Исследование системы «автор – издатель»: подход на основе временных рядов

Ф.В.Краснов, А.В.Дименто́в, М.Е.Шварцман

26 сентября 2019 г.

Аннотация

Один и тот же автор можем быть первым в списке авторов одной статьи и последним в другой. Порядковый номер автора изменяется со временем и отражает вклад автора в исследование. В данном исследовании изучен феномен изменения порядкового номера авторов от статьи к статье. Для этого предложена новая методика на основе анализа и сравнения временных рядов. Основным результатом исследования является методика выявления кластеров авторов характеризующих их научную карьеру и издательскую политику журнала.

Ключевые слова: соавторства, иерархический кластерный анализ, динамическое программирование, поиск экстремальных значений.

1 Введение

В данном исследовании был создан подход к выявлению скрытых структур в развитии публикационной активности авторов журнала. Для редакции представляет интерес кто из авторов составляет основу журнала – публикуется регулярно и пишет на характерные для журнала темы. Таких авторов можно условно назвать *ядром*. Выявить такое *ядро авторов* представляется достаточно сложной задачей, так как непонятны критерии отнесения отдельного автора к такому ядру. Сопутствующей задачей является выделение авторов с нестандартной публикационной активностью. Например, авторов пишущих каждый раз в разных исследовательских коллективах.

Для проведения анализа нужна количественная характеристика публикационной активности автора. Такая характеристика автора изменя-

ется со временем и может быть проанализирована с помощью аппарата анализа временных рядов и методов выявления аномалий.

Сами журналы так же изменяются – приходят и уходят главные редакторы, изменяется редакционная политика, появляются новые требования от ВАК и многое другое. Поэтому только рассмотрение поведения системы «автор – издатель» во времени позволяет делать выводы о таких изменениях.

Данная статья состоит из описания методов и результатов эксперимента.

2 Методы

В некоторых академических областях основные участники исследования располагаются первыми в списке авторов. Однако эта практика может легко привести к конфликту. Поэтому многие издатели рекомендуют перечислять авторов в алфавитном порядке, чтобы обеспечить равное упоминание всех авторов. Это обычная практика в некоторых академических областях, таких как математика или экономика. Поэтому для предлагаемого подхода к исследованию порядкового номера автора необходима проверка на «алфавитный порядок». Если в наборе данных выявлено использование «алфавитного порядка», то рассматриваемый в данном исследовании подход на основе временных рядов не применим.

Временной ряд – это собранный в разные моменты времени статистический материал о значении каких-либо параметров исследуемого процесса. В рассматриваемой системе «автор – издатель» процессом являются публикации (научные статьи), а параметром этого процесса порядковый номер автора в публикации.

Существует множество методик анализа публикационной активности на основании графов соавторства [1, 2, 3]. В этих исследованиях используется дополнительная информация об авторах из графов соавторства – центральность, клики, Pagerank и другие. В данном исследовании предлагается рассмотреть порядок авторов в статье, как источник дополнительной информации о развитии автора. Для этого вводится понятие *профиля автора*, как временного ряда с целью выявления его внутренних структур.

Профиль автора строится на основании последовательности порядковых номеров при соавторстве по каждой публикации. Например, такая последовательность для автора (автор$_1$) может выглядеть так:

$5, \ldots, 5, 4, \ldots, 4, 3, \ldots, 3, 2, \ldots, 2, 1, \ldots, 1.$

Это означает, что автор$_1$ был 5-м соавтором в нескольких статях,

потом стал публиковаться 4-м и так далее. Пример для трех авторов приведен на рисунке 1 слева.

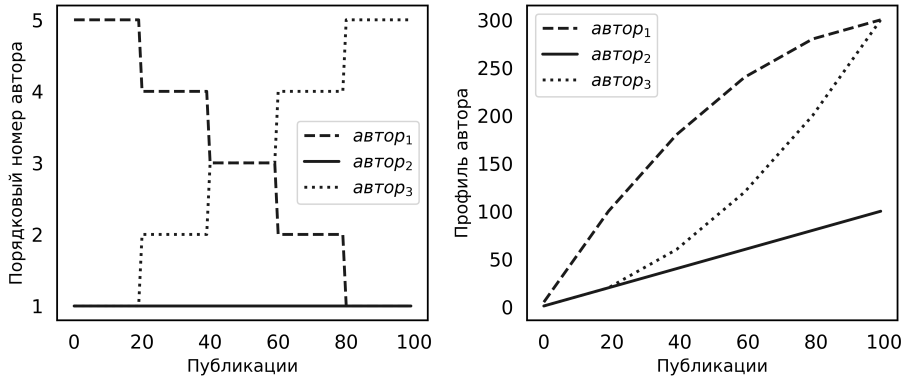

Рис. 1: Примеры профилей авторов

Профиль автора (рис. 1 справа) является интегральным показателем места автора в публикациях. В самом простом варианте профиль автора – это вектор в котором элементом с номером i является сумма порядковых номеров во всех предыдущих публикациях. У такого определения профиля автора есть интуитивно-понятное объяснение – действительно вклад автора в исследование может только накапливаться со временем.

Для выявления скрытых структур среди профилей авторов с помощью методов обучения без учителя необходимо определить расстояние между двумя профилями. Для вычисления отклонения бывает достаточно простого измерения расстояния между компонентами двух последовательностей (евклидово расстояние). Однако часто две последовательности имеют приблизительно одинаковые общие формы, но эти формы не выровнены по оси X. Чтобы определить подобие между такими последовательностями, мы должны «деформировать» ось времени одной (или обеих) последовательностей, чтобы достигнуть лучшего выравнивания. Алгоритм динамической трансформации временной шкалы (DTW-алгоритм, от англ. dynamic time warping) – это алгоритм, позволяющий найти оптимальное соответствие между временными последовательностями [4].

Рассмотрим два временных ряда – X длины n и Y длины m.

$$X = x_1, x_2, \ldots, x_i, \ldots, x_n \tag{1}$$

$$Y = y_1, y_2, \ldots, y_j, \ldots, y_m \tag{2}$$

Первый этап алгоритма состоит в следующем: Мы строим матрицу d порядка $n \times m$ (матрицу расстояний), в которой элемент $d_{i\,j}$ есть расстояние $d(x_i, y_j)$ между двумя точками x_i и y_j. Обычно используется евклидово расстояние: $d(x_i, y_j) = (x_i - y_j)^2$, или $d(x_i, y_j) = |x_i - y_j|$. Каждый элемент (i, j) матрицы соответствует выравниванию между точками x_i и y_j. На втором этапе строим матрицу трансформаций (деформаций) D, каждый элемент которой вычисляется исходя из следующего соотношения:

$$D_{i\,j} = d_{i\,j} + min(D_{i-1\,j}, D_{i-1\,j-1}, D_{i\,j-1}) \tag{3}$$

После заполнения матрицы трансформации, мы переходим к заключительному этапу, который заключается в том, чтобы построить некоторый оптимальный путь трансформации (деформации) и DTW расстояние. Путь трансформации W – это набор смежных элементов матрицы, который устанавливает соответствие между X и Y. Он представляет собой путь, который минимизирует общее расстояние между X и Y. k-ый элемент пути W определяется как $w_k = (i, j)_k, \quad d(w_k) = d(x_i, y_j) = (x_i - y_j)^2$.

Таким образом:

$$W = w_1, w_2, \ldots, w_k, \ldots, w_K \tag{4}$$

$$max(m, n) \leqslant K < m + n, \text{ где } K \text{ — длина пути.} \tag{5}$$

DTW расстояние между двумя последовательностями рассчитывается на основе оптимального пути трансформации с помощью формулы:

$$DTW(X, Y) = min \left\{ \frac{\sum_{k=1}^{K} d(w_k)}{K} \right\} \tag{6}$$

K в знаменателе используется для учета того, что пути трансформации могут быть различной длины. Важно отметить, что DTW не является метрикой, так как из $DTW(X, Y) = 0$ не следует, что $X = Y$.

На рисунке 2 приведен пример профилей авторов. Можно видеть, что авторы начинают публикации в разное время и на разных позициях. Таким образом имеет место быть та самая невыравненность по оси времени (абсциссе).

На рисунке 3 отображен пример матрицы расстояний между профилями авторов.

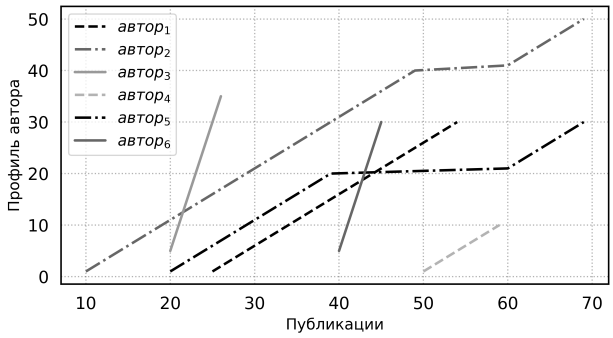

Рис. 2: Профили авторов

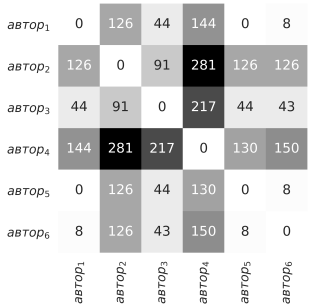

Рис. 3: Матрица расстояний DTW

Определив расстояния в пространстве профилей авторов, мы получили возможность выделить редкие профили авторов. Для этой цели будем использовать направление Кибернетики, названное «Поиск аномалий». Выделяют следующие типы методов поиска аномалий: методы машинного обучения, методы подмены задачи, метрические методы [5], модельные тесты, статистические тесты [6]. В методах статистических тестов выделяют направление связанное с поиском экстремальных значений (extreme value analysis,EVA) [7]. В рамках EVA мы будем отталкиваться от предположения, что распределение расстояний между профилями авторов носит нормальный характер и будем считать экстремальными значения, выходящие за пределы $\pm 3\sigma$.

Для решения задачи поиска структур в профилях авторов мы применили агломеративные методы из иерархического кластерного анализа. Одним из преимуществ иерархической кластеризации является то, что вам не нужно заранее знать количество кластеров в ваших данных. В частности, метод Уорда (англ. Ward's method) для оценки расстояний между кластерами использует дисперсионный анализ. В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центра кластера, получаемого в результате их объединения. Такой подход позволяет в дальнейшем оперировать усредненными профилями авторов для каждого кластера. Но с методической точки зрения мы должны оценить какая из стратегий сортировки для иерархического кластерного анализа будет давать оптимальный результат. Для этой оценки мы использовали коэффициент копенетической корреляции.

Предположим что набор данных X_i в процессе моделирования кластеров произвел следующую дендрограмму T_i – упрощенную модель, в которой «близкие» данные были сгруппированы в иерархическое дерево. Тогда можно определить следующие меры расстояния.

- $x(i,j) = |X_i - X_j|$, обычное евклидово расстояние между i-м и j-м наблюдениями.
- $t(i,j) =$ дендрограмматическое расстояние между точками модели T_i и T_j. Это расстояние - высота узла в дереве, на котором эти две точки первый раз соединяются вместе.

Тогда коэффициент копенетической корреляции будет вычисляться по следующей формуле (7):

$$c = \frac{\sum_{i<j}(x(i,j) - \bar{x})(t(i,j) - \bar{t})}{\sqrt{[\sum_{i<j}(x(i,j) - \bar{x})^2][\sum_{i<j}(t(i,j) - \bar{t})^2]}} \tag{7}$$

Коэффициент копенетической корреляции является мерой того, насколько точно дендрограмма сохраняет попарные расстояния между исходными немоделированными точками данных. Хотя он наиболее широко применяется в области биостатистики [8] (обычно для оценки кластерных моделей последовательностей ДНК или других таксономических моделей), его также можно использовать в других областях исследований [9], где необработанные данные имеют тенденцию встречаться в виде кластеров. Этот коэффициент также был предложен для использования в качестве теста для вложенных кластеров.

3 Результаты

3.1 Данные

Мы использовали общедоступные архивы пяти журналов из различных областей науки (экономика, библиотечное дело, компьютерные науки). Всего было проанализировано более 500 выпусков. В частности, например, мы выполнили анализ 195 выпусков архива журнала «Вопросы экономики» за период с января 2003 года по апрель 2019 года. Так как, фокусом исследования было выбрано изучение авторства научных статей во времени, для этого был создан набор данных, содержащий поля: Автор, Статья, Порядковый номер автора, Дата, Журнал. Полученный набор данных содержит более 5000 авторов и 8000 статей.

3.2 Эксперимент

Результаты эксперимента приведены для архива журнала «Вопросы экономики».

3.3 Поиск редких профилей авторов

В соответствии с изложенной выше методикой были построены временные ряды позиций авторов (профилей авторов) и матрицы расстояний. Основное распределение профилей авторов соответствует Гауссовскому распределению с $\mu = 0.92$ и $\sigma = 0.09$. На основании сравнения с модельным Гауссовским распределением были выявлена группа (A0) профилей, вписывающихся в диапазон трех σ. Группа профилей, выходящая за диапазон 3σ содержит аномальные профили – обозначенные как группы A1, A2 на рисинке 4. Аномальные группы профилей авторов были выделены из набора данных и рассмотрены отдельно.

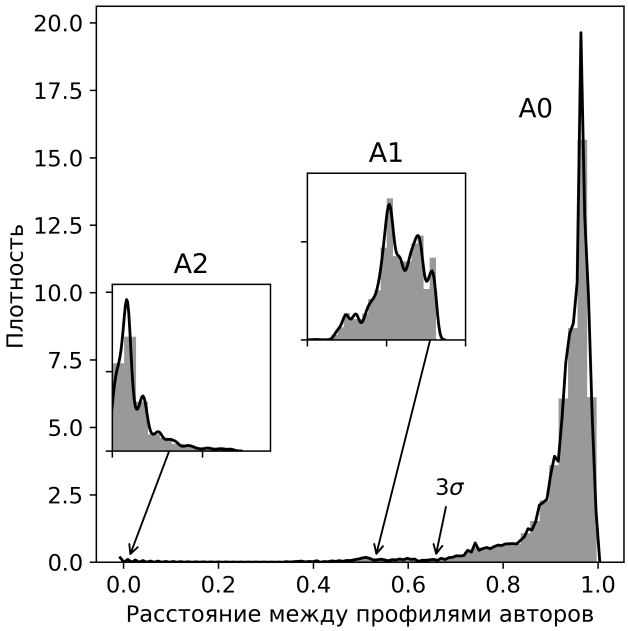

Рис. 4: Аномалии в профилях авторов

Основная группа профилей авторов (A0) была изучена с помощью методов кластерного анализа.

3.4 Кластерный анализ

В агломеративных методах иерархического кластерного анализа новые кластеры создаются путем объединения более мелких кластеров и, таким образом, дерево создается от листьев к стволу. Поиск оптимальной стратегии сортировки был произведен с помощью максимизации коэффици-

ента копенетической корреляции 7. Результаты представлены в таблице 1.

Таблица 1: Сравнение значений коэффициента копенетической корреляции

Название метода	Коэффициент копенетической корреляции
Центроидный метод	0.91
Метод средней связи	0.71
Метод Уорда	0.64
Метод одиночной связи	0.89
Метод полной связи	0.88

В результате применения центроидного метода сортировки были выделены кластеры профилей авторов. Усредненные профили для каждого кластера представлены на рисунке 5.

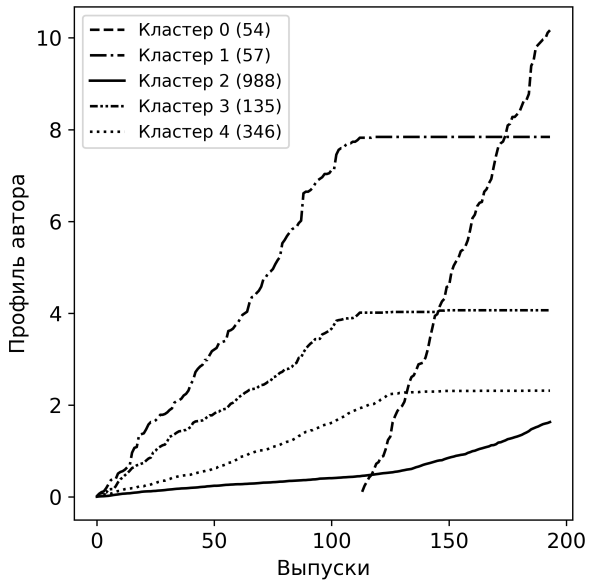

Рис. 5: Кластеры профилей авторов

Из рисунка 5 мы видим следующие особенности поведения системы «автор – издатель»:

- В районе 110-ого выпуска в редакции журнала произошли изменения. Авторы, публиковавшие статьи до 110-ого выпуска, выделены

в кластеры 1,3 и 4. После 110-ого выпуска эти авторы перестали публиковаться. С другой стороны кластеры авторов 0 и 2 наоборот начали свои публикации, а ранее не публиковались.

- Кластер 0 имеет самую большой угол наклона к оси X. Это означает, что авторы из кластера 0 участвуют в статьях не на первых позициях. Среднее значение позиции автора для кластера 0 равно 5 ± 3.

- Кластер 2 самый большой по размеру – 988 (62.5%) авторов. Среднее значение позиции в кластере 2 равно 1. Это говорит о том, что авторы из кластера 2 регулярно публикуются на первых позициях. Кластер 2 можно назвать – «ядром авторов».

Приведенные выводы визуального наблюдения могут быть получены формальным путем с помощью дифференцирования профилей соавторов. Знак второй производной по времени от среднего профиля в каждом кластере характеризует профиль как увеличивающий или уменьшающий порядковый номер автора. Вклад в исследования является наибольшим у меньших порядковых номеров исследователей. Так же, по определению, когда профиль автора становится параллельным оси времени это означает, что новые публикации отсутствуют.

4 Заключение

Авторами разработан подход к анализу системы «автор – издатель» для больших архивов журналов.

Для исследования была использована информация о порядковом номере автора в заголовке статьи, ранее не использовавшаяся в исследованиях публикационной активности.

С помощью метода остаточной дисперсии были выявлены пять кластеров профилей авторов и даны их качественные характеристики. На основании характеристик кластеров было выделено «ядро авторов» – наиболее устойчивый и продуктивный кластер.

На фоне стабильного поведения «ядра авторов» авторами сделаны суждения о развитии издателя, которые согласуются с реальностью. Так изменение редакционной политики, произошедшее в истории журнала «Вопросы экономики», совпало с характерными изменениями в выделенных кластерах профилей авторов.

Для других журналов, участвующих в эксперименте, так же были выявлены аналогичные особенности, которые нашли подтверждение.

Разработанная в данном исследовании методика анализа публикационной активности, была использована в составе комплекса методов для сравнительного анализа коллекций журналов [10, 11].

Список литературы

[1] Co-author relationship prediction in heterogeneous bibliographic networks / Yizhou Sun, Rick Barber, Manish Gupta et al. // Proceedings - 2011 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011. — 2011. — P. 121–128.

[2] The effects of funding and co-authorship on research performance in a small scientific community / Franc Mali, Toni Pustovrh, Rok Platinovšek et al. // Science and Public Policy. — 2016. — Vol. 44, no. 4. — P. 486–496.

[3] Co-authorship networks in the digital library research community / Xiaoming Liu, Johan Bollen, Michael L. Nelson, Herbert Van De Sompel // Information Processing and Management. — 2005. — Vol. 41, no. 6. — P. 1462–1480.

[4] Berndt Donald J, Clifford James. Using dynamic time warping to find patterns in time series. // KDD workshop / Seattle, WA. — 10 no. 16. — 1994. — P. 359–370.

[5] LOF: identifying density-based local outliers / Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, Jörg Sander // ACM sigmod record / ACM. — Vol. 29. — 2000. — P. 93–104.

[6] Smith Richard L et al. Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone // Statistical Science. — 1989. — Vol. 4, no. 4. — P. 367–377.

[7] De Haan Laurens, Ferreira Ana. Extreme value theory: an introduction. — Springer Science & Business Media, 2007.

[8] Farris James S. On the cophenetic correlation coefficient // Systematic Zoology. — 1969. — Vol. 18, no. 3. — P. 279–285.

[9] Carvalho Priscilla Ramos, Munita Casimiro Sepúlveda, Lapolli André Luiz. Validity studies among hierarchical methods of cluster analysis using cophenetic correlation coefficient // Brazilian Journal of Radiation Sciences. — 2019. — Vol. 7, no. 2A.

[10] Krasnov Fedor, Khasanov Mars. Unsupervised Co-Authorship Based Algorithm for Clustering of R&D Trends at Science and Technology Centers in Oil and Gas Industry. // AIST (Supplement). — 2018. — P. 1–12.

[11] Краснов Федор Владимирович, Шварцман Михаил Ефремович, Дементов Александр Владимирович. Сравнительный анализ коллекций научных журналов // Труды СПИИРАН. — 2019. — Vol. 18, no. 3. — P. 767–793.