. «Социтирование социтирований»: структурирование массива научных публикаций, отражающих современные библиометрические методы	3 (2) 1
Кохно П.А. Поиск и аналитическая переработка информации для ее внедрения в промышленное производство	3 (2) 9
Бетин В.Н., Ивашенко В.А., Супрун А.П. Использование регулярных особенностей частично определенных функциональных нейронных сетей при поиске решения	4 (2) 1
Белоусов К.И., Сазина Д.А., Рябинин К.В., Брохин Л.Ю. Сенсорная технолингвистика: о механизмах передачи мультимодальных сообщений в перцептивно-когнитивных интерфейсах	4 (2) 9
Аветисян Т.В., Меньяйлов Д.В., Преображенский А.П. Оценка подхода к обнаружению и мониторингу событий в социальных сетях для борьбы с природными катастрофами	4 (2) 18

Россихина Л.В., Орлова Д.Е., Чертов В.А. Информационная поддержка принятия решений при управлении региональными инвестиционно-строительными проектами: системный подход	6 (2) 1	Аментес А.В., Забейхайло М.И. Об опыте разработки атласа биомаркеров исхода нейрохирургических операций	8 (2) 1
Мыльников Л.А., Салтыкова А.Д., Аврамович З.Ж. Нотация <i>Event-Driven Process Methodology</i> для исследования информационных процессов	7 (2) 19	Ефимова Е.А. Алгебраическая спецификация правдоподобных рассуждений	10 (2) 22
Сарин К.С., Коломников Р.Е., Светлаков М.О., Ходашинский И.А. Нечеткий классификатор инкрементного обучения типа Min-Max для решения задач кибербезопасности	9 (2) 11	Шестерникова О.П., Фабрикантова Е.Ф., Романов Д.В. Об отражении в ДСМ-методе автоматизированной поддержки исследований причинности в медицине (на примере установления связи негативной шизофрении с генетическими параметрами)	12 (2) 9
		ИНФОРМАЦИОННЫЕ СИСТЕМЫ	
Шведенко В.В. Использование метода роевого управления для поддержания гомеостаза сложной системы на основе взаимодействующих информационных объектов	9 (2) 22	Пителинский К.В., Бритвина В.В., Александрова А.В. Цифровые двойники и базовый цифровой профиль обучающегося	2 (2) 9
Родионов А.Н. Композиционные зависимости и формулы высокопорядковых отношений реляционных моделей данных	10 (2) 1	Цветкова В.А., Махно О.О. Модель и принципы построения автоматизированной библиотечно-информационной системы	4 (2) 30
Мыльников Л.А., Салтыкова А.Д. Исследование производительности информационных процессов в нотации <i>Event-Driven Process Methodology (EPM)</i>	10 (2) 15	Шведенко В.Н., Алексеев Д.С. Способ получения дополнительной информации на основании изучения волновых явлений в аппаратной части информационной системы управления объектами предметной области	5 (2) 15
Сидняев Н.И., Синева Е.Е. Информация о фреймах и математическое моделирование баз знаний	11(2) 1	ИНФОРМАЦИОННЫЙ ПОИСК	
Мыльников Л.А., Салтыкова А.Д. Структурно-функциональное моделирование в нотации <i>Event-driven Process Methodology</i> как инструмент для генерации кода программ, имитирующих процессную деятельность	12 (2) 1	Сидняев Н.И., Бутенко Ю.И., Синева Е.Е. Стратегии поиска в пространстве состояний баз знаний	6 (2) 25
ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ		ИНФОРМАЦИОННЫЙ АНАЛИЗ	
Финн В.К. Об эмпирических закономерностях ранга r в ДСМ-методе автоматизированной поддержки исследований	1 (2) 11	Федорец О.В., Солошенко Н.С. О практической значимости информетрической модели Брэдфорда для прогнозирования рассеяния статей и оптимизации отбора журналов	1 (2) 34
Забейхайло М.И. Интеллектуальный анализ данных как инструмент доказательной медицины	3 (2) 14	Калачихин П.А. Соответствие между иерархическими классификаторами знаний	2 (2) 1
Виноградов Д.В., Якимова Л.А. Вероятностный подход к «доброму старомодному» искусственному интеллекту	3 (2) 21	Васильцов В.С., Крылова Н.П., Сушинская А.В., Красовский М.С. Информационный анализ в управлении рисками бизнес-процессов (на примере металлургических предприятий)	3 (2) 27
Забейхайло М.И., Михеенкова М.А., Трунин Ю.Ю. О небинарном варианте отношения причинности в интеллектуальном анализе онкологических данных	6 (2) 13	Ходашинский И.А. Управление популяционным разнообразием алгоритма «стаи ласточек» для решения задачи нечёткой классификации	5 (2) 26
Гусакова С.М. Особенности обнаружения эмпирических закономерностей в задачах с операцией сходства, соответствующей глобальному сходству	6 (2) 21		

Тимошенко И.В. О моделировании информационной деятельности современных библиотек	7 (2) 31	Егорова А.Ю., Зацман И.М., Романенко В.О. Метод оценки воспроизводимости результатов машинного перевода с помощью ChatGPT	12 (2) 36
Боярский К.К., Каневский Е.А. Инструментарий для анализа документов стратегического планирования регионов	8 (2) 15	СПРАВОЧНО-ИНФОРМАЦИОННЫЙ РАЗДЕЛ	
Жебит В.А. На пути к сознанию машин: идентификация скрытых системных свойств материальных объектов	8 (2) 24		
Кузнецова А.С., Шефер О.Р., Лебедева Т.Н. Визуализация информации как неотъемлемый этап компьютерного моделирования астрофизических объектов	9 (2) 31	Антопольский А.Б., Вигурский К.В., Коргунюк Ю.Г. О модернизации информационной системы «ПартАрхив»	1 (1) 42
Булдакова Т.И., Вишневская Ю.А., Джалолов А.Ш., Суятинов С.И. Оценка социально-экономического состояния региона на основе модели Хакена	10 (2) 33	Гоннова С.М., Чуйкова Н.А., Разуваева Е.Ю. Электронные системы мониторинга – инструмент управления реализацией основополагающих документов СНГ	2 (1) 36
Жебит В.А. На пути к сознанию машин: о ментальности физических систем	10 (2) 42	Иншакова Н.Г., Панкеев И.А. Технический писатель как коммуникационная специальность: редакторская трактовка профессиограммы	2 (2) 37
Мамедова Н.А., Борцов А.С. Автоматизированное решение задачи качественного анализа изображений на основе вариации Лапласа для разработки мобильного приложения	11 (2) 21	Антопольский А.Б., Гук Д.Ю. О подготовке русского перевода онтологии по культурному наследию CIDOC CRM	3 (1) 24
Чиняков Н.А. Модель генерации текстового описания мультимодальных данных	11 (2) 35	Сухоручкина И.Н., Сухоручкина А.А. Классификация разработанных и используемых в России программ автоматического и автоматизированного перевода научно-технических документов на основе искусственного интеллекта	5 (1) 22
Федорец О.В., Солошенко Н.С. Исследование тематической и терминологической трансформации коллекций научных статей по междисциплинарным областям знания	12 (2) 15	Плешкевич Е.А. Важный вклад в теорию документоведения и документологии	8 (1) 26
Водяхо А.И., Жукова Н.А., Ананьева В.Я. Использование цифровых нитей и систем цифровых двойников в социо-кибер-физических системах	12 (2) 26	Шамаев В.Г., Горшков А.Б. Полнотекстовый архив «Акустического журнала» в интернет-портале «Акустика» (русскоязычные источники)	9 (1) 31
АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА		Сухоручкина И.Н., Сухоручкина А.А. Информационное обеспечение научно-технического сотрудничества Российской академии наук и Китайской академии наук в рамках БРИКС	10 (1) 31
		Антопольский А.Б., Вигурский К.В., Коргунюк Ю.Г. Информационная система «ПартАрхив» как инструмент развития методологической базы политологических исследований	11 (1) 44
Бабина О.И. Тематическое моделирование корпуса текстов отзывов для извлечения аспектов оценки из мнений	2 (2) 21	Чулок А.И. Рецензия на опытно-конструкторскую разработку программного обеспечения	12 (1) 32
Бутенко Ю.И. Метод выравнивания многокомпонентных англо- и русскоязычных терминов в параллельном корпусе на основе пакета SimAlign	8 (2) 29		
Исмагулов З.С., Косяков Д.В., Гуськов А.Е. Сравнение подходов к извлечению математических методов из научных текстов	11 (2) 40		

Авторский указатель

Аветисян Т.В.	4 (2) 18	Егорова А.Ю.	12 (2) 36	Мироненко А.Ю.	5 (1) 13
Аврамович З.Ж.	7 (2) 19	Еркимбаев А.О.	5 (2) 1	Миронов В.В.	4 (1) 23
Александрова А.В.	2 (2) 9	Ермолаев В.А.	4 (1) 15	Михеенкова М.А.	6 (2) 13
Алексеев Д.С.	5 (2) 15	Ефимова Е.А.	10 (2) 22		9 (2) 3
	7 (1) 7			Мосеева Д.С.	3 (1) 8
Аментес А.В.	8 (2) 1	Жебит В.А.	8 (2) 24	Мустафаев Ф.Ф.	12 (1) 23
Ананьева В.Я.	12 (2) 26		10 (2) 42	Мыльников Л.А.	7 (2) 19
Анисимов И.В.	9 (1) 25	Жукова Н.А.	12 (2) 26		10 (2) 15
Антопольский А.Б.	1 (1) 42	Журкович И.К.	3 (2) 1		12 (2) 1
	3 (1) 24			Мыльникова А.В.	2 (1) 26
	11 (1) 44	Забежайло М.И.	3 (2) 14		
			6 (2) 13	Орлова Д.Е.	6 (2) 1
			8 (2) 1		
Бабина О.И.	2 (2) 21	Зацман И.М.	12 (2) 36	Панкеев И.А.	2 (2) 37
Баканов А.С.	9 (1) 11	Зицерман В.Ю.	5 (2) 1	Пивинская И.И.	6 (1) 22
Белоусов К.И.	4 (2) 9				11 (1) 1
Бескаравайная Е.В.	7 (1) 24	Ибрагимов Р.С.	7 (1) 1	Пителинский К.В.	2 (2) 9
	10 (1) 10	Иващенко В.А.	4 (2) 1		11 (1) 12
Бетин В.Н.	4 (2) 1	Иншакова Н.Г.	2 (2) 37	Плешакова М.А.	12 (1) 12
Блинов П.Ю.	2 (1) 1	Исмагулов З.С.	11 (2) 40	Плешкевич Е.А.	8 (1) 26
Богатова Р.С.	7 (1) 13			Потапов И.В.	5 (1) 13
Борцов А.С.	11 (2) 21	Кажберова В.В.	3 (1) 14	Преображенский А.П.	4 (2) 18
Боярский К.К.	8 (2) 15	Калачихин П.А.	2 (2) 1		
Бритвина В.В.	2 (2) 9		9 (1) 3	Разуваева Е.Ю.	2 (1) 36
Брохин Л.Ю.	4 (2) 9	Калюжная Т.А.	12 (1) 12		6 (1) 22
Булдакова Т.И.	10 (2) 33	Каневский Е.А.	8 (2) 15		11 (1) 1
Бутенко Ю.И.	6 (2) 25	Кобзев Г.А.	5 (2) 1	Редькина Н.С.	6 (1) 1
	8 (2) 29	Коломников Р.Е.	9 (2) 11		9 (1) 18
		Коргунюк Ю.Г.	1 (1) 42	Родионов А.Н.	10 (2) 1
Васильева В.С.	1 (1) 30		11 (1) 44	Романенко В.О.	12 (2) 36
Васильева И.Н.	7 (1) 13	Корнеева Н.Д.	7 (1) 13	Романов Д.В.	12 (2) 9
Васильцов В.С.	3 (2) 27	Косяков Д.В.	2 (1) 1	Россихина Л.В.	6 (2) 1
Вигурский К.В.	1 (1) 42	Косяков Д.В.	11 (2) 40	Рубан Д.А.	4 (1) 15
	11 (1) 44	Кохно П.А.	3 (2) 9	Рябинин К.В.	4 (2) 9
Виноградов Д.В.	3 (2) 21		4 (1) 1		
Вишневская Ю.А.	10 (2) 33		8 (1) 7	Сазина Д.А.	4 (2) 9
Водяхо А.И.	12 (2) 26	Красовский М.С.	3 (2) 27	Салтыкова А.Д.	7 (2) 19
		Крылова Н.П.	3 (2) 27		10 (2) 15
Гербина Т.В.	12 (1) 1	Крымская А.С.	4 (1) 31		12 (2) 1
Гоннова С.М.	2 (1) 36	Кузнецова А.С.	9 (2) 31	Сарин К.С.	9 (2) 11
	6 (1) 22	Курмышева Л.К.	8 (1) 14	Светлаков М.О.	9 (2) 11
	11 (1) 1			Сидняев Н.И.	6 (2) 25
Горшков А.Б.	9 (1) 31	Лебедева Т.Н.	9 (2) 31		11 (2) 1
Гребенщикова Е.Г.	11 (1) 39			Синева Е.Е.	6 (2) 25
Губа К.С.	5 (1) 13	Мазов Н.А.	8 (1) 14		11 (2) 1
Гук Д.Ю.	3 (1) 24		10 (1) 17	Синявина С.В.	10 (1) 17
Гуреев В.Н.	8 (1) 14	Макеев М.И.	9 (1) 25	Слащева Н.А.	11 (1) 33
	10 (1) 17	Маковой С.О.	11 (1) 12	Смирнова О.В.	5 (1) 1
Гусакова С.М.	6 (2) 21	Максимов Н.В.	1 (1) 1	Солошенко Н.С.	1 (2) 34
Гуськов А.Е.	11 (2) 40		7 (2) 1		12 (2) 15
		Мамедова Н.А.	11 (2) 21	Супрун А.П.	4 (2) 1
Данилина Я.В.	12 (1) 1	Махно О.О.	4 (2) 30	Сухоручкина А.А.	5 (1) 22
Дворовенко О.В.	5 (1) 7	Мельникова Е.В.	2 (1) 19		10 (1) 31
Джабаров С.Г.	12 (1) 23	Мельникова Е.В.	8 (1) 1	Сухоручкина И.Н.	5 (1) 22
Джалолов А.Ш.	10 (2) 33	Меняйлов Д.В.	4 (2) 18		10 (1) 31
Дмитриева Е.Ю.	5 (1) 1	Метлова А.В.	5 (1) 1	Сушинская А.В.	3 (2) 27
Дьяченко Е.Л.	5 (1) 13	Мильман Б.Л.	3 (2) 1	Суятинов С.И.	10 (2) 33

Съедин Д.Ю.	1 (2) 1	Харыбина Т.Н.	7 (1) 24	Шамаев В.Г.	9 (1) 31
Сюнтюренко О.В.	1 (1) 15		10 (1) 10	Шведенко В.В.	6 (1) 13
		Ходашинский И.А.	5 (2) 26		9 (2) 22
Тимошенко И.В.	7 (2) 31		9 (2) 11	Шведенко В.Н.	5 (2) 15
Тодосийчук А.В.	3 (1) 1	Цапин Д.М.	11 (1) 12	Шевченко Л.Б.	11 (1) 25
	10 (1) 1	Цветкова В.А.	4 (2) 30	Шестерникова О.П.	12 (2) 9
Трищенко Н.Д.	9 (1) 25		11 (1) 33	Шефер О.Р.	9 (2) 31
Трунин Ю.Ю.	6 (2) 13			Шорин О.Н.	1 (1) 21
		Червинская Н.В.	1 (1) 15	Шрайберг Я.Л.	3 (1) 8
Ударцева О.М.	4 (1) 6		5 (1) 1		5 (1) 1
Фабрикантова Е.Ф.	12 (2) 9	Чернышева П.А.	2 (1) 26	Щербина А.В.	4 (1) 15
Федорец О.В.	1 (2) 34	Чертов В.А.	6 (2) 1		
.	12 (2) 15	Чиняков Н.А.	11 (2) 35	Якимова Л.А.	3 (2) 21
Финн В.К.	1 (2) 11	Чуйкова Н.А.	2 (1) 36	Якубов А.В.	12 (1) 29
Фролова Т.И.	3 (1) 14		6 (1) 22	Яшалова Н.Н.	4 (1) 15
		Чулок А.И.	12 (1) 32		7 (1) 1

УВАЖАЕМЫЕ КОЛЛЕГИ!

ВИНИТИ РАН предлагает Вашему вниманию Реферативный Журнал в электронной форме

РЖ в электронной форме (ЭлРЖ) выпускается по всем разделам естественных, технических и точных наук.

Каждый номер ЭлРЖ является полным аналогом печатного номера РЖ по составу описаний документов, их оформлению и расположению. Он сопровождается оглавлением, указателями.

ЭлРЖ представляет собой информационную систему, снабженную поисковым аппаратом и позволяющую пользователю на персональном компьютере:

- читать номер РЖ, последовательно листая рефераты;
- просматривать рефераты отдельных разделов по оглавлению;
- обращаться к рефератам по указателям авторов, источников, ключевых слов;
- проводить поиск документов по словам и словосочетаниям;
- выводить текст описаний документов во внешний файл.

ЭлРЖ могут быть:

- записаны на DVD-ROM;
- передаваться через FTP-сервер (клиенту предоставляется логин и пароль с доступом к FTP-серверу ВИНТИ, с которого он скачивает заказанные журналы).

Электронные реферативные журналы можно заказать за текущий год с любого номера, а также за предыдущие годы.

Подробную информацию Вы можете получить:

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

Телефон: 8 499-152-62-11

E-mail: feo@viniti.ru

ВИНИТИ РАН

Центр научно-информационного обслуживания

Информационные услуги, предоставляемые ЦНИО ВИНТИ РАН:

- проведение тематического поиска и консультации поисковых экспертов;
- подготовка списков научной литературы;
- подбор, копирование полнотекстовых материалов из первоисточников на бумажном носителе и в электронном виде;
- библиометрическая оценка публикационной активности исследователей и научных организаций с использованием российских и зарубежных баз данных;
- информационное обеспечение информационно-аналитической деятельности по подготовке и предоставлению аналитических обзоров и других научных материалов.

ВИНИТИ РАН располагает следующими информационными ресурсами:

- фондом НТЛ, включающим более 2,5 млн. отечественных и иностранных журналов, книг, депонированных рукописей, авторефератов диссертаций и другой научной литературы, ретроспектива – с 1991 года;
- базами данных и Интернет-ресурсами: БД ВИНТИ (разработка ВИНТИ), БД SCOPUS, БД Questel (патенты) и другими реферативными ресурсами;
- полнотекстовыми электронными ресурсами (статьи, патенты, материалы конференций).

Ознакомиться с информацией о доступных полнотекстовых и реферативных ресурсах можно на сайте ВИНТИ РАН www.viniti.ru

К услугам пользователей – **Электронный Каталог ВИНТИ** <http://catalog.viniti.ru>
и служба электронной доставки документов.

Осуществляется платное информационное обслуживание по разовым заказам и на договорной основе с предоставлением всех необходимых финансовых документов.

Проводится индивидуальное обслуживание пользователей в читальном зале ЦНИО ВИНТИ РАН.

Подробную информацию Вы можете получить:

Адрес: 125190, Россия, г. Москва, ул. Усиевича, 20, ВИНТИ РАН;

Телефоны: 499-155-42-17, 499-155-42-43;

E-mail: cnio@viniti.ru

УВАЖАЕМЫЕ КОЛЛЕГИ!

ВИНИТИ РАН предлагает Вашему вниманию База данных (БД) ВИНИТИ РАН в режиме online

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 более 600 000 документов в год.

БД ВИНИТИ РАН включает 26 тематических фрагментов, состоящих из более чем 190 разделов.

Документы БД содержат библиографию, ключевые слова, рубрики и реферат первоисточника.

На основе БД ВИНИТИ пользователям доступны следующие продукты:

- online доступ к базе данных круглосуточно, без выходных;
- выполнение тематического поиска специалистом ВИНИТИ по запросу заказчика;
- по заявкам предоставляются любые наборы тематических фрагментов БД ВИНИТИ или их разделов на любых видах электронных носителей, или через FTP-сервер;
- для ознакомления с возможностями поиска имеется демо-версия базы данных bd.viniti.ru.

База данных ВИНИТИ зарегистрирована Российским агентством по правовой охране программ для ЭВМ, баз данных и топологий интегральных микросхем (РосАПО) (Свидетельство № 960034 от 23.09.1996г.)

Подробную информацию Вы можете получить:

Адрес: 125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН

Телефон: 8 499-152-54-81

E-mail: feo@viniti.ru