УДК 004.89 **DOI** 10.15622/sp.2019.18.3.766-792

Ф.В. Краснов, А.В. Дименто в, М.Е. Шварцман

СРАВНИТЕЛЬНЫЙ АНАЛИЗ КОЛЛЕКЦИЙ НАУЧНЫХ ЖУРНАЛОВ

Краснов Ф.В, Дименто в А.В, Шварцман М.Е. **Сравнительный анализ коллекций научных журналов.**

Аннотация. Разработан подход для сравнительного анализа коллекций научных журналов на основе анализа графа соавторств и модели текста. Использование временных рядов метрик графа соавторства позволило провести анализ тенденций в развитии коллабораций авторов журнала. Модель текста была построена с помощью методов машинного обучения. При помощи модели текста была произведена классификация контента журналов для выявления степени аутентичности различных журналов и различных выпусков одного журнала. Разработана метрика Коэффициент контентной аутентичности, позволяющая количественно оценивать аутентичность коллекций журналов в сравнении. Сравнительный тематический анализ коллекций журналов выполнен с использованием тематической модели с аддитивной регуляризацией. На основании созданной тематической модели авторами построены тематические профили архивов журналов в едином тематическом базисе. Разработанный подход был применен к архивам двух журналов по тематике *Ревматология* за период 2000 – 2018 гг. В качестве эталона для сравнения метрик соавторств были взяты публичные наборы данных научной лаборатории SNAP Стендфордского университета. Проведено сравнение коллабораций соавторов журналов по тематике *Ревматология* с эталонными коллаборациями авторов. Произведено количественное сопоставление больших объемов текстов и метаданных научных статей. В результате проведенного авторами эксперимента с использованием разработанных методик показано, что контентная аутентичность выбранных журналов составляет 89%, соавторства в одном из журналов имеют ярко выраженную центральность, что является отличительной чертой редакционной политики. Наглядность и непротиворечивость полученных результатов подтверждает эффективность предложенного подхода. Разработанный в ходе эксперимента код на языке программирования Python может быть применен для сравнительного анализа других коллекций журналов на русском языке.

Ключевые слова: сравнительный тематический анализ, сравнительная модель текста, глубокий анализ текста, анализ социальных сетей, метрики графов.

1. Введение. Сама возможность исследования соавторств связана с появлением электронных научных библиотек. В 1994 году цифровые библиотеки стали видимыми благодаря совместной инвестиционной программе Национального научного фонда (NSF, USA) и Управления перспективными исследовательскими проектами Министерства обороны США (DARPA) [1]. В 2005 году создана Российская Ассоциация электронных библиотек. Инициаторами стали Российская государственная библиотека, Библиотека по естественным наукам РАН и другие организации.

Значительное количество исследований сетей соавторов [2–6] направлено на улучшение понимания структуры научного сотрудничества.

Соавторские сети являются важным классом социальных сетей и широко используются для определения структуры научного сотрудничества и статуса отдельных исследователей. Хотя соавторство несколько похоже на изученные сети цитирования в научной литературе [7], оно подразумевает гораздо более сильную социальную связь. Цитирование может происходить без ведома авторов, и могут изменяться во времени. Соавторство подразумевает временные и коллегиальные отношения, которые ставят его более прямо в область анализа социальных сетей.

Цифровые издательские платформы хранят архивы журналов и могут выполнять роль аналитического инструмента для редакций и для новых авторов. Новые авторы заинтересованы в том, чтобы найти наилучший журнал для публикации своей статьи. Например, у Elsevier есть Fingerprint Engine, помогающая автору определиться с журналом для публикации. Престиж журнала определяется метрикам цитируемости. Существуют различные метрики качества журналов. Например, метрика CiteScore, в которую включено около 40 тысяч журналов. CiteScore, по сути, показывает среднее цитирование публикаций издания за трехлетний период. Конечно, получить для журнала высокое значения метрики CiteScore является целью любой редакции. Например, журнал «Progress in Materials Science» обладает значением CiteScore 30.87 – это очень высокое значение. Но чем принципиально отличается журнал «Progress in Materials Science» от журнала «Materials Science and Engineering» со значением CiteScore 0.46? CiteScore – максимально прозрачная метрика, для ее подсчета не используются никакие специальные алгоритмы, но расчет производится по определенному списку журналов. Редакция журнала, стремящегося попасть в этот список, встает перед задачей приведения терминологии к мировым стандартам. Что тоже обязывает анализировать контент автоматизированными средствами и проводить сравнение с другими журналами.

В первом приближении можно представить журнал как набор статей – текст и мета информация. Для простоты дальнейшего изложения обозначим набор рассматриваемых архивов журналов как множество J, тогда один архив журнала будем обозначать как j. Каждый архив журнала состоит из множества статей p. В свою очередь, статья p состоит из текста (d), информации об авторах (a), дате выпуска (y) и других метаданных. Математическая модель для нескольких архивов журналов может быть записана в следующем виде:

$$J = (j_0, \ldots, j_N), \tag{1}$$

$$\text{где } j_i = \left[D_{j_i}\left(Y_{j_i}\right), A_{j_i}\left(Y_{j_i}\right) \right];$$
$$D_{j_i}\left(Y_{j_i}\right) = \left(d_0, \ldots, d_K\right);$$
$$A_{j_i}\left(Y_{j_i}\right) = \left(a_0, \ldots, a_L\right);$$
$$Y_{j_i} = \left(y_0, \ldots, y_k\right).$$

В выражении (1) $D_{j_i}(\circ)$ обозначает множество текстов научных статей из архива журнала j_i, $A_{j_i}(\circ)$ обозначает множество авторов статей из архива журнала j_i, а Y_{j_i} обозначает множество дат выпусков из архива журнала j_i. Для данного исследования важно подчеркнуть, что рассмотрение $D_{j_i}(\circ)$ и $A_{j_i}(\circ)$ производится в контексте определенного множества Y_{j_i}, что также может быть истолковано как зависимость $D_{j_i}(\circ)$ и $A_{j_i}(\circ)$ от времени.

Для сравнения двух коллекций журналов в самом общем случае нам необходимо определить метрику близости этих коллекций $Sim(\circ)$. Имея на входе модель коллекций J, метрику близости $Sim(J)$ должна выдавать количественную характеристику обратного расстояния между коллекциями в пространстве, определяемом моделью J. Значения $Sim(J)$ после нормировки находятся в интервале от 0 до 1. Значения $Sim(J) <= 0.5$ называют неточными, так как отличить сравниваемые в такой модели коллекции не представляется возможным. Значения $Sim(J) > 0.5$ называют точными. Так как $Sim(\circ)$ определена, а изменяются только рассматриваемые модели J, то будем далее говорить о точности модели.

Сформулируем исследовательский вопрос *B1* :

B1 : Какова будет точность модели

$$J = (j_{ProgressInMaterialsScience}, j_{MaterialsScienceAndEngineering})$$

при определении различий журналов? В частности, с какой точностью по произвольной статье $\left(\hat{d}\right)$ мы сможем определить, к какому из двух журналов эта статья относится?

Ответ на вопрос *B1* важен для определения контентной аутентичности журнала: насколько содержания двух журналов похожи. В настоящее время вопросы сравнительного анализа текстов получили развитие во многих исследованиях [8–10] и добились высокой точности в сравнении текстов на разных языках и сравнения эмоциональной окраски текстов. А в современном исследовании [11] рассматривается вопрос идентифика-

ции текстов, сгенерированных автоматически (искусственно) с помощью программных алгоритмов. Данная задача является актуальной в связи с распространением таких текстов в Интернете и даже в научных электронных библиотеках.

Кроме различий между журналами представляет ценность и понимание того, как контент одного журнала эволюционирует во времени. Отсюда следует вопрос *B2*:

> *B2* : Пусть даны два непересекающихся временных промежутка $Y^1_{j_0}$ и $Y^2_{j_0}$ такие, что $Y^1_{j_0} \in Y_{j_0}$ и $Y^2_{j_0} \in Y_{j_0}$. С какой точностью модель на основе J сможет отличить, к какому из временных промежутков относится статья $\left(\hat{d}\right)$?

Ответы на исследовательские вопросы *B1-2* дают количественные оценки и могут служить для редакций журналов побуждением к действиям. Но способы улучшения контентной аутентичности журнала остается на усмотрение редакции журнала. Нет универсальных рекомендаций по продвижению журнального контента. Поэтому точность ответов на вопросы *B1-2* не всегда должна быть высокой. Возможно, что существуют определенные временные циклы, когда с точки зрения редакции целесообразно возвращаться к определенным тематикам. И тогда точность ответов на вопрос *B2* будет невысокой, то есть, например, невозможно будет определить год выпуска журнала по произвольной статье. Или другой случай, если редакция журнала j_0 определила, что хочет быть похожа по тематикам на какой-то определённый журнал j_1, то тогда точность в ответе на вопрос *B1* для этих журналов будет невысокой: $Sim(J(j_0, j_1) < 0.5)$. Все эти нюансы определяются редакционной стратегией и ответы на вопросы *B1-2* полезны как для определения стратегии, так и для мониторинга ее выполнения.

Авторов A_{j_i} статей, входящих в архив журнала j_i, принято рассматривать как профессиональное сообщество $G(A)$. К такому представлению сообщества $G(A)$ применимы методы анализа социальных сетей (SNA – Social network analysis). Обозначим множество этих методов как M_i, к таким методам относятся: выявления сообществ, определение метрик графов, определение лидеров мнений, кластеризация сетей соавторов. Понимание структурных особенностей коллаборации своих авторов также представляется важным для редакции журнала, сосредоточенной на развитии. Очевидно, что высокоцитируемые авторы делают журналы высокоцитируемыми. Но для журнала, стремящегося к увеличению цитируемости, целесообразно «создавать» новых высокоцитируемых авторов.

С другой стороны, высокоцитируемые журналы должны обладать особенностями в структуре авторского сообщества, которые необходимо изучать. И соавторские сообщества важно изучать не только сами по себе, но и в сравнении друг с другом. Таким образом, третий исследовательский вопрос (*B3*) будет следующим:

> *B3* : Пусть даны два сообщества соавторов: $G(A_{j_0})$ и $G(A_{j_1})$. Какие методы M_i из набора средств для анализа социальных сетей являются наиболее полезными для редакций при сравнении $G(A_{j_0})$ и $G(A_{j_1})$?

Ответы на вопросы *B1-3* сами по себе являются информативными, но не дают конкретных рекомендаций по достижению целей редакции, направленных на развитие журнала. Постановка задачи по выработке рекомендаций может быть сделана в виде определения целевого состояния журнала, генерации стратегий трансформации, симуляции развития и решения оптимизационной задачи по приближению настоящего состояния журнала к целевому. Решение задачи по выработке рекомендаций выходит за рамки данного исследования.

Последующие разделы исследования содержат описание методики построения модели на основе J в сравнении с уже сделанными исследованиями в этом направлении; описание набора данных; результаты проведения цифрового эксперимента и заключение.

2. Методика исследования. Простая и широко используемая сетевая модель соавторства основана на ненаправленном двоичном графе G, в котором каждое ребро представляет отношение соавторства.

В данном исследовании авторы использовали более сложную сетевую модель соавторства, основанную на ориентированном взвешенном мультиграфе [12] с весами.

Двоичное представление графа сети соавторства опускает ряд факторов, которые формируют паттерны сотрудничества между авторами. Есть много случаев, когда двоичная сеть не соответствует здравому смыслу. Например, если два автора совместно публикуют много статей, следует ли считать связь между ними более важной, чем связь между случайными соавторами? Кроме того, если в одной статье два автора, а в другой – сто авторов, следует ли считать авторов первой статьи более связанными, чем авторов второй статьи? Чтобы разрешить эти противоречия, сеть соавторов представляют в виде ориентированного взвешенного мультиграфа. Граф соавторства G обозначается $G = (A, C, W)$, где A – множество узлов (авторов), C – множество ребер (отношения соавторов между авторами), а W – множество весов w_{ik}, которые связаны с каждым

ребром, соединяющим пару авторов (a_i, a_k). Тогда связи между двумя авторами будут определять следующие факторы:

1. Частота соавторства: авторы, которые часто являются соавторами, должны иметь более высокий вес соавтора.

2. Общее количество соавторов статей: если в статье много авторов, каждая отдельная связь соавтора должна иметь меньший вес.

Теперь мы можем определить вес соавторских связей. Для этого рассмотрим множество статей с двумя и более соавторами. Пусть множество из L авторов, как и ранее, будет обозначено как $A = (a_0, \ldots, a_L)$. Пусть множество из N статей обозначено как $P = (p_0, \ldots, p_n, \ldots, p_N)$, а $f(p_n)$ – число авторов статьи p_n. Тогда в пространстве (a_i, a_k, p_n) для каждой точки можно определить метрику E по следующей формуле:

$$E(a_i, a_k, p_n) = \frac{1}{f(p_n) - 1}. \tag{2}$$

Метрика $E(a_i, a_k, p_n)$ будет отражать степень, в которой авторы a_i и a_k имеют исключительное право на соавторство для конкретной статьи. Определение (2) для метрики $E(a_i, a_k, p_n)$ придает больший вес отношениям соавторов в статьях с меньшим количеством соавторов, чем статьям с большим числом соавторов. Другими словами, метрика $E(a_i, a_k, p_n)$ взвешивает отношение соавтора с точки зрения того, насколько оно исключительно. Поэтому метрику E принято называть *Исключительностью*.

Введем понятие *Частоты соавторства* $F(a_i, a_k)$ как суммы всех значений $E(a_i, a_k, p_n)$ для всех статей, созданных в соавторстве авторами a_i и a_k (3).

$$F(a_i, a_k) = \sum_{n=1}^{N} E(a_i, a_k, p_n). \tag{3}$$

Определение метрики *Частоты соавторства* (3) придает больший вес авторам, которые совместно публикуют больше статей и делают это только вдвоём. Более универсальной величиной является *Нормированная частота соавторства (НЧС)* (4), определенная на пространстве пар соавторов.

$$\mathcal{F}(a_i, a_k) = \frac{F(a_i, a_k)}{\sum_{k=1}^{L} F(a_i, a_k)}. \quad (4)$$

Отметим, что согласно определению (4), *НЧС* не является симметричной метрикой. Нормировочный член $\sum_{k=1}^{L} F(a_i, a_k)$ для автора a_i с большим количеством соавторов будет больше, чем для автора a_i с одним соавтором. Таким образом, *НЧС* делает сумму всех весов исходящих соавторств для одного автора равной единице.

Важными характеристиками структуры графа соавторства являются метрики, основанные на центральности: *degree centrality* (DC), *closeness centrality* (CC) и *betweenness centrality* (BC) [13]. Чтобы использовать эти метрики для сравнительного анализа коллекций научных журналов, мы провели их адаптацию. Основное содержание проведенной адаптации состоит в нормировке вышеперечисленных метрик для ориентированного взвешенных мультиграфов.

Метрика *degree centrality* для узла определяется как общее количество ребер, которые примыкают к этому узлу. Метрика *degree centrality* представляет собой простейшее воплощение понятия центральности, поскольку оно измеряет только то, сколько связей связывают авторов с их непосредственными соседями в сети. Тем не менее авторы могут быть хорошо связаны с их непосредственными соседями, но быть частью относительно изолированной группы. В таком случае получается, что хотя локально авторы хорошо связаны, но средняя величина метрики *degree centrality* будет не высока. Поэтому метрика *closeness centrality* расширяет определение метрики *degree centrality*, фокусируясь на том, насколько близок автор ко всем остальным авторам. Чтобы вычислить *closeness centrality* для узла, необходимо определить расстояния по кратчайшему пути до всех авторов в сети и инвертировать эти значения в метрику близости. Таким образом, автор, обладающий большим значением метрики *closeness centrality*, характеризуется множеством коротких связей с другими авторами в сетях.

Метрика *betweenness centrality* характеризует другой смысл центральности. Он основан на определении того, как часто конкретный узел находится на кратчайшем пути между любой парой узлов в сети. Узлы, которые часто находятся на кратчайшем пути между другими узлами, учитываются с большим весом, поскольку они контролируют поток информации в сети. Метрика *betweenness centrality* может использоваться в автономных сетях, однако высока вероятность что она будет генерировать

большое количество узлов с нулевой центральностью, поскольку многие узлы могут не действовать в качестве соединительных узлов в сети. Хотя обсуждаемые метрики центральности могут быть распространены на направленные и взвешенные сети, этому уделяется меньше внимания [13,14]. В этой статье мы сосредоточимся на их использовании в ориентированных взвешенных сетях.

Для анализа контента архивов научных журналов необходимо проанализировать текст.

Задачи по обработке текста были поставлены в 60-70 годах XX века при обработке естественного языка [15,16]. Нужно было приводить текст к более удобной для последующего анализа форме. Эту процедуру общепринято называть *нормализацией текста*. Для нормализации текста использовались регулярные выражения (regular expressions), концепцию которых разработал С. К. Клини [17]. Одним из первых, кто использовал регулярные выражения в работе с тестом был К. Томсон [18].

Важным этапом в нормализации текста является лексический анализ. Задача лексического анализа состоит в разделение текста на части: предложения, слова, буквы. Иногда лексический анализ называют токенизацией от английского слова (*tokenizing*) [19].

Другая задача нормализации текста состоит в определении слов с единой основой и называется лемматизацией. Основа слова не обязательно совпадает с морфологическим корнем слова. Лемматизация для русского языка отличается от лемматизации для английского [20–22]. Поэтому для английского языка используют процедуру лемматизации на основе частотных алгоритмов [23, 24], также называемую стемминг (от английского слова *stemming*). Но для других языков лемматизация использует еще более сложные алгоритмы. Например, есть стемминг для Древнегреческого языка [25].

Таким образом, нормализация текста состоит из трех этапов:

1. Выделения слов из текста.
2. Приведения слов к более общим формам.
3. Выделении предложений.

Для автоматизации задач нормализации текста используют библиотеки на языке программирования Python. Например, библиотеку NLTK [26], содержащую огромное количество различных алгоритмов обработки текста для построения моделей текста.

Модели, которые присваивают вероятности словам в последовательностях слов называются вероятностными моделями текста. Математически это определение можно записать в виде уравнения. Допустим, у нас есть вероятность последовательности из n слов $P(w_1, \ldots, w_n)$, такая, что

вероятность третьего слова $P(w_3)$ равна $P(w_3|w_1, w_2)$. Тогда следующее выражение определяет вероятностную модель текста.

$$P(w) = P(w_1, w_2, \ldots, w_n) = \prod_{i^n} P(w_i|w_1, w_2, \ldots, w_{i-1}). \quad (5)$$

Так как вычисление $P(w)$ представляет сложность O^n, то современные исследования текста используют представление $P(w)$ как однородной Цепи Маркова и строят приближенные модели [27]:

1. Униграмная модель $P(w_1, w_2, \ldots, w_n) \approx \prod_i P(w_i)$.
2. Биграмная модель $P(w_i|w_1, w_2, \ldots, w_{i-1}) \approx \prod_i P(w_i|w_{i-1})$.

Можно так же рассматривать n–граммные модели для большего охвата контекста, как в работах [28, 29].

В последние годы бурно развиваются методики тематического моделирования. Недавние исследования привели к развитию нескольких основных направлений: вероятностного [30] на основе SVD [31] и генеративного [32]. Тематическое моделирование определяет каждую тему как распределение некоторого количества слов с определенными вероятностями. Большинство современных тематических моделей строятся на основе распределения Дирихле (LDA, Latent Dirichlet Allocation) [33]. Трудно представить, что настолько универсальное распределение, как LDA будет одинаково хорошо работать для любых текстов. Необходимы тонкие настройки алгоритма на конкретный проблемный домен.

Формальная постановка задачи тематического моделирования следующая. Пусть зафиксирован словарь терминов W, из элементов которого складываются документы, и дана коллекция D документов $d \in D$. Для каждого документа d известна его длина n_d и количество n_{dw} использований каждого термина w. Пусть $\Phi = (\varphi_{wt})$ – матрица распределений терминов (w) в темах (t), а $\Theta = (\theta_{td})$ – матрица распределений тем (t) в документах (d). Тогда задача тематического моделирования состоит в том, чтобы найти такие матрицы Φ и Θ для выполнения равенства (6).

$$p(w|d) = \sum_{t \in T} \varphi_{wt}\theta_{td}, \quad (6)$$

где φ_{wt} – вероятности терминов w в каждой теме t, θ_{td} – вероятности тем t в каждом документе d, а $p(w|d)$ – вероятность появления термина w в документе d.

Уравнение (6) можно представить в матричном виде $\Phi \cdot \Theta$. При этом легко показать, что данная задача имеет много решений (7).

$$\Phi \cdot \Theta = \Phi \cdot \Lambda \cdot \Lambda^{-1} \cdot \Theta = \hat{\Phi} \cdot \hat{\Theta}, \tag{7}$$

где $\hat{\Phi} = \Phi \cdot \Lambda$, а $\hat{\Theta} = \Lambda^{-1} \cdot \Theta$.

Из уравнения (7) следует, что матрицы $\hat{\Phi}$ и $\hat{\Theta}$ также будут являться решениями уравнения (6). Но не все матрицы Φ и Θ будут содержать хорошо интерпретируемые тематики. Таким образом, в задачу (6) необходимо ввести условия, способствующие получению адекватных и интересных тематик. Образно можно сказать, что необходимо оцифровать специфику предметной области текста для встраивания в алгоритм поиска оптимальных матриц Φ и Θ. Отметим, что при использовании LDA для создания тематической модели такой настройки на предметную область не производится. Для решения подзадачи настройки тематической модели на предметную область авторами использован механизм регуляризаторов; последовательность применения регуляризаторов в процессе обучения тематической модели называется стратегией реуляризации. Разработка стратегии регуляризации для построения информативной тематической модели текста является трудоемкой оптимизационной задачей.

Для решения задач классификации текста широко применяются методы машинного обучения с учителем. Разметка коллекции текстов для обучения классификатора означает отнесение текста к определенному классу; например, классом для научной статьи может быть год ее выпуска или журнал, в котором она вышла. Разметка коллекции не обязательно должна производится в ручную. Для научных статей метаинформация может быть источником для выбора класса. Процесс обучения в задачах машинного обучения страдает от двух явлений: недообученности и переобученности. К сожалению, появление классификаторов на основе искусственных нейронных сетей не избавило нас от необходимости уделять внимания этим явлениям: даже нейронные сети глубокого обучения переобучиваются. Поэтому коллекцию текстов в процессе обучения разбивают на несколько случайных выборок: обучающую и проверочную. Комбинируя обучение на одной из выборок и проверку классификатора на другой, подбирают параметры классификатора и количество итераций обучения так, чтобы максимизировать обученность и минимизировать эффект от переобучения.

Основываясь на изложенных выше подходах по анализу сетей соавторства и текстов статей, проведен цифровой эксперимент сравнения двух коллекций научных журналов.

3. Эксперимент. В исследовании участвуют архивы двух журналов: j_0 и j_1. На рисунке (Рисунок 1) изображено распределение количества статей для каждого журнала по годам, а на гистограмме (Рисунок 2) распределение статей по авторам. По оси x на гистограмме (Рисунок 2) отложены количества статей на одного автора, а по оси y количества ученых с таким количеством соавторств.

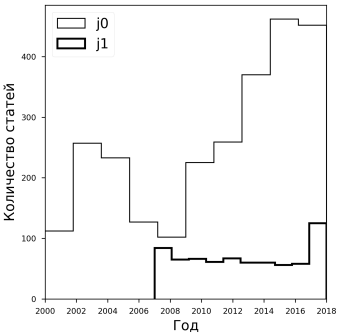

Рис. 1. Распределение статей по годам

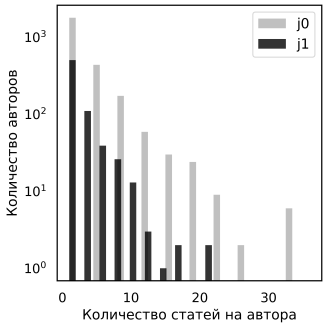

Рис. 2. Распределение статей по авторам

Общее количество статей для каждого журнала различается: j_0 издается дольше, чем j_1, имеет больше авторов и статей. В дальнейшем исследовании нужно учитывать, что эти журналы находятся на разных этапах жизненного цикла. В подтверждение этого, на рисунке 2 приведено распределение количества статей по авторам для обоих журналов. Мы видим, что в обоих журналах есть авторы с количеством статей больше 10, что составляет больше одной статьи в год.

Для каждого журнала были построены ориентированные взвешенные графы соавторств. Для сравнения мерик *Betweenness Centrality* и *Degree Centrality* кроме распределений для журналов j_0 и $j1$ были использованы метрики для сообществ Arxiv GR [1] и HepTh [2]. На рисунках 3, 4 отображены распределения метрик *Betweenness Centrality* и *Degree Centrality* для построенных графов соавторств.

Arxiv GR и HepTh – это соавторства сообществ по тематикам «Общая теория относительности и квантовая космология» и «Теоретическая физика высоких энергий» , построенные в лаборатории SNAP [34]

[1] https://snap.stanford.edu/data/ca-GrQc.html
[2] https://snap.stanford.edu/data/ca-HepTh.html

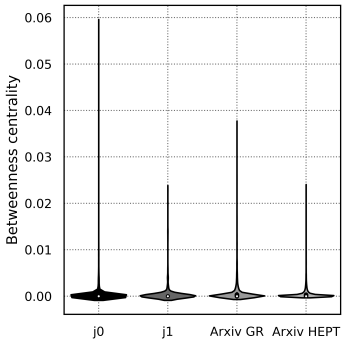

Рис. 3. Распределение метрики Betweenness Centrality

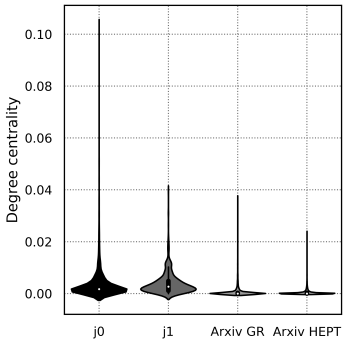

Рис. 4. Распределение метрики Degree Centrality

на данных из электронной цифровой библиотеки arXiv. Описательные данные по всем графам приведены в таблице 1.

Таблица 1. Описательные данные графов соавторств

Название	Узлы	Рёбра
ca-GrQc	5 242	14 496
ca-HepTh	9 877	25 998
j_0	2 940	13 380
j_1	1 858	2 956

В таблице 1 приведены описательные данные графов соавторств, построенных по следующему алгоритму: если автор a_i является соавтором статьи с автором p_k, то граф содержит ненаправленное ребро от i до k. Если статья написана в соавторстве с L авторами, то создается полностью связанный подграф на L узлах.

Журнал j_0 обладает наибольшим значением метрики *Betweenness Centrality* из рассматриваемых журналов. Этот факт говорит о том, что в j_0 присутствуют авторы, участвующие в большем количестве публикаций.

Проведем кластеризацию авторов на основании значений метрик центральности и Нормированной Частоты Соавторства (НЧС) (4). Для кластеризации будем использовать метод *KMeans++* из библиотеки *sklearn* [35]. Поскольку мы не знаем априори, сколько классов авторов следует ожидать (неконтролируемая классификация), нам нужно найти способ получить оценку количества кластеров. Для этого мы использовали

«метод локтя» [36] для зависимости параметра «инерции» (*inertia*) от количества кластеров.

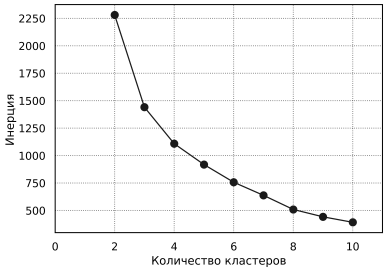

Рис. 5. Зависимость количества кластеров от параметра «инерции» алгоритма *KMeans++*

Из зависимости (рисунок 5) мы определили, что количество кластеров равно 4. Для наглядности мы присвоили кластерам интуитивно-понятные имена: студенты, аспиранты, научные сотрудники, руководители. И провели кластеризацию построенных графов соавторства.

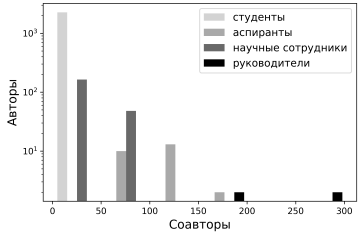

Рис. 6. Кластеры авторов для журнала j_0.

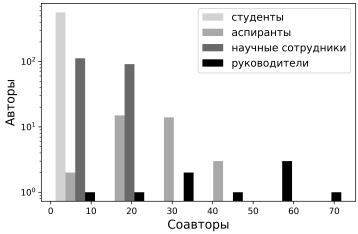

Рис. 7. Кластеры авторов для журнала j_1.

Результаты проведенной кластеризации отображены на гистограммах (рисунки 6 и 7). Мы видим, что в кластер «руководители» попали авторы с самым большим количеством соавторов. Но в журнале j_0 среднее значение в распределении по количеству соавторов для кластера «руководители» еще не настолько велико, как в журнале j_1. Это свидетельствует о том, что есть авторы, проводящие исследования с большим количеством разных соавторов. Такие авторы являются идеологами, лидерами мнений данного сообщества. Другой полюс – кластер «студенты» представлен одинаково для обоих журналов это соавторы с небольшим количеством статей, написанных в одиночку.

Перейдем к рассмотрению эволюции журналов во времени. Наглядной метрикой для для рассмотрения в виде временного ряда может служить плотность графа (*Density*) [37], вычисляемая как число узлов L к количеству всех комбинаций из N ребер (8).

$$Density = \frac{2L}{N(N-1)}. \tag{8}$$

Для эволюции журнала увеличение плотности графа соавторства означает, что новые авторы пишут в соавторстве с уже опубликовавшими свои статьи авторами. И наоборот, уменьшение плотности означает, что новые авторы пишут самостоятельно, без привлечения в соавторы ранее публиковавшихся авторов данного журнала. Для интуитивного понимания плотности графа на диаграммах (рисунки 8 и 9) приведены графы соавторов одного автора с большим значением метрики *Degree*.

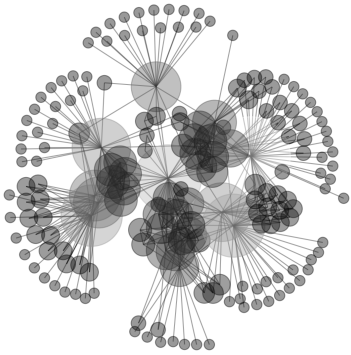

Рис. 8. Соавторы для узла *a12* (j_0)

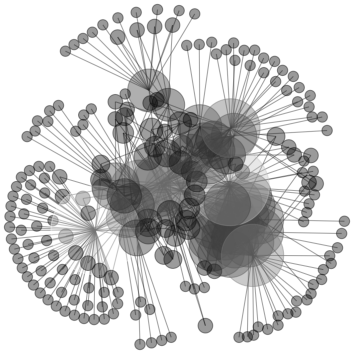

Рис. 9. Соавторы для узла *a16* (j_1)

Более точное распределение изменения плотности графов для каждого журнала отображено на рисунке 10. Для получения этой зависимости на основании мета информации о статьях были построены графы соавторства для каждого года. Граф соавторства для определенного года строится на основании всех статей, опубликованных до этого года включительно.

Из зависимостей, отображенных на рисунке 10 можно сделать вывод о том, что журнал j_0 более сосредоточен на повторных статьях от сформировавшегося круга авторов. А журнал $j1$ активнее привлекает новых авторов.

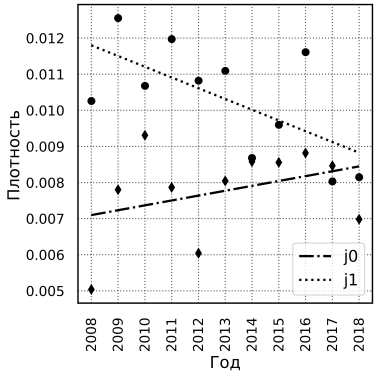

Рис. 10. Распределение плотности графов соавторств по годам

Рассмотрим контентное наполнение журналов. Контент журнала – это набор статей. Каждая статья представляет из себя текст на русском языке. Для работы с текстом мы используем вероятностную модель текста.

В исходном виде, архивы журналов были представлены в формате «Дублинского ядра» (Dublin Core). В ходе предварительной обработки архивов была извлечена мета-информация о статьях, текст приведен к нормальному виду, удалены высокочастотные и редкие слова, создан словарь для биграммной модели, емкость которого составила 22.7 тысяч терминов. Для анализа текста мы выбрали инструмент на основе тематической модели текста с аддитивной регуляризацией, описанный в [38]. Настройка тематической модели текста на данную коллекцию научных статей состоит из подбора оптимальных параметров тематической модели и выбора стратегии регуляризации. Одним из наиболее важных параметров тематической модели является количество тематик. Так как тематическая модель сама не может определить количество тематик, то воспользовались методикой, разработанной в исследовании [39]. Суть этой методики состоит в поиске максимума метрики *cDBI*, характеризующей качество модели в зависимости от количества кластеров (Рисунок 11).

На рисунке 11 нам важно, что максимум качества тематическая модель достигает при 22 темах. Для выбора стратегии регуляризации была использована методика, разработанная в исследовании [40]. В основе этой методики лежит принцип кластеризации тематик на базе плотности. В результате получены два кластера тематик: основные (sbj_i) и шумовые(nz_i). Основные тематики статьи отражают ее главные темы, отличающие эту статью от других статей. Таких тематик в статье может быть одна или

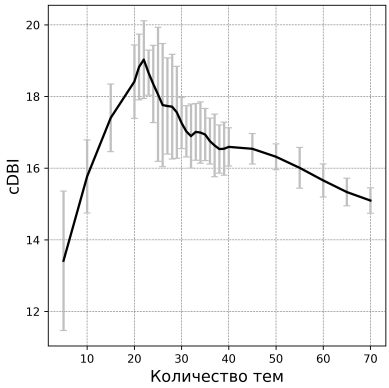

Рис. 11. Распределение метрики качества тематической модели (*cDBI*) в зависимости от количества тематик

несколько. Шумовые тематики отражают общие для всех статей тематики. Например, цитирование основных исследований, сделанных в данном направлении.

Из определения (5) следует способ для проверки качества вероятностной модели текста. Для этого пользуются метрикой Перплексия (*Perplexity*) : $\mathcal{P} \approx \sqrt{\frac{1}{P(w)}}$, где $P(w)$ – это полная вероятность из выражения (6). На графике (рисунок 12) показано, как метрика *Perplexity* уменьшается с обучением модели. На интуитивном уровне уменьшение *Perplexity* тематической модели означает, что растет упорядоченность модели.

Метрика *Perplexity* не имеет интерпретируемых абсолютных значений. Из наблюдений известно, что в моделях на русском языке она выше, чем в моделях на английском. Поэтому при обучении тематической модели этап уменьшения *Perplexity* обычно предшествует регуляризации. На рисунке 13 отображены зависимости матриц Θ и Φ из выражения (7) от итераций обучения модели. После достижения *Perplexity* полого уменьшения на 10 итерации включается стратегия разреживания основных тематик до 20 итерации и уплотнения значений шумовых тематик до 30 итерации обучения. Количество итераций было подобрано так, чтобы разреженность значений обоих матриц для основных тематик превысила 60%. Такова была стратегия последовательной регуляризации тематической модели. Получившаяся при такой стратегии матрица Θ для каждого журнала отображена на рисунках 14 и 15.

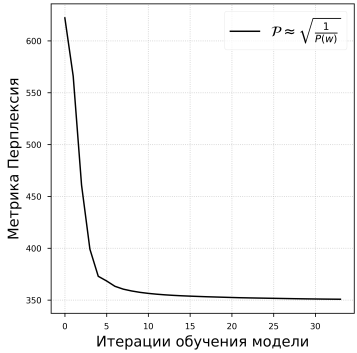

Рис. 12. Зависимость Перплексии тематической модели от итераций обучения модели

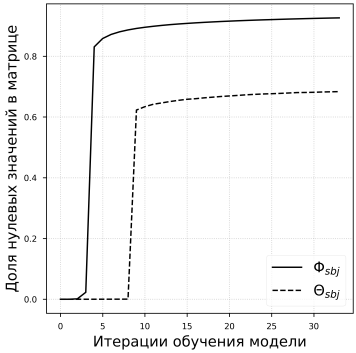

Рис. 13. Доля нулевых значений в тематической модели в зависимости от итераций обучения

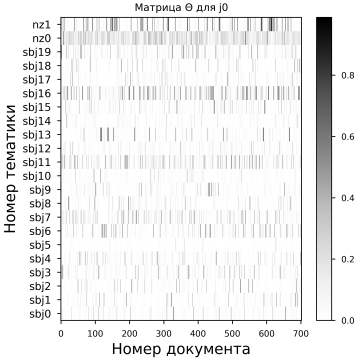

Рис. 14. Матрица Θ для j_0

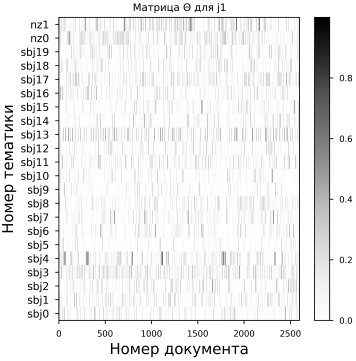

Рис. 15. Матрица Θ для j_1

Значения матрицы Θ для каждого журнала соответствуют вероятностям появления тематик в данном документе. Значение матрицы Φ представляют распределения вероятностей для терминов всего словаря. Матричное произведение Φ и Θ дает, согласно уравнению (7), векторное представление текста коллекции. Приведем пример тематики для одного документа №601 с названием «Распространенность и клинические особенности подагры и болезни депонирования пирофосфата кальция у пациентов с острым артритом». Этот документ полностью (0.998) посвящен тематике sbj6. Наибольшими вероятностями у тематики sbj6

обладают следующие термины: *подагра*, *больной подагра*, *нпвп*, *кристалл*, *мужчина*, *аллопуринол*, *тофус*.

Из рисунков 14 и 15 видно, что тематики в коллекциях j_0 и j_1 представлены по разному. Для количественной оценки этого соотношения был построен профиль тематик каждой коллекции. Так как коллекции содержат разное количество выпусков, то профили были нормированы (Рисунок 16). Теперь слало более наглядно, что, например, тематике sbj5 в обоих журналах уделено достаточно мало внимания. Возможно, эта тематика содержит потенциал для научных исследований.

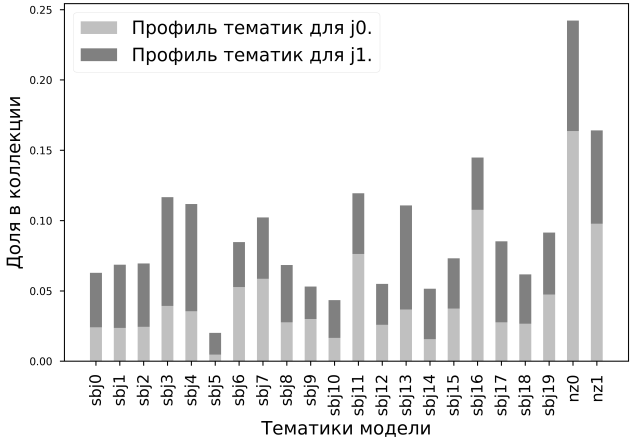

Рис. 16. Профили тематик.

В таблице 2 приведена расшифровка кодов тематик: sbj_i и nz_i. Для краткости по каждому коду тематики приведены топ-5 терминов с наибольшими значениями вероятностей.

Отметим, что термины в описании тематик в таблице 2 приведены в простых словоформах. Как мы видим, среди терминов встречаются униграммы и биграммы. Корректность терминов в тематиках точнее может оценить только эксперт в данной предметной области.

Перейдем к рассмотрению эволюционного развития контента журналов. Для этого выберем за один шаг развития один год. Выбор меньшего шага не будет информативным, так как, например, при выборе в качестве шага одного выпуска мы не сможем оценить изменения из-за небольшого объема выпусков. Рассмотрим более подробно задачу определения контентной аутентичности (*B2*). Для выявления контентной аутентичности журналов использовался метод прямого измерения. Постановка задачи

Таблица 2. Таблица терминов тематик

sbj0	ребенок	ювенильный	взрослый	орло	вариант	детский	юиа
sbj1	атеросклероз	артерия	артериальный	кардиоваскулярный	сосудистый	ибс	ссз
sbj2	ссд	легкое	легочный	ипл	больной ссд	склеродермия	поражение легкое
sbj3	балл	шкала	опросник	здоровье	соэ	активность заболевание	ваш
sbj4	тромбоз	афс	сердце	мутация	клапан	ных	ной
sbj5	врс	сердце	память	инвалидность	ритм	инвалид	интервал
sbj6	подагра	больной подагра	нпвп	кристалл	мужчина	анкилозировать	анкилозировать спондилит
sbj7	гибп	ремиссия	бпвп	монотерапия	ада	инф	неделя
sbj8	пес	псориаз	увеит	заболеваемость	псориатический	население	псориатический артрит
sbj9	инфекция	депрессия	расстройство	антибиотик	сон	реа	сутки
sbj10	операция	эндопротезирование	коленный	хирургический	тазобедренный	коленный сустав	тцз
sbj11	коленный	коленный сустав	остеоартроз	неделя	ваш	хрящ	гонартроз
sbj12	экспрессия	хрящ	ген	хондроцит	синовиальный	коллаген	рост
sbj13	научный	ревматологический	профессор	центр	здравоохранение	страна	наука
sbj14	ртм	гибп	ремиссия	достижение	курс	ритуксимаба	бпвп
sbj15	позвоночник	кость	спа	мрт	отдел	костный	спондилит
sbj16	нпвп	жкт	диклофенак	нимесулид	средство	безопасность	нестероидный
sbj17	цитокин	концентрация	активация	рецептор	иммунный	аутоиммунный	провоспалительный
sbj18	скв	больной скв	волчанка	красный	системный красный	красный волчанка	беременность
sbj19	инфекция	васкулит	кожа	вирус	лихорадка	узел	гепатит
nz0	должный	фно	ревматол	например	побочный	частность	появление
nz1	перелом	костный	мпк	остеопороз	бедро	кость	отдел

выглядит следующим образом: найти точность классификации фрагмента текста достаточной длины из архива журнала относительно двух классов – j_0 или j_1? В результате такого измерения мы выясним, насколько массив текста j_0 отличается от j_1. Данная задача может быть решена с помощью методов машинного обучения с учителем. Мы разделяем весь массив текста на три части: обучающая выборка, проверочная выборка, отложенная выборка. На обучающей и проверочной выборке мы производим настройку классификатора, а на отложенной выборке проверяем насколько хорошо классификатор справляется с новыми данными. Отложенная выборка необходима для контроля за переобученностью (overfitting) классификатора. В качестве векторной модели текста авторы использовали модель TF-IDF. Моделью классификатора был выбран RandomForest с 300 эстиматорами. Метрикой качества для модели была выбрана Accuracy (точность). На проверочной выборке была получено значение 0.90, а на отложенной выборке 0.89. Также был проведен анализ классификации по «кривой ошибок» – ROC-анализ. На рисунке 17 представлена ROC-кривая («кривая ошибок») для рассматриваемой модели классификации.

По оси x отложена *Специфичность алгоритма* классификации, также называемая false positive rate (FPR). А по оси y отложена *Чувствительность алгоритма* классификации, также называемая true positive rate (TPR). Количественную интерпретацию ROC дает показатель AUC (англ. area under ROC curve, площадь под ROC-кривой) — площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций. Чем выше показатель AUC, тем качественнее классификатор, при этом значение 0,5 демонстрирует непригодность выбранного метода классификации (соответствует случайному гаданию). Значение менее 0,5 говорит, что классификатор действует с точностью до наоборот: если положительные назвать отрицательными и наоборот, классификатор будет

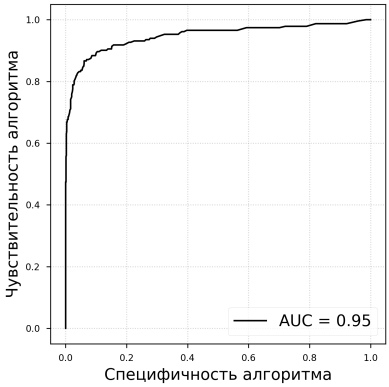

Рис. 17. ROC-кривая для классификации контента журналов

работать лучше. В рассматриваем результате эксперимента оба журнала легко (95%) отличить друг от друга по содержанию с помощью алгоритмов машинного обучения.

Далее разобьем архивы журналов по годам. Проверка будет состоять в том, чтобы измерить точность отнесения тестового контента к определенному году выпуска. Такая постановка относится к задачам машинного обучения с учителем и решается с помощью классификации. Классами будут являться пара из двух лет. Например, пара 2005 и 2017. Таким образом, у нас будет определена размеченная выборка для обучения классификатора. Процесс обучения будет состоять в выборе коэффициентов классификатора, которые, получив на вход текст, будут выдавать год, к которому этот текст относится. Если точность такой классификации для определенной пары лет меньше 0.5, то это означает, что отличить тексты разных лет не представляется возможным. Далее будем называть значение точности классификации, выраженное в процентах коэффициентом аутентичности контента (КАуК).

На рисунках 18 и 19 представлены матрицы аутентичности контента для каждого журнала. Для каждой пары лет вычислен коэффициент аутентичности при помощи изложенного выше подхода. Видно, что в среднем контент журнала имеет высокую аутентичность для j_0 — 92%, j_1 — 97%. Но есть пары лет в которых контент повторялся. Например, для j_0 в паре 2009 – 2010 КАуК = 50%, а для j_1 в паре 2010 – 2011 КАуК = 67%.

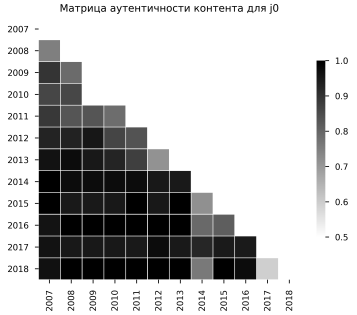

Рис. 18. Матрица аутентичности контента для j_0

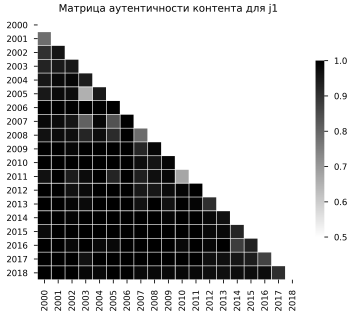

Рис. 19. Матрица аутентичности контента для j_1

4. Заключение. В данном исследовании разработана новая методика анализа архивов журналов при помощи комбинированного подхода на основании Теории графов и Компьютерной лингвистики. Основной акцент в исследовании авторы поставили на сравнительном анализе архивов журналов, что потребовало создания единого аналитического пространства и разработки метрик этого пространства. Сложность данной задачи обусловлена большим объемом информации и невозможна без автоматизированных процедур. В исследовании авторами задействованы не только текст научных статей, но и метаинформация из научных статей. Авторы не ограничились описательным сравнением, а исследовали глубинные структурные различия двух коллекций журналов. Научная новизна данного исследования состоит во взамимоувязке нескольких теоретических разработок с целью выработки новой методики прикладного анализа. Практическая ценность настоящей статьи состоит в разработке и экспериментальной проверке нового подхода к редакционному анализу контента научных периодических журналов. В исследовании сформулированы вопросы (*B1–3*), имеющие практическую ценность для стратегии управления контентом научного журнала, и найдены экспериментальные ответы. А именно:

– проведено сравнение количественных метрик журналов на основании графов соавторств. Показаны различия в структуре соавторств журналов. Выявлены стратегические принципы выбора авторов редакциями журналов

– выполнена кластеризация авторов журналов. Предложено единое кластерной пространство и интуитивно-понятные названия кластеров. На

основании выделенных кластеров построен инструмент для влияния на определенные группы соавторов для усиления их позиций на общем фоне авторов

– проведен анализ эволюции контента журналов во времени на основе предложенной авторами метрики «плотности». Показаны тенденции в развитии профессионального сообщества авторов для каждого журнала во времени. Сделано заключении о последствиях такой редакционной политики

– проведен контентный сравнительный анализ на основе тематической модели. Разработана стратегия регулиризации тематической модели для получения информативных и хорошо интерпретируемых тематик. Показаны потенциалы развития тематик. Проведен сравнительный анализ профилей тематик для двух журналов тематики

– выполнен анализ контентной аутентичности на основании предложенной авторами новой метрики: коэффициента аутентичности контента (КАуК).

Полученные результаты выполнены с использованием высоконагруженного вычислительного кластера с привлечением современных свободно распространяемых библиотек на языке программирования Python. Доработки существующих алгоритмов выполнены также на языке программирования Python, что позволяет говорить о возможности независимой воспроизводимости результатов и возможности проведения экспериментов с другими коллекциями научных журналов.

Результаты данного исследования создают основу для выработки изменений в редакционной политике рассматриваемых журналов и позволяют ранжировать эффективность мер по совершенствованию контента.

Литература

1. *Wiederhold G*. Intelligent integration of information// ACM SIGMOD Record. 1993. vol. 22. no. 2. pp. 434-437.
2. *Newman M.E.J*. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality// Physical review E. 2001. Vol. 64. no. 1. pp. 016131.
3. *Smeaton A.F. et al*. Analysis of papers from twenty-five years of SIGIR conferences: what have we been doing for the last quarter of a century?// ACM SIGIR Forum. 2002. vol. 36. pp. 39–43.
4. *Farkas I. et al*. Networks in life: Scaling properties and eigenvaluespectra// Physica A: Statistical Mechanics and its Applications. 2002. vol. 314. no. 1-4. pp. 25–34.
5. *Cunningham S.J., Dillon S.M*. Authorship patterns in information systems// Scientometrics. 1997. Vol. 39, no. 1. pp. 19.
6. *Egghe L., Rousseau R., Van Hooydonk G*. Methods for accrediting publications to authors or countries: Consequences for evaluation studies// Journal of the American Society for Information Science. 2000. vol. 51, no. 2. pp. 145–157.
7. *Garfield E*. Is citation analysis a legitimate evaluation tool?//Scientometrics. 1979. vol. 1, no. 4. pp. 359–375.

8. *Witten I.H., Frank E., Hall M.A., Pal C.J.* Data Mining: Practical machine learning tools and techniques// Morgan Kaufmann. 2016. 558 p.
9. *Lucas C. et al.* Computer-assisted text analysis for comparative politics//Political Analysis. 2015. vol. 23, no. 2. pp. 254-277.
10. *Zhao W.X. et al.* Comparing twitter and traditional media using topic models // European conference on information retrieval. 2011. pp. 338–349.
11. *Шумская А.О.* Метод определения искусственных текстов на основе расчета меры принадлежности к инвариантам//Труды СПИИРАН. 2016. Вып 6(49). С.104-121.
12. *Bondy J.A., Murty U.S.R. et al.* Graph theory with applications // London: Macmillan. 1976. vol. 290. 270 p.
13. *Wasserman S., Faust K.* Social network analysis: Methods and applications // Cambridge university press. 1994. vol. 8. 857 p.
14. *Newman M.E.J.* Analysis of weighted networks // Physical review E. 2004. vol. 70. no. 5. pp. 056131.
15. *Weizenbaum J.* ELIZA — a computer program for the study of natural language communication between man and machine // Communications of the ACM. 1966. vol. 9. no. 1. pp. 36–45.
16. *Kucera H., Francis W.N.* Computational analysis of present-day American English // Dartmouth Publishing Group. 1967. 424 p.
17. *Kleene S.C.* Representation of events in nerve nets and finite automata // RAND PROJECT AIR FORCE SANTA MONICA CA. 1951. 101 p.
18. *Thompson K.* Programming techniques: Regular expression search algorithm // Communications of the ACM. 1968. vol. 11. no. 6. pp. 419–422.
19. *Lovins J.B.* Development of a stemming algorithm // Mech. Translat. & Comp. Linguistics. 1968. vol. 11. no. 1-2. pp. 22–31.
20. *Segalovich I.* A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // International Conference on Machine Learning; Models, Nechnologies and Applications (MLMTA). 2003. pp. 273–280.
21. *Sharoff S., Nivre J.* The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge // 2011 Russian Conference on Computational Linguistics. 2011. 14 p.
22. *Korobov M.* Morphological analyzer and generator for Russian and Ukrainian languages // International Conference on Analysis of Images, Social Networks and Texts. 2015. pp. 320–332.
23. *Willett P.* The Porter stemming algorithm: then and now // Program: electronic library and information systems. 2006. vol. 40. no. 3. pp. 219–223.
24. *Porter M.F.*Snowball: A language for stemming algorithms. 2001. URL: http://snowball.tartarus.org/texts/introduction.html (дата обращения: 15.02.2019).
25. *Packard D.* Computer-assisted morphological analysis of ancient Greek // Proceedings of the International Conference on Computational Linguistics (COLING-1973). 1973. vol. 2. 14 p.
26. *Bird S., Klein E., Loper E.* Natural language processing with Python: analyzing text with the natural language toolkit // O'Reilly Media, Inc. 2009. 504 p.
27. *Schwenk H., Gauvain J.L.* Connectionist language modeling for large vocabulary continuous speech recognition // 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2002. vol. 1. pp. I-765–I-768.
28. *Teahan W.J., Cleary J.G.* The entropy of English using PPM-based models // Proceedings of Data Compression Conference-DCC'96. 1996. pp. 53–62.

29. *Teahan W.J., Cleary J.G.* Models of English text // Proceedings DCC'97. Data Compression Conference. 1997. pp. 12–21.
30. *Hofmann T.*. Probabilistic latent semantic indexing // ACM SIGIR Forum. 2017. vol. 15. no. 2. pp. 211–218.
31. *Lu X., Zheng X., Li X.* Latent semantic minimal hashing for image retrieval // IEEE Transactions on Image Processing. 2016. vol. 26. no. 1. pp. 355–368.
32. *Law J.* Latent Topical Skip-Gram for mutually learning topic model and vector representations // arXiv preprint arXiv:1702.07117. 2017.
33. *Blei D.M., Ng A.Y., Jordan M.I.* Latent dirichlet allocation // Journal of machine Learning research. 2003. vol. 3. pp. 993–1022.
34. *Leskovec J., Kleinberg J., Faloutsos C.* Graph evolution: Densification and shrinking diameters // ACM Transactions on Knowledge Discovery from Data (TKDD). 2007. vol. 1. no. 1. pp. 2.
35. *Arthur D., Vassilvitskii S.* k-means++: The advantages of careful seeding // Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. 2007. pp. 1027–1035.
36. *Bholowalia P., Kumar A.*. EBK-means: A clustering technique based on elbow method and k-means in WSN // International Journal of Computer Applications. 2014. vol. 105. no. 9. pp. 17–24.
37. *Alba R.D.* A graph-theoretic definition of a sociometric clique // Journal of Mathematical Sociology. 1973. vol. 3. no. 1. pp. 113–126.
38. *Воронцов К.В., Потапенко А.А.* Тематические модели с аддитивной регуляризацией // Machine Learning. vol. 101. no. 3. pp. 303–323.
39. *Krasnov F., Sen A.* The Number of Topics Optimization: Clustering Approach // Machine Learning and Knowledge Extraction. 2019. vol. 1. no. 1. pp. 416–426.
40. *Краснов Ф.В., Ушмаев О.С.* Разведка скрытых направлений исследований в нефтегазовой отрасли с помощью анализа библиотеки OnePetro // International Journal of Open Information Technologies. 2018. vol. 6. no.5. pp. 7–14.

Краснов Федор Владимирович – канд. техн. наук, эксперт, блок Научного Инжиниринга, ГазпромНефть Научно-Технический Центр. Область научных интересов: интеллектуальная аналитика текстов. Число научных публикаций — 58. krasnov.fv@gazpromneft-ntc.ru; наб. реки Мойки, 75–79, литер Д, 190000, Санкт-Петербург, Российская Федерация; р.т.: +7 (812)313-6924; факс: +7 (812)313-6924.

Шварцман Михаил Ефремович – заместитель директора, дирекция, Национальный электронно-информационный консорциум; начальник отдела, Отдел исследования компьютерных систем, ФГБУ Российская государственная библиотека. Область научных интересов: электронные библиотеки, анализ текста. Число научных публикаций — 91. shvar@neicon.ru; Летниковская, 5, 115114, Москва, Российская Федерация; р.т.: +79031995708; факс: +7(499)754-99-94.

Дементов Александр Владимирович – начальник ИТ отдела, ИТ отдел, Национальный электронно-информационный консорциум. Область научных интересов: информатика, наукометрия и библиометрия. Число научных публикаций — 9. dimentov@neicon.ru; Летниковская, 5, 115114, Москва, Российская Федерация; р.т.: 7(499)754-99-94; факс: +7(499)754-99-94.

DOI 10.15622/sp.2019.18.3.766-792

F.V. KRASNOV, A.V. DIMENTOV, M.E. SHVARTSMAN

COMPARATIVE ANALYSIS OF SCIENTIFIC JOURNALS COLLECTIONS

Krasnov F.V., Dimentov A.V., Shvartsman M.E. **Comparative analysis of scientific journals collections.**

Abstract. The authors developed an approach to comparative analysis of scientific journals collections based on the analysis of co-authors graph and the text model. The use of time series of co-authorship graphs metrics allowed analysis of trends in the development of journal's authors. The text model was built using machine learning techniques. The journals content was classified to determine the authenticity degree of various journals and different issues of a single journal via a text model. A developed metric of Content Authenticity Ratio allows quantifying the authenticity of journal collections in comparison. Comparative thematic analysis of journals collections was carried out using the thematic model with additive regularization. Based on the created thematic model, thematic profiles of the journals archives in a single thematic basis were constructed. The approach developed by the authors was applied to archives of two journals on the Rheumatology for the period from 2000 to 2018. As a benchmark for comparing the co-author's metrics, public data sets from the SNAP research laboratory at Stanford University were used. As a result, the existing examples of the effective functioning of the authors collaborations in order to improve the work of journals' editorial staff were adapted. Quantitative comparison of large volumes of texts and metadata of scientific articles was carried out. As a result of the experiment conducted using the developed methods, it was shown that the content authenticity of the selected journals is 89%, co-authorships in one of the journals have a pronounced centrality, which is a distinctive feature of the editors' policy . The clarity and consistency of the results confirm the effectiveness of the approach proposed by the authors. The code developed in the course of the experiment in the Python programming language can be used for comparative analysis of other collections of journals in the Russian language.

Keywords: Comparative Thematic Analysis, Comparative Text Model, Deep Text Analysis, Social Network Analysis, Graph Metrics.

Krasnov Fedor Vladimirovich – Ph.D., Expert, Science Engineering Department, Gazprom Neft Science and Technology Center. Research interests: text mining. The number of publications — 58. krasnov.fv@gazpromneft-ntc.ru; 75–79, литер Д, Moika River emb., 190000, St Petersburg, Russian Federation; office phone: +7 (812)313-6924; fax: +7 (812)313-6924.

Shvartsman Mikhail Efremovich – deputy director, Directorate, National Electronic Information Consortium; head of department, R&D Department, Russian State Library. Research interests: Digital Library, text mining. The number of publications — 91. shvar@neicon.ru; 5, Letnikovskaya, 115114, Moscow, Russian Federation; office phone: +79031995708; fax: +7(499)754-99-94.

Dimentov Alexander Vladimirovich – head of the information department, IT Department, National Electronic Information Consortium. Research interests: Quantitative Science of Science: informetrics, scientometrics, bibliometrics. The number of publications — 9. dimentov@neicon.ru; 5, Letnikovskaya, 115114, Moscow, Russian Federation; office phone: 7(499)754-99-94; fax: +7(499)754-99-94.

References

1. Wiederhold G. Intelligent integration of information. *ACM SIGMOD Record*. 1993. vol. 22. no. 2. pp. 434–437.
2. Newman M.E.J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality.*Physical review E*. 2001. vol. 64. no. 1. pp. 016131.
3. Smeaton A.F. et al. Analysis of papers from twenty-five years of SIGIR conferences: what have we been doing for the last quarter of a century?. *ACM SIGIR Forum*. 2002. vol. 36. pp. 39–43.
4. Farkas I. et al. Networks in life: Scaling properties and eigenvalue spectra. *Physica A: Statistical Mechanics and its Applications*. 2002. Vol. 314, no. 1-4. pp. 25-34.
5. Cunningham S.J., Dillon S.M. Authorship patterns in information systems. *Scientometrics*. 1997. vol. 39, no. 1. pp. 19.
6. Egghe L., Rousseau R., Van Hooydonk G. Methods for accrediting publications to authors or countries: Consequences for evaluation studie.*Journal of the American Society for Information Science*. 2000. vol. 51. no. 2. pp. 145–157.
7. Garfield E. Is citation analysis a legitimate evaluation tool? *Scientometrics*. 1979. vol. 1, no. 4. pp. 359-375.
8. Witten I.H., Frank E., Hall M.A., Pal C.J. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. 2016. 558 p.
9. Lucas C. et al. Computer-assisted text analysis for comparative politics. *Political Analysis*. 2015. vol. 23, no. 2. pp. 254-277.
10. Zhao W.X. et al. Comparing twitter and traditional media using topic models. *European conference on information retrieval*. 2011. pp. 338–349.
11. Shumskaya A.O. [Method of the Artificial Text Identification based on the Calculation of the Belonging Measure to the Invariants]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2016. vol. 6(49). pp. 104–121. (In Russ.).
12. Bondy J.A., Murty U.S.R. Graph theory with applications. London: Macmillan. 1976. vol. 290. 270 p.
13. Wasserman S., Faust K. Social network analysis: Methods and applications. Cambridge university press. 1994. vol. 8. 857 p.
14. Newman M.E.J. Analysis of weighted networks. *Physical review E*. 2004. vol. 70. no. 5. pp. 056131.
15. Weizenbaum J. ELIZA — a computer program for the study of natural language communication between man and machine. *Communications of the ACM*.1966. vol. 9. no. 1. pp. 36–45.
16. Kucera H., Francis W.N. Computational analysis of present-day American English. Dartmouth Publishing Group. 1967. 424 p.
17. Kleene S.C. Representation of events in nerve nets and finite automata. RAND PROJECT AIR FORCE SANTA MONICA CA. 1951. 101 p.
18. Thompson K. Programming techniques: Regular expression search algorithm. *Communications of the ACM*. 1968. vol. 11. no. 6. pp. 419-422.
19. Lovins J.B. Development of a stemming algorithm. *Mech. Translat. & Comp. Linguistics*. 1968. vol. 11. no. 2. pp. 22-31.
20. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. International Conference on Machine Learning; Models, Nechnologies and Applications (MLMTA). 2003. pp. 273–280.
21. Sharoff S., Nivre J. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. 2011 Russian Conference on Computational Linguistics. 2011. 14 p.

22. Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages. International Conference on Analysis of Images, Social Networks and Texts. 2015. pp. 320–332.
23. Willett P. The Porter stemming algorithm: then and now. Program: electronic library and information systems. 2006. vol. 40. no. 3. pp. 219–223.
24. Porter M.F. Snowball: A language for stemming algorithms. 2001. Available at: http://snowball.tartarus.org/texts/introduction.html (accessed: 15.02.2019).
25. Packard D. Computer-assisted morphological analysis of ancient Greek. Proceedings of the International Conference on Computational Linguistics (COLING-1973). 1973. vol. 2. 14 p.
26. Bird S., Klein E., Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc. 2009. 504 p.
27. Schwenk H., Gauvain J.L. Connectionist language modeling for large vocabulary continuous speech recognition. 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2002. vol. 1. pp. I-765–I-768.
28. Teahan W.J, Cleary J.G. Models of English text. Proceedings DCC'97. Data Compression Conference. 1997. pp. 12–21.
29. Teahan W.J, Cleary J.G. Models of English text. Proceedings DCC'97. Data Compression Conference. 1997. pp. 12–21.
30. Hofmann T. Probabilistic latent semantic indexing. ACM SIGIR Forum. 2017. vol. 15. no. 2. pp. 211–218.
31. Lu X., Zheng X., Li X. Latent semantic minimal hashing for image retrieval. *IEEE Transactions on Image Processing*. 2016. vol. 26, no. 1. pp. 355-368.
32. Law J. Latent Topical Skip-Gram for mutually learning topic model and vector representations. arXiv preprint arXiv:1702.07117. 2017.
33. Blei D.M., Ng A.Y., Jordan M.I. Latent dirichlet allocation. *Journal of machine Learning research*. 2003. vol. 3. pp. 993–1022.
34. Leskovec J., Kleinberg J., Faloutsos C. Graph evolution: Densification and shrinking diameters. ACM Transactions on Knowledge Discovery from Data (TKDD). 2007. vol. 1. no. 1. pp. 2.
35. Arthur D., Vassilvitskii S. k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. 2007. pp. 1027–1035.
36. Bholowalia P., Kumar A. EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications*. 2014. vol. 105. no. 9. pp. 17–24.
37. Alba R.D. A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*. 1973. vol. 3. no. 1. pp. 113–126.
38. Vorontsov K.V., Potapenko A.A. [Additive regularization of topic models]. *Machine Learning*. vol. 101. no. 3. pp. 303–323. (In Russ.).
39. Krasnov F., Sen A. The Number of Topics Optimization: Clustering Approach. *Machine Learning and Knowledge Extraction*. 2019. vol. 1. no. 1. pp. 416–426.
40. Krasnov F.V., Ushmaev O.S. [Exploration of Hidden Research Directions in Oil and Gas Industry via Full Text Analysis of OnePetro Digital Library]. *International Journal of Open Information Technologies*. 2018. vol. 6. no. 5. pp. 7–14. (In Russ.).