Формирование научного ландшафта в области сельскохозяйственных наук

Дмитрий Девяткин [a]
Младший научный сотрудник, devyatkin@isa.ru

Елена Нечаева [b]
Секретарь совета по реализации Федеральной научно-технической программы развития сельского хозяйства на 2017–2025 годы, valkmanalena@rambler.ru

Роман Суворов [a]
Младший научный сотрудник, rsuvorov@isa.ru

Илья Тихомиров [a]
Заведующий лабораторией «Интеллектуальные технологии и системы», tih@isa.ru

[a] Федеральный исследовательский центр «Информатика и управление» Российской академии наук, 117312, Москва, проспект 60-летия Октября, 9

[b] Администрация Президента Российской Федерации, 103132, г. Москва, Старая пл., 4

Аннотация

Формирование научных ландшафтов способствует повышению эффективности управления наукой и результативности деятельности ученых. В статье предлагается новая методика построения ландшафтов, опирающаяся на обработку естественного языка и тематическое моделирование. Методика особенно актуальна для ряда отраслей науки, слабо представленных в зарубежных наукометрических базах данных. Она предусматривает сбор полных текстов из достоверных источников, выделение исследовательских направлений с помощью тематического моделирования, полуавтоматическую привязку документов к показателям из наукометрических баз данных и статистический анализ. Полнотекстовая аналитика необходима ввиду низкой представленности некоторых направлений российских исследований в зарубежных наукометрических базах, недостаточного качества данных в Российском индексе научного цитирования (РИНЦ) и отсутствия единой системы классификаторов для сопоставления материалов из разных источников. Среди преимуществ методики — снятие аналитических ограничений, накладываемых составленными вручную тематическими таксономиями, интеграция разнородных данных. Она дополняет традиционный подход, основанный на использовании баз данных Scopus и Web of Science. Методика прошла апробацию на примере публикаций молодых российских ученых по сельскохозяйственным наукам, однако возможности ее применения не ограничиваются одной предметной областью. С помощью тематического моделирования выявлены шесть ключевых направлений исследований. Они заметно различаются по наукометрическим показателям, что говорит о необходимости дифференцированного подхода к оценке научной деятельности.

Дальнейшие перспективы развития представленной методики включают ее применение в других предметных областях, интеграцию нескольких источников научно-технической документации, построение сводных ландшафтов, учитывающих разнородную информацию.

Ключевые слова: анализ текстов; тематическое моделирование; научные ландшафты; сельскохозяйственные науки; публикационная активность; наукометрические базы данных; молодые ученые; РИНЦ.

Цитирование: Devyatkin D., Nechaeva E., Suvorov R., Tikhomirov I. (2018) Mapping the Research Landscape of Agricultural Sciences. *Foresight and STI Governance*, vol. 12, no 1, pp. 69–78. DOI: 10.17323/2500-2597.2018.1.69.78

Научный ландшафт (*research landscape*) раскрывает структуру и основные тенденции развития области знаний [*Oldham et al.,* 2012; *Кристофилопулос, Манцанакис,* 2016]. Процесс его построения часто называют картированием науки (*mapping studies*). Эти термины аналогичны общепринятым понятиям «патентный ландшафт» и «патентное картирование»[1], с тем отличием, что применяются не к патентам, а к научным публикациям.

Традиционно для построения патентных и научных ландшафтов используются информационно-аналитические системы — Google Patents, PatSearch, Exactus Patent, Scopus, Web of Science и т. п. Необходимое условие их применения — наличие репрезентативной базы документов, что в силу ряда причин не всегда возможно. Так, российские исследования в области сельскохозяйственных наук мало отражены в зарубежных наукометрических базах. По данным InCites, число проиндексированных Web of Science Core Collection (WoS) статей по сельскохозяйственным наукам в период с 2012 по 2016 г. не превышает 2000 единиц (в зависимости от выбранной схемы классификации данные разнятся). Наиболее популярным и быстрорастущим направлением является почвоведение. Кроме того, сельскохозяйственные науки представлены в 10 раз слабее, чем медицина и науки о здоровье. Выборочная проверка показала, что немногие российские исследователи, защитившие диссертации по рассматриваемому направлению, имеют хотя бы одну публикацию в WoS и Scopus. Следовательно, для построения научных ландшафтов в области сельскохозяйственных наук России эти наукометрические базы неприменимы.

Альтернатива международным наукометрическим базам — Российский индекс научного цитирования (РИНЦ), содержащий наиболее полную информацию о структуре отечественной науки [*Зибарева и др.,* 2015]. Однако РИНЦ в его текущем состоянии по ряду причин невозможно использовать как единственный источник данных для построения научного ландшафта [*Еременко,* 2014; *Фрадков,* 2015]. Серьезную проблему представляют неточность связывания публикаций и авторов, сложность исправления этих привязок, включение в РИНЦ всех научных изданий безотносительно их качества (что потенциально позволяет «накручивать» показатели). Например, не так давно из РИНЦ были исключены свыше 300 низкокачественных изданий, при этом качество оставшихся не оценивалось [*Экономов,* 2017]. Для борьбы с искусственным завышением наукометрических показателей в РИНЦ введены более устойчивые версии популярных наукометрических показателей: h-индекс без самоцитирований, h-индекс по ядру РИНЦ, индекс Херфиндаля для журналов. Однако качество исходных данных (в том числе соответствия авторов и публикаций) по-прежнему оставляет желать лучшего. В ходе нашего исследования выявлен ряд других проблем, которые будут описаны далее. Упомянутые пробелы могут объясняться слабой публикационной активностью, неразвитой культурой цитирования в сельскохозяйственных науках и недостатками самого РИНЦ. К аналогичным выводам приходят и другие исследователи [*Сидорова,* 2016].

Для того чтобы преодолеть хотя бы часть названных проблем, предлагается использовать дополнительную информацию, в частности данные Высшей аттестационной комиссии (ВАК) о защитах диссертаций. В соответствии с российским законодательством квалификация научных работников оценивается государственной системой аттестации, предусматривающей присуждение ученых степеней кандидата и доктора наук согласно номенклатуре научных специальностей. Можно утверждать, что исследователь, защитивший диссертацию в соответствующей области науки, обладает определенным уровнем компетенции в рамках тематики его диссертационного исследования, что, как ожидается, должно повысить качество научного ландшафта. Наряду с этим в базе данных ВАК имеется информация о возрасте, что позволяет учитывать его, например, при оценке деятельности молодых ученых.

Другая распространенная проблема картирования науки — наличие различных классификаторов, которые призваны содействовать аналитической работе, структурируя предметные области, но по факту ее затрудняют. Для анализа сельскохозяйственных наук можно было бы использовать российский Государственный рубрикатор научно-технической информации (ГРНТИ), универсальную десятичную классификацию (УДК) и номенклатуру научных специальностей ВАК. Однако эти классификаторы не связаны между собой и сильно отличаются друг от друга. Кроме того, их изменение со временем приводит к упразднению одних кодов и появлению новых. К тому же анализируемая предметная область далеко не всегда вписывается в таксономию, предлагаемую авторами классификаторов. В результате привязка конкретных исследований к кодам бывает весьма спорна, а зачастую и некорректна. Решить перечисленные проблемы можно автоматическим анализом текстов, в частности, за счет кластеризации и тематического моделирования [*Shvets et al.,* 2015]. Эти методы позволяют структурировать предметные области без опоры на какие-либо заранее составленные классификаторы. В результате применения таких инструментов исходное множество делится на группы объектов, объединенных по некоторому набору признаков.

В настоящей работе на основе анализа полных текстов мы группируем авторефераты кандидатских и докторских диссертаций российских исследователей. Получившиеся в результате группы трактуются как ***направление исследований***.

Построение ландшафта состоит из трех основных этапов:

- формирование исходного списка ученых и показателей их деятельности;
- группирование ученых по направлениям исследований;
- сравнение групп ученых и анализ динамики показателей.

[1] Приказ Роспатента № 8 от 03.01.2017 г. Режим доступа: http://www.consultant.ru/document/cons_doc_LAW_212062/, дата обращения 15.12.2017.

Предложенная методика, апробируемая на примере молодых ученых и сельскохозяйственных наук, особенно актуальна для областей знаний, слабо представленных в Scopus и WoS. Лежащие в ее основе средства автоматического анализа текстов на естественном языке дают возможность агрегировать данные о публикационной и диссертационной активности ученых из различных источников независимо от степени их структурированности.

Исходные данные

Информационной базой исследования служили авторефераты кандидатских и докторских диссертаций, информация Минобрнауки России об исследователях, получивших ученую степень по сельскохозяйственным наукам, в том числе фамилия, имя и отчество (ФИО), организация и год рождения соискателя (данные ВАК), а также наукометрические показатели из РИНЦ (http://elibrary.ru).

Исходная выборка охватила ученых в возрасте до 40 лет (на момент проведения исследования — 2016 г.), защитивших диссертации по сельскохозяйственным наукам в период с 2008 по 2015 г. включительно. Ограничение по периоду защиты обусловлено отсутствием в ВАК необходимой информации за более ранние годы. При анализе использовалась информация о 2572 молодых ученых.

Всего в представленных ВАК данных присутствуют четыре кода научных специальностей верхнего уровня: 06 (сельское хозяйство), 05 (технические науки), 03 (биологические науки) и 25 (науки о Земле); 10 кодов второго уровня и 32 — третьего (полные идентификаторы специальностей). Примечательно, что часть кодов относятся к старым версиям классификатора ВАК, а часть — к новой редакции, поэтому его применение для построения научного ландшафта без дополнительной обработки сильно затруднено.

Для выявления структуры предметной области использовалась коллекция полных текстов авторефератов кандидатских и докторских диссертаций, собранных на сайте Российской государственной библиотеки. Привязка автореферата к данным РИНЦ и ВАК осуществлялась по ФИО и наименованию организации, где выполнено диссертационное исследование.

Наукометрические показатели ученых собирались в автоматизированном режиме. Выполнялся поиск в авторском указателе eLIBRARY по ФИО исследователя, затем осуществлялся отбор по месту работы (сравнивались соответствующие поля страницы автора в РИНЦ и в выгрузке из базы данных ВАК). Дополнительно учитывались год рождения и дата первой публикации для исключения профилей с некорректно привязанными публикациями. Если описанные условия выполнялись, запускался программный модуль, копировавший необходимые наукометрические показатели в базу данных с исходной информацией для построения ландшафта. Работа с информационной системой РИНЦ выявила ряд ее недостатков:

- низкий процент заполненных профилей (только 56% авторов имеют заполненные профили);
- отсутствие информации о возрасте исследователей (что, например, не позволяет построить отдельную аналитику по молодым ученым);
- нехватка достаточно мощных аналитических инструментов или средств выгрузки данных для анализа с помощью других инструментов;
- малая информативность h-индекса по ядру РИНЦ (для 75% рассмотренных ученых его значение не превышает 1);
- наличие только одного тематического рубрикатора, что не позволяет провести сегментацию и анализ направлений исследований с нескольких точек зрения. Для сравнения: в той же WoS имеется возможность использовать различные классификаторы для анализа структуры заданной области.

Тест хи-квадрат выявил отсутствие статистически значимых различий в распределении автоматически выделенных направлений между общей совокупностью из 2572 исследователей и выборкой из 56% авторов с успешно сопоставленными профилями в РИНЦ. Это дает основание считать, что полученная выборка профилей является репрезентативной для исходного списка ученых, полученного из базы ВАК. Все дальнейшие операции с наукометрическими показателями выполнялись только в отношении 56% исследователей, для которых удалось успешно сопоставить профиль в РИНЦ.

Собранный массив исходных данных оформлялся в виде таблицы со следующими столбцами:

- ФИО;
- год рождения;
- ученая степень;
- год защиты;
- код специальности;
- организация;
- идентификатор автореферата;
- число публикаций;
- h-индекс;
- h-индекс без учета самоцитирований;
- h-индекс по ядру РИНЦ;
- суммарное число цитирований;
- средневзвешенный импакт-фактор журналов, опубликовавших статьи;
- средневзвешенный импакт-фактор журналов, процитировавших статьи;
- научные направления.

Каждая строка в этой таблице соответствует одному исследователю. Поле «научные направления» заполнялось по результатам определения структуры предметной области с помощью тематического моделирования коллекции полных текстов авторефератов.

Определение тематической структуры предметной области

Для выявления направлений исследований внутри сельскохозяйственных наук на основе коллекции полных текстов авторефератов молодых ученых по дан-

ным ВАК строилась вероятностная тематическая модель латентного размещения Дирихле (*Latent Dirichlet Allocation,* LDA) [*Blei et al.,* 2003]. Используемый нами метод неоднократно обеспечивал достоверную интерпретацию результатов группировки текстов, близких по тематике, что отражено в ряде отечественных и зарубежных работ [*Garousi, Mäntylä,* 2016; и др.]. Он позволяет получать обобщенное описание тематической структуры анализируемой коллекции текстов. В ходе эксперимента использовалась программная реализация LDA из библиотеки Gensim [*Rehurek, Sojka,* 2010].

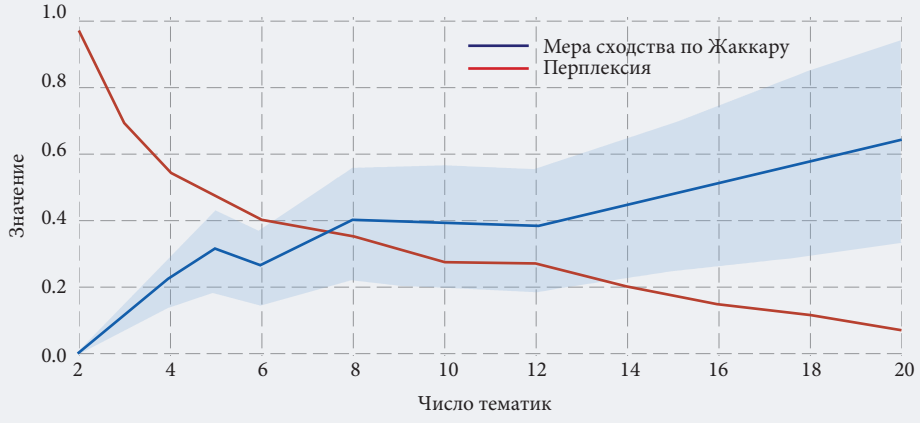
Набор признаков формировался на основе подхода, называемого «мешком слов» (*Bag-of-Words*). Ввиду отсутствия в авторефератах каких-либо авторских ключевых слов или словосочетаний выявить лексику анализируемых работ возможно только путем анализа их полных текстов. Для построения слов и словосочетаний по полным текстам авторефератов применялся лингвистический анализатор, способный выполнять морфологический, синтаксический и семантический анализ, в том числе разметку семантических ролей (*semantic role labeling*) текстов на русском и английском языках. Морфология русскоязычных текстов обрабатывалась методами, реализованными в библиотеке АОТ [*Sokirko,* 2001], синтаксис — посредством программы MaltParser [*Nivre et al.,* 2007], предварительно обученной на размеченном корпусе SynTagRus [*Nivre et al.,* 2008]. Аналогичные процедуры для англоязычных публикаций выполнялись с помощью инструментария библиотеки Freeling [*Padró, Stanilovsky,* 2012]. Методика реляционно-ситуационного анализа, созданная в Институте системного анализа ФИЦ ИУ РАН [*Osipov et al.,* 2013], стала основой для обработки семантики. Оценки качества анализа на размеченном корпусе представлены в работе [*Shelmanov, Smirnov,* 2014]. Эффективность используемого анализатора косвенно подтверждается опытом его применения в прикладных целях, включая выявление текстовых заимствований (PAN CLEF-2014) [*Zubarev, Sochenkov,* 2014] и полнотекстовый поиск (РОМИП-2008) [*Смирнов и др.,* 2008]. С помощью упомянутого инструмента выделяются отдельные слова и именные группы, объединяющие синтаксически связные слова, в которых главная роль принадлежит существительному [*Suvorov, Sochenkov,* 2015]. Нормализация полученных словосочетаний осуществляется путем приведения главного слова к нормальной форме, после чего из всех возможных вариантов выбирается форма зависимого слова, согласованная по морфологическим признакам с главным.

Основным критерием для оценки качества тематической модели выбрана перплексия [*Hofmann,* 1999]. Она характеризует способность модели восстанавливать исходное вероятностное распределение лексики (слова и словосочетания) по документам в анализируемой коллекции: чем ниже значение этого показателя, тем лучше модель описывает данные. Самым существенным параметром при построении модели является число тематик (или, другими словами, классов). С ростом этого показателя гибкость модели увеличивается, поэтому она начинает все лучше описывать исходные данные (следовательно, величина перплексии падает). Однако если задать чрезмерно большое значение, то тематики по содержанию начнут сильно пересекаться, и подобное разбиение трудно интерпретировать. Таким образом, встает вопрос выбора количества тематик. В настоящей работе для оценки близости тематик использовалась мера Жаккара [*Manning et al.,* 2008] между списками из 30 наиболее характерных для каждой из них слов и словосочетаний. Оптимально число тематик, при котором перплексия достаточно низкая, а степень их близости не слишком высокая.

Для обозначения результатов тематического моделирования, выполненного по вышеописанной схеме, воспользуемся термином *направления исследований* (или просто *направления*). По итогам данной процедуры некоторые авторефераты могут быть отнесены сразу к нескольким направлениям, в итоге появляется возможность выявлять и анализировать группы междисциплинарных работ.

На рис. 1 приведены графики нормированного значения перплексии тематической модели и нормированного среднего значения меры сходства по Жаккару между

Рис. 1. **Перплексия и мера сходства по Жаккару**



Источник: составлено авторами.

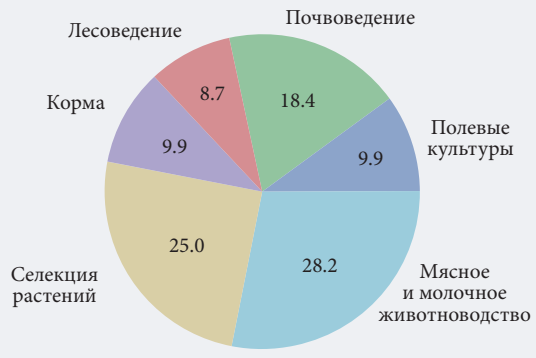
Табл. 1. **Направления исследований, автоматически выделенные на основе полных текстов авторефератов за 2008–2015 гг.**

Направление исследований	Ключевые слова и словосочетания (выделены автоматически)
Мясное и молочное животноводство	Корова, животное, молоко, скот, живая масса, лактация, бычок, удой, генотип, молочная продуктивность, мес, индекс, живой, группа коров, комлацкий, матка, помесь, молочный, линия, гол, сверстница, тушь, месячный возраст, кобыла, степной, черно-пестрая порода, кровь, стадо, баранчик, телка, первотелка, рождение, молодняк, группа животных, мясная продуктивность, животный, голова, сверстник, мясо
Селекция растений	Гибрид, линия, балл, плод, стандарт, сортообразец, популяция, болезнь, подвой, приморский, ранний, комбинация, побег, масса зерна, новый сорт, колос, генотип, декада, плодовый, селекционный, изменчивость, хозяйственный, цветение, яблоня, масса семян, посадка, картофель, скрещивание, черенок, создать, коллекция, исходный материал, поражение, диаметр, грунт, созревание, питомник, полегание, сад
Корма	Рацион, опытная группа, контрольная группа, комбикорм, живая масса, добавка, кормление, животное, бройлер, протеин, переваримость, живой, организм, препарат, цыпленок-бройлер, птица, кровь, кормовой кальций, цыпленок, поросенок, сохранность, яйцо, протеин сырой, скармливание, голова, аналог, энергия обмена, научно-хозяйственный опыт, свинья, молодняк, гол включение, свинка, кура, клетчатка, мясо, период опыта, вещество переваримость, метионин
Лесоведение	Насаждение, дерево, лесной, сосна, лес, земля, рубка, дуб, диаметр, древостой, запас, склон, класс, полоса, сосняк, ярус, уход, ельник, древесина, покров, агроландшафт, житняк, распределение, ель, подрост, крона, лесная полоса, тип леса, шишка, угодье, горизонт, модель, овсяница, травостой, северный, пашня, сосна обыкновенная
Почвоведение	Орошение, численность, полив, край, препарат, раствор, металл тяжелый, потеря, фракция, июль, рис, режим орошения, корневая гниль, вредитель, тля, кислота, стимул, кубань, амарант, мина, экологически, личинка, скорость, сентябрь, биологически, водный режим, горизонт
Полевые культуры	Озимая пшеница, минеральное удобрение, ячмень, яровая пшеница, кг/га, севооборот, ц/га, норма высева, посев срок, зеленая масса, трава, рожь, гербицид, пар, предшественник, внесение удобрения, кукуруза, смесь, озимая рожь, обработка семян, горох, навоз, овес, обработка почва, люцерна, препарат, многолетний трава, азотное удобрение, чернозем, вспашка, почва слой, соя, доза удобрения, качество зерна, клевер, всхожесть, л/га, урожайность зерна, подкормка, способ посева

Источник: составлено авторами.

направлениями, а также дисперсии этих величин. При увеличении количества направлений с шести до восьми дисперсия меры сходства по Жаккару значительно возрастает — более чем в полтора раза. Появляются отдельные пары направлений, имеющие много общей лексики, а перплексия снижается незначительно. Исходя из этого, а также из интерпретации тематического разбиения, становится очевидно, что следует ограничиться числом направлений, равным шести. Именно при таком значении обеспечиваются интерпретируемость тематического группирования, приемлемый баланс перплексии и меры сходства по Жаккару.

Рис. 2. **Распределение числа защит диссертаций по направлениям исследований** (%)



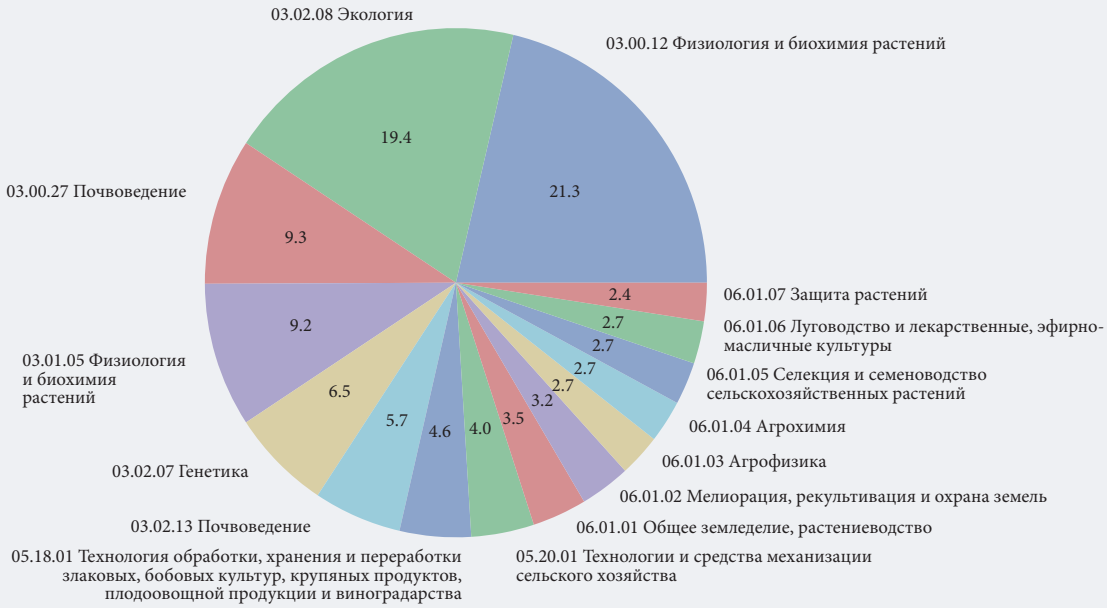
Источник: составлено авторами.

Наименования направлений и соответствующие им характерные ключевые слова и словосочетания приведены в табл. 1. Они сформулированы авторами статьи на основе списков ключевых слов и словосочетаний, созданных автоматически в процессе обучения тематической модели LDA.

На рис. 2 изображено соотношение числа защит диссертаций по каждому из автоматически выделенных исследовательских направлений. Для сравнения на рис. 3 приведено распределение данного показателя по 15 кодам номенклатуры научных специальностей в области сельскохозяйственных наук, отличающихся наибольшей активностью.

Распределение, отраженное на рис. 3, характеризуется скошенностью (малое число популярных специальностей и длинный «хвост» специальностей с незначительным количеством защит). Сравнительный анализ направлений исследовательской активности при таком распределении затруднен по нескольким причинам. Ввиду того что список научных специальностей содержит как старые, так и новые коды номенклатуры ВАК, их сопоставление — сложная задача, не имеющая однозначного решения. К тому же ввиду значительно различающихся объемов выборок по каждой специальности оценки параметров распределений наукометрических показателей характеризуются различными доверительными интервалами (причем для менее популярных специальностей они становятся все шире). Для наиболее редких направлений оценивание параметров распределений и вовсе не имеет смысла, что делает невозможным их сравнение с более представленными специальностями.

Рис. 3. **Распределение числа защит по топ-15 кодам номенклатуры научных специальностей** (%)



Источник: составлено авторами.

Из сравнения рис. 2 и 3 можно заключить, что распределение числа защит диссертаций по автоматически выделенным направлениям получилось менее скошенным по сравнению с распределением по кодам научных специальностей. Таким образом, дальнейшая процедура тематического анализа упрощается, подтверждается его полезность для выявления структуры предметной области и совместной оценки показателей публикационной активности. Существенным отличием направлений исследований от кодов научных специальностей является их пересечение: любой ученый может относиться одновременно к нескольким областям. Тем самым в исследовании учитывается междисциплинарность — важная характеристика развития современной науки.

На рис. 4 приведено сравнение количества защит по сельскохозяйственным наукам за каждый год с 2008 по 2015 г. включительно. Все столбцы раскрашены согласно соотношению числа защит в соответствующий год по каждому направлению. Из рисунка следует вывод, что акцент смещается с полевых культур и почвоведения в сторону селекции растений.

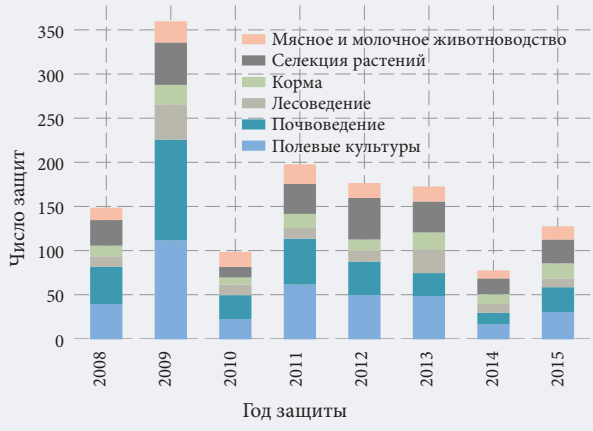
Анализ наукометрических показателей молодых ученых

Наукометрические показатели анализировались с применением простых гистограмм, расчетов плотности распределения с гауссовским сглаживанием [*Scott,* 1992]; графики сравнения представлены в виде диаграмм-скрипок (*violin plots*) [*Hintze, Nelson,* 1998]. Для обработки данных использовались язык программирования Python 2.7 и интерактивная среда разработки Jupyter, графики построены с помощью библиотек Matplotlib [*Hunter,* 2007] и Seaborn (https://seaborn.pydata.org/).

Как упоминалось, менее 56% российских молодых ученых в области сельскохозяйственных наук имеют заполненные профили в РИНЦ. Это означает, что они уделяют крайне мало внимания отслеживанию своей публикационной активности. После связывания профилей в eLIBRARY и данных ВАК записи со всеми известными параметрами, пригодные для дальнейшего анализа, остались лишь по 1419 ученым. Выяснилось, что индекс Хирша, равный или превосходящий 5, имеют всего 107 человек (из более чем 2500 ученых в первоначальном списке). Если взять индекс Хирша без учета самоцитирований, то молодых ученых с h-индексом 5 и выше выявлено всего 78 человек.

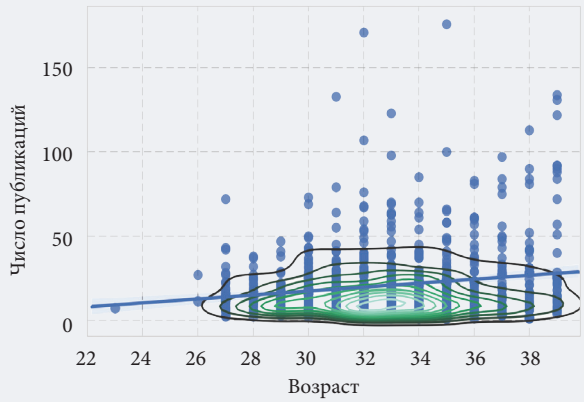
На рис. 5 изображена зависимость между числом публикаций и возрастом ученого. Каждая точка соот-

Рис. 4. **Динамика диссертационной активности по направлениям исследований**



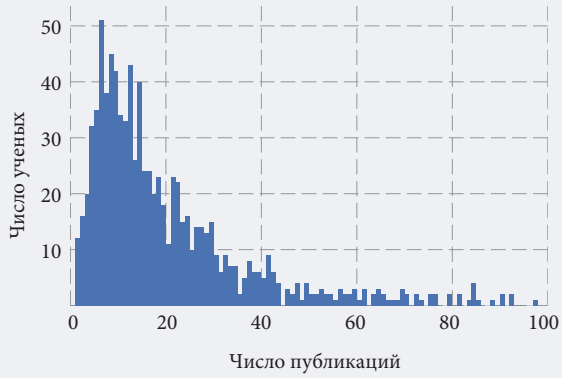
Источник: составлено авторами.

Рис. 5. **Линейная корреляция и оценка плотности совместного распределения числа публикаций в РИНЦ и возраста ученого**



Источник: составлено авторами.

Рис. 6. **Гистограмма распределения числа публикаций по ученым**



Источник: составлено авторами.

ветствует конкретному ученому (координата по горизонтали — его возраст, а по вертикали — число публикаций на момент проведения исследования). Прямой линией изображен линейный тренд (корреляция). Овальные линии уровня описывают плотность совместного распределения возраста и числа публикаций.

С демографической точки зрения из приведенного на рис. 5 графика видно, что наиболее продуктивный возраст ученого в области сельскохозяйственных наук наступает в 32–34 года. Как правило, это молодой кандидат наук, активно работающий и публикующийся. Всплеск также наблюдается к 39–40 годам, которые предположительно являются наиболее продуктивным возрастом молодых докторов либо кандидатов наук, приближающихся к защите докторской. Возрастной отметке 30 лет предшествует плавный рост публикационной активности, который в последующие пять лет стабилизируется.

На рис. 6 приведена гистограмма распределения числа публикаций (с ограничением до 100). Основная часть ученых имеют от восьми до 25 работ, наибольшее значение — 176.

На рис. 7 приведена гистограмма распределения индекса Хирша без учета самоцитирований, из которой следует, что подавляющее количество ученых имеют h-индекс от 1 до 3. Наибольший индекс Хирша в выборке равен 22.

На рис. 8–11 приведены диаграммы-скрипки, иллюстрирующие различия эмпирических распределений ряда наукометрических показателей между автоматически выделенными направлениями исследований. Каждая фигура представляет собой график ядерной оценки плотности вероятности (симметричный относительно вертикальной оси) с гауссовским сглаживанием. Чем шире фигура, тем выше доля ученых с соответствующим значением показателя. Пунктирные линии внутри фигур обозначают 25, 50 и 75-й перцентили. Для оценки статистической значимости различий в распределении показателя между направлениями использовался U-тест Манна–Уитни [*Mann, Whitney,* 1947]. Цель проверки — убедиться в том, что в одной области показатели в среднем выше, чем в другой, поэтому применение именно данного критерия представляется оправданным. Значение критерия и соответствующая p-величина вычислялись для всех возможных пар направлений исследований (всего таких пар 15). Статистическая значимость (критическое значение p) установлена на уровне 0.05.

На рис. 8 приведено сравнение распределений индекса Хирша без учета самоцитирований между автоматически выделенными научными направлениями. Ширина фигур пропорциональна численности ученых с соответствующим значением h-индекса. Исследователи с индексом Хирша менее 5 на графике не отображены. Это отсечение сделано, чтобы показать различия в цитируемости ведущих ученых по каждому направлению. График иллюстрирует заметные различия между направлениями по представленности учеными мирового уровня (с h-индексом от 10 и более). Применение кри-

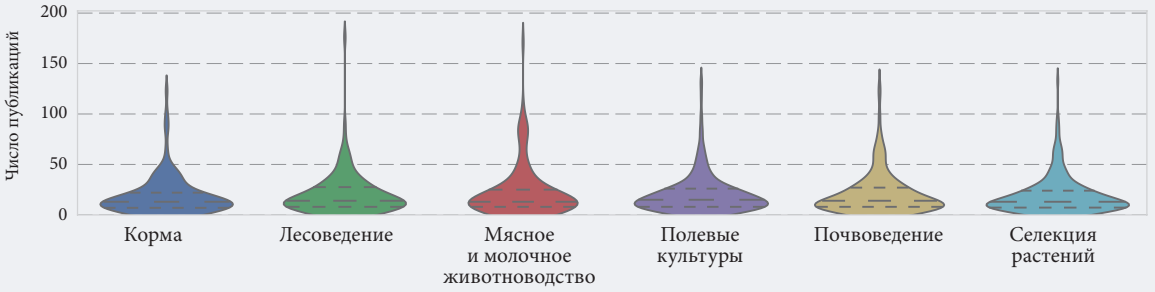
Рис. 7. **Гистограмма распределения индекса Хирша без учета самоцитирований**

Источник: составлено авторами.

Рис. 8. Сравнение распределений h-индекса без учета самоцитирований

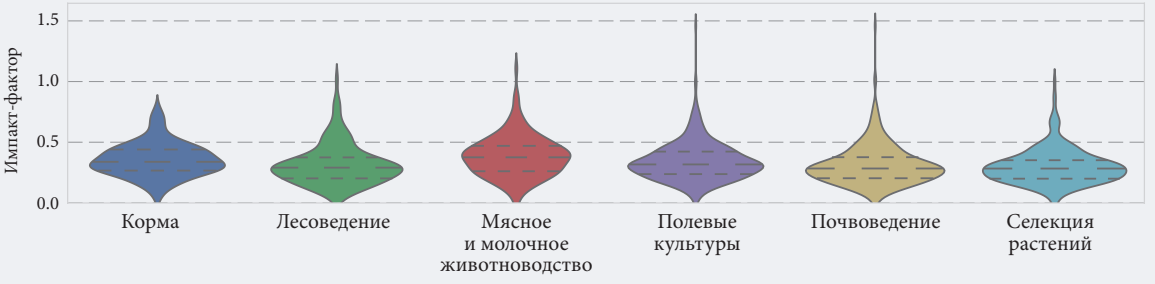
Источник: составлено авторами.

Рис. 9. Сравнение распределений количества публикаций



Источник: составлено авторами.

Рис. 10. Сравнение распределений средневзвешенных импакт-факторов журналов, в которых были опубликованы статьи ученых



Источник: составлено авторами.

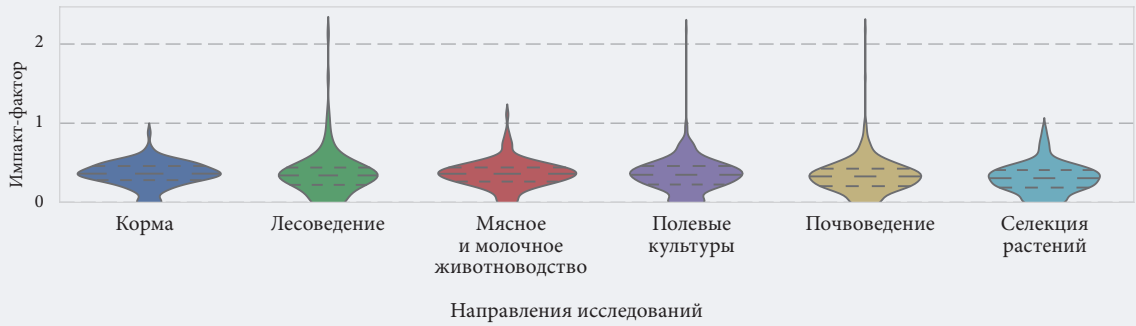
терия Манна–Уитни[2] позволило установить, что при уровне значимости 0.05 в 40% случаев (шести парах) по одному направлению из пары h-индекс в среднем оказывался выше, чем у другого. Наиболее существенные расхождения наблюдаются в парах «лесоведение — мясо-молочное животноводство» (p=0.007), «мясо-молочное животноводство — селекция растений» (p=0.0003), «полевые культуры — селекция растений» (p=0.0006). Наименьшими различиями характеризуется связка «корма — мясо-молочное животноводство» (p=0.44).

На рис. 9 сравниваются распределения количества публикаций ученых по каждому из направлений. Как видим, публикационная активность всех исследователей находится примерно на одном уровне. Статистически значимых различий среднего числа публикаций между направлениями не выявлено (среднее p=0.28).

На рис. 10 и 11 приведено сравнение по импакт-факторам научных журналов, в которых авторы публикуют свои работы, с изданиями, цитирующими эти статьи. Графики свидетельствуют, что эти распределения значительно различаются между направлениями. Статистически значимые разбросы величин импакт-фактора между журналами, в которых статьи были опубликованы, и процитировавшими их изданиями выявлены между десятью и семью парами направлений соответственно (67 и 47% пар). Наиболее

[2] Критерий Манна–Уитни вычислялся на основе всех значений h-индекса (в том числе меньше 5).

Рис. 11. **Сравнение распределений средневзвешенных импакт-факторов журналов, из которых были процитированы статьи ученых**



Источник: составлено авторами.

заметны расхождения средних значений показателя в парах «мясо-молочное животноводство — селекция растений» (p=0.0005), «корма — селекция растений» (p=0.0004), «полевые культуры — селекция растений» (0.0008). Наименьшие различия присущи комбинациям «корма — мясо-молочные животноводство» (p=0.37) и «лесоведение — полевые культуры» (p=0.32).

Представляют интерес различия по импакт-факторам журналов, опубликовавших статьи, и процитировавших их изданий в пределах одного направления. Из графиков следует, что, например, в отличие от иных областей, у журналов, цитирующих работы по лесоведению и почвоведению, рейтинг существенно выше, чем у тех, на чьих страницах они были опубликованы. Это подтверждается U-тестом Манна–Уитни с p=0.03 и p=0.02 соответственно.

Из сказанного с высокой долей уверенности вытекает вывод о существовании определенных различий как в наукометрических показателях, так и в культуре цитирования в разных направлениях сельскохозяйственных наук. В его пользу говорят результаты статистического U-теста Манна-Уитни. Тем самым подтверждается необходимость дифференцированного подхода к анализу результативности ученых по различным направлениям исследований в рамках одной отрасли наук.

Заключение

В статье предложена и апробирована новая методика построения научных ландшафтов. Новизна заключается в объединении полнотекстовой аналитики и традиционной статистической обработки наукометрических данных для улучшения достоверности, устойчивости и интерпретируемости научного ландшафта. Представленный подход восполняет недостатки вручную составленных таксономий, которые ведут к излишней детализации и затрудняют сравнение направлений. Он особенно актуален для областей науки, слабо представленных в международных наукометрических базах данных Scopus и WoS. Отметим, что рассматриваемый способ объединения различных источников данных имеет альтернативы, но именно он, по мнению авторов, решает ряд обозначенных в статье проблем.

Предложенный инструментарий не заменяет традиционные наукометрические инструменты, а дополняет их, формируя более целостную и интерпретируемую картину для отдельных задач анализа и оценки развития науки. Среди его преимуществ — возможность выявления перспективных междисциплинарных исследовательских направлений, сформировавшихся на практике (подход «от данных»). Благодаря этому он может применяться, например, в рамках работы с масштабными вызовами, обозначенными в Стратегии научно-технологического развития РФ[3].

В результате экспериментальной апробации методики впервые был получен научный ландшафт в области сельскохозяйственных наук на примере молодых ученых в возрасте до 40 лет. Выделены шесть крупных научных направлений: корма, лесоведение, мясное и молочное животноводство, полевые культуры, почвоведение, селекция растений. Самые высокие значения библиометрических показателей отмечены в области почвоведения. Существенные наукометрические различия между рассматриваемыми направлениями указывают на необходимость дифференцированного подхода к оценке результатов научной деятельности.

Перспективы дальнейших работ по совершенствованию методики авторы связывают с построением научных ландшафтов для других предметных областей, созданием алгоритмов выявления перспективных направлений исследований, составлением тематических рейтингов ученых и организаций. Предстоит изучить возможности применения методов текстового анализа для сопоставления научно-технических документов из разнородных источников (баз научных статей, патентов и др.), которые невозможно автоматически связать друг с другом ввиду отсутствия сквозной классификации. Подобное сопоставление могло бы лечь в основу комплексного анализа отдельных научно-технологических областей и построения более сложных научных ландшафтов.

Работа выполнена при поддержке РФФИ, грант № 14-29-05008.

[3] Утверждена Указом Президента РФ № 642 от 01.12.2016 г. Режим доступа: http://publication.pravo.gov.ru/Document/View/0001201612010007, дата обращения 14.11.2017.

Библиография

Еременко Г. (2014) Во всем виноват РИНЦ? // Троицкий вариант. № 163. С. 7.

Зибарева И.В., Солошенко Н.С. (2015) Тематическая структура российского сегмента научных журналов в глобальных и национальных информационных ресурсах // Материалы Третьей международной конференции НЭИКОН «Электронные научные и образовательные ресурсы: создание, продвижение и использование». М.: НП НЭИКОН. С. 255–259.

Кристофилопулос Э., Манцанакис С. (2016) Китай-2025: научный и инновационный ландшафт // Форсайт. Т. 10. № 3. С. 7–16.

Сидорова В.В. (2016) Использование РИНЦ для оценки научной деятельности гуманитариев // Сибирские исторические исследования. № 1. С. 27–39.

Смирнов И.В., Соченков И.В., Муравьев В.В., Тихомиров И.А. (2008) Результаты и перспективы поискового алгоритма Exactus // Труды российского семинара по оценке методов информационного поиска (РОМИП) 2007–2008. СПб: НУ ЦСИ. С. 66–76.

Фрадков А. (2015) РИНЦ продолжает врать // ТрВ-Наука. Режим доступа: http://trv-science.ru/2015/09/08/risc-prodolzhaet-vrat/, дата обращения 15.05.2017.

Экономов И. (2017) Мусорная наука // ТрВ-Наука. Режим доступа: http://trv-science.ru/2017/04/25/musornaya-nauka/, дата обращения 15.05.2017.

Blei D.M., Ng A.Y., Jordan M.I. (2003) Latent Dirichlet Allocation // Journal of Machine Learning Research. Vol. 3. P. 993–1022.

Garousi V., Mäntylä M.V. (2016) Citations, research topics, and active countries in software engineering: A bibliometrics study // Computer Science Review. Vol. 19. P. 56–77.

Hintze J.L., Nelson R.D. (1998) Violin plots: A box plot-density trace synergism // The American Statistician. Vol. 52. № 2. P. 181–184.

Hofmann T. (1999) Probabilistic latent semantic analysis // Proceedings of the XV Conference on Uncertainty in Artificial Intelligence. Berlington, MA: Morgan Kaufmann Publishers, Inc. P. 289–296.

Hunter J.D. (2007) Matplotlib: A 2D graphics environment // Computing in Science and Engineering. Vol. 9. № 3. P. 90–95.

Mann H.B., Whitney D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other // The Annals of Mathematical Statistics. Vol. 18. № 1. P. 50–60.

Manning C.D., Prabhakar R., Schütze H. (2008) Introduction to information retrieval. Cambridge: Cambridge University Press.

Nivre J., Boguslavsky I.M., Iomdin L.L. (2008) Parsing the SynTagRus treebank of Russian // Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008). Stroudsburg, PA: Association for Computational Linguistics (ACL). P. 641–648.

Nivre J.H., Hall J., Nilsson J., Chanev A., Eryiğit G., Kübler S., Marinov S., Marsi E. (2007) MaltParser: A language-independent system for data-driven dependency parsing // Natural Language Engineering. Vol. 13. № 2. P. 95–135.

Oldham P., Hall S., Burton G. (2012) Synthetic biology: Mapping the scientific landscape // PLoS One. Vol. 7. № 4. P. e34368. Режим доступа: https://www.ncbi.nlm.nih.gov/pubmed/22539946, дата обращения 24.05.2017.

Osipov G., Smirnov I., Tikhomirov I., Shelmanov A. (2013) Relational-situational method for intelligent search and analysis of scientific publications // Proceedings of the Integrating IR Technologies for Professional Search Workshop, Moscow, Russian Federation, March 24, 2013 / Eds. M. Lupu, M. Salampasis, N. Fuhr, A. Hanbury, B. Larsen, H. Strindberg. P. 57–64. Режим доступа: http://ceur-ws.org/Vol-968/irps_10.pdf, дата обращения 24.05.2017.

Padró L., Stanilovsky E. (2012) Freeling 3.0: Towards wider multilinguality // Proceedings of the International Conference on Language Resources and Evaluation "LREC2012", Istanbul, 2012. P. 2473–2479. Режим доступа: http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf, дата обращения 24.05.2017.

Rehurek R., Sojka P. (2010) Software framework for topic modelling with large corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, May 22. P. 51–56. Режим доступа: http://www.lrec-conf.org/proceedings/lrec2010/workshops/W10.pdf, дата обращения 24.05.2017.

Scott D.W. (1992) Multivariate Density Estimation: Theory, Practice, and Visualization. New York; Chicester: John Wiley & Sons.

Shelmanov A.O., Smirnov I.V. (2014) Methods for semantic role labeling of Russian texts // Computational Linguistics and Intellectual Technologies. Proceedings of International Conference Dialog. Vol. 13. № 20. P. 607–620.

Shvets A., Devyatkin D., Sochenkov I., Tikhomirov I., Popov K., Yarygin K. (2015) Detection of Current Research Directions Based on Full-Text Clustering // Proceedings of 2015 Science and Information Conference (SAI 2015), July 28-30, 2015, London, United Kingdom. Piscataway, NJ: Institute of Electrical and Electronics Engineers (IEEE). P. 483–488.

Sokirko A. (2001) A short description of Dialing Project. Режим доступа: http://www.aot.ru/docs/sokirko/sokirko-candid-eng.html, дата обращения 15.05.2017.

Suvorov R.E., Sochenkov I.V. (2015) Establishing the similarity of scientific and technical documents based on thematic significance // Scientific and Technical Information Processing. Vol. 42. P. 321–327.

Zubarev D., Sochenkov I. (2014) Using sentence similarity measure for plagiarism source retrieval // Working Notes for CLEF 2014 Conference. P. 1027–1034. Режим доступа: https://pdfs.semanticscholar.org/4556/08d685695c1a7f05ffd8257fae79e1f64593.pdf, дата обращения 15.05.2017.