

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

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

Издается с 1961 г.

№ 12

Москва 2019

ОБЩИЙ РАЗДЕЛ

УДК [002:620.9](470)–047.44

В.А. Маркусова, Л.Э. Миндели, Д.А. Рубвальтер, А.В. Золотова, М.А. Акоев,
А.Н. Либкинд

Динамика публикационной активности России в области энергии и топлива за 2008-2017 гг. по данным Web of Science и InCites^{*}

Представлен анализ динамики публикационной активности России в области энергетики и топлива (Energy&Fuels) за 2008-2017 гг. В качестве источников данных были использованы семь БД Web of Science Core Collection (WoS), аналитический инструмент InCites и национальная база данных Russian Science Citation Index – Clarivate (RSCI-C), расположенная на платформе WoS. Показано, что мировой поток по разделу Energy&Fuels (E&F) за 2008-2017 гг. увеличился в 3,1 раза, тогда как по всем разделам – только в 1,4 раза. Объем массива публикаций в мире по E&F составил более 406,2 тыс. единиц, а в России – 6,7 тыс. Несмотря на снижение ранга РАН среди лидирующих организаций по E&F в мире с 5-го в 2008 г. до 18-го в 2017г., доля её публикаций в журналах первого и второго квартилей составила 80% от её массива по этой проблеме в 2017 г. Значения показателей цитируемости публикаций России по E&F существенно ниже, чем соответствующие значения для лидирующих по E&F стран. Значительное различие было выявлено в распределении ведущих организаций по данным WoS-CC и RSCI-C соответственно. Негативным

^{*} Статья подготовлена при финансовой поддержке РФФИ: проекты 17-02-00157 и 17-02-00078

фактором является небольшое количество публикаций российских авторов с традиционными соавторами из Германии и США. Международное научное сотрудничество России в области E&F менее интенсивно, чем в целом по РФ.

Ключевые слова: публикационная активность, энергетика и топливо, индикаторы, *InCites*, *WoS*, международное научное сотрудничество

ВВЕДЕНИЕ

Научная политика России и широкое использование библиометрических показателей для оценки научных исследований стали объектом нескольких недавно опубликованных исследований [1-5]. 7 мая 2018 г. Президент Российской Федерации подписал Указ № 204 «О национальных целях и стратегических задачах развития Российской Федерации до 2024 г.». По мнению экономистов, сформулированные в Указе цели могут быть достигнуты и задачи решены «консолидацией и сплочением действий стратегического планирования, обеспечивающих развитие социально-экономической, научной и технологической сфер деятельности» [6]. Для этого в соответствии с Указом были разработаны и утверждены паспорта 13 национальных программ, в том числе паспорт Национальной программы «Наука». Среди показателей для оценки результатов были выполнения этой программы приняты три индикатора публикационной активности отечественных исследователей, основанные на данных информационной системы *Web of Science*. Россия, Саудовская Аравия и США входят в тройку крупнейших производителей сырой нефти. Россия – второй по величине экспортер нефти в мире. Официальная статистика свидетельствует, что на российскую нефтегазовую отрасль приходится четверть ВВП страны. Однако, если учесть ряд дополнительных факторов [7], то российская экономика оказывается еще более зависимой от углеводородов. Согласно оценке А. Tekingunduz [8], экспорт нефти и газа составляет 40% от общих доходов федерального бюджета России, экспортирующей в основном сырую нефть, что объясняется прежде всего значительной долей устаревшего оборудования для ее переработки. После экономического кризиса 2008 г. многие страны обратили свое внимание на исследования и разработки в области энергетических ресурсов, что выразилось в значительном росте числа мировых публикаций по энергетике и топливу уже в 2014 г. по сравнению с 2009 г. [9]. В 2009 г. министерством энергетики США (*Department of Energy*) была принята и ежегодно финансируется специальная Программа создания Центров передовых исследований по энергетике (*Energy Frontiers Research Centers – EFRC*). В деятельности этих центров выделено специальное направление *Basic Science* (фундаментальные науки), в рамках которого распределяются гранты по физике конденсированных сред и материалов, химии, геологическим наукам и тем аспектам физико-биологических наук, которые направлены на создание новых материалов и на разработку новых химических процессов. Это затрагивает практически все аспекты энергетических ресурсов – производства, преобразования, передачи, хранения, эффективности и минимизации отходов. Оценка результатов началь-

ного этапа создания *EFRC* за 2009-2014 гг. показала, что в этих центрах, несмотря на отсутствие формальных механизмов координации исследований, увеличилось соавторство между членами *EFRC*, появились новые авторы и повысилось качество исследований [10].

Цель настоящей работы – сопоставительный библиометрический анализ эволюции отечественных фундаментальных исследований по энергетике и топливу (*Energy&Fuels*) по сравнению с ведущими странами мира. Кроме того, были рассмотрены модели национального и международного сотрудничества (соавторства) в этой области исследований.

МЕТОДИКА ИССЛЕДОВАНИЯ

В качестве основных источников данных в настоящем исследовании были использованы информационные продукты *Clarivate Analytics: Web of Science Core Collection (WoS CC)*, состоящая из *Science Citation Index Expanded (SCI-E)*, *Social Science Citation Index (SSCI)* и *Art&Humanities Citation Index (A&HCI)*, *Conference Proceeding Citation Index-Science (CPCI-S)*, *Conference Proceeding Citation Index-Social Sciences&Humanities (CPCI-SSH)*, *Book Citation Index-Science (BKCI-S)*, *Book Citation Index-Social Sciences&Humanities (BKCI-SSH)*; аналитическая база данных *InCites*, в которую с 2016 г. входит БД *Journal Citation Reports (JCR)*. Отметим, что при подготовке первых трех из этих указателей (баз данных) рассматриваются 164 наименования отечественных журналов. В качестве дополнительного источника использована национальная БД *Russian Science Citation Index (RSCI-C)* – совместный продукт *eLibrary* и *Clarivate Analytics*. Эта БД аналогична другим национальным базам данных: *Clarivate's Korean, Chinese and Latino Citation Index*, размещенным на платформе *WoS*, однако их журналы-источники не индексируются для подготовки *WoS*, а также для указателя содержания журналов *Current Contents* [10]. В *RSCI-C* индексируются 628 наименований отечественных журналов с 2005 г. по настоящее время.

В качестве количественного показателя нами была принята публикационная активность, включающая число всех видов публикаций (страны, региона, организации) и широко используемая в библиометрических исследованиях. Хронологические рамки охватывают 2008-2017 гг. Классификатором направлений исследований была выбрана одна из классификаций *WoS*, известная как *Research Area*. Все документы в *WoS* классифицируются по 152 рубрикам. Опция «*Research Analysis*» была использована для классификации документов. В обширном библиометрическом исследовании [11] отмечался значительный рост отечественных трудов конференций и их практически нулевая цитируемость. Поскольку статьи и

обзоры являются наиболее информационно-значимыми видами документов, то для сопоставительной оценки качественных показателей научной деятельности, которыми являются разнообразные индикаторы цитируемости, из массивов публикаций по *E&F* в *WoS* и в *InCites* были выбраны только статьи и обзоры (*Articles&Reviews – A&R*). Мировой массив *A&R* по этому направлению составил 246152 единицы за 2008-2017 гг., а отечественный – 4237 единиц.

Для оценки качества исследований, результаты которых опубликованы в статьях, были использованы различные индикаторы цитируемости, в том числе: средняя цитируемость публикаций страны в определенной области (предметной категории) по сравнению со средней цитируемостью данной области в мировом потоке статей; доля международного сотрудничества в написании статей; доля публикаций, выполненных исследовательскими организациями и/или университетами совместно с промышленными предприятиями (доля научно-производственных публикаций)»; доля публикаций страны в журналах первого и второго квартилей данной категории; относительный импакт-фактор публикаций страны по отношению к мировому потоку статей; нормализованный показатель цитируемости публикаций страны; доля публикаций страны в 1% и 10% самых цитируемых публикаций. Для оценки влияния отечественной науки на мировую и для сопоставления России с другими странами по различным показателям цитирования была получена статистика из *InCites*. С целью нивелирования (сглаживания) колебаний годового числа публикаций данные были проанализированы за два пятилетних периода: 2008-2012 и 2013-2017 гг. Международное и национальное сотрудничество оценивалось по числу совместных публикаций отечественных и иностранных ученых. Особое внимание было уделено национальному научному сотрудничеству между регионами. К нему обычно относят публикации, в которых участвовали ученые из двух или более организаций, расположенных в различных регионах данной страны. В нашем случае, к такому типу сотрудничества относились публикации, в которых не менее двух соавторов работали в разных организациях из двух или более различных субъектов Российской Федерации.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ

Публикационная активность в мире по данным семи баз данных *WoS CC* выросла в 1,4 раза в 2017 г. по сравнению с 2008 г. (с 2,01 млн до 2,8 млн публикаций соответственно). В России этот рост был выше в 1,85 раза (с 34,4 тыс. до 63,8 тыс. публикаций). Этому росту способствовала государственная инвестиционная политика, направленная на стимулирование фундаментальных исследований в университетах. Более низкие темпы количества российских статей и обзоров по сравнению с темпами роста всего потока публикаций были отмечены в работе Х. Муда [11], посвященной сопоставительному анализу темпов роста российских публикаций в *WoS* и *Scopus* за

2000-2016 гг., в которой он предположил, что это произошло за счет высокого темпа роста числа российских трудов конференций в *WoS* за тот же период.

Мировой массив всех видов документов по *E&F* вырос в 3,1 раза с 21,2 тыс. до 65,2 тыс. за 2008–2017 гг. Темп роста статей и обзоров в мире по *E&F* был немного выше и составил 3,4 раза (с 11,7 тыс. до 39,9 тыс.). Массив российских публикаций по энергии и топливу за 2008-2017 гг. насчитывает 6,7 тыс., что составляет 1,65% от соответствующего мирового массива (406180 публикаций). В целом для всех видов российских публикаций по *E&F* увеличение за этот период составило 2,8, а для статей и обзоров – 2,1. Распределения видов российских и мировых публикаций по *E&F* достаточно близки: статьи – 61,5% (Россия) и 57,5% (мир); материалы конференций – 38,3% (Россия) и 37,9% (мир).

Тенденции роста мирового массива публикаций по *E&F* за 2008-2017 гг. в зависимости от типа документа представлены на рис. 1. Сопоставительный анализ вклада стран, лидирующих в исследованиях по *E&F*, выявил значительные перемены в их позициях. Данные приведены в табл. 1 (страны расположены по числу публикаций в 2008 г.).

Как следует из табл. 1, КНР в 2017 г. вытеснила США и заняла 1-е место по количеству публикаций. Темп роста публикационной активности в США оказался ниже чем в России и в мире в целом. Ранг России незначительно понизился с 14-го в 2008 г. до 15-го в 2017 г. Самые высокие темпы роста продемонстрировали Саудовская Аравия, Норвегия, Польша и Малайзия. Значительно понизили свои ранги в 2017 г. такие страны как Франция, Турция, Греция, Дания, Сингапур. Произошли изменения и в рангах 50-ти лидирующих организаций по *E&F* в 2017 г. по сравнению с 2008 г. В табл. 2 представлен список 14 лидирующих организаций по публикациям в области *E&F* за 2008 и 2017 гг.

Китайская академия наук (*CAS*) с 354 публикациями (второй ранг) в 2008 г. вытеснила Министерство энергетики США (*DOE*) и заняла первое место в 2017 г. (2099 публикаций). Российская академия наук (РАН) занимавшая 5-е место (243 публикации) в 2008 г. снизила свою позицию до 18-го места (447 записей) в 2017 г. Темпы роста РАН (в 1,8 раза) были значительно ниже по сравнению с темпами роста Общества им. Гельмгольца (в 5 раз), *CAS* (в 5,8 раза), *Tsinghua University* (в 5,7 раза) и многих китайских университетов.

В структуре отечественной науки раздел *E&F* улучшил свои позиции, переместившись с 32-го места в 2008 г. на 19-е место в 2017 г. Российские публикации по *E&F* охватывали такие предметные категории *WoS*, как прикладная физика (*Physics Applied*), переработка нефти (*Oil Processing*), эффективность транспортировки нефти и газа (*Efficiency of Oil and Gas Transportation*), технические проблемы нефтехимии (*Engineering Petroleum Chemical*).

Динамика числа российских публикаций по *E&F*, заиндексированных в *WoS* (с учетом всех видов документов и только в *A&R*), и в *RSCI* в 2008-2017 гг., представлена на рис. 2.

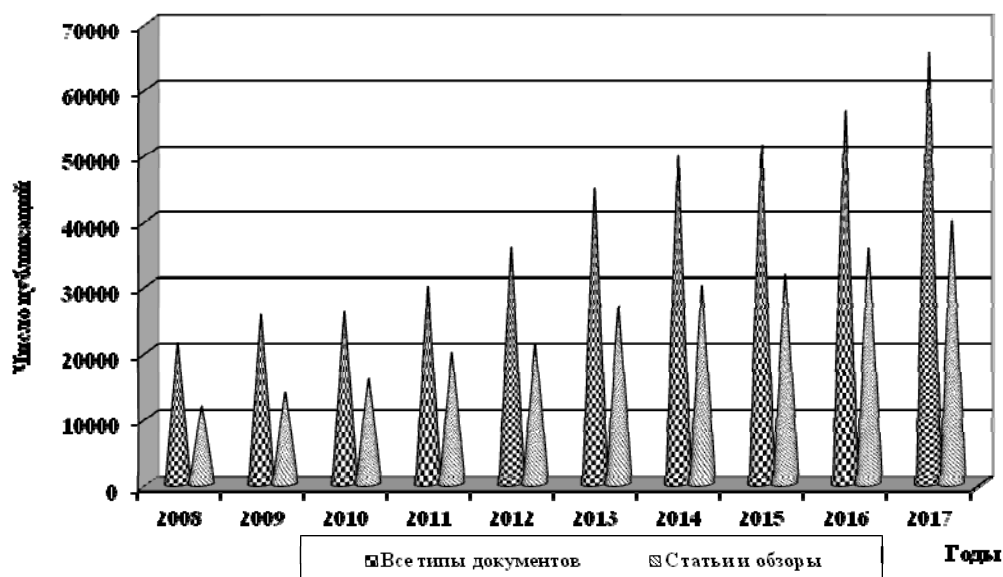


Рис. 1. Тенденции роста количества статей и обзоров по *E&F* в мире в зависимости от типа документа, *Web of Science*, 2008-2017 гг.

Таблица 1

Количество публикаций лидирующих стран по *E&F* в 2008 г. и 2017 г.

Страны	Ранги		Публикации		% от 21238	
	2008	2017	2008	2017	2008	2017
Мировой массив			21238	65226	100	100
США	1	2	3671	9136	17,29	14,01
КНР	2	1	2932	19667	13,81	30,15
Япония	3	10	977	2094	4,60	3,21
Индия	4	3	943	3624	4,44	5,56
Канада	5	11	887	1956	4,18	3,00
Германия	6	4	824	3258	3,88	5,00
Англия	7	5	683	2995	3,22	4,59
Франция	8	13	635	1868	2,99	2,86
Южная Корея	9	9	624	2103	2,94	3,22
Италия	10	6	596	2746	2,81	4,21
Испания	11	8	578	2152	2,72	3,30
Турция	12	20	574	949	2,70	1,46
Россия	13	14	539	1502	2,54	2,30
Тайвань	14	24	497	788	2,34	1,21
Иран	15	12	480	1901	2,26	2,91
Австралия	16	7	377	2266	1,78	3,47
Бразилия	17	15	369	1286	1,74	1,97
Швеция	18	18	258	1068	1,22	1,64
Нидерланды	19	22	251	840	1,18	1,29
Малайзия	20	16	238	1116	1,12	1,71
Греция	21	35	237	479	1,12	0,73
Польша	22	17	218	1068	1,03	1,64
Швейцария	23	26	177	779	0,83	1,19
Португалия	24	28	173	654	0,82	1,00
Таиланд	25	33	163	504	0,77	0,77
Дания	28	19	155	988	0,73	1,52
Сингапур	29	21	144	846	0,68	1,30
Норвегия	30	25	119	781	0,56	1,20
Саудовская Аравия	42	23	64	795	0,30	1,22

Таблица 2

Количество публикаций лидирующих организаций по E&F в 2008-2017 гг., Web of Science

Организации	Ранги		Публикации		% от 21238	% от 65226
	2008	2017	2008	2017	2008	2017
Мировой массив			21238	65226	100	100
Министерство энергетики США (DOE)	1	2	413	1521	1,95	2,33
Китайская академия наук (CAS)	2	1	354	2099	1,67	3,22
Centre national de recherches scientifiques (CNRS), Франция	3	5	304	810	1,43	1,24
Indian Institute of Technology, Индия	4	3	290	961	1,37	1,47
Российская академия наук	5	18	243	447	1,14	0,69
Tsinghua University, КНР	6	4	162	934	0,76	1,43
North China Electric Power University, КНР	7	11	160	598	0,75	0,92
University of California System, США	8	15	158	491	0,74	0,75
Shanghai Jiao Tong University, КНР	9	13	144	508	0,68	0,78
Общество им. Гельмгольца, Германия	10	8	133	672	0,63	1,03
China University of Petroleum, КНР	17	6	109	763	0,51	1,17
Xi'an Jiaotong University, КНР	26	7	90	746	0,42	1,14
University of Chinese Academy of Sciences, КНР	35	9	78	629	0,37	0,96
Tianjin University, КНР	14	10	117	615	0,55	0,94

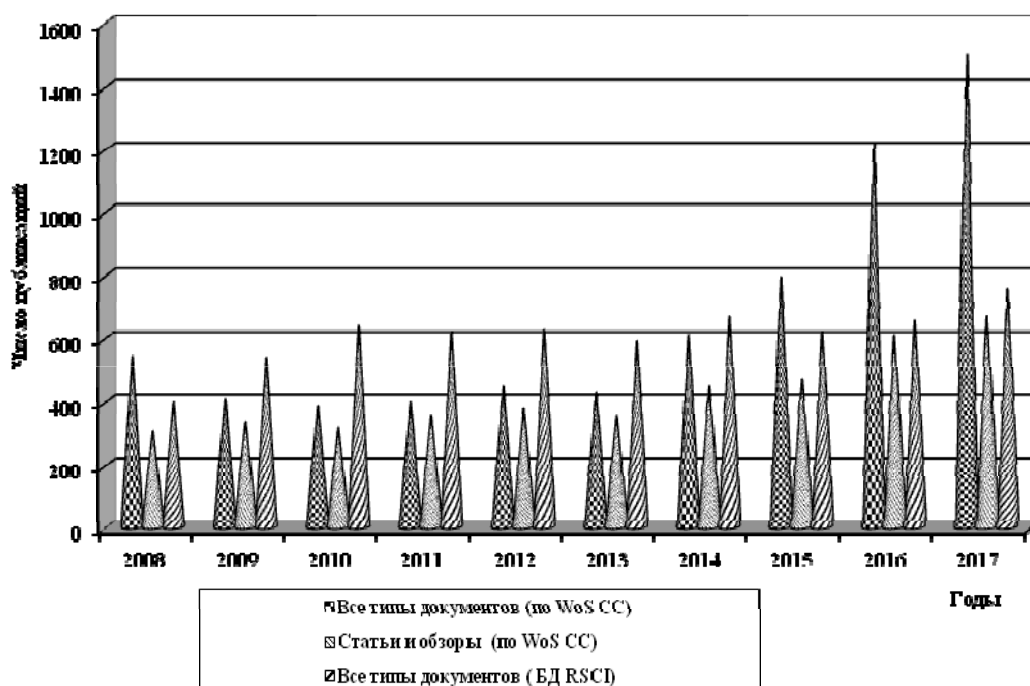


Рис. 2. Тенденции роста количества статей и обзоров российских авторов по E&F, WoS и RSCI-C, 2008-2017 гг.

Анализ показал, что список лидирующих организаций, полученный по семи базам данных *WoS CC*, достаточно слабо коррелирован с аналогичным списком, полученным по БД *RSCI-C*. По данным семи баз *WoS CC* Российская академия наук в течение всего исследуемого периода занимала лидирующие позиции по *E&F* в России, однако её доля в мировом потоке сократилась с 50,1% до 41,2%. Примечательно, что в 2017 г. 80% статей РАН по этой тематике было опубликовано в журналах первого и второго квартилей (по данным *JCR*, 2017). Это почти в два раза превышает значение аналогичного показателя для общероссийского массива публикаций – 42,2 %. Было также установлено, что после 2013 г. заметно вырос вклад национальных исследовательских университетов.

Как было отмечено, публикационная активность стран и организаций – это количественный

показатель. Качество научных исследований определяют различные показатели цитируемости. Рост мультидисциплинарных научных исследований и появление новых научных направлений привели к разработке нормализованного комплексного показателя цитируемости, независимого от области знания и типа документов (*Category Normalized Citation Impact – CNCI*). Статистика по значению этого показателя для стран, организаций и предметных категорий доступна на платформе *InCites* (<http://help.incites.clarivate.com/inCites2Live/indicatorsGroup/aboutHandbook/usingCitationIndicatorsWisely/normalizedCitationImpact.html>). Для того чтобы устранить влияние годичных флуктуаций цитируемости, нами были сопоставлены два пятилетних периода: 2013-2017 гг. (табл. 3) и 2008-2012 гг. (табл. 4).

Таблица 3

**Цитируемость публикаций стран,
ранжированных по числу публикаций по *E&F*, *InCites*, 2013-2017 гг.**

Название страны	Число публикаций	Нормализованный импакт-фактор	Доля публикаций в топ 1%	Доля публикаций в топ 10%	Импакт-фактор цитируемости	Доля научно-производственных публикаций	Доля публикаций в журналах Q1	Доля публикаций с МНС
Мировой поток	181422	1,23	1,99	17,92	18,16	3,36	54,59	23,29
КНР	51326	1,49	2,87	23,24	20,72	4,47	56,65	25,78
США	29927	1,42	2,99	21,36	21,27	4,39	57,45	42,76
Великобритания	10682	1,32	2,38	19,57	19,32	3,17	55,11	53,44
Индия	8977	1,08	1,21	15,06	16,75	0,91	57,18	21,83
Англия	8970	1,36	2,55	20,01	19,82	3,44	55,35	53,01
Германия	8952	1,34	2,48	19,12	19,61	5,91	60,89	47,10
Южная Корея	8472	1,28	2,04	17,78	19,28	5,29	62,76	28,32
Испания	7702	1,18	1,22	16,23	18,79	2,30	60,27	43,03
Канада	7317	1,17	1,87	16,77	17,54	3,62	50,94	46,34
Австралия	6958	1,50	2,77	22,10	21,88	2,14	58,02	57,69
Италия	6860	1,37	2,10	19,37	19,24	2,74	56,48	39,50
Япония	6563	1,25	1,92	15,42	18,72	5,93	61,70	42,31
Иран	6467	2,19	2,12	18,80	14,36	0,37	39,16	26,09
Франция	6073	1,16	1,73	16,37	17,31	7,67	56,70	52,69
Турция	3807	0,96	0,92	13,63	13,79	2,57	45,51	21,62
Тайвань	3741	1,04	1,34	12,99	16,93	2,32	59,30	25,50
Малайзия	3510	1,31	2,76	20,48	24,02	0,85	62,06	50,97
Швеция	3366	1,25	1,54	18,12	18,74	4,43	60,30	53,65
Бразилия	3350	0,91	1,60	10,51	13,10	4,87	51,67	34,60
Россия	2926	0,48	0,31	5,06	6,08	1,40	18,88	24,74
Голландия	2913	1,36	1,54	20,87	20,69	5,73	58,27	60,18
Дания	2621	1,31	1,91	21,02	19,40	5,04	54,72	53,22
Саудовская Аравия	2449	1,50	3,92	22,50	20,74	2,86	54,15	70,11
Сингапур	2273	2,26	7,26	34,49	35,95	2,60	72,59	58,51
Швейцария	1989	1,86	4,58	25,34	26,59	6,04	64,79	59,23
Норвегия	1613	1,01	0,68	12,83	13,52	21,02	42,93	53,19
Общий массив России	255814	0,83	0,87	6,55	4,78	0,97	26,54	30,23

**Цитируемость публикаций стран,
ранжированных по числу публикаций по E&F, InCites, 2008-2012 гг.**

Название страны	Число публикаций	Нормализованный импакт-фактор	Доля публикаций в топ 1%	Доля публикаций в топ 10%	Импакт-фактор цитируемости	Доля научно-производственных публикаций	Доля публикаций в журналах Q1	Доля публикаций с МНС
Мировой поток	125785	1,20	1,93	16,24	22,04	2,95	55,56	14,07
КНР	28531	0,95	1,39	13,15	18,20	3,20	59,67	13,30
США	24154	1,40	2,92	19,39	24,27	3,87	56,60	23,42
Германия	6178	1,64	3,50	20,62	21,75	5,83	55,15	34,23
Япония	6122	1,17	1,62	15,65	19,64	5,55	67,95	23,55
Канада	5357	1,39	2,28	18,98	24,73	2,80	50,61	32,44
Индия	5054	1,07	1,25	15,43	28,67	0,89	63,51	15,75
Англия	4977	1,57	3,11	22,00	28,29	5,73	47,64	39,18
Испания	4358	1,42	1,74	20,40	32,92	2,66	61,89	31,76
Франция	4195	1,60	2,86	21,24	25,67	7,75	57,39	39,83
Южная Корея	4128	1,13	1,38	15,60	27,07	6,78	76,57	22,84
Италия	3804	1,46	2,21	19,43	26,60	2,42	61,49	29,02
Австралия	3282	1,64	3,17	22,67	29,68	2,35	58,18	40,65
Тайвань	3177	1,12	1,83	14,60	24,17	2,17	71,68	14,51
Турция	3064	1,13	1,66	16,68	29,26	2,61	39,73	14,03
Иран	2959	0,99	1,18	14,19	18,29	0,44	33,74	15,55
Бразилия	2216	0,95	0,63	12,23	20,82	3,20	58,21	22,70
Россия	2171	0,42	0,28	5,57	7,39	1,06	18,74	19,12
Нидерланды	2008	1,78	2,99	25,40	31,20	5,48	57,44	44,82
Швеция	1826	1,89	2,63	23,55	30,41	5,64	59,01	43,54
Малайзия	1501	1,51	1,87	23,25	36,53	1,13	65,35	30,71
Дания	1368	2,40	5,92	31,43	43,76	5,85	59,17	45,25
Греция	1353	1,35	2,59	19,73	33,37	1,48	51,49	28,16
Польша	1344	0,72	0,60	7,96	14,70	1,79	52,17	20,83
Норвегия	1245	2,71	5,14	29,56	24,30	22,25	46,87	35,90
Общий массив России	176022	0,61	4,91	9,40	1,00	22,54	22,54	30,65

Сопоставление различных показателей цитируемости публикаций России по E&F показало незначительный рост за 2013-2017 гг. практически всех показателей, хотя величина комплексного нормализованного показателя *CNCI* России самая низкая (0,5) среди всех 25 стран. Этот факт, в частности, может быть связан с самой низкой долей отечественных публикаций как в журналах первого квартиля (18,88%), так и с самой низкой долей (0,27%) публикаций в топ 1% самых цитируемых статей рассматриваемых стран. Самые высокие значения *CNCI* имеют Сингапур (2,26) и Иран (2,19), показатель цитируемости которого выше, чем каждой из стран Большой Семерки. Отметим, что у обеих стран очень высокая доля публикаций в журналах первого квартиля (*Q1*): 72,6% у Сингапура и 61,7% у Ирана.

Рост международного научного сотрудничества является одной из особенностей глобализации. В об-

щем массиве публикаций России доля этого сотрудничества была стабильной и составляла около 30% в обоих обследуемых периодах. Хотя в массиве публикаций по E&F наблюдался рост доли сотрудничества до 24,7% в 2013-2017 гг. по сравнению с 2008-2012 гг. (18,5%), эта доля ниже, чем в общем массиве публикаций России. Многочисленные исследования показали, что международное соавторство из разных стран оказывает положительное влияние на цитируемость публикаций. Более низкая доля такого сотрудничества в исследуемой области также оказала влияние на более низкое значение комплексного показателя *CNCI*. Страны Большой семерки имели этот показатель в пределах 50% и выше. Известен факт, что страны, имеющие низкий научный потенциал, в последнее время усиленно развивают иностранное соавторство. Поэтому не удивительно, что у Саудовской Аравии зафиксирована самая высокая

его доля – свыше 70%. Этот фактор также способствовал высокому значению показателя *CNCI*.

Список стран, сотрудничающих по *E&F*, расширился за десятилетний период, но модель сотрудничества была спорадической. Необходимо подчеркнуть, что в течение последних двадцати лет США, Германия и Франция были ведущими нашими партнерами, доля каждой из этих стран в сотрудничестве с Россией составляла от 9% до 10% от общего массива публикаций. Однако в обследуемом массиве публикаций по *E&F* было установлено, что абсолютное количество совместных публикаций с пятью ведущими странами (США, Германия, Франция, Япония и Китай) было низким. Так в 2008 г. выявлено не более пяти-семи совместных публикаций с каждой из этих стран, а в 2017 г. – не более 10-15. Причина этого явления заслуживает отдельного исследования. Статистика международного сотрудничества по *E&F* в 2008-2017 гг. по данным *SCI-E WoS* приведена в табл. 5 (приведены данные по девяти странам, которые наиболее активно сотрудничали с Россией по *E&F*). По диагонали представлено общее число совместных публикаций авторов соответствующей страны с российскими авторами, а в клетке, находящейся на пересечении строки и столбца, – число тех публикаций, которые были подготовлены совместно учеными из соответствующих двух стран с участием авторов из России. Следует учесть, что табл. 5 составлена по данным только одной базы данных *WoS – SCI-E*, в которой по *E&F* за период 2008-2017 гг. заиндексированы 4127 российских публикаций.

Региональное научное сотрудничество является важным фактором передачи знаний в промышлен-

ность и для развития инновации. Его доля была ниже по сравнению с аналогичным показателем для всего массива российских публикаций (между 15% и 18%), представленных в *WoS* (24%) [12]. Сотрудничество между организациями было в основном двусторонним и стабильным в течение 2008-2017 гг. В среднем в каждой совместной (подготовленной в соавторстве ученых из разных организаций) участвовали 2,7 организации. Список регионов (субъектов РФ), участвующих в сотрудничестве, увеличился с 31 в 2008 г. до 43 в 2017 г. В течение этих десяти лет в этот список были вовлечены 63 региона. Девять регионов, лидирующих по количеству совместных публикаций по *E&F*, представлены в табл. 6. По диагонали указано общее количество публикаций по *E&F* данного региона, а на пересечении строк и столбцов, количество тех публикаций, которые были выполнены в соавторстве ученых из двух соответствующих регионов. Как и в случае международного сотрудничества табл. 6 составлена по данным только одной базы данных *WoS – SCI-E*.

Сотрудничество с промышленностью никогда не было сильной стороной российской науки. Однако в 2013–2017 гг. наблюдалась положительная тенденция усиления сотрудничества по сравнению с предыдущим пятилетним периодом: рост до 1,4% по сравнению с 1,1%. Этот показатель также выше, чем в целом по России – 0,89% в 2013-2017 гг. Норвегия демонстрирует самые сильные связи с промышленностью (около 21,1%, 27-й ранг), за ней следуют Франция (7,8%) и Нидерланды (6,8%) по данным *InCites*.

Таблица 5

Количество статей международного научного сотрудничества по *E&F*, 2008-2017 гг., *WoS*

	США	Франция	Германия	КНР	Англия*	Казахстан	Япония	Нидерланды	Италия
США	112	4	3	14	5	1	3	2	1
Франция		85	7	1	2	0	2	3	5
Германия			56	4	1	0	2	4	3
КНР				51	2	0	5	0	1
Англия*					38	0	0	2	1
Казахстан						36	0	0	0
Япония							32	0	1
Нидерланды								29	4
Италия									27

* Англия, Шотландия, Уэльс и Северная Ирландия в *WoS* представлены как отдельные страны, а объединенная статистика по Великобритании имеется только в *InCites*.

Количество статей регионального научного сотрудничества по E&F 2008-2017 гг., WoS

Регион	Москва	Новосибирская область	Московская область	Томская область	Республика Татарстан	Санкт-Петербург	Республика Башкирия	Кемеровская область	Свердловская область
Москва	1642	23	94	8	16	20	16	0	1
Новосибирская область		676	4	42	0	2	1	14	5
Московская область			345	9	5	2	1	1	4
Томская область				345	0	0	1	12	1
Республика Татарстан					209	1	3	0	4
Санкт-Петербург						204	1	0	4
Республика Башкирия							126	0	0
Кемеровская область								118	1
Свердловская область									111

Общее количество российских публикаций по E&F, проиндексированных в RSCI-C, составило 6,3 тыс. записей. Вклад сектора высшего образования был доминирующим (около 65,3%) среди организаций, проиндексированных RSCI-C в 2017 г. По данным этой БД ведущей организацией оказался Государственный университет нефти и газа им. И.В. Губкина (доля 4,8%). Доля научно-производственных статей в МГУ им. М.В. Ломоносова по E&F была низкой (около 1%), а среди всех российских организаций, проиндексированных в WoS в течение последних двадцати лет – 4,5%. Мы предполагаем, что исследователи из организаций РАН, как и из МГУ, предпочли публиковаться в журналах, индексируемых только в WoS. Этому частично способствовала политика Министерства науки и образования РФ, стимулирующая организации и исследователей к опубликованию результатов исследований в журналах первого и второго квартилей (Q1 и Q2) по данным Journal Citation Reports (JCR).

Существует значительное различие в списках российских организаций, опубликовавших статьи по E&F, между WoS и RSCI-C. Мощные российские нефтегазовые компании практически отсутствуют среди организаций, проиндексированных в WoS. Влияние промышленных агентств заметно в RSCI-C. Выявлена сеть региональных филиалов компании «Лукойл», которые расположены в регионах, где есть места разведки, добычи и переработки нефти. Массив публикаций по E&F в RSCI-C представлен 11-ю российскими научными журналами. Речь идет о журналах, которые ежегодно индексируются в RSCI в те-

чение всего анализируемого периода. Девять из этих журналов переводятся на английский язык, а два – публикуют только содержание, реферат и списки цитируемой литературы по-английски. Журнал «Вестник Южно-Уральского государственного университета» (русский) с 2012 г. вошел в систему Открытого доступа (Gold Open Access). Список журналов, опубликовавших статьи по E&F, приведен в Приложении.

ВЫВОДЫ

За период с 2008-2017 гг. только две области исследований мировой науки (по классификации Web of Science), а именно: «Наука, технологии и т.п.» (Science, Technology and Other Topics) и E&Fuel продемонстрировали поразительный рост (в 3 раза), тогда как ежегодное общее число мировых публикаций по данному 7-ми БД WoS CC увеличилось за этот период всего в 1,4 раза. Ежегодное число российских публикаций по E&F за этот период увеличилось в два раза. Россия значительно отстает по этому показателю от ряда стран, в частности, от Ирана (рост в 6 раз), от всех стран БРИК и Малайзии.

Сопоставление различных индексов цитируемости публикаций России по E&F показало незначительный рост за 2013-2017 гг. по сравнению с 2008-2012 гг., хотя величина комплексного нормализованного показателя CNCI для России самая низкая (0,5) среди всех 25 стран, лидирующих по числу публикаций. Этот факт, в частности, может быть связан с самой низкой долей отечественных публикаций как в журналах первого квартиля (18,88%), так и с самой низкой долей (0,27%) в топ 1% самых цитируемых

статей среди рассматриваемых стран. Несмотря на снижение позиции РАН среди лидирующих организаций по *E&F* в мире с 5-го ранга в 2008 г. до 18-го в 2017 г., её доля публикаций в журналах *Q1* и *Q2* составила 80%. Негативным фактором является более низкая доля (около 19%) международного научного сотрудничества по *E&F* в сравнении со стабильной долей (более 30%) по массиву российских публикаций в целом.

Мощные отечественные нефтегазовые компании практически отсутствуют среди организаций, проиндексированных в базах данных *WoS* *CC*. Напротив, национальная российская база данных *RSCI-C* позволила выявить сеть региональных филиалов нефтяной компании «Лукойл», сотрудники которых являются авторами публикаций, представленных в этой БД. Однако эти публикации индексируются в менее известных и информационно-значимых журналах. Список этих отечественных журналов стабилен и составил 11 наименований журналов.

СПИСОК ЛИТЕРАТУРЫ

- Abbott A., Schiermeier Q. Central Europe up-close // *Nature*. – 2014. – № 515. – P. 2-4.
- Radocevic R., Yoruk L. Are there global shifts in the world science base? Analysing the catching up and falling behind of world Regions // *Scientometrics*. – 2014. – Vol. 101, №2. – P. 1897–1924. DOI 10.1007/s11192-014-1344-1.
- Karaulova M., Abdullah G., Shackleton O., Shapira P. Science System Pass-Dependencies and their influences: nanotechnology research in Russia // *Scientometrics*. – 2016 - Vol. 100, № 3. – P. 365-383. DOI: 10.1007/s11192-016-1916-3.
- Turko T., Bakhturin G., Bagan V., Poloskov S., Gudym D. Influence of the program "5-top 100" on the publication activity of Russian universities // *Scientometrics*. – 2016 – Vol. 109. – P. 769-782. DOI 10.1007/s11192-016-2060-9.
- Гиляревский Р.С., Либкинд А.Н., Маркусова В.А. Динамика публикационной активности России в 1993-2017 гг. по данным *Web of Science* // Научно-техническая информация. Сер. 2. – 2019. – №3. – С. 1-13; Gilyarevskii R.S., Libkind A.N., Markusova V.A. Dynamics of Russian publications activity 1993-2017 based on Web of Science data // *Automatic Documentation and Mathematical Linguistics*. – 2019. – Vol. 53, № 2. – P. 51–63.
- Миндели Л.Э., Остапюк С.Ф., Фетисов В.П. Об организации долгосрочного прогнозирования фундаментальных и поисковых научных исследований // *Экономика и математические методы*. – 2019. – Т. 55, №1. – С. 56-67.
- Мовчан А.А. – URL: <https://carnegieendowment.org/experts/1057>.
- Tekingunduz A. How oil prices impact Russia's economy. – URL: <https://www.trtworld.com/europe/how-oil-prices-impact-russia-s-economy-22067>. Nov.29. – 2018.
- Маркусова В.А., Золотова А.В., Котельникова Н.А., Степанец О.М., Шухаева А.Н. Динамика в приоритетных областях развития мировой науки и стран БРИК // Научно-техническая информация. Сер.1. – 2016. – № 4. – С. 17-25.
- Smith A. M., Lai S.Y., Bea-Taylor J., Hill R.B.M., Kleinhenz N. Collaboration and change in the research networks of five energy frontier research centers // *Research Evaluation*. – 2016. – №4. – С. 472-485.
- Moed H., Markusova V., Akoev M. Trends in Russian research output indexed in Scopus and Web of Science // *Scientometrics*. – 2018 – Vol. 118, № 2. 1153-1180. – URL: <https://doi.org/10.1007/s11192-018-2769-8>.
- Миндели Л.Э., Иванов В.В., Либкинд А.Н., Маркусова В.А. Библиометрический подход к анализу национального научного сотрудничества на основе соавторства: Web of Science за 2006-2013 гг. // Научно-техническая информация. Сер. 1. – 2016. – № 8. – С.13–22.
- Иванов В.В., Маркусова В.А., Миндели Л.Э. Государственные инвестиции и публикационная активность вузов: библиометрический анализ // Вестник РАН. – 2016. – Т. 86, № 7. – С.611-619; Ivanov V.V., Markusova V.A., Mindeli L.E. Government investments and the publishing activity of higher educational institutions: Bibliometric analysis // *Herald of the Russian Academy of Sciences*. – 2016 – Vol. 86, № 4. – P. 314-321

ПРИЛОЖЕНИЕ

Список отечественных журналов, опубликовавших статьи по направлению исследований «Энергия и топливо», *Russian Science Citation Index-C*, 2008-2017 гг.

Наименование и язык опубликования журнала	Общее число публикаций	Доля в массиве публикаций
Общее число публикаций за 2008 -2017 гг.	6106	100%
Атомная энергетика (английский)	1250	20,40
Геология и геофизика и разработка нефтяных месторождений (русский)	1082	17,72
Известия РАН Сер. Энергетика (английский)	738	12,08
Химия и технология нефти и газа (английский)	684	11,20
Труды российского гос. ун-та нефти и газа им. И.В.Губкина (русский)	559	9,15
Геология нефти и газа (английский)	557	9,12

Наименование и язык опубликования журнала	Общее число публикаций	Доля в массиве публикаций
Технологии нефти и газа (английский)	495	8,10
Вестник Южно-Уральского гос. ун-та (русский)	384	6,29
Химия твердого тела (английский)	142	2,32
Теплоэнергетика (английский)	121	1,99
Кокс и химия (английский)	94	1,54

Материал поступил в редакцию 14.08.19.

Сведения об авторах

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

МИНДЕЛИ Леван Элизбарович – член-корреспондент РАН, доктор экономических наук, профессор, научный руководитель ИПРАН РАН, Москва
e-mail: l.mindeli@issras.ru

РУБВАЛЬТЕР Дмитрий Александрович – доктор экономических наук, профессор, ТЕМОЦентер, Москва
e-mail: dmitry.rubvalter@yandex.ru

ЗОЛотова Анна Витальевна – старший научный сотрудник ВИНТИ РАН, Москва
e-mail: korablikanna@mail.ru

АКОЕВ Марк Анатольевич - заведующий лабораторией по наукометрии Уральского Федерального Университета им. Б.Н. Ельцина, г. Екатеринбург
e-mail: m.a.akoev@urfu.ru

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

К проблеме доказательности медицинского диагноза: интеллектуальный анализ эмпирических данных о пациентах в выборках ограниченного размера*

Обсуждается возможность расширения представлений об обоснованности врачебных решений диагностического характера, принимаемых в рамках так называемой доказательной медицины. Предложен подход, позволяющий строить в процессе интеллектуального анализа накапливаемых эмпирических данных специальные – характеризующие каузальность возникновения диагностируемого эффекта – логические условия (характеристические функции), которые принимают значение «истина» на всех случаях наличия целевого эффекта и значение «ложь» на всех случаях его отсутствия в обучающей выборке прецедентов. Для решения этой задачи на расширяющихся последовательностях обучающих выборок используются: а) формальное уточнение понятия сходства описаний прецедентов как бинарной алгебраической операции и б) математическая техника порождения эмпирических зависимостей в стиле ДСМ-метода автоматизированной поддержки научных исследований. Особенности и возможности развиваемого подхода детализированы на примере решения задачи анализа причин и прогнозирования псевдопрогрессии опухолей головного мозга.

Ключевые слова: интеллектуальный анализ данных, доказательная медицина, сходство как алгебраическая операция, ДСМ-метод, диагностика, псевдопрогрессия опухолей головного мозга

ВВЕДЕНИЕ

Сегодня подход, определяемый как «доказательная медицина (Evidence-Based Medicine – EBM)», – это не только очевидный шаг вперед в повышении качества и результативности принимаемых медицинскими специалистами решений, но и стандарт *de facto* в обработке медицинских данных. Эффективность EBM-подходов по сравнению с известными уже ранее подходами и методами не вызывает сомнений. Однако на данный момент доминирующими в рамках EBM средствами анализа эмпирических данных и поддержки принятия решений являются статистические методы, которые, будучи, безусловно, результативными в широком спектре значимых приложений, тем не менее, не лишены в силу «природы» и специфики заложенных в них математических оснований ряда существенных недостатков. Так, с позиции естественных (нацеленных на успех соответствующих медицинских воздействий) требований к результатам, порождаемым статистическими мето-

дами анализа данных, вряд ли можно считать абсолютно неоспоримым положение дел с:

- *точностью* формируемого в результате анализа данных именно такими методами заключения (например, диагноза в случае конкретного пациента в его конкретном состоянии), характеризующего *вероятностью* ошибки¹;
- *персонализацией* диагностического заключения, т.е. ориентированностью выносимого заключения именно на конкретного пациента в его конкретном состоянии, а не только с фокусировкой на *статистически значимую группу* лиц, к которой может быть отнесен данный пациент;
- *неформальной интерпретируемостью* (объясняемостью в содержательных медицинских терминах, понятиях и категориях, а не в терминологии процедурно-алгоритмической схемы², используемой

¹ Так, в частности, с точки зрения теории вероятностей событие, имеющее *нулевую* вероятность, может тем не менее иметь место *любое конечное* число раз.

² Т.е. объяснениях, которые учитывают то, как устроен *реальный «мир»*, а не используемая для его анализа математическая модель.

* Работа выполнена при частичной поддержке Центра анализа и хранения больших данных МГУ им. М.В. Ломоносова

для математического анализа данных) результатов анализа данных, лежащих в основе соответствующего диагностического заключения.

ДОКАЗАТЕЛЬСТВО КАК НЕОСПОРИМОСТЬ И ОТКРЫТЫЕ ТЕОРИИ

Одним из подходов, позволяющих преодолевать обозначенные выше ограничения, является (хорошо известный в математике, в частности – в математической логике) вариант уточнения понятия *доказуемости* утверждения как демонстрации его *неоспариваемости* на доступных данных (см., [1, 2] и др.). В самом общем виде используемая здесь понятийная конструкция выглядит следующим образом. Если предметом исследования является *открытая* – допускающая пополнение новыми данными – *предметная область*, то для ее формализованного описания естественно использовать так называемые *открытые теории*³,

что на «процедурном» уровне позволяет:

- накапливать и описывать коллекции эмпирических фактов (базы фактов – БФ);
- рассматривать расширяющиеся, т.е. пополняемые новыми фактами, последовательности БФ;
- выделять эмпирические зависимости (ЭЗ) – истинные на имеющейся БФ формулы;
- выделять в множествах этих ЭЗ те, которые сохраняют свою истинность при расширении текущей базы фактов новыми данными;
- для каждого текущего состояния БФ строить эмпирическую теорию **Т(БФ)**, представляющую собой такое множество ЭЗ, истинных в данной БФ, что:

$$\mathbf{T(БФ)} = \{\mathbf{ЭЗ(БФ)}\},$$

причем такую, что каждый факт φ , имеющий место в данной **БФ**, является следствием в формируемой теории **Т(БФ)**:

$$\forall \varphi [(\varphi \in \mathbf{БФ}) \supset (\mathbf{T} \vdash \varphi)].$$

По-видимому, не требуется специальной аргументации утверждению, что медицина представляет открытую предметную область, для описания которой используются именно расширяемые коллекции эмпирических фактов, а выявляемые в ней эмпирические зависимости (в отличие, например, от физики, где доминирующую роль играют универсальные закономерности – так называемые *законы сохранения* и т.п.) имеют преимущественно *локальный* и *контекстно-зависимый* характер. Именно по этим причинам в медицине, в частности, при решении задач диагностического характера, представляется естественным для анализа эмпирических данных и поддержки принятия решений применять математическую технику открытых теорий. В рамках этого подхода существенным уточнением представленной выше понятий-

ной конструкции является использование так называемой *аргументационной* схемы установления (проверки) *неоспариваемости* формируемых в процессе анализа данных эмпирических заключений (ЭЗ), предполагающей формирование на имеющихся эмпирических данных как «доводов» в пользу конкретного заключения (аргументов «ЗА» – *pro*), так и «доводов» против этого же конкретного заключения (аргументов «ПРОТИВ» – *contra*), с дальнейшим сравнением «по силе» аргументов «ЗА» и «ПРОТИВ» и последующим принятием одного из четырех типов результирующего решения:

О положительное заключение, в ситуации, когда доводы «ЗА» доминируют над доводами «ПРОТИВ»;

О отрицательное заключение, если доводы «ПРОТИВ» доминируют над доводами «ЗА»;

О эмпирические противоречия (недостаточно исходных данных для *неоспариваемого* заключения), если доводы «ЗА» и «ПРОТИВ» равносильны по значимости;

О отказ от принятия решения (нет данных для принятия неоспариваемого заключения), если доводы как «ЗА», так и «ПРОТИВ» на имеющихся в текущей базе фактов эмпирических данных отсутствуют.

АНАЛИЗ ПРИЧИННЫХ ВЛИЯНИЙ И ЗАДАЧИ ДИАГНОСТИКИ

Рассмотрим один из вариантов реализации представленной выше процедурной схемы интеллектуального анализа медицинских эмпирических данных. В ее основу положен следующий комплекс неформальных соображений:

■ если на заданном прецеденте целевой исследуемый эффект имеется, то в его описании (среди значений признаков, характеризующих именно этот прецедент) непременно есть *необходимые факторы*, определяющие («вынуждающие» в смысле «причинности») возникновение этого целевого эффекта. «Значимые» (для характеристики целевого эффекта) комбинации необходимых факторов и описывают «причины» (или же их составные части, выражаемые в используемом языке представления знаний) возникновение изучаемого целевого эффекта;

■ «значимые» комбинации *необходимых факторов* (значений необходимых признаков) отражаются сходствами описаний прецедентов;

■ для представления таких «значимых» комбинаций естественно использовать максимальные по размеру, т.е. уже не пополняемые⁴ дополнительными признаками «замкнутые» множества *необходимых факторов*;

■ из всех порождаемых (причем отдельно – на примерах, т.е. прецедентах наличия целевого эффекта и контрпримерах, т.е. прецедентах его отсутствия) на исходной базе фактов значимых комбинаций *необходимых факторов* представляется естественным оставить лишь *непротиворечивые* комбинации необходимых факторов: такие *значимые сходства* (комбинации факторов), формируемые на примерах, которые не содержатся ни в одном из контрпримеров

³ Открытые (пополняемые при появлении новых данных) множества таких истинных (на имеющихся данных) формул, что любой из имеющихся на текущий момент в рассматриваемой предметной области факт может быть представлен как *логическое следствие* этого множества формул.

⁴ Разумеется, с точностью до используемого в текущей Базе Фактов перечня прецедентов.

(действительно, ведь наличие «причины» должно «вызывать» целевой эффект, а на контрпримерах баз фактов его нет).

Следует заметить, что данная процедурная схема учитывает содержательную «асимметрию» примеров и контрпримеров по отношению к существованию целевого эффекта: его наличие есть естественное следствие «причинного» влияния необходимых факторов. А вот его отсутствие может быть обусловлено двумя принципиально разными основаниями. Как *a*) наличием «контрвлияний» (например, результативной активностью иммунной системы пациента, эффективностью лечебных мероприятий и т.п.), так и *b*) положением, когда ни «влияний» *pro*, ни «влияний» *contra* на рассматриваемом прецеденте попросту нет (например, в ситуации, когда пациент здоров).

К сожалению, в подавляющем большинстве практически значимых случаев корректно разделить в текущей базе фактов ситуации типа *a* и *b* не представляется возможным. Как следствие, обучаться на всех контрпримерах (а не исключительно на выделенных в особую подвыборку случаях класса *a*) значит допустить внесение не имеющего содержательной интерпретации «шума» в результаты выполняемого «причинного» анализа.

Математическая формализация предлагаемой процедурной схемы интеллектуального анализа данных выполняется следующим образом:

- *сходство* (описаний) прецедентов формализуется как *алгебраическая операция* (см., например, [3–5] и др.). С помощью такой операции (выбирая лишь непустые результаты ее применения) определяется соответствующее отношение схождения на парах прецедентов текущей базы фактов;

- для каждого прецедента текущей базы фактов формируются *классы схождения* (т.е. множества всех прецедентов, сходных с ним в смысле задействованных операции и отношения схождения), а также *классы эквивалентности* (подклассы соответствующих классов схождения, определяемые фиксированными результатами применения операции схождения) на примерах и на контрпримерах. При этом каждый из классов эквивалентности характеризуется соответствующей комбинацией конкретных значений конкретных признаков – *факторов влияния*, с формальной точки зрения представляющей собою неподвижную точку некоторого замыкания Галуа [6], т.е. содержащей *все те и только те* атрибуты, которые взаимно-однозначно соответствуют именно этому классу эквивалентности прецедентов текущей базы фактов;

- непротиворечивость (см. выше) комбинаций признаков, которые характеризуют каждый класс эквивалентности (прецедентов анализируемой базы фактов), формируемый на примерах, обеспечивается выполнением специального условия *запрета на контрпримеры*. Это условие гарантирует, что представляющая сходство примеров в соответствующем классе эквивалентности прецедентов комбинация значений признаков – *факторов влияния*, обеспечивающих наличие целевого эффекта, – не «вкладывается» ни в один из сформированных на контрпримерах классов эквивалентности;

- найденные таким образом комбинации необходимых *факторов влияния* используются для формирования характеристических функций (ХФ) текущей базы фактов – таких логических условий, которые на всех примерах (прецедентах наличия изучаемого целевого эффекта) принимают значение «истина», а на всех контрпримерах (прецедентах отсутствия целевого эффекта) значение «ложь». Семейства ХФ, сформированные комбинациями необходимых факторов – инструмент описания «причинности» возникновения изучаемого эффекта, причем с точностью до структуры (информационной значимости, репрезентативности, ...) используемой при этом базы фактов;

- каузально корректное разделение исходной базы фактов средствами каждой из таких ХФ (непустота формируемого семейства ХФ) представляет условие (критерий) достаточности (репрезентативности, ...) текущей базы фактов для ее корректного описания «причинными» зависимостями необходимых *факторов влияния*.

Таким образом, предлагаемые характеристические функции могут рассматриваться как «индикатор» достаточности текущего «наполнения» базы фактов для представимости анализируемого целевого явления (диагностируемой патологии) уже обсуждавшимися выше средствами открытой теории $T(БФ) = \{ЭЗ(БФ)\}$, формируемой на накапливаемых эмпирических данных. При этом каждая из таких ХФ представляет собою лишь частично определенную функцию: ее значения зафиксированы лишь на прецедентах текущей БФ. На всех оставшихся комбинациях значений параметров (значений признаков, используемых для описания прецедентов) каждая такая ХФ может принимать значения «истина» или «ложь», причем какое именно – определяется лишь природой «причинности» анализируемого целевого явления. В процессе расширения текущей базы фактов описаниями новых прецедентов семейство характеристических функций, корректным образом описывающих как предыдущие, так и текущее состояние БФ, естественным образом сокращается за счет снятия неопределенности в выборе соответствующих истинностных значений на новых прецедентах. При этом очевидный интерес представляет выделение таких подсемейств ХФ, которые сохраняют корректность описания на всей формируемой в ходе реального исследования (и аккумулирующей его результаты) расширяющейся последовательности баз фактов.

Можно показать, что емкость семейства характеристических функций, формируемых на текущей базе фактов, в общем случае растет *экспоненциально* быстро с *линейным* увеличением числа прецедентов в текущем описании БФ. (Соответствующая переборная задача принадлежит классу #P как называемых перечислительно полных комбинаторных проблем [7–9]). Тем не менее, условие непустоты семейства характеристических функций, формируемых на данной БФ, может быть проверено быстрым алгоритмом, требующим объемов вычислений, полиномиально быстро растущих при линейном увеличении размеров исходной БФ. Наконец, при пополнении текущей базы фактов данными о новых прецедентах задача проверки непустоты подсемейства характери-

стических функций, сохраняющих корректность описания базы фактов и на предыдущих, и на текущем ее состоянии (не смотря на экспоненциально быстро растущую в общем случае емкость таких расширений) также является эффективно разрешимой: существует алгоритм ее реализации, требующий полиномиально быстро растущих объемов вычислений при линейном увеличении размеров БФ.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ И ЗАДАЧИ МЕДИЦИНСКОЙ ДИАГНОСТИКИ

Покажем, как представленная техника интеллектуального анализа данных работает в задачах медицинской диагностики. Сделаем это на примере идентификации и прогнозирования возникновения эффекта так называемой *псевдопрогрессии (ПсП)* опухолей головного мозга (см., например, [10–12]). С *псевдопрогрессией* обычно ассоциируют состояние, характеризующееся фиксируемыми посредством МРТ/КТ *временными изменениями* в опухоли и перифокальных тканях, сопровождающееся увеличением объема опухоли и степени ее контрастирования, возникающее, как правило, в ранние сроки после лучевого лечения, *спонтанно регрессирующие* без дополнительного противоопухолевого лечения. Назовем характеризующие *псевдопрогрессию* изменения, фиксируемые объективными средствами:

- усиление накопления контрастного вещества с расширением границы или без (T1 к/у);
- усиление отека в опухоли и перифокальных тканях (T2/Flair);
- увеличение объема опухоли за счет кисты и/или *солидного* компонента.

Некоторые примеры визуализации временных динамических изменений опухолей у пациентов с *ПсП* можно найти, в частности, в демонстрационных материалах к докладу авторов настоящей работы на Конгрессе «Информационные технологии в медицине» – ИТМ-2019 [13]).

С точки зрения анализа данных и поддержки принятия медицинских решений псевдопрогрессия специфична, прежде всего, отсутствием точных данных, которые позволяют надежно определять:

- сроки ее возникновения;
- длительность ее течения;
- стандарты *диагностики* как собственно псевдопрогрессии, так и ее *дифференциальной диагностики* с прогрессией опухоли;
- рекомендации по ведению пациентов с *ПсП*;
- *прогностические факторы* развития псевдопрогрессии.

Представляемое исследование эффекта псевдопрогрессии (*ПсП*) описанными выше средствами интеллектуального анализа данных было ориентировано, в первую очередь, на три последних (из пяти приведенных выше) цели – диагностика эффекта *ПсП*, включая дифференциальную диагностику с прогрессией опухоли, а также выделение так называемых *маркеров* развития *псевдопрогрессии*, позволяющих иметь *неоспариваемое обоснование* диагностического заключения о наличии *ПсП* у конкретного нового пациента.

Исходные данные (база фактов) выполненного исследования – это описания 410 пациентов Национального медицинского исследовательского центра нейрохирургии имени академика Н.Н. Бурденко (НМИЦ нейрохирургии имени академика Н.Н. Бурденко), накапливавшиеся в течение примерно 15 лет. Используемая база фактов содержит данные по 67 прецедентам эффекта *ПсП* (примерам в принятой нами выше терминологии) и 343 прецедентам его отсутствия (контрпримерам), каждый из которых характеризуется значениями (булевскими, числовыми, ...) более чем 150 параметров, при этом допускается отсутствие данных по некоторым параметрам у некоторых пациентов. Аккумулирующая имеющиеся эмпирические данные база фактов постоянно *расширяется*, однако в силу естественных причин с достаточно ограниченной скоростью – в несколько десятков описаний новых пациентов ежегодно. Таким образом, в части математических моделей и методов компьютерного анализа данных приходится иметь дело с обучающей выборкой *ограниченного* размера, однако при этом порождающей необходимость оперировать при ее изучении с реальными Big_Data-эффектами. Действительно, даже если ограничиться лишь булевым случаем представления данных и знаний при характеристике прецедентов, т.е. ситуациями, когда значимый для описания *ПсП* признак либо задействован (имеет «влияние»), либо нет, то придется рассматривать более чем 2^{150} вариантов комбинирования возможных значений задействованных параметров, выделяя из них *значимые* комбинации *факторов влияния* – скрытые в описаниях имеющихся в текущей базе фактов прецедентов *эмпирические зависимости*, которые и определяют (см. выше) возникновение целевого эффекта – *ПсП* опухоли.

Некоторые знания о природе возникновения псевдопрогрессии в рассматриваемой ситуации позволяет получить использование традиционного (для обработки медицинских эмпирических данных) статистического «инструментария». Так, например, в процессе анализа рисков развития *ПсП* установлено (см., в частности, [14–16]), что достаточно высокие для статистических средств анализа данных показатели достоверности характеризуют эмпирическую зависимость (корреляцию), связывающую на анализируемой базе фактов у пациентов *старше 11 лет* наличие кисты в опухоли, которая имеет *инфратенториальную локализацию*. Однако, подход, ориентированный на поиск статистических гипотез в процессе анализа множеств корреляций значений параметров в описаниях прецедентов текущей БФ, к сожалению, не снимает уже обсуждавшихся ранее (см. выше) вопросов:

- как быть с *точностью* и *персонализацией* диагностических заключений и рекомендаций, формируемых в процессе анализа возможных корреляций наблюдаемых значений параметров;
- как обеспечить содержательную *объяснимость* и неформальную *интерпретируемость* диагностических заключений, формируемых на базе выявленных корреляций?

$$F = F_1 \& [F_2 \vee F_3 \vee F_4]$$

где

$$F_1 = f_{145}(I) \& [(f_{137}(0) \& f_{138}(0) \& f_{147}(I)) \vee f_{141}(0)]$$

«покрывает» все **67 пациентов** с эффектом **ПсП** – см. (i) в Таблице

$$F_2 = f_{51}(I)$$

а **51_Киста (I)** – «покрывает» **56** из **67 пациентов** с эффектом **ПсП**

$$F_3 = f_{43}(I) \& f_{44}(0) \& \dots \& = f_{143}(7)$$

т.е. конъюнкция выделенных значений **16** признаков из п. ii в Таблице

$$F_4 = f_{42}(I) \& f_{43}(0) \& \dots \& = f_{132}(0)$$

т.е. конъюнкция выделенных значений **18** признаков из п. iii в Таблице

Пример характеристической функции в задаче диагностики псевдопрогрессии на данных НМИЦ нейрохирургии им. академика Н.Н. Бурденко.

Структура значений параметров характеристической функции, представленной на рисунке

(i) комбинация 5 признаков, покрывающая все 67 пациентов с эффектом ПсП :		
145 БСВ 0-1 (I)	137 Рецидив (0-нет) 138 Рецидив + МТС (0-нет)	147 БСВ 0-1ПП-2Рец+мтс (I) 141 жив\нет (0-да)
(ii) Сходство 16 признаков: 43 N очагов (1) 44 МТС (0) 51_Киста (1) 58 Мезенцефальный с-м Парино (огр.взор + в.Ну) (0 – нет) 60 Пираминая с-ка (0) 62 Чувствительные нарушения (0- нет) 72 Карновский (80-90) 72_1 Карновский, группы (1-2) 87 Катамнез да/нет (0 – нет) 120 Мезенцефальный с-м Парино (огр.взор + в.Ну) (0 – нет) 121 Мезенцефальный парез III н. (0 – нет) 122 Пираминая с-ка (0 – нет) 124 Чувствительные нарушения (0- нет) 134 Карновский (80-100) 135 Карновский, группы (1-2) 143 Событие (7 - псевдопрогрессия) «покрывает» 15 пациентов (номера 34, 52, 61, 149, 151, 177, 188, 247, 264, 291, 297, 317, 339, 343 и 366) из 21 пациента , оставшихся «непокрытыми» с помощью F₂		(iii) Сходство 18 признаков: 42 NF1 (0) 43 N очагов (1) 44 МТС (0) 56 Мост (гор.взор) (0 – нет) 65 OD (1 - норма 0,9-1,0) 66 OS (1 - норма 0,9-1,0) 67 Поля (1-Норма) 68 Атрофия ЗН OD (1 – Норма) 69 Атрофия ЗН OS (1 – Норма) 86 % уменьшения объема (0) 117 Бульбарный с-м (0 – нет) 118 Мост (гор.взор) (0 – нет) 119 Мост (VII н. ХаусБракман) (1 – норма) 127 OD (1 - норма 0,9-1,0) 128 OS (1 - норма 0,9-1,0) 130 Атрофия ЗН OD (1 – Норма) 131 Атрофия ЗН OS (1 – Норма) 132 Эндокринные функции (0-нет нарушений) «покрывает» еще 6 пациентов (номера 48, 82, 193, 149, 216, 267 и 365) из 21 пациента , оставшихся «непокрытыми» с помощью F₂

В свою очередь, процедурная схема интеллектуального анализа данных, использующая математическую технику характеристических функций, позволила формировать логические условия, принимающие одновременно значение «истина» на всех примерах и значение «ложь» – на всех контрпримерах анализируемой базы фактов так, что каждый шаг порождения соответствующей характеристической функцией (как выделение сходства описаний прецедентов обрабатываемой базы фактов) легко интерпретируем в терминах и понятиях анализируемой

предметной области. Именно этим – пошаговой объясняемостью и интерпретируемостью промежуточных выводов интеллектуального анализа данных – обусловлены также объясняемость и интерпретируемость конечных результатов такого анализа (конкретной структуры характеристической функции), формируемых на этой базе фактов. Одним из установленных эмпирическим путем результатов выполненного интеллектуального анализа данных оказалась общая для всех формируемых на данной базе фактов структура характеристической функции, в

которой объединяются две группы *факторов влияния*: наличие тех «причинных» влияний, которые «вынуждают» возникновение псевдопрогрессии (наличие своего рода аргументов *pro* в анализе достаточности оснований для принятия результатов выполняемого интеллектуального анализа данных – см. выше) и отсутствие тех факторов влияния, которые указывали бы на какие-либо отклонения от «характерного» для эффекта псевдопрогрессии (*ПсП*) развития событий (отсутствие актуальных аргументов *contra* в оценке достаточности оснований считать приемлемыми порождаемые результаты интеллектуального анализа данных – см., в частности, комбинации индикаторов нормального состояния для соответствующих параметров в примере характеристической функции в таблице).

Установленная эмпирическим путем непустота семейства характеристических функций, формируемых на анализируемой базе фактов, демонстрирует ее достаточную *представительность (репрезентативность)* для изучения эффектов *ПсП* и диагностики новых пациентов. Особого внимания при этом заслуживает *детерминистский, персонализированный и содержательно интерпретируемый* характер диагностических заключений, формируемых в рамках предложенной схемы интеллектуального анализа данных. Дополнительные возможности (в том числе, в части сокращения объемов комбинаторного перебора в процессе формирования характеристических функций) дает интеграция в рамках развиваемого здесь подхода статистических и детерминистских средств анализа эмпирических данных: значимые корреляционные зависимости между некоторыми факторами влияния могут быть использованы как своего рода «точка входа» в выполняемый в рамках предложенной схемы интеллектуального анализа данных процесс порождения классов эквивалентности описаний прецедентов (см. выше), обеспечивая отсечение, по крайней мере, части артефактов⁵ (содержательно не интерпретируемых, однако формальной корректных эмпирических зависимостей). Пример характеристической функции такого вида, дополняющий соответствующими детерминистскими заключениями уже приведенные выше соображения корреляционного характера относительно роли *кисты* и *локализации опухоли* в диагностике *ПсП*, приведен на рисунке.

Не менее интересна возможность для отсечения артефактов, характерных для интеллектуального анализа данных и машинного обучения, *содержательных* (неформально *интерпретируемых*) механизмов фальсификации (в смысле К.Р. Поппера [18, 19]) порождаемых эмпирических зависимостей. Примером такой процедуры может быть использование эффекта уменьшения опухоли прямо в процессе штатной серии сеансов лучевой терапии. Построенные на аккумулярованных в текущей базе фактов данных о та-

ком явлении соответствующие характеристические функции в этой ситуации представляют уже (фиксируемый объективными средствами) эффект уменьшения опухоли, рассматриваемый как контр-эффект по отношению к увеличению опухоли, наблюдаемому в случае псевдопрогрессии. Полученные в рамках анализа этого контр-эффекта характеристические функции могут быть использованы как естественный инструмент отсечения (по крайней мере – некоторых) артефактов в процессе проверки на фальсифицируемость получаемых результатов интеллектуального анализа данных о псевдопрогрессии.

* * *

Предложен и успешно апробирован при решении задачи диагностики псевдопрогрессии (в том числе, дифференциальной диагностики *ПсП* с прогрессией опухоли, а также – выделения прогностических факторов развития *ПсП*) новый подход к анализу эмпирических данных, расширяющий доступные на сегодняшний день в массовом порядке возможности доказательной медицины в части решения задач диагностического характера. Выделим наиболее значимые характерные особенности развиваемого подхода:

- возможности надежно (доказуемо⁶ корректно) оперировать в том числе и *малыми (статистически незначимыми)* выборками описаний прецедентов;
- *персонализация* формируемых в процессе интеллектуального анализа данных диагностических заключений и рекомендаций, аргументируемая точными детерминистскими средствами;
- возможности обеспечивать единую (общую) схему интеллектуального анализа данных для *расширяющихся* (постоянно пополняемых) новыми элементами выборок эмпирических данных;
- *пошаговая интерпретируемость*, неформальная (использующая содержательные термины и понятия исследуемой предметной области) *объясняемость* а также содержательная *фальсифицируемость* диагностических заключений, формируемых в рамках предложенной схемы интеллектуального анализа данных;
- эффективная (обеспечивающая практические возможности оперировать в условиях открытости – постоянной пополняемости новыми элементами – анализируемых коллекций эмпирических данных) *аргументационная схема* оценки достаточности оснований для принятия формируемых в процессе интеллектуального анализа данных диагностических заключений, в основе которой лежит неклассическая теория истины, ориентированная на работу в условиях характерного для *Big Data* эффекта *Open*⁷.

Предложенный подход к организации интеллектуального анализа эмпирических данных позволяет порождать практически значимые результаты при решении социально-значимых задач медицинской диагностики.

⁵ Неустранимость которых в процессе интеллектуального анализа данных и машинного обучения в общем случае продемонстрировали, в частности, К.В. Воронцов и Д.В. Виноградов (см. анализ эффектов так называемого переобучения, например, в [17] и др.)

⁶ В понимании *доказуемости* как *неоспариваемости* на имеющихся эмпирических данных.

⁷ Эффекта *незамкнутости* описаний анализируемой предметной области.

СПИСОК ЛИТЕРАТУРЫ

1. Есенин-Вольпин А.С. Об антитрадиционной (ультраинтуиционистской) программе оснований математики и естественнонаучном мышлении // Семиотика и информатика. – 1993. – Вып. 33. – С. 13-67. – URL: <https://istina.msu.ru/journals/504250/>
2. Многозначные логики и их применения: в 2-х т. / под ред. Финн В.К. – М.: УРСС Изд-во ЛКИ, 2008. – Т. 2: Логики в системах искусственного интеллекта. – 240 с.
3. Финн В.К. Индуктивные методы Д.С. Милля в системах искусственного интеллекта. Часть I // Искусственный интеллект и принятие решений. – 2010. – №3. – С. 3-21; Часть II // Там же. – №4. – С. 14-40.
4. Забейхайло М.И. О некоторых оценках сложности вычислений в ДСМ-рассуждениях Часть I // Искусственный интеллект и принятие решений. – 2015. – №1. – С. 3-17; Часть II // Там же. – №2. – С. 3-17.
5. Забейхайло М.И. О некоторых возможностях управления перебором в ДСМ-методе // Искусственный интеллект и принятие решений. – 2014. – № 3. – С. 3-21.
6. Кон П.М. Универсальная алгебра. – М.: Мир, 1968. – 359 с.
7. Simon J. On the difference between one and many // Lecture Notes in Computer Science. – 1977. – Vol.52. – P. 480-491.
8. Valiant L.G. The complexity of enumeration and reliability problems // SIAM J. Comput. – 1979. – Vol. 8. – P. 410-421.
9. Valiant L.G. The complexity of computing the permanent // Theoretical Computer Science. – 1979. – Vol. 8. – P. 189-201.
10. Нуднов Н.В., Желудкова О.Г., Мнацаканова И.В., Сидорова Е.В., Подоксенова Т.В., Шевцов А.И. Псевдопрогрессия у больного анапластической эпендимомой после лучевой терапии // Медицинская визуализация. – 2018. – № 2. – P. 18-24.
11. Hygino da Cruz L.C., Rodriguez I., Domingues R.C., Gasparetto E.L., Sorrensen A.G. Pseudoprogression and pseudoresponse: imaging challenges in the assessment of posttreatment glioma // American Journal of Neuroradiology. – 2011. – Vol. 32(11). – P. 1978-1985.
12. Parvez K., Parvez A., Zadeh G. The diagnosis and treatment of pseudoprogression, radiation necrosis and brain tumor recurrence // International Journal of Molecular Sciences. – 2014. – Vol. 15(7). – P. 11832-11846.
13. Трунин Ю.Ю., Голанов А.В., Костюченко В.В., Галкин М.В., Хухлаева Е.А., Коновалов А.Н. Псевдопрогрессия доброкачественной глиомы на примере пилоидной астроцитомы среднего мозга. Клиническое наблюдение // Онкологический журнал: лучевая диагностика, лучевая терапия. – 2018. – Т.1, № 1. – С. 94-97
14. Трунин Ю.Ю., Голанов А.В., Костюченко В.В., Галкин М.В., Хухлаева Е.А., Коновалов А.Н. Увеличение объема пилоидной астроцитомы среднего мозга: рецидив или псевдопрогрессия? Клинич. наблюдение // Опухоли головы и шеи. – 2016. – Т.6, № 1. – С. 68-75.
15. Trunin Y., Golanov A.V., Kostjuchenko V.V., Galkin M.V., Konovalov A.N. Pilocytic Astrocytoma Enlargement Following Irradiation: Relapse or Pseudoprogression? // Cureus. – 2017. – URL: <https://www.cureus.com/articles/3962-pilocytic-astrocytoma-enlargement-following-irradiation-relapse-or-pseudoprogression>.
16. Забейхайло М.И., Трунин Ю.Ю. О доказательности медицинского диагноза: интеллектуальный анализ эмпирических данных о пациентах в выборках ограниченного размера // Слайды к докладу на Конгрессе «Информационные технологии в медицине – ИТМ-2019» (Москва, 11 октября 2019 г.). – URL: https://itmcongress.ru/itm2019/agenda/O_dokazatelnosti_meditsinskogo_diagnoza_intellektualnyy_analiz_empiricheskikh_dannykh_o_patsientakh/
17. Виноградов Д.В. Вероятностно-комбинаторный формальный метод обучения, основанный на теории решеток: дис. ... д-ра физ.-мат. наук. – М.: ФИЦ ИУ РАН, 2018. – 131 с. – URL: http://www.frccsc.ru/diss-council/00207305/diss/list/vinogradov_dv
18. Popper K.R. The Logic of Scientific Discovery. – London–New York: Routledge Classics, 1959. – xxvii+513 p.
19. Popper K.R. Conjectures and Refutations. The Growth of Scientific Knowledge. – New York: Basic Books, 1962. – xii+412 p.

Материал поступил в редакцию 24.10.19.

Сведения об авторах

ЗАБЕЖАЙЛО Михаил Иванович – доктор физико-математических наук, заведующий отделом Федерального исследовательского центра «Информатика и управление» Российской академии наук, Москва
e-mail: m.zabezhailo@yandex.ru

ТРУНИН Юрий Юрьевич – кандидат медицинских наук, старший научный сотрудник Отделения радиологии и радиохирургии Национального медицинского исследовательского центра нейрохирургии им. академика Н.Н.Бурденко Минздрава России, Москва
e-mail: ytrunin@nsi.ru

Методические основы формирования информационного пространства и цифровых двойников объектов умного дома

Умный дом рассматривается как полиструктурная система, в которой происходят информационные и энергетические взаимодействия между её компонентами. Представлен анализ компонентов полиструктуры умного дома и энерго-информационные процессы, такие как энергопотребление, управляемость, надежность, экономичность, позволяющие в комплексе осуществлять эффективное управление этой полиструктурой, и перспективные методы оценки её функционирования.

Ключевые слова: умный дом, компоненты умного дома, информационные взаимодействия компонентов полиструктуры, управление компонентами полиструктуры, энергоэффективность

ВВЕДЕНИЕ

Энергетика нового поколения диктует свои условия перспективного развития традиционных энергосистем и образования новых рынков в сфере электроэнергетики [1, 2]. Новые энергетические рынки влияют на формирования у конечного потребителя иного подхода к использованию энергии в бытовых целях. Человек из пассивного участника энергетического рынка постепенно превращается в активного потребителя. Причина эволюционных процессов в поведении потребителей электроэнергии кроется в широком распространении технологий беспроводной связи, микроконтроллеров и энергоэффективной техники, способной интегрироваться в локальные компьютерные сети. Отдельного внимания при изучении энергосистемы бытовых потребителей, как основы жизнеобеспечения полиструктуры жилого пространства человека, заслуживает технология умного дома.

Идея умного дома будет способствовать развитию рынка B2C (бизнес для клиента), изменению топологии построения электрических сетей (децентрализация производства и энергопотребления), а также даст конкурентные преимущества развитию передовых технологий для построения современной энергосистемы. В настоящее время недостаточно развита методология построения системы умного дома. Такая система должна рассматриваться как открытая полиструктурная система с возможностью включения в нее новых компонентов. В умном доме появляются объекты, управление которыми рационально осуществлять посредством цифровых двойников. Поэтому развитие методологических основ создания инфор-

мационного пространства умного дома для управления цифровыми двойниками его объектов является целью настоящей статьи.

ПОЛИСТРУКТУРА УМНОГО ДОМА

Умный дом – многогранное понятие, в состав которого входит множество различных компонентов, отличающихся физическими свойствами, функциональным назначением, особенностями внутренней организации. Под умным домом будем понимать место проживания человека, объекты социальной и культурной жизни (общественные поликлиники, детские сады, школы и т. п.). При внедрении информационных технологий, используемых потребителем, возникает ситуация, связанная с финансовыми рисками. Такое состояние может существенно повлиять на эффективность работы всех компонентов системы. В настоящее время выделяют два направления отбора парадигм построения системы умного дома [3]: вследствие разрушения непригодной парадигмы; при чрезвычайно быстром прохождении через неустойчивое состояние инфраструктуры. В первом варианте, когда складываются условия, приводящие к кризису существующей концепции, выход из него осуществляется принятием радикальных мер, а именно полным вытеснением старой парадигмы новой. Во втором варианте замена концептуальной схемы построения инфраструктуры происходит по мере накопления знаний и в силу научно-технического прогресса. Принимая во внимание величину финансовых вложений в инфраструктуру умного дома, можно сделать вывод, что второй вариант развития более предпочтителен. В

связи с этим анализ функционирования энергообеспечения умного дома и его влияние на жизнеспособность компонентов полиструктуры является актуальной задачей.

Обычно потребителей, использующих технологию умного дома, интересует как ее максимальная эффективность, так и относительно устойчивое состояние инфраструктуры. При этом интересно такое состояние [4], при котором инфраструктура приспособлена к изменениям внешней среды, и, в то же время, до некоторой степени нечувствительна к таким изменениям. Это состояние подобно состоянию частичного гомеостаза. Для успешного управления инфраструктурой умного дома необходимо локализовать связи, поддерживающие такой гомеостаз, а также найти механизмы, позволяющие поддерживать её дальнейшее развитие. Полиструктурную систему нельзя стабилизировать одной обратной связью, поэтому, скорее всего это будет комплекс гомеостатических связей и механизмов. Поиск такого комплекса можно свести к следующим задачам: определение потенциалов системы, поиск напряжения в системе, нахождение обратных связей, расчёт потребностей.

Рассмотрим оптимальную инфраструктуру умного дома как полиструктурную систему, состоящую из нескольких подсистем, генерирующих разнородные по природе ресурсы. В основном под полиструктурной системой понимают множество разнообразных компонентов, отличающихся физическими свойствами, функциональным назначением, сложностью внутренней структуры, представляющих собой единое целое. Такая система обладает множеством подсистем, имеющих собственные оригинальные структуры, которые интегрируются в общую полиструктуру сложной системы. Например, в работе [5] полиструктура описана так: «...система человеческой социальной деятельности оказывается полиструктурной, т. е. состоит из многих как бы наложенных друг на друга структур, а каждая из них, в свою очередь, состоит из многих частных структур, находящихся в иерархических отношениях друг с другом».

В [6] приводится понятие категории структуры и другое определение полиструктуры, согласно которому структура – это есть система в плане синхронии, и система обретает свою сущность только когда функционирует, т. е. когда она является динамичной. Поэтому структура представляет собой некий одномоментный фиксированный слепок системы. Если рассмотреть временной период ее функционирования, то можно получить полиструктуру системы на этом временном периоде.

Умные дома сегодня позволяют не только облегчить быт, но и добиться оптимального потребления ресурсов, что становится возможным благодаря обмену информацией между различными компонентами полиструктурной системы.

КОМПОНЕНТЫ ПОЛИСТРУКТУРЫ УМНОГО ДОМА И ИХ ВЗАИМОДЕЙСТВИЯ

К компонентам умного дома относятся высокотехнологическое аппаратное и программное обеспечение, средства управления компонентами системы, средства обеспечения отказоустойчивости, доступности

и информационной безопасности. Как правило, на первых этапах проектирования полиструктура умного дома выглядит упрощенно, но по мере добавления компонентов возникают более сложные задачи автоматизации прикладных процессов, которые предъявляют дополнительные требования к надежности полиструктуры.

Под термином умный дом обычно понимают интеграцию в единую полиструктурную систему управления зданием следующих компонентов [7, 8]:

- системы управления и связи (отвечают за удаленное управление всеми электроприборами, приводами механизмов, механизацию здания, а также управление телефонной связью и локальной сетью здания);
- система отопления, вентиляции и кондиционирования (обеспечивает благоприятный климат в доме в заданном режиме);
- система освещения (экономит электроэнергию за счет рационального использования естественного освещения);
- система электропитания здания (обеспечивает бесперебойное питание, автоматически переключая на альтернативные источники электроснабжения);
- система безопасности и мониторинга (включает систему видеонаблюдения, систему контроля доступа в помещения, охранно-пожарную сигнализацию, автоматическую блокировку водоснабжения при протечке и заливе помещения, удаленное управление системами дома через телефон)

Современные программно-аппаратные комплексы умного дома позволяют оперативно подстраиваться под возникающие потребности пользователя, предоставляя ему множество возможностей участия в работе рынков энергии и энергосервисов [9, 10]. Структурная схема энергетических и информационных взаимосвязей компонентов полиструктуры умного дома представлена на рис. 1.

Каждый компонент полиструктуры умного дома обладает определенными свойствами (физическими, энергетическими и т. д.), функциональным назначением (освещение, отопление, кондиционирование и т. д.), показателем энергопотребления, эффективностью управления, надежностью всей полиструктуры. Связь между компонентами внутри полиструктуры осуществляется в виде взаимодействия «энергия – энергия», «информация – информация», «энергия – информация» и «информация – энергия». Компоненты полиструктуры умного дома и их энерго-информационные взаимодействия представлены в таблице.

Система электропитания – жизнеобеспечивающая система умного дома, связанная со всеми другими системами в виде взаимодействия «энергия – энергия». Ключевыми свойствами системы является безопасность, отказоустойчивость, экономичность и управляемость. В состав системы электропитания входят: датчики движения и света, выключатели, диммеры, силовой шкаф, резервные источники питания, выпрямители, аккумуляторные батареи, инверторы. Цепи электропитания подходят ко всем компонентам полиструктуры умного дома, включая охранно-пожарную сигнализацию, системы связи и видеонаблюдение, технологическое оборудование, такое как, например, водона-

гревательный котел или насосы водоснабжения, а также ответственны за питание шкафа с централизованной системой управления. Это позволяет контролировать потребление электроэнергии и, соответственно, уменьшить в несколько раз затраты на ее оплату. Применяемая при этом электрическая проводка отличается от стандартной, в первую очередь, наличием дополнительных сигнальных и

управляющих цепей к датчикам и исполнительным механизмам, что дает возможность потребителю отключить от питания все системы в доме, находясь вдали от него. Поэтому компоненты полиструктуры умного дома информационно связаны с энергетической системой через взаимодействия «энергия – информация», «информация – информация», «информация – энергия».



Рис. 1. Структурная схема взаимодействия компонентов полиструктуры умного дома (энергия – энергия, информация – информация, энергия – информация, информация – энергия)

Компоненты полиструктуры умного дома в разрезе их энерго-информационных взаимодействий

Компоненты полиструктуры умного дома и их назначение	Элементы системы	Энерго-информационные взаимодействия компонентов полиструктуры
Система электропитания (требуется для обеспечения надежным электроснабжением потребителей электроэнергии)	Силовые шкафы, резервные источники питания, выпрямитель, инвертеры, аккумуляторные батареи, дизель-генератор, электрические кабели, измерительные трансформаторы тока и напряжения, счетчики электрической энергии	Энергетическое взаимодействие с остальными компонентами полиструктуры для обеспечения надежного и бесперебойного электроснабжения. Информационное взаимодействие с технологическими системами: водонагревательными котлами, насосами водоснабжения, а также системами охранно-пожарной сигнализации, связи, видеонаблюдения для мониторинга величины потребления ресурса и эффективного управления полиструктурой умного дома
Система освещения (требуется для эффективного управления освещением жилого пространства)	Осветительные приборы, датчики движения, датчики света, выключатели, диммеры	Энергетическое взаимодействие с системой электропитания, устройствами управления освещением. Информационное взаимодействие с компонентами системы мониторинга, безопасности посредством встроенных в осветительные приборы модулей беспроводной связи.
Система отопления, вентиляции и кондиционирования (обеспечивает микроклимат помещений умного дома)	Система кондиционирования, вентиляционная система, радиаторы электрического или водяного отопления, датчики температуры, датчики влажности, датчики протечек воды, датчики давления в трубах, управляющие вентили, счетчики воды, газа, тепла	Энергетическое взаимодействие с системой электропитания, центрального водоснабжения и отопления, автономного отопления и электроснабжения. Информационное взаимодействие с системами мониторинга, безопасности, системой связи за счет встроенных модулей беспроводной связи и информационных портов на технологическом оборудовании отопления, вентиляции и кондиционирования.

Компоненты полиструктуры умного дома и их назначение	Элементы системы	Энерго-информационные взаимодействия компонентов полиструктуры
Система безопасности и мониторинга (обеспечивает наблюдение за компонентами полиструктуры умного дома, защиту от вторжения через физическое и виртуальное пространство)	Охранно-пожарная сигнализация, системы автоматического пожаротушения, датчики дыма/огня/газа, датчики вскрытия двери, датчики разбитого стекла, датчики движения, видеонаблюдение	Энергетическое взаимодействие с центральной и автономной системами электроснабжения. Информационное взаимодействие со всеми компонентами полиструктуры для консолидации информации о состоянии систем, их доступности и управляемости
Система связи (обеспечивает транспортировку информации внутри локальной вычислительной сети и выход в глобальную сеть)	Оборудование связи: активное и пассивное сетевое оборудование, проводные и беспроводные каналы связи	Энергетическое взаимодействие с центральной и автономной системами электроснабжения. Информационное взаимодействие со всеми компонентами полиструктуры имеющими модули проводной и беспроводной связи для консолидации информации о состоянии систем, их доступности и управляемости

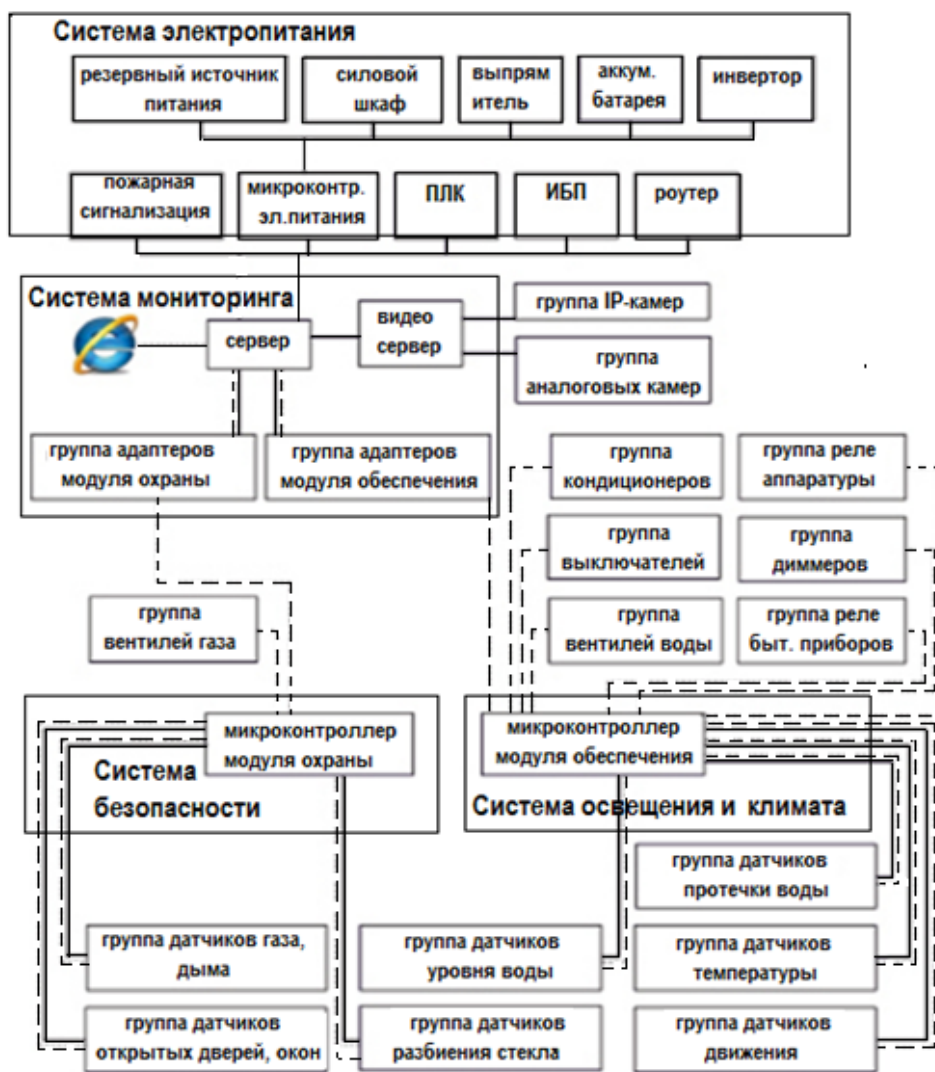


Рис. 2. Функциональная схема модулей управления компонентов полиструктуры «умного дома»

Интеграция системы электроснабжения в общую схему домашней автоматизации осуществляется при замене обычной аппаратуры распределения электроэнергии и защиты на специализированную [11, 12] (не только со штатными функциями, но и с коммуникационными возможностями), либо при дооснаще-

нии стандартной системы электроснабжения аппаратурой, имеющей возможности интеллектуального управления и мониторинга. При модернизации стандартной электрической сети появляется возможность контролировать не только оборудование, включенное в сеть, но и режимы подачи электроэнергии.

Функциональная схема модулей управления компонентов полиструктуры умного дома представлена на рис. 2. Сплошные линии на схеме отражают физические связи объектов, пунктирные – информационные связи объектов. Каждый элемент, представленный на схеме, входит в полиструктуру умного дома, и, значит, энергетически и информационно связан с системой электропитания, а также взаимодействует с одним или несколькими информационными элементами из множества информационных элементов полиструктуры умного дома [13].

СИСТЕМА УПРАВЛЕНИЯ ПОЛИСТРУКТУРОЙ УМНОГО ДОМА

Система управления полиструктурой умного дома использует набор показателей, организованных в соответствии со структурой процессов и применяемых в них информационных объектов для контроля достижения цели управления во взаимосвязанных подсистемах. Система управления полиструктурой задействована в оперативном, тактическом и стратегическом контурах принимаемых решений, настраиваемых в соответствии с установленным деревом целей. Настройка параметров (ключевых ориентиров) дерева целей зависит от условий микро и макросреды объекта, а также долгосрочных интересов его развития. В зависимости от контура система управления полиструктурой активизирует ряд ключевых показателей, необходимых для принятия управленческих решений. В соответствии с требованиями системы управления показатели связываются между собой через процессы и осуществляют мониторинг достижения цели управления. Таким образом, для эффективного управления полиструктурой выбирается цель управления, задается набор процессов, объектов, их свойств, связей между объектами, процессами и контурами управления, а также методы работы с ними [14, 15].

Цели управления полиструктурой умного дома – достижение максимальной энергоэффективности системы и высокой степени надежности её компонентов, адаптивность к текущим требованиям потребителей. Выделим в качестве ключевых критериев эффективности функционирования полиструктуры умного дома его энергопотребление, надежность, экономичность и управляемость. Показатель эффективности управления полиструктурой умного дома представим в виде формулы:

$$\alpha_{\text{компан}} = \alpha_{\text{потр}} + \alpha_{\text{надежн}} + \alpha_{\text{эконом}} + \alpha_{\text{упр}} \quad (1)$$

где: $\alpha_{\text{компан}}$ – обобщенный критерий эффективности управления компонента полиструктуры;

$\alpha_{\text{потр}}$ – критерий энергопотребления компонента;

$\alpha_{\text{надежн}}$ – критерий надежности компонента;

$\alpha_{\text{эконом}}$ – критерий экономичности компонента;

$\alpha_{\text{упр}}$ – критерий управляемости компонента.

Значимость каждого критерия в эффективности управления компонента полиструктуры настраивается, исходя из весовых показателей, выставляемых экспертами системы и функциональным заказчиком. Таким образом, система становится гибкой к возможным изменениям компонентов, а также к меняющимся требованиям потребителей. В процессе управления важно учитывать как можно большее количество факторов, оказывающих влияние на устойчивость системы. Для повышения эффективности управления процессом необходимо увеличивать его информативность за счет усложнения структур информационных объектов и методов работы с ними. Каждое свойство элемента полиструктуры может считаться показателем исполнения процесса, для чего анализируются изменения его значений. Показатель рассматривается как количественное или качественное отражение состояния свойства элемента системы или связи между элементами, которые измеряются и представляются наборами данных. Каждый показатель может быть получен из материальной системы, из другой информационной системы, рассчитан через группу иных показателей или вычислен собственным методом, который представляет собой процесс преобразования данных об объектах, их свойствах и связях между объектами [16].

ЗАКЛЮЧЕНИЕ

Умный дом рассмотрен как полиструктурная система, состоящая из компонентов энергетически и информационно связанных между собой и внешними системами. Даны базовые понятия полиструктуры умного дома. Выделены компоненты полиструктуры умного дома образующие единое энерго-информационное пространство: система электропитания, освещения, микроклимата, безопасности и связи. Энергетическая система является элементом поддержки жизнеспособности компонентов умного дома, а также рассматривается как элемент полиструктуры сложных технических и социальных систем. Взаимодействие компонентов системы между собой и с энергетической системой внутри умного дома происходит в виде связей типа «энергия – энергия», «энергия – информация», «информация – информация», «информация – энергия». Полиструктура умного дома позволяет создавать открытые системы с возможностью встраивания новых элементов для обеспечения эффективности ее управления.

Эффективность управления полиструктурой в целом определяется деревом целей (показателей), учитывающим такие критерии эффективности как энергопотребление, управляемость, надежность, экономичность. Значимость каждого критерия в эффективности управления компонента полиструктуры настраивается, исходя из весовых показателей, выставляемых экспертами системы и функциональным заказчиком. Целью управления полиструктурой умного дома стоит считать достижение максимальной энергоэффективности системы, высокой степени надежности её компонентов, адаптивность к потребностям активных потребителей способных гибко изменять настройки полиструктурной системы.

СПИСОК ЛИТЕРАТУРЫ

1. Цифровой переход в электроэнергетике России: Центр стратегических разработок, экспертно-аналитический доклад // Центр стратегических разработок. – М., 2017. – 47 с. – URL: https://csr.ru/wp-content/uploads/2017/09/Doklad_energetika-Web.pdf (дата обращения: 01.09.2019).
2. Мозохин А.Е., Дроздов В.Г., Старовров Б.А. Энергетика нового уклада (EnergyNet): проектирование интеллектуальных цифровых систем на электрических подстанциях: учеб. пособие. – Кострома: Изд-во Костром. гос. ун-та, 2018. – 70 с.
3. Петрова С.Ю. Общая задача управления полиструктурной системой // Вестник Новгородского государственного университета. – 2009. – № 50. – С. 35-39.
4. Щекочихин О.В., Шведенко В.В., Шведенко П.В. Онтология понятий информационной системы с поведением // Научно-технический вестник Поволжья. – 2016. – №5. – С. 223-226.
5. Перов В.И., Парахина В.Н. Формирование полиструктурных социально-экономических систем и особенности управления ими. – М.: Изд-во «НЕФТЬ и ГАЗ» РГУ нефти и газа им. И.М. Губкина, 2007. – 210 с.
6. Шведенко В.Н., Щекочихин О.В., Ровинская Т.И. Интеллектуальный анализ данных в интегрированной информационной системе, обладающей свойством поведения // Информатизация и связь. – 2018. – № 2. – С. 30-34.
7. Rau N.S., Taylor B. A central inventory of storage and other technologies to defer distribution upgrades-optimization and economics // IEEE Transactions on Power Delivery. – 1998. – Vol. 13, № 1. – P. 194-202. DOI: 10.1109/61.660878
8. Subudhi S., Agarwal P., Ghose T. A multi-stage concept for distribution system planning // Proceedings of First International Conference on Automation, Control, Energy and Systems. Hooghy, 2014. – P. 1-6. DOI: 10.1109/ACES.2014.6808024
9. Masera M., Bompard E.F., Profumo F., Hadjsaid N. Smart (electricity) grids for smart cities: assessing roles and societal impacts // Proceedings of the IEEE. – 2018. – Vol. 106, № 4. – P. 613-625. DOI: 10.1109/JPROC.2018.2812212
10. Cheng Z., Duan J., Chow M-Y. To centralize or to distribute: that is the question: a comparison of advanced microgrid management systems // Industrial Electronics Magazine IEEE. – 2018. – Vol. 12, № 1. – P. 6-24. DOI: 10.1109 / MIE.2018.2789926.
11. Du Y., Hao T., Lukic S., Lubkeman D., Dubey A., Karsai G. Development of a controller hardware-in-the-loop platform for microgrid distributed control applications // Electronic Power Grid (eGrid) 2018 IEEE. – 2018. – P. 1-6. DOI: 10.1109/eGRID.2018.8598696
12. Zhao C., Chen J., He J., Cheng P. Privacy-preserving consensus-based energy management in smart grids // Signal Processing IEEE Transactions on. – 2018. – Vol. 66, № 23. – P. 6162-6176. DOI: 10.1109/TSP.2018.2872817.
13. Wang K., Hu X., Li H. et al. A survey on energy internet communications for sustainability // IEEE Transactions on Sustainable Computing. – 2017. – Vol. 2, № 3. – P. 231-254. DOI: 10.1109/TSUSC.2017.2707122.
14. Alahakoon D., Yu X. Smart Electricity Meter Data Intelligence for Future Energy Systems: A Survey // IEEE Transactions on Industrial Informatics. – 2016. – Vol. 12, № 1. – P. 425-436. DOI: 10.1109/TII.2015.2414355
15. Philip BV., Alpcan T., Jin J., Palaniswami M. Distributed real-time iot for autonomous vehicles // Industrial Informatics IEEE Transactions on. – 2019. – Vol. 15, № 2. – P. 1131-1140. DOI: 10.1109/TII.2018.2877217
16. Islam T., Hashem M.M.A. A big data management system for providing real time services using fog infrastructure // Proceedings of the IEEE Symposium on Computer Applications & Industrial Electronics. Penang, 2018. – P. 85-89. DOI: 10.1109/ISCAIE.2018.8405449.

Материал поступил в редакцию 10.10.19.

Сведения об авторах

ШВЕДЕНКО Владимир Николаевич – доктор технических наук, профессор, ведущий научный сотрудник ВИНТИ РАН, Москва
e-mail: vv_shved@mail.ru

МОЗОХИН Андрей Евгеньевич – кандидат технических наук, заместитель начальника отдела эксплуатации автоматизированных систем управления филиала ПАО «МРСК Центра» – «Костромаэнерго», Кострома
e-mail: mozokhin@mail.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81'322.2:004.65

И.В. Селиванова, Д.В. Косяков, А.Е. Гуськов

Классификация научных текстов на основе компрессии аннотаций публикаций

Исследуется возможность установления смысловой близости научных текстов методом их автоматической классификации, основанным на сжатии аннотаций. Идея метода состоит в том, что алгоритмы компрессии типа PPM (prediction by partial matching) сжимают терминологически близкие тексты существенно лучше, чем далекие. Если для каждой классифицируемой тематики будет сформировано ядро публикаций (аналог обучающей выборки), то наилучшая доля сжатия будет указывать на принадлежность классифицируемого текста к соответствующей тематике. Было определено 30 тематических категорий, каждой из них в базе данных Scopus получены аннотации около 500 публикаций, из которых разными способами выбирались 100 аннотаций для ядра и 20 аннотаций для тестирования. Установлено, что построение ядра на основе высокоцитируемых публикаций выявляет до 12% ошибок против 32% при случайной выборке. На качество классификации влияет и изначальное количество категорий: чем меньше категорий участвует в классификации и чем больше терминологические различия между ними, тем выше её качество.

Ключевые слова: классификация научных текстов, сжатие текстов, библиографические базы данных, Scopus

ВВЕДЕНИЕ

В последние десятилетия в связи с ростом количества научных публикаций проблема классификации текстов становится особенно актуальной. При этом классифицируются как целые художественные или поэтические тексты [1, 2], так и короткие текстовые сообщения, например, СМС или твиты [3-5]. Но единого метода классификации, который бы сразу показал и высокую эффективность, и низкие трудозатраты, до настоящего момента найдено не было.

Одной из областей, где эта задача имеет особую важность, является классификация научных документов. Неверно классифицированные публикации затрудняют поиск необходимых ученому статей, что является причиной потери актуальных исследований в интересующей его области наук.

В работах [6, 7] был предложен метод классификации научных текстов, основанный на сжатии информации. Он показал более чем 90%-эффективность и низкие трудозатраты при классификации англоязычных текстов Архива научных публикаций *arXiv.org*. Одним из недостатков этих работ являлось применение метода только к полнотекстовым документам. Полные тексты статей, из-за их объема, часто содержат много лишнего фраз, что может привести

к некоторым ошибкам при классификации. Аннотации публикаций, наоборот, должны содержать только ключевые моменты, используемые в статье [8, с. 35], что может облегчить автоматическую классификацию научных работ. Более того, во многих научных библиографических базах данных (ББД) не всегда удается получить полный текст статьи и часто доступна только ее аннотация.

Цель нашего исследования – применение метода, основанного на алгоритмах сжатия, к классификации аннотаций публикаций, индексируемых в международных ББД. Первоначальные результаты этой работы были представлены на XVII Российской конференции «Распределенные информационно-вычислительные ресурсы» *DICR-2019*.

Наиболее авторитетными ББД для ученых являются *Web of Science (WoS)* и *Scopus*. Они служат источником данных для множества наукометрических исследований, различных рейтингов университетов, а также используются при оценке публикационной активности исследователей, организаций и стран.

Тем не менее, классификация публикаций в этих библиографических базах данных вызывала большое количество вопросов. Одним из главных ее недостатков являлось то, что как в *WoS*, так и в *Scopus* опре-

деление тематики публикации происходило только на уровне журнала, в котором она издана [9, 10]. Особенно негативно это сказывалось на статьях из мультидисциплинарных журналов. Публикация, которая могла иметь только одно направление, была отнесена ко всем тем же предметным рубрикам, что и этот журнал. Это создавало «замусоренность» научных направлений публикациями, которые, возможно, не имели к ним никакого отношения.

Для улучшения качества классификации в *Scopus* в 2017 г. введена новая система классификации *Topic Prominence in Science* [11], которая была построена подобно методике, описанной в работе [12]. Авторы предложили способ присвоения публикациям отдельных категорий, состоящий из трех этапов: на первом этапе происходило определение связанности публикаций путем использования прямого цитирования; на втором – происходила иерархическая кластеризация публикаций; третий этап был посвящен названию получившихся кластеров метками на основе терминов заголовков и аннотаций публикаций, находящихся в этой группе.

Важную роль для улучшения точности классификации научных документов играет выбор системы классификации. Эти системы могут быть разделены на несколько типов:

1) Библиотечные классификаторы (например, УДК). За основу этой системы взята десятичная классификация, разработанная в 1876 г. американским библиографом М. Дьюи. Центральная часть УДК – основные таблицы, охватывающие всю совокупность знаний и построенные по иерархическому принципу деления от общего к частному с использованием цифрового десятичного кода [13]. Также к библиотечным классификаторам относится ББК – национальная классификационная система России. Принцип ее формирования аналогичен УДК.

2) Национальные классификаторы, например, классификация *Fields of research (FOR)* из *Australian and New Zealand Standard Research Classification (ANZSRC)*, представляющая собой иерархическую классификацию с тремя уровнями: на первом – находятся обширные научные направления, на втором – связанные с ними группы, – на третьем расположены более узкие области. *FOR* применяется, например, в платформе *Dimensions* [14]. Другими примерами национальных классификаторов является Номенклатура научных специальностей России, используемая Высшей аттестационной комиссией Минобрнауки (ВАК) [15], Общероссийский классификатор специальностей по образованию (ОКСО), который сопоставлен с Международной стандартной классификацией образования (МСКО) [16] и Государственный рубрикатор научно-технической информации (ГРНТИ) [17].

3) Международные классификаторы, к которым относится *Field of science and technology (FOS)* – система классификации, опубликованная Организацией экономического сотрудничества и развития в 2002 г. и измененная в 2007 г. [18]. В основе этой классификации лежат 6 научных направлений, разделенных на 42 области. Другим примером международного классификатора является номенклатура *UNESCO* – сис-

тема, разработанная ЮНЕСКО для классификации научных работ и диссертаций [19].

4) Классификаторы в международных ББД. Эти классификаторы, как и предыдущие, построены по иерархическому принципу. В *WoS* классификатор состоит из трех уровней. На основном уровне расположены шесть областей наук, которые разделены на 39 подуровней. На третьем подуровне классификатора *WoS* содержится 253 категории [20]. В *Scopus* используется *All Science Journals Classification (ASJC)*. Она включает издания, распределенные по четырем общим научным направлениям: биологические науки, физические науки, медицина, социальные и гуманитарные науки. Эти направления разделены на 27 крупных предметных областей и более 300 узких категорий [21]. Несмотря на широкий охват тематик, у *ASJC* есть существенные недостатки. Например, в двух разных научных областях встречаются две близкие категории: *Language and Linguistics* (код – 1203, область – *Arts and Humanities*) и *Linguistics and Language* (код – 3310, область – *Social Sciences*). Эта проблема была отмечена в 2016 г. в работе [22], где авторы предлагали либо слить эти категории, либо обозначить более четкие различия между ними.

Важность выбора системы классификации отмечается в работе [23]. При рассмотрении двух португальских систем классификации: *FCT* и *DeGóis platform*, было обнаружено, что область науки *Nursing* полностью отсутствует в первой системе, а во второй она нечетко определена. В работе [24] показаны две основные проблемы номенклатуры *UNESCO*: первая связана с тем, что в этой классификации в крупных научных областях могут быть потеряны более мелкие категории, вследствие чего классификация будет недостаточно полной; вторая – заключается в том, что в этой классификации отсутствуют области наук, которые появились недавно, что также существенно ухудшает качество классификации.

Для решения задачи классификации текстовых документов применяется множество различных методов. Одним из наиболее часто используемых алгоритмов является метод *k*-ближайших соседей и его модификации, где классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки. На этом методе базируется, например, работа [25]. Другим алгоритмом является байесовская классификация, которая работает на вычислении апостериорных вероятностей классов. Этот алгоритм применен в работе [26]. Представителем линейных классификаторов служит метод опорных векторов, который заключается в построении гиперплоскости, разделяющей объекты выборки наиболее оптимальным способом. Сравнение алгоритмов байесовской классификации и метода опорных векторов при классификации заголовков статей приводится в работе [27].

В последнее время для решения задачи классификации все чаще применяются нейронные сети. В статье [28] для решения задачи классификации текстов предлагается использовать рекуррентные сверточные нейронные сети. Ее авторы приходят к выводу, что применение нейронных сетей при классификации текстовых документов поможет избежать проблемы

разреженности данных, а также собрать больше контекстуальной информации о сущностях по сравнению с традиционными методами. Сверточные нейронные сети показали высокую точность (83,98%) и при классификации патентных документов [29].

В среднем точность различных алгоритмов классификации текстовой информации варьируется от 70% до 86% [30, 31].

Основное преимущество классификации научных публикаций – общность терминов, понятий и оборотов, используемых в текстах одной и той же области наук. При этом, чем более узконаправленной является научная область, тем более специфичной является лексика относящихся к ней статей.

Для классификации научных публикаций могут быть использованы методы, базирующиеся на:

- цитировании, которое включает прямое цитирование, ко-цитирование и библиографическое сочетание [32-34]. Чаще всего источником данных для таких исследований является БД *WoS*;
- метаданных публикаций (списков соавторов, названий публикаций, ключевых слов и др.). Такой метод применяется в работе [35] при кластеризации биомедицинских публикаций в базе данных *MEDLINE*;
- комбинации использования цитирования и метаданных. Например, в работе [36] авторы применяют этот подход для улучшения качества классификации документов в *ACM Digital Library*;
- аннотациях и полных текстах публикаций. Классификацией аннотаций публикаций из научных направлений *Materials science*, *Physics u Chemistry* БД *Scopus* занимались авторы статьи [37]. Проверка качества классификации медицинских и биологических публикаций из базы данных *PubMed* путем сравнения результатов применения методов *k*-ближайших соседей, байесовской классификации и опорных векторов к аннотациям рассмотрена в работе [38]. Полнотекстовые научные документы классифицировались в работах [7, 39].

1. МЕТОДЫ И ДАННЫЕ КЛАССИФИКАЦИИ

1.1. Метод

Рассмотрим подробнее метод, основанный на алгоритмах сжатия. Пусть есть n научных областей $X_1, \dots, X_n$, для каждой из которых определен характерный для нее набор текстов (файлы, с которыми составляют «ядро» это области). Также есть некоторый тестовый файл u , тематику которого нужно определить, и архиватор φ , который может быть применен для сжатия любого множества текстов.

Работа метода состоит в том, что тестовый файл u начинает последовательно сжиматься с каждым из n ядер при помощи архиватора φ . В итоге, область тестового файла u определяется тем ядром, с которым он имеет наилучшее сжатие.

Для работы метода были выбраны следующие параметры:

- Архиватор – WinRAR при максимальном значении памяти 128 Мбайт, алгоритм сжатия – RPPMd [6, 7];

- Объем ядра – 100 файлов. Этот объем является рекомендованным в работе [7], так как при большем количестве файлов в ядре качество классификации практически не улучшается, а время работы алгоритма увеличивается.

1.2. Данные

Источником информации для настоящего исследования служит БД *Scopus*. Процесс извлечения данных состоит из трех этапов.

1. Извлечение информации о названии журнала, *eid* публикации, цитировании, категории через *Scival* – аналитический инструмент компании *Elsevier*, основанный на данных *Scopus*. Данные по 30 категориям были выбраны за период 2009-2018 гг. приведены в табл. 1.
2. Формирование файлов аннотаций путем получения их текстов через *Scopus Abstract Retrieval API*
3. Удаление файлов аннотаций с отсутствующим текстом

Всего было выгружено 15 тыс. файлов (по 500 для каждой категории).

1.3. Процесс классификации

Для проверки того, насколько эффективно метод, основанный на алгоритмах сжатия, работает на аннотациях публикаций, индексруемых в БД *Scopus*, классификация проводилась для:

- 1) тестовых файлов с одной категорией,
- 2) тестовых файлов с несколькими категориями.

В первом случае осуществлялась проверка работы метода путем классификации файлов с одной категорией, а также подбор оптимальных по составу ядер.

На втором этапе классификация проводилась для тестов с несколькими категориями, ядра для которой были подобраны на первом этапе. Этот случай полезен при проверке качества классификации публикаций из мультидисциплинарных изданий.

Рассмотрим эти два этапа подробнее.

1.3.1. Классификация тестовых файлов с одной категорией

При классификации тестовых файлов с одной категорией введены три типа ошибок определения категории:

- Ошибки I типа связаны с неправильным определением категории внутри научного направления, например, вместо категории *Aquatic Science* определялась *Plant Science* из той же области *Agricultural and Biological Sciences*.
- К ошибкам II типа отнесены те тестовые файлы, у которых определилась другая область при одном научном направлении. Например, вместо нужной категории *Cell Biology* из области *Biochemistry, Genetics and Molecular Biology* определялась категория *Pharmacology* из области *Pharmacology, Toxicology and Pharmaceutics*. При этом общее научное направление *Life Sciences* сохранилось.
- III тип ошибок – самый серьезный, к нему отнесены те тестовые файлы, у которых ошибка в классификации произошла на самом верхнем уровне. Например, вместо научного направления *Physical Sciences* определилось *Social Sciences*.

Исследуемые области наук по уровням классификации

Направление	Область	Категория
Life Sciences	Agricultural and Biological Sciences	Animal Science, Zoology
Life Sciences	Agricultural and Biological Sciences	Aquatic Science
Life Sciences	Agricultural and Biological Sciences	Plant Science
Social Sciences	Arts and Humanities	History
Social Sciences	Arts and Humanities	Literature, Literary Theory
Life Sciences	Biochemistry, Genetics and Molecular Biology	Cell Biology
Life Sciences	Biochemistry, Genetics and Molecular Biology	Endocrinology
Social Sciences	Business, Management and Accounting	Marketing
Physical Sciences	Chemical Engineering	Catalysis
Physical Sciences	Chemistry	Inorganic Chemistry
Physical Sciences	Chemistry	Organic Chemistry
Physical Sciences	Computer Science	Artificial Intelligence
Physical Sciences	Computer Science	Computer Vision and Pattern Recognition
Physical Sciences	Computer Science	Hardware and Architecture
Physical Sciences	Earth and Planetary Sciences	Geology
Physical Sciences	Earth and Planetary Sciences	Oceanography
Physical Sciences	Mathematics	Algebra and Number Theory
Physical Sciences	Mathematics	Geometry and Topology
Physical Sciences	Mathematics	Logic
Physical Sciences	Mathematics	Numerical Analysis
Physical Sciences	Mathematics	Statistics and Probability
Health Sciences	Medicine	Ophthalmology
Health Sciences	Medicine	Surgery
Life Sciences	Pharmacology, Toxicology and Pharmaceutics	Pharmacology
Physical Sciences	Physics and Astronomy	Astronomy and Astrophysics
Physical Sciences	Physics and Astronomy	Condensed Matter Physics
Physical Sciences	Physics and Astronomy	Nuclear and High Energy Physics
Social Sciences	Psychology	Social Psychology
Social Sciences	Social Sciences	Library and Information Sciences
Social Sciences	Social Sciences	Sociology and Political Science

Таблица 2

Пример расчета нормированного коэффициента сжатия

Области теста	Algebra and Number Theory, %	Geometry and Topology, %	мин, %	Algebra and Number Theory	Geometry and Topology
Algebra and Number Theory, Geometry and Topology	26,70	26,80	26,70	OK	OK
Algebra and Number Theory, Geometry and Topology	33,25	32,85	32,85	missed	OK

Классификация осуществлялась с использованием ядер двух типов:

1) **произвольные ядра** были составлены из аннотаций публикаций, отобранных случайным образом.

2) **подобранные ядра** были составлены из 100 самых высокоцитируемых публикаций в каждой области наук. Такой подход, предположительно, увеличивает качество ядра в связи с тем, что публикации из той же области, цитирующие отобранные в ядро

статьи, с высокой долей вероятности наследуют и характерную лексику.

В обоих случаях использовался один и тот же набор произвольных тестовых файлов, для каждой из 30 категорий было отобрано по 20 тестовых файлов, суммарно 600 тестов.

Также было проверено влияние на качество классификации наличия стоп-слов [40] и названий издательств, присутствующих в текстах аннотаций (на-

пример, © 2009 *The American Physical Society*). Создание ядер и подготовка тестовых файлов были выполнены как с оригинальными текстами аннотаций, так и с удалением стоп-слов (например, *always*, *every*, *just* и т.п.), заменой заглавных букв на строчные и удалением всех символов, кроме цифр, букв и знаков препинания: ., !, ?, :, ,, -.

1.3.2. Классификация тестовых файлов с несколькими категориями

Для классификации тестовых файлов с несколькими категориями использовались только ядра с самыми высокоцитируемыми публикациями.

Отбор 20 тестов осуществлялся произвольным образом из публикаций, у которых по меньшей мере две категории совпадало с категориями из табл. 1. Суммарно, как и в предыдущих случаях, было отобрано 600 тестов.

Введем следующие группы результатов:

- у тестового файла верно определилось не менее 50% указанных категорий. Например, у теста было указано 4 категории: *Algebra and Number Theory*, *Numerical Analysis*, *Geometry and Topology*, *Discrete Mathematics and Combinatorics*. В число исследуемых нами категорий входят только первые три. Соответственно, чтобы попасть в эту группу, у тестового файла должны определиться минимум два из трех первых. Верно определенными считаются категории, у которых нормированный коэффициент сжатия (процент сжатия за вычетом минимального процента сжатия по всем категориям) меньше, чем минимальный процент сжатия со всеми категориями * 0,50% (при большем пороге количество ошибок почти не изменялось). В качестве примера такого расчета рассмотрим два тестовых файла с категориями *Algebra and Number Theory* и *Geometry and Topology* (табл. 2). Минимальный процент сжатия у первого теста определился с категорией *Algebra and Number Theory*. При этом если в качестве порогового значения выбирать не только минимальный процент сжатия, а минимальный процент сжатия со всеми категориями * 0,50%, то с этим тестом правильно будет определена и вторая указанная категория: *Geometry and Topology*. У второго же тестового файла была определена только категория *Geometry and Topology*;

- у тестового файла определилась хотя бы одна из указанных категорий;

- все категории тестового файла определились неверно. К этому случаю будут отнесены те тестовые файлы, у которых определилась какая-то другая категория

При этом один тестовый файл может относиться только к одной из этих групп.

1.4. Влияние количества категорий на качество классификации

В работе [30] авторы приходят к выводу, что количество категорий влияет на классификацию, и если объединить категории со схожими терминами в одну, то качество классификации улучшится.

Для оценки влияния количества категорий на качество классификации аннотаций нами был проведен эксперимент с последовательным увеличением коли-

чества рассматриваемых категорий с 5 до 30 с шагом в 5 (5, 10, 15, 20, 25, 30). Для тестовых файлов были отобраны случайным образом по 20 публикаций из первых 5 категорий (всего 100 файлов).

2. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ

2.1. Классификация тестовых файлов с одной категорией

2.1.1. Классификация при произвольных ядрах

В табл. 3 приведены результаты классификации тестовых файлов с одной категорией при произвольных ядрах по типам ошибок.

Доля ошибочно определенных тестовых файлов составила 32% от общего количества (192 из 600 тестов).

Чаще всего определение неверного научного направления происходит из-за категорий, близких по терминологии. Например, такими категориями являются *Aquatic Science* и *Oceanography*. Но иногда характер ошибки определить не удастся, например, в случае с публикацией с *eid=2-s2.0-67651018249* из категории *Condensed Matter Physics*, вместо которой определилась категория *Marketing*. Визуально определить причину по тексту аннотации не удалось (рис. 1).

Нами было проведено попарное сжатие файлов из ядер этих двух категорий и тестового файла с *eid=2-s2.0-67651018249*. Почти по всем отдельным 100 файлам из ядра категории *Marketing* тест с *eid=2-s2.0-67651018249* показывает лучшее сжатие. При этом средний нормированный коэффициент сжатия тестового файла с категорией *Condensed Matter Physics* составляет 9,76%, а с категорией *Marketing* – 9,13%.

В топ-10 файлов, с которыми произошло наилучшее сжатие этого теста, вошли 3 файла из категории *Condensed Matter Physics* и 7 из категории *Marketing* (табл. 4).

Тексты двух аннотаций категории *Marketing*, с которыми у тестового файла произошло лучшее попарное сжатие, приведены на рис. 2, 3. У этих файлов полностью различается терминология как между собой, так и с исследуемым тестовым файлом. Таким образом, результаты позволяют предположить, что при определении категории тестового файла с *eid=2-s2.0-67651018249* категории *Condensed Matter Physics* ошибка может быть связана с работой метода.

2.1.2. Классификация при подобранных ядрах

В табл. 5 приведены результаты классификации тестовых файлов с одной категорией при ядрах, в состав которых входят самые высокоцитируемые публикации.

Использование подобранных ядер улучшило результаты классификации на 20%. Число ошибок III типа уменьшилось в 3 раза. В основном такие ошибки возникали из-за находящихся в разных научных направлениях категорий или файлов, в которых применяются схожие термины. Например, у теста из категории *Sociology and Political Science* неверно определилась категория *Aquatic Science*, но в тексте этой аннотации применяется много терминов, используемых в категории *Aquatic Science* (рис. 4).

**Результаты классификации тестовых файлов при произвольных ядрах
с различным числом цитирования**

Категория	Общее кол-во тестов	Кол-во ошибок	Ошибки		
			I типа	II типа	III типа
Algebra and Number Theory	20	7	5	2	0
Animal Science and Zoology	20	1	0	1	0
Aquatic Science	20	12	11	0	1
Artificial Intelligence	20	8	6	2	0
Astronomy and Astrophysics	20	1	1	0	0
Catalysis	20	5	0	5	0
Cell Biology	20	1	0	0	1
Computer Vision and Pattern Recognition	20	3	3	0	0
Condensed Matter Physics	20	11	2	4	5
Endocrinology	20	3	1	2	0
Geology	20	0	0	0	0
Geometry and Topology	20	11	9	2	0
Hardware and Architecture	20	0	0	0	0
History	20	8	1	7	0
Inorganic Chemistry	20	13	4	1	8
Library and Information Sciences	20	12	2	8	2
Literature and Literary Theory	20	11	5	5	1
Logic	20	2	0	1	1
Marketing	20	2	0	2	0
Nuclear and High Energy Physics	20	4	4	0	0
Numerical Analysis	20	0	0	0	0
Oceanography	20	11	0	1	10
Ophthalmology	20	9	7	0	2
Organic Chemistry	20	3	0	2	1
Pharmacology	20	18	0	18	0
Plant Science	20	20	13	7	0
Social Psychology	20	1	0	1	0
Sociology and Political Science	20	8	0	7	1
Statistics and Probability	20	6	0	2	4
Surgery	20	1	0	0	1
Общее количество	600	192	74	80	38
Доля от общего количества, %		32	12	13	6

Two-particle dispersion is of central importance to a wide range of natural and industrial applications. It has been an active area of research since Richardson's (1926) seminal paper. This review emphasizes recent results from experiments, high-end direct numerical simulations, and modern theoretical discussions. Our approach is complementary to Sawford's (2001), whose review focused primarily on stochastic models of pair dispersion. We begin by reviewing the theoretical foundations of relative dispersion, followed by experimental and numerical findings for the dissipation subrange and inertial subrange. We discuss the findings in the context of the relevant theory for each regime. We conclude by providing a critical analysis of our current understanding and by suggesting paths toward further progress that take full advantage of exciting developments in modern experimental methods and peta-scale supercomputing. Copyright © 2009 by Annual Reviews. All right reserved. All rights reserved.

Рис. 1. Аннотация публикации с eid=2-s2.0-67651018249

Топ-10 файлов, с которыми произошло наилучшее сжатие исследуемого теста
с eid=2-s2.0-67651018249 категории *Condensed Matter Physics*

Неверно определившийся тест из <i>Condensed Matter Physics</i>	Идентификатор файла	Категория файла	Нормированный процент сжатия, %
2-s2.0-67651018249	2-s2.0-70350534620	Condensed Matter Physics	0,00
2-s2.0-67651018249	2-s2.0-80052140988	Marketing	1,17
2-s2.0-67651018249	2-s2.0-67149130202	Marketing	1,47
2-s2.0-67651018249	2-s2.0-79960889541	Marketing	2,70
2-s2.0-67651018249	2-s2.0-70449090433	Condensed Matter Physics	2,94
2-s2.0-67651018249	2-s2.0-70449127336	Condensed Matter Physics	3,16
2-s2.0-67651018249	2-s2.0-79959944133	Marketing	3,20
2-s2.0-67651018249	2-s2.0-67149101079	Marketing	3,49
2-s2.0-67651018249	2-s2.0-78650307261	Marketing	3,88
2-s2.0-67651018249	2-s2.0-78751585438	Marketing	3,90

Franchisee selection is a major input for franchising success. In this article, we argue that franchisee selection criteria do not differ between social and commercial franchising. They may be even more relevant for obtaining social franchising success. We discuss criteria for franchisee selection and present details of our multiple case study research to support the argument. Our study finds that evolved social franchisors do adopt similar selection criteria as commercial franchisees. In addition, constraints faced with franchisee selection among commercial franchisors are reflected also among social franchisors. We contribute to franchising literature by extending commercial franchisee selection criteria to social franchisee selection. A major managerial implication of this research is that existing franchising professionals could easily assist new social franchisors in developing their social franchisees. Future research could be study criteria weights and methodology adopted for making final selection. A new research direction could involve studying if selection criteria would differ based on (a) social cause and (b) franchisee location. © Taylor & Francis Group, LLC.

Рис. 2. Аннотация публикации категории *Marketing* с eid=2-s2.0-80052140988

Despite the popularity of online digital music and the broad application of digital music sampling, in the existing literature, there is a lack of substantial studies that examine online digital music sampling. This study uses a laboratory experiment to explore the determinants of the five effectiveness dimensions, i.e., evaluation, Willingness-to-Pay (WTP), perceived sampling usefulness, sampling cost and the likelihood of being a free rider, of online digital music sampling. Digital music samples with a higher quality and longer segments were found to increase the sampler's music evaluation and make the evaluation process more useful. Also, the sampler's music evaluation significantly determines his/her WTP. Higher music evaluations not only decrease the sampler's sampling cost during the sampling process, but also reduces the probability that the sampler will take the music sample as a substitute for the original music. This study also shows that the current practice of online digital music sampling is not ideal and music retailers could improve their music sampling strategies by providing digital music samples with longer segments and of higher quality. All of these findings have significant implications for music retailers to use digital music sampling strategies better. Copyright © 2009, Inderscience Publishers.

Рис. 3. Аннотация публикации категории *Marketing* с eid=2-s2.0-67149130202

Результаты классификации тестовых файлов при ядрах с самыми высокоцитируемыми публикациями

Категория	Общее кол-во тестов	Кол-во ошибок	Ошибки		
			I типа	II типа	III типа
Algebra and Number Theory	20	8	7	1	0
Animal Science and Zoology	20	7	6	1	0
Aquatic Science	20	2	2	0	0
Artificial Intelligence	20	5	2	2	1
Astronomy and Astrophysics	20	0	0	0	0
Catalysis	20	4	0	4	0
Cell Biology	20	4	1	3	0
Computer Vision and Pattern Recognition	20	3	3	0	0
Condensed Matter Physics	20	2	0	2	0
Endocrinology	20	1	0	1	0
Geology	20	0	0	0	0
Geometry and Topology	20	3	3	0	0
Hardware and Architecture	20	0	0	0	0
History	20	4	2	1	1
Inorganic Chemistry	20	3	0	3	0
Library and Information Sciences	20	3	1	1	1
Literature and Literary Theory	20	2	1	1	0
Logic	20	3	2	1	0
Marketing	20	1	0	1	0
Nuclear and High Energy Physics	20	3	3	0	0
Numerical Analysis	20	0	0	0	0
Oceanography	20	4	0	0	4
Ophthalmology	20	0	0	0	0
Organic Chemistry	20	1	0	0	1
Pharmacology	20	2	0	2	0
Plant Science	20	2	1	1	0
Social Psychology	20	2	0	2	0
Sociology and Political Science	20	2	1	0	1
Statistics and Probability	20	1	0	1	0
Surgery	20	0	0	0	0
Общее количество	600	72	35	28	9
Доля от общего количества, %		12	6	5	2

*This paper seeks to understand how the Brazilian Amazon, which many thought unsuitable for **agricultural development**, has yielded to a dynamic cattle **economy** in only a few decades. It does so by embedding the Thunian model of **location** rents within the regime of **capital accumulation** that has driven the Brazilian **economy** since the mid-20th century. The paper addresses policies that have created location rents in **Amazônia**, the effect of these rents on **land managers**, and the **spatial implications** of their **behavior** on **forests**. Thus, the paper connects macro-processes and structures to agents on the **ground**, in providing a political **ecological explanation** relevant to **land change science**. The policy discussion focuses on **reductions** in transportation costs, improvements in **animal health**, and monetary and **trade** reforms. To illustrate the impact of policy, the paper presents data on the **geography** of Amazonian herd **expansion**, on the growth of Amazonian exports, and on the profitability of the **region's** cattle **economy**. It follows the empirical presentation with more abstract consideration of the spatial relations between cattle ranching and **soy farming**, and implications for **deforestation**. The paper concludes on a speculative note by considering the likelihood of **forest transition** in the **region**, given the transformation of **Amazônia** into a global **resource frontier**. © 2008 Elsevier Ltd. All rights reserved.*

Рис. 4. Аннотация теста с eid=2-s2.0-70449527784 из категории *Sociology and Political Science* (**жирным** выделены термины, часто встречающиеся в категории *Aquatic Science*)

This chapter provides a tutorial overview of distributed optimization and game theory for decision-making in networked systems. We discuss properties of first-order methods for smooth and non-smooth convex optimization, and review mathematical decomposition techniques. A model of networked decision-making is introduced in which a communication structure is enforced that determines which nodes are allowed to coordinate with each other, and several recent techniques for solving such problems are reviewed. We then continue to study the impact of noncooperative games, in which no communication and coordination are enforced. Special attention is given to existence and uniqueness of Nash equilibria, as well as the efficiency loss in not coordinating nodes. Finally, we discuss methods for studying the dynamics of distributed optimization algorithms in continuous time. © 2010 Springer London.

Рис. 5. Текст аннотации публикации с eid=2-s2.0-77958562700 из категории Library and Information Sciences

The horse skeleton found in the autumn of 1958 at the fortress of Buhen in northern Sudan has become one of the most prominent, but also one of the most enigmatic equid remains from the second millennium BC: Firstly, because of its assumed early date of c. 1675 BC, deduced by W.B. Emery after analysing the stratigraphical data, This - according to our knowledge at the time - being several decades before the oldest known equid remains in Egypt. Secondly, because of wear on the lower left second premolar (LP2), which has led to the conclusion that it was most probably caused by bit-wear. Since the 1960s, both conclusions have been subject to criticism. The purpose of this study is to provide a review of the history of research and reception of the Buhen horse in its interdisciplinary context over the last fifty years with the result that only modern scientific techniques might be able to solve some of the outstanding questions. © 2009 Brill.

Рис. 6. Тестовый файл с eid=2-s2.0-77951062083 из категории History

Некоторые неверно определенные тесты, возможно, связаны с неверной изначальной классификацией. Так, в случае категории *Library and Information Sciences* тестовый файл отнесся к категории *Artificial Intelligence*. Текст аннотации содержит значительно количество терминов, характерных для области *Computer Science* (рис. 5).

В некоторых закономерностях не было выявлено определения неверной категории. Например, тестовый файл из *History* неверно отнесся к *Condensed Matter Physics* (рис. 6).

Стоит отметить, что неверно определившийся файл из пункта 2.1.1. с eid=2-s2.0-67651018249 при подобранных ядрах определился верно. Таким образом, состав ядра оказывает большое влияние на качество классификации.

2.1.3. Влияние на классификацию стоп-слов и названий издательств

Для изучения влияния присутствия названий издательств в аннотациях на качество классификации использовались подобранные ядра (см. п. 2.1.2).

В табл. 6 приведено сравнение качества классификации аннотаций со стоп-словами и названиями издательств и без них. При удалении названий издательств количество ошибок возросло на 3%. Почти в

2 раза увеличилось количество ошибок в категориях *History*, *Geometry and Topology*, *Literature and Literary Theory*, *Sociology and Political Science*. Возможно, это связано с тем, что чаще всего высокоцитируемые публикации печатаются в одних издательствах, в названиях которых указаны важные термины для категории. Например, в одном из неверно определившихся после удаления издательства теста раньше встречалась следующая строка: © 2010 English Literary Renaissance Inc. Published by Blackwell Publishing Ltd.

При удалении стоп-слов из аннотаций, где присутствовали названия издательств, количество ошибок уменьшилось до 11%. Однако это уменьшение произошло не равномерно по всем категориям: если в категории *Cell Biology* удаление стоп-слов повлияло положительно на качество классификации, то в категории *Literature and Literary Theory* количество ошибок увеличилось в 2 раза.

В случае удаления как стоп-слов, так и названий издательств, количество ошибок увеличилось до 16%. Аналогично предыдущему случаю на ряд категорий удаление стоп-слов и названий издательств повлияло положительно, тогда как другие стали определяться ошибочно. Так, например, в категории *Geometry and Topology* тест с eid= 2-s2.0-84055189802 при удалении стоп-слов определился верно, а тест с

eid= 2-s2.0-77956268008, который раньше определялся верно, теперь отнесся к категории *Numerical Analysis*. Возможно, это связано с тем, что при удалении стоп-слов длина аннотации уменьшается.

Таким образом, отсутствие названий издательств в текстах аннотаций негативно влияет на качество классификации. Про влияние стоп-слов однозначного же вывода сделать нельзя.

Таблица 6

Влияние стоп-слов и присутствия названий издательств на качество классификации

Категория	Общее кол-во тестов	Кол-во ошибок	Количество ошибок		
			со стоп-словами, без названий издательств	без стоп слов, с названиями издательств	без стоп-словом и названий издательств
Algebra and Number Theory	20	8	8	8	7
Animal Science and Zoology	20	7	7	6	8
Aquatic Science	20	2	2	1	2
Artificial Intelligence	20	5	6	6	7
Astronomy and Astrophysics	20	0	0	0	0
Catalysis	20	4	3	5	5
Cell Biology	20	4	3	2	3
Computer Vision and Pattern Recognition	20	3	3	1	3
Condensed Matter Physics	20	2	2	1	1
Endocrinology	20	1	2	1	2
Geology	20	0	0	0	0
Geometry and Topology	20	3	6	2	6
Hardware and Architecture	20	0	0	0	0
History	20	4	7	4	8
Inorganic Chemistry	20	3	5	3	5
Library and Information Sciences	20	3	4	3	5
Literature and Literary Theory	20	2	4	4	4
Logic	20	3	3	3	4
Marketing	20	1	1	1	1
Nuclear and High Energy Physics	20	3	3	2	5
Numerical Analysis	20	0	1	0	1
Oceanography	20	4	4	3	5
Ophthalmology	20	0	1	0	1
Organic Chemistry	20	1	2	1	2
Pharmacology	20	2	3	2	2
Plant Science	20	2	3	2	2
Social Psychology	20	2	1	1	3
Sociology and Political Science	20	2	3	2	4
Statistics and Probability	20	1	2	1	2
Surgery	20	0	0	0	0
Общее количество	600	72	89	65	98
Доля от общего количества, %		12	15	11	16

2.2. Классификация тестовых файлов с несколькими категориями

Результаты классификации тестовых файлов с несколькими категориями приведены в табл. 7.

Ошибочно определелись 23% (140 из 600) тестовых файлов. При этом неверно определилось научное направление у 6 % (37 из 600) тестов. Стоит отметить, что в некоторых случаях эта ошибка возникала из-за категорий, близких по терминологии, но находящихся в разных научных направлениях. Например, категория Aquatic Science из направления Life Sciences и категория Oceanography из Physical Sciences.

В качестве примера приведем тестовый файл с eid= 2-s2.0-57649228732, у которого указаны две категории: Aquatic Science и Plant Science. Метод определил категорию Oceanography. Текст аннотации приведен на рис. 7.

В других случаях ошибки её характер определить не удалось. Так, у тестового файла с eid= 2-s2.0-79451471007

вместо категорий *Library and Information Sciences* и *History* из направления *Social Sciences* определилась категория *Aquatic Science* научного направления *Life Sciences* (рис. 8).

2.3. Влияния количества ядер на качество классификации

На рис. 9 изображена зависимость количества ошибок от количества ядер, участвующих в классификации. Результаты показывают, что при расширении числа категорий также увеличивается и число ошибок.

Если при 5 ядрах ошибочно определился только один тест из категории *Algebra and Number Theory*, то при 30 категориях количество ошибок возросло до 13. При этом как при 5, так и при 30 ядрах безошибочно определялись тесты из категории *Surgery*.

Таким образом, изначальная выборка количества и состава категорий влияет на точность классификации.

Таблица 7

Результаты классификации файлов с несколькими категориями

Группа теста	Количество тестов	Доля от 600 тестов, %
Определилось не менее 50% категорий	413	69
Определилась хотя бы одна категория	47	8
Все категории определелись неверно	140	23

*In California, the toxic algal species of primary concern are the dinoflagellate *Alexandrium catenella* and members of the pennate diatom genus *Pseudo-nitzschia*, both producers of potent neurotoxins that are capable of sickening and killing marine life and humans. During the summer of 2004 in Monterey Bay, we observed a change in the taxonomic structure of the phytoplankton community-the typically diatom-dominated community shifted to a red tide, dinoflagellate-dominated community. Here we use a 6-year time series (2000-2006) to show how the abundance of the dominant harmful algal bloom (HAB) species in the Bay up to that point, *Pseudo-nitzschia*, significantly declined during the dinoflagellate-dominated interval, while two genera of toxic dinoflagellates, *Alexandrium* and *Dinophysis*, became the predominant toxin producers. This change represents a shift from a genus of toxin producers that typically dominates the community during a toxic bloom, to HAB taxa that are generally only minor components of the community in a toxic event. This change in the local HAB species was also reflected in the toxins present in higher trophic levels. Despite the small contribution of *A. catenella* to the overall phytoplankton community, the increase in the presence of this species in Monterey Bay was associated with an increase in the presence of paralytic shellfish poisoning (PSP) toxins in sentinel shellfish and clupeoid fish. This report provides the first evidence that PSP toxins are present in California's pelagic food web, as PSP toxins were detected in both northern anchovies (*Engraulis mordax*) and Pacific sardines (*Sardinops sagax*). Another interesting observation from our data is the co-occurrence of DA and PSP toxins in both planktivorous fish and sentinel shellfish. We also provide evidence, based on the statewide biotoxin monitoring program, that this increase in the frequency and abundance of PSP events related to *A. catenella* occurred not just in Monterey Bay, but also in other coastal regions of California. Our results demonstrate that changes in the taxonomic structure of the phytoplankton community influences the nature of the algal toxins that move through local food webs and also emphasizes the importance of monitoring for the full suite of toxic algae, rather than just one genus or species. © 2008 Elsevier B.V.*

Рис. 7. Аннотация публикации с eid= 2-s2.0-57649228732

Current records management methodologies and practices suffer from an inadequate understanding of the 'human activity systems' where records managers operate as 'mediators' between a number of complex and interacting factors. Although the records management and archival literature recognizes that managing the active life of the records is fundamental to their survival as meaningful evidence of activities, the context where the records are made, captured, used, and selectively retained is not explored in depth. In particular, the various standards, models, and functional requirement lists, which occupy a vast portion of that literature, especially in relation to electronic records, do not seem to be capable of framing records-related 'problems' in ways that account for their dynamic and multiform nature. This paper introduces the idea that alternative, 'softer' approaches to the analysis of organizational functions, structures, agents, and artifacts may usefully complement the 'hard', engineering-like approaches typically drawn on by information and records specialists. Three interrelated theoretical and methodological frameworks—namely, Soft Systems Methodology, Adaptive Structuration Theory, and Genre Theory—are discussed, with the purpose of highlighting their contributions to our understanding of the records context. © 2010 Springer Science+Business Media B.V.

Рис. 8. Аннотация публикации с eid= 2-s2.0-79451471007

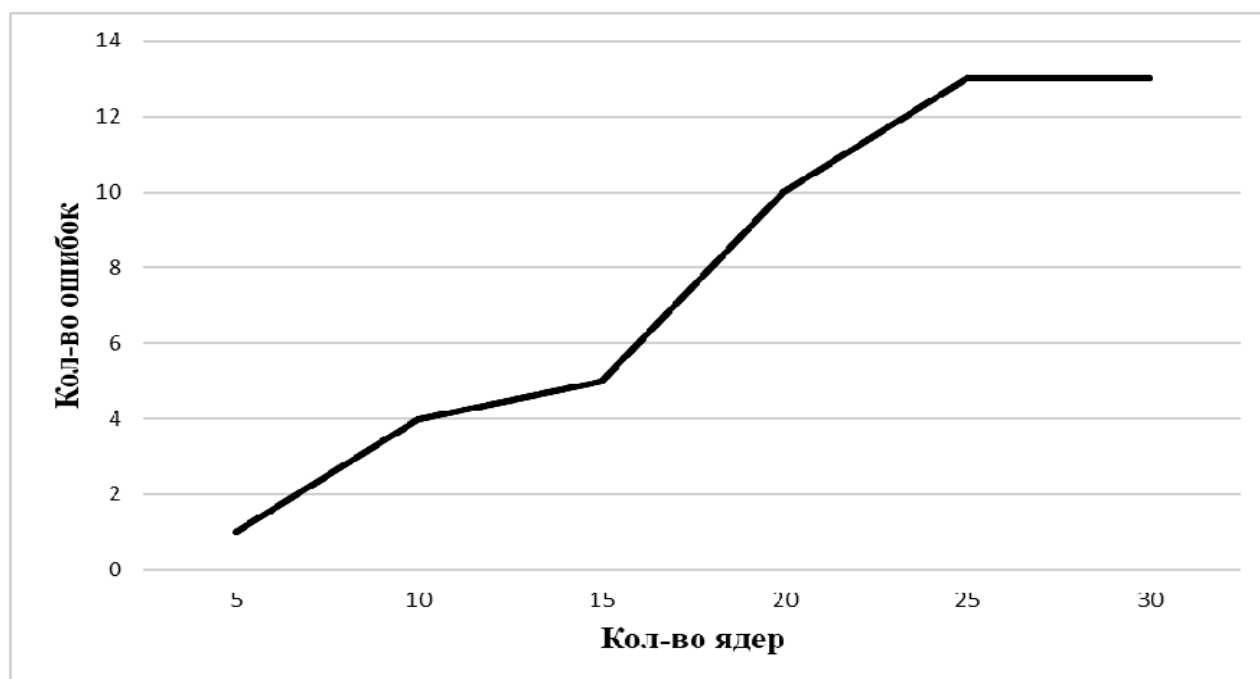


Рис. 9. Зависимость качества классификации от количества ядер

ЗАКЛЮЧЕНИЕ

Исследование показало, что метод классификации научных текстов, основанный на алгоритмах сжатия информации, показывает разную эффективность в зависимости от условий, в которых проводится классификация. При произвольных аннотациях публикаций, индексируемых в ББД *Scopus*, в ядре количество ошибок классификации составило 32%. Подбор ядер из аннотаций высокоцитируемых публикаций позволил сократить их количество до 12%. Это, предположительно, связано с тем, что в таких аннотациях используется более характерная

для области наук лексика, которая наследуется и остальными публикациями.

Удаление стоп-слов и названий издательств в большинстве случаев негативно сказывается на качестве классификации. Возможно, это связано с тем, что в названии издательств используются специальные термины, а также с тем, что длина аннотации уменьшается.

Помимо состава ядер на качество классификации влияет и изначальное количество категорий: чем меньше категорий участвует при классификации и чем больше терминологическое различие между ними, тем выше качество этой классификации.

Тем не менее, исходная классификация статей была сделана на основе журнальной классификации *Scopus*, которая также имеет ошибки, и их количество крайне трудно измерить. Это вносит свою объективную погрешность в наши исследования.

СПИСОК ЛИТЕРАТУРЫ

1. Барахнин В.Б., Кожемякина О.Ю., Пас-тушков И.С., Рычкова Е.В. Автоматизированная классификация русских поэтических текстов по жанрам и стилям // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. – 2017. – Т.15, №3. – С. 13-23.
2. Батура Т.В. Формальные методы определения авторства текстов // Вестник НГУ. Серия: Информационные технологии. – 2012. – Т.10, №4. – С. 81-94.
3. Dos Santos C.N., Gatti M. Deep convolutional neural networks for sentiment analysis of short texts // COLING 2014 – 25th International Conference on Computational Linguistics, Proceedings of COLING 2014: Technical Papers. – 2014. – P. 69-78.
4. Sriram B., Fuhry D., Demir E., Ferhatos-manoglu H., Demirbas M. Short text classification in twitter to improve information filtering // SIGIR 2010 Proceedings – 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2010. – P. 841-842.
5. Kiritchenko S., Zhu X., Mohammad S.M. Sentiment analysis of short informal texts // Journal of Artificial Intelligence Research. – 2014. – Vol.50. – P. 723-762.
6. Рябко Б.Я., Гуськов А.Е., Селиванова И.В. Теоретико-информационный метод классификации текстов // Пробл. передачи информ. – 2017. – Т. 53, №3. – С. 100–111; Ryabko B.Y., Gus'kov A.E., Selivanova I.V. Information-Theoretic Method for Classification of Texts // Problems of Information Transmission. – 2017. – Vol.53, Iss. 3. – P.294-304. – URL: <https://link.springer.com/article/10.1134/S0032946017030115>.
7. Селиванова И.В., Рябко Б.Я., Гуськов А.Е. Классификация посредством компрессии: применение методов теории информации для определения тематики научных текстов // Научно-техническая информация. Сер. 2. – 2017. – № 6. – С.8-15; Selivanova I.V., Ryabko B.Ya., Guskov A.E. Classification by Compression: Application of Information-Theory Methods for the Identification of Themes of Scientific Texts // Automatic Documentation and Mathematical Linguistics. – 2017. – Vol. 51, № 3. – P.120-126.
8. Hall G.M. How to write a paper. – A John Wiley & Sons, Ltd., Publication, 2013. – 170 c.
9. Perianes-Rodriguez A., Ruiz-Castillo J. A comparison of the Web of Science and publication-level classification systems of science // Journal of Informetrics. – 2017. – Vol. 11, Iss.1. – P.32-45.
10. Shu F., Julien C.A., Zhang L., Qiu J., Zhang J., Lariviere V. Comparing journal and paper level classifications of science // Journal of Informetrics. – 2019. – Vol.13, Iss.1. – P.202-209.
11. Topic Prominence in Science стал доступен пользователям SciVal. – URL: <http://elsevierscience.ru/news/428/topic-prominence-in-science-stali-dostupny-polzovatelyam-scival> (дата обращения: 14.10.2019).
12. Waltman L., van Eck N.J. A new methodology for constructing a publication-level classification system of science // Journal of the American Society for Information Science and Technology. – 2012. – Vol.63, Iss.12. – P.2378-2392.
13. УДК, ББК, ISBN – обязательные элементы выходных сведений издания. – URL: <https://www.ipu.ru/structure/information-services/polygraphy/20804> (дата обращения: 14.10.2019).
14. 1297.0 – Australian and New Zealand Standard Research Classification (ANZSRC), 2008. – URL: <https://www.abs.gov.au/Ausstats/abs@.nsf/Latestproducts/1297.0Main%20Features32008?opendocument&tabname=Summary&prodno=1297.0&issue=2008> (дата обращения: 14.10.2019).
15. Паспорта научных специальностей. – URL: <http://arhvak.minobrnauki.gov.ru/316> (дата обращения: 14.10.2019).
16. ОККО – Общероссийский классификатор специальностей по образованию. – URL: <https://classifikators.ru/okso> (дата обращения: 14.10.2019).
17. ГРНТИ – Государственный рубрикатор научно-технической деятельности 2019. – URL: <http://grnti.ru/> (дата обращения: 14.10.2019).
18. Revised field of science and technology (FOS) classification in the Frascati Manual. – URL: <http://www.oecd.org/science/inno/38235147.pdf> (дата обращения: 14.10.2019).
19. Proposed international standard nomenclature for fields of science and technology. – URL: <https://unesdoc.unesco.org/ark:/48223/pf0000082946> (дата обращения: 14.10.2019).
20. Парфенова С.Л., Долгова В.Н., Богатов В.В., Халтакшинова Н.В., Коробатов В.Я. Методический подход к формированию рубрикаторов-переходников для анализа направлений *Web of Science* и *Scopus* в разрезе приоритетов Стратегии научно-технологического развития РФ // Экономика науки. – 2018. – Т.4, №2. – С.143-153.
21. Scopus. Руководство по охвату контента. – URL: http://elsevierscience.ru/files/Scopus_Content_Guide_Rus_2017.pdf. – С. 21 (дата обращения: 14.10.2019).
22. Wang Q., Waltman L. Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus // Journal of Informetrics. – 2016. – Vol.10, Iss.2. – P.347-364.
23. Mendes A.C. Science classification, visibility of the different scientific domains and impact on scientific development Scopus // Revista de Enfermagem Referência. – 2016. – Vol.10, Iss.4. – P.143-149.
24. Martínez-Frías J., Hochberg D. Classifying science and technology: Two problems with the UNESCO system // Interdisciplinary Science Reviews. – 2007. – Vol.32, Iss.4. – P.315-319.
25. Tan S. Neighbor-weighted K-nearest neighbor for unbalanced text corpus // Expert Systems with Applications. – 2005. – Vol.28, Iss.4. – P.667-671.

26. Jiang L., Li C., Wanga S., Zhanga L. Deep feature weighting for naive Bayes and its application to text classification // *Engineering Applications of Artificial Intelligence*. – 2016. – Vol.52. – P.26-39.
27. Wang S., Manning C.D. Baselines and bigrams: Simple, good sentiment and topic classification // 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference. – 2012. – Vol.2. – P.90-94.
28. Lai S., Xu L., Liu K., Zhao J. Recurrent convolutional neural networks for text classification // In *Twenty-Ninth AAAI Conference on Artificial Intelligence*. – 2015. – P. 2267-2273.
29. Li S., Hu J., Cui Y., Hu J. DeepPatent: patent classification with convolutional neural networks and word embedding // *Scientometrics*. – 2018. – Vol.117, Iss.2. – P.721-744.
30. Li Y.H., Jain A.K. Classification of Text Documents // *The Computer Journal*. – 1998. – Vol.41, Iss.8. – P.537-546.
31. Xia R., Zong C., Li S. Ensemble of feature sets and classification algorithms for sentiment classification // *Information Sciences*. – 2011. – Vol.181, Iss.6. – P.1138-1152.
32. Šubelj L., van Eck N.J., Waltman L. Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods // *PLoS ONE*. – 2016. – Vol.11, Iss.4. – P.1-23.
33. Liu X., Yu S., Moreau Y., Janssens F., Moor B.D., Glänzel W. Hybrid Clustering by Integrating Text and Citation Based Graphs in Journal Database Analysis // *IEEE International Conference on Data Mining Workshops*, Miami. – 2009. – P.521-526.
34. Waltman L., Boyack K.W., Colavizza G., van Eck N.J. A principled methodology for comparing relatedness measures for clustering publications // *arXiv:1901.06815*. – URL: <https://arxiv.org/ftp/arxiv/papers/1901/1901.06815.pdf> (дата обращения: 14.10.2019).
35. Boyack K.W., Newman D., Duhon R.J., Klavans R., Patek M., Biberstine J.R., Schijvenaars B., Skupin A., Ma N., Börner K. Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches // *PLoS ONE*. – 2011. – Vol.6, Iss.6. – P.1-11.
36. Zhang B., Chen Y., Fan W., Fox E.A., Gonçalves M.A., Cristo M., Calado P. Intelligent GP fusion from multiple sources for text classification // *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management*, Bremen, Germany, October 31 – November 5. – 2005.
37. Tshitoyan V., Dagdelen J., Weston L., Dunn A., Rong Z., Kononova O., Persson K.A., Ceder G., Jain A. Unsupervised word embeddings capture latent knowledge from materials science literature // *Nature*. – 2019. – Vol.571. – P.95-98.
38. Borrajo L., Romero R., Iglesias E.L., Redondo Marey C.M. Improving imbalanced scientific text classification using sampling strategies and dictionaries // *Journal of Integrative Bioinformatics*. – 2011. – Vol.8, Iss.3. – P.1-15.
39. Sinclair G., Webber B. Classification from full text: A comparison of canonical sections of scientific papers // In *Proc. of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, Geneva, Switzerland. – 2004. – P. 66-69.
40. Riloff E. Little words can make a big difference for text classification // *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, Seattle, Washington, USA. – 1995. – P. 130-136.

Материал поступил в редакцию 15.10.19.

Сведения об авторах

СЕЛИВАНОВА Ирина Вячеславовна – младший научный сотрудник ГПНТБ СО РАН, г. Новосибирск
e-mail: selivanova@spsl.nsc.ru

КОСЯКОВ Денис Викторович – заместитель директора по развитию ГПНТБ СО РАН
e-mail: kosyakov@spsl.nsc.ru

ГУСЬКОВ Андрей Евгеньевич – кандидат технических наук, директор ГПНТБ СО РАН
e-mail: guskov@spsl.nsc.ru

Указатель статей, опубликованных в сборнике «Научно-техническая информация», и Авторский указатель за 2019 год

Указатель статей

ДСМ-МЕТОД АВТОМАТИЗИРОВАННОЙ ПОДДЕРЖКИ ИССЛЕДОВАНИЙ И ЕГО ПРИМЕНЕНИЕ В ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМАХ ДЛЯ МЕДИЦИНЫ		Таран В.В. О развитии концепции Всемирной паутины	5 (2)	1
Финн В.К. Об эвристиках ДСМ- исследований (дополнения к статьям)		Алейников А.В., Милецкий В.П., Пименов Н.П., Стребков А.И. Феномен «фейк-ньюс» и трансформация информационных стратегий в цифровом обществе	6 (1)	1
Чебанов Д.К., Михайлова И.Н. Интеллектуальный анализ данных пациентов с меланомой для поиска маркеров заболевания и значимых генов	10 (2) 1* 10 (2) 35	Крымская А.С. Сэмюэль Клеменет Брэдфорд (к 140-летию со дня рождения)	7 (1)	1
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	10 (2) 41	Берестова Т.Ф. О создании дефиниции информации на основе выявления её сущности	8 (2)	5
ОБЩИЙ РАЗДЕЛ		Серов Н.В. Концептуализация предикатов спора Гёте с Ньютоном о цвете	9 (1)	1
Семенюк Э.П. Информационный эффект трансдисциплинарности в концепции устойчивого развития		Сюнтюрено О.В., Дмитриева Е.Ю. Государственная система научно- технической информации в структуре задач цифровой экономики	10 (1)	1
Калачихин П.А. Моделирование динамики исследовательских фронтов	1 (1) 1 1 (2) 1	Соколова И.С. Научно-популярные ресурсы России: журналы и интернет- медиапроекты	11 (1)	1
Астахова Л.В. Визуализация информационных ресурсов в условиях цифровизации сферы познания (Обзор)	2 (1) 1	Астахова Л.В., Медведев И.А. Инструментальный мониторинг соответствия веб-сайтов требованиям законодательства в области персональных данных	12 (1)	1
Урсул А.Д. Информационная природа эволюции и освоения мира: концептуальная гипотеза	2 (2) 1	Гоннова С.М., Разуваева Е.Ю. Национальные системы научно- технической информации – потенциал для развития научной дипломатии в СНГ	12 (2)	1
Нестеров А.В. Об унификации концептуальной модели мета-онтологии	3 (1) 1	Маркусова В.А., Миндели Л.Э., Рубвальтер Д.А., Золотова А.В., Акоев М.А., Либкинд А.Н. Динамика публикационной активности России в области энергии и топлива за 2008-2017 гг. по данным <i>Web of Science</i> и <i>InCites</i>		
Гиляревский Р.С., Либкинд А.Н., Маркусова В.А. Динамика публикационной активности России в 1993-2017 гг. по данным <i>Web of Science</i>	3 (2) 1	ОРГАНИЗАЦИЯ ИНФОРМАЦИОННОЙ РАБОТЫ		
Редькина Н.С. Современные тенденции в управлении исследовательскими данными	4 (1) 1	Соколов И.А., Григорьев О.Г., Тихомиров И.А., Девяткин Д.А., Суворов Р.Е., Жебель В.В. О создании национальной системы выявления приоритетов научных исследований и разработок	1 (1)	14
Астахова Л.В. Информационно- психологическая теория духовного развития личности в эпоху цифровой культуры. К 95-летию со дня рождения Ю.С. Зубова	5 (1) 1			

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

Крулев А.А. Роль наукометрии в стратегическом планировании научной деятельности	1 (1) 21	Демшина Н.В., Мосунова Л.А. Изучение уровней смыслового восприятия информации в дополнительном образовании	5 (1) 29
Кий М.И. Взаимодействие библиотеки с пользовательской аудиторией средствами виртуальной среды	1 (1) 27	Антопольский А.Б., Ефременко Д.В. О создании современной цифровой инфраструктуры для хранения и анализа научно-технической информации	6 (1) 8
Комарица В.Н., Малинина В.Н. Опыт формирования международного редакционного совета научного журнала	1 (1) 31	Шемберко Л.В., Уварова Т.Б. Информационно-аналитическое обеспечение археологических исследований на основе цифровых технологий	6 (1) 18
Ибраев А.Ж., Пономарева Н.И., Козбагарова Г.А. Анализ публикационной активности государств – членов Евразийского экономического союза	2 (1) 10	Гиляревский Р.С., Мельникова Е.В. Национальные электронные библиотеки России и США	7 (1) 8
Сладкова О.Б., Пирумова Л.Н., Пирумов А.А. Современные средства поиска для удовлетворения информационных потребностей специалистов (на примере агропромышленного комплекса)	2 (1) 18	Сикорская О.Н., Бовкунович М.А. О направлениях научно-технического развития Республики Беларусь	7 (1) 14
Дорофеева В.И., Никольский Д.Н., Федяев Ю.С. Информационная система мониторинга научно-исследовательской деятельности в инновационной структуре вуза	2 (1) 23	Иншакова Н.Г., Панкеев И.А. Редактирование научного произведения: нормативно-правовые аспекты	7 (1) 22
Анисимова А.Э., Рязанова А.А. О методике сравнительного квалификационного анализа требований к профессиональным навыкам с целью коррекции национальных образовательных программ	2 (1) 29	Шамаев В.Г., Шуко Ю.Н. Банк данных ВИНТИ РАН. Проблемы и перспективы	8 (1) 15
Ловцов А.И., Угринович Е.В. Интеграция национальных репозиторий научной информации открытого доступа стран – членов Международного центра научно-технической информации	3 (1) 6	Парамонова И.Е. Информационное взаимодействие: критерии выбора коммуникационных каналов в научно-технической библиотеке	8 (1) 21
Мохначева Ю.В., Бескаравайная Е.В. Профессиональная деятельность за рубежом научных диаспор Пущинского научного центра РАН	3 (1) 13	Саркисян Д.Б. Цифровой идентификатор DOI – инструмент навигации по научным публикациям в Интернете	8 (1) 27
Зубехина Т.В., Колесник А.В., Маркивская Л.Л. Технология веб-квест в электронном обучении	3 (1) 20	Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А., Косинов А.В. Курирование цифровых научных данных	9 (1) 12
Сюнтюрченко О.В. Цифровая среда: аналитическая постобработка информации с использованием методов наукометрии и анализа данных	4 (1) 8	Селиванова И.В., Косяков Д.В., Гуськов А.Е. Влияние ошибок в базе данных <i>Scopus</i> на оценку результативности научных исследований	9 (1) 25
Неретин О.П., Лопатина Н.В., Зубов Ю.С. Цифровизация сферы интеллектуальной собственности: от научного обоснования к практической реализации	4 (1) 17	Яшалова Н.Н., Шрейдер Н.В., Яковлева Е.Н. Цифровая грамотность общества: ситуация, проблемы и перспективы на современном этапе научно-технического прогресса	10 (1) 6
Грузова А.А. Преодоление информационных барьеров в технической коммуникации	5 (1) 8	Сысоев А.Н., Корнилова М.Б., Толмачёв А.О. Научная поддержка действующей классификации: на примере УДК	10 (1) 12
Бескаравайная Е.В., Харыбина Т.Н. Создание модели мониторинга научных исследований	5 (1) 19	Мельникова Е.В. Опыт формирования цифровых коллекций ведущих библиотек России и США	11 (1) 6
		Алевизаки О.Р., Кара-Мурза Е.С., Ломыкина Н.Ю., Агафонова М.А. Маркетинг в социальных сетях для продвижения брендов средств массовой информации	11 (1) 12
		Шевченко Л.Б. Дизайн и удобство использования библиотечных веб-сайтов	11 (1) 21
		Калачихин П.А. Оценка исследователей по критериям репутационной ответственности	12 (1) 19

ДОКУМЕНТАЛЬНЫЕ ИСТОЧНИКИ ИНФОРМАЦИИ

- Куш Г.А., Батюшко А.А.** 3 (1) 26
О распределении отечественных патентных документов, вошедших в базу перспективных изобретений, по приоритетным направлениям и критическим технологиям
- Трищенко Н.Д.** 4 (1) 23
Трансформация системы научной коммуникации под влиянием открытого доступа: текущий статус, предпосылки перемен, эффекты и перспективы развития
- Мохначева Ю.В., Цветкова В.А.** 6 (1) 28
Представленность статей российских авторов в мировом потоке научных публикаций по *Web of Science Core Collection* (2010-2017)
- Гуреев В.Н., Лакизо И.Г., Мазов Н.А.** 10 (1) 19
Неэтичное авторство в научных публикациях (обзор проблемы)

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

- Артемова С.В., Артемов А.А., Каменская М.А.** 2 (2) 9
Интеллектуальная информационно-управляющая система энергосберегающего управления
- Толчеев В.О.** 5 (2) 10
Исследование и анализ предметной области «глубокое обучение»
- Гусакова С.М., Охлупина А.Н.** 6 (2) 1
Интеллектуальная ДСМ-система как средство автоматизированной поддержки научных исследований в почерковедении
- Гусакова С.М., Добрынин Д.А., Харчевникова Н.В.** 9 (2) 1
Сравнение языков представления данных в задаче «структура–активность»
- Забейайло М.И., Трунин Ю.Ю.** 12 (2) 12
К проблеме доказательности медицинского диагноза: интеллектуальный анализ эмпирических данных о пациентах в выборках ограниченного размера

ИНФОРМАЦИОННЫЕ СИСТЕМЫ

- Абрамский М.М., Циммерман А.М., Альмухаметова А.А., Алтынбаева Д.Т.** 3 (2) 14
Онтологический подход к проектированию образовательных программ в цифровых средах
- Водяхо А.И., Осипов В.Ю., Жукова Н.А., Червонцев М.А.** 4 (2) 1
Когнитивные технологии в управлении мониторингом
- Шведенко В.Н., Волков А.А.** 5 (2) 22
Метод создания цифрового двойника на основе агрегации информационных объектов

- Грушо А.А., Забейайло М.И., Писковский В.О., Сенчило В.В., Тимонина Е.Е.** 6 (2) 9
Подходы к интеграции информационных систем цифровой экономики
- Фаткиева Р.Р.** 8 (2) 18
Система показателей информационной безопасности промышленных предприятий
- Съедин Д.Ю.** 11 (2) 1
Программно-вычислительный комплекс для решения задачи интеграции атрибутивных данных информационных объектов
- Шведенко В.Н., Мозохин А.Е.** 12 (2) 19
Методические основы формирования информационного пространства и цифровых двойников объектов умного дома

ИНФОРМАЦИОННЫЙ ПОИСК

- Бениаминов Е.М.** 7 (2) 1
Алгебра скрытых отношений как средство моделирования статистических баз данных

ИНФОРМАЦИОННЫЙ АНАЛИЗ

- Брумштейн Ю.М., Васильев Н.В.** 1 (2) 9
Анализ структуры сайтов академий наук некоторых зарубежных стран
- Лимарев П.В., Лимарева Ю.А.** 1 (2) 25
Многофакторная модель информационного рынка
- Жукова Н.А., Андриянова Н.Р.** 2 (2) 18
Проблема когнитивного мониторинга распределенных объектов
- Виноградов Д.В.** 2 (2) 30
Субмультипликативность и остановка спаривающей цепи Маркова для ВКФ-метода
- Шведенко В.Н., Шведенко В.В., Щекочихин О.В.** 3 (2) 21
Использование структурного и параметрического полиморфизма при создании цифровых двойников
- Кулямина Е.Ю., Еркимбаев А.О.** 3 (2) 25
Электронная библиотека как средство публикации научных результатов в сети: опыт распространения данных по аномалиям термического расширения вещества
- Брумштейн Ю.М., Васильев Н.В.** 4 (2) 13
Академии наук стран американского континента: анализ информационного присутствия в интернет-пространстве
- Максимов Н.В., Голицына О.Л., Монанков К.В., Лебедев А.А., Баль Н.А., Кюрчева С.Г.** 7 (2) 8
Средства семантического поиска, основанные на онтологических представлениях документальной информации
- Калачихин П.А.** 8 (2) 28
Сопоставление результатов исследований по научным сообществам

Виноградов Д.В. Ещё один вероятностный алгоритм для вычисления сходств	9 (2) 10	Фесенко В.П. Винительный падеж как маркер определенности объекта при переходном глаголе с отрицанием (на примере терминов родства)	7 (2) 20
Брумштейн Ю.М., Васильев Н.В. Академии наук стран Западной Европы: анализ направлений деятельности, информационного наполнения сайтов и их вебометрических показателей	9 (2) 13	Полохин С.Н., Рябко Б.Я., Савина Н.Н. Развитие теоретико-информационного метода классификации текстов и его применение к анализу авторства	7 (2) 33
Нестерова Е.И. Особенности онтологии исходных данных и метаданных в области систем виртуальной реальности	11 (2) 10	Боярский К.К., Каневский Е.А. Семантика устойчивых словосочетаний с глаголами	11 (2) 23
Жукова Н.А. Общая и частные задачи многоуровневого синтеза моделей объектов мониторинга	11 (2) 16	Селиванова И.В., Косяков Д.В., Гуськов А.Е. Классификация научных текстов на основе компрессии аннотаций публикаций	12 (2) 25

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

Кустова Г.И. Парадоксы метатекстовых конструкций с ментальными предикатами	1 (2) 29
Хайруллин В.И. Перевод и локализация: о переводческой компрессии и расширении информации	2 (2) 34
Гольдштейн С.Л., Кудрявцев А.Г., Грицюк Е.М., Ходенева М.А. Фильтры текстов, отличающихся по смыслу от интересов пользователя	2 (2) 38
Шереметьева С.О. Экстракция многокомпонентных терминов и ключевых слов из многоязычной патентной документации	4 (2) 25
Кустова Г.И. Относительные наречия и относительные прилагательные по данным национального корпуса русского языка	4 (2) 34
Егорова М.А., Егоров А.А. Прародина народов носителей праиндоевропейского языка: математические модели для исследования лингвистической информации	5 (2) 27
Краснов Ф.В., Шварцман М.Е., Диментов А.В., Сень А.И. Тематическая когерентность двуязычного корпуса научных статей (на примере нефтегазовой отрасли)	6 (2) 22
Петрова Е.Т., Петров Т.Г., Чебанов С.В., Мошкин С.В. Метод кодирования многокомпонентных объектов (<i>RHA</i>) и его применение для упорядочения шрифтов прямого начертания	6 (2) 28

СПРАВОЧНО-ИНФОРМАЦИОННЫЙ РАЗДЕЛ

Воробьев А.А., Горский А.В., Козырев В.А. Информационные ресурсы для методики оценки показателей надежности тягового подвижного состава	3 (2) 32
Арутюнов В.В. Об итогах Второй международной научно-практической конференции «Информационная безопасность: вчера, сегодня, завтра»	7 (1) 26
Джиги А.А., Майстрович Т.В. Новый уровень понимания библиотечно-информационной услуги	9 (1) 33
Лесников С.В. Виды разметок текстовых корпусов русского языка	9 (2) 27
Смирнов Ю.В. О классификаторе программ для электронных вычислительных машин и баз данных	11 (2) 32

НАМ ПИШУТ

Ханжин А.Г., Кожокару А.А. О сохранности информации	2 (1) 36
Миралиев К.Х., Алибаева М.М., Азизова Р.Н. Информационная обеспеченность науки в Республике Таджикистан	12 (1) 28

Авторский указатель

Абрамский М.М.	3 (2) 14	Девяткин Д.А.	1 (1) 14	Лебедев А.А.	7 (2) 8
Агафонова М.А.	11 (1) 12	Демшина Н.В.	5 (1) 29	Лесников С.В.	9 (2) 27
Азизова Р.Н.	12 (1) 28	Джиги А.А.	9 (1) 33	Лесько К.А.	10 (2) 41
Алевизаки О.Р.	11 (1) 12	Диментов А.В.	6 (2) 22	Либкинд А.Н.	3 (2) 1
Алейников А.В.	6 (1) 1	Дмитриева Е.Ю.	9 (1) 1		12 (2) 1
Алибаева М.М.	12 (1) 28	Добрынин Д.А.	9 (2) 1	Лимарев П.В.	1 (2) 25
Алтынбаева Д.Т.	3 (2) 14	Дорофеева В.И.	2 (1) 23	Лимарева Ю.А.	1 (2) 25
Альмухаметова А.А.	3 (2) 14			Ловцов А.И.	3 (1) 6
Акеев М.А.	12 (2) 1	Егоров А.А.	5 (2) 27	Ломыкина Н.Ю.	11 (1) 12
Андриянова Н.Р.	2 (2) 18	Егорова М.А.	5 (2) 27	Лопатина Н.В.	4 (1) 17
Анисимова А.Э.	2 (1) 29	Еркимбаев А.О.	3 (2) 25		
Антопольский А.Б.	6 (1) 8		9 (1) 12	Мазов Н.А.	10 (1) 19
Артемов А.А.	2 (2) 9	Ефременко Д.В.	6 (1) 8	Майстрович Т.В.	9 (1) 33
Артемова С.В.	2 (2) 9			Максимов Н.В.	7 (2) 8
Арутюнов В.В.	7 (1) 26	Жебель В.В.	1 (1) 14	Малинина В.Н.	1 (1) 31
Астахова Л.В.	2 (1) 1	Жукова Н.А.	2 (2) 18	Маркивская Л.Л.	3 (1) 20
	5 (1) 1		4 (2) 1	Маркусова В.А.	3 (2) 1
	11 (1) 1		11 (2) 16		12 (2) 1
Баль Н.А.	7 (2) 8	Забейайло М.И.	6 (2) 9	Медведев И.А.	11 (1) 1
Батюшко А.А.	3 (1) 26		12 (2) 12	Мельникова Е.В.	7 (1) 8
Бениаминов Е.М.	7 (2) 1	Зицерман В.Ю.	9 (1) 12		11 (1) 6
Берестова Т.Ф.	8 (1) 5	Золотова А.В.	12 (2) 1	Милецкий В.П.	6 (1) 1
Бескаравайная Е.В.	3 (1) 13	Зубехина Т.В.	3 (1) 20	Миндели Л.Э.	12 (2) 1
	5 (1) 19	Зубов Ю.С.	4 (1) 17	Миралиев К.Х.	12 (1) 28
Бовкунович М.А.	7 (1) 14			Михайлова И.Н.	10 (2) 35
Боярский К.К.	11 (2) 23	Ибраев А.Ж.	2 (1) 10	Мозохин А.Е.	12 (2) 19
Брумштейн Ю.М.	1 (2) 9	Иншакова Н.Г.	7 (1) 22	Монанков К.В.	7 (2) 8
	4 (2) 13			Мосунова Л.А.	5 (1) 29
	9 (2) 13	Калачихин П.А.	1 (2) 1	Мохначева Ю.В.	3 (1) 13
Варварина Г.Г.	10 (2) 41		8 (2) 28		6 (1) 28
Васильев Н.В.	1 (2) 9		12 (1) 19	Мошкин С.В.	6 (2) 28
	4 (2) 13	Каменская М.А.	2 (2) 9	Неретин О.П.	4 (1) 17
	9 (2) 13	Каневский Е.А.	11 (2) 23	Нестеров А.В.	3 (1) 1
Виноградов Д.В.	2 (2) 30	Кара-Мурза Е.С.	11 (1) 12	Нестерова Е.И.	11 (2) 10
	9 (2) 10	Кий М.И.	1 (1) 27	Никольский Д.Н.	2 (1) 23
Винокурова Л.В.	10 (2) 41	Кобзев Г.А.	9 (1) 12		
Водяхо А.И.	4 (2) 1	Кожокару А.А.	2 (1) 36	Осипов В.Ю.	4 (2) 1
Волков А.А.	5 (2) 22	Козбагарова Г.А.	2 (1) 10	Охлупина А.Н.	6 (2) 1
Воробьев А.А.	3 (2) 32	Козырев В.А.	3 (2) 32		
		Колесник А.В.	3 (1) 20	Панкеев И.А.	7 (1) 22
Гиляревский Р.С.	3 (2) 1	Комарица В.Н.	1 (1) 31	Парамонова И.Е.	8 (1) 21
	7 (1) 8	Корнилова М.Б.	10 (1) 12	Петров Т.Г.	6 (2) 28
Голицына О.Л.	7 (2) 8	Косинов А.В.	9 (1) 12	Петрова Е.Т.	6 (2) 28
Гольдштейн С.Л.	2 (2) 38	Косяков Д.В.	9 (1) 25	Пименов Н.П.	6 (1) 1
Гоннова С.М.	12 (1) 1		12 (2) 25	Пирумов А.А.	2 (1) 18
Горский А.В.	3 (2) 32	Краснов Ф.В.	6 (2) 22	Пирумова Л.Н.	2 (1) 18
Григорьев О.Г.	1 (1) 14	Крулев А.А.	1 (1) 21	Писковский В.О.	6 (2) 9
Грицюк Е.М.	2 (2) 38	Крымская А.С.	7 (1) 1	Полохин С.Н.	7 (2) 33
Грузова А.А.	5 (1) 8	Кудрявцев А.Г.	2 (2) 38	Пономарева Н.И.	2 (1) 10
Грушо А. А.	6 (2) 9	Кулямина Е.Ю.	3 (2) 25		
Гуреев В.Н.	10 (1) 19	Кустова Г.И.	1 (2) 29	Разуваева Е.Ю.	12 (1) 1
Гусакова С.М.	6 (2) 1		4 (2) 34	Редькина Н.С.	4 (1) 1
	9 (2) 1	Куц Г.А.	3 (1) 26	Рубвальтер Д.А.	12 (2) 1
Гуськов А.Е.	9 (1) 25	Кюрчева С.Г.	7 (2) 8	Рябко Б.Я.	7 (2) 33
	12 (2) 25	Лакизо И.Г.	10 (1) 19	Рязанова А.А.	2 (1) 29

Савина Н.Н.	7 (2) 33	Толмачёв А.О.	10 (1) 12	Цветкова В.А.	6 (1) 28
Саркисян Д.Б.	8 (1) 27	Толчеев В.О.	5 (2) 10	Циммерман А.М.	3 (2) 14
Селиванова И.В.	9 (1) 25	Трищенко Н.Д.	4 (1) 23	Чебанов С.В.	6 (2) 28
	12 (2) 25	Трунин Ю.Ю.	12 (2) 12	Чебанов Д.К.	10 (2) 35
Семенюк Э.П.	1 (1) 1	Тюляева Е.Ю.	10 (2) 41	Червонцев М.А.	4 (2) 1
Сенчило В.В.	6 (2) 9				
Сень А.И.	6 (2) 22	Уварова Т.Б.	6 (1) 18	Шамаев В.Г.	8 (1) 15
Серов Н.В.	8 (2) 5	Угринович Е.В.	3 (1) 6	Шварцман М.Е.	6 (2) 22
Сикорская О.Н.	7 (1) 14	Урсул А.Д.	2 (2) 1	Шведенко В.В.	3 (2) 21
Сладкова О.Б.	2 (1) 18			Шведенко В.Н.	3 (2) 21
Смирнов Ю.В.	11 (2) 32	Фаткиева Р.Р.	8 (2) 18		5 (2) 22
Соколов И.А.	1 (1) 14	Федяев Ю.С.	2 (1) 23		12 (2) 19
Соколова И.С.	10 (1) 1	Фесенко В.П.	7 (2) 20	Шевченко Л.Б.	11 (1) 21
Стребков А.И.	6 (1) 1	Финн В.К.	10 (2) 1	Шемберко Л.В.	6 (1) 18
Суворов Р.Е.	1 (1) 14		10 (2) 41	Шереметьева С.О.	4 (2) 25
Сысоев А.Н.	10 (1) 12			Шестерникова О.П.	10 (2) 41
Съедин Д.Ю.	11 (2) 1	Хайруллин В.И.	2 (2) 34	Шрейдер Н.В.	10 (1) 6
Сюнтюренко О.В.	4 (1) 8	Ханжин А.Г.	2 (1) 36		
	9 (1) 1	Харчевникова Н.В.	9 (2) 1	Щекочихин О.В.	3 (2) 21
		Харыбина Т.Н.	5 (1) 19	Шуко Ю.Н.	8 (1) 15
Таран В.В.	5 (2) 1	Ходенева М.А.	2 (2) 38		
Тимонина Е.Е.	6 (2) 9			Яковлева Е.Н.	10 (1) 6
Тихомиров И.А.	1 (1) 14			Яшалова Н.Н.	10 (1) 6