Опыт получения и использования наукометрической информации в системах управления научной деятельностью*

С. П. Белов, Е. Л. Плискин, А. В. Усков

Аннотация. Управление наукой опирается на объективные количественные показатели о научных исследованиях. Необходимо собирать данные о людях, которые поддерживаются грантами, о том, какие исследования проводятся, вместе с кем и где, и какие результаты проистекают от исследований. Продолжать возлагать всю отчетность на исследователей непродуктивно. Научно-технические агентства в различных странах разрабатывают информационные платформы для выявления и характеристики научных результатов. Особенный интерес представляет бразильский многолетний опыт национальной веб-платформы LATTES, собирающей в обязательном порядке биографические очерки всех получающих государственную поддержку исследователей любого возраста и положения, от студентов младших курсов до академиков. Административные отчеты описывают, *кто* ведет исследования и *вместе с кем*. В статье описаны некоторые методы и алгоритмы теории решений, которые целесообразно использовать для управления наукой, а также опыт применения многокритериального анализа для оценки результативности проектов целевых фундаментальных исследований.

Ключевые слова: *наукометрические показатели, теория принятия решений, многокритериальный анализ, экспертная оценка, биографические очерки, тематическое моделирование*.

Введение

Управление наукой предполагает умение отвечать на следующие вопросы: Сколько средств целесообразно тратить на науку в стране? На какую именно науку? В каком соотношении должны использоваться государственные и частные источники финансирования? Указывают ли запросы со стороны потенциальных исполнителей научных программ на недостаточный объем финансирования науки или на избыток количества самих исследователей?

Эти вопросы выявляют необходимость в таких объективных количественных показателях, на которые, с одной стороны, смогут опираться органы, принимающие решения, а с другой стороны, которыми смогут оперировать эксперты в поддержку тех или иных направлений.

Во многих областях государственного управления решения принимаются на основе хорошо обоснованных индикаторов. Например, политика в области трудовых ресурсов во многом зависит от такого показателя, как уровень безработицы. Экономическая политика опирается на множество показателей, включая рост ВВП. Политика в области образования использует результаты экзаменов. Хотя эти показатели могут быть несовершенны, но широко применяются, и их теоретическое обоснование хорошо известно лицам, принимающим решения. Иначе в научной политике: здесь ощущается недостаток инфраструктуры.

Несмотря на широкое распространение таких научных метрик, как h-индекс (индекс Хирша) и импакт-фактор, им часто недостает иного обоснования, чем само наличие данных, на основе которых рассчитываются эти метрики. Смысл этих показателей часто непонятен лицам, принимающим решения, для которых эти метрики генерируются. Когда эти метрики принимаются за основу научной политики, то возникают нежелательные искажения [8]:

- Эти метрики отдают предпочтение пожилым исследователям перед молодыми.
- Они не учитывают такие процессы передачи знаний, как преподавание.
- Они способствуют «игре» с показателями ради выигрыша.

* Работа выполнена при частичной поддержке РФФИ (гранты № 14–29–05047, № 14–29–05048).

Некоторые методы сбора объективной информации о научно-технической деятельности описаны в п. 2. Особенный интерес представляет бразильский многолетний опыт национальной веб-платформы LATTES, собирающей в обязательном порядке биографические очерки всех получающих государственную поддержку исследователей любого возраста и положения, от студентов младших курсов до академиков.

В тех случаях, когда эксперты дают множественные оценки предлагаемым проектам по различным критериям, а решение о поддержке того или иного проекта в конечном счете принимается лицом или агентством, находящимся вне экспертной инфраструктуры, для выбора используются методы многокритериального анализа, который развился в теории принятия решений [1]. В п. 1 описаны некоторые методы и алгоритмы теории решений, которые целесообразно использовать для управления наукой.

В п. 3 описано применение многокритериального анализа для оценки результативности проектов целевых фундаментальных исследований, поддержанных Российским фондом фундаментальных исследований (РФФИ) и выполняемых в интересах федеральных агентств и ведомств.

1. Методы АРАМИС и ПАКС

Метод АРАМИС (Агрегирование и Ранжирование Альтернатив около Многопризнаковых Идеальных Ситуаций) для коллективного упорядочения многокритериальных вариантов разработан А. Б. Петровским [1]. Метод основан на оценке близости вариантов к наилучшему и к наихудшему из возможных вариантов. Каждый вариант решения оценивается каждым экспертом по всем критериям. Каждый критерий может иметь собственную шкалу оценок.

Метод АРАМИС включает следующие шаги. В пространстве вариантов оценок задаются опорные точки A+ и A-, которые соответствуют наилучшему и наихудшему из возможных вариантов оценок. Для каждого варианта A в пространстве мультимножеств оценок по основной метрике вычисляется расстояние P1 до наилучшего опорного варианта A+ и расстояние P2 до наихудшего опорного варианта A-. Метрика может учитывать важность различных критериев для ЛПР. Заметим, что выбор метрики может повлиять на итоговое упорядочивание вариантов. Для каждого варианта A вычисляется показатель относительной близости к наилучшему варианту A+ по формуле: (P1 / (P1 + P2)). Показатель изменяется от 0 до 1. Варианты упорядочиваются по возрастанию данного показателя. Лучший вариант A* определяется минимальным показателем близости к лучшему варианту A+.

Сравнение, упорядочение или классификация многопризнаковых объектов может быть весьма трудоемкой процедурой и требовать специальных методов опроса ЛПР. Когда количество сравниваемых объектов мало (3–5), а количество признаков велико (десятки, сотни), возникает необходимость в специальных подходах для сокращения размерности признакового пространства и использовании психологически корректных операций получения информации от ЛПР и экспертов. Многоэтапная интерактивная технология ПАКС (Последовательное Агрегирование Классифицируемых Ситуаций) [10] предназначена для сравнения, упорядочения и классификации многопризнаковых объектов по их свойствам и включает три этапа. На первом этапе, исходя из предпочтений ЛПР, проводится снижение размерности признакового пространства путем построения иерархической системы составных критериев. На втором этапе, используя различные методы вербального анализа решений, последовательно формируются шкалы всех составных критериев. На третьем этапе выполняется окончательное решение рассматриваемой задачи выбора в полученном пространстве составных критериев меньшей размерности с помощью того или иного метода принятия решений.

2. Методы сбора объективной информации о научно-технической деятельности

2.1. Эмпирическая основа

Успех научных измерений зависит от сбора данных. Необходимо собирать данные о людях, которые поддерживаются грантами, о том, какие исследования проводятся, вместе с кем и где, и какие результаты проистекают от исследований. Административные отчеты, такие как в американской программе STAR METRICS [9], описывают, КТО ведет исследования и ВМЕСТЕ С КЕМ. Без информации обо всех получателях научного финансирования результаты инвестиций в науку остались бы освещенными не полностью. Критически важны сведения о студентах, поскольку многие результаты инвестиций проистекают из студенческих достижений, в том числе на этапе внедрения научных знаний в деловую практику. Равно важны также и знания о проектных командах, поскольку наука все более делается такими командами. Из биографических очерков и из других источников выясняется, КАКИЕ РЕЗУЛЬТАТЫ получены вследствие финансирования.

Однако продолжать возлагать всю отчетность на исследователей непродуктивно. По некоторым оценкам, 42 % времени исследователей с государственным финансированием в США тратится на админи-

стративные обязанности, а не на само исследование. Научно-технические агентства в различных странах разрабатывают информационные платформы для выявления и характеристики научных результатов. Множество результатов доступны в цифровой форме и обрабатываются такими поисковыми службами, как «CiteSeerX», «Google Scholar» и «Microsoft Academic Search». Разрабатываются все более точные методы для надежной атрибуции научных продуктов авторам, что является нетривиальной задачей вследствие значительной неоднозначности имен авторов.

Результаты исследований становятся все более доступными через Интернет, включая метаданные, а часто и полный текст многих исследовательских публикаций. Появляются каталоги наборов данных. Собственные странички исследователей в сети и биографические очерки (резюме) также служат источниками информации. Здесь можно указать на американский проект SciENcv [2] и бразильскую веб-платформу LATTES [3, 4], которые позволяют исследователям описывать свои научные результаты. Цифровые данные анализируются различными современными методами, такими как тематическое моделирование [5, 6].

2.2. Бразильская национальная система LATTES

Платформа Латтес [3, 4] представляет собой доступный через Интернет набор баз данных и программных средств, поддерживаемый бразильским министерством науки и технологии, и содержит широкий спектр фактов и сведений об исследованиях и разработках. Платформа Латтес полностью встроена в бразильское академическое сообщество. Центральной частью платформы Латтес является база данных биографических очерков (анкет) пользователей (curriculum vitae, CV). Она включает сведения практически обо всех людях в Бразилии, причастных к исследованиям, начиная от студентов младших курсов до академиков, а также сведения о некоторых иностранных ученых, принимающих участие в финансируемых бразильской стороной проектах.

Структура анкеты исследователя в Латтес была разработана в 1970-х гг. бразильским национальным советом по науке и технологическому развитию (Brazilian National Council of Scientific and Technological Development, CNPq). Первоначальные данные собирались в бумажном виде, а первая автоматизированная версия появилась в 1990-х гг. Сейчас анкета заполняется пользователем через Интернет. Эти анкеты используются всеми финансирующими науку бразильскими агентствами для оценки программ исследований и учебных курсов, а также для оценки проектных предложений. Например, заявки на гранты подаются федеральным агентствам в режиме онлайн и в обязательном порядке снабжаются гиперссылкой на анкеты всех участников в Латтес, чтобы помочь экспертам оценить проектную команду. Благодаря столь широкому распространению платформы Латтес в Бразилии бразильские исследователи постоянно обновляют свои анкеты сведениями о своей работе, новых публикациях и т. п. Поэтому платформа Латтес представляет собой ценный источник для анализа динамики различных областей знания.

В структуре анкеты Латтес предусмотрен ввод информации о многих возможных видах научно-технической деятельности: не только о статьях в научных журналах, но также о прошлых и текущих проектах, руководстве студентами и аспирантами, о патентах, публикациях научно-популярного характера, созданных видеофильмах или об участии в спектаклях. Исходя из столь подробной информации, можно рисовать портреты исследовательской динамики на уровне индивидуумов, групп или целых областей знания. Заметим, что для некоторых областей знания, таких как «литература и искусство», трудно найти иной столь полный и стандартизованный каталог продукции.

Платформа Латтес является самодостаточной и не предполагает обязательной регистрации публикаций в других базах данных, таких как Scopus, Web of Science, или SciELO. Однако в процессе заполнения анкеты имеется возможность поиска публикаций в этих источниках. В сентябре 2013 г. в базе данных Латтес имелось 3,2 млн анкет пользователей, доступных через Интернет. Эта база данных используется бразильскими учеными в различных областях науки для получения информации о профиле исследований, о междисциплинарном сотрудничестве, а также об исследовательских трендах.

Форма анкеты пользователя в системе Латтес включает следующие разделы: общие сведения; образование; опыт работы; проекты; продукция, в том числе библиографическая и техническая, а также другие художественные и культурные произведения; патенты; просвещение и популяризация науки и техники; участие и организация конференций, выставок, ярмарок и олимпиад; научное руководство студентами и аспирантами: участие в научных комитетах, советах, жюри; цитируемость. Для каждой публикации могут указываться следующие признаки: является ли эта публикация одной из пяти ведущих публикаций данного исследователя; носит ли публикация образовательный или научно-популярный характер. Полная печатная форма анкеты может достигать 200 страниц, смотря по тому, сколько информации о себе и своих работах ввел пользователь. В типичном случае ученые заполняют разделы «общие сведения», «опыт работы», сведения об исследовательской продукции (включая публикации) и о научном руководстве.

Форма описания исследовательского проекта в системе Латтес включает следующие разделы: общие сведения» о проекте, проектная команда, включая количество студентов различных уровней подготовки и количество аспирантов; финансирование (сумма финансирования не публикуется в Интернете); описание научно-технической продукции; научное руководство студентами и аспирантами.

2.3. Тематическое моделирование

Тематическое моделирование (topic modeling) [5, 6] представляет собой вероятностную модель, которая автоматически обучается множеству тем (категорий) на множестве документов, анализируя слова в документах. Каждый документ связывается с небольшим количеством тем. Каждая тема характеризуется некоторым набором слов. Гранулярность тем зависит от целей анализа. Важно, что в этом подходе темы документов вычисляются автоматически, исходя например, из текста заявок на гранты, а не вручную, при помощи таксономий или ключевых слов.

Вероятностные тематические модели осуществляют «мягкую» кластеризацию документов. Каждый документ или термин может относиться к нескольким темам с различными вероятностями. Вероятностные тематические модели описывают каждую тему дискретным распределением на множестве терминов, а каждый документ — дискретным распределением на множестве тем. С математической точки зрения предполагается, что коллекция документов — это последовательность терминов, выбранных случайно и независимо из смеси таких распределений, и ставится задача восстановления компонент смеси по выборке.

Одна из самых распространенных тематических моделей — это латентное размещение Дирихле (LDA), эта модель была разработана Дэвидом Блейем. Алгоритм LDA реализован на языке Java в программе MALLET (MAchine Learning for LanguagE Toolkit) [7].

Задача построения тематической модели состоит в следующем. Задана коллекция текстовых документов. Каждый документ из коллекции представляет собой последовательность слов. Предполагается, что каждый документ может относиться к одной или нескольким темам. Темы отличаются друг от друга различной частотой употребления слов. Требуется определить число тем, характерные для каждой темы частотные распределения слов и тематику каждого документа: в какой степени он относится к каждой из тем. Целью построения тематической модели может быть как непосредственно выявление множества латентных тем, так и решение различных дополнительных задач. Примеры дополнительных задач:

- ранжировать документы по степени релевантности заданной теме (тематический поиск);
- ранжировать документы по степени тематического сходства с заданным документом или его фрагментом;
- построить иерархический тематический каталог коллекции документов и выработать правила каталогизации новых документов;
- определить, как темы изменялись со временем (предполагается, что для каждого документа известно время его создания);
- определить тематику авторов (предполагается, что для каждого документа известен список авторов);
- определить тематику различных сущностей, связанных с документами (например, журналов, конференций, организаций, стран);
- разбить документ на тематически однородные фрагменты.

2.4. Проект STAR METRICS — административная отчетность

Американский проект STAR METRICS представляет собой совместный проект между американскими федеральными агентствами по науке и исследовательскими институтами [9]. Система STAR METRICS черпает информацию из зарплатных ведомостей, где имеются сведения о занимаемых должностях. Это позволяет напрямую вычислять распределение финансирования по категориям получателей. Например, в типичной лаборатории может быть широкий спектр должностей, включая техников, студентов младших и старших курсов, а также вспомогательный исследовательский персонал. Финансирование студентов и ранний опыт исследований оказывают важное влияние на склонность к научным и техническим областям, поскольку студенты «впитывают» методы исследований. Тем самым воспитывается следующее поколение ученых и инженеров. Данные системы STAR METRICS также позволяют получать информацию о частично занятых участниках исследований (ЧЗУ). Эти данные оказываются гораздо выше, чем количество полностью занятых участников (ПЗУ). В некоторых проектах на каждого ПЗУ приходятся по несколько студентов младших курсов, техников и клиницистов, ассистентов и обслуживающего персонала, получающих часть проектного финансирования.

Исследовательские институты предоставляют в программу STAR METRICS обезличенные статистические данные из своих административных информационных систем. Сведения собираются поквартально. На основании собираемых данных программа STAR METRICS формирует отчеты и присылает их обратно в исследовательскую организацию. Отчеты показывают динамику количества и качества рабочих мест, созданных при помощи полученных грантов.

Состав собираемых данных в программе STAR METRICS включает сведения о гранте; размер накладных расходов, удержанных институтом из гранта в отчетный период; обезличенные сведения о людях, включая внутренний табельный номер исполнителя в институте (но не глобальный персональный номер, такой как страховой); категория персонала: штатный научный сотрудник, студент младших или старших курсов и т. п.; процент занятости участника: целая ставка или полставки; доля гранта, доставшаяся данному сотруднику в отчетный период, от 0 до 1; косвенные затраты на персонал помимо почасовой оплаты, такие как сверхурочные и премии; суммы платежей поставщикам; а также сведения о «дочерних» грантах сторонним организациям за счет основного гранта, полученного институтом.

3. Оценка результативности научных проектов

Рассмотрим задачу многокритериальной оценки результативности проектов целевых фундаментальных исследований, поддержанных Российским фондом фундаментальных исследований (РФФИ) и выполняемых в интересах федеральных агентств и ведомств. При формализации результативности целевых фундаментальных исследований за исходные показатели были приняты критерии, содержащиеся в анкете РФФИ для экспертной оценки отчета о выполнении проекта, которые учитываются при формировании интегрального показателя результативности проекта [11].

Анкета экспертизы отчета состоит из двух разделов: оценка полученных результатов и ожидаемые результаты завершающей стадии проекта. Раздел «Оценка полученных результатов проекта» включает 4 критерия:

- Критерий K1. Степень выполнения заявленных задач проекта. Оценки: q11 — задачи выполнены полностью; q12 — задачи выполнены частично, имеющееся отставание несущественно; q13 — задачи выполнены частично, имеется существенное отставание.
- Критерий K2. Оценка научного уровня полученных результатов. Оценки: q21 — превосходит уровень имеющихся решений; q22 — находится на уровне имеющихся решений; q23 — уступают уровню имеющихся решений.
- Критерий K3. Патентоспособность полученных результатов. Оценки: q31 — получены охраноспособные результаты; q32 — патентование нецелесообразно.
- Критерий K4. Перспективы использования полученных результатов. Оценки: q41 — результаты работ уже используются; q42 — идет подготовка к использованию и коммерциализации результатов; q43 — перспективы использования и коммерциализации результатов неясны.

Раздел «Ожидаемые результаты завершающей стадии проекта» характеризует возможности практической реализации проекта и состоит из 4 критериев:

- Критерий K5. Ожидаемые результаты завершающего этапа выполнения проекта. Оценки: q51 — соответствуют заявленной цели проекта; q52 — не соответствуют заявленной цели проекта.
- Критерий K6. Решение задач, поставленных в завершающей части проекта. Оценки: q61 — реально; q62 — не реально.
- Критерий K7. Наличие трудностей в работе по проекту. Оценки: q71 — есть; q72 — нет.
- Критерий K8. Взаимодействие с организациями, в которых предполагается использовать результаты проекта (заполняется только для итогового отчета). Оценки: q81 — имеются договоры о взаимодействии с организациями; q82 — имеются договоры о взаимодействии, но деловые контакты развиты недостаточно; q83 — взаимоотношения документально не оформлены, взаимодействие слабое.

Формально множество научных проектов описывается восемью показателями (критериями) K1, …, K8, которые имеют следующие шкалы: $X1 = \{0, 1, 2\}$; $X2 = \{0, 1, 2\}$; $X3 = \{0, 1\}$; $X4 = \{0, 1, 2\}$; $X5 = \{0, 1\}$; $X6 = \{0, 1\}$; $X7 = \{0, 1\}$; $X8 = \{0, 1, 2\}$, где 0 обозначает лучшую оценку, 1 — среднюю (или худшую), 2 — худшую. Таким образом, размерность исходного признакового пространства $X1 \times \ldots \times X8$ равна 1296. Критерием верхнего уровня является «Результативность проекта», градации оценок по шкале которого (наивысшая, высокая, средняя, низкая, неудовлетворительная) определяют упорядоченные классы решений D1, …, D5. Требуется разбить множество комбинаций градаций оценок на пять классов результативности $D1f \ldots fD5$. Очевидно, что непосредственная классификация 1296 вариантов связана с существенными трудозатратами ЛПР.

Построение интегрального показателя результативности научного проекта рассматривается как решение задачи многокритериальной порядковой классификации по иерархической системе критериев, которая строится с помощью технологии ПАКС путем снижения размерности признакового пространства. В качестве многопризнаковых объектов выступают комбинации градаций оценок проектов по критериям, агрегированные показатели играют роль классов решений. ЛПР имеет возможность различным образом формировать понятие «результативность проекта» и сравнивать интегральные показатели, сконструированные различными способами.

Литература

1. *Петровский А. Б.* Теория принятия решений. М.: Издательский центр «Академия», 2009. 400 с.
2. Проект SciENcv на сайте американского национального центра по биотехнологической информации (NCBI). http://www.ncbi.nlm.nih.gov/books/NBK154494/.
3. Бразильская национальная платформа Латтес (LATTES). http://lattes.cnpq.br.
4. *Pacheco R. C. S.* The role of Lattes Platform in the Brazilian Innovation System. http://www.nsf.gov/attachments/123272/public/1.Pacheco.pdf
5. Сайт Д. Блея по тематическому моделированию. http://www.cs.princeton.edu/~blei/topicmodeling.html.
6. *Blei D. M.* Probabilistic Topic Models // Communications of the ACM. April 2012. V. 55. № 4. P. 77–84.
7. MAchine Learning for LanguagE Toolkit. http://mallet.cs.umass.edu/topics.php.
8. *Foster I., Lane J.* Science Based Measures of Science Investments. https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=ESAMACE2014&paper_id=114
9. Science and Technology for America's Reinvestment Measuring the EffecTs of Research on Innovation, Competitiveness and Science (STAR METRICS). https://www.starmetrics.nih.gov/.
10. *Петровский А. Б.*, *Ройзензон Г. В.* Многокритериальный выбор с уменьшением размерности пространства признаков: многоэтапная технология ПАКС // Искусственный интеллект и принятие решений. 2012. № 4. С. 88–103.
11. *Петровский А. Б.*, *Ройзензон Г. В.*, *Балышев А. В., Тихонов И. П.* Ретроспективный анализ результативности научных проектов // International Journal „Information Models and Analyses“. 2012. V. 1. № 4. P. 349–356.

Белов Сергей Павлович. Декан НИУ Белгородского ГУ. С. н. с. Д. т. н. Окончил в 1975 г. Харьковское высшее военное командно-инженерное училище им. Маршала Советского Союза Крылова. Количество печатных работ: 82. Область научных интересов: информационные технологии и вычислительные системы. E-mail: belov@bsu.edu.ru

Плискин Евгений Львович. В. н. с. ИСА РАН. К. т. н. Окончил в 1982 г. МФТИ. Количество печатных работ: 14. Область научных интересов: автоматизированные информационные системы. E-mail: pliskin@cognitive.ru

Усков Анатолий Васильевич. Зав. лабораторией ИСА РАН. К. ф.-м. н. Окончил в 1961 г. МГУ. Количество печатных работ: более 50 (в том числе 3 монографии). Область научных интересов: искусственный интеллект и системное программирование. E-mail: uskov@cognitive.ru