

УДК 001.38
ББК 72.4+73.4

ОБ ОДНОМ ПОДХОДЕ К ТИПИЗАЦИИ УЧЕНЫХ ПО БИБЛИОМЕТРИЧЕСКИМ ДАННЫМ

Васильев И. И.¹,

(Московский физико-технический институт, Москва)

Чеботарев П. Ю.²

(Учреждение Российской академии наук

Институт проблем управления РАН, Москва)

Предложен набор показателей для решения задачи типизации ученых по библиометрическим данным методами кластерного анализа. Алгоритм иерархической кластеризации Уорда применен к множествам математиков, физиков и психологов с высокими показателями цитируемости. Анализ полученных результатов позволяет не только выделить устойчивые типы ученых, но и исследовать отличия разбиений ученых на группы в различных научных дисциплинах.

Ключевые слова: типизация ученых, наукометрия, библиометрия, индексы цитирования, кластерный анализ, Google Scholar.

1. Введение

Количество ученых в мире продолжает расти быстрыми темпами. Еще быстрее растет число публикуемых научных работ. Однако, прирост количества не означает прирост качества.

¹ *Илья Игоревич Васильев, бакалавр, студент магистратуры (ilya.vasilev@phystech.edu).*

² *Павел Юрьевич Чеботарев, доктор физико-математических наук, заведующий лабораторией (Москва, ул. Профсоюзная, д. 65, тел. (495)335-18-05; pavel4e@gmail.com).*

Вопрос качества научных работ является центральным для принятия решений в области научной политики [5, 6]. Большой интерес представляют задачи оценки сравнительной влиятельности ученых, их вклада в развитие науки, типизации ученых и др. [9, 10, 13].

В данной работе мы сосредоточимся в основном на последней задаче. Можно ли выделить отчетливые «типы» ученых, характеризующиеся стилем их работы, динамикой восприятия их научным сообществом, влиянием их публикаций и т.п.; каковы доли ученых, относящихся к каждому из типов; сильно ли эта типизация зависит от области науки?

Цель статьи – выяснить, могут ли осмысленные ответы на эти вопросы быть получены на основании простейшей информации о цитировании публикаций, предоставляемой библиографическими базами данных, такими как [Web of Science](#), [Scopus](#), [Microsoft Academic Search](#), [Google Scholar](#) и др. В качестве источника данных в статье используется Google Scholar.

Применяемый подход основан на следующей идее. Типы ученых могут не задаваться извне, а выявляться в результате кластеризации множества исследователей по показателям цитируемости и интерпретации найденных кластеров, то есть посредством машинного обучения без учителя.

Аналізу научной активности на основании библиометрической информации посвящено большое и быстро растущее число исследований (см., например, [1, 3, 5–8]). Для оперативного отслеживания потока публикаций по наукометрии может быть рекомендован Интернет-ресурс [Sciencemetrics](#).

Одной из наиболее близких к данной статье по постановке задачи является работа [13]. В ней для кластеризации ученых используется евклидово расстояние между индивидуальными профилями цитируемости, нормализованное делением на общие количества ссылок в каждом профиле. Эта метрика относится к интегральным. Для более тонкого учета особенностей зависимости цитируемости от времени в настоящей работе используется ряд дифференциальных характеристик. Сравнимый подход был применен в [9], где в дополнение к общему числу ссылок и индексу Хирша для характеристики кривых цитируемости использовался

так называемый индекс перфекционизма, штрафующий за низкоцитируемые статьи, причем кластеризация проводилась в группах ученых, однородных по академическому стажу, и отдельно для разных периодов времени. Однако, фокусом исследования в [9] была не типизация ученых, как в настоящей статье, а оценка их влиятельности. Рафинированную дифференциальную меру, а именно, фрактальную размерность кривой цитируемости, те же авторы предлагают и исследуют в следующей своей статье [10], но эта мера пока не использовалась для кластеризации ученых.

Отметим интересный содержательный подход к ранжированию ученых, предложенный в [4, 12] и использующий метод Linstrat. Определенным ограничением, связанным с этой методикой, является относительная сложность первичной подготовки данных об ученых. А именно, требуется специальное «кодирование» трудов каждого автора с использованием общих или специальных таксонометрических справочников (например, ACM CCS 2012).

В отличие от этого, подход, используемый в данной работе, не требует никакой специальной (и не полностью формализованной) подготовки данных.

2. Основные понятия и используемые показатели

Под библиометрией понимается применение математических и статистических методов к изучению печатных изданий разного рода. В данной работе используются лишь библиометрические показатели [1, 7, 8], характеризующие цитирование работ ученых (и косвенно – их публикационную активность). Источником данных является платформа Google Scholar, предоставляющая открытый доступ к информации о научных публикациях и их цитировании. Каждый ученый, зарегистрированный в этой системе, указывает до пяти ключевых слов («тегов»), характеризующих область его научной деятельности. К профилю ученого в Google Scholar подключены библиографические описания его публикаций, для каждой публикации – информация о работах, где имеются ссылки на нее, а также ряд интегральных показателей: гистограмма ссылок на работы ученого по годам, индекс Хирша, индекс Хирша за последние пять лет, индекс $i10$ (число работ, цитируемых не менее десяти раз), общее число ссылок на

работы автора, число ссылок за последние 5 лет.

В число используемых для кластеризации показателей имеет смысл включить следующие индексы цитируемости, которые будем называть *кумулятивными* показателями, поскольку они являются неубывающими функциями от времени:

1. Индекс Хирша – наибольшее число h публикаций автора, на каждую из которых имеется не менее, чем h ссылок;
2. Индекс $i10$ – число публикаций автора, на которые имеется не менее десяти ссылок;
3. $Citat$ – общее количество ссылок на работы автора.

Кроме классических индексов цитируемости, в работе используются несколько производных показателей, которые будем называть *структурными*.

Идея их введения состоит в следующем. Одной из целей исследования является идентификация таких типов научной активности, которые долгое время могут сохраняться на протяжении исследовательской карьеры – даже при заметном росте кумулятивных показателей. Естественным при этом является свойство *масштабируемости*, т.е. сохранения значения показателя при пропорциональном изменении годовых количеств ссылок на работы автора за весь рассматриваемый период.

Здесь понадобятся специальные термины. Под *убыванием цитируемости за год* будем понимать уменьшение числа ссылок на работы автора за год по сравнению с числом ссылок за предыдущий год более чем на 4%. Аналогично *приращение цитируемости за год* – это увеличение годового числа ссылок на работы автора более чем на 4% по сравнению с предыдущим годовым числом ссылок. Использование этих терминов позволяет незначительные колебания цитируемости трактовать как стабильность.

Ниже описанию каждого из предлагаемых структурных показателей предшествует его обозначение.

1. Add (от «addition») – сумма годовых *приращений цитируемости* с 2000 по 2015 г., отнесенная к среднегодовому числу ссылок. Если имеется начальный отрезок этого периода, за который работы ученого не цитировались, то он исключается из периода усреднения: этот отрезок, скорее всего, предшествует профессиональной «инициации» ученого.

Данный показатель характеризует относительный рост цитируемости за исследуемый период, достигнутый за счет «существенных» (более 4%) годовых увеличений.

2. *Ded* (от «*deduction*») – аналогичный предыдущему показатель, измеряющий относительную величину убывания за рассматриваемый период. Он равен сумме годовых *убываний цитируемости* с 2000 по 2015 г., отнесенной к среднегодовому числу ссылок автора. Если имеется начальный отрезок этого периода, за который работы ученого не цитировались, то он исключается из периода усреднения.
3. *Ssc* («*sign switching count*») – количество *смен знака* разности смежных годовых «существенных» приращений цитируемости в течение зафиксированной карьеры ученого (не ранее, чем с 1977 г.) до 2015 года. Указанную смену знака можно также охарактеризовать как локальный экстремум годовой цитируемости.
4. *Mis* (от «*maximal increasing series*») – максимальное число лет, идущих подряд, за период исследования, на протяжении которых наблюдалось *приращение* годового числа цитирований.
5. *Mds* (от «*maximal decreasing series*») – аналогично предыдущему показателю, максимальная длина последовательности *убываний* годовой цитируемости.

3. Постановка задачи

Далее в работе решается следующая задача: используя представленный набор библиометрических показателей, провести иерархическую кластеризацию нескольких множеств ученых (относящихся к разным наукам), предложить интерпретацию полученных кластеров и выяснить, выявляет ли кластеризация существенные различия между рассматриваемыми множествами ученых (и науками).

Полученная кластеризация должна давать возможность «на лету» относить ученых, не входящих в исходные кластеризуемые множества, но специализирующихся в тех же науках, к определенным классам и, тем самым, формулировать гипотезы об их роли в научном сообществе.

4. Применяемые методы

4.1. ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ МЕТОДОМ УОРДА

Напомним, что кластеризацией называют разбиение множества на подмножества схожих элементов. Каждое такое подмножество называют кластером. В отличие от классификации, при кластеризации подмножества не имеют ни априорных описаний, ни заранее известных представителей.

Среди алгоритмов иерархической кластеризации [2, 11] выделяются два основных типа: «восходящие» и «нисходящие» алгоритмы. Нисходящие алгоритмы работают по принципу «сверху-вниз»: вначале все объекты помещаются в один кластер, который затем разделяется на все более мелкие подкластеры. Более распространены восходящие алгоритмы, которые вначале помещают каждый объект в свой индивидуальный кластер, а затем объединяют кластеры во все более крупные, пока число кластеров не уменьшится до двух. И в том, и в другом случае строится система вложенных разбиений. Результаты таких алгоритмов обычно представляют в виде дерева, которое называют *дендрограммой*. Пример дендрограммы, полученной в данном исследовании, показан на Рис. 1.

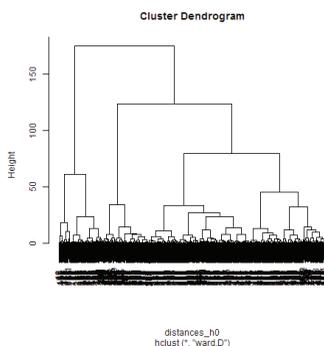


Рис. 1. Пример дендрограммы иерархической кластеризации. Вертикальные линии соответствуют выделенным кластерам. По вертикали – номер итерации.

Важным элементом любого алгоритма иерархической кластеризации является выбор метрики – функции, рассчитывающей расстояние между кластерами на каждой итерации. Примеры используемых метрик:

1. Евклидово расстояние: $\rho(x, x') = \sqrt{\sum_i^n (x_i - x_i')^2}$;
2. Квадрат евклидова расстояния: $\rho(x, x') = \sum_i^n (x_i - x_i')^2$;
3. Манхэттенская метрика (расстояние «городских кварталов»): $\rho(x, x') = \sum_i^n |x_i - x_i'|$;
4. Расстояние Чебышева: $\rho(x, x') = \max\{|x_i - x_i'|\}$.

В работе используется метод Уорда [14], признанный одним из лучших методов иерархической кластеризации. В нем в качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате объединения этих кластеров. На каждом шаге алгоритма объединяются два кластера, самые близкие в указанной метрике. В методе Уорда важную роль играет так называемая «стоимость слияния», которая для двух множеств А и В рассчитывается по формуле

$$\Delta(A, B) = \frac{n_A n_B}{n_A + n_B} \|\vec{m}_A - \vec{m}_B\|^2, \text{ где } \vec{m}_j - \text{центр кластера } j.$$

На первой итерации сумма квадратов равна нулю, так как каждый объект образует отдельный кластер. Затем по мере объединения кластеров эта сумма растет: на каждом шаге объединяются два кластера, «стоимость слияния» которых минимальна.

4.2. ИСПОЛЬЗУЕМЫЕ СРЕДСТВА РАЗРАБОТКИ

Анализ данных разделяется в работе на два этапа. Первый – сбор и первичная обработка данных, второй – кластеризация и интерпретация результатов.

Модули сбора данных реализованы на языке программирования Python с использованием библиотек BeautifulSoap и Selenium. Хранение организовано в базе данных Postgre, находящейся в свободном доступе.

Анализ данных проводится с использованием языка программирования R со встроенными библиотеками кластеризации, построения диаграмм и др.

5. Методика исследования

Для исследования отобраны четыре множества ученых. Они составлены из наиболее цитируемых ученых – тех, которые образуют начальные фрагменты выдачи при поиске в системе Google Scholar по тегам «Mathematics», «Physics» и «Psychology». Таким образом получены следующие множества:

- a. «Математики» – первые 500 авторов при поиске по тегу «Mathematics».
- b. «Математики+» – 543 ученых, полученных поиском по тегу «Mathematics» со сдвигом на 198 позиций. Тем самым это множество включает 302 «последних» ученых из множества «Математики» и отличается от него примерно на 45%.
- c. «Физики» – 515 ученых с тегом «Physics». 99,2% из них не входят в множество «Математики».
- d. «Психологи» – 556 ученых с тегом «Psychology». 99,9% из них не входят в множество «Математики».

Данные были получены из системы Google Scholar с помощью скрипт-файла на языке Python. Имена ученых не анализировались и не сохранялись; записи, идентифицированные индивидуальными кодами, заносились в базу данных Postgre.

Отметим, что полученные выборки можно использовать для типизации не всех ученых, а лишь весьма успешных – тех, чьи работы получили заметное признание. Более тонкая особенность этих выборок: в них входят ученые, в числе тегов указавшие не только частные разделы своей науки, но и название науки в целом.

Для работы с данными использовался язык R. Выбранные показатели имели существенно разные шкалы, поэтому для сопоставимости преобразованием сжатия они были приведены к единичной дисперсии. При представлении профилей цитируемости на диаграммах годовые значения цитируемости ученых нормировались на их средние значения по индивидуальным профилям.

Далее методом Уорда выполнялась иерархическая кластеризация для каждого из множеств ученых. Она строилась в двух вариантах: (1) по кумулятивным и структурным показателям; (2) только по структурным показателям. Поскольку индекс Хирша имеет довольно высокую корреляцию с показателем i_{10} ,

использовался лишь первый. Таким образом, кумулятивные показатели включали индекс Хирша (обозначение: «h_index»), индекс Хирша за период¹ 2011-2015 гг. («h_index_last») и число цитирований за тот же период («Citat_last»).

Сначала строилась кластеризация с тремя кластерами, затем их число последовательно увеличивалось до шести. Особое внимание уделялось интерпретации результатов и сравнению разных множеств и кластеризаций с разными наборами показателей.

6. Результаты и их интерпретация

Ниже приведены результаты кластерного анализа для рассмотренных множеств авторов.

6.1. КЛАСТЕРИЗАЦИЯ МАТЕМАТИКОВ

Методом Уорда множество «Математики» было разделено на три достаточно компактные группы (Рис. 2).

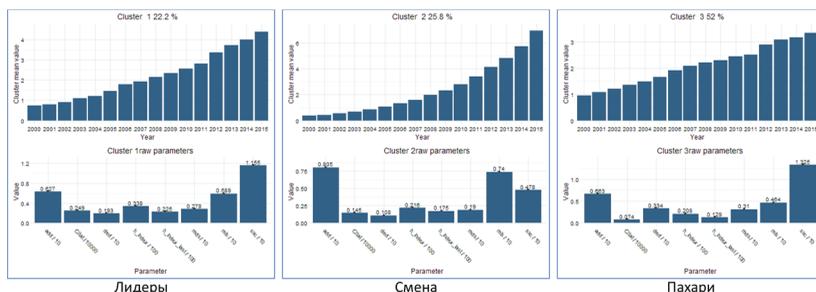


Рис. 2. «Математики»: три кластера.

Первый кластер, составляющий примерно половину множества «Математики», характеризуется следующими признаками.

1. Давнее начало карьеры (в 2000 г. уже достаточно высокий уровень цитируемости).

¹ Отметим, что если бы последний пятилетний период рассматривался в работе как «плавающий», то индексы за этот период нельзя было бы считать кумулятивными, т.к. он допускали бы убывание.

2. Рост цитируемости в среднем – не очень быстрый и по преимуществу линейный, у части – с насыщением. Средние значения некоторых показателей этому по кластеру¹:
 - a. $h_index_last/h_index = 0,62$;
 - b. $Citat_last = 740$.
3. Убывание цитируемости по сравнению с предыдущим годом встречается регулярно и имеет заметную амплитуду:
 - a. $ssc = 13,25$;
 - b. $ded/add = 0,5$ (существенное убывание составляет половину от существенного приращения);
 - c. $mds/mis = 0,68$ (максимальные последовательности убывания составляют чуть более $2/3$ от максимальных последовательностей возрастания).

Второй кластер включает чуть менее четверти множества и характеризуется следующими признаками.

1. Также достаточно давнее начало карьеры.
2. Рост годовой цитируемости более быстрый, чем в первом кластере и нелинейный – функция выпукла вниз; средние:
 - a. $h_index_last/h_index = 0,67$ – несколько выше, чем для первого кластера; $2/3$ текущего значения индекса Хирша накоплены² за последнюю пятилетку;
 - b. $Citat_last = 2490$ – в 3,4 раза выше, чем для первого кластера.
3. Убывание цитируемости встречается реже, чем в первом кластере и имеет существенно меньшую амплитуду:
 - a. $ssc = 11,55$;
 - b. $ded/add = 0,26$;
 - c. $mds/mis = 0,47$.

Оставшаяся четверть множества «Математики» (кластер 3) – достаточно молодые и весьма успешные ученые.

1. В 2000 г. их цитируемость еще незначительна, а чаще отсутствует.

¹ Знак « \approx » используем как для точных, так и для приблизительных равенств.

² Отметим, что индекс Хирша, в отличие от общего числа цитирований, не аддитивен по времени.

2. Рост цитируемости быстрый и близкий к квадратичному;
 - a. $h_index_last/h_index = 0,81$ – существенно бóльшая доля накоплена за последние годы.
 - b. $Citat_last = 1450$ – по общему числу цитирований они вдвое обгоняют кластер 1 и постепенно приближаются к более опытным ученым из кластера 2.
3. Убывание годовой цитируемости встречается редко и несравнимо по амплитуде с возрастанием:
 - a. $ssc = 4,78$;
 - b. $ded/add = 0,14$;
 - c. $mds/mis = 0,26$ – все эти показатели примерно вдвое ниже, чем в предыдущем кластере.

Для краткости выделенные три группы ученых можно охарактеризовать следующим образом.

Кластер 1 – «пахари»: ученые, добившиеся высоких показателей многолетним результативным трудом, не принесшим, однако, широко признанных достижений (для таких достижений характерен квадратичный, «вирусный» рост цитируемости).

Кластер 2 – «лидеры»: опытные ученые, имеющие достаточно широко известные достижения, обеспечившие им рост цитируемости, опережающий публикационную активность автора и в среднем близкий к квадратичному.

Кластер 3 – «смена», будущие лидеры, уже имеющие яркие достижения, позволившие в среднем обогнать «пахарей» по общему числу цитирований.

Анализ значений индексов по кластерам свидетельствует в пользу осмысленности введенных условных наименований. Но, строго говоря, они служат всего лишь мнемоническими метками разных типов профилей показателей. Вопрос, насколько семантика меток соответствует этим профилям, может быть предметом обсуждения. Однако, основных результатов работы такая дискуссия не затронет, поскольку при интерпретации результатов метки могут быть заменены номерами кластеров, и останутся лишь четкие утверждения о соотношении исследуемых показателей для полученных кластеров.

6.2. ДРОБЛЕНИЕ КЛАСТЕРОВ МАТЕМАТИКОВ

Метод Уорда позволяет «продолжить» кластеризацию, в результате чего полученные кластеры в определенном порядке «разделяются»¹ на компактные подкластеры. В работе проводились три итерации такого дробления.

Для множества «Математики» на первой итерации кластер «пахарей» разделился на два подкластера в количественном соотношении 2,6 : 1 (Рис. 3). Меньший подкластер (примерно 1/7 исходного множества) объединяет математиков, у которых было высокое цитирование еще в 2000 г., в 2008-2011 гг. – в среднем наблюдается стабилизация годовой цитируемости, в 2012 г. – некоторый рост, а затем намечается спад. Часть из них, по-видимому, завершает карьеру, причем влияние их результатов со временем снижается. У них фактически $ded = add$, $ssc = 18,12$. После выделения их в подкластер у оставшихся – более высокое по сравнению с ними цитирование на конец периода исследования.

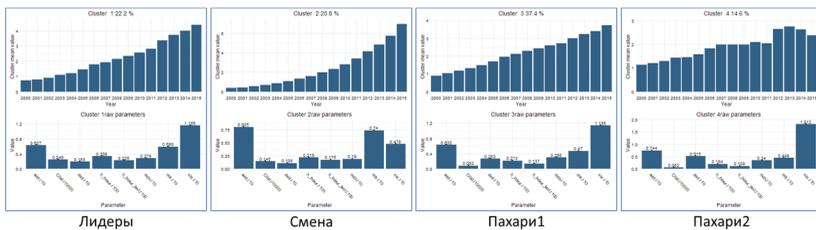


Рис. 3. «Математики»: четыре кластера.

Далее разделяется кластер «смена» (Рис. 4). Из него выделяются 4,2% (здесь и далее проценты указываются от всего исследуемого множества) самых молодых и успешных. В 2000-2002 ни у кого из них еще не было цитирований. Им абсолютно незнано убывание цитируемости: $ded = mds = ssc = 0$. Не претендуя на буквальность, эту группу можно назвать «акселераты».

¹ Кавычки напоминают, что логика метода Уорда обратна логике изложения: кластеры не разделяются, а объединяются, что полезно учитывать и далее.

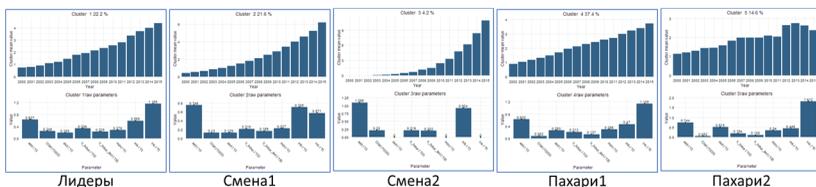


Рис. 4. «Математики»: пять кластеров.

Наконец, на третьей итерации разделяется (Рис. 5) кластер «лидеров» (исходно он составлял 22,2%). Из него выделяются 6,2% ученых, у которых почти не было убывания цитируемости: $ded/add = 0,09$; $ssc = 6,19$ (лишь немногим больше, чем в исходном кластере молодых и высоко-успешных, и его менее «взрывной» части). В случае яркого продолжения карьеры в этот подкластер по прошествии лет попадут ученые из кластера «4,2% самых молодых и успешных», выделившегося на предыдущем шаге.

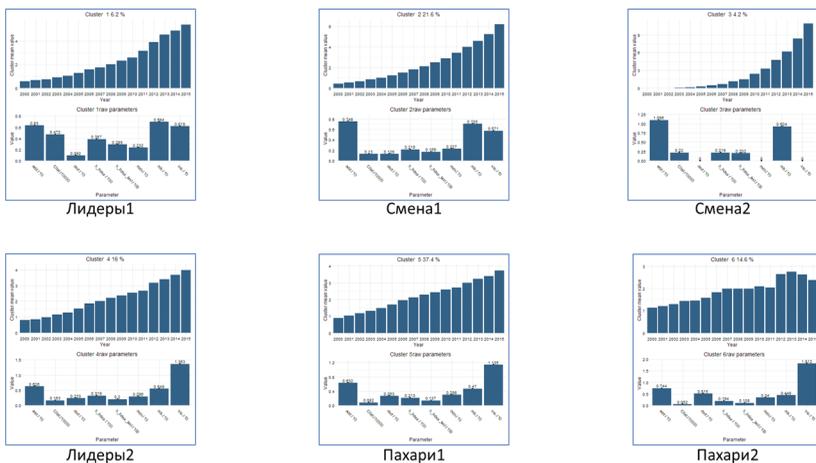


Рис. 5. «Математики»: шесть кластеров.

А по показателям он похож на менее «взрывную» часть молодых, но заметно обгоняет ее по общему числу ссылок и индексу Хирша – в силу большего научного стажа.

6.3. КЛАСТЕРИЗАЦИЯ «МАТЕМАТИКОВ» БЕЗ УЧЕТА ОБЩЕГО ЧИСЛА ССЫЛОК И ИНДЕКСА ХИРША

Для краткости мы называем общее число ссылок и индекс Хирша *кумулятивными показателями* (сокращенно *к.п.*). Представляет большой интерес вопрос о сравнении кластеризации, полученной выше, с кластеризацией без учета кумулятивных показателей – на основе лишь структурных показателей.

В Таблица 1 и Таблица 2 собраны значения показателей по трем начальным кластерам для выборок математиков, физиков и психологов, а именно, для «top-списков» тех, кто включил название одной из этих дисциплин в набор своих тематических тегов.

Можно заметить, что присутствие кумулятивных показателей помогает классифицировать ученых с характерным возрастающим трендом на более и менее опытных авторов (Таблица 1, кластеры 1 и 3 соответственно).

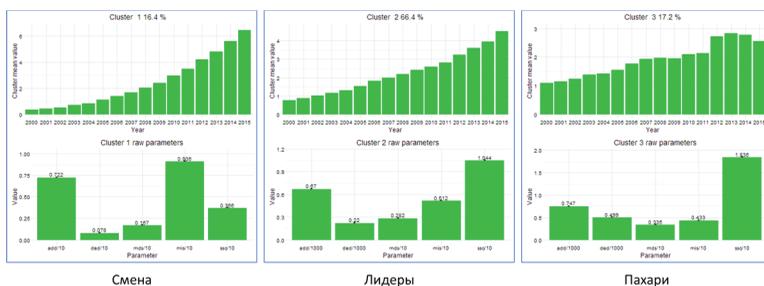


Рис. 6. «Математики»: три кластера без к.п.

При отсутствии кумулятивных показателей (Рис. 6) самым большим из исходных кластеров оказывается кластер «лидеров». Он пополняется теми, кто похож на лидеров скорее структурно (по форме зависимости цитируемости от времени), чем количественно. В кластере «пахарей» теперь в среднем наблюдается не просто насыщение, а даже небольшой спад цитируемости в конце. Более того, их средний тренд цитируемости имеет характерную особенность: насыщение к 2009-2011 г., затем заметный рост в 2012 г. и новое насыщение со спадом в 2015 г. Таким образом, отказ от кумулятивных показателей приводит к кластерам, отличия профилей в которых проявляются более рельефно.

При иерархической кластеризации первым разделяется кластер «лидеров» – в пропорции 4 : 3 (Рис. 7). Его подкластеры объединяют соответственно авторов с медленнее и быстрее растущей цитируемостью. У первых этот рост фактически линейный. При кластеризации с кумулятивными показателями эти ученые попали в кластер «пахарей» – теперь же они вошли в кластер «лидеров». У вторых рост вогнутый (выпуклый вниз): цитируемость растет быстрее, чем число собственных работ, поскольку количество труда переходит в более высокое качество результатов.

Тем самым без кумулятивных показателей в кластер «пахарей» попадают лишь те, чье влияние со временем в среднем не растет. К таким должны, в частности, относиться ученые, переставшие производить существенно новое; некоторые из них «исписались» – продолжают выдавать рутинную продукцию, не вызывающую высокого интереса коллег. Их можно условно назвать «инерционными».

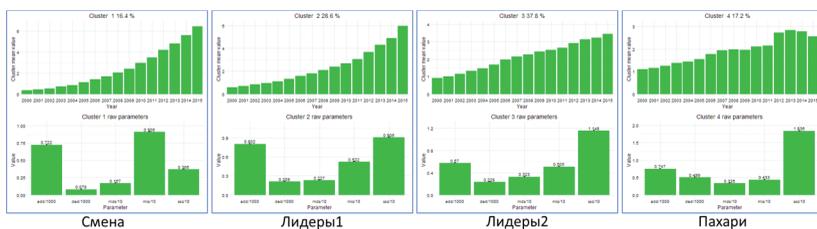


Рис. 7. «Математики»: четыре кластера без к.п.

Затем разделяется кластер «смены» (Рис. 8): из него выделяются 4,4% самых молодых и «взрывных» (аналогично кластеризации с кумулятивными показателями) и 12% ученых, чья работа в среднем началась раньше, а динамика цитируемости в последние пять лет сменила вогнутый рост на линейный.

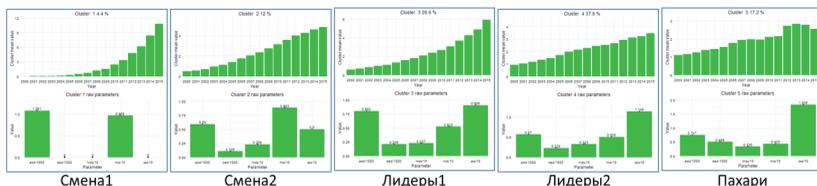


Рис. 8. «Математики»: пять кластеров без к.п.

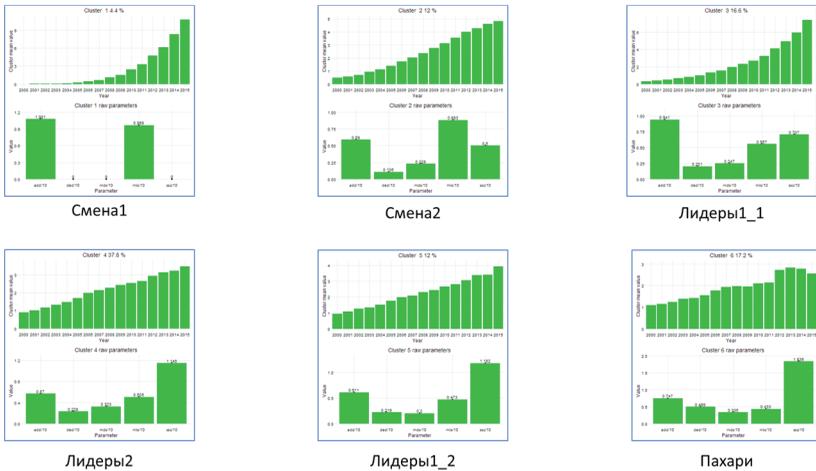


Рис. 9. «Математики»: шесть кластеров без к.п.

Наконец, третьим разделяется не кластер «пахарей», как можно было ожидать, а подкластер «лидеров с вогнутым ростом», составлявший 28,6% (Рис. 9). Из него выделяются ученые (16,6%) с самым быстрым вогнутым ростом и 12% с несколько более высоким показателем неустойчивости роста (ssc) и всплеском роста лишь в последний год периода наблюдения.

Таким образом, приходим к выводу, что отказ от кумулятивных показателей изменяет распределение ученых по кластерам, иначе устанавливая линии разграничения между ними. Это приводит к выделению уже на первом этапе кластера «инерционных», который в присутствии кумулятивных показателей выделялся лишь на втором шаге. Ученые с быстрым линейным ростом цитируемости в отсутствие кумулятивных показателей попадают в кластер «лидеров», а в присутствии их – в кластер «пахарей». Тем самым в конечном итоге выделяются примерно те же подгруппы ученых (что свидетельствует в пользу их действительной компактности), но в ином порядке.

6.4. КЛАСТЕРИЗАЦИЯ НАБОРА «МАТЕМАТИКИ+»

«Математики+» – это множество «Математики», рассмотренное ранее, к которому добавлено чуть больше половины его

численности – ученые, стоящие далее в списке, упорядоченном по убыванию цитируемости, – и несколько меньший начальный отрезок множества исключен из рассмотрения. Тем самым набор данных обновлен на 48,5%.

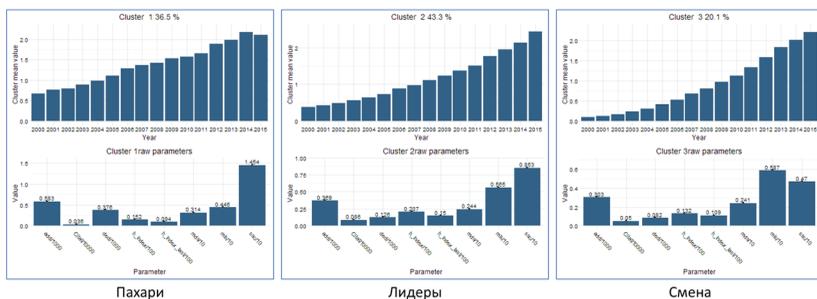


Рис. 10. «Математики+»: три кластера.

Средний научный стаж в этом множестве ниже, чем в исходном: примерно 40% в 2000 г. имели нулевые либо пренебрежимо малые значения цитируемости. Как изменятся типы ученых, выделяющиеся в этом множестве?

При кластеризации множества «Математики+» на три группы (Рис. 11 Рис. 10) доля кластера «лидеры» выше, чем для исходного множества математиков, а доля «пахарей» ниже. Фактически размеры кластеров здесь – средние между значениями при кластеризации множества «Математики» с учетом и без учета кумулятивных показателей.

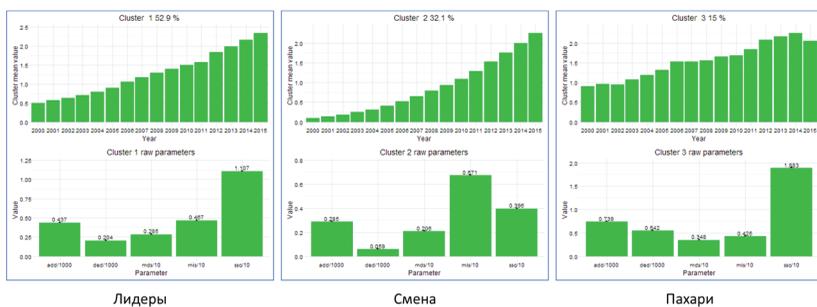


Рис. 11. «Математики+»: три кластера без к.п.

При кластеризации «Математиков+» без кумулятивных показателей (Рис. 11 Рис. 11) более половины элементов кластера «пахарей», полученного с учетом кумулятивных показателей, переходят в кластеры «лидеров» и «смены». Остаются лишь те, у кого показатель годовой цитируемости характеризуется насыщением.

6.5. РАЗДЕЛЕНИЕ КЛАСТЕРОВ НАБОРА «МАТЕМАТИКИ+»

Первым разделяется кластер «лидеры» (Рис. 12), из которого выделяется подкластер опытных авторов с ростом, близким к линейному (2/3 «лидеров»), и дополняющий его подкластер, объединяющий наиболее молодых лидеров с быстрым вогнутым ростом цитируемости (одна треть).

Далее происходит разделение кластера «пахарей» (Рис. 13). В обоих подкластерах, как и в совокупном кластере, в последние годы наблюдается насыщение показателя цитируемости.

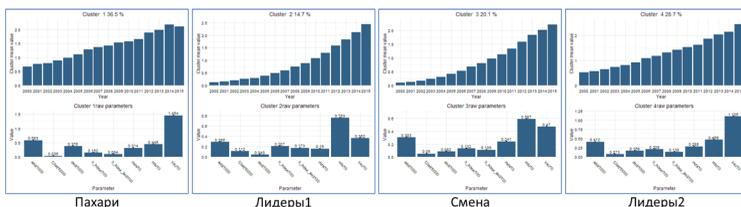


Рис. 12. «Математики+»: четыре кластера.

Первый подкластер объединяет чуть более трети авторов с очень большим стажем: у них почти вдвое выше значения показателей add и ded, чем в дополняющем подкластере, куда входят также опытные, но более молодые. Кроме того, в первом подкластере в 1,6 раз выше показатель ssc, т.е. заметно чаще меняется знак приращения годовой цитируемости.

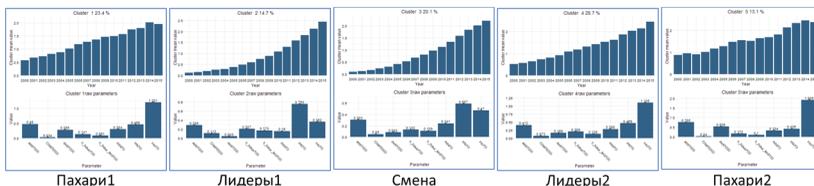


Рис. 13. «Математики+»: пять кластеров.

Наконец, на третьем шаге разделяется не кластер «смена», а выделившийся ранее 15-процентный подкластер «наиболее молодых лидеров» (Рис. 14). Из него выделяются 3,9% самых молодых и «взрывных» ($h_{\text{index}_{\text{last}}}/h_{\text{index}} = 0,95$), не знакомых с убыванием цитируемости. Дополняющая его часть в 2,7 раза больше, в ней стаж ученых выше ($h_{\text{index}_{\text{last}}}/h_{\text{index}} = 0,81$), а показатели убывания (*ded*, *mds*, *ssc*) хотя и малы, но достаточны, чтобы сделать рост цитируемости менее крутым. Показатель *mds* здесь равен 2,19, т.е. в среднем участник этой подгруппы имел хотя бы один двухлетний период убывания цитируемости. Эта подгруппа также состоит из достаточно молодых авторов, и профиль ее похож на профиль кластера «смена», выделившегося вначале. Вместе с тем эти авторы более опытные, чем «смена», и кумулятивные показатели цитируемости у них выше. Именно различие кумулятивных показателей в основном отличает эти две подгруппы.

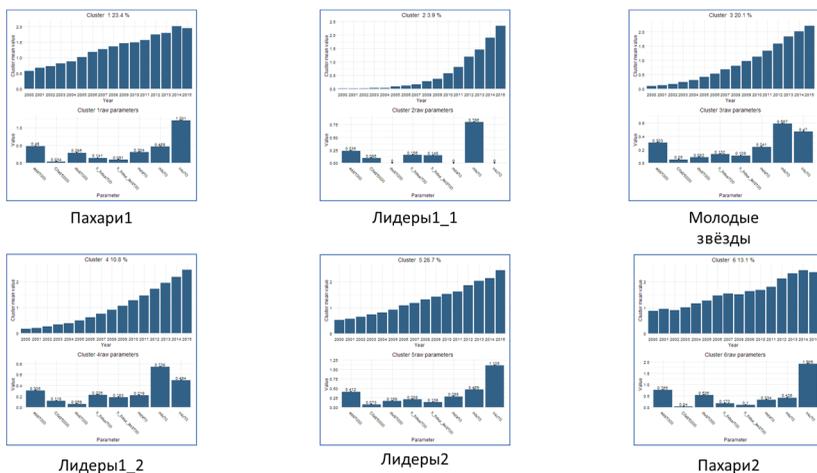


Рис. 14. «Математики+»: шесть кластеров.

Интересно отметить, что подкластер «самых молодых и взрывных лидеров» выделился из кластера «лидеры», а не из кластера «смена». Это значит, что в случае учета кумулятивных показателей у авторов из этого подкластера качества «лидеров» проявляются сильнее, чем отличительные особенности «смены».

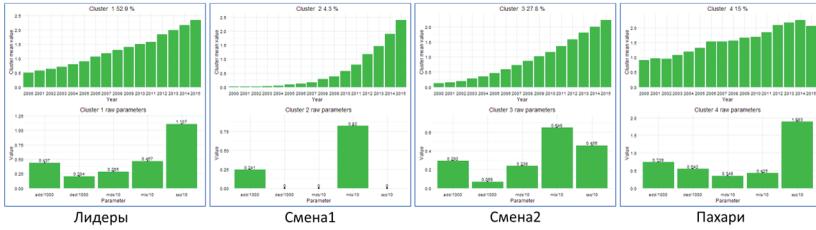


Рис. 15. «Математики+»: четыре кластера без к.п.

В отсутствие кумулятивных показателей первым разделяется кластер «смена» (Рис. 15). Как и при к.п., он образует подкластеры авторов, не знакомых с убыванием цитируемости и авторов с более медленным ростом цитируемости. Однако соотношение размеров этих подкластеров в данном случае равно примерно 1 : 6,5 против 1 : 3 ранее.

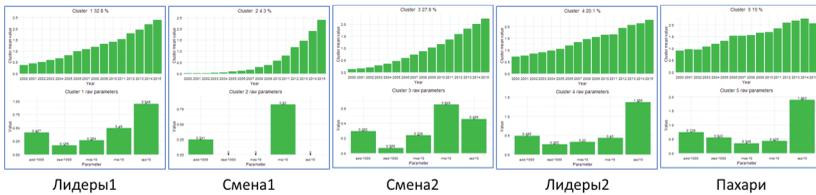


Рис. 16. «Математики+»: пять кластеров без к.п.

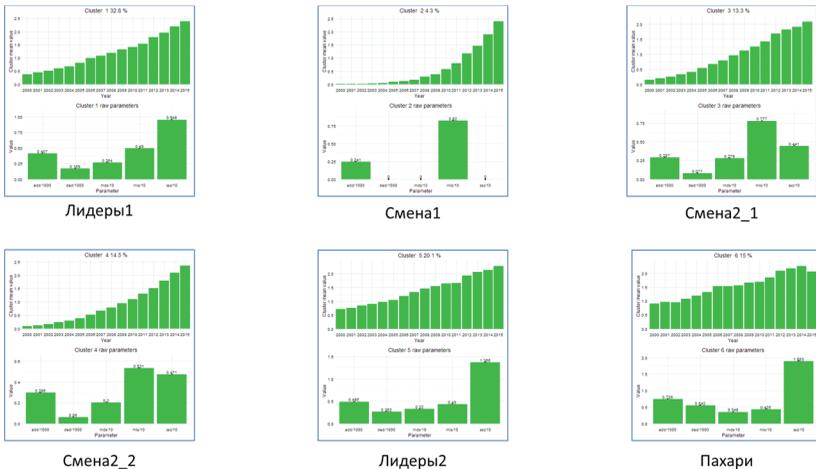


Рис. 17. «Математики+»: шесть кластеров без к.п.

Далее происходит разделение кластера «лидеров» как обычно: на более молодых с быстрым ростом и более опытных с умеренным ростом (Рис. 16). Другое заметное отличие образовавшихся подкластеров – в разнице средних показателей ssc: он в 1,4 раза выше для группы, которая в 1,6 раза меньше.

Наконец, большой подкластер кластера «смена» разделяется почти пополам (Рис. 17). В чуть большей подгруппе рост более быстрый и соответствующая функция более вогнута. Цитируемость в этой подгруппе до 2013 г. была ниже, а теперь выше.

Подробный анализ полученных результатов проводится в разделе 7.

6.6. КЛАСТЕРИЗАЦИЯ МНОЖЕСТВА «ФИЗИКИ»

Исходные кластеры физиков (Рис. 18) «пахари», «лидеры» и «смена», полученные при кластеризации по полному набору показателей, составляют соответственно 48,4%, 19,8% и 31,6%, что достаточно близко к значениям для исходного множества математиков (52%, 22,2%, 25,8%). Характеристики групп тоже вполне сравнимы с полученными для математиков.

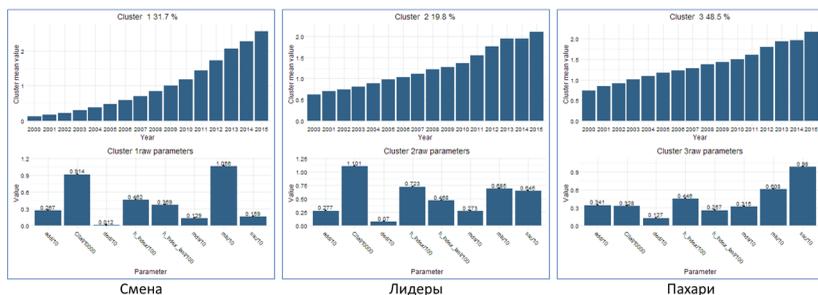


Рис. 18. «Физики»: три кластера.

При кластеризации без учета кумулятивных показателей (Рис. 19) результат также структурно похож на полученный для математиков, но если для последних соотношение численностей было 1 : 4 : 1, то для физиков – 1 : 1,7 : 1, иными словами, среди физиков, кластеризуемых структурно (без к.п.), «пахарей» и «молодых» больше, чем среди математиков за счет меньшего количества «лидеров». Это подтверждает представление о физике как о более «коллективной» науке, чем математика.

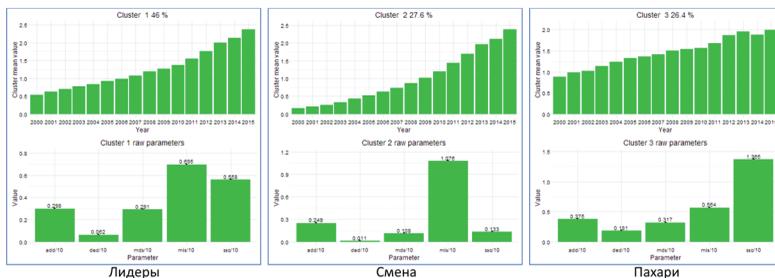


Рис. 19. «Физики»: три кластера без к.п.

6.7. РАЗДЕЛЕНИЕ КЛАСТЕРОВ ФИЗИКОВ

Несмотря на разницу в пропорциях кластеров, «разделение» кластеров физиков происходит в целом аналогично случаю математиков.

Первым разделяется кластер «пахарей» (Рис. 20): на устойчиво наращивающих годовую цитируемость и «насыщающихся». Последних в 2,5 раза меньше.

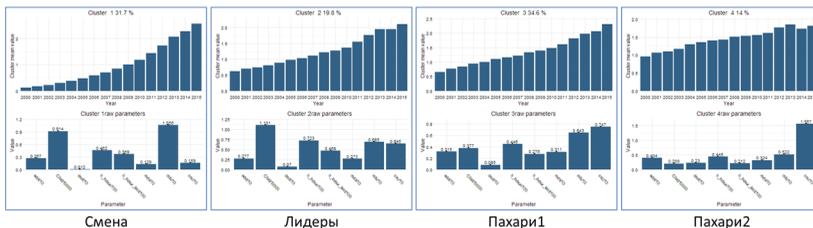


Рис. 20. «Физики»: четыре кластера.

Далее кластер «смена» распадается (Рис. 21) на тех, чей профиль цитируемости растет медленнее и даже демонстрирует признаки насыщения и ученых, чей профиль продолжает расти по вогнутому закону; первых в 1,5 раза больше.

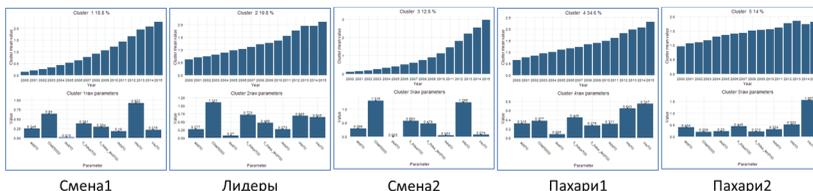


Рис. 21. «Физики»: пять кластеров.

Наконец, из кластера «лидеров» выделяется 4,3% сравнительно молодых, добившихся быстрого роста цитируемости с 2010 г. и тех, у кого цитируемость в последние годы растет значительно медленнее, порой с насыщением (Рис. 22). Последних в 3,6 раза больше.

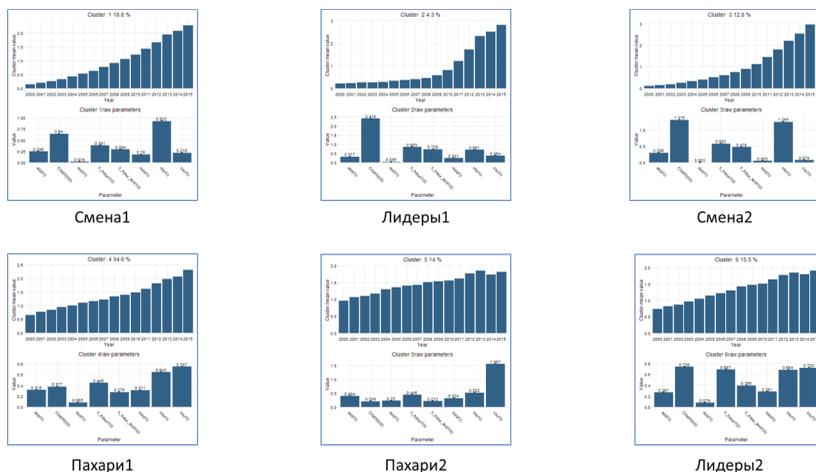


Рис. 22. «Физики»: шесть кластеров.

При кластеризации **без кумулятивных показателей** сначала, как и в случае математиков, делится кластер «смена» (Рис. 23): на 11,5% (от общего множества) тех, у кого цитируемость в среднем вогнуто растет практически без убывания и 16,1% тех, у кого после 2012 г. – явное насыщение, хотя убывание случается редко.

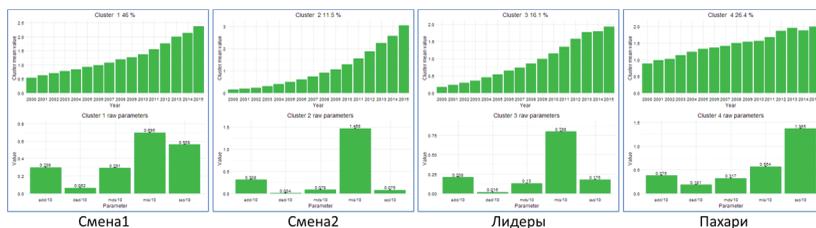


Рис. 23. «Физики»: четыре кластера без к.п.

Далее разделяется кластер «лидеров» (Рис. 24): на тех, для кого после 2010 года наблюдаются признаки насыщения и тех,

для кого рост цитируемости после 2010 г. имеет высокий темп. Соотношение численности этих групп: 1,8 : 1.

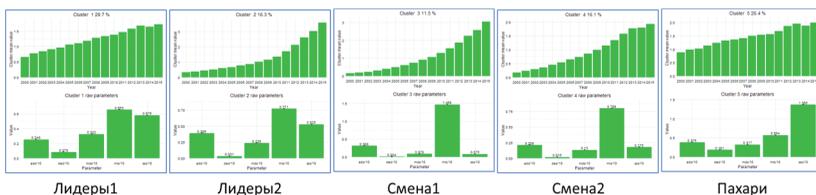


Рис. 24. «Физики»: пять кластеров без к.п.

Наконец, кластер «пахари» делится в соотношении 2 : 1 на тех, чей профиль цитируемости в среднем сохраняет линейность роста на протяжении исследуемых 15 лет и ученых, чей профиль показывает тенденцию к снижению (Рис. 25). Анализ – в разделе 7.

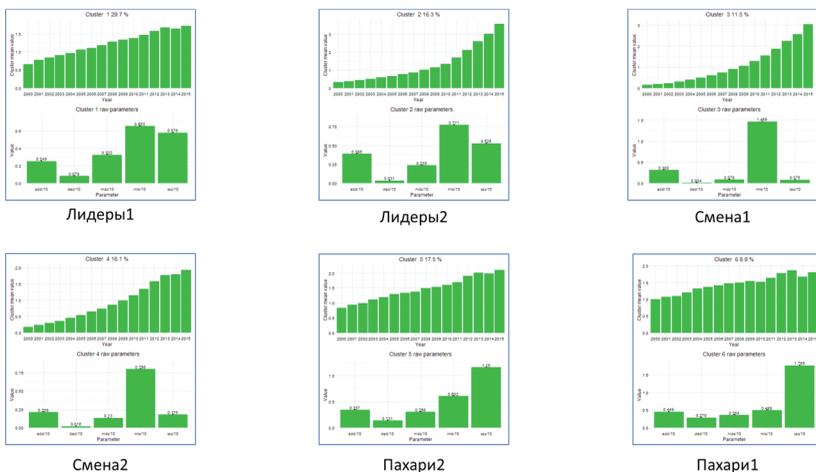


Рис. 25. «Физики»: шесть кластеров без к.п.

6.8. КЛАСТЕРИЗАЦИЯ ПСИХОЛОГОВ

При иерархической кластеризации ученых с тегом «Psychology» на первом этапе наблюдаем выделение трех типов авторов, описанных выше: «пахари» (47,7% / 37,2%), «лидеры» (34% / 42,4%) и «смена» (18,3% / 20,3%) – в скобках сначала приведены данные кластеризации с к.п., потом – без них (Рис.

26, Рис. 27). Кластер «смена» в данном случае выделяется необычайно ярко: нулевыми значениями показателей убывания цитирования – как в присутствии, так и в отсутствие к.п. Размер этого кластера во втором случае лишь незначительно больше. А вот кластер «лидеры» оказывается без к.п. больше на четверть.

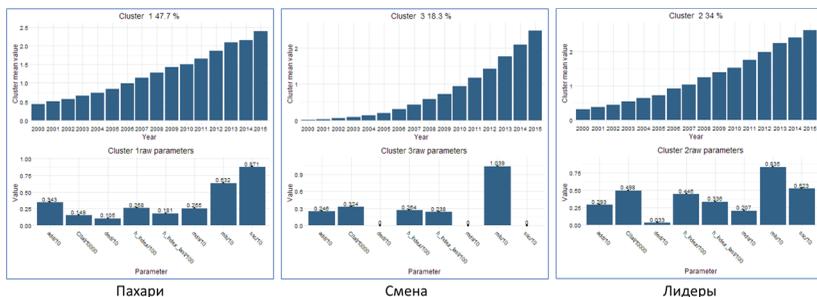


Рис. 26. «Психологи»: три кластера.

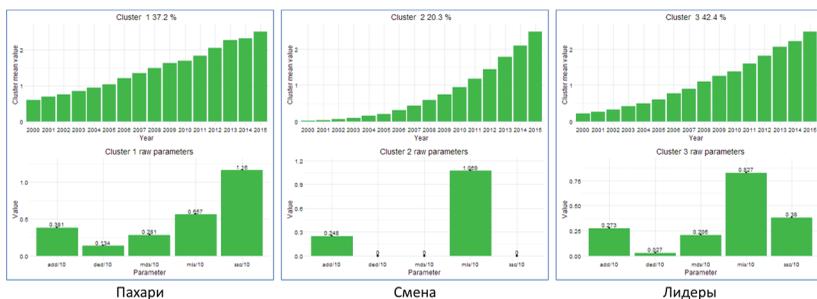


Рис. 27. «Психологи»: три кластера без к.п.

Данное отличие может быть объяснено тем, что выбранные кумулятивные показатели помогают точнее выделить опытных ученых с нелинейным (вогнутым) ростом цитируемости в последние 10 лет (к.п. проясняют, насколько давно автор стал известен и как изменилась его «плотность цитируемости» в последний период). На основе этой информации в кластер «лидеры» включаются лишь наиболее успешные, а не достигшие соответствующего порога попадают в кластер «пахари».

Кластер «смена» здесь соответствует группе «акселераты» у математиков: в него входят молодые ученые с вогнутым ростом

годовой цитируемости, не знакомые с ее убыванием. Но акселераты-психологи оказываются старше акселератов-математиков: первых начинают «точно» цитировать в 2001-2002 г., а последних – только в 2004 г. Это подтверждает репутацию математики как области, где успеха часто добиваются уже в юности. В кластере «смена» $h_{\text{index}_{\text{last}}}/h_{\text{index}} = 0,9$ (у «акселератов»-математиков это значение равно 0,92).

Достигнутый в 2015 г. уровень цитирования «смены» обгоняет уровень «пахарей» и приближается к уровню «лидеров».

6.9. РАЗДЕЛЕНИЕ КЛАСТЕРОВ ПСИХОЛОГОВ

При кластеризации по полному набору показателей первым разделяется самый большой кластер «пахари»: на треть самых опытных и две трети более молодых (Рис. 28). У последних гораздо меньше случаев убывания цитируемости.

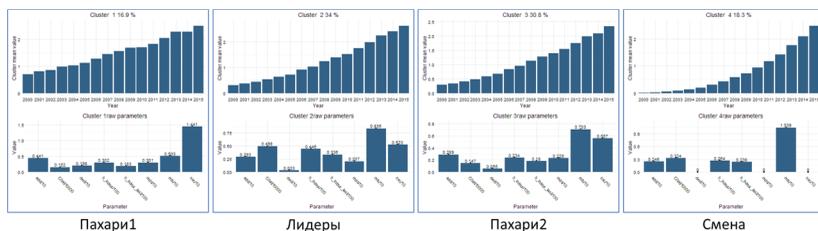


Рис. 28. «Психологи»: четыре кластера.

Затем делится кластер «лидеры» – почти пополам: на более молодых и более опытных (Рис. 29). В остальном больших отличий между ними нет.

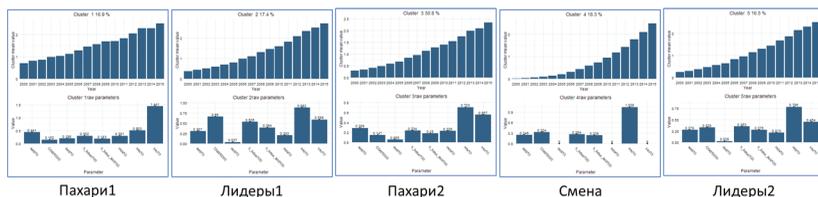


Рис. 29. «Психологи»: пять кластеров.

На третьем шаге разделяется подкластер «опытных лидеров»

(Рис. 30): из него выделяется треть очень опытных ученых, чья годовая цитируемость в последние годы уже почти не растет. После их вычленения у оставшихся двух третей наблюдается вогнутый рост цитируемости с малым числом случаев убывания.

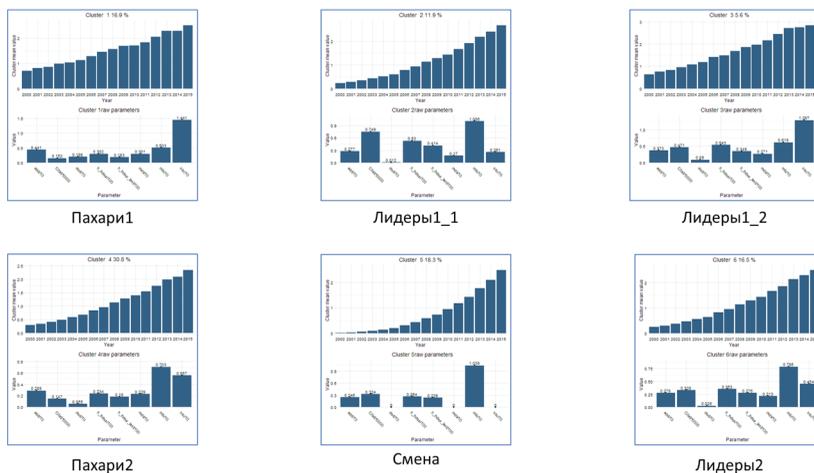


Рис. 30. «Психологи»: шесть кластеров.

Таким образом, в отличие от предыдущих результатов кластеризации, группа «смена» в случае психологов остается цельной после трех шагов разделения кластеров.

Без кумулятивных показателей первым разделяется кластер «пахари» (Рис. 31): из него выделяется одна шестая часть очень опытных авторов, у которых в 2009-2010 гг. заметно насыщение цитируемости, а позже вновь восстанавливается рост.

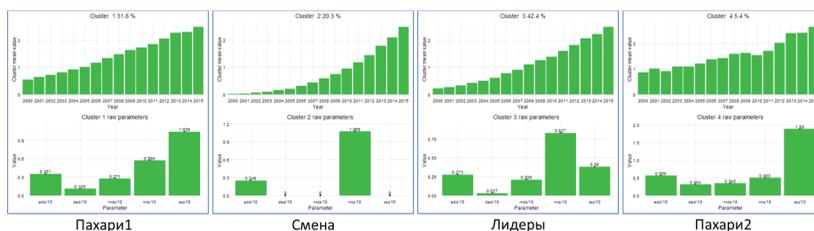


Рис. 31. «Психологи»: четыре кластера без к.п.

После ее вычленения у оставшихся 5/6 в среднем наблюдается линейный рост цитируемости с различной тенденцией к насыщению в конце.

Затем кластер «смена», куда входят молодые авторы, не знакомые с убыванием цитируемости, делится (Рис. 32) на схожие по форме профиля подкластеры (в соотношении 1 : 2), где в первом стаж работы выше на 60%, а цитируемость – на треть.

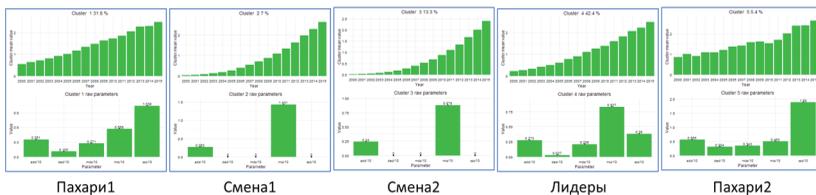


Рис. 32. «Психологи»: пять кластеров без к.п.

Наконец, последним почти поровну делится кластер «лидеры» (Рис. 33). Чуть больший подкластер показывает вогнутый рост цитируемости, чуть меньший – почти линейный рост.

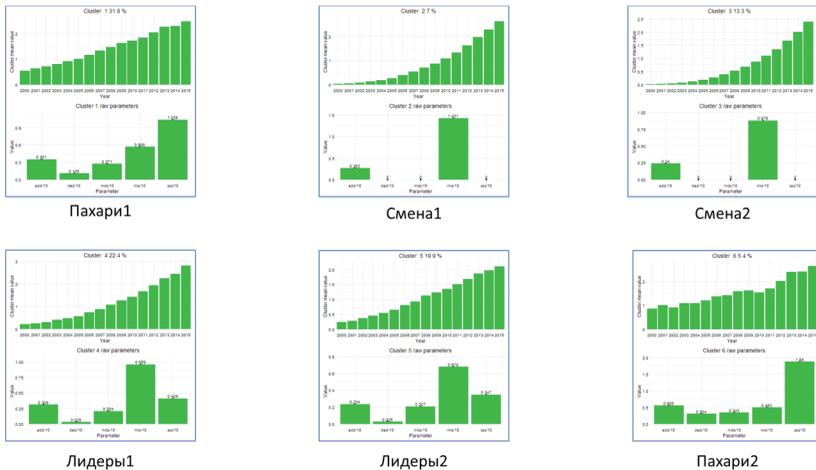


Рис. 33. «Психологи»: шесть кластеров без к.п.

7. Анализ результатов

Анализу результатов предположем табличное описание исходной типизации ученых, которая далее будет уточнена.

7.1. РАЗДЕЛЕНИЕ НА ТРИ КЛАСТЕРА: ТАБЛИЧНОЕ ПРЕДСТАВЛЕНИЕ

В Таблица 1 и Таблица 2 представлены разделения рассмотренных множеств ученых на три кластера и приведены средние значения ряда численных показателей по кластерам.

Таблица 1. Результаты разделения на три кластера с кумулятивными показателями (рисунки Рис. 2, Рис. 10, Рис. 18, Рис. 26).

Кластеры: условные названия и размеры	Фактор научного стажа	Динамика годовой цитируемости		Колебания цитируемости		
		$\frac{h_{index\ last}}{h_{index}}$	$Citat_{last}$	ssc	$\frac{ded}{add}$	$\frac{mds}{mis}$
«Математики» (500 ученых)						
«Пахари» 52%	Давнее начало (в 2000 г. уже довольно высокая цитируемость)	Рост годовой цитируемости в среднем линейный и достаточно медленный, у многих с насыщением		Убывание годовой цитируемости встречается нередко и имеет заметную амплитуду		
		0,62	740	13,25	0,5	0,68
«Лидеры» 22,2%	Давнее начало, заметная цитируемость в 2000 г.	Рост цитируемости быстрее, чем у «пахарей» и нелинейный – вогнутый		Убывание – реже и имеет существенно меньшую амплитуду, чем у «пахарей»		
		0,67	2490	11,55	0,26	0,47
«Смена» 25,8%	В 2000 г. цитируемость низкая либо нулевая	Рост годовой цитируемости быстрый и вогнутый		Убывание встречается редко и несравнимо с возрастом		
		0,81	1450	4,78	0,14	0,26
«Математики+» (543 ученых)						
«Пахари» 36,5%	Довольно высокая	Рост цитируемости близкий к линейному, с насыщением в конце		Убывания чаще и интенсивнее, чем для «пахарей» в множестве		

	цитируе- мость в 2000 г.			«Математики»		
		0,62	360	14,54	0,62	0,70
«Лидеры» 43,3%	Средняя цитируе- мость в 2000 г.	Рост вогнутый		Убывания чуть больше по относительной амплитуде и чуть менее продолжительны, чем для «лидеров» в множестве «Математики»		
		0,72	860	8,53	0,30	0,43
«Смена» 20,1%	Низкая либо нуле- вая цитиру- емость в 2000 г.	Рост вогнутый, быст- рее, чем у «лидеров»		Редкие колебания; убывания по размаху и времени несколько больше, чем для «смены» в множестве «Математики»		
		0,83	500	4,7	0,26	0,41
«Физики» (515 ученых)						
«Пахари» 48,4%	Давнее начало, много ссы- лок в 2000 г.	Рост линейный; в 2014 г. признаки насыщения		Довольно частые коле- бания с существенной амплитудой		
		0,58	3280	9,8	0,34	0,52
«Лидеры» 19,8%	В 2000 г. немалое число ссы- лок	До последних лет рост вогнутый		Колебания реже и с меньшей амплитудой, чем у «пахарей»		
		0,65	11010	6,45	0,22	0,399
«Смена» 31,6%	В 2000 г. ссылок очень мало	Вогнутый рост годо- вой цитируемости		Редкие колебания, убывание несущественно		
		0,799	9140	1,59	0,04	0,12
«Психологи» (556 ученых)						
«Пахари» 47,6%	Много ссы- лок в 2000 г.	Почти линейный рост, признаки насыщения в 2014 г.		Колебания реже и сла- бее, чем у «пахарей» в точных науках		
		0,70	1490	8,71	0,29	0,40
«Лидеры» 33,9%	В 2000 г. цитируе- мость ниже, чем у «пахарей»	Вогнутый рост		Колебания реже и сла- бее, чем у «пахарей»; убывания меньше, чем у «смены» в множестве «Математики»		

		0,75	4980	5,23	0,09	0,25
«Смена» 18,3%	В 2000 г. цитируемость ничтожная	Быстрый вогнутый рост, много цитат в поздние годы		Строго возрастающее годовое цитирование		
		0,90	3240	0	0	0

Таблица 2. Результаты разделения на три кластера без кумулятивных показателей (рисунки Рис. 6, Рис. 11, Рис. 19, Рис. 27).

Кластеры: условные названия и размеры	Научный стаж и динамика	Колебания цитируемости		
		ssc	$\frac{ded}{add}$	$\frac{mds}{mis}$
«Математики»				
«Пахари» 17,2%	Много ссылок в 2000 г., насыщение в 2007-11, затем скачок и спад	Частые колебания, убывания сравнимы с приращениями		
		18,36	0,66	0,77
«Лидеры» 66,4%	Достаточно известны в 2000 г., далее вогнутый рост	Колебания реже, приращения доминируют над убываниями		
		10,44	0,34	0,55
«Смена» 16,4%	Низкая или нулевая цитируемость в 2000 г., затем быстрый вогнутый рост	Редкие колебания, убывания незначительны		
		3,66	0,08	0,18
«Математики+»				
«Пахари» 15%	Высокий уровень цитирований в 2000 г.; рост годовой цитируемости с насыщениями и убыванием в конце	Частые колебания, убывания почти не отстают от приращений		
		18,83	0,71	0,82
«Лидеры» 52,8%	Заметный уровень цитирований в 2000 г.; рост слабо-вогнутый	Колебания реже, приращения вдвое превосходят убывания		
		11,07	0,44	0,61
«Смена» 32%	Очень низкая цитируемость в 2000 году; быстрый вогнутый рост	Редкие колебания, убывания в профилях годовой цитируемости незначительны		

		3,96	0,18	0,31
«Физики»				
«Пахари» 26,4%	Много цитирований в 2000 г., далее медленный практически линейный рост	Довольно частые колебания, убывания лишь вдвое отстают от приращений		
		13,65	0,46	0,56
«Лидеры» 45,9%	Заметная цитируемость в 2000 г., слабо-вогнутый рост, приближаемый сплайном двух линейных участков	Редкие колебания, убывания малозначимы		
		5,58	0,19	0,42
«Смена» 27,5%	Низкая цитируемость в 2000 г., затем вогнутый рост	Колебания пренебрежимо малы		
		1,3	0,05	0,1
«Психологи»				
«Пахари» 37,2%	Высокая цитируемость в 2000 г., затем линейный рост	Приращения доминируют над убываниями; показатели колебаний примерно на уровне «лидеров» в множестве «Математики»		
		11,6	0,31	0,50
«Лидеры» 42,4%	В среднем невысокая цитируемость в 2000 г., затем слабо-вогнутый рост	Редкие колебания, убывания незначительны; показатели – на уровне групп «смена» в множествах математиков		
		3,8	0,09	0,25
«Смена» 20,3%	Крайне низкая цитируемость в 2000 г., быстрый вогнутый рост	Участков убывания годовой цитируемости нет		
		0	0	0

Дальнейший анализ позволит уточнить эти кластеризации с помощью ранее полученных разделений ученых на 4-6 групп.

7.2. «МАТЕМАТИКИ» И «МАТЕМАТИКИ+»

При кластеризации математиков на шесть групп во всех случаях выделяется небольшая группа «самых молодых и взрывных». Без претензии на буквальность она была названа группой «акселератов». Их начинают цитировать не ранее 2004 г., но благодаря нелинейному (вогнутому) росту цитируемости они быстро догоняют по этому показателю значительно более опытных ученых. Им незнакомо убывание цитируемости. В исходном множестве математиков размер этой группы 4,2% при кластеризации с кумулятивными показателями и 4,4% при кластеризации без них. В множестве «Математики+» – соответственно 3,9% и 4,3%. Наличие кумулятивных показателей «отрывает» от этой группы некоторых авторов, чье текущее значение цитируемости ниже, чем у остальных.

Наиболее удаленный от «акселератов» кластер – очень опытные авторы, которых активно цитировали еще в 2000 г., затем их годовая цитируемость линейно и не слишком быстро росла, но в последние годы у них наметилось насыщение или снижение этого показателя. Эта группа была условно названа «инерционные». Она составляет 14,6% (17,2%) от исходного множества математиков и 13,1% (15%) от множества «Математики+» где первое число относится к кластеризации с кумулятивными показателями, а второе – к кластеризации без них. Мы видим, что данная крайняя группа, как и предыдущая, при отсутствии кумулятивных показателей (к.п.) показывает бóльшую численность, однако эта разница не принципиальна. Последнее свидетельствует о том, что границы этих групп достаточно отчетливы: в «буферные зоны» попадают немногие.

Оставшиеся четыре кластера объединяют ученых с устойчивым, но не самым быстрым ростом цитируемости. В каждом случае можно выделить группу довольно молодых (впрочем, часть из них цитировалась уже в 2000 г.), но не самых «взрывных». Убывание цитируемости у них случается редко. Данные по численности групп будем далее приводить в формате («Математики» с к.п. / «Математики» без к.п. / «Математики+» с к.п. / «Математики+» без к.п.): (27,8% / 28,6% / 30,8% / 27,8%). Эта группа (назовем ее «молодые»), как показывают приведенные цифры, достаточно устойчива, и она всегда делится на две подгруппы А

и В, характеризующиеся: (А) меньшим опытом и более быстрым и вогнутым ростом цитируемости; (В) более значительным стажем и более медленным ростом цитируемости. Границы этих подгрупп более размыты, чем границы объединяющей группы. Численности их приведем в том же формате, где первое слагаемое – размер подгруппы А: (21,6% + 6,2% / 16,6% + 12% / 20,1% + 10,7% / 14,5% + 13,3%). Легко заметить, что при учете к.п. более динамичная подгруппа А больше (21,6%, 20,1%) против соответственно 16,6% и 14,5% без учета к.п.

Наконец, последняя группа объединяет опытных ученых с достаточно устойчивым ростом цитируемости. Назовем их «корифеями». Численность ее в указанном выше формате составляет (53,4% / 49,8% / 51,9% / 52,8%). Как и предыдущая, эта группа делится на подгруппы А (более динамичная) и В, границы которых менее четки, чем границы группы. Их численности: (А + В) = (16% + 37,4% / 12% + 37,8% / 28,6% + 23,3% / 32,7% + 20,1%). Здесь впервые заметно существенное отличие множеств «Математики» и «Математики+». Если в первом компактно отделяется небольшая подгруппа А более динамичных ученых (16%, 12%), то во втором она насчитывает более половины группы «корифеев».

Главный вывод: в обоих множествах каждым из методов кластеризации выделяются четыре группы с достаточно четкими естественными границами и очень небольшими транзитными зонами: «акселераты», «молодые», «корифеи» и «инерционные». Их приблизительные размеры составляют соответственно 4%, 29%, 52% и 15% и практически не зависят ни от выбранного множества, ни от способа кластеризации. Они получаются группировкой подкластеров исходных трех групп, которые были названы «лидеры», «пахари» и «смена». Исходные группы также представляют интерес, но, по-видимому, они менее «сущностны»: их относительные размеры и границы довольно сильно отличаются для множеств математиков и разных способов кластеризации.

7.3. «ФИЗИКИ»

Как уже было отмечено, кластеризация физиков имеет много общего с кластеризацией математиков.

Рассмотрим аналоги групп «акселераты», «молодые», «корифеи» и «инерционные», которые в случае математиков имеют весьма четкие границы.

При кластеризации на шесть групп по полному набору показателей выделяются 14% «инерционных» (у математиков их было от 13 до 17 процентов). Их характеризует высокая цитируемость еще в 2000 г., затем медленный линейный рост и убывание цитируемости в последние два года рассматриваемого периода.

Группа «корифеи» составляет 50% (у математиков в среднем 52%) и состоит из подгрупп А (более динамичные авторы) – их 34,5% и В – 15,5%.

Группа «молодые» составляет 31,6% (у математиков – от 27,8% до 30,8%) и состоит из подгрупп А (более динамичные) и В численностью соответственно 12,8% и 18,8%. В отличие от математиков, подгруппа А меньше подгруппы В.

В то же время мы не видим здесь явной группы «молодых и взрывных», имевших нулевую цитируемость в начале 2000-х и совершенно не знакомых с убыванием цитируемости. Наименьшие средние показатели убывания имеет подкластер «смена-2», составляющий 12,8% (выше он упомянут как подгруппа А группы «молодые»). Кроме того, выделяется не встречавшийся ранее подкластер в 4,3% (размер, характерный для математиков-«акселератов»), состоящий из ученых, имевших заметную, но довольно низкую и почти не растущую годовую цитируемость до середины 2000-х, а с тех пор показывающих быстрый, хотя и замедляющийся рост. Наличие этого кластера гипотетически может быть объяснено следующим отличием физики от математики. В физике меньше ярких одиночек, которые могут прославиться 2-3 достижениями. Работа здесь чаще выполняется командно. Авторы, вошедшие в указанный подкластер, не очень молоды, они добились определенных результатов уже к началу 2000-х. После этого им посчастливилось войти в сильные команды, и дела у них «пошли в гору». Разумеется, подтверждение этой гипотезы требует дополнительных исследований.

Таким образом, главный вывод из полученной кластеризации физиков состоит в следующем. Среди физиков, как и среди математиков, выделяются группы «молодые» (32%), «корифеи» (50%) и «инерционные» (14%). В множестве математиков средние доли

этих групп были соответственно 29%, 52% и 15%. Группа «акселераты» (самые молодые, но уже добившиеся больших успехов) среди физиков не выделяется (вероятно, она «растворена» в группе «молодые»-А и может проявиться при росте числа кластеров). Вместо них выделяется группа не очень молодых авторов, имевших заметную, но невысокую и почти не растущую цитируемость в начале 2000-х и быстрый, но замедляющийся рост цитируемости с конца 2000-х. Эта группа (которую можно назвать «попавшие в струю»), как и «акселераты»-математики, составляет 4%. Данная «подмена группы» может быть объяснена отличием физики от математики: в первой больше удельный вес командной работы, во второй – ярких одиночек.

Оценим теперь, меняется ли картина при кластеризации физиков без к.п. Здесь мы видим группу «молодые» (27,5%), состоящую из подгрупп А (более динамичные, с очень низкими показателями убывания) и В численностью соответственно 11,4% и 16,1%, что весьма похоже на кластеризацию без к.п. Но на этом сходство заканчивается. Далее мы видим подкластер «лидеры»-2, куда входят сравнительно молодые, но добившиеся максимального среди всех кластеров признания. Эта группа «молодых корифеев» объединяет части прежних групп «корифеи» и «молодые». Ее размер 16,3%. Далее идет группа «корифеи» (47,1%) с подгруппами «корифеи-А» (29,7%) и «корифеи-В» (17,4%). Первая из них характеризуется более устойчивым ростом цитируемости (показатель ssc у нее вдвое ниже), у второй же выше средняя цитируемость на протяжении всего периода наблюдений 2000-2015 гг. В последние годы в подгруппе «корифеи-В» видны признаки насыщения. Наконец, оставшаяся группа – это «инерционные» (8,9%). Тем самым в данной кластеризации наблюдается сдвиг границ групп: между группами «молодые» и «корифеи» вклинились «молодые корифеи» в результате чего «корифеи», почти сохранив размер группы (47,1%), включили в себя часть авторов с насыщением цитируемости, в результате чего группа «инерционные» уменьшилась в 1,6 раза.

7.4. «ПСИХОЛОГИ»

В множестве «Психологи» выделяется отчетливый кластер «смена», аналогичный группе «акселераты» у математиков. Это молодые ученые, имеющие в среднем вогнутый рост годовой цитируемости и не знакомые с ее убыванием. Но если у математиков эта группа составляла лишь 4%, то у психологов в нее с/без к.п. входит 18,3% / 20,3%. Ее участники имеют в среднем больший научный стаж, чем «акселераты»-математики.

При кластеризации без к.п. эта группа делится на две подгруппы: более молодую (13,3%) и более опытную с более высокой цитируемостью (7%). При кластеризации с к.п. после трех «разделений» кластеров она остается цельной, но выделяется объемная группа более или менее молодых ученых, которую можно назвать «в расцвете сил» (59%). По успешности она делится на три подгруппы А (11,8%), В (16,5%) и С (30,7%): отношение их среднего индекса Хирша – как 2 : 1,3 : 1, отношение числа цитирований за последние 5 лет – как 5 : 2,2 : 1. Отношение значений показателя ssc (количество смен знака годовых приращений цитируемости) – как 1 : 1,7 : 2,1. При кластеризации без к.п. этой группе соответствует кластер «лидеры» (42,4%), имеющий две подгруппы: А (22,4%) и В (19,9%). Для первой характерен вогнутый рост годовой цитируемости, для второй – линейный. Их средние показатели add относятся как 3 : 2.

Психологи, у которых была довольно высокая цитируемость еще в 2000 г., образуют группу «опытные», которая с/без к.п. составляет 22,5% / 37,2%. Она разделяется на подгруппы А (5,6% / 31,8%) и В (16,9% / 5,4%). Отношения показателей add и ssc для этих подгрупп составляют соответственно 2,3 : 1 / 1,2 : 1 и 1 : 1,1 / 1 : 1,8.

В множестве психологов не выделяется группа «инерционные», где средняя цитируемость заметно снижалась бы в последние годы. Такие ученые есть, но они «спрятаны» в кластерах, где в той или иной мере проявляется тенденция к насыщению. При кластеризации с к.п. кластер такого рода имеет размер 5,6%, без к.п. 5,4%, но в последнем заметен рост в 2015 г.

Отметим, что после выделения кластеров любой ранее не учтенный ученый может быть отнесен к одной из групп методом ближайших соседей или иным методом классификации.

8. Общие выводы

Результаты кластеризации рассмотренных множеств ученых обсуждались в разделе 7. Здесь приведем наиболее общие выводы.

Иерархическая кластеризация математиков (два множества, два набора параметров кластеризации) показала наличие четырех устойчивых групп: «акселераты» (4%), «молодые» (29%), «корифеи» (52%) и «инерционные» (15%) – указаны средние размеры, разброс относительно этих средних значений невелик.

При кластеризации с к.п. среди физиков, как и среди математиков, выделяются группы «молодые» (31,7%), «корифеи» (48,5%) и «инерционные» (19,8%) – относительные численности их примерно такие же, как у математиков. Вместо группы «акселераты» выделяется группа «попавшие в струю» размером также 4%.

При кластеризации физиков без к.п. выделяются группы: «молодые» (27,6%) – состоит из двух подгрупп разной успешности; «молодые корифеи» (16,3%); «корифеи» (47,2%) – состоит из двух подгрупп разной успешности; «инерционные» (8,9%).

Психология – наука более гуманитарная, и для нее результат кластеризации несколько иной. Здесь выделяется группа «смена-акселераты» (18% / 20% при кластеризации с / без к.п.), аналогичная группе акселератов-математиков, но не такая молодая. Кроме того, выделяется группа «в расцвете сил» (59% / 42,4%), разделяющаяся на три / две подгруппы разной успешности и группа «опытные» (22,5% / 37,2%), разделяющаяся на две подгруппы разной успешности. И здесь случай, когда границы групп существенно зависят от того, используются ли кумулятивные показатели. В множестве психологов не выделяется группа «инерционные», где средняя годовая цитируемость заметно снижалась бы в последние годы. Вообще, рост цитируемости показывает в психологии большую стабильность, чем в математике и физике.

В целом математика во многом похожа на физику (другую точную науку): также выделяются группы «молодые», «корифеи» и «инерционные» с примерно одинаковой численностью. Но наличием группы «акселераты» она похожа на психологию. Действительно, математика, согласно одной из точек зрения, есть наука, занимающая промежуточное положение между естественными и гуманитарными дисциплинами.

9. Заключение

В статье предложен набор библиометрических показателей для типизации успешных ученых посредством иерархической кластеризации. Набор включает стандартные индексы цитирования и пять структурных показателей, характеризующих динамику цитируемости ученого.

Годовое число ссылок есть разностное приближение первой производной зависимости общего числа ссылок от времени. Используемые структурные показатели определяются в терминах приращений годовой цитируемости, т.е. второй производной числа ссылок, характеризующей выпуклость этой функции.

Особенность структурных показателей – их масштабируемость: пропорциональное изменение цитируемости не меняет их величины. Таким образом, стандартные индексы цитируемости выражают масштаб, структурные показатели – особенности роста числа ссылок.

Один из результатов работы состоит в том, что кластеризации, полученные с использованием индексов масштаба и без них часто дают близкие результаты. Это не может быть объяснено однородностью выборки по индексам масштаба: средняя цитируемость в некоторых кластерах отличается почти на порядок.

Другой результат: группы, слабо чувствительные к способу кластеризации (и отчасти к выборке), могут формироваться объединением кластеров более дробной кластеризации. Для двух множеств математиков эти группы были названы «акселераты», «молодые», «корифеи» и «инерционные», для физиков – «молодые», «корифеи» и «инерционные» (причем доли этих групп для физиков и математиков близки), для психологов – «смена-акселераты», границы же групп «в расцвете сил» и «опытные» при откате от индексов масштаба меняются. Для физиков при учете индексов масштаба выделяется группа «попавшие в струю», а без их учета – «молодые корифеи».

Исследование показало, что методы кластерного анализа, примененные к библиометрическим данным, помогают не только при решении задачи типизации ученых, но и при исследовании отличий между научными дисциплинами.

Литература

1. БОРОВСКИЙ А. *Основные библиометрические показатели для оценки эффективности научной работы.* – Пермь: Изд-во Пермского национального исследовательского политехнического университета, 2012.
2. ВОРОНЦОВ К.В. *Алгоритмы кластеризации и многомерного шкалирования.* – М.: МГУ, 2007.
3. МИРКИН Б.Г. *О понятии научного вклада и его измерителях // Управление большими системами.* – 2013. – Т. 44 – С. 292–307.
4. МИРКИН Б., ОРЛОВ М. *Методы многокритериальной стратификации и их экспериментальное сравнение / Препринт WP7/2013/06.* – М.: ВШЭ, 2013.
5. ЧЕБОТАРЕВ П.Ю. *Наукометрия: как с ее помощью лечить, а не калечить? // Управление большими системами.* – 2013. – Т. 44 – С. 14–31.
6. ЧЕБОТАРЕВ П.Ю. *Оценка ученых: пейзаж перед битвой // Управление большими системами.* – 2013. – Т. 44 – С. 506–537.
7. CRONIN B., SUGIMOTO C.R. (EDS.) *Scholarly Metrics under the Microscope: from Citation Analysis to Academic Auditing.* – Medford, NJ: ASIS&T, 2014.
8. DE BELLIS N. *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics.* – Lanham, MD: Scarecrow Press, 2009.
9. GOGGLOU A., SIDIROPOULOS A., KATSAROS D., MANOLOPOULOS Y. *A Scientist's impact over time: The predictive power of clustering with peers // Proceedings of the 20th International Database Engineering & Applications Symposium.* – ACM, 2016. – P. 334–339.
10. GOGGLOU A., SIDIROPOULOS A., KATSAROS D., MANOLOPOULOS Y. *The fractal dimension of a citation curve: quantifying an individual's scientific output using the geometry of the entire curve // Scientometrics.* – 2017. To appear. DOI: 10.1007/s11192-017-2285-2.

11. JAIN A.K., MURTY M.N., FLYNN P.J. *Data clustering: a review* // ACM Computing Surveys (CSUR). – 1999. – Vol. 31. – № 3. – P. 264–323.
12. MIRKIN B., ORLOV M. *Research impact: level of results, citation, merit* // Working paper WP7/2014/09. – Moscow: HSE, 2014.
13. OSBORNE F., PERONI S., MOTTA E. *Clustering citation distributions for semantic categorization and citation prediction* // Proceedings of the 4th International Conference on Linked Science – Making Sense Out of Data (LISC2014). Volume 1282. – CEUR-WS.org, 2014. – P. 24–35.
14. WARD, JR J.H. *Hierarchical grouping to optimize an objective function* // Journal of the American Statistical Association. – 1963. – Vol. 58. – № 301. – P. 236–244.

MAKING A TYPOLOGY OF SCIENTISTS ON THE BASIS OF BIBLIOMETRIC DATA

Ilya Vasilyev, Moscow Institute of Physics and Technology, Moscow, student (ilya.vasilev@phystech.edu).

Pavel Chebotarev, Institute of Control Sciences of RAS, Moscow, Doctor of Science, head of laboratory (Moscow, Profsoyuznaya st., 65, (495) 335-18-05, pavel4e@gmail.com).

Abstract: In this paper, we propose a set of indicators for solving the problem of differentiation and stratification of scientists on the basis of bibliometric data using cluster analysis. Ward's hierarchical clustering algorithm is applied to some sets of mathematicians, physicists, and psychologists with high citation rates. The analysis of the obtained results allows one not only to describe several stable types of scientists, but also to study the differences between distinct groups of scientists in various scientific disciplines.

Keywords: typology of scientists, scientometrics, bibliometrics, citation indices, cluster analysis, Google Scholar.

Статья представлена к публикации

членом редакционной коллегии ...заполняется редактором...

Поступила в редакцию ...заполняется редактором...

Опубликована ...заполняется редактором...