ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ

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

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 5** **Москва 2020**

ИНФОРМАЦИОННЫЕ СИСТЕМЫ

УДК 004.91

В.С. Егоров, Е.С. Козлова, К.Е. Ломотин, О.В. Федорец, А.В. Филимонов, А.В. Шапкин

Система автоматической классификации текстов для обработки потока научных публикаций в ВИНИТИ РАН*

Представлены результаты разработки и тестирования системы автоматической классификации научных текстов, позволяющей определять тематику текстов по трём классификационным схемам в пакетном и диалоговом режимах. Описаны структурно-функциональные компоненты, используемые методы оценки качества классификации, методика обучения, выбор оптимальной модели классификации, основные направления внедрения автоматического классификатора в технологию обработки электронного документального потока в ВИНИТИ РАН.

Ключевые слова: *автоматическая классификация текста, Word2Vec, машинное обучение, персептрон, логистическая регрессия, обработка естественного языка, производственная технология информационного центра*

DOI: 10.36535/0548-0027-2020-05-1

* Работа выполнена в рамках Госзадания № 0003-2019-0002 "Разработка, лексикографическое и понятийно-терминологическое развитие системы взаимосвязанных классификаций научно-технической информации (ГРНТИ, УДК, ВАК, Scopus, WoS, MSC, МКС). Развитие научно-методической и программно-технологической компонент информационной технологии формирования и анализа качества реферативной базы данных ВИНИТИ. Разработка и экспертиза стандартов системы СИБИД".

ВВЕДЕНИЕ

Технологической основой любого информационного комплекса (СМИ, сайты Интернета, библиотеки, центры научно-информационного обслуживания и т. д.) является производственная система, обеспечивающая для пользователей удобную и оперативную навигацию доступа к создаваемым информационным массивам. Традиционно построение подобной системы базируется на реализации тематической классификации обрабатываемых объектов. Для ВИНИТИ объекты обработки – это научные публикации, т. е. текстовая информация.

Определимся с терминологией, которую будем использовать. Под классификацией текста мы подразумеваем его индексирование по рубрикатору. Здесь термин "рубрикатор" – это синоним термина "классификационная схема"; соответственно, термин "рубрика" – синоним термина "класс". Индексирование по рубрикатору (рубрицирование) – процедура присвоения объекту индексов рубрик (одной или нескольких), взятых из заранее определённого списка тематических рубрик. Типичные примеры – индексирование литературы в библиотеках, или документов информационных центрах. Наиболее известные классификаторы – Универсальная десятичная классификация, Десятичная классификация Дьюи, Библиотечно-библиографическая классификация, Государственный рубрикатор научно-технической информации, Международная патентная классификация.

Присвоение документу классификационного индекса – это весьма дорогостоящая операция, поскольку требует привлечения квалифицированных специалистов в предметной области.

Успехи развития современных компьютерных технологий в области интеллектуального анализа данных позволяют создавать программы автоматической классификации, "обученные" анализировать тексты с целью отнесения их к тем или иным классам конкретного рубрикатора − с той или иной степенью вероятности.

Разработке подобных систем посвящается значительное число теоретических и прикладных исследований в области информационных технологий [1–4]. Примеры практического использования систем автоматической классификации текстов в области обработки научно-технической информации можно найти в литературе последних лет [5–9], где выделяются общие обсуждаемые проблемы:

- выбор моделей классификации и программных инструментов, позволяющих выполнять "интеллектуальную" обработку текстов;
- выбор методов предварительной обработки текста (нормализация, векторизация и пр.);
- подбор гиперпараметров модели классификации, наилучшим образом соответствующих характеру обрабатываемых текстов и используемых классификационных схем;
- обучение автоматического классификатора на коллекциях текстов в заданной области знания;
- оценка качества классификации и достоверности результатов.

В настоящей статье представлена система автоматической тематической классификации текстов (АТКТ), разработанная в ВИНИТИ РАН с использованием известных на настоящее время методов обработки документов, доступных инструментов и программных продуктов, пригодных для индексации [10-12]. Мы постарались дать общие сведения о построении системы АТКТ и особенностях ее реализации в аспекте вышеуказанных проблем, а также показать начальный опыт ее использования для обработки потока поступающих в ВИНИТИ документов.

Отметим некоторые черты технологического процесса обработки научно-технической литературы в ВИНИТИ, знание которых потребуется для понимания целей и задач рассматриваемой разработки.

ТЕХНОЛОГИЯ ИНДЕКСАЦИИ ДОКУМЕНТОВ В ВИНИТИ РАН

Сегодня в ВИНИТИ РАН работают 16 профилированных по тематике отделов – научно-редакционных подразделений: Автоматика и радиоэлектроника, Астрономия, Биология, География, Геология, Информатика, Математика, Машиностроение, Металлургия, Механика, Охрана окружающей среды, Транспорт, Физика, Химия, Экономика промышленности, Электротехника.

Общий объем входного потока документов в настоящее время составляет более 1 млн в год. При этом доля русскоязычных документов составляет 29%, англоязычных − 63%. Ежемесячно формируются и издаются около 190 тематических выпусков Реферативного журнала (РЖ), при этом в разных отделах число выпусков варьируется от 2-3-х до нескольких десятков. В процессе обработки документы проходят три стадии тематического классифицирования.

На первой стадии технологического процесса выполняется разметка входного потока по областям науки и техники − для каждого документа определяется научно-тематический отдел, специалисты которого будут его обрабатывать.

На второй стадии (в отделах) документы в соответствии с тематикой распределяются по выпускам Реферативного журнала. Направление документа в соответствующий выпуск РЖ, как правило, означает определение научного сотрудника, ответственного за его обработку.

На третьей стадии обработки документов научные редакторы − специалисты в своих предметных областях − осуществляют глубокое индексирование документов по Рубрикатору ВИНИТИ, который представляет собой иерархический классификатор, включающий более 50 тыс. рубрик.

Таким образом, задача тематического классифицирования решается в ВИНИТИ последовательно на трех уровнях: *отдел → выпуск РЖ → индекс Рубрикатора*. Применительно к системе АТКТ это означает, что она должна "уметь" определять тематику документа в соответствии с тремя классификационными схемами, которые будем называть "вариантами классификации" и для краткости обозначать:

1) *КС-ОНТ* − области науки и техники − список отделов (16 классов);
2) *КС-РЖ* − выпуски РЖ − список из 194 классов;

3) *КС-ГРНТИ* − тематические классы Рубрикатора; при этом на данном этапе решается упрощенная задача – грубое определение классов на 2-м уровне ГРНТИ[1], включающем около 800 индексов (из них тематике ВИНИТИ соответствуют лишь 460). Это обусловлено практической невозможностью обучить систему работе на всём 50-тысячном множестве классов Рубрикатора ВИНИТИ.

Конкретные классы будем называть "рубриками". Каждая рубрика имеет два атрибута − уникальный индекс и название. Примеры рубрик:

Рубрикатор	*Индекс*	*Название*
КС-ОНТ	e3	Биология
	f7	Химия
КС-РЖ	04Р1	Биотехнология. Бионанотехнологии. Бионаноматериалы
	04Т4	Токсикология
	04Т6	Фармакология
КС-ГРНТИ	62.09	Сырье и продуценты для биотехнологического производства
	62.13	Биотехнологические процессы и аппараты
	62.33	Клеточная инженерия
	62.35	Технологическая биоэнергетика
	62.37	Прикладная генетическая инженерия
	62.39	Инженерная энзимология
	62.41	Иммунобиотехнологические методы анализа

Этот пример, в частности, демонстрирует различную глубину определения тематики в рассматриваемых рубрикаторах.

Материалом для работы системы АТКТ являются русско- и англоязычные научные тексты, извлекаемые из метаданных научно-технической литературы, поступающей в ВИНИТИ во входном потоке. Существенными для классификации элементами данных являются оригинальное заглавие, авторская аннотация и ключевые слова. Результат работы АТКТ − это оценка вероятности соответствия заданного текста рубрикам в различных вариантах классификации.

Построение АТКТ основано на технологии машинного обучения, для чего используются коллекции документов из политематической базы данных ВИНИТИ. Обучающие выборки содержат результаты ручной обработки документов, выполненной профильными специалистами при формировании информационных продуктов. Для обучения используются заглавия, рефераты и ключевые слова документов, снабжённых кодами отделов, кодами выпусков РЖ, индексами Рубрикатора ГРНТИ.

В оперативном доступе имеется несколько миллионов документов, обработанных за последние годы, что вполне достаточно для обучения системы и исследований. Причем эти документы в наибольшей мере соответствуют решаемой задаче, так как позволяют обучать систему автоклассификации в условиях, максимально приближенных к области её планируемого применения. Немаловажно и то, что большой объем имеющихся материалов позволяет добиваться достаточной представительности рубрик в обучающих выборках, варьировать в широком диапазоне структуру выборок для обеспечения нужной равномерности представления рубрик.

Методы машинного обучения предполагают многократность их выполнения с последующим анализом результатов, внесением изменений в управляющие параметры системы и корректировкой исходной выборки. Мониторинг соответствия результатов ручного и автоматического индексирования является важным средством оценки качества АТКТ и выработки рекомендаций по её совершенствованию. Применительно к ВИНИТИ правильность автоматически получаемых рубрик (отдел, выпуск РЖ, индекс ГРНТИ) может быть проверена по результатам ручной обработки документов в процессе формирования Реферативного журнала.

ПОСТРОЕНИЕ АВТОКЛАССИФИКАТОРА. ОБУЧЕНИЕ И НАСТРОЙКА

Основными структурно-функциональными компонентами программного комплекса АТКТ являются текстовый предобработчик, векторайзер и классификатор.

Модуль предобработчика текста решает задачу очистки входного документа от элементов разметки, служебных символов, а также выполняет лемматизацию, т. е. приведение всех слов к начальной форме. В результате текст переводится в универсальное внутреннее представление, независимое от входного формата. Для учета специфики русско- и англоязычных текстов в АТКТ предусмотрены соответствующие средства.

Векторайзер преобразует текст в его векторную модель заданной размерности. Для этого в АТКТ используется технология *Word2Vec* [13], что позволяет представить текст как числовой вектор, сохранив при этом его семантику в виде наиболее весомых или усредненных тематических признаков слов, составляющих текст. Основной параметр *Word2Vec* − это размерность получаемого вектора; она определяет детальность выражения смысла в тематическом пространстве.

Классификатор на основе векторного представления текста принимает решение о том, к каким рубрикам относится текст, и оценивает его вероятность для каждой рубрики. Классификатор формирует тематический профиль текста − по всем вариантам классификационных схем, работе с которыми он обучен. В этом модуле реализована возможность применения нескольких наиболее известных в настоящее время алгоритмов машинного обучения и моделей принятия решений, которые используют методы обработки естественного языка на основе компьютерного анализа с применением средств искусственного интеллекта и компьютерной лингвистики.

[1] ГРНТИ – Государственный рубрикатор научно-технической информации России. Имеет три уровня иерархии. Рубрикатор ВИНИТИ фактически является его развитием до 9-го уровня.

Программное приложение АТКТ реализовано на платформонезависимом интерпретируемом языке программирования *Python* версии 3.6 в среде 64-разрядной ОС *Windows* версии не ниже 7. При разработке нами использованы следующие компоненты:

1) *Scikit-Learn* – модуль Python, который содержит модели машинного обучения и средства обработки данных;
2) *Gensim* – библиотека, реализующая модель *Word2Vec*;
3) *Pandas* – библиотека, предоставляющая контейнер, построенный по реляционному принципу и используемый для операций с данными;
4) *PyQt5* – интерфейс фреймворка *Qt* для *Python*; применяется для реализации событийно-ориентированной архитектуры и межпотокового взаимодействия, а также для создания пользовательского интерфейса;
5) *PyMystem3* – интерфейс библиотеки *Mystem* для *Python*; служит для лемматизации текста;
6) *Textblob lemmatizer* – модуль библиотеки обработки естественного языка *Textblob*, предназначенный для лемматизации;
7) *NLTK* – набор инструментов для обработки естественного языка; включает, в частности, стеммер Портера, стеммер Ланкастера, лемматизатор на основе базы лексики *WordNet*.

Работоспособность системы АТКТ определяется, в частности, правильным выбором настроек ее функционирования. В нашем случае определяющими среди них можно считать такие факторы:

- выбор алгоритмов предобработки текстов (очистки, лемматизации и пр. – с учётом особенностей естественных языков);
- параметры векторизации текста для модели *Word2Vec*;
- оценка эффективности различных моделей машинного обучения, доступных для классификации; оптимизация их настроек (также называемых гиперпараметрами) в соответствии с применяемыми классификаторами;
- формирование "правильных" обучающих выборок − достаточного объёма, сбалансированных по классам, достаточными размерами текстов и пр.

Точно оценить вклад каждого фактора в конечный результат не представляется возможным, но можно его оценить, т. е. измерить качество результатов автоматического индексирования. Решение таких задач обеспечивает специально построенная система обучения АТКТ, которая позволяет в автоматизированном режиме проводить сравнительные исследования алгоритмов, заложенных в АТКТ, с использованием различных обучающих выборок и различных параметров настройки.

Для оценки качества автоматической классификации используются изобретённые в прошлом веке для оценки результатов информационного поиска [14–16] традиционные меры:

- доля правильных ответов (*accuracy*);
- точность (*precision*)
- полнота (*recall*);
- *F*-мера (F_1-мера, *F-measure*, f_1 *score*).

Эти меры вычисляются по отдельности для каждой рубрики во всех вариантах классификации. Для обобщённой оценки качества индексирования используется макро-усреднение (*macro-averaging*) и микро-усреднение (*micro-averaging*) перечисленных мер. Оценка качества базируется на сравнении результатов автоматического и экспертного индексирования.

Прежде чем перейти к математическим формулировкам, необходимо подчеркнуть принципиальную разницу между множеством рубрик, присвоенных документу компьютерной программой (алгоритмом), и множеством рубрик, присвоенных документу экспертом (человеком). Алгоритм классифицирует согласно заложенной в него математической модели, поэтому выдаёт нечёткое множество упорядоченных пар (код рубрики, релевантность рубрики). По сути, релевантность является функцией принадлежности рубрики нечёткому множеству, так как указывает, в какой степени элемент принадлежит множеству. Как и функция принадлежности, релевантность принимает значения в числовом отрезке [0, 1].

В отличие от алгоритма, человек назначает документу обычное (чёткое) множество рубрик, в котором функция принадлежности каждого элемента равна единице. Теоретически экспертизу можно усложнить: попросить экспертов упорядочить рубрики по убыванию их соответствия тематике документа, т. е. ранжировать. Эту информацию можно было бы использовать для более точного сравнения результатов автоматического и экспертного индексирования. Однако, не имея возможности постфактум получить от экспертов ранги, считаем все рубрики эксперта равнозначными.

Таким образом, необходимо сравнивать четкие и нечёткие множества рубрик. Самый очевидный подход – избавиться от нечёткости, получив чёткое множество рубрик из нечёткого множества, присвоенного документу алгоритмом. С этой целью применяются два метода отбора рубрик чёткого множества – пороговый и ранговый.

Пороговый метод использует пороговое значение функции принадлежности, в нашем случае это пороговое значение релевантности. Пусть A нечёткое множество с функцией принадлежности $\mu_A(x)$=$relevant_A(x)$:

$$A = \left\{ \left(x, \mu_A(x)\right) \mid x \in X \right\}, \text{ где } \mu_A(x) \in [0,1],$$

тогда чёткое множество A_α, которое называется альфа-срезом (α-срезом) нечёткого множества A, определяется следующим образом:

$$A_\alpha = \left\{ x \in X \mid \mu_A(x) \geq \alpha \right\},$$

где α – пороговое значение функции принадлежности.

Ранговый метод использует пороговое значение ранга. Ранг элемента x в множестве B – $rank_B(x)$ – это порядковый номер элемента x в нечётком множест-

ве B, упорядоченном по убыванию функции принадлежности. Тогда, если B упорядоченное нечёткое множество:

$$B = \left\{ \left(x, \mu_B(x), rank_B(x) \right) \mid x \in X \right\},$$

где $rank_B(x) > 0$,

то чёткое множество B_r ранга r (ранговый r-срез нечёткого множества B) получаем следующим образом:

$$B_r = \left\{ x \in X \mid rank_B(x) \le r \right\},$$

где r – пороговое значение ранга.

Мы используем оба метода отбора элементов из нечёткого множества рубрик. Ранговый метод применяется для вычисления точности, полноты и F_1-меры; пороговый метод – для вычисления меры сходства двух альтернативных множеств рубрик, присвоенных одному документу, если одно из них нечёткое, а другое чёткое.

Приведем математические формулировки точности, полноты, F_1-меры и меры сходства.

Пусть D – множество документов, отнесённых алгоритмом к рубрике; E – множество документов, отнесённых экспертами к рубрике, а величины tp, fp, fn, tn определены как мощности конечных множеств:

$$tp = \left| \left\{ x \mid x \in D \,\&\, x \in E \right\} \right|$$
$$fp = \left| \left\{ x \mid x \in D \,\&\, x \notin E \right\} \right|$$
$$fn = \left| \left\{ x \mid x \notin D \,\&\, x \in E \right\} \right|$$
$$tn = \left| \left\{ x \mid x \notin D \,\&\, x \notin E \right\} \right|,$$

тогда доля правильных ответов (*accuracy*), точность (*precision*) и полнота (*recall*) вычисляются по формулам:

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$
$$precision = \frac{tp}{tp + fp}$$
$$recall = \frac{tp}{tp + fn}.$$

Обобщённой мерой качества индексирования будем считать F_1-меру (f_1 *score*), которая равна среднему гармоническому точности и полноты:

$$f_1score = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 \cdot precision \cdot recall}{precision + recall}.$$

Такую меру иногда называют сбалансированной F-мерой, так как значения *precision* и *recall* имеют одинаковый вес.

Описанные меры (*accuracy*, *precision*, *recall*, f_1*score*) рассчитываются по каждой рубрике и после завершения обучения выдаются в отчёт системой АТКТ, что позволяет сразу оценить его успешность.

Обычно обучение длится долго и запускается редко, поэтому дополнительно, для оперативного контроля качества индексирования между обучениями, нужны другие меры, способные вычисляться на любом подмножестве документов, т. е. не только по рубрикам. Например, можно оценивать качество индексирования по отраслям науки, по видам документов (статьи, книги, патенты и т.д.), по издательствам, по статьям из заданного списка научных журналов и т. п. С этой целью удобно измерить точность, полноту и F_1-меру для каждого документа в подмножестве, а затем использовать средние значения этих мер для оценки всего подмножества.

Введём понятие точности и полноты индексирования для документа с идентификатором *ID*. Пусть D_{ID} – множество рубрик, присвоенных алгоритмом документу *ID*, E_{ID} – множество рубрик, присвоенных экспертами документу *ID*. Тогда точность ($precision_{ID}$), полнота ($recall_{ID}$) и F_1-мера (f_1score_{ID}) для документа *ID* имеют следующий вид:

$$precision_{ID} = \frac{\left| D_{ID} \cap E_{ID} \right|}{\left| D_{ID} \right|}$$

$$recall_{ID} = \frac{\left| D_{ID} \cap E_{ID} \right|}{\left| E_{ID} \right|}$$

$$f_1score_{ID} = \frac{2 \cdot precision_{ID} \cdot recall_{ID}}{precision_{ID} + recall_{ID}}. \qquad (1)$$

Нетрудно убедиться, что формула (1) преобразуется следующим образом:

$$f_1score_{ID} = \frac{2 \cdot \left| D_{ID} \cap E_{ID} \right|}{\left| D_{ID} \right| + \left| E_{ID} \right|}. \qquad (2)$$

Эта бинарная мера сходства предложена датским биологом Торвальдом Сёренсеном (Thorvald Sørensen) в 1948 г. В русскоязычных источниках она обычно называется мерой или коэффициентом Сёренсена [17, 18].

На сегодняшний день нами проведены исследования семи алгоритмов машинного обучения и влияния их гиперпараметров на качество индексирования. Это логистическая регрессия (*logistic regression*), метод опорных векторов (*svm*), персептрон (*perceptron*), метод *k*-ближайших соседей (*knn*), дерево решений (*decision tree*), случайный лес (*random forest*), адаптивный бустинг (*AdaBoost*).

Эксперименты по выбору моделей и подбору значений гиперпараметров проводились на обучающих выборках небольшого объёма (менее 100 тыс. документов). Такой объём выборки позволяет обучить несколько моделей в течение 1-2 недель на сервере с 8 Гбайт ОЗУ и двумя 4-ядерными процессорами *Intel Xeon* с тактовой частотой 2,5 ГГц.

Скорость обучения зависит не только от объёма обучающей выборки, но и от выбранной размерности векторного пространства модели *Word2Vec*, а также от параметров настройки моделей классификации. При слишком больших значениях размерности или гиперпараметров процесс обучения выполняется слишком долго, поэтому его приходится останавливать, уменьшать значения и запускать заново. Именно по причине низкой скорости обучения пришлось отказаться от метода опорных векторов (*svm*), несмотря на его высокие показатели качества при малых размерностях *Word2Vec* (50-100). Логистическая регрессия и персептрон демонстрировали близкие к *svm* показатели качества индексирования наряду с высокой скоростью обучения. Поэтому было решено сосредоточиться на этих двух моделях.

По завершении обучения мы проводили тестирование полученной модели – выполнялось индексирование документов тестовой выборки.

В обучающие и тестовые выборки отбираются документы, опубликованные в тематических выпусках Реферативного журнала ВИНИТИ. Для обучения используются названия, рефераты и ключевые слова документов, снабженные кодами отделов, выпусков РЖ, индексами Рубрикатора ВИНИТИ и ГРНТИ. В среднем объем документа составляет 850 символов; число тематических кодов – от 1 до 3 для каждой классификационной схемы.

При формировании выборок соблюдались правила: обучающая и тестовая выборки не пересекаются; объём тестовой выборки не менее 1/3 объёма обучающей выборки; год издания документов обеих выборок относится примерно к одному хронологическому периоду длительностью 2-3 года (допускается сдвиг хронологического отрезка тестовой выборки вперёд на 1 год).

По окончании цикла обучения-тестирования формируются отчёты − по каждой модели и для каждого из трех вариантов классификации (ОНТ, РЖ, ГРНТИ). Отчеты включают меры *Accuracy*, *Precision*, *Recall*, f_1 *score*, исходные значения переменных *tp*, *fp*, *fn*, *tn*, средние значения мер с использованием макро- и микро-усреднения (*macro-average* и *micro-average*). Статистика выдаётся по пяти ранговым срезам: $r \in \{1, 2, 3, 4, n\}$, где n означает "все ответы классификатора". Для примера в табл. 1 показана статистика по двум ранговым срезам для варианта *КС-ОНТ + percepron*[200] – при размерности *Word2Vec,* равной 400; объёмы обучающей и тестовой выборок равны 90 тыс. и 45 тыс. документов, соответственно.

Из табл. 1 видно, как возрастает полнота (*recall*) и убывает точность (*precision*) при увеличении порогового значения ранга с 1-го до 2-х, а также, что мера *accuracy* бесполезна на тестовых выборках, в которых количество документов по рубрикам существенно различается. Обучающие и тестовые выборки должны быть репрезентативными, поэтому они, хотя и в сглаженном (сбалансированном) виде, должны отражать неравномерность частотного распределения научных публикаций по тематикам науки. Для *КС-ОНТ* мы считаем выборку сбалансированной, если количество документов в рубриках различается не более чем в 10 раз. На практике количество документов по рубрикам различается более значительно. Мы видим, что у двух рубрик, наиболее слабо представленных в потоке документов (это ″информатика" и "астрономия"), оказались самые высокие значения *accuracy,* что является завышенной оценкой качества классификации. На самом деле, согласно более объективной оценке f_1*score*, ″информатика" стоит только на пятом месте, а "астрономия" вообще имеет f_1*score* ниже среднего.

Таблица 1

Статистика тестирования модели персептрон[200] для *КС-ОНТ*

Отдел	Один ответ классификатора (r = 1)				Два ответа классификатора (r = 2)				*tp*+*fn*
	accuracy	*precision*	*recall*	f_1*score*	*accuracy*	*precision*	*recall*	f_1*score*	
Авт. и радиоэлектр.	0,963	0,756	0,794	0,774	0,881	0,399	0,941	0,560	4558
Астрономия	0,991	0,788	0,760	0,774	0,979	0,478	0,889	0,622	1097
Биология	0,989	0,970	0,947	0,958	0,925	0,638	0,985	0,774	7338
География	0,983	0,870	0,778	0,822	0,938	0,449	0,934	0,607	2883
Геология	0,978	0,775	0,844	0,808	0,931	0,438	0,936	0,597	3104
Информатика	0,995	0,820	0,886	0,851	0,979	0,418	0,946	0,580	883
Математика	0,992	0,915	0,817	0,863	0,968	0,504	0,910	0,649	1852
Машиностроение	0,968	0,789	0,753	0,771	0,905	0,422	0,918	0,579	4031
Металлургия	0,975	0,733	0,742	0,737	0,941	0,437	0,906	0,590	2653
Механика	0,984	0,724	0,662	0,692	0,958	0,373	0,834	0,516	1514
Охрана окр. среды	0,986	0,775	0,881	0,825	0,912	0,290	0,964	0,446	2080
Транспорт	0,978	0,743	0,871	0,802	0,942	0,471	0,949	0,630	2945
Физика	0,972	0,850	0,828	0,839	0,909	0,493	0,946	0,648	4992
Химия	0,954	0,827	0,883	0,854	0,820	0,458	0,966	0,622	8642
Экономика пром-сти	0,983	0,894	0,848	0,871	0,965	0,665	0,943	0,780	3766
Электротехника	0,969	0,846	0,715	0,775	0,927	0,501	0,877	0,638	4155
Среднее									
macro	0,979	0,817	0,813	0,813	0,930	0,465	0,928	0,615	3530,8
micro	0,979	0,830	0,830	0,830	0,930	0,469	0,939	0,626	-

Примечание: *percepron*[200] – это модель персептрон с 200-ми элементами в одном скрытом слое

Таблица 2

Влияние размерности Word2Vec на качество классификации

Модель классификатора	Классификационная схема	кол-во рубрик	max(*macro* f_1 *score*) / max(*micro* f_1 *score*)		
			***vector_dim*=50**	***vector_dim*=100**	***vector_dim*=400**
Perceptron[100]	*КС-ОНТ*	16	0,63 / 0,67	0,69 / 0,72	0,75 / 0,77
Logistic regression	*КС-ОНТ*	16	0,60 / 0,64	0,61 / 0,66	0,67 / 0,70
Perceptron[100]	*КС-РЖ*	196	0,31 / 0,39	0,55 / 0,55	0,56 / 0,58
Logistic regression	*КС-РЖ*	196	0,40 / 0,42	0,54 / 0,54	0,81 / 0,78
Perceptron[100]	*КС-ГРНТИ*	436	0,12 / 0,38	0,62 / 0,62	0,67 / 0,69
Logistic regression	*КС-ГРНТИ*	436	0,25 / 0,36	0,68 / 0,62	0,95 / 0,86

Примечание: В этой таблице используются оба метода усреднения *F*-меры (макро- и микро-).

В результате серии экспериментов мы наблюдали интересный эффект. Если количество рубрик классификационной схемы невелико, то увеличение размерности *Word2Vec* не оказывает существенного влияния на качество индексирования. Но если количество рубрик превышает размерность *Word2Vec*, то это существенно улучшает качество индексирования.

Приведём конкретный пример. Список отделов ВИНИТИ содержит только 16 позиций, поэтому размерность 50-100 оказывается вполне достаточной. Варианты классификации *КС-РЖ* и *КС ГРНТИ* содержат более 100 рубрик – 196 и 436 соответственно. Видимо поэтому для них размерность *Word2Vec*, равная 50, оказалась явно недостаточной. Табл. 2 демонстрирует, какое влияние на *F*-меру оказывает увеличение размерности векторного представления слов (*vector_dim*) на примере обучения двух моделей: персептрона с одним скрытым слоем (100 элементов), и логистической регрессии.

Из табл. 2 следует простое эмпирическое правило: чем больше классов в классификационной схеме, тем больше должна быть размерность модели *Word2Vec*. При переходе от размерности 50 к размерности 100 качество индексирования по выпускам РЖ и по ГРНТИ существенно повысилось у обеих моделей. При переходе от размерности 100 к размерности 400 качество индексирования существенно повысилось только у логистической регрессии. Возможно, персептрону недостаточно 100 ассоциативных элементов в скрытом слое, чтобы улучшить результат при увеличении количества входных элементов до 400.

Последующие эксперименты с *КС-ОНТ* показали, что увеличение количества ассоциативных элементов со 100 до 200 улучшает показатели max(*macro* f_1 *score*) и max(*micro* f_1 *score*) с 0,75 и 0,77 до 0,81 и 0,83, соответственно. С использованием *КС-РЖ* и *КС-ГРНТИ* подобные эксперименты пока не проводились. Настройка гиперпараметров нейронной сети для автоматической индексации текстов при изменении размерности векторного представления слов и количества классов – это отдельная тема, которая находится за рамками настоящей статьи.

В общем виде технология выбора модели индексирования и оптимизации значений гиперпараметров похожа на конкурс, который проходит в несколько туров.

В первом туре эксперименты проводятся с небольшими обучающими выборками (менее 100 тыс. документов) и небольшой размерностью *Word2Vec* (50-100 координат), так как на этом этапе важно иметь высокую скорость обучения для перебора большого количества моделей и комбинаций гиперпараметров.

В следующие туры проходят наиболее перспективные модели, с которыми проводятся эксперименты по увеличению объёма обучающей выборки и размерности *Word2Vec*, а также по дальнейшей настройке гиперпараметров. Такой подход базируется на предположении, что модели, показавшие относительно низкое качество индексации на маленьких выборках и размерностях векторов, покажут относительно низкое качество и при их увеличении. Не для всех моделей и комбинаций гиперпараметров это соотношение верно, но полная проверка этого предположения требует огромных вычислительных ресурсов. На наших выборках наиболее перспективными оказались логистическая регрессия и персептрон. От метода опорных векторов пришлось отказаться из-за низкой скорости обучения, остальные модели показали невысокое качество индексирования на небольших выборках и размерностях *Word2Vec*.

В итоге для каждого варианта индексирования выбиралась своя модель принятия решений с набором гиперпараметров, оптимизирующим целевую метрику. В нашем случае точность и полнота индексирования одинаково важны, поэтому максимизируется F_1-мера. Аналогичным образом можно максимизировать точность, полноту, несбалансированную F_β-меру – выбор целевой метрики зависит от приоритетов внедрения автоклассификатора в конкретную информационную систему.

Вначале опишем алгоритм выбора модели в общем виде, затем приведём конкретный пример. Обозначим $F(i, A_j)$ значение целевой функции, вычисленное на срезе A_j нечёткого множества рубрик, присвоенных документам при использовании *i*-й модели. Соберём все значения целевой функции в матрицу $(a_{i,j})$ размера $m \times n$, где $a_{i,j} = F(i, A_j)$, $i = \overline{1, m}$, $j = \overline{1, n}$, m – количество исследуемых моделей, n – количество ранговых *r*-срезов нечёткого множества рубрик. Оптимальной назовём модель, которой соответствует строка с номером s, если $a_{s,j} = \max\limits_{i,j} a_{i,j}$.

Таблица 3

Матрица выбора оптимальной модели для *КС-ОНТ*

Модель классификатора	*macro* f_1 *score* на трёх ранговых срезах		
	***r*=1**	***r*=2**	***r*=3**
Perceptron[100]	0,752	0,595	0,462
Perceptron[200]	**0,813**	0,615	0,471
Logistic regression	0,665	0,568	0,465

Таблица 4

Средний коэффициент Сёренсена при различных пороговых значениях *α*

Классификационная схема	α=0,0	α=0,1	**α=0,2**	α=0,3	α=0,4	α=0,5	α=0,6	α=0,7	α=0,8	α=0,9
КС-ОНТ	0,461	0,649	**0,651**	0,640	0,620	0,594	0,560	0,522	0,475	0,407
КС-РЖ	0,295	0,481	**0,505**	0,486	0,437	0,364	0,288	0,217	0,145	0,072
КС-ГРНТИ	0,150	0,410	**0,427**	0,415	0,383	0,331	0,276	0,223	0,164	0,097

Описанный алгоритм продемонстрируем на примере варианта классификации *КС-ОНТ*. Обучающая выборка 90 тыс. документов, размерность *Word2Vec* равна 400. Исследуются три модели на трёх ранговых срезах. В качестве целевой функции используется *macro* f_1 *score*. Полученные значения целевой метрики сведены в матрицу размера 3×3 (табл. 3), где видно, что максимальное значение целевой функции (0,813) дает модель *perceptron*[200]. Таким образом, именно эту модель следует использовать в варианте *КС-ОНТ*.

Для изучения сходства результатов автоматического и экспертного индексирования системой АТКТ мы проиндексировали более 1 млн документов, опубликованных в РЖ ВИНИТИ в 2017-2018 гг. При этом для *КС-РЖ* и *КС-ГРНТИ* применялась логистическая регрессия, для *КС-ОНТ* – персептрон. В качестве меры сходства был использован коэффициент Сёренсена, вычисленный для каждого документа при пороговых значениях релевантности от 0,0 до 0,9 с шагом 0,1.

Коэффициент Сёренсена для документа вычисляется по формуле (2) и совпадает с F_1-мерой, вычисляемой по формуле (1). Средние значения коэффициентов собраны в табл. 4, позволяющую выбрать оптимальное пороговое значение релевантности − в отдельности для каждого варианта. Для этого выбирается значение α, при котором средний коэффициент Сёренсена достигает максимального значения.

Согласно табл. 4, оптимальное пороговое значение релевантности равно 0,2, причём для всех трёх вариантов классификации.

ОПЫТ ПРАКТИЧЕСКОГО ПРИМЕНЕНИЯ АВТОКЛАССИФИКАТОРА В ВИНИТИ РАН

Система АТКТ работает в ВИНИТИ с начала 2019 г. Она установлена на специально выделенном сервере, который предоставляет внешним приложениям услугу: в ответ на предъявленный текст выдаёт взвешенный список тематических рубрик по любой из трех классификационных схем − отделы, выпуски РЖ, индексы ГРНТИ. Это позволяет использовать систему в производственном процессе и рассматривать перспективные направления её применения.

Реализуемую в ВИНИТИ систему АТКТ удобно представить в виде многоярусной схемы, при которой на каждом ярусе используются результаты работы предыдущего яруса:

- на первом ярусе входной поток документов (естественно, его часть в электронной форме) подвергается автоматической индексации по всем предусмотренным классификационным схемам; результаты складываются в таблицы Единой технологической базы данных (ЕТБД), доступные для всех приложений;
- на втором ярусе результаты автоиндексации могут использоваться как рекомендации при тематическом индексировании документов − на разных стадиях обработки;
- на третьем ярусе можно проводить сравнение результатов автоиндексации с результатами ручной индексирования, выполненного научными сотрудниками при формировании информационных продуктов ВИНИТИ;
- на четвертом ярусе можно определять недостатки автоиндексации и выдвигать предложения по необходимости переобучения системы АТКТ или по совершенствованию моделей и алгоритмов АТКТ.

Функции 1-го яруса выполняет программное приложение "Электронный эксперт", которое управляет процессом автоиндексации: формирует потоки данных для АТКТ, осуществляет запуск системы АТКТ, загрузку и обработку результатов ее работы в ЕТБД. Программа работает ежедневно по расписанию и обрабатывает документы входного потока, поступившие в ВИНИТИ за последние сутки. В результате формируется массив, в котором отображаются показатели вероятности отнесения документа к той или иной рубрике классификационной схемы.

Программа оптимизирована для обработки больших массивов документов. Оптимизация состоит в разделении входного потока документов для АТКТ на порции и реализации изолированной параллельной обработки порций. Производительность Электронного эксперта оценивалась на выборке порядка миллиона документов при различных значениях количества документов в порции и количества одновременно запущенных процессов АТКТ, на различных конфигурациях компьютеров. Объемы порции варьировались от 100 до 10000 документов, а количество потоков - от 1 до количества ядер компьютера. Эксперименты показали, что на 64-разрядной рабочей станции с частотой 2,5 ГГц с 4 ядрами и 8 Гб памяти среднее время классификации (равное 30 мс) достигается при использовании параллельно двух процессов АТКТ и размере порции 2000 документов. При сокращении размера порции до 1000 документов и увеличении до шести количества параллельных процессов АТКТ среднее время индексирования будет равно 10 мс. При такой скорости можно классифицировать 2,9–8,6 млн документов в сутки по одной классификационной схеме. Поскольку в ВИНИТИ каждый документ индексируется по трём схемам, производительность Электронного эксперта на практике составляет 1–3 млн документов в сутки, что значительно превышает потребности Института.

Первое, наиболее очевидное, применение результаты работы Электронного эксперта нашли в Отделении обработки входного потока научно-технической литературы, где осуществляется научная систематизация (тематическая разметка) документов для направления в отделы научной информации ВИНИТИ. Здесь результаты классификации в варианте *КС-ОНТ* используются в качестве "подсказок" для разметчика. Соответствующие функции внедрены в автоматизированное рабочее место для поддержки принятия решений. В дальнейшем предполагается продвинуться в сторону автоматической разметки, когда некоторые (пока ограниченные, специально выделенные) материалы входного потока можно направлять в тематические отделы, целиком полагаясь на результат автоиндексации, т. е. без участия оператора-разметчика. Автоматический режим можно предположить осуществимым, например, для отдельных научных журналов, литературы некоторых издательств и определённых тематик и т. п.

Другое направление использования результатов АТКТ связано с повышением уровня представления в продуктах ВИНИТИ документов входного потока. Как отмечалось выше, традиционные продукты Института охватывают не более 60% поступающих документов, а 40% документов не проходят никакой обработки и не отражаются в РЖ и БД. Между тем пользователям могут быть полезны просто библиографические сведения о новых публикациях. И здесь весьма уместно использовать результаты автоиндексации по ГРНТИ и предоставить пользователям услугу просмотра новых поступлений литературы по тематическим рубрикам – с существенно более широким охватом, чем в традиционных продуктах. Дополнительным бонусом для пользователей может быть и то, что автоиндексатор дает более широкий спектр тематических принадлежностей публикаций, чем это делаются при ручной обработке узкопрофильные специалисты (это особенно важно в связи с возрастанием доли многодисциплинарных публикаций, на что не успевают должным образом реагировать устоявшиеся технологии). Работа над такими продуктами ведётся, и мы рассчитываем в ближайшее время представить их на суд потребителей.

Третье направление применения АТКТ – это внедрение системы в практику работы научных сотрудников, осуществляющих аналитико-синтетическую переработку документов в отраслевых отделах Института. В настоящее время проводятся эксперименты по подключению данных Электронного эксперта в качестве рекомендаций при индексировании документов по Рубрикатору ВИНИТИ. Кроме того, представляется перспективным дать научным сотрудникам возможность поиска во входном потоке документов, которые в действительности соответствуют тематике отраслевых отделов или выпусков РЖ, но по каким-либо причинам (возможно ошибочно) не были направлены в отдел в результате ручной разметки.

Еще раз подчеркнем, что результаты работы Электронного эксперта носят консультационный характер; в конечном счёте, ответственность за правильность научной систематизации несет человек. Понятно, что уровень доверия к результатам автоиндексации зависит как от субъективных предпочтений оценивающего, так и от объективных факторов, определяемых степенью совершенства алгоритмов системы АТКТ.

Качество работы системы АТКТ подвергается непрерывному мониторингу посредством анализа результатов Электронного эксперта. С этой целью в технологический контур внедрён комплекс процедур на языке Transact SQL, которые в автоматическом режиме обеспечивают предоставление информации о работе Электронного эксперта: количество обработанных документов, оперативные оценки полноты и точности классификации, оповещения о выходе этих показателей за допустимые пределы. Для анализа используются результаты ручного индексирования документов, выполненного научными сотрудниками ВИНИТИ при формировании выпусков РЖ. По данным мониторинга принимаются решения о переобучении системы АТКТ.

В частности, анализ показывает явную неравномерность качества автоиндексации при обработке документов из различных областей науки и техники. Если для публикаций по биологии, экономике промышленности, математике, физике, химии система АТКТ дает хорошие чётко выраженные ответы, то для других областей результаты получаются более размытыми, с большим количеством альтернатив, без явных приоритетов. Это можно объяснить двумя причинами.

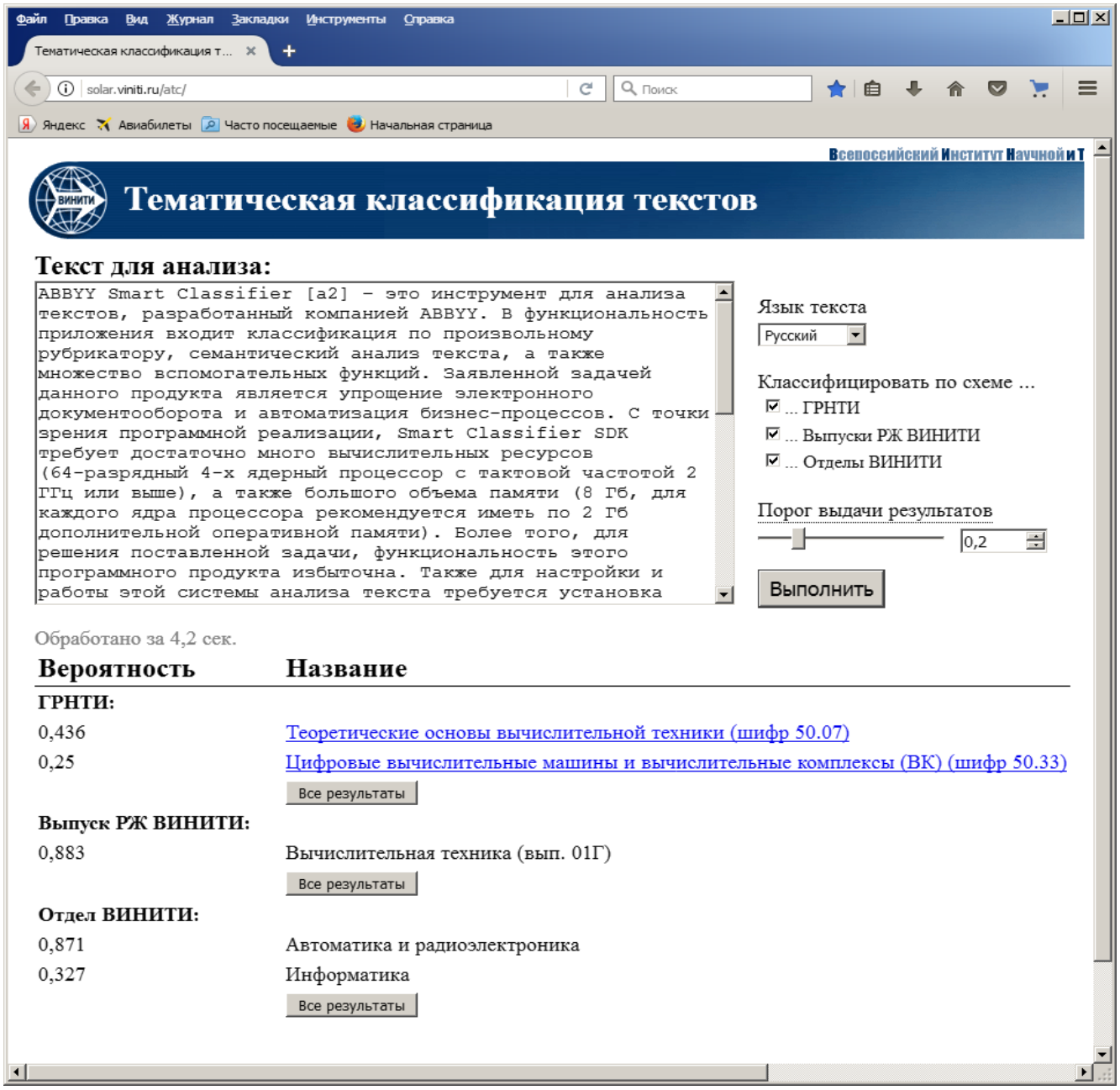
Во-первых, – свойствами текстов. Например, публикации по биологии или химии содержат существенно больше характерных терминов, чем, скажем, публикации по машиностроению или транспорту.

Во-вторых, – недостатками классификационных схем, связанными с условностями деления на отрасли

в политематических и междисциплинарных областях. Например, классификационные схемы некоторых выпусков РЖ по транспорту, машиностроению, автоматике и радиоэлектронике в значительной степени пересекаются.

Из этих наблюдений очевидно, что степень практического применения методов автоиндексации может существенно различаться в подразделениях ВИНИТИ, осуществляющих обработку документов. По одним тематическим направлениям можно в большой мере доверять результатам автоиндексации и активно использовать их при формировании традиционных информационных продуктов и даже для создания новых продуктов и услуг. По другим тематическим направлениям результаты автоиндексации менее достоверны, поэтому не могут конкурировать с ручным индексированием (хотя могут использоваться в качестве чернового материала). Здесь для повышения достоверности автоиндексации необходимы серьёзные усилия по совершенствованию моделей принятия решений, самих классификационных схем и технологий обучения.

Интересным средством демонстрации и ручного контроля возможностей системы АТКТ является *web*-услуга "Тематическая классификация текстов"[2] Она позволяет сотрудникам Института самостоятельно оценивать работу автоклассификатора. Пример показан на рисунке.



Интерфейс Тематического классификатора текстов

[2] В создании этого продукта принимают участие Б.В. Крутиков − ведущий программист Управления информационных систем, и А.В. Ефимов − младший научный сотрудник Отделения обработки входного потока ВИНИТИ РАН.

Введя любой текст, пользователь может получить результаты его классификации − по отраслям науки и техники (отделы), по номенклатуре выпусков Реферативного журнала ВИНИТИ, по рубрикам ГРНТИ 2-го уровня. В качестве результатов анализа текста выдаются рубрики заданных классификационных схем со значениями их веса.

Пока эта услуга доступна только в Интранете ВИНИТИ. В будущем, при положительной оценке результатов тестирования и наличии достаточной вычислительной мощности, возможен доступ к этому сервису внешних пользователей через Интернет.

ВЫВОДЫ

В заключение отметим, что в ближайшем будущем вряд ли можно планировать отказ от ручного индексирования научных публикаций силами квалифицированных научных сотрудников, и автоматическое индексирование не заменит полностью интеллектуальную работу специалистов. Однако применение методов автоиндексации на определенных стадиях обработки документов представляется весьма перспективным – если не сказать неизбежным – в условиях нарастающей лавины документального потока и существующего запроса пользователей в оперативном информировании о поступлении документов по их тематическим профилям. Задача в том, чтобы при формировании информационных услуг находить адекватные механизмы сосуществования и взаимного дополнения людей и программ-роботов, выполняющих интеллектуальный анализ текстов.

Авторы настоящей статьи, не являясь профессиональными разработчиками собственно средств искусственного интеллекта, рассчитывают на совершенствование своей системы за счет применения новых инструментов обработки текстов, которые постоянно появляются в мире компьютерных технологий и становятся доступными для использования.

При любых подходах технология машинного обучения не может достичь высоких результатов без хороших целенаправленно сформированных обучающих выборок (подобно тому, как ученики не могут добиться успеха без учителей по своим предметам). Участие человека в обработке текстов необходимо, так как именно результаты ручной индексирования позволяют обучать и регулярно переобучать программу-робот. ВИНИТИ РАН обладает поистине огромным запасом "хорошо индексированных" документов, и этот запас постоянно пополняется и может использоваться для совершенствования средств автоиндексации.

СПИСОК ЛИТЕРАТУРЫ

1. Паттерсон Дж., Гибсон А. Глубокое обучение с точки зрения практика / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2018. – 418 с.
2. Aghaebrahimian A., Cieliebak M. Hyperparameter Tuning for Deep Learning in Natural Language. – URL: CEUR-WS.org/Vol-2458/paper5.pdf (дата обращения: 05.05.2020)
3. A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval [Основа для оценки автоматической индексации или классификации в контексте поиска] // Journal of the association for information science and technology – 2015. DOI: 10.1002/asi
4. Altınel B., Can Ganiz M. Semantic text classification: A survey of past and recent advances // Information Processing & Management. – 2018 November. – Vol. 54(6). – P. 1129-1153. DOI: 10.1016/j.ipm.2018.08.001
5. Golub K., Hagelbäck J., Ardö A. Automatic Classification Using DDC on the Swedish Union Catalogue. – URL: https://www.semanticscholar.org/, (дата обращения: 05.05.2020).
6. Arash Joorabchi, Abdulhussain E. Mahdi. Classification of scientific publications according to library controlled vocabularies: A new concept matching-based approach. – URL: https://www.semanticscholar.org/ (дата обращения: 05.05.2020). DOI: 10.1108/LHT-03-2013-0030
7. Some Thoughts on Preserving Functions of Library Catalogs [Размышления о сохранении функций библиотечных каталогов в сетевых средах] // Bulletin of the Association for Information Science and Technology – October/November 2016 – Vol. 43, Number 1
8. 12 years on - Is the NLM medical text indexer still useful and relevant? // Journal Biomed Semantics. – 2017 Feb 23/ – Vol. 8(1). – P. 8. – URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5324252/. DOI: 10.1186/s13326-017-0113-5.,
9. Создание базы данных ресурсов AgNIC с использованием полуавтоматической индексации материала // Journal of Agricultural & Food Information. – 2014. – № 15. – P. 159–179.
10. Romanov A., Lomotin K., Kozlova E. Application of Natural Language Processing Algorithms to the Task of Automatic Classification of Russian Scientific Texts // Data Science Journal. – 2019. – Vol. 18. – №. 1.
11. Козлова Е.С., Ломотин К.Е., Романов А.Ю. Применение алгоритмов обработки естественного языка: инструмент для автоматической классификации текста // Цифровые Трансформации и Глобальное Общество (DTGS-2018): материалы международной конференции. – Хам, Швейцария: Springer, 2018. – С. 310-323. DOI: 10.1007/978-3-030-02846-6_25.
12. Ломотин К.Е., Козлова Е.С., Романов А.Ю. Сравнительный анализ методов автоматической классификации для генерации кода УДК для научных статей // Информационные Инновационные Технологии: материалы международной научно-практической конференции. – М.: Ассоциация выпускников и сотрудников ВВИА имени профессора Н.Е. Жуковского, 2017. – С. 359-363.

13. Mikolov T. et al. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. – 2013. – P. 3111-3119.
14. Информационный поиск // Википедия. [2018—2018]. Дата обновления: 29.06.2018. – URL: https://ru.wikipedia.org/?oldid=93657750 (дата обращения: 26.07.2019).
15. Precision and recall // Wikipedia, The Free Encyclopedia. – URL: https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=900147451 (дата обращения: 26.07.2019).
16. Documentation of scikit-learn 0.21.2. User Guide. 3.3. Model evaluation: quantifying the quality of predictions. – URL: https://scikit-learn.org/stable/modules/model_evaluation.html (дата обращения: 26.07.2019).
17. Коэффициент сходства // Википедия. [2019—2019]. Дата обновления: 08.01.2019. –URL: https://ru.wikipedia.org/?oldid=97352771 (дата обращения: 26.07.2019).
18. Коэффициент Сёренсена // Википедия. [2017—2017]. Дата обновления: 22.03.2017. –URL: https://ru.wikipedia.org/?oldid=84432232 (дата обращения: 26.07.2019).

Материал поступил в редакцию 27.04.20.

Сведения об авторах

ЕГОРОВ Виктор Серафимович – заместитель начальника Управления информационных систем ВИНИТИ РАН
e-mail: vs-egorov@viniti.ru

КОЗЛОВА Екатерина Сергеевна – программист Отдела программных систем, ВИНИТИ РАН
e-mail: hse.kozlovaes@gmail.com

ЛОМОТИН Константин Евгеньевич – программист Отдела программных систем ВИНИТИ РАН,
e-mail: ke.lomotin@gmail.com

ФЕДОРЕЦ Олег Владимирович – кандидат технических наук, начальник Отдела программных систем ВИНИТИ РАН
e-mail: ovf@viniti.ru

ФИЛИМОНОВ Алексей Викторович – главный специалист Отдела программных систем ВИНИТИ РАН
e-mail: afil@viniti.ru

ШАПКИН Александр Владимирович – кандидат технических наук начальник Управления информационных систем ВИНИТИ РАН
e-mail: ss@viniti.ru

УДК 004.832.2

В.Н. Бетин, С.Э. Лукьянов, А.П. Супрун

Механизм поиска решения в формализме функциональных нейронных сетей

Описывается развитие формализма сетей функциональных нейронов (ФН-сетей), разработанного для создания спектра информационных систем, реализующих интеллектуальную компьютерную обработку разнородных данных от различных информационных источников в автоматизированных системах поддержки принятия решений. Рассмотрены проблемы, влияющие на скорость поиска решения задач. Для ее повышения рекомендовано использовать архив ранее найденных решений. Предложен метод обобщения решений, позволяющий уменьшить объем архива и расширить применимость найденных обобщений к более широкому классу возможных задач.

Ключевые слова: *искусственный интеллект, системы поддержки принятия решений, ФН-сеть, автоматический поиск решений, автоматический синтез алгоритмов, системы понимания текста, смысловая обработка информации, информационно аналитические системы, машины логического вывода, обобщение знаний*

DOI: 10.36535/0548-0027-2020-05-2

ВВЕДЕНИЕ

При создании информационных систем могут применяться два подхода: первый – разработка специализированных систем [1–4]; второй – создание адаптивных систем, синхронизирующих свои собственные изменения с изменениями в окружающем мире таким образом, чтобы в результате система приобретала желаемую функциональность (наделение системы интеллектом). Если ранее доминировал первый подход, то в современных условиях рост издержек, связанных с сопровождением специализированных систем, приводит к активному внедрению информационных систем с элементами искусственного интеллекта. В первую очередь это относится к системам автоматического перевода, реферирования, поиска информации и различным системам поддержки принятия решений, включая системы защиты информации [5–7]. Общим в них является наличие наращиваемой базы знаний и использование машинного вывода для подготовки выходных данных. Традиционно подобные системы строились в формализме семантических сетей [8], для которого разработан ряд программных инструментов, позволяющих уже сейчас создавать практические приложения [9,10], однако современные информационные системы должны уметь осуществлять интеллектуальную обработку сырых данных, что предполагает автоматическое обобщение и слияние новых знаний со знаниями, добытыми ранее [11–13]. К сожалению, реализация этих механизмов в формализме семантических сетей неудобна [10]. Эффективность агрегирования знаний определяется структурами для их представления и алгоритмами разрешения противоречий в сырых данных, что делает важным поиск иных формализмов. Один из них – это функциональные нейронные сети (ФН-сети) [14]. Этот формализм получен в поисках ответа на вопрос: какие структуры данных и какие математические операции в отношении этих структур необходимы, чтобы обеспечить функциональность, присущую интеллекту, независимо от физического способа его реализации. Настоящая работа опирается на результаты, изложенные в [11–17], и посвящена повышению эффективности механизма поиска решения в формализме ФН-сетей путем использования информации о ранее решенных задачах.

ТЕРМИНОЛОГИЯ И ОБСУЖДЕНИЕ ПРОБЛЕМ

Формализм предназначен для создания и обобщения функциональных моделей любой предметной деятельности в произвольных, плохо формализуемых областях. Предметная деятельность предполагает выполнение субъектами операций так, чтобы достичь нужных целей. Операция – это любая деятельность субъекта, меняющая свойства объектов. Объекты, необходимые для выполнения операции, образуют ее вход, а объекты, полученные в результате выполнения – ее выход. Задачей, возникшей перед субъектом,

будем называть ситуацию, когда субъекту известна цель в виде совокупности объектов, которые надо получить, и он располагает некоторыми объектами (ресурсами), которые может использовать, но не знает как это сделать. Решить задачу – значит найти ранее неизвестную схему выполнения операций, организующую операции в пространстве и во времени так, чтобы преобразовать ресурсы в цель. В зависимости от предметной области, схему выполнения операций можно интерпретировать как программу, алгоритм или технологический процесс [14].

Деление сущностей на объекты, субъекты, операции условно и зависит от взгляда аналитика на проблему. Например, субъект токарь, выполняющий технологическую операцию, является объектом воздействия для руководства цеха, а технологическая операция объектом проектирования для технолога. Поэтому для формального представления объектов, субъектов и операций в формализме используется единая и универсальная информационная структура – функциональный нейрон (ФН), – содержащая уникальное название класса, к которому относиться описываемая физическая сущность (характеристику в виде списка параметров, различающую сущности из разных классов) и, в случае если сущность обозначает операцию, – списки классов входных и выходных объектов [15]. Для целей настоящей статьи функциональный нейрон удобно интерпретировать, как многополюсник ("белый ящик" с известной структурой), имеющий N $(0 \le N < \infty)$ входов и M $(0 \le M < \infty)$ выходов. Если ФН не имеет входов и выходов ($N = 0$ и $M = 0$), то такие ФН служат для обозначения субъектов, выполняющих операции, или объектов, которые служат предметом операций. Функциональный нейрон может иметь сложную внутреннюю структуру, моделируемую системой других взаимодействующих ФН.

Функциональный нейрон имитирует выполнение произвольных операций во времени путем изменения статусов объектов на входе и выходе. Статус S – это маркер объекта, изменяющийся во времени в процессе поиска решения задачи. С его помощью субъект, решающий задачу, различает ресурсы и цели. Он имеет значение «есть» на интервале времени t^E, если субъект знает, что объект является ресурсом или станет доступен на интервале t^E после выполнения некоторой промежуточной операции. Статус «нужно» на интервале t^H характеризует объекты, про которые субъект знает, что они является конечной или промежуточной целью. Неопределенное значение статуса $S = "?"$ означает, что на интервале времени $t^?$ у субъекта нет информации о достижимости или принадлежности объекта к ресурсам или целям. Наконец, статус «решение» на интервале t^P имеют объекты, которые одновременно доступны и необходимы для достижения конечной цели.

Каждая сущность предметной области, моделируемая функциональным нейроном, имеет уникальное имя и относится к некоторому классу, описанному в базе знаний. Класс имеет уникальное имя в этой базе и ассоциирует общие свойства группы сущностей. К таким свойствам относятся: общее устройство, набор параметров и связанные операции [14]. На множестве классов введено отношение наследования, что позволяет уменьшить объем базы знаний. Будем говорить, что класс F является производным (наследником) класса R (обозначая это в виде $F \leftarrow R$), если будут выполняться следующие условия:

- объекты/субъекты класса F содержат все параметры сущности класса R;
- субъекты класса F могут выполнять все операции субъектов из класса R;
- если объект/субъект класса R служит параметром для некоторого класса A, то в качестве значения этого параметра может использоваться и объект/субъект класса F;
- если некоторая операция применима к объекту класса R, то она также применима и к объектам класса F;
- если объект класса F может быть получен в результате выполнения некоторой операции, то и объект класса R может быть получен с помощью той же операции.

ФН-сеть – это множество функциональных нейронов (ФН), в котором ФН, играющие роль операций, связаны входами и выходами с ФН, выступающими в роли объектов, образуя граф функциональных связей. Причем с каждым входом или выходом операции в одну и ту же единицу времени может быть связан только один объект. Объект может быть связан одновременно с произвольным числом операций. На рис. 1 приведен пример ФН-сети, описывающей постановку задачи поиска алгоритма решения алгебраического уравнения $a \times x + b \times x + c = d$, $(a + b) \neq 0$ с неопределенными, но известными коэффициентами a, b, c, d относительно неизвестного x [16]. Соответственно буквой F обозначены имена классов, а буквой f – имена конкретных сущностей.

На рис. 2 приведен функциональный нейрон «Пожарить картофель» и его декомпозиция на множество взаимодействующих функциональных нейронов в виде схематичной ФН-сети [17], в которой опущены имена классов и состояния.

На рис. 3 приведена схема ФН-сети из базы знаний, определяющая понятие «дистрибутивность» [16].

Реальные операции требуют времени для своего выполнения (задержка), а связи объектов меняются во времени. Однако исходная задача, при ее расширении метаданными о временных интервалах, задержках и правилах их вычисления, может быть сведена к задаче с идеальными операциями без задержек, в которой однажды установленная связь между операцией и объектами будет существовать всегда [17]. Таким идеальным операциям соответствуют логические связки и математические выражения, используемые в рассуждениях и при выводе формул. Поэтому далее будем рассматривать только такие идеальные операции и связи.

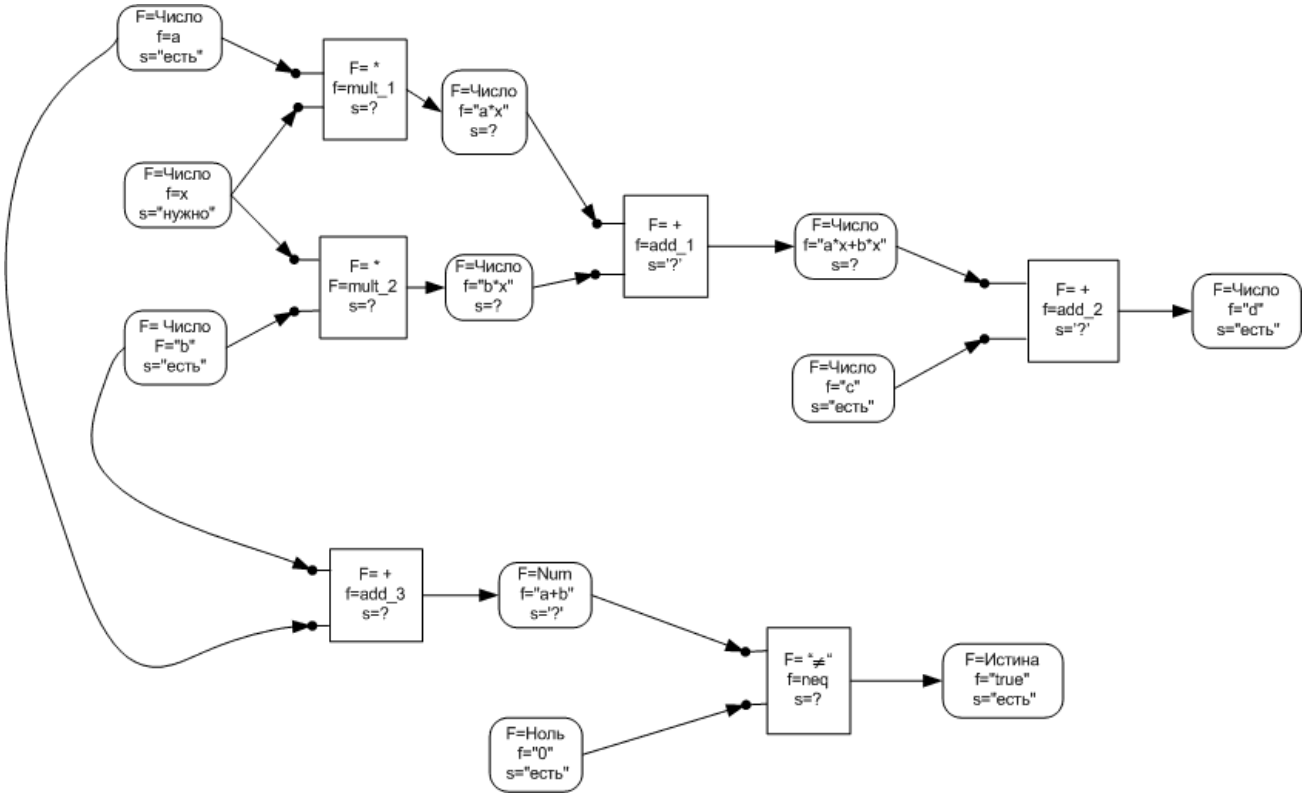


Рис. 1. Исходная конфигурация ФН-сети, соответствующая задаче поиска алгоритма решения уравнения $a \times x + b \times x + c = d, \quad (a+b) \neq 0$.

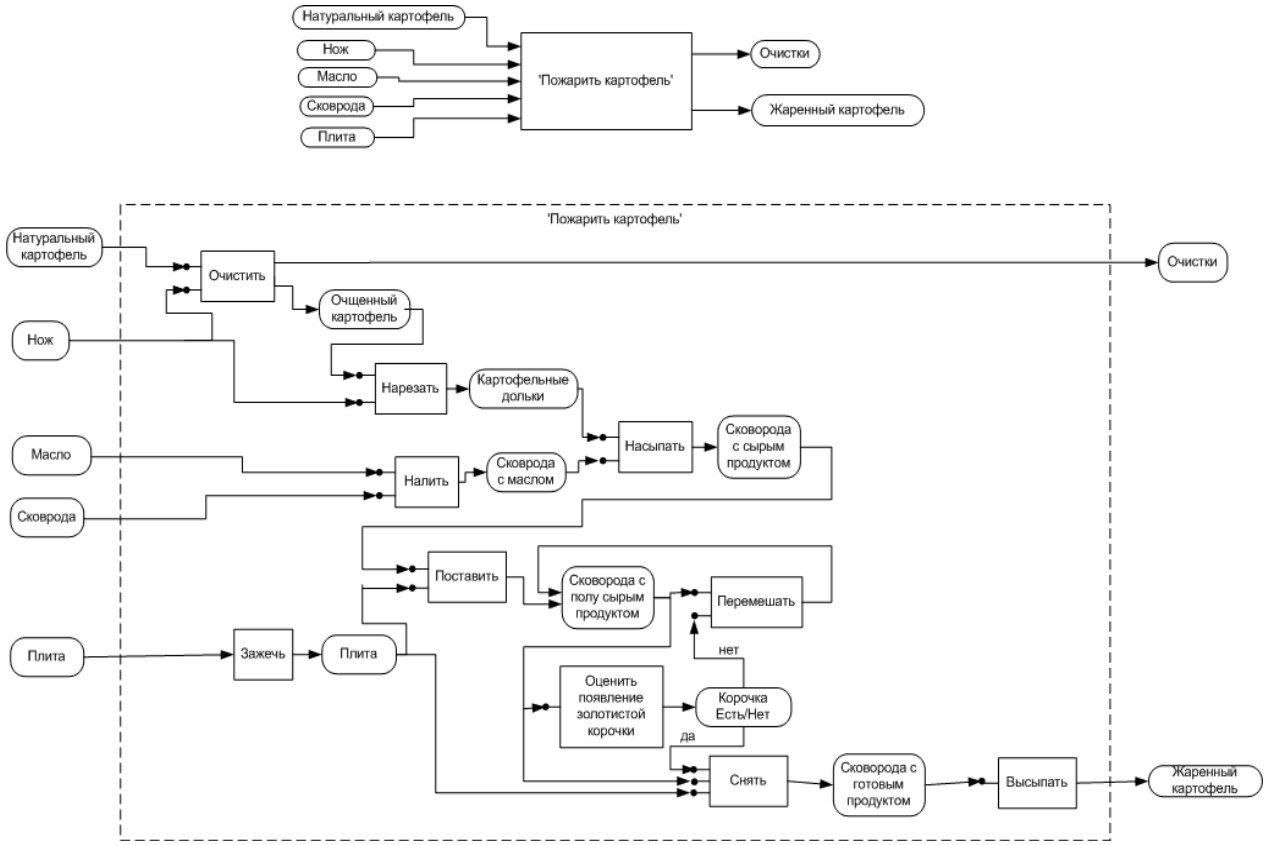


Рис. 2. Декомпозиция ФН «Пожарить картофель».

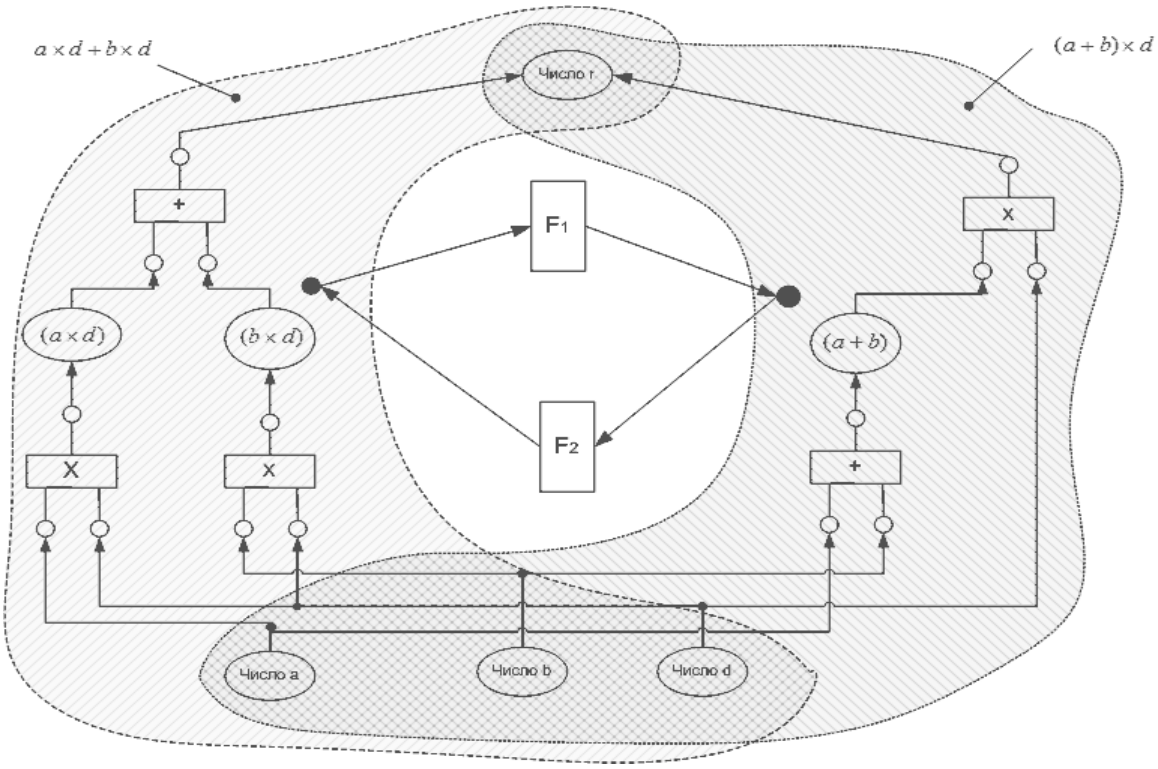


Рис. 3 Фрагмент сцены «Поле чисел» в БЗ, определяющий «дистрибутивность» в виде взаимообратных подстановок $F_1 : a \times d + b \times d \rightarrow (a + b) \times d$ и $F_2 : (a + b) \times d \rightarrow a \times d + b \times d$.

Пусть все объекты на входе некоторого функционального нейрона, имитирующего операцию *f*, имеют статус *«есть»*. Тогда операция *f* потенциально выполнима. Выполнив ее, субъект получит в свое распоряжение объекты на ее выходе. Следовательно, эти объекты из неопределенного статуса можно перевести в статус *«есть»*, а из статуса *«нужно»* в статус «*решение*». Аналогично, если объект в статусе *«нужно»* имеет связь с выходом операции *f*, то все объекты на ее входе с неопределенным статусом «?» представляют промежуточные цели и им следует присвоить статус *«нужно»*. Пусть ФН-сеть Z_0 описывает исходную задачу. Пусть X_0 множество объектов Z_0 в статусе «есть» (ресурсы) и Y_0 –в статусе «нужно» (цели). Остальные объекты в Z_0 имеют неопределенный статус *«?»*. Пересчитаем статусы объектов в Z_0 для целей Y_0. Будем это повторять для новых объектов, переходящих в статус "нужно" пока статусы в Z_0 не перестанут меняться. Затем выполним пересчет статусов "есть" в Z_0 для ресурсов X_0. Будем повторять его пока все цели в Y_0 не изменят значение статуса со значения «нужно» на значение «решение», или в Z_0 не исчерпаются возможности для изменения статусов. В первом случае Z_0 содержит искомую схему выполнения операций, и она может быть выделена путем отбора функциональных нейронов, связывающих объекты в состоянии «решение». Во втором случае Z_0 пополняется сценами из базы знаний. Сцена – это именованная ФН-сеть в базе знаний, определяющая классы объектов, субъектов и операций [15]. Для пополнения выбирается сцена с максимальным количеством элементов, сходных по связям и типу (с учетом наследования [11, 15]) элементам Z_0, изменившим статус при пересчете состояний. После этого в расширенной ФН-сети Z_1 с новыми элементами и новыми связями осуществляется очередной перерасчет статусов. Изменение статусов инициирует новое расширение Z_2 и так далее, пока не будет получено решение, либо процесс не будет остановлен по каким-то другим причинам.

Процесс расширения ФН-сети включает в себя анализ изоморфизма между графами связей в сцене из базы знаний, отобранной для расширения, и текущей конфигурацией ФН-сети, а это затратная по времени процедура. Уменьшить число сравнений сцен можно:

- при помощи сокращения количества сцен в базе знаний путем замены множества частных понятий их обобщением (этот механизм подробно рассмотрен в [11]);
- при использовании механизма отложенного расширения, контролирующего рост ФН-сети при добавлении подстановок из базы [15]. Идея этого метода заключена в выделении повторяющихся изоморфных фрагментов, к которым применим одинаковый набор подстановок, применении подстановки к единственному фрагменту и переносе результатов выполнения подстановки на все эквивалентные фрагменты без выполнения подстановок к этим фрагментам.
- путем рационализации выбора подстановок для расширения путем анализа сходства решаемой задачи с задачами, решенными ранее. Этот путь рассматривается в настоящей статье.

РЕШЕНИЕ ПОСТАВЛЕННОЙ ЗАДАЧИ

Рассмотрим различные варианты использования знаний о ранее решенных задачах. Пусть $\{z\} = \{< X_0, Y_0, Z_0, Z_P >\}$ – множество решенных задач, где Z_P – ФН-сеть, полученная на завершающем шаге поиска решения (контекст решения), содержащая искомую схему выполнения операций $Z_{CBO} \subset Z_P$. Эта схема состоит из элементов в состоянии "решение" и операций, которые их связывают. Контекст решения Z_P помимо схемы выполнения операций также содержит множество расширяющих подстановок $\{T\}$ с добавленными сценами из базы знаний, фрагменты которых могут и не входить в решение.

Пусть $z' = \{< X_0', Y_0', Z_0' >\}$ – новая задача и ФН-сеть Z_0', на которой она определена, имеет непустое пересечение $\Delta Z = Z_0' \cap Z_{CBO}$ по сходству с решением старой задачи (напомним, что пересечение по сходству предполагает изоморфизм функциональных подграфов и эквивалентность объектов в узлах по типам с учетом наследования типов). Заметим, что сравнение новой задачи Z_0' с контекстом ранее решенной задачи Z_P на данном шаге анализа нецелесообразно, так как Z_P может содержать много лишнего, не имеющего отношения к задаче Z_0'. Пусть ΔZ содержит: часть исходных данных $\Delta X = X_0' \cap \Delta Z_{CBO}$; фрагмент того что требуется $\Delta Y = Y_0' \cap \Delta Z_{CBO}$; фрагмент функционального графа $\Delta Z_{CBO} \subset Z_{CBO}$, связывающий ΔX и ΔY. Возможны варианты:

- одновременно все эти фрагменты не пусты ($\Delta X \neq \varnothing$, $\Delta Y \neq \varnothing$ и $Z_{CBO} \neq \varnothing$). Это значит, что в рамках новой задачи удалось выделить подзадачу, которую известно как решать, так как контекст решения Z_P содержит все подстановки, использованные для расширения исходной ФН-сети Z_0, и они могут быть легко выделены.
- $\Delta Y \neq \varnothing$, но там либо нет данных ($\Delta X = \varnothing$), либо в $\Delta Z_{CBO} \subset Z_{CBO}$ отсутствуют операции, соединяющие исходные данные ΔX и требуемый результат ΔY. В таком случае имеет смысл сравнить Z_0' с Z_P. Если для $\Delta Z = Z_0' \cap Z_p$ не удается выделить подзадачу, как в ранее рассмотренном случае, то можно попытаться переформулировать проблему, достраивая подстановками из базы знаний фрагмент ФН – сети $Z_p \setminus Z_0'$.

Из приведенного анализа следует, что старый опыт решения позволяет рационализировать перебор вариантов подстановок из базы знаний, однако по мере роста «архива» решений будет происходить и рост числа сравнений, который можно ограничить путем обобщения найденных решений. Для формирования "архива" обобщений рассмотрим множество всех пар решенных задач. Пусть в паре решений Z_{SBO}^1 и Z_{SBO}^2 нашлись фрагменты ФН-сетей $\Delta Z_{SBO}^1 \subset Z_{SBO}^1$ и $\Delta Z_{SBO}^2 \subset Z_{SBO}^2$, функциональные графы которых изоморфны и могут различаться лишь типами объектов в узлах. Эти фрагменты определяют сходные подзадачи. Выделим элементы в списках целей Y_0^1 и Y_0^2, для получения которых используются объекты из ΔZ_{SBO}^1 и ΔZ_{SBO}^2. Это можно сделать, двигаясь из объектов в ΔZ_{SBO}^1 и ΔZ_{SBO}^2 по графу связей вход-выход операций в ΔZ_{SBO}^1 и ΔZ_{SBO}^2, до достижения объектов в Y_0^1 и Y_0^2. Пусть $\hat{Y}_0^1 \subset Y_0^1$ и $\hat{Y}_0^2 \subset Y_0^2$ подмножества таких объектов. Теперь, двигаясь от целей из $\hat{Y}_0^1$ и $\hat{Y}_0^2$ в обратном направлении, выделим для них ближайшие по графу связей подмножества объектов в ΔZ_{SBO}^1 и ΔZ_{SBO}^2, которые обозначим E_1 и E_2.

Рассмотрим пару изоморфных элементов $e_1 \in E_1$ и $e_2 \in E_2$, которые имеют в общем случае разные типы G_1 и G_2, унаследованные от некоторых предков. Пусть они являются входными для операций $f_1^1, f_2^1, .. f_j^1$ из Z_{SBO}^1 и $f_1^2, f_2^2, .. f_j^2$ из Z_{SBO}^2 типов $F_1, F_2, ... F_j$ (j – их количество). Напомним, что чем более общий (универсальный) тип (выше в иерархии наследования), тем более бедным набором свойств он обладает, т.е. к более общему объекту применим меньший набор типов операций. Пусть типы G_1 и G_2 имеют общих предков из списка $\{R\}$, к которому одновременно применимы операции типа $F_1, F_2, ... F_j$. Тогда при формировании единого универсального представления для ΔZ_{SBO}^1 и ΔZ_{SBO}^2 целесообразно заменить в нем объекты e_1 и e_2 типов G_1 и G_2 объектом e^* с общим типом из списка $\{R\}$ максимально высокого уровня в дереве классификации типов, к которому применим набор операций типов $F_1, F_2, ... F_j$. Аналогично поступим с другими объектами в ΔZ_{SBO}^1 и ΔZ_{SBO}^2, которые являются входными для операций, где e_1 и e_2 аналогично служат выходом. Будем продолжать этот процесс до исчерпания ΔZ_{SBO}^1 и ΔZ_{SBO}^2. В случае успеха вместо двух частных решений будет получено одно типовое решение ΔZ_{SBO}^*, в котором частные объекты заменены объектами более общих типов, применимое к большему числу вариантов возможных задач, сходных по структуре и отличающихся типами объектов.

Очевидно, что предложенный процесс может быть многократно повторен уже для пар ранее полученных обобщений с целью выделения более мелких подзадач. Этот процесс следует продолжать до тех пор, пока будут находиться общие подзадачи. Конечным результатом будет иерархия вложенных обобщенных типовых задач, для каждой из которых известен набор подстановок из базы знаний, полезных для поиска решений. Полученная иерархия обобщенных решений позволяет оптимизировать порядок их рассмотрения, при котором первыми рассматриваются самые мелкие ранее решенные подзадачи, с последующим переходом к более крупным подзадачам, согласно с полученной иерархией.

Это делает возможным использовать на каждом шаге сравнения результаты предыдущего сравнения и существенно сократить время анализа.

ВЫВОД

Предложенный метод обобщения решений в архиве ранее найденных решений, строящий вложенную иерархию обобщенных подзадач, сокращает объем архива типовых решений, время их анализа, и повышает эффективность автоматического поиска решения.

СПИСОК ЛИТЕРАТУРЫ

1. Пошатаев О.Н., Съедин Д.Ю. Информационная система ЕГИСУ НИОКТР, как инструмент мониторинга и анализа работ в научно-технической сфере // Информатизация и связь. – 2016. – №4. – С. 46-52.
2. Съедин Д.Ю. Разработка и реализация алгоритма связывания данных в государственной информационной системе гражданского назначения // Научно-техническая информация. Сер. 2. – 2018. – №7. – С. 32-39; S'edin D.Yu. The development and implementation of the data-binding algorithm in the state civil information system // Automatic Documentation and Mathematical Linguistics. – 2018. – Vol. 52, № 4. – P. 195-202.
3. Романенко Г.С., Толчков А.Н., Чумичкин А.А. Моделирование информационных систем // Информатизация и связь. – 2020. – № 1. – С. 112-117.
4. Ковтун И.И., Романенко Г.С. Математическое моделирование, численные методы и комплексы программ в задачах повышения эффективности многономенклатурных производств // Информатизация и связь. – 2020. – № 1. – С. 27-33.
5. Силаев Ю.В., Тхорь В.А. К вопросу о создании интеллектуальной системы мониторинга и обеспечения информационной безопасности для использования в автоматизированных системах // Информатизация и связь. – 2017. – № 1. – С. 119-121.
6. Огарок А.Л., Селиванов С.А. «Кибербезопасность сложных информационных и управляющих систем» // Информатизация и связь. – 2020. – № 1. – С. 40-46.
7. Старовойтов А.В., Богданов Ю.М., Огарок А.Л., Селиванов С.А. Анализ эвристического и нейроматематического подходов к разработке алгоритмов и созданию систем // Информатизация и связь. – 2018. – №1. – С. 7-13.
8. Sowa J.F. Conceptual structures: information processing in mind and machine. Reading. – MA: Addison – Wesley, 1984.
9. Gorshkov S. Building ontologies for agent-based simulation //Advances in Swarm and Computational Intelligence, 6th International Conference, ICSI 2015 held in conjunction with the Second BRICS Congress, CCI 2015, Beijing, China, June 25-28, 2015, Proceedings, Part III – P. 185-193.
10. Matuszek C., Witbrock M., Kahlert R., Cabral J., Schneider D., Shah P., Lenat D. Searching for common sense: populating cyc from the web // Proceedings of the Twentieth National Conference on Artificial Intelligence, Pittsburgh, Pennsylvania. – July 2005. – PA: AAAI Press, 2005. – P. 1430-1435.
11. Бетин В.Н., Лукьянов С.Э., Супрун А.П. Обработка и обобщение знаний в интеллектуальной системе поддержки принятия решений ситуационного центра, построенной на базе сетей функциональных нейронов // Информатизация и связь. – 2013. – № 3. – С.10-15.
12. Бетин В.Н., Лукьянов С.Э., Супрун А.П. Выделение знаний из текстов на естественном языке в интеллектуальной аналитической системе // Информатизация и связь. – 2011. – № 6. – С. 51-54.
13. Бетин В.Н., Супрун А.П. Ситуационные центры: методика грамматического обучения интеллектуальной аналитической системы // Научно-техническая информация. Сер. 2. – 2013. – № 9. – С. 30-34.
14. Бетин В.Н. Дедуктивный вывод в системах автоматизированного проектирования и машинного перевода // Научно-техническая информация. Сер. 2 . – 2004. – № 7. – С. 20
15. Бетин В.Н., Лукьянов С.Э., Супрун А.П. Оптимизация алгоритмов поиска решения в системах поддержки принятия решений, реализованных в формализме функциональных нейронных сетей // Информатизация и связь. – 2016. – № 4. – С. 37-45.
16. Бетин В.Н., Лукьянов С.Э., Супрун А.П. Повышение эффективности подсистемы принятия решений в ситуационных центрах путем сжатия пространства решений // Информатизация и связь. – 2012. – № 8. – С. 38-40.
17. Бетин В.Н., Лукьянов С.Э., Супрун А.П. Использование метазнаний в системе поддержки принятия решений, реализованной в формализме сетей функциональных нейронов // Научно-техническая информация. Сер. 2. – 2016. – № 1. – С.16-20.

Материал поступил в редакцию 27.02.20.

Сведения об авторах

БЕТИН Владимир Николаевич – кандидат технических наук, ведущий научный сотрудник отдела разработки СПО автоматизированных систем Федерального государственного автономного научного учреждения "Центр информационных технологий и систем органов исполнительной власти" (ФГАНУ ЦИТиС), Москва
e-mail: betin_v@imail.ru, betin@inevm.ru

ЛУКЬЯНОВ Станислав Эмильевич – кандидат технических наук, главный специалист НТЦ перспективных технологий информационных процессов ФГАНУ ЦИТиС
e-mail:lukyanov@inevm.ru

СУПРУН Антон Павлович – начальник отдела разработки СПО автоматизированных систем ФГАНУ ЦИТиС
e-mail: suprun@inevm.ru

УДК [004.89:510.64]:616–006

Д.К. Чебанов, И.Н. Михайлова

Интеллектуальная система для анализа онкологических данных, реализующая ДСМ-метод автоматизированной поддержки исследований*

Проведено исследование генетических, клинических и иммунных данных пациентов с меланомой, с целью предсказания степени агрессивности заболевания, что позволит организовать персонализированный лечебный процесс исходя из индивидуального риска наступления ремиссии или рецидива. Одновременно выявлены комбинации генетических мутаций, которые могут служить маркерами таких состояний, что позволит создавать тест-системы без необходимости определения полного перечня генов. Приведены описание интеллектуальной системы на основе ДСМ-метода автоматизированной поддержки исследований, подробно разобраны её процедуры и стратегии анализа данных, а также результаты работы системы с примерами из исходных данных.

Ключевые слова: *искусственный интеллект, интеллектуальная система, онкология, меланома, генетические данные, мутации, иммунные данные, ДСМ-метод АПИ*

DOI: 10.36535/0548-0027-2020-05-3

ВВЕДЕНИЕ

В основе любого онкологического заболевания лежат нарушения в работе ДНК, которые в большинстве случаев являются мутациями. Такие мутации вызывают повреждение функции содержащего их гена. Не все мутации ДНК являются онкогенными, т. е. вызывающими возникновение и развитие злокачественных опухолей: существуют различные степени клинической значимости нарушений ДНК. В настоящее время в онкологии и молекулярной биологии работы по выявлению новых онкогенных мутаций продолжаются с помощью лабораторных экспериментов или компьютерного моделирования. В этой связи в современной медицинской науке возникает важный вопрос: каков вклад каждой мутации в прогноз заболевания? Ответ на него осложняется тем фактом, что чаще всего действие мутаций обусловлено их комбинацией, что подразумевает либо взаимное усиление онкогенного эффекта, либо его частичную или полную нейтрализацию. В связи с этим практикующих врачей интересует: можно ли, имея данные о мутациях, определить степень агрессивности заболевания, в частности, спрогнозировать наступление ремиссии или рецидива, чтобы точнее планировать лечебный процесс, его продолжительность и интенсивность.

В настоящей работе авторами решаются две актуальные задачи:

1. Выявление пациентов, степень агрессивности заболевания у которых предполагает наступление ремиссии, отделив их от пациентов с продолжающимся заболеванием, или рецидивом.

2. Определение комбинаций нарушений работы генов, по которым можно сделать прогноз заболевания (наступление ремиссии либо рецидива заболевания), и которые, соответственно, являются онкогенными или иммуногенными (влияют на иммунный ответ). Эта задача подразумевает также выявление таких мутаций, которые ранее не считались онкогенными, но по итогам исследования могут быть отнесены к клинически значимым для образования и развития опухолей.

В работе [1] авторами был применен ДСМ-метод автоматизированной поддержки исследований (ДСМ-метод АПИ) к данным пациентов с меланомой, содержащим генетическую информацию. В настоящей статье на базе представленного ранее исследования реализовано создание полноценной интеллектуальной системы: описаны ее компоненты, расширен инструментарий интеллектуального анализа данных, увеличен массив данных как в части числа пациентов, так и в количестве описывающих их признаков, а также уделено повышенное внимание методу исследования, в частности подробно описаны процедуры анализа данных.

* Работа выполнена при частичной финансовой поддержке РФФИ (проект № 18-29-03063)

ОПИСАНИЕ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ

Созданная нами интеллектуальная система основана на ДСМ-методе автоматизированной поддержки исследований (ДСМ-метод АПИ), реализующем такие компоненты интеллектуальной системы как: пользовательский интерфейс, базу фактов (БФ), ДСМ-решатель [3]. ДСМ-метод АПИ обнаруживает закономерности в сложноструктурированных эмпирических данных, которые содержат в неявном виде причинно-следственные зависимости, и обеспечивает формализацию знаний предметной области средствами многозначной логики, для чего обобщает в гипотезах информацию, полученную из обучающей выборки, затем применяет эти гипотезы для предсказания исследуемого эффекта неизвестных объектов. ДСМ-метод АПИ имеет критерий достаточного основания правдоподобного вывода [3].

Интеллектуальная система (ИС) работает на локальном компьютере пользователя. Все необходимые для работы системы данные также хранятся в памяти компьютера. ИС реализована на языке программирования python 3.7.

Система состоит из шести основных компонентов, входящих в её состав (рис.1):

1. Модуль процедур ДСМ-метода (Решатель ДСМ-системы)
2. Базовая инфраструктура
3. Модуль подготовки базы фактов
4. Модуль пользовательского интерфейса
5. Файловая система базы фактов
6. Модуль расшифровки результатов.

Решатель ДСМ-системы реализует синтез трёх познавательных процедур: индукции (для анализа данных), аналогии (для предсказания изучаемых эффектов) и абдукции (для принятия порождённых гипотез с использованием объяснения) [3].

Вычислительные процессы в Решателе реализованы параллельно на трёх уровнях:

1) обработки (+)-примеров (пациентов с эффектом «ремиссия», имеющих истинностное значение «фактически истинно») и (–)-примеров (пациентов с эффектом «меланома», имеющих истинностное значение «ложно»). Эти процессы выполняются на текущем уровне независимо друг от друга;

2) обработки всех возможных перестановок расширений БФ: в каждом из потоков обрабатывается отдельная последовательность расширений БФ (16 потоков);

3) выполнения стратегий: каждая стратегия выполнялась в отдельном потоке (16 потоков), независимо от остальных.

Решатель выполнялся на кластере Amazon AWS EC2 Instance r5.8xlarge с 32-я виртуальными ядрами и 256 ГБ оперативной памяти, однако оптимизация, примененная в ходе программной реализации интеллектуальной системы, позволяет успешно выполнять все процедуры также и на персональном компьютере: время расчета в таком случае составляет около 10 часов.

Базовая инфраструктура включает установленный на компьютере пользователя интерпретатор python 3.7 с пакетами pandas, multiprocessing, csv, time, PyQt.

Модуль подготовки БФ состоит из скриптов, которые приводят исходные табличные данные формата .xls(x) к виду, необходимому для работы Решателя, и сохраняют в формате .csv.

Пользовательский интерфейс реализован с помощью фреймворка Qt и библиотеки PyQt. Он предназначен для взаимодействия пользователя с исходной базой фактов, настройками ДСМ-решателя и результатами исследования, и позволяет выполнить следующие операции:

- подготовку и загрузку данных (исходной БФ);
- запуск процесса интеллектуального анализа данных (проведение ДСМ-исследования, т. е. обнаружение эмпирических закономерностей в последовательно расширяемых БФ посредством применения ДСМ-рассуждения);
- получение результатов исследования: гипотезы о причинах и гипотезы о предсказаниях;
- протоколирование процесса исследования, в том числе отслеживание возникающих ошибок.

Файловая система включает совокупность файлов с исходными данными, подготовленной БФ, а также файлов, полученных в ходе эксперимента: результаты в виде списков гипотез и эмпирических закономерностей, протоколы работы системы, вспомогательные файлы для работы процедур.

Модуль расшифровки результатов позволяет представить полученные гипотезы и закономерности в обозначениях и терминах исходной задачи, что дает возможность пользователю-эксперту проанализировать результат.

Рис. 1. Архитектура интеллектуальной системы

Перед проведением процедуры исследования необходимо подготовить базу фактов:

1) определить операции сходства для примеров исходных данных. В рассматриваемом наборе данных результатом применения операции сходства будет являться совпадение признаков у тех объектов, на которых определяется сходство. В случае, если признаки не совпадают, результатом операции сходства будет пустой элемент;

2) выделить объекты с наличием исследуемого эффекта и объекты с его отсутствием, что в терминологии ДСМ-метода означает присваивание оценки истинностных значений: «+1» (истинностное значение «фактически истинно») – для примеров первой группы, что дает (+)-примеры, и «-1» (истинностное значение «ложно») – для примеров второй группы, что дает (−)-примеры;

3) выделить примеры, о которых неизвестны данные о наличии или отсутствии эффекта, – для формирования контрольной группы для предсказания, объекты из этой группы имеют эффект «истинностное значение неопределенности» (обозначаются «τ»).

В Решателе реализованы эвристики ДСМ-метода АПИ, которые были применены в работах [1, 4, 5]. Решатель реализует следующие процедуры:

1. Атомарный ДСМ-метод АПИ (изучаемый эффект состоит из одного признака), который включает:

- предикаты сходства: простое сходство (обозначается *a*), сходство с запретом на контр-примеры (обозначается *b*), упрощённый метод сходства-различия (обозначается *d*0);
- 16 стратегий, каждая из которых определена парой предикатов сходства: для (+)-примеров и для (–)-примеров (M^+-предиката и M^--предиката соответственно) [2]: $Str_{x,y}$, $x \in \{a^+, (ab^+), (ad0^+), (ad0b^+)\}$, $y \in \{a^-, (ab^-), (ad0^-), (ad0b^-)\}$

Используемые стратегии с присвоенными им номерами перечислены в Листинге 1.

1) $Str_{a,a}$, 2) $Str_{ab,a}$, 3) $Str_{ad0,a}$, 4) $Str_{ad0b,a}$, 5) $Str_{a,ab}$,
6) $Str_{ab,ab}$, 7) $Str_{ad0,ab}$, 8) $Str_{ad0b,ab}$, 9) $Str_{a,ad0}$,
10) $Str_{ab,ad0}$, 11) $Str_{ad0,ad0}$, 12) $Str_{ad0b,ad0}$,
13) $Str_{a,ad0b}$, 14) $Str_{ab,ad0b}$, 15) $Str_{ad0,ad0b}$,
16) $Str_{ad0b,ad0b}$.

Листинг 1. Список стратегий, реализованных в системе

2. Задание количества и объема расширений БФ. В настоящем исследовании установлено 4 последовательных расширения, каждое из которых насчитывает 100 примеров. Процедура обеспечивает n! перестановок расширений, таким образом в проведенном рассуждении обработано 16 перестановок расширений БФ. Перестановки минимизируют случайность в организации БФ, и повышают тем самым устойчивость системы.

ОПИСАНИЕ БАЗЫ ФАКТОВ

В распоряжении исследователей имелись генотипические (мутации в ДНК), фенотипические (общие клинические данные) и иммунные (наличие мутаций в генах, ассоциированных с иммунным ответом при меланоме [9]) данные пациентов с меланомой кожи: одни из них имеют состояние ремиссии, другие – действующий диагноз. Сведения были взяты из базы данных TCGA Research Network (https://www.cancer.gov/tcga).

Выбор в качестве признаков комбинации генотипических, фенотипических и иммунных свойств обусловлен тем, что, согласно биологическим представлениям, именно в этих свойствах содержатся причины возникновения ремиссии, как результата имеющихся в опухоли мутаций, свойств организма, и ответа иммунной системы на присутствие опухоли в организме. Проведение интеллектуального анализа данных призвано выделить из представленных свойств именно те, которые являются признаками ремиссии, а также, с высокой вероятностью, и ее фундаментальными первопричинами. Такой подход позволит не только разработать тест-систему для прогнозирования развития заболевания, но и приблизиться к разработке новых методов лечения.

Одной из задач, поставленной в исследовании, было определение влияния иммунных данных на предсказание. С этой целью были проведены два эксперимента: вначале без иммунных данных в исходном массиве, затем с полным набором данных.

Массив исходных данных, или первичная база фактов в настоящей работе дополнен 57 случаями по сравнению с данными, использованными нами в предыдущем ДСМ-исследовании [1], и составляет 414 пациентов. Полный перечень признаков БФ приведен в *Приложении 1*. БФ содержит 216 (+)-примеров – пациентов, у которых отмечено состояние ремиссии не менее 2 лет, и 198 (–)-примеров – пациентов у которых данное состояние не отмечено. Таким образом, изучаемым эффектом примеров является наличие либо отсутствие у пациента ремиссии. Из 414 пациентов 400 были выделены для обучения интеллектуальной системы, остальные – оставлены на предсказание.

Количество признаков составило 342 – без иммунных данных и 491 – с иммунными данными, из них: генотипических – 335, фенотипических – 7, иммунных – 149. Каждый из перечисленных признаков является бинарным, т. е. принимает значение булевской переменной (0 или 1). Так, для генетических данных значением является наличие или отсутствие патогенной мутации в гене; для количественных клинических данных, таких как возраст, значение «1» является превышением медианного значения для всех пациентов, в то время как «0» присваивается значениям ниже медианы; для иммунных данных значение «1» присваивается в случае превышения порогового значения клинической нормы, «0» – если порог нормы не превышен.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Было проведено ДСМ-рассуждение с использованием 16 стратегий ДСМ-метода АПИ. Исследование проводилось на двух БФ: в первой использовались исключительно генотипические и фенотипические данные, во второй – те же данные с добавлением признаков иммунных данных.

Правильные предсказания и ошибки для всех стратегий сведены в табл. 1 и 2, в которых приведены следующие количественные показатели: колонка ***l0*** – правильные предсказания; колонка ***a*** – критичные ошибки, при которых (+)-примеры предсказаны как (–)-примеры и наоборот; колонка ***b*** – некритичные ошибки, когда пример предсказан с эффектом «0» (истинностное значение «противоречие»); колонка ***c*** – отказы от предсказания.

Таблица 1

Таблица правильных предсказаний и ошибок для стратегий при исследовании БФ без иммунных данных

Стратегия	*l0*		*a*		*b*		*c*	
	+	-	+	-	+	-	+	-
1	5	5	2	2				
5,13	7			6				1
9	5	4	1	2	1	1		
2,4,10,12	2	6	5	1				
6,8	6	2	1	4				1
14,16	6	2		4	1			1
3	5	4	2	2		1		
7,15	7			7				
11	5	4	1	3	1			
объединение	6	5	1	2				

Таблица 2

Таблица правильных предсказаний и ошибок для стратегий при исследовании БФ с иммунными данными

Стратегия	*l0*		*a*		*b*		*c*	
	+	-	+	-	+	-	+	-
1,3,9,11	6	4	1	3				
5,7,13,15	7			6				1
2,4,10,12	2	6	5	1				
6,8	6	2	1	3		1		1
14,16	6	2	1	4				1
объединение	6	5	1	2				

Таблица 3

Правильные предсказания примеров стратегиями (обозначены «1»), для БФ без иммунных данных

Пример	Стратегия																
	1	5	9	13	2	6	10	14	3	7	11	15	4	8	12	16	объединение
392		1		1						1		1					
393		1		1		1		1		1		1		1		1	1
394	1	1	1	1		1		1	1	1	1	1		1		1	1
395	1		1		1	1	1	1	1		1		1	1	1	1	1
396					1		1						1		1		
397	1	1	1	1		1		1	1	1	1	1		1		1	1
398	1	1	1	1		1		1	1	1	1	1		1		1	1
399	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
400	1		1		1		1		1		1		1		1		1
401	1		1		1		1		1		1		1		1		1
402	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
403	1		1		1		1		1		1		1		1		1
404																	
405	1				1	1	1	1					1	1	1	1	1
Всего верных предсказаний	**10**	**7**	**9**	**7**	**8**	**8**	**8**	**8**	**9**	**7**	**9**	**7**	**8**	**8**	**8**	**8**	**11**

Таблица 4

Количество гипотез и эмпирических закономерностей, участвовавших в правильных предсказаниях

Пример	Фактический знак примера	(+)-гипотезы		(–)-гипотезы	
		вкладывающиеся в пример	в т.ч. гипотезы, являющиеся ЭЗК	вкладывающиеся в пример	в т.ч. гипотезы, являющиеся ЭЗК
393	1	400	19	351	0
394	1	4902	50	2533	3
395	-1	71	1	81	1
397	1	68	2	35	0
398	1	308	21	170	1
399	1	316	8	43	0
400	-1	45	1	57	2
401	-1	266	1	303	0
402	1	305	6	56	0
403	-1	1	1	14	1
405	-1	36	1	48	0

Наилучший результат, составляющий 10 верных предсказаний из 14, был достигнут при применении стратегии, в которой оба предиката сходства используют метод простого сходства.

Добавление иммунных данных позволяет лучше работать стратегиям с упрощенным методом сходства-различия: результативность таких стратегий становится наивысшей (10 верных предсказаний из 14).

Таблицы 1 и 2 показывают сильную эквивалентность стратегий, что в терминах ДСМ-метода АПИ означает совпадение, как правильных предсказаний, так и ошибок у примененных стратегий.

Правильные предсказания для каждого из примеров представлены в табл. 3 (примеры правильных предсказаний каждого знака представлены в *Приложении 2*). Их анализ показывает, что всего один пример ни разу не был предсказан корректно (неправильные предсказания представлены в *Приложении 3*). Остальные примеры были правильно предсказаны не менее 4 стратегиями, и 2 примера предсказывались верно всеми 16-ю стратегиями. Таким образом, имеет место пересечение правильных предсказаний у различных стратегий, что также отразилось в объединении стратегий: оно было организовано объединением гипотез, порожденных всеми стратегиями, с учетом принципа непротиворечивости, и последующим предсказанием примеров с помощью полученного набора стратегий. При помощи объединения стратегий достигнут результат в 11 верных предсказаний, что превышает максимальный результат стратегий, взятых в отдельности.

Объединение гипотез, полученных при применении различных стратегий, для обеих БФ обеспечило улучшение качества предсказания неизвестных примеров за счет правильного предсказания 11-ти примеров из 14-ти, в то время как лучший результат при использовании одиночных стратегий составлял 10 верных предсказаний.

Согласно терминологии ДСМ-метода АПИ, гипотезы, сохраняющиеся при всех возможных перестановках последовательных расширений, являются эмпирическими закономерностями (ЭЗК). В проведенном нами исследовании получены ЭЗК в количестве 73 для знака (–) и 139 для знака (+). При этом наиболее ценными являются такие ЭЗК, которые участвуют в правильных предсказаниях неизвестных примеров: они считаются эмпирическими законами (ЭЗ) [6]. В проведенном исследовании для БФ без иммунных данных было обнаружено 43 эмпирических закона для знака (+) и 5 – для знака (–), они представлены в *Приложении 4*. Для базы фактов с иммунными данными было обнаружено 21 эмпирических закона для знака (+) и 3 – для знака (–), они представлены в *Приложении 5*.

В табл. 4 показано участие ЭЗК в правильных предсказаниях.

Эмпирические законы, которые могут быть использованы в качестве маркеров рецидива, имеют знак (–). Среди них, после оценки экспертом, было выявлено 3 эмпирических закона, имеющие физический смысл: (ZNRF3, RPS6, CCND1), (DLL4, CCND1, CDK4, IFNA1), (ZNRF3, CCND1, IFNA1). Согласно результату ДСМ-рассуждения, мутации в этих генах означают высокий риск наступления рецидива, и соответственно являются онкогенными. Таким образом, решается вторая поставленная перед исследованием задача, а именно нахождение новых мутаций с онкогенными свойствами, которые не были ранее квалифицированы как таковые в известных источниках.

Так, из полученных для негативного эффекта законов возможно выделить мутированные гены, которые ранее не были отмечены как патогенные для текущего заболевания. Этими генами являются RPS6, DLL4, IFNA1, которые входят в сигнальные каскады вместе с генами с подтверждённой степенью онкогенности, однако они не входят в перечни онкогенов и генов-супрессоров опухолевого роста.

ОБСУЖДЕНИЕ

Добавление иммунных данных повысило устойчивость системы в части предсказания примеров: это выражается в том, что наивысшего показателя достигла не одна стратегия со слабыми предикатами (сходства), как в эксперименте без иммунных данных, а четыре, включая стратегии с упрощенным методом сходства-различия и запретом на контр-примеры.

Анализ эмпирических законов, полученных в ходе исследования базы фактов с иммунными данными, показывает наличие признаков, которые в соответствии с клиническими представлениями достоверно влияют на прогноз заболевания (стадия, возраст, пол, появление новых образований). Это свидетельствует о дополнительном доверии к примененному методу интеллектуального анализа данных.

Обнаруженные гены могут быть проработаны с точки зрения включения и в тест-системы для определения прогноза заболевания, либо рассмотрения их в качестве терапевтических мишеней при разработке препаратов.

Пример, который всеми стратегиями предсказывался неверно (*Приложение 3*), имеет 34 признака: это высокое значение, так как 74% пациентов имеют меньшее количество признаков. Этот факт дает основания полагать, что внутри данных признаков могут находиться фрагменты совокупности причин, находящиеся в противоречии с гипотезами, и именуемые в терминологии ДСМ-метода «тормозами» [7].

Второй пример, предсказанный неправильно 12 стратегиями из 16-ти, имеет незначительный перевес в количестве содержащихся в нем гипотез каждого из знаков (20 против 24), что говорит о том, что результат предсказания близок к фактическом противоречию.

Третий неправильно предсказанный пример имеет в качестве признаков информацию о наличии всего 3-х мутаций, что значительно меньше, чем количество генотипических данных у большинства примеров, таким образом, возможности по вложению гипотез в данный пример ограничены. Это может быть вызвано некорректным лабораторным исследованием, – в любом случае пример является нетипичным для имеющейся базы фактов.

Объединение стратегий является действенной мерой, повышающей результативность системы в части количества правильных предсказаний.

Для устранения ошибок предсказаний планируется пересмотр набора признаков, входящего в исходную базу фактов.

Развитие представленной системы предполагается в следующих направлениях:

1) доработка пользовательского интерфейса с целью автоматизации ряда процедур, в частности подготовки БФ;

2) добавление комплексного значения эффекта примеров;

3) анализ ошибок и улучшение показателей предсказания.

ЗАКЛЮЧЕНИЕ

Меланома кожи – один из наиболее опасных онкологических заболеваний. Для клинических специалистов дополнительная сложность в лечении меланомы заключается в невозможности предсказать, у кого из пациентов будет происходить обострение заболевания после удаления первичного опухолевого образования, что позволило бы заблаговременно предпринять превентивные терапевтические меры.

В настоящей работе предпринята попытка при помощи интеллектуальной системы, реализующей ДСМ-метод автоматизированной поддержки исследований, отделить пациентов, у которых ожидается рецидив, от пациентов с продолжающейся ремиссией. При помощи интеллектуального анализа данных корректно предсказано состояние 11 пациентов из 14. Кроме того, выявлены генетические факторы, вносящие вклад в характер развития заболевания у пациентов.

Использование системы ДСМ-метода АПИ подразумевает наличие в качестве исходных данных информации, которая может быть получена во многих современных научно-клинических учреждениях. При последующем пополнении базы фактов ожидается выявление специфичных маркеров, которые могут заменить полный перечень признаков без потери точности прогноза.

Системы ДСМ-метода с расширяемыми базами фактов для целей прогнозирования развития меланомы необходимо внедрять в клиническую и научно-исследовательскую практику, для выявления групп риска среди пациентов и последующей реализации концепции персонализированной медицины, а также для получения новых знаний о механизмах онкологических заболеваний.

* * *

Авторы выражают благодарность Михаилу Ивановичу Забежайло за ценные рекомендации и идеи в процессе работы.

СПИСОК ЛИТЕРАТУРЫ

1. Чебанов Д.К., Михайлова И.Н. Интеллектуальный анализ данных пациентов с меланомой для поиска маркеров заболевания и значимых генов // Научно-техническая информация. Сер. 2. – 2019. – № 10 – С. 35-40; Chebanov D.K., Mikhailova I.N. Intellectual Mining of Patient Data with Melanoma for Identification of Disease Markers and Critical Genes // Automatic Documentation and Mathematical Linguistics. – 2019. – Vol.53, № 5. – P. 283-287.
2. Финн В.К. Дистрибутивные решетки индуктивных ДСМ-процедур // Научно-техническая информация. Сер. 2. – 2014. – № 11. – С.1-36: Finn V.K. Distributive Lattices of Inductive JSM Procedures // Automatic Documentation and Mathematical Linguistics. – 2014. – Vol. 48, № 6. – P. 264-295.
3. ДСМ-метод автоматического порождения гипотез: Логические и эпистемологические основания / сост. О.М. Аншаков, Е.Ф. Фабрикантова; под. общ. ред. О.М. Аншакова. – М.: ЛИБРОКОМ, 2009. – 433 с.
4. Шестерникова О.П., Финн В.К., Винокурова Л.В., Лесько К.А., Варварина Г.Г., Тюляева Е.Ю. Интеллектуальная система для диагностики заболеваний поджелудочной железы // Научно-техническая информация. Сер. 2. – 2019. – № 10. – С. 41-48: Shesternikova O.P., Finn V.K., Vinokurova L.V., Les'ko K.A., Varvanina G.G., Tyulyaeva E.Yu. An Intelligent System for Diagnostics of Pancreatic Diseases // Automatic Documentation and Mathematical Linguistics. – 2019. – Vol.53, № 5. – P. 288-294.

5. Финн В.К. Об эвристиках ДСМ-исследований (дополнения к статьям) // Научно-техническая информация. Сер. 2. – 2019. – № 10. – С. 1-34; Finn V.K. On the Heuristics of JSM Research (Additions to Articles) // Automatic Documentation and Mathematical Linguistics. – 2019. – Vol. 53, № 5. – P. 250-282.
6. Финн В.К. Об определении эмпирических закономерностей посредством ДСМ - метода автоматического порождения гипотез // Искусственный интеллект и принятие решений. – 2010. – № 4. – С. 41-48.
7. Финн В.К., Шестерникова О.П. О новом варианте обобщенного ДСМ-метода автоматизированной поддержки научных исследований // Искусственный интеллект и принятие решений. – 2016. – № 1. – С. 57-64.
8. Zaretsky J.M. et al. Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma // The New England Journal of Medicine. – 2016. – Vol. 375, № 9. – P. 819-829.

ПРИЛОЖЕНИЕ 1

Список признаков БФ

1.1. Генотипические признаки: мутации, вызывающие изменение белка, в следующих генах:
DLL4, MGA, NRAS, ZNRF3, ERBB4, ROS1, BRAF, STK3, TLE1, YAP1, FAT3, KSR2, NCOR2, CNTN1, FLT3, JAG2, NTRK3, AXIN1, LLGL1, KSR1, NOTCH3, CBLC, INSRR, CRB1, PSEN2, WNT3A, FZD7, DNER, RASGRP3, SOS1, CNTN6, FAT4, DCHS2, APC, HEY2, NOTCH4, MET, WNT2, HIPK2, MLXIPL, TGFBR1, MAPKAP1, NOTCH1, SHC3, NCOR1, DVL2, TP53, RASGRP4, INSR, NOTCH2, MYCN, FAT2, FGFR2, PTEN, KRAS, CHD8, SOS2, TSC2, ERBB2, SMAD2, SMAD4, CRB3, MTOR, SHC1, NTRK1, WNT9A, AKT3, HDAC1, IRS1, CBLB, WWTR1, FBXW7, FAT1, MXD4, RBPJ, MXD3, RICTOR, THBS2, EGFR, FGFR1, RAPGEF1, AMER1, RASGRP2, PPP1CA, NF1, LZTR1, RCE1, FZD4, WIF1, CREBBP, RPTOR, ITCH, MRAS, WNT7A, RASA1, DLL1, HEY1, ARRDC1, HRAS, LRP5, PEBP1, WNT10B, ACVR1B, IRS2, DLK1, RASGRP1, ADAM10, RASGRF1, IGF1R, DLL3, NUMBL, TEAD2, HMCN1, CHEK2, DEPDC5, MFNG, ACVR2A, RAF1, TGFBR2, PLXNB1, LEF1, RAPGEF2, PDGFRA, KIT, PDGFRB, PIK3R1, SCRIB, DAB2IP, TSC1, CTBP2, RET, FRAT2, ATM, PTPN11, DTX1, RASAL1, FZD10, TEAD4, LATS2, NUMB, SHC4, MLST8, MNT, TAOK1, STK11, HEYL, ICMT, E2F1, ALK, PIK3CA, HES1, RASGRF2, CUL1, RAC1, SFRP1, ABL1, JAK2, NTRK2, CDKN2C, KDM5A, PSENEN, SPEN, FZD9, TLE4, ARAF, SOST, LLGL2, MYCL, WWC1, MAP2K1, DKK3, ERBB3, STK4, EGFL7, RASAL2, WNT8A, AKT1, SPRED2, DKK2, TAOK2, MAPK1, HDAC2, SFRP5, MAP2K2, DTX3L, LIMD1, MST1, INPP4B, SHOC2, CBL, CHD4, MDM2, MAX, SPRED1, SMAD3, RFNG, CSNK1D, ARHGAP35, PPP2R1A, WNT4, HES3, JAG1, EP300, TCF7L1, WNT5A, CTBP1, MOB1B, LATS1, WNT16, TAOK3, FZD1, CDKN2A, CCNE1, CTNNB1, E2F3, DTX2, FZD8, WNT5B, PSEN1, TLE2, MAML1, RHEB, LFNG, FHL1, RB1, MLXIP, LRP6, SAV1, AXIN2, KAT2B, FGFR4, NOV, CRB2, KEAP1, MYC, DKK1, DTX3, RASA3, AJUBA, MLX, PDK1, WNT1, FGFR3, WNT8B, RASAL3, PIN1, RASA2, TEAD3, RPS6, CCND1, CDK4, MAML3, MFAP5, NCSTN, PTPN14, RSPO1, PIK3R3, MXD1, ADAM17, DVL3, SFRP2, EIF4EBP1, NDP, DCHS1, NPRL3, AKT1S1, CSNK1E, FZD6, FZD3, WNT11, FZD2, RBPJL, PIK3CB, SNW1, SPRED3, TCF7L2, WTIP, ERF, DVL1, LGR4, TLE3, NF2, MAML2, AKT2, TAZ, PORCN, HES2, FNTB, MAPK3, CCND3, MXI1, RPS6KB1, RIT1, FNTA, ERRFI1, CUL3, TEAD1, CCND2, WNT7B, CIR1, NFE2L2, SKP1, MFAP2, GSK3B, DEPTOR, LGR5, CDK6, CDKN1A, WNT10A, WNT6, RPS6KA3, SFRP4, SHC2, PIK3R2, APH1B, WNT9B, DKK4, DTX4, MDM4, RNF43, SAP30, CDKN1B, HES4, TCF7, FZD5, CDKN2B, GRB2, APH1A, PEA15, CDK2, RBX1, NPRL2, FRAT1, NRARP, MOB1A

1.2. Фенотипические признаки:
- Возраст (до/свыше 58 лет)
- Пол
- Стадия по AJCC (I-II / III-IV)
- Неоадьювантная терапия
- Появление новых опухолей после первичного лечения
- Радиотерапия
- Общая выживаемость (до / свыше 36 мес)

1.3. Иммунные данные:
TNFSF11, LAMA2, TREML2, TNFRSF11A, TNFRSF1B, CCR5, MKI67, CD19, CXCR6, IL12RB2, CD276, CDCA3, TNFRSF10B, TNFRSF12A, PTGS2, STAT3, CDH4, ENTPD1, CCRL2, SIGLEC14, FOXP3, NGFR, CCL5, VCAM1, CD8B, JAK2, CDH2, STAT4, BMPR1B, ULBP1, MSR1, EDAR, CD28, ITGB1, FAS, LGALS9, NOS2, HRH4, CD40, SEMA3G, VEGFA, CD86, TDO2, HAVCR1, ARG1, IFNB1, TNFRSF19, CCL4, TNFRSF4, TGFBR1, ADAM12, CDH1, FASLG, CPE, RAET1E, CD40LG, GREB1, IFNG, MICB, NCAM1, LRRC32, TNFRSF13B, CTLA4, SIRPB1, CLEC5A, RELT, TGFB1, CD4, HAVCR2, TGFB2, CCL20, STAT1, CD8A, HLA-B, TNFRSF11B, AHR, TIMD4, TNFSF15, TGFB3, TGFBR2, EDA2R, IDO1, IL10, IL17RB, ICOS, CLEC2D, TNFSF14, CCR4, SLAMF1, EDA, TNFSF13B, CCR10, TNFRSF1A, LGALS3, ANXA2, LOXL1, SOCS1, CCL22, CDH3, TNFRSF18, TNFRSF8, TNFSF18, TNFSF4, TNFRSF14, TNFRSF25, TNFRSF6B, IL12A, TNFSF10, CCR6, TNF, HLA-DRA, TREM1, TNFRSF21, IL6, TNFRSF10C, TNFSF8, CD274, PDCD1LG2, CCL17, TNFSF9, IFNA1, IL23A, TMEM45A, TNFRSF10D, CD3D, HLA-DRB1, LTBR, CD247, TNFRSF9, SFRP1, CD3E, CD80, LAG3, CD27, HLA-C, HLA-E, ULBP3, LTA, LTB, TNFSF12, BTLA, TNFRSF17, CCL2, MICA, HLA-A, CD70, CLEC12B, CD68, ULBP2, CD3G, TNFRSF13C, TNFSF13, IFNA2

ПРИЛОЖЕНИЕ 2

Примеры правильных предсказаний

	Пример №402	**Пример №399**	**Пример №395**	**Пример №402**
Признаки	Мутация NRAS Мутация WNT3A Мутация FAT2 Мутация CHD8 Мутация RAC1 Мутация CHD4 Мутация PPP2R1A Мутация DKK1 Мутация RASA3 Возраст – свыше 58 лет Пол – мужской Общая выживаемость – свыше 36 мес. Зафиксирована прогрессия заболевания	Мутация ROS1 Мутация BRAF Мутация FAT3 Мутация PTEN Мутация HMCN1 Мутация PDGFRA Мутация TAOK1 Мутация ALK Мутация DTX4 Пол: мужской Общая выживаемость – свыше 36 мес.	Мутация BRAF Мутация FAT4 Мутация NOTCH4 Мутация HIPK2 Мутация NF1 Мутация MFNG Возраст – свыше 58 лет Пол - мужской Стадия по AJCC – III или выше	Мутация ROS1 Мутация KSR2 Мутация HMCN1 Пол – мужской Стадия по AJCC – III или выше
Число стратегий, предсказавших пример правильно	16	16	12	8
Фактический эффект примера	(+)	(+)	(-)	(-)
Число гипотез верного знака, вложившихся в пример	305	316	81	14
Число гипотез противоположного знака, вложившихся в пример	56	43	71	1
Число ЭЗК, участвовавших в предсказании	(+): 6 (-): 0	(+): 8 (-): 0	(+): 1 (-): 1	(+): 0 (-): 1

ПРИЛОЖЕНИЕ 3

Неправильные предсказания

	Пример №392	**Пример №396**	**Пример №404**
Признаки	Мутация DLL4 Мутация NRAS Мутация DVL1 Мутация LGR5 Возраст – свыше 58 лет Пол – мужской Стадия по AJCC – III или выше Зафиксирована прогрессия заболевания	Мутация NCOR1 Мутация NF1 Мутация TAOK2 Возраст – свыше 58 лет Пол – мужской Стадия по AJCC – III или выше Зафиксирована прогрессия заболевания	Мутация MGA Мутация ROS1 Мутация BRAF Мутация TLE1 Мутация FAT3 Мутация CRB1 Мутация FAT4 Мутация DCHS2 Мутация DVL2 Мутация FGFR2 Мутация CHD8 Мутация CRB3 Мутация CBLB Мутация THBS2 Мутация TSC1 Мутация SPEN Мутация ARAF Мутация LLGL2 Мутация STK4 Мутация SPRED2 Мутация INPP4B

	Пример №392	Пример №396	Пример №404
			Мутация WNT5A Мутация CTBP1 Мутация WNT16 Мутация CCNE1 Мутация LRP6 Мутация AJUBA Мутация RPS6 Мутация CDK4 Мутация DCHS1 Мутация RBX1 Пол: мужской Общая выживаемость – свыше 36 мес.
Число стратегий, предсказавших пример неправильно	12	12	16
Фактический эффект примера	(+)	(–)	(–)
Число гипотез противоположного знака, вложившихся в пример	20	103	2923
Число гипотез верного знака, вложившихся в пример	24	72	464

ПРИЛОЖЕНИЕ 4

Обнаруженные эмпирические законы, участвовавшие в правильных предсказаниях, и имеющие физический смысл по результатам анализа экспертом (БФ без иммунных данных)

3.1. (–)-ЭЗ:

- ZNRF3, IFNA1
- ZNRF3, RPS6, CCND1
- DLL4, CDK4, IFNA1
- DLL4, CCND1, CDK4, IFNA1
- ZNRF3, CCND1, IFNA1

3.2. (+)-ЭЗ:

- ZNRF3, STK3, CCND1, IFNA1, MFAP5
- STK3, CCND1, IFNA1, MFAP5
- STK3, LLGL1, CCND1, MFAP5
- STK3, IFNA1, MFAP5
- STK3, LLGL1, MFAP5
- PTGS2, STK3, MFAP5
- TNFRSF11B, IFNA1, MFAP5
- ZNRF3, CCND1, IFNA1, MFAP5
- LLGL1, PSEN2, MFAP5
- ZNRF3, PSEN2, MFAP5
- ZNRF3, STK3, IFNA1, MFAP5
- STK3, LLGL1, PSEN2, MFAP5
- STK3, WIF1, MFAP5
- TNFSF13B, MFAP5
- PSEN2, CCND1, IFNA1, MFAP5
- TNFRSF11B, CCND1, MFAP5
- ZNRF3, PSEN2, CCND1, MFAP5
- WNT3A, CCND1, MFAP5
- ZNRF3, IFNA1, MFAP5
- STK3, PSEN2, MFAP5
- STK3, LLGL1, PSEN2, CCND1, MFAP5
- DLL4, PSEN2, MFAP5
- STK3, LTB, MFAP5
- PSEN2, RPS6, MFAP5
- LTB, CCND1, MFAP5

Обнаруженные эмпирические законы, участвовавшие в правильных предсказаниях, и имеющие физический смысл по результатам анализа экспертом (БФ с иммунными данными)

3.1. (+)-ЭЗ:

- MGA, Общая выживаемость – свыше 36 мес
- CRB1, FAT4, Общая выживаемость – свыше 36 мес
- ROS1, FAT3, Общая выживаемость – свыше 36 мес
- FAT3, CRB1, FAT4, Общая выживаемость – свыше 36 мес
- FAT4, FAT1, Общая выживаемость – свыше 36 мес
- CDH4, FAT4, Общая выживаемость – свыше 36 мес
- STAT4, Общая выживаемость – свыше 36 мес
- NF1, Общая выживаемость – свыше 36 мес
- SPEN, Общая выживаемость – свыше 36 мес
- FAT3, CRB1, Мужской пол, Общая выживаемость – свыше 36 мес
- BRAF, FAT1, Общая выживаемость – свыше 36 мес
- PLXNB1, Общая выживаемость – свыше 36 мес
- FAT3, HMCN1, Общая выживаемость – свыше 36 мес
- FAT3, CRB1, FAT4, Мужской пол, Общая выживаемость – свыше 36 мес
- DCHS2, Общая выживаемость – свыше 36 мес
- NRAS, FAT4, Общая выживаемость – свыше 36 мес
- FAT3, SPEN, Общая выживаемость – свыше 36 мес
- FAT3, CRB1, Общая выживаемость – свыше 36 мес
- BRAF, FAT4, Мужской пол, Общая выживаемость – свыше 36 мес
- FAT3, FAT4, Общая выживаемость – свыше 36 мес
- HMCN1, Общая выживаемость – свыше 36 мес

3.1. (–)-ЭЗ:

- BRAF, Возраст свыше 58 лет, Мужской пол
- NRAS, AJCC выше III, Возникновение образований после первичного лечения
- NRAS, Мужской пол, AJCC выше III, Возникновение образований после первичного лечения

Материал поступил в редакцию 07.04.20.

Сведения об авторах

ЧЕБАНОВ Дмитрий Константинович – генеральный директор ООО «ОнкоЮнайт Клиникс»
e-mail: chebanov.dk@gmail.com

МИХАЙЛОВА Ирина Николаевна – доктор медицинских наук, ведущий научный сотрудник НМИЦ онкологии им. Н.Н. Блохина МЗ РФ,
e-mail: irmikhaylova@gmail.com

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 050: [002:001.8]

А.Н. Либкинд, В.А. Маркусова, И.А. Либкинд

К вопросу определения динамики показателей периода полужизни журналов по *Journal Citation Reports**

Обсуждаются проблемы, возникающие при определении значений показателей периода полужизни журнальных статей по ежегодным выпускам Journal Citation Reports за период 1997-2018 гг. В выпусках до 2017 г. точное значение этих показателей не указывалось, когда оно превышало 10 лет. Доля журналов с такими показателями нередко достигает нескольких десятков процентов. Предлагаются приемы, позволяющие обойти это ограничение и достоверно оценивать динамику периода полужизни статей по категориям Web of Science, что важно при неправомерном значении, придаваемом библиометрическим показателям.

Ключевые слова: *старение литературы, период полужизни, Cited Half-life, Citing Half-life, Journal Citation Reports*

DOI: 10.36535/0548-0027-2020-05-4

ВВЕДЕНИЕ

Возможность более точно измерять многие закономерности, присущие журнальным публикациям, в наукометрии появилась после создания Ю. Гарфилдом в 1964 г. «Указателя библиографических ссылок в естественных науках». Однако многие из этих закономерностей начали изучаться значительно раньше. Обсуждая способы оценки старения литературы в определенных областях знания, Дж. Бернал ещё в 1958 г. предложил для этого термин «период полужизни» журнальных статей по аналогии с периодом полураспада радиоактивных веществ. Его идею использовали американские ученые Р. Бартон и Р. Кеблер [1, с. 20]. Они разработали в 1960 г. способ вычисления этого показателя, который измеряли интервалом времени, в течение которого была опубликована половина всей используемой в настоящее время литературы по какой-либо отрасли или предмету. На современном языке это означает, что они имели в виду время опубликования половины статей, цитируемых в журналах данного года. По мере совершенствования библиометрического инструментария стал приниматься во внимание и возраст библиографических ссылок в самих журналах этого года, т.е. цитирующих статей.

Со временем выяснилось, что период полужизни отражает не только старение научной литературы, но и её рост, о чём выразительно написал Д. Прайс: «В течение нескольких лет после публикации спрашиваемость статьи или её относительная цитируемость уменьшается крайне медленно (по параболе, если считать по логарифмам прошедших лет). Даже через столетие возможность цитирования уменьшается только на порядок. Большинство ссылок падает на работы последних лет потому, что этих работ большинство, и очень сомнительно, чтобы это вызывалось эффектом немедленности, связанным с быстрым старением» [2, с. 292]. Эту проблему старения подробно изучали ученые-информатики: английский – М. Лайн [3] и российский – В.М. Мотылев [4]. М. Лайн писал: «Период полужизни литературы вынужден быть тем короче, чем быстрее она растёт, если среднее число цитирований на одну статью за это время не уменьшается. Если каждая статья имеет одинаковую вероятность быть использованной или процитированной, более новая литература используется чаще просто потому, что ее больше».

Для теоретиков информатики и историков науки важно учитывать старение литературы в чистом виде, а для информационных работников и библиотекарей период полужизни является важным практическим показателем и продолжает широко использоваться. В наше время, когда библиометрические показатели служат (хотя и не всегда оправдано) оценкой научной деятельности ученых, организаций и даже стран,

* Работа выполнены при поддержке Российского фонда фундаментальных исследований (проекты РФФИ 20-07-00014 и 20-010-00179).

важно понимать, как вычисляется и что реально значит такой показатель, как период полужизни журнальных статей [5, 6].

В настоящем исследовании в качестве исходных данных нами были использованы ежегодные выпуски «*Journal Citation Reports – Science Edition» (JCR-SE)* и «*Journal Citation Reports – Social Science Edition» (JCR-SSE)* за период 1997–2018 гг. Этот аналитический инструмент располагается на платформе *Web of Science (WoS)* компании *Clarivate Analytics*. Среди ряда показателей, которые приводятся в каждом ежегодном выпуске *JCR* для каждого включенного в этот выпуск журнала, нас, прежде всего, интересовали показатели, известные как показатели «периода полужизни» – *Cited Half-life* и *Citing Half-life*, и кроме того принадлежность журнала к той или иной тематической категории *WoS (subject categories WoS или WoS categories)*, а также соответствующие этим категориям показатели «периода полужизни».

Напомним основные определения, связанные с понятием «период полужизни», воспользовавшись определениями, содержащимися в разделе «*Help*» информационной платформы *WoS.* (https://help.incites.clarivate.com/ incitesLiveJCR/ overviewGroup/ overviewJCR. html)[1]:

- цитируемый период полужизни некоторого журнала (*Cited Half-life – CdHL*) представляет собой медианный возраст (в годах) массива тех статей из этого журнала, которые в заданном году выпуска «*Journal Citation Report» (JCR)* были процитированы. Это значит, что половина процитированных статей из указанного журнала были опубликованы ранее цитируемого периода полужизни этого журнала;
- цитирующий период полужизни журнала (*Citing Half-life – CgHL*) некоторого журнала представляет собой медианный возраст (в годах) массива тех статей, которые цитировал этот журнал в данном году выпуска *JCR*. Это значит, что половина статей, которые цитировал данный журнал, были опубликованы ранее периода полужизни;
- агрегированный цитируемый период полужизни (*Aggregate Cited Half-life – AgrCdHL*), характеризующий некоторую тематическую категорию *Web of Science (WoS category)*, представляет собой медианный возраст в годах публикаций в журналах, соответствующих этой категории (значения этого показателя получаются путем агрегирования соответствующих журнальных данных);
- агрегированный цитирующий период полужизни (*Aggregate Citing Half-life – AgrCgHL*), характеризующий некоторую тематическую категорию *WoS*, представляет собой медианный возраст статей, которые в год выпуска *JCR* цитировали журналы, соответствующие данной категории (значения этого показателя получаются путем агрегирования соответствующих журнальных данных).

ПРОБЛЕМЫ ПРИ ОПРЕДЕЛЕНИИ ПЕРИОДА ПОЛУЖИЗНИ ЖУРНАЛОВ

Приступая к задаче определения динамики показателей *Cited Half-life* (*CdHL*) и *Citing Half-life* (*CgHL*), мы были вынуждены учитывать ряд проблем. Во-первых, единственным источником статистических данных для решения задачи определения динамики показателей периода полужизни журналов может послужить информация, содержащаяся в ежегодных выпусках информационной системы *Journal Citation Reports* (*JCR*). Во-вторых, выполненный нами предварительный анализ *JCR* за период 1997–2018 гг. показал, что значения показателей *CdHL* и *CgHL* для каждого журнала, включенного в соответствующий ежегодный выпуск *JCR*, обычно приводятся в числовой форме с точностью до 0,1 года. При этом в ежегодных выпусках *JCR* за 2017 г. и 2018 г., данные о значениях *CdHL*/*CgHL* для каждого из журналов приводятся полностью, т.е. в числовой форме. К сожалению, во всех остальных ежегодных выпусках *JCR* (за период 1997–2016 гг.) картина иная. А именно, если значение соответствующего показателя для некоторого журнала превышает 10 лет, то для этого журнала вместо точного числового значения указывается только текстовое выражение вида «>10.0». Попытки определения динамики значений рассматриваемых показателей при использовании только тех совокупностей журналов, для которых значения показателей приводятся в числовой форме, приведут, в лучшем случае, к существенным искажениям. Действительно, использовать «напрямую» текстовые данные вида «>10.0» при соответствующих вычислениях не удастся и, следовательно, этими данными придется просто пренебречь. И это притом, что доля журналов, у которых значения *CdHL* больше 10 лет в выпусках *JCR-SE*, значительна и находится в пределах 15,7–21,5%, и со временем она возрастает. Так, нижний предел этой доли (15,7%) соответствует выпуску 1997 г., а верхний (21,5%) – выпуску 2016 г. В случае *CgHL* в тех же выпусках *JCR-SE* доля таких журналов еще выше и находится в пределах 29–31%. Доля журналов, для которых приводятся такие неточные данные, в случае *JCR-SSE* для *CdHL* находится в пределах 11,9–31%, а для *CgHL* – в пределах 25–50,9% (1997 г. и 2016 г. соответственно). Таким образом, приходится констатировать, что подавляющее большинство (20 из 22) выпусков *JCR* содержат неполные данные о значениях *CdHL* и *CgHL* конкретных журналов.

К сожалению, ситуация с такими неполными данными для тематических категорий *WoS* еще сложнее. Действительно, для всех без исключения ежегодных выпусков *JCR* и всех категорий *WoS* за весь период 1997–2018 гг. при значениях показателей, превы-

[1] Cited Half-life some journal is median age of the articles from this journal that were cited in the JCR year. It means that half of a journal's cited articles were published more recently than the cited half-life.

Citing Half-life is median age of articles cited by the journal in the JCR year. It means that the articles cited by this journal were published more recently than the citing half-life.

Aggregate Cited half-life is the median age, in years, of items in any journal in the category that were cited during the JCR year. Shows the distribution by cited year of citations to articles published in journals in the category in the JCR year.

Aggregate Citing half-life is the median age of articles cited by journals in the category in the JCR Year based on aggregated journal data. Shows the distribution by cited year of citations from journals in the category made in the JCR year.

шающих 10 лет, в *JCR* приводятся не числовые, а текстовые значения вида «>10.0», что не позволяет использовать эти значения для дальнейших вычислений[2].

Отметим, что если игнорировать журналы, для которых в *JCR* в качестве значений *CdHL* (*CgHL*) указаны текстовые выражения «>10», то с каждым годом в расчетные значения этих показателей будут вноситься всё бо́льшие искажения. В итоге, при положительной динамике показателей результаты расчетов могут дать динамику отрицательную, что совершенно недопустимо.

Если рассмотреть эту ситуацию с методической точки зрения, то здесь необходимо указать следующее. Казалось бы, исходя из чисто статистических соображений, даже с указанными выше потерями можно было бы пренебречь, если бы не одно существенное обстоятельство. А именно, при попытке ограничиться только выборкой журналов, которые имеют числовые значения показателей, мы тем самым отбрасываем именно те журналы, которые характеризуются самыми большими значениями показателей. Таким образом, здесь мы имеем дело не с обычной случайной выборкой из генеральной совокупности и, следовательно, чисто статистический подход здесь недопустим.

Из всего изложенного следует, что прежде чем приступить к решению проблемы определения динамики показателей *CdHL* и *CgHL*, как для журналов, так и для категорий *WoS*, необходимо решить задачу оценки средних (точнее, средневзвешенных) значений этих показателей для каждого года из рассматриваемого периода**.** То есть**,** нам необходимо каким-либо образом восполнить недостающие данные в годовых распределениях журналов по этим показателям за 1997–2016 гг. Именно решению этой задачи и, в конечном счете, разработке методов для определения динамики значений показателей периода полужизни, характеризующих соответствующие совокупности журналов, посвящена настоящая статья. Прежде чем перейти к непосредственному рассмотрению этой задачи, а также соответствующих совокупностей журналов и их значений по показателям *CdHL/CgHL*, введем необходимые понятия и определения.

Определение 1. Будем называть распределением журналов по значениям показателя *CdHL* таблицу, в первой графе (колонке) которой последовательно в порядке возрастания приводятся значения *CdHL*, а во второй графе против каждого значения *CdHL* из первой колонки указано число журналов, каждый из которых имеет именно это значение *CdHL*. В дальнейшем для краткости вместо выражения «распределение журналов по значениям показателя» будем применять – «распределение».

Характер рассматриваемых нами совокупностей журналов и соответствующих им показателей позволяет классифицировать эти совокупности по следующим основаниям (признакам):

1) принадлежность журнала к заданному (i-му) ежегодному выпуску (году опубликования) *JCR-SE/ JCR-SSE*;

2) принадлежность журнала к данной тематической категории *Web of Science* (*subject category WoS – WoS Category*) и/или к заданному набору *WoS Categories*;

3) принадлежность журнала той или иной стране (страна издания журнала);

4) факт присутствия журнала в i-м выпуске *JCR-SE/JCR-SSE* и в следующем $i+1$ выпуске. Такие журналы будем называть *«Относительно сохранившимися журналами»* или *«Относительно постоянными журналами»* – "*Relatively regular journals (RR Journals)"* or *"Relatively preserved journals" (RP Journals)* or *"Relatively retentive journals (RR Journals)"*;

5) факт присутствия журнала во всех без исключения ежегодных выпусках *JCR-SE/ JCR-SSE* за заданный период (в нашем случае за 1997-2018 гг.). Эти журналы назовем *«Абсолютно сохранившимися журналами»* или *«Всегда присутствующими журналами»* – "*Absolutely preserved journals* or *always present journals (AP Journals)"* or "*Absolutely retentive journals (AR Journals)"* [7]. В качестве исходной точки на шкале времени в этом случае будет принят 1997 г.;

6) факт присутствия журнала в i+1 выпуске при обязательном отсутствии этого журнала в i-м выпуске. Такие журналы будем называть относительно новыми – *relatively new journals (RN Journals)*;

7) факт присутствия журнала в i-м выпуске *JCR-SE/JCR-SSE* при обязательном его отсутствии во всех предшествующих выпусках. Такие журналы назовем абсолютно новыми – *absolutely new journals (AN Journals).*

В настоящей работе будут рассмотрены те совокупности журналов и категорий *WoS*, которые могут быть сформированы с помощью признаков, указанных в пунктах (1)–(3) и (5). Для описания и сопоставления различных распределений журналов по *CdHL* и *CgHL* воспользуемся понятиями, применяемыми для аналогичных целей в статистике: среднее значение распределения (*mv – mean value*); мода распределения (*mode – Md*); медиана распределения (*median – Mn*); полная ширина на уровне половины максимального значения (*FWHM – full width at half maximum*); асимметрия распределения[3].

[2] Число и доля таких категорий со временем возрастает. Так, если в 2003 г. в выпуске в *JCR-SE* доли категорий со значением *Aggregate Cited Half-life (AgrCdHL)* и *Aggregate Citing Half-life (AgrCgHL),* составляли 4,7% и 11,2 (%), то в 2018 г. эти показатели достигли 13,4% и 16,3% соответственно. Аналогичная, но еще более ярко выраженная картина наблюдается и для выпуска *JCR-SSE*: если в 2003г. доля таких категорий для *AgrCdHL* и для *AgrCgHL* составляла 14,8% и 16,7% соответственно, то в 2018 г. эти значения достигли 43,1% и 41,3%.

[3] *Среднее значение i-го распределения* (*mv – mean value*) – сумма значений всех элементов *i*-го распределения, деленная на число этих элементов – является аналогом математического ожидания в теории вероятностей. В нашем случае элементы – это значения *CdHL/CgHL* (годы), а значения элементов – это число журналов, соответствующих конкретному году *CdHL /CgHL*. *Мода распределения (mode – Md)* – значение, которое в *i*-м распределении встречается наиболее часто. В нашем случае мода – это то значение *CdHL/CgHL* (в годах), которому соответствует наибольшее число журналов в этом распределении. *Медиана распределения (median - Mn)* – такое значение в рас-

Замечание. В дальнейшем в целях упрощения изложения в тех случаях, когда это не будет вызывать путаницы, мы будем упоминать только показатель *CdHL*, полагая при этом, что все соображения, определения и вычисления, которые приводятся далее в отношении показателя *CdHL*, в полной мере касаются и показателя *CgHL*.

Определение 2. Те журналы, каждый из которых в данном выпуске *JCR* охарактеризован значением *CdHL*, не превышающем 10 лет, будем называть *основными журналами распределения*, а ту часть распределения, которая соответствуют этим журналам, – *«Основной частью распределения»*.

Определение 3. Журналы, каждый из которых в выпуске *JCR* охарактеризован значением *CdHL*, превышающем 10 лет, будем называть *журналами хвоста распределения*, а ту часть распределения, которая соответствует этим журналам, – *«Хвостом распределения журналов» или просто «Хвостом распределения»*. Это определение в значительной степени совпадает с понятием «хвост распределения», которое обычно применяется в статистике в теории распределений. При этом предложенное здесь *ad hoc* определение хвоста распределения журналов, несмотря на частный характер, имеет свои достоинства – с его помощью мы всегда однозначно можем указать на начало хвоста распределения.

Определение 4. Распределение, у которого каждому журналу хвоста распределения в *JCR* для показателя *CdHL* приводится числовое значение, будем называть полным распределением, и, соответственно, распределение, у которого журналам хвоста в *JCR* приводятся текстовые выражения вида «>10.0», – усеченным (неполным).

Важное замечание. Все предлагаемые далее приемы и методы ориентированы, прежде всего, на превращение усеченных распределений в полные. Следующий шаг – вычисление средневзвешенных значений *CdHL* для такого превращенного распределения. При этом мы не ставим перед собой задачу определения действительных числовых значений *CdHL* для тех журналов, у которых в *JCR* в качестве значения показателя *CdHL* указано текстовое выражение «>10.0». Еще раз подчеркнем: в данном случае речь идет не о журналах как таковых, а о совокупностях журналов и соответствующих им распределениях.

ВОЗМОЖНЫЕ ПРИЕМЫ ОЦЕНКИ ДИНАМИКИ ПЕРИОДА ПОЛУЖИЗНИ

Вернемся к особенностям исходных данных. Как уже отмечалось, при построении распределений, а также при вычислении средних/средневзвешенных значений *CdHL* и *CgHL*, неполные, по сути, текстовые выражения вида «>10.0» не удается корректно использовать для численной обработки, по крайней мере, без применения некоторых искусственных приемов.

Прием 1(способ). Каждому журналу и категории *WoS*, у которых для *CdHL*/*CgHL* в *JCR* вместо числовых значений указано «>10.0», припишем значения 10,1, что позволит при соответствующих вычислениях учитывать также те журналы, у которых рассматриваемые показатели имеют значения, превышающие 10 лет. Насколько такой прием смещает (естественно, в сторону уменьшения) вычисляемые значения показателей, можно оценить сравнив вычисленные двумя различающимися способами средневзвешенные значения[4] *CdHL* для одного и того же набора журналов, который, к тому же, соответствует одному и тому ежегодному выпуску *JCR*. В качестве такого набора используем выпуск *JCR* за 2018 г., где каждому журналу, в том числе и журналу с *CdHL*, превышающему 10 лет, приписано числовое значение.

При вычислении средневзвешенного значения *CdHL* этим способом каждому журналу, у которого *CdHL* превышает 10 лет, вместо реальных числовых значений припишем значение 10,1. При другом способе будем использовать реальные числовые значения *CdHL*, в том числе и для тех журналов, у которых *CdHL* превышает 10 лет[5]. Понятно, что чем больше доля журналов, для которых *JCR* приводит текстовые значения вида «>10.0», тем менее точными оказываются получаемые средневзвешенные значения *CdHL*/*CgHL*. Однако при отсутствии точных исходных данных и учитывая, что этот прием применяется к каждому ежегодному выпуску *JCR*, все же можно будет с определенной точностью определить тенденции изменения значений этих показателей. Действительно, в случае применения к каждому годовому выпуску *JCR* этого приема мы при вычислении вносим если не неизменную от выпуска к выпуску, то близкую по величине систематическую ошибку. При сопоставлении показателей в динамике на эту ошибку можно делать поправку, а в некоторых случаях её можно просто игнорировать. Остается добавить, что вычисленное значение систематической ошибки (0,91)[6] в случае выпусков *JCR-SE* является, скорее всего, её верхней оценкой. Дело в том, что именно выпуск 2018 г., для которого была осуществлена эта

пределении, что ровно половина из значений l_j в распределении больше или равна этому значению ($l_j \le Md$), а другая половина меньше или равна этому значению. *Полная ширина на уровне половины максимального значения* (*FWHM – full width at half maximum*) – разность между правой и левой координатами на оси абсцисс, при условии, что на оси ординат эти координаты соответствуют половине максимального значения в распределении. *Асимметрия распределения* – характеризует степень его отклонения от распределения симметричного. Если правый хвост распределения длиннее левого то говорят, что распределение характеризуется положительной (правой) асимметрией (*right asymmetry/positive asymmetry*), если левый хвост длиннее правого, то будем говорить, что распределение характеризуется отрицательной (левой) асимметрией *(left asymmetry/negative asymmetry*).

[4] Средневзвешенное значение *CdHL* вычислялось путем деления {суммы [произведений (числа журналов в данной группе *CdHL*) на (значение *CdHL* этой группы)]} на {общее число журналов в распределении}.

[5] В первом случае средневзвешенное значение CdHL = 8,94, во втором – этот показатель оказывается равным 9,85. Указанный прием приводит к тому, что вычисленное с его помощью средневзвешенное значение получилось заметно ниже того, которое вычислено с использованием реальных значений: 9,85–8,94=0,91.

[6] См. сноску 5.

оценка, характеризуется максимальным значением доли журналов, у каждого из которых *CdHL* больше 10 лет. Аналогичная ситуация характерна и для средневзвешенных значений *CgHL*.

К сожалению, при построении распределений журналов по показателям *CdHL/CgHL*, применение описанного приема не решает проблему с неточными данными. Действительно, непосредственное включение в распределение журналов, для каждого из которых указано «>10.0», оказывается невозможным. Такие журналы из соответствующих распределений приходится исключать. В противном случае на графике после значения «10» возникнет почти вертикальный отрезок с длиной, равной доле этих журналов. Таким образом, предложный *Прием 1* при построении распределений оказывается недостаточным, а также не отличается и особой точностью при вычислении средневзвешенных значений соответствующих показателей. Очевидно, что для более полноценного решения этих задач необходимо разработать более эффективный способ.

Прием 2 (способ). Попытаемся «преобразовать» усеченное распределение в полное. Усеченным будем называть распределение, из которого исключены журналы со значениями *CdHL/CgHL*, превышающими 10 лет, а полным распределение – из которого такие журналы не исключены, при условии, что каждому журналу сопоставлено числовое, а не текстовое значение указанных показателей. Для предполагаемого преобразования некоторого усеченного распределения в полное попытаемся использовать данные другого (реперного) распределения при условии, что оно само является полным. Для реализации этого приема попытаемся сформулировать рабочую гипотезу, предварительно введя некоторые необходимые определения и выполнив необходимый анализ ряда распределений журналов по значениям показателей *CdHL* и *CgHL*.

Определение 5. Группой журналов в данном распределении будем называть такую совокупность журналов, которые имеют одинаковые (совпадающие, равные) значения данного показателя, в частности, – равные значения *CdHL* или *CgHL*. Для краткости группу журналов просто группой. Например, группой являются все журналы, значения *CdHL* каждого из которых составляют 7,5 лет.

На основе этого определения и используя данные ежегодных выпусков *JCR*, по значениям *CdHL* и *CgHL* были построены соответствующие графики распределений журналов по этим показателям: на оси абсцисс отложены группы журналов (значения *CdHL* или *CgHL*), а по оси ординат – число (доля) тех журналов, каждый из которых характеризуется значением *CdHL/CgHL*, соответствующим этой группе.

Определение 6. Однотипными (распределениями, принадлежащими определенному классу распределений) будем называть распределения, соответствующие одному и тому же показателю. Так, однотипными распределениями являются все распределения журналов по значениям показателя *CdHL* – их отнесем к одному классу, а другой класс однотипных распределений представляют распределения журналов по значениям показателя *CgHL*.

Особенности распределений *CdHL* и *CgHL* и их динамику иллюстрируют графики на рис. 1 и 2. Для того чтобы на одном рисунке можно было совместить и сопоставить несколько различных графиков распределений, шкала значений по оси ординат приводится в долях числа журналов от их общего числа. На рис. 1 представлены распределения, соответствующие и *CdHL,* и *CgHL*, т.е. распределения, принадлежащие двум различным классам. Рассмотрим распределения по *CdHL*. Несмотря на то, что эти распределения соответствуют разным тематическим выпускам *JCR* (*JCR-SE и JCR-SSE*) и различным ежегодным (2017 г. и 2018 г.) выпускам этого ресурса, по своей форме они близки. Ещё более ярко выражено сходство для распределений по *CgHL*, которые хотя и относятся к одному и тому же выпуску *JCR-SE* (2018 г.), однако соответствуют разным наборам журналов: одно из этих распределений построено с учетом всех журналов, а другое – с учетом только постоянно сохраняющихся журналов, т.е. тех, которые присутствовали в каждом выпуске *JCR-SE* в течение всего периода наблюдений (1997–2018 гг.).

Чтобы не затемнять изображение, на рис. 2 приведены лишь линии трендов, а не конкретные значения каждого из распределений по *CdHL/CgHL*. Судя по значениям R^2 (коэффициент детерминации – *determination coefficient*) линии трендов достаточно хорошо аппроксимируют реальные распределения (значения R^2 близки к 1). Из рис. 2 видно, что значения *Cited Half-life* и *Citing Half-life* со временем увеличиваются. В частности, заметно увеличение значений моды соответствующих распределений. Так, для случая распределений по *CdHL*: 2003 г. мода равна 5,8 лет; 2010 г. – 6,1 лет; 2017 г. – 6,4 года. Соответственно, для случая распределений по *CgHL* мода составляет: 2003 г. – 7,7 лет; 2017 г. – 8,3 года. Кроме того, наблюдается увеличение доли журналов (приведены вверху рис. 2 после значений R^2), у которых значения соответствующих полупериодов жизни превышает 10 лет. Для случая *CdHL* этот показатель составлял: в 2003 г. – 16,1%; в 2010 г. – 17,0%; в 2017 г. – 22,6%. В случае *CgHL* – в 2003 г. – 31,2%; в 2017 г. – 32,4%[7].

Суммируя результаты анализа, и сопоставляя графики, часть из которых представлена на рис. 1 и 2, можно заключить:

- формы графиков распределений по *CdHL* близки друг к другу;
- формы графиков распределений по *CgHL* близки друг к другу
- распределения журналов по *CdHL* и по *CgHL* характеризуются положительной (правой) асимметрией (*right asymmetry*);

[7] При рассмотрении графиков, изображенных на рис. 2, может сложиться ошибочное впечатление, что на долю журналов, у которых значения соответствующих полупериодов жизни превышает 10 лет, приходится всего несколько процентов. Это связано с тем, что на оси ординат отложены значения долей (%), рассчитанные исходя из общего числа только тех журналов, у которых значения соответствующих показателей не превышают 10 лет. Это значит, что журналы, у которых значения полупериода жизни превышает 10 лет, в распределениях, представленных на рис. 2 просто не учитывались.

- значения моды для распределения по *CgHL* заметно больше значений моды соответствующих распределений по *CdHL* (распределение по *CgHL* смещено вправо по отношению к соответствующему распределению по *CdHL*);
- величина *FWHM* в случае *CgHL* всегда меньше, чем в случае *CdHL*. Визуально это выглядит следующим образом: распределение по *CgHL* в заданном году всегда выше и уже распределения по *CdHL* в этом же году.

Помимо подобия графиков распределений, принадлежащих одному и тому же классу, можно отметить и различия между распределениями, принадлежащих к разным классам. Так, на рис. 2 видно, что значения моды у распределений *CgHL* существенно больше, чем в случае распределений *CdHL*, а значение *FWHM* на графиках *CgHL* меньше, чем этот параметр для графиков *CdHL*.

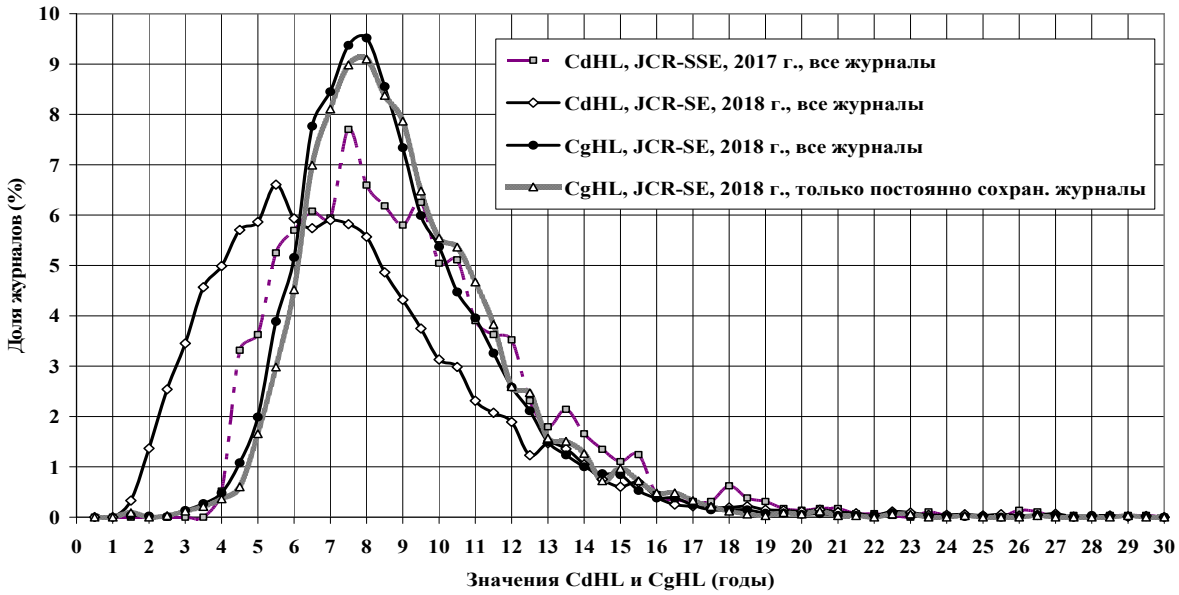


Рис. 1. Полные (не усеченные) распределения журналов по *Cited Half-life и Citing Half-life*
Две кривые, которым соответствуют первая и вторая строка легенды (считая сверху) соответствуют распределениям журналов по *CdHL,* другие две кривые (третья и четвертая строка легенды) – распределениям по CgHL.

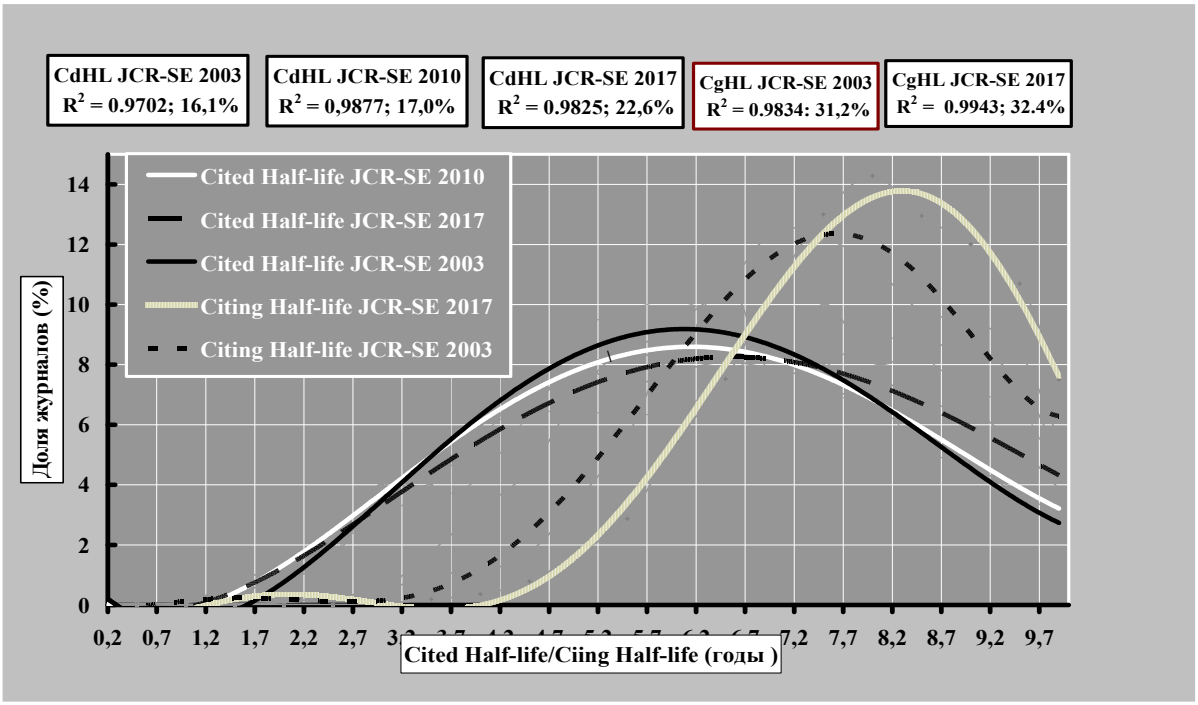


Рис. 2. Примеры усеченных распределений *Cited Half-life* и *Citing Half-life* за 2003, 2010, 2017 гг.

Исходя из вышеизложенного, можно сформулировать следующие утверждения.

Утверждение 1. Однотипные распределения, даже если они получены за различные моменты времени и на различающихся массивах журналов, по своей форме и основным характеристикам достаточно близки друг к другу.

Утверждение 2. Рассмотрим два распределения: *C* и *D*, и пусть наборы значений *CdHL* (группы журналов) в этих распределениях полностью совпадают. Далее, пусть распределение *C* содержит больше журналов, чем распределение *D*. Можно предположить, что в распределении, которое характеризуется бо́льшим числом журналов (распределение *C),* в некоторую группу попадет больше журналов, чем в такую же группу распределения, которое имеет меньшее числом журналов (распределение *D).*

Утверждение 3. Рассмотрим два распределения: *F* и *H*. Пусть общее число журналов распределения *F* равно общему числу журналов распределения *H,* а число журналов в какой-либо части (в основной части или в хвосте) распределения *F* больше, чем число журналов в соответствующей части распределения *H*. И пусть некоторая группа m_F находится в заданной части распределения *F*, а некоторая группа m_H – в такой же части распределения *H*. При этом потребуем, чтобы по значению показателя *CdHL* эти группы совпадали, т.е. выполнялось условие: $m_F = m_H$. Можно предположить, что при выполнении всех сформулированных нами условий и требований в том распределении, у которого в заданной части содержится бо́льшее число журналов, в группу m_F попадет больше журналов, чем в равную ей группу m_H того распределения, которое содержит меньшее число журналов в той же части (в распределении *H*).

Утверждение 4. Суммируя Утверждения 2 и 3, можно заключить, что чем больше журналов в некотором распределении и чем бо́льшая доля данной части (основной или хвоста) распределения, тем выше вероятность попадания журналов в ту или иную группу, находящуюся в этой части распределения. Другими словами – число журналов в некоторой группе (положительно) зависит от общего числа журналов в распределении, а также находится в некоторой положительной зависимости от того, какую долю от общего числа журналов занимает та часть распределения, в которой находится данная группа.

Рабочая гипотеза. Утверждения 1 и 4, которые, по сути, являются взаимно дополняющими, будем рассматривать в качестве рабочей гипотезы, которая в более компактном виде может быть сформулирована следующим образом: чем больше журналов, распределенных по значениям показателя *CdHL/CgHL,* и чем большая доля той части распределения, в которой находится заданная группа журналов с определенным значением соответствующего показателя, тем выше вероятность попадания журналов в эту группу. Другими словами: число журналов в некоторой группе (положительно) зависит от общего числа журналов в распределении, а также находится в некоторой положительной зависимости от того, какую долю от общего числа журналов занимает та часть распределения, в которой находится эта группа. Проверка и использование этой гипотезы (в случае подтверждения её справедливости), как мы надеемся, позволит с достаточной степенью точности, преобразовывать усеченные распределения в полные. Это, в свою очередь, даст возможность «восстанавливать» данные о числе (доле) журналов в группах, находящихся в хвосте соответствующих распределений (именно эта часть распределения нас, прежде всего, интересует, хотя, в принципе, положения рабочей гипотезы в равной мере распространяются и на основную часть распределений). Напомним, что под группой журналов мы понимаем численное значение (измеряемое в годах) показателя *CdHL/CgHL*, а под числом журналов в данной группе – количество журналов, каждый из которых характеризуется значением показателя *CdHL/CgHL*, соответствующим этой группе.

Определение 7. Реперным (базовым) будем называть распределение, которое является полным и с помощью которого предполагается дополнить распределение, являющееся неполным (усеченным). Основное требование к реперному распределению – все его группы должны иметь числовое значение, т.е. отсутствует группа с текстовым значением вида «>10.0».

Попытаемся формализовать рабочую гипотезу, чтобы с её помощью можно было рассчитывать число журналов в соответствующих группах. Рассмотрим пару, образованную двумя распределениями журналов по *CdHL*, каждое из которых соответствует одному из ежегодных выпусков *JCR*: распределение *G*, соответствующее выпуску *JCR* за год *g*, и распределение *I*, соответствующее выпуску *JCR* за год *i*. При этом потребуем, чтобы в качестве первого элемента в любой такой паре всегда выступало реперное (полное) распределение, а в качестве второго – усеченное, неполное[8], т.е. из рассмотрения будут исключены все пары, состоящие из двух неполных распределений. Причем формируя очередную пару распределений, мы каждый раз будем выбирать в качестве реперного одно и то же распределение – соответствующее 2018 г.[9]. Такое постоянство в выборе реперного распределения будет способствовать бо́льшей сопоставимости результатов.

Оценим общее число пар, которое, возможно, потребуется обработать для решения задач, поставленных в настоящей статье. Исходя из имеющихся у нас данных и, несмотря на сформулированные ограничения на состав пар, общее число таких пар может насчитывать многие сотни и даже тысячи. Поясним, на чем основывается эта оценка. Анализируемый нами период охватывает 1997-2018 гг., следовательно, обработке и анализу могут быть подвергнуты 22 ежегодных выпуска *JCR*. Получаем 21 пару. Каждый выпуск *JCR* состоит из двух изданий: *JCR-SE* (охватывает тематику по естественным, точным и техническим наукам) и *JCR-SSE* (охватывает тематику по обществен-

[8] Последнее требование не является обязательным: возможна пара из распределений, соответствующих выпускам *JCR* за 2018 г. и 2017 г. Последнее также является полным.

[9] При этом у нас сохраняется возможность выбрать в качестве реперного другое распределение, а именно то, которое соответствует 2017 г., (см. сноску 8).

ным наукам) – это значит, что полученное число удвоится и составит 42 пары. Поскольку речь идет о двух показателях (*CdHL* и *CgHL*), то это число снова удвоится и составит уже 84 пары. В каждом из ежегодных тематических изданий *JCR* мы выделяем две совокупности журналов: (1) совокупность всех журналов, содержащихся в данном годовом тематическом выпуске; (2) совокупность только тех журналов, которые присутствуют во всех без исключения соответствующих ежегодных тематических изданиях в течение всего 22-х летнего периода. В итоге получаем 168 пар, и это только те пары, которые можно сформировать, рассматривая совокупности журналов без учета того, каким тематическим категориям *WoS* соответствует тот или иной журнал. Если же сформировать совокупности журналов, соответствующие тем или иным наборам категорий *WoS*, а также совокупности журналов, соответствующие конкретным странам, то число пар возрастет на порядки. Столь обширный экспериментальный материал позволяет надеяться на получение достаточно надежных результатов, которые с высокой степенью вероятности позволит принять или отбросить рабочую гипотезу.

Преобразуем распределение G в распределение *I*.

a. Выберем в качестве реперного распределения G, а именно, то распределение, которое соответствует выпуску *JCR* за 2018 г. Этот выпуск содержит все (полные) численные данные о значениях интересующего нас параметра для каждого журнала. Нам известно общее число журналов N_g в распределении G и число тех журналов n_g, каждый из которых в этом распределении имеет значение *CdHL*, превышающее 10 лет (хвост распределения), а также известно число журналов l_g, образующих основную часть распределения, т.е. число журналов у которых значение *CdHL* не превышает 10 лет.

b. Выберем в качестве второго элемента пары такое распределение *I*, которое соответствует выпуску *JCR*, содержащему неполные данные (т.е. одному из ежегодных выпусков за период 1997–2016 гг.). Это значит, что в распределении *I* приводятся численные данные о значении *CdHL* только для тех журналов, у которых *CdHL*≤10, а против каждого журнала, у которого *CdHL*>10, указано только «*CdHL*>10» и не более того. Несмотря на указанную неполноту данных в распределении *I*, нам все же известно (как и в случае распределения *G*) общее количество журналов N_i и число тех журналов n_i, каждый из которых характеризуется значением *CdHL*>10 (хвост распределения), а также известно число журналов l_g, образующих основную часть распределения, т.е. число журналов у которых значение *CdHL* не превышает 10 лет.

c. Вычислим отношение общего числа журналов в распределении *I* к общего числа журналов в распределении *G*. Обозначим это отношение через $\alpha_{i/g}$:

$$\alpha_{i/g} = \frac{N_i}{N_g} \qquad (1)$$

Определим также соответствующие доли хвостов (β_g и β_i) в распределениях *G* и *I*:

$$\beta_g = \frac{n_g}{N_g} \qquad (2)$$

$$\beta_i = \frac{n_i}{N_i} \ . \qquad (3)$$

Следовательно, доля основных частей этих распределений составит: для G – $(1-\beta_g)$, а для I – $(1-\beta_i)$.

d. Вычислим коэффициент $k_{g\to i}$, который назовем коэффициентом преобразования (трансформации) распределения *G* в распределение *I*.

Для хвоста этот коэффициент будет выглядеть следующим образом:

$$k_{g_t\to i_t} = \alpha_{i/g} * \frac{\beta_i}{\beta_g} \ , \qquad (4)$$

где $k_{g_t\to i_t}$ – коэффициент преобразования в случае хвоста распределения, индекс t при символах g и i в выражении $k_{g_t\to i_t}$ указывает на то, что речь идет именно о хвосте распределения.

Для основной части коэффициент преобразования соответственно выглядит следующим образом:

$$k_{g_b\to i_b} = \alpha_{i/g} * \frac{1-\beta_i}{1-\beta_g} \ , \qquad (5)$$

где $k_{g_b\to i_b}$ – коэффициент преобразования в случае основной части распределения, индекс b при символах g и i в выражении $k_{g_b\to i_b}$ указывает на то, что речь идет именно об основной части распределения.

e. Вычислим функцию преобразования для каждой группы m_g распределения *G* в соответствующую группу m_i распределения *I*, т.е. для каждой группы m_i распределения *I* вычислим число журналов r_{m_i}, которое ей соответствует (в ней содержится). В общем виде эта функция выглядит следующим образом:

$$f(r_{m_g} \to r_{m_i}) = k_{g\to i} * r_{m_g} \ , \qquad (6)$$

где: r_{m_g} – число журналов в группе m_g распределения *G*;

r_{m_i} – число журналов в группе m_i распределения *I*;

$f(r_{m_g} \to r_{m_i})$ – функция преобразования числа журналов r_{m_g} в группе m_g распределения *G* в число журналов r_{m_i} в группе m_i распределения *I*, при усло-

вии, что обе эти группы по значению *CdHL* равны друг другу, т.е. $m_i = m_g$.

h. Для групп, принадлежащих хвосту распределения, функция преобразования выглядит:

$$f_t(r_{m_g} \to r_{m_i}) = \alpha_{i/g} * \frac{\beta_i}{\beta_g} * r_{m_g} , \qquad (7)$$

а для группы, принадлежащей основной части распределения, функция преобразования примет вид:

$$f_b(r_{m_g} \to r_{m_i}) = \alpha_{i/g} * \frac{1-\beta_i}{1-\beta_g} * r_{m_g} . \qquad (8)$$

Упрощая формулу (7), получим, что число журналов r_{m_i} в заданной группе m_i, находящейся в хвосте распределения *I*, рассчитывается по формуле:

$$r_{m_i} = \alpha_{i/g} * \frac{\beta_i}{\beta_g} * r_{m_g} . \qquad (9)$$

Аналогично, после упрощения формулы (8), для основной части получим:

$$r_{m_i} = \alpha_{i/g} * \frac{1-\beta_i}{1-\beta_g} * r_{m_g} . \qquad (10)$$

Приведенные вычисления необходимо сделать последовательно столько раз, сколько групп p_g содержит распределение *G*. Причем каждый раз в формуле (9) и, соответственно, в формуле (10) указывается то число журналов r_{m_g}, которое соответствует текущей группе m_g распределения *G*.

Для определения числа итераций для преобразования усеченных распределений в полные введем следующие обозначения: p_g – число групп в реперном распределении *G;* p_{g_t} – число групп в хвосте распределения *G*.

Отметим, что если в распределении *G* содержится p_g групп, то количество итераций вычисления по формуле (9) необходимо выполнить p_{g_t} раз, а по формуле (10) необходимо выполнить $p_g - p_{g_t}$ раз, а, в целом, для всего распределения такие вычисления должны быть выполнены p_g раз.

В результате этих вычислений вместо исходного усеченного распределения *I* мы должны получить теоретическое распределение I_{theor}, близкое исходному, при этом не совсем с исходным совпадающее. Основное отличие полученного распределения I_{theor} от исходного *I*, состоит в том, что I_{theor} является полным распределением, тогда как исходное – усеченным, т.е. эти распределения различаются своими хвостовыми частями. Однако и основные части этих распределений также будут различаться: расчетные значения далеко не всегда могут совпадать со значениями, соответствующими исходному распределению. Забегая вперед, отметим, что именно степень различия (совпадения) между основными частями распределений *I* и I_{theor} может служить одним из критериев для оценки степени справедливости принятой рабочей гипотезы. В общем, расчетное (теоретическое) распределения I_{theor} в некотором смысле является виртуальным. Действительно, при всех приведенных выше расчетах, мы требуем, чтобы и количество групп, и сами группы (значения *CdHL*) в распределении I_{theor} были равны числу групп и соответствующим им значениям *CdHL* реперного распределения *G*. Это требование вытекает из сформулированной нами рабочей гипотезы, точнее из той ее части, в которой утверждается подобие (близость) форм графиков различных (но однотипных) распределений. При этом сопоставляемые графики соответствуют различающимся распределениям, т. е. распределениям, полученным из разных ежегодных выпусков *JCR*.

Различие между реперным распределением *G* и распределением I_{theor}, «восстановленным» с помощью реперного и в результате использования формул (9) и (10), будет состоять только в различии числа (доли) журналов в группах.

Анализ результатов этих вычислений и построений должен подтвердить или опровергнуть рабочую гипотезу. В случае ее подтверждения нам останется рассчитать средневзвешенные значения *CdHL/CgHL* для соответствующих совокупностей журналов, полученных (совокупностей) по данным различающихся по времени опубликования того или иного ежегодного выпуска *JCR*. Средневзвешенные значения можно будет вычислить по следующей несложной схеме:

(a) умножить значение данной группы (значение показателя) на число журналов в этой группе;

(b) просуммировать полученные произведения (сумму взять по всему распределению);

(c) разделить результат действия *(a)* на результат действия *(b)*. Результат действия *(c)* и будет искомым средневзвешенным значением показателя *CdHL/CgHL*, которое характеризует данную совокупность журналов в данный момент времени.

Формализуя эту схему, получим:

$$W\_CHL_i = (\sum_1^p m_i * r_{m_i}) / \sum_1^p r_{m_i} , \qquad (11)$$

где W_CHL_i – средневзвешенное значение *CdHL/CgHL*, характеризующее совокупность журналов, соответствующую распределению *I*.

Полученные значения для каждого распределения журналов следует нанести на график, на оси ординат которого будут отложены средневзвешенные значения *CdHL/CgHL*, а на оси абсцисс – годы опубликования выпусков *JCR*.

ЗАКЛЮЧЕНИЕ

В настоящей работе поставлена задача разработки методики и способов определения динамики значения показателей периода полужизни *Cited Half-life* (*CdHL*) и *Citing Half-life* (*CgHL*). С этой целью по данным ежегодных выпусков «*Journal Citation Reports – Science Edition» (JCR-SE)* и «*Journal Citation Reports – Social Science Edition» (JCR-SSE)* за 22–х летний период (1997-2018 гг.) сформированы и соответствующим образом проанализированы массивы мировых журналов. Установлено, что для первых 20 ежегодных выпусков JCR (1997-2016 гг.) существует проблема неполноты (неточности) данных. Для этих выпусков каждого из журналов, у которого значение показателя CdHL (CgHL) превышает 10 лет, в JCR приводится не конкретное численное значение соответствующего показателя, а текстовое выражение вида «>10.0».

Для решения проблемы неполноты данных нами было предложено два метода, один из которых потребовал разработки рабочей гипотезы, состоящей в следующем. Чем больше журналов, представленных в виде распределения по значениям показателя *CdHL/CgHL,* и чем большая доля той части распределения, в которой находится некоторая группа журналов с определенным значением соответствующего показателя, тем выше вероятность попадания журналов в эту группу. Другими словами: число журналов в некоторой группе (положительно) зависит от общего числа журналов в распределении, а также находится в некоторой положительной зависимости от того, какую долю от общего числа журналов занимает та часть распределения, в которой находится эта группа. Напомним, что под группой журналов мы понимаем численное значение (измеряемое в годах) показателя *CdHL/CgHL*, а под числом журналов в данной группе – количество журналов, каждый из которых характеризуется значением показателя *CdHL/CgHL*, соответствующим этой группе. Предложенная гипотеза основывается на ряде полученных в настоящем исследовании результатов. В частности, на наблюдении, согласно которому однотипные распределения, даже если они получены за разные моменты времени и на различающихся массивах журналов, по своей форме и основным характеристикам достаточно близки друг к другу.

Последующая формализация этой гипотезы привела к разработке несложного математического аппарата, который в случае подтверждения справедливости предложенной гипотезы позволит «восстанавливать» недостающие данные и с помощью этих данных рассчитывать средневзвешенные значения показателей *CdHL* и *CgHL*. Что, в свою очередь, даст возможность определять тенденции в изменении этих значений для заданного набора журналов в заданные интервалы времени.

* * *

Авторы выражают глубокую признательность профессору Р.С. Гиляревскому за ряд ценных советов и критических замечаний, касающихся методики настоящего исследования, истории рассматриваемых показателей и их истолкования. Авторы также благодарны сотруднику компании *Clarivate Analytics* В.Г. Богорову за активное участие в обсуждении рассматриваемых в статье проблем и консультации, касающиеся структуры данных *JCR*.

СПИСОК ЛИТЕРАТУРЫ

1. Burton R.E., Kebler R.W. The “half-life” of some scientific and technical literature // American Documentation. – 1960. – Vol. 11, № 1. – P. 18-22.
2. Price D. General theory of bibliometrical and other cumulative processes // Journal of the American Society for Information Science. – 1976. – Vol. 29, № 5. – P. 292-206.
3. Line M.B. The “half-life” of periodical literature: apparent and real obsolescence / note by B.C. Vickery // Journal of Documentation. – 1970. – Vol. 26, № 1. – P. 46–54.
4. Мотылев В.М. Старение научно-технической литературы. – Л.: Наука, 1986. – 159 с.
5. Михайлов А.И., Черный А.И., Гиляревский Р.С. Основы информатики / изд. 2-е перераб. и доп. – М.: Наука, 1968. – С. 97–98.
6. Михайлов А.И., Черный А.И., Гиляревский Р.С. Научные коммуникации и информатика. – М.: Наука, 1976. – С. 173–176.
7. Либкинд А.Н, Маркусова В.А., Либкинд И.А., Янц М., Иванов К.Н. Моделирование динамики процесса сохранения журналов в качестве наиболее авторитетных научных изданий // Научно-техническая информация. Сер. 2. – 2013. – № 3. – С. 9-34; Libkind A.N., Markusova V.A., Libkind, I.A. Jansz M., Ivanov K.N. Modeling the dynamics of the retentivity process of journals among the most authoritative scientific serials // Automatic Documentation and Mathematical Linguistics – 2013 – Vol. 47, № 2. – P. 69–92

Материал поступил в редакцию 27.02.20.

Сведения об авторах

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

МАРКУСОВА Валентина Александровна – доктор педагогических наук, ведущий научный сотрудник ВИНИТИ РАН, Москва
e-mail: markusova@viniti.ru

ЛИБКИНД Илья Александрович – Ведущий программист, PerformIT, Москва
e-mail: anliberty@mail.ru

СПРАВОЧНО-ИНФОРМАЦИОННЫЙ РАЗДЕЛ

УДК 002:001.89

С.М. Гоннова

Национальные особенности публикационной активности ученых Китая, США, России

Рассматриваются некоторые особенности измерения эффективности публикационной активности учёных в Китае и США. Обсуждается дискуссионный вопрос внедрения Методики расчета качественного показателя государственного задания "Комплексный балл публикационной результативности", утвержденной 30 декабря 2019 года Минобрнауки России.

Ключевые слова: *публикационная активность, научный результат, результативность, эффективность науки, наукометрические показатели, система, методика расчета, нацпроект «Наука»*

DOI: 10.36535/0548-0027-2020-05-5

В национальном проекте «Наука» главный показатель – количество научных публикаций в изданиях, индексируемых в международных базах данных.

30 декабря 2019 г. заместитель Министра науки и высшего образования Российской Федерации С.В. Кузьмин утвердил (письмо от 14.01.2020 г. № МН-8/6-СК Минобрнауки России) Методику расчета качественного показателя государственного задания "Комплексный балл публикационной результативности" для научных организаций, подведомственных Минобрнауки России (далее – Методика) [URL: http://www.sib-science.info/ru/news/predlozhennaya-09022020]. А в середине января 2020 г. научные организации страны были проинформированы о начале мероприятий по корректировке государственных заданий на 2020 г. в соответствии с этой Методикой. Главное отличие новой Методики – переход от количественных показателей к комплексному баллу публикационной результативности (КБПР). В частности, вводится система определения качества статей через баллы, которые учитывают уровень журнала, в котором опубликована статья, если статья подготовлена коллективом авторов, то каждому соавтору засчитывается только доля от общего вклада.

Основная особенность утвержденной Методики – использование фракционного счета-метода, который разделяет вклад авторов и организаций в научный результат, и такой категории как качество журналов, которую определяет уровень цитируемости журнала. Методика учитывает различные типы изданий, включая публикации в журналах, индексируемых *Web of Science*, *Scopus*, статьи в журналах списка ВАК, а также монографии, зарегистрированные в Российской книжной палате.

При этом если посмотреть на проблему российской науки шире и глубже, то по мнению Главы профильного комитета Госдумы Вячеслава Никонова, российская наука и образование недофинансированы. Из этого можно сделать вывод, как бы Минобрнауки не оценивал науку, при неудовлетворительном финансировании данного сегмента, благоприятных результатов не будет [1].

Сегодня анонсировано, что национальная система "Индекса научных цитирований" с китайскими характеристиками и учетом международного опыта будет создана в Китае.

После многих лет принуждения китайских исследователей к публикации своих научных исследований в престижных международных журналах Министерство образования Китая и Министерство науки и техники Китая объявили о выпуске документов, направленных на ослабление "чрезмерной зависимости" от Индекса научного цитирования (*SCI*) документов академических акций, предложений о работе и финансирования научных исследований [2]. Отметим, что в настоящее время Китай занимает второе место в мире по научно-исследовательским статьям, опубликованным в международных журналах, уступая лишь США.

Согласно этим документам, импакт-фактор журнала (*JIF*) и Индекс научного цитирования (*SCI*) не должны использоваться организациями в качестве

наиболее важных критериев при наборе и продвижении персонала. Университеты и научно-исследовательские институты не имеют права предоставлять денежные средства для публикации в журналах, индексируемых *SCI*. Метрики, связанные с *SCI*, запрещены к использованию для составления рейтинга университетов или дисциплин.

«Меньше, но лучше» – новое правило публикации в нашей области, комментируют это китайские исследователи в социальных сетях.

В документах, изданных китайскими министерствами, выдвигается несколько критериев оценки различных типов исследований: прикладные исследования должны быть сосредоточены на фактическом вкладе результатов научного исследования в реальную жизнь, а не на количестве работ, опубликованных исследователем; в теоретических областях, не рассчитанных на немедленное внедрение, в списках литературы ученые должны приводить не более пяти репрезентативных работ, чтобы доказать ценность своих результатов, и, по крайней мере, три из них должны быть опубликованы в китайских журналах, если эти ученые хотят подать заявку на финансирование своих исследований на национальном уровне или награды [2].

В новых министерских руководящих документах отмечено, что университеты уделяли слишком много внимания *SCI*, а некоторые сделали наличие большого числа документов в *SCI* своим главным приоритетом. *SCI*, принадлежащий *Clarivate Analytics*, охватывает более 9 тыс. публикаций китайских ученых, в основном на английском языке. Этот показатель был одним из наиболее важных для определения уровня исследований в Китае в течение примерно двух десятилетий.

Этот шаг двух министерств, поддержанный Министерством финансов Китая, которое отвечает за финансирование исследований на национальном уровне, является радикальным изменением публикационной культуры страны. В первую очередь – это отказ от принципа «Публикуй или проиграешь», придерживаться которого принуждали ученых в Китае, чтобы они ориентировали свои статьи на международные издания.

В официальном заявлении *«Global Times»* подчеркнуто, что документы направлены на то, чтобы ослабить феномен *SCI*-превосходства, который признается односторонним, чрезмерным и известен среди специалистов библиометрии в университетах и академических институтах Китая как искажение информации.

Быстрым ответом на новую политику представления работ для опубликования стало то, что Национальный фонд естественных наук Китая изменил свои правила для ежегодного инновационного исследовательского группового проекта и Фонда для выдающихся молодых ученых в 2020 г. Кандидаты больше не обязаны при подаче предложений перечислять свои индексированные публикации с оценками цитирования.

Футао Хуан, профессор педагогики высшего образования в японском Университете Хиросимы, который изучал исследовательскую культуру молодых китайских ученых, выразил мнение, что отказ от приверженности к международным изданиям внесет огромные изменения в систему оценки исследований в Китае.

Новая публикационная политика запрещает университетам требовать от учащихся, аспирантов публиковать свои исследования в качестве условия для получения дипломов, поэтому публикационная нагрузка на авторов будет снижена. Старших и ординарных профессоров реализация новой политики освободит от дилеммы «Публикуй или проиграешь». Поскольку качество их публикаций превосходит количество, они смогут приложить больше усилий в высококачественных исследованиях, которым обычно нужно больше времени, чтобы получить инновационные результаты.

Дисциплинарные различия также очевидны. Социальные и гуманитарные науки имеют меньше международных связей и сетей, и английский язык не является для них такой основой общения, как для естественных наук.

Новая система оценки научной деятельности, ориентированная на метрические показатели, направлена на поощрение публикаций статей в китайских научных журналах, и, как правило, приветствуется исследователями в области социальных и гуманитарных наук. Один из социологов пишет, что это благоприятная политика, особенно для социальных наук: она игнорирует английские журнальные метрики и предлагает качественные методы оценки, такие как рецензирование и экспертизу. Исследователи прикладных наук также поддерживают новое движение, потому что они тоже боролись с доминированием системы *SCI* и импакт-фактором. Они признают новую политику как здоровое развитие науки и исследований в Китае. Например, врачи поддерживают эту новую политику, потому что она дает больше шансов продвинуться. Они не могут публиковать статьи, когда приходится делать по две операции в день.

Цель публикационной политики состоит в том, чтобы создать новую систему оценки наиболее актуальных для нужд Китая работ и специальных исследований, которые могут быть использованы для решения китайских проблем и ориентировать исследователей в Китае на вопросы, связанные с национальными приоритетами.

Ученые по-прежнему могут публиковать работы в ведущих международных журналах – таких как «*Nature*», «*Science*» и «*Cell*», – но исследования, которые появятся в менее влиятельных журналах, индексируемых в указателе *SCI*, больше не будут служить обязательным условием государственного финансирования (согласно руководящим принципам).

Нет никаких сомнений в том, что новая система оценки научной деятельности и новые требования к более широкой публикации в китайских журналах могут совершенствовать научные исследования Китая, а также развивать международное научное сотрудничество.

Если эта политика будет полностью реализована университетами и институтами Китая, то потребность китайских исследователей в публикации статей

в низкокачественных журналах будет быстро снижаться. Журналы, которые благодаря прежним метрикам имеют большее цитирование, будут проигнорированы, так как общественная репутация издания станет более важной для китайских исследователей при принятии решения о публикации. Хищные журналы потеряют свой рынок в Китае, потому что публикация в журналах из черных списков будет сурово наказываться. Авторитетные и топ-рейтинговые международные журналы по-прежнему лучший выбор, который будет осуществляться опытными руководителями исследователей, и конкуренция за публикацию в таких журналах станет ожесточеннее.

В соответствии с новой политикой система оценки научной деятельности должна быть ориентирована на оригинальность и научную ценность научных работ. Для лучших академических издателей, как китайских, так и международных, предоставление качественных услуг рецензирования всегда является важной стратегией выживания и развития. Новая политика поощряет исследователей публиковать или представлять свои наиболее важные работы в отечественных китайских журналах с международным влиянием и в материалах ведущих научных конференций.

Председатель КНР Си Цзиньпин во время национальной образовательной конференции в 2018 г. впервые заявил об отказе от международных научно-исследовательских изданий, а также о том, что академические стандарты в высших учебных заведениях Китая не могут существенно зависеть от западных идей или стандартов. В своем выступлении он подчеркнул, что Китай должен иметь свои собственные национальные академические стандарты и нормы.

Новые руководящие принципы предписывают не использовать показатели, связанные с *SCI*, для установления рейтинга университета или дисциплины при награждении, присуждении профессиональных званий, приглашении преподавателей, оценки их производительности и распределении ресурсов. Университетам запрещается использовать цитирование в *SCI* в качестве предварительного условия при наборе персонала. Академические учреждения больше не могут вознаграждать отдельных лиц и департаменты только на основе показателей *SCI*, а дипломы выпускников не должны ограничиваться указанием количества работ в журналах с высоким импакт-фактором и индексируемым в *SCI*.

В документах китайских министерств подчеркнуто, что будет создана китайская система "Индекса научных цитирований" с китайскими характеристиками и международным влиянием. Будут поощряться публикации об исследованиях, финансируемых государством, в высококачественных отечественных научно-технических журналах.

Новая система оценки научной деятельности, детали которой еще предстоит «обкатать», для исследований в фундаментальных дисциплинах должна быть сосредоточена на оригинальности и ценности научных работ, а не на их количестве и индексировании в *SCI*.

Руководством Китая предписывается что – исследования в области технологических инноваций должны быть сосредоточены на их фактическом вкладе в реальную жизнь, а не на количестве опубликованных работ. Университеты Китая не должны указывать публикации, индексируемые в *SCI,* в требованиях для аспирантов, чтобы получить докторские степени.

Китай является неотъемлемой частью мирового научного сообщества, и китайские исследователи все больше и больше участвуют в международных научных исследованиях. Развитие здоровой академической системы оценки научной деятельности в Китае принесет пользу всей международной экосистеме академической коммуникации и может изменить глобальную картину публикаций Международной ассоциации научных, технических и медицинских издателей (*International Association of Scientific, Technical and Medical Publishers – STM*).

В США, одном из мировых научных лидеров, нет фетиша по поводу публикаций. Проблема с имеющимися базами данных (*Web of Science*, *Scopus*), которые ставятся во главу угла в новой Методики Минобрнауки России, заключается в том, что они представляют собой динамичные массивы информации, в которых журналы постоянно включаются и исключаются из системы цитирования. Первоначально Институт научной информации США (*Institute for Scientific Information*) разработал индикатор влиятельности журнала (*Journal Impact Factor*), который послужил основой для отбора порядка 3,5 тыс. журналов, вошедших в указатель научного цитирования (*Science Citation Index*). После покупки Института научной информации фирмой *Thomson Reuters* в начале 1990-х гг. и появления базы данных *Scopus*, созданной компанией *Elsevier* в 2000-х гг., коммерческие критерии стали преобладающими, что привело к резкому росту количества индексируемых журналов. Когда ученый в США решает, куда конкретно отправить рукопись, он исходит из потенциальной аудитории, а не из того, в какой указатель цитирования входит журнал [3].

В американских университетах каждая кафедра самостоятельно решает, как оценивать работу преподавателей с учетом особенностей разных дисциплин. В отношении фундаментальной науки ситуация намного сложнее, поскольку существующие библиометрические индикаторы (количество публикаций и цитирований, индексы цитирования) вызывают дискуссии среди специалистов.

В настоящее время в России идет дискуссия о внедрении утвержденной Методики расчета качественного показателя государственного задания "Комплексный балл публикационной результативности" для научных организаций, подведомственных Минобрнауки России.

Счетная палата РФ выпустила доклад, в котором среди прочего призвала создать систему мониторинга научных исследований. Одновременно Министерство науки и высшего образования РФ утвердило Методику, которая предписывает научным организациям увеличивать публикационную активность. Минобрнауки России будет оценивать по Методике

выполнение госзаданий подведомственными научными организациями.

11.03.2020 г. в Минобрнауки России прошло расширенное заседание Рабочей группы по установлению единых требований к порядку формирования и утверждения государственного задания на проведение фундаментальных, поисковых и прикладных научных исследований за счет бюджетных ассигнований федерального бюджета. Главной темой заседания было обсуждение расчета комплексного балла публикационной результативности. Министр науки и высшего образования РФ В.Н. Фальков в ходе заседания сообщил, что Методика оценки выполнения подведомственными научными организациями государственных заданий должна быть скорректирована с учетом специфики различных наук и утверждена не позднее середины апреля [4]. Министр рекомендовал рабочей группе расширить состав, создав внутри секции по направлениям – гуманитарные и общественные науки, сельскохозяйственные науки, медицинские науки, а также секцию, которая будет заниматься публикационной активностью университетов.

Кроме вопросов о своевременности разработки национальных обоснованных методик оценки публикационной активности, траты денег из бюджета на публикации в западных журналах (а не на повышение качества российский научных журналов) возникает вопрос к Минобрнауки России по научному подходу к разработке и апробации Методики. А именно по срокам (начало разработки Методики – 2016 г., ФАНО России), по общей сумме потраченного бюджетного финансирования (за 4 года), по исполнителям и соисполнителям (разработчики по госзаказам и конкурсам), по масштабу опытного внедрения и по публичному обсуждению результатов исследований предварительного варианта Методики и т.д.

Возвращаясь к вопросу о главном показателе в национальном проекте «Наука» – количество научных публикаций в изданиях, индексируемых в международных базах данных, – будем надеяться, что выполнение плана по «комплексным» баллам в госзаданиях подведомственных Минобрнауки России научных организаций гарантирует выполнение показателей нацпроекта «Наука», а также обеспечение стране научно-технологический прорыва.

СПИСОК ЛИТЕРАТУРЫ

1. Колесников А. (специально для портала «Научные публикации»). НАУКОМЕТРИЯ: сторонники и противники. – URL: https://научные-публикации.рф/naukometriya/naukometriya-storonniki-i-protivniki (дата обращения: 15.03.2020).
2. Как Новая политика Китая может изменить публикационное поведение исследователей «Scholarly Kitchen» март 3, 2020. – URL: n.sspnet.org/2020/03/03/guest-post-how-chinas-new-policy-may-change-researchers-publishing-behavior/?informz=1 (дата обращения: 03.03.2020).
3. Костяев С. (преподаватель Университета Ратгерса, США). Уравниловка и утечка мозгов. Как в США оценивают продуктивность ученых. Новая Газета № 22 от 2 марта 2020. – URL: https://novayagazeta.ru/articles/2020/03/02/84133-uravnilovka-i-utechka-mozgov (дата обращения: 15.03.2020).
4. Заседание рабочей группы по расчету комплексного бала публикационной активности (прямая трансляция). – URL: https://www.poisknews.ru/science-territory/zasedanie-rabochej-gruppy-po-raschetu-kompleksnogo-bala-publikaczionnoj-aktivnosti-pryamaya-translyacziya/ (дата обращения: 11.03.2020).

Материал поступил в редакцию 19.03.20.

Сведения об авторе

ГОННОВА Светлана Михайловна – начальник отдела инноваций и перспективных разработок ВИНИТИ РАН, Москва
e-mail: gonnova@viniti.ru

УДК 007.52:004.896

А.М. Петрина

Функционирование и перспективы коллаборативных роботов: обзор инноваций

Представлен обзор результатов исследования состояния и перспектив применения коллаборативной робототехники. Определены понятия: робот, искусственный интеллект, киберфизическая система и коллаборативный робот. Основное внимание уделено разработкам коллаборативным роботам и их применению при модернизации производства. Показано, что основные направления развития робототехники ориентированы не только на автоматизацию производственных процессов, но и на цифровые технологии, подлежащие внедрению на передовых производствах, а также, что одним из основных компонентов производственных систем цифрового производства станут робототехника и роботизированные технологические комплексы, поддерживающие интеграцию технологических, технических, программных и других средств и систем, автоматизирующих этапы разработки и изготовления изделий машиностроения.

Ключевые слова: *роботы коллаборативные, роботизированные технологические комплексы, автоматизация, цифровое производство*

ВВЕДЕНИЕ

Коллаборативная робототехника – это новый этап развития робототехники, в основу которой заложен искусственный интеллект (ИИ) и киберфизические системы, стимулирующие своим появлением самые разнообразные виды деятельности человека. Учитывая большое количество научных работ и реальных проектов, отражаемых в Реферативном журнале ВИНИТИ «Робототехника» и связанных с разным толкованием понятия «коллаборативная робототехника», приведем уточнение этого понятия.

Базовые понятия коллаборативной робототехники – это «робот», «искусственный интеллект» и «киберфизическая система». Как правило, они смешиваются друг с другом, потому что провести границу между этими понятиями бывает достаточно сложно, но можно выделить ряд признаков, присущих каждому из них, которые позволяют провести границы между этими понятиями.

Основной отличительный признак робота то, что он представляет собой некое устройство или механизм. Данный признак позволяет отличать робота от программного обеспечения, с которым его очень часто отождествляют. Таким образом, робот – это механизм и ему присуще физическое начало. Кроме того, робот создан искусственным способом и, с биологической точки зрения, в нём отсутствует жизнь, однако он обладает способностью к анализу информации и взаимодействию с окружающей средой [1].

В отличие от робота, искусственный интеллект, главным образом, – это программа или алгоритм и его аппаратное воплощение не является главным. При этом такая программа позволяет анализировать окружающую среду и непрерывно самообучаться, т. е. мыслить, развиваться и действовать подобно человеку.

Ключевая особенность понятия «киберфизическая система» заключается в том, что эта система объединяет информационную и физическую составляющие, а также аппаратное воплощение. Иными словами, это понятие является родовым по отношению к понятию «робот». Например, разновидностями киберфизических систем являются беспилотные летательные аппараты, автономные высокоавтоматизированные транспортные средства и другие средства научно-технического прогресса.

Термин «коллаборативные роботы» (*collaborative robot* – коллективный робот) был введен для обозначения класса промышленных роботов, предназначенных для совместной и безопасной производственной работы с человеком в общей рабочей зоне. Концепция коллаборативных роботов (сокращенно – коботов) появилась в 1973 г. в научно-фантастическом фильме «Westwordl» [2]. Далее в рамках исследовательского проекта, возглавляемого Фондом *General Motors*, в 1995 г. возникла идея – создать в первую очередь коботов, безопасных для человека, работающего совместно с роботами в общей рабочей зоне в промышленном производстве, и помочь человеку решить сложные задачи, которые нельзя полностью автоматизировать [3]. Например, в нефтегазовой отрасли: в трубопроводной промышленности – оценка состояния инфраструктуры, охрана трубопроводов; в добывающей промышленности – поиск, разведка и контроль добычи; в промышленности по переработке

полезных ископаемых – контроль количества и качества запасов, мониторинг их состояния; в машиностроении – при точной сборке и контроле качества продукции. В результате коботы стали демонстрировать «умное» и безопасное поведение на производстве.

КОЛЛАБОРАТИВНАЯ РОБОТОТЕХНИКА

Использование коботов – это постоянно развивающееся перспективное направление, которое еще долго будет находиться на передовых позициях средств обследования и контроля в различных отраслях промышленности.

К любой коллаборативной робототехнической системе, независимо от её типа, предъявляются два основных требования: (1) для человека должна быть гарантирована полная безопасность со стороны робота, функционирующего в той же рабочей зоне, и (2) управление роботом должно быть настолько простым, чтобы можно было обойтись без специальной подготовки оператора.

Кобот отличается от классических промышленных роботов своими чрезвычайно удобными программными инструментами и интегрированными сенсорными функциями, которые определяют местоположение и интегрируются со складскими и другими системами. Из-за ручного обучения кобот внедряется в производственный процесс в течение нескольких часов с момента доставки. Сотрудники предприятия могут сами его запускать без привлечения системного интегратора, что сокращает стоимость внедрения.

В настоящее время более чем в 50-ти странах мира установлено свыше 8400 коботов. При их продвижении компании-производители успешно акцентируют свое внимание на потребности рынка в удобных в использовании коботах, которые могут работать вместе с людьми и имеют быструю окупаемость. Сегодня коботы работают совместно с человеком на многих предприятиях и внедряются в новые технологии. Результатом этого взаимодействия становится повышение гибкости управления, качества и производительности систем человек – робот. Коботы могут применяться на производстве в спектре таких задач, как сборка, обслуживание станков, сварочное производство, окрашивание, полирование, контроль качества, манипулирование, упаковка, тестирование, складирование готовой продукции, шлифование, сверление, логистика, включающая в себя оптимизацию интегрированного движения материалов, а также товаров и управление информационными потоками. Особенно успешно коботы применяются при упаковке товаров в картонные коробки и укладке их на поддоны. По данным Международной федерации робототехники, за период с 2007 по 2017 гг. применение коботов в промышленности возросло в 50 и 30 раз [4].

Расширенное внедрение коботов способствует обеспечению максимальной безопасности человека на производстве из-за перегородок или решеток. В отличие от промышленных роботов, коботы специально разработаны для совместной работы с людьми, и нет необходимости закрывать их в защищенном ограждении. На сегодня сложилось три подхода к решению проблемы охраны человека: 1) ограничение сил и моментов, развиваемых коботом; 2) немедленная контролируемая остановка кобота при контакте с человеком; 3) постоянный контроль скорости кобота и его расстояния от человека.

Конструктивно стандартная модель кобота поступает на производство в комплекте со встроенной камерой и интегрированной с ней системой технического зрения, которая используется в связке с мобильными решениями, и позволяет выполнять следующие задачи: обнаружение объектов, считывание штрих-кодов, классификация цвета, а также функции контроля и измерения, распознавания текста, передачи информации на другие уровни предприятия и т.п. Коботы легко интегрируются в многочисленные проекты. Установленные на мобильных платформах они становятся элементами инновационных логистических решений. Коботы могут привезти комплектные узлы и полуфабрикаты для сборочных станций, готовые для размещения на складе или на станциях контроля качества.

Совсем недавно использование беспилотных летательных аппаратов (БПЛА) в промышленности было достаточно новым явлением. Сегодня значительно удешевилась конструкция аппаратов, добавились новые их возможности, упростился процесс управления, усовершенствовалась технология сбора и обработки полученных данных. Использование беспилотников для нужд промышленности стало почти повсеместным. БПЛА состоит из летательного аппарата, представленного двумя типами: первым – это уменьшенная копия самолета, и второй, названный коптером, основанный на технологиях вертолетной тяги. Коптеры могут иметь от трех до восьми винтов. Наиболее распространенные модели – с четырьмя винтами, их называют квадрокоптерами. Главное отличие беспилотника самолетного типа от беспилотника вертолетного типа – высокая скорость и дальность полета. Зато коптеры вертолетного типа могут вертикально подниматься, зависать и вертикально снижаться. Вторая составляющая БПЛА – это оборудование, которое устанавливается на аппарат для получения необходимых пользователю данных. Помимо традиционных видеокамер на беспилотник устанавливают лазерные сканеры, ультрафиолетовые камеры, газоанализаторы, оборудование для расчета координат, а также другое оборудование, позволяющее пользователям решать свои специфические задачи.

Особенно быстро сейчас развивается недавно появившаяся технология использования защищенных беспилотных летательных аппаратов, устойчивых к столкновениям с препятствиями и безопасных при контакте с человеком [5]. Идея таких беспилотников основана на способности насекомых сохранять устойчивость после столкновения в полете. Беспилотники с функцией защиты позволяют решать важные проблемы, связанные с полетом в сложных и ограниченных пространствах, в опасных производственных объектах и в помещениях с людьми, при этом исключаются последствия столкновения и возможные травмы. Реализация принципа устойчивости и безопасности при столкновениях без применения систем распознавания препятствий и предотвращения столкновений с ними позволяет БПЛА, устойчивым к столкновениям, обеспечивать высокий уровень

надежности, требуемый в отраслях, где невозможен простой оборудования, который ведет к большим затратам. Поскольку полеты выполняются внутри производственных объектов, такому БПЛА не требуется согласования и получения полетных разрешений.

Коботы могут быть интегрированы с мобильными роботами, выполняющими общую задачу. Здесь выделяются роботы, действующие в экстремальных условиях, например, участвующие в спасательных и восстановительных операциях. Такой робот *Txplore* был разработан концерном *ABB* для проведения осмотров активных частей внутри масляных баков без слива масла [6]. Он совместим со всеми марками трансформаторного масла, а также с синтетическими жидкими диэлектриками. Радиоуправляемый робот механически и электрически автономен, оснащен собственным двигателем, системой освещения инспектируемого места, датчиками, передающей и принимающей радиосистемой и видеокамерой. Такой робот прошел опытную эксплуатацию на трансформаторном заводе в г. Сент-Луис (штат Миссури, США) и на одной из трансформаторных подстанций компании *American Electric Power*.

Ещё один пример использования кобота: компанией *Chevron* с помощью кобота *Flyability* выполнены осмотры резервуаров под давлением. Один из пилотов контролирует сам беспилотник, другой – прямую видеопередачу, чтобы обеспечить лучшее качество изображения места контроля даже в сложных ситуациях недостаточности освещения. Ранее для таких работ использовались робототехнические решения – механическая рука и роботы на магнитных колесах. Однако большинство из них имели ограничения при работе в локальном пространстве и при столкновении с препятствиями.

Устойчивость беспилотных летательных аппаратов и безопасность при столкновениях без применения систем распознавания препятствий и предотвращения столкновений с ними, позволяет таким беспилотникам обеспечивать высокий уровень эксплуатационной надежности, требуемый в отраслях, где простой оборудования невозможен или ведет к большим затратам.

Кобот *Elios* разработан швейцарской компанией *Flyability*, основанной в 2014 г. в Лозанне [5]. Этот защищенный от столкновений с препятствиями робот способен выдерживать столкновения на скоростях до 4м/с с твердыми поверхностями благодаря внешнему защитному каркасу, изготовленному по принципу пчелиных сот из углеродного волокна с мягким внешним покрытием. Ячейки этого каркаса могут легко заменяться в случае повреждений. Он предохраняет и защищает установленный на борту подвес, состоящий из двух видеокамер с разными частотой и разрешениями. Каркас предохраняет людей от травм при соприкосновении с лопастями винтов. Защитный каркас позволяет коботу перемещаться, катясь по поверхности объекта как колобок для контроля с близкого расстояния. Технология конструкции кобота *Elios* непрерывно совершенствуется. Так, по сравнению с базовой версией, в конструкцию внесены существенные изменения, позволяющие проводить визуальный контроль промышленных объектов быстрее, проще и эффективнее. Этот защищенный от столкновений с препятствиями кобот *Elios* пока еще высокотехнологичная новинка, но он уже испытан и взят на вооружение технологическими и добывающими мировыми компаниями *Shell*, *Chevron*, *Exxon Mobil*, *BP* и др. Его используют в тех местах, где обследование традиционными способами с помощью инспекции сложны, опасны и даже невозможны и ведут к большим затратам.

В качестве примера можно привести инспекцию надземных резервуаров для хранения углеводородов и химических веществ. Контроль проводился для нефтяных компаний *BR* и *Avia*. Обследовались резервуары высотой 25 м и диаметром 18 м. Работы проводились на высоте 25 м над землей, в условиях кромешной тьмы, в пространстве с опасной для здоровья атмосферой. А кобот *Elios* совершал осмотр одного резервуара за 5-10 рейсов, каждый из которых занимал около 10 мин. Было обследовано около 100 резервуаров, что позволило сэкономить значительные ресурсы и средства.

Коллаборативная робототехника осваивает сравнительно новое направление – многоагентные робототехнические системы и робототехнические комплексы военного назначения, которые могут качественно изменить современную технологию в ряде производственных отраслей. Для этого проводятся новые теоретические исследования, поскольку управление группой однородных и разнородных коботов принципиально отличается от управления одним «интеллектуальным» роботом. Коботы, образующие гетерогенную группу, обладают различными техническими возможностями. Они должны самостоятельно внутри группы самоорганизовываться, не допуская «конфликта интересов» и обеспечивая эффективное решение задачи группы в целом.

С развитием технологий обработки речи, изображений и видео, взаимодействие человека с робототехническими системами выходит на новый уровень и стремится стать похожим на общение человека с человеком. Интерфейсы взаимодействия человека и робота можно разделить на две категории: удаленные, использующие жесты и голос, и физические – такие как обучающие пульты, тактильные интерфейсы, а также системы обратной связи [7].

При использовании группы автономных необитаемых подводных аппаратов (АНПА) необходимо решить целый комплекс задач, связанных с информационным обменом между коботами. Так, в рамках проекта *WiMUST* [8], поддержанного Евросоюзом и посвященного использованию группы АНПА для ведения сейсморазведки морского дна, выделено три направления, в которых остро стоит проблема взаимодействия коботов: (1) навигация, (2) координация и (3) реконфигурация, что неразрывно связано с развитием алгоритмов коммуникации автономных необитаемых подводных аппаратов [9].

При этом существенно меняются и функции человека-оператора, который по существу переходит от управления одним мобильным роботом к управлению целыми коллективами роботов.

ЗАКЛЮЧЕНИЕ

Развитие робототехники и искусственного интеллекта может изменить жизнь людей, повысить эффективность и уровень безопасности работ в промышленности, сократить затраты, обеспечить качество услуг и не только на производстве, и в таких сферах, как транспорт, здравоохранение, спасательные операции, образование, сельское хозяйство, торговля и др. Применение роботов позволит избежать необходимости работать в условиях, угрожающих жизни и здоровью людей. Несмотря на преимущества, которые коботы привносят на производства, многие задачи в обозримом будущем предполагается выполнять с участием человека с его когнитивными способностями.

Анализ модернизации производства показывает, что основные направления его развития ориентированы не только на автоматизацию технологических процессов, но и на производство изделий с использованием цифровых информационных технологий (Индустрия 4.0) [10], где одним из ведущих компонентов являются робототехника и роботизированные технологические комплексы.

СПИСОК ЛИТЕРАТУРЫ

1. Яковлев К.С., Боковой А.В., Кашкин С.Ю. Анализ терминологических и содержательных аспектов понятий «Искусственный интеллект» и «Робототехника» в свете необходимости их правового регулирования // Всероссийский научно-практ. семинар «Беспилотные транспортные средства с элементами ИИ» (БТС-ИИ-2019, Казань, 22-24 мая 2019: Труды семинара. С.-Петербург, Институт информатики и автоматизации РАН. – Переяславль-Залесский, 2019. – С. 253-212.
2. Tilo Michal. Cobots sind langst unter uns // Maschinenmarkt. – 2019. – Vol. 125, № 20. – P. 28-32.
3. Пашина М. Зачем вашей компании Кобот? // Оборудование, разработки, технологии. – 2019. – №12(138). – С. 27-30.
4. Einfach automatisiert mit Cobots // Nechn. Logist. – 2019. – Vol. 59, № 3. – P. 42- 43.
5. Чихунов Д.А. Летающий кроулер. Противоударный дрон-робот для тепловизионного обследования труднодоступных и опасных объектов // Территория NDT.- 2019 апрель-июнь. – С. 55-57. (Реф.19.12-37.47).
6. Stapleton Jamie. Robotics redefine transformer inspection // Mod. Power Syst. – 2019. – Vol. 39, №1. – P. 22-24.
7. Рудианов Н.А., Хрущев В.С. Функциональный подход к проектированию специализированных робототехнических комплексов // Изв. ЮФУ. Техн. н. – 2019. – № 1. – С. 18-27.
8. Жданова М.М., Воронин В.В., Сизякин Р.А., Гапон Н.В., Егизарян К.О. Система управления коллаборативными робототехническими комплексами на основе методов бесконтактного распознавания действий человека // 4-я Международная научная конференция «Моделирование нелинейных процессов и систем», Москва, 15-17 окт., 2019: Сборник тезисов / Моск. Гос. Технол. Ун-т «Станкин». – М., 2019. – С. 83 – 84.
9. Мартынова Л.А. Обмен информацией между автономными необитаемыми подводными аппаратами при их взаимодействии в группе // Вопросы оборонной техники. Сер. 16. – 2018. – № 11-12. – С. 135-144.
10. Шукалов А.В., Заколдаев Д.А., Жаринов И.О. От индустрии 3.0 к индустрии 4.0: Обзор инноваций // Там же. – С. 153-159.

Материал поступил в редакцию 28.01.20.

Сведения об авторе

ПЕТРИНА Алла Макаровна – кандидат технических наук, научный редактор РЖ «Робототехника» ВИНИТИ РАН, Москва
e-mail: viniti@mach04.ru

ВНИМАНИЮ ЧИТАТЕЛЕЙ!
УНИВЕРСАЛЬНАЯ ДЕСЯТИЧНАЯ КЛАССИФИКАЦИЯ (УДК)

НОВОЕ ИЗДАНИЕ
УДК. ИЗМЕНЕНИЯ И ДОПОЛНЕНИЯ.
Выпуск 7

Содержание выпуска:

В настоящем электронном издании помещены **изменения и дополнения**, опубликованные Консорциумом УДК в выпусках 32 и 33 «Extentions and corrections to the UDC»:

ИЗМЕНЕНИЯ И ДОПОЛНЕНИЯ К ТАБЛИЦАМ ОБЩИХ ОПРЕДЕЛИТЕЛЕЙ

– Опубликованы изменения к **Таблице IG.** Общие определители времени

ИЗМЕНЕНИЯ И ДОПОЛНЕНИЯ К ОСНОВНЫМ ТАБЛИЦАМ УДК

Опубликованы изменения к классам:

– **2 Религия. Богословие**
– **33 Экономика. Народное хозяйство. Экономические науки**
– **582 Систематика растений**
– **551.7 Историческая геология.**

Для удобства пользователей издание открывает Общая методика применения Универсальной десятичной классификации.

Для подписки необходимо направить заявку по адресу:
125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН
Телефоны: 499-155-42-52, 499-155-42-85, 499-151-78-61
E-mail: typo@viniti.ru, feo@viniti.ru
http://www.udcc.ru

ВНИМАНИЮ ЧИТАТЕЛЕЙ!

ИЗДАНИЕ УДК

УНИВЕРСАЛЬНАЯ ДЕСЯТИЧНАЯ КЛАССИФИКАЦИЯ

АЛФАВИТНО-ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

в 2-х томах

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

Том I содержит АПУ от буквы А до Н;

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

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

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

Для подписки необходимо направить заявку для оформления счета по адресу:

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

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

E-mail: feo@viniti.ru

http://www.udcc.ru