УДК 004.65 + 005 + 001.5

ИДЕНТИФИКАЦИЯ АВТОРОВ В РАМКАХ ПРЕДМЕТНОЙ ОБЛАСТИ В СЕМАНТИЧЕСКОЙ БИБЛИОТЕКЕ

О. М. Атаева[1, [0000-0003-0367-5575]], В. А. Серебряков[2, [0000-0003-1423-621X]], Н. П. Тучкова[3, [0000-0001-6518-5817]]

[1,2,3]*Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН, г. Москва*

[1]oli@ultimeta.ru, [2]serebr@ultimeta.ru, [3]natalia_tuchkova@mail.ru

Аннотация

Рассмотрены особенности задачи идентификации авторов и определения авторского вклада в публикации в цифровых библиографических коллекциях. Особенности проблемы недостаточной идентификации проявляются в повторах информации, двойниковании, наличии авторов с полностью совпадающими именами, самоцитировании, автоплагиате и собственно плагиате. Предлагается использовать информацию о публикациях, которая уже накоплена в цифровой библиотеке в виде связанных данных предметной области и множества данных тезауруса адресата, как автора и пользователя библиотеки. Эта информация содержит связи, благодаря которым для идентификации авторства можно использовать контексты ключевых слов, множества соавторов и ассоциативные связи терминов в словарях и тезаурусах. Важно, что рассматривается массив научных публикаций, поскольку они имеют сложившуюся традиционную структуру, что позволяет сравнивать фиксированные элементы текста (аннотации, ключевые слова, коды классификаторов и т. д.). Таким образом, даже при полном совпадении имен в публикациях можно ставить вопрос об авторстве, если в цифровой библиотеке публикации соответствуют различным предметным областям. Разрешение таких противоречий осуществляется путем оценки множества связей всех элементов вторичной информации о публикации. Результатом сравнения может быть добавление автора в некоторую предметную область, т. е. расширение тезауруса адресата и персонального тезауруса автора, или появление в библиотеке полных тезок, но из разных областей знаний. Показано, что современные средства анализа данных

позволяют оценить вклад автора в публикацию, несмотря на то, что конечно, реальный вклад в научное исследование может оценить только научное сообщество.

Ключевые слова*: сравнение научных текстов, семантический поиск, тезаурус для онтологии знаний*, *информационный запрос с помощью тезауруса, семантические библиотеки, способы идентификации авторов, тезаурус адресата, вторичная информация, частотный словарь индивидуума***,** *LibMeta*.

ВВЕДЕНИЕ

Проблемы определения того, кто заслуживает быть автором научной статьи и каков его вклад в коллективную публикацию, если в цифровой коллекции нет достоверной информации, разрешаются различными способами. В основном выполняются сопоставление близких по тематике статей и опрос зарегистрированных авторов, как в ResearchGate. С вопросами, связанными с идентификацией авторов в библиографических системах, сталкиваются практически все цифровые ресурсы, известные на сегодняшний день. При обновлении информации могут появиться «спорный» автор, полный тезка, «старый» автор с другой транскрипцией в написании фамилии и т. д. Всем известны трудности собственной идентификации даже в таких авторитетных базах данных, как WoS и Scopus, когда несмотря на все выставленные фильтры, получаем в результате поиска список из «смеси» своих и чужих работ, что отражено, например, в публикации [1]. Нередко приходится вручную формировать необходимый список, несмотря на существующий в этих системах (как и во многих других) механизм автоматического формирования авторского указателя. Исключение составляют публикации и издания, в которых изначально требуется задать ORCID автора. Собственные идентификаторы ввели также elibrary (SPIN-код автора), система ИСТИНА (IstinaResearcherID, IRID), Scopus (Scopus Author ID), Web of Science ResearcherID, Google Scholar Citation ID. Чем больше индексов указывает автор при регистрации в этих системах и статьях при передаче их издательствам, тем точнее он идентифицируется, естественно. Некоторые издательства делают обязательными ссылки на индексы авторов соответствующих баз данных, с которыми эти издательства сотрудничают. Тот факт, что идентификаторы авторов сопровождают публикации, говорит о том, что другие

способы, несмотря на принятые правила идентификации, оказываются недостаточно надежными.

Существует ряд требований к статьям и авторам в отдельных специфических предметных областях, и они были утверждены, например, для авторства в медицинских исследованиях, но стали впоследствии общепринятыми. Автор – это тот, кто участвует в развитии идеи, сборе и анализе данных, написании работы, внесении в текст актуальных и идеологически оправданных изменений.

Тем не менее, для определения вклада автора в коллективные исследования этих средств недостаточно, на что указывается, например, в работе [2]. Более того, в цифровой век в некоторых научных сообществах существуют варианты: обсуждение коллегами вклада авторов в исследования; предоставление издательствам права высказывать мнение об авторстве на основе накопленной информации. Это нарушает традиционные нормы, принятые ранее [3].

Изменился уровень достоверности, прозрачности и документирования данных об авторах. Таким образом, проблема авторства ставится шире и не ограничивается вторичной информацией при индексации в базах данных. Эта проблема включает человеческий фактор, опрос экспертов, редакторов и соавторов. В целом отмечается тенденция увеличения числа соавторов за последние 30 лет [4], хотя для отечественных научных работников это ведет к известным проблемам в отчетности перед фондами и министерствами.

В настоящей работе рассматриваются варианты использования данных, которые имеются в арсенале современных информационных технологий для индексации публикаций, авторов и их вклада в коллективные работы в цифровых библиографических системах.

1. О СРЕДСТВАХ ИДЕНТИФИКАЦИИ АВТОРОВ

1.1. Множества данных для идентификации авторов

Структура научной публикации – это особенность научных статей, вполне устоявшаяся для многих отечественных и международных журналов. Строгость, которой предлагается придерживаться авторам в соответствии с инструкцией от издателей, продиктована в какой-то мере процессом оцифровки публикаций для последующей их индексации в библиографических базах данных. В 1070-х годах появилось семейство стандартов для машиночитаемой каталогизации (*Machine-*

Readable Cataloging, MARC) [5] с дальнейшей разработкой стандарта ISO 2709 (ГОСТ 7.14-84 (СТ СЭВ 4269-83) СИБИД и ГОСТ 7.14-98 СИБИД). Эти стандарты первоначально были предложены Библиотекой конгресса США в качестве форматов межбиблиотечного обмена библиографическими данными, а позднее адаптировались для национальных библиотек и стали в той или иной форме использоваться во всех англоязычных библиотечных системах. Естественным образом стандартные поля библиографических записей для машиночитаемой каталогизации стали компонентами и фиксированными позициями в структуре научных статей.

Таким образом, был сформирован список обязательных полей вторичной информации о документе «научная статья»: автор(ы), аффиляция автора(ов), название, ключевые слова, классификаторы (MSC, UDC и/или специализированные), выходные данные (издательство, страницы, год). В дальнейшем добавились аннотация, список цитируемой литературы и идентификаторы, такие как ORCID и др. Все эти поля используются для индексирования публикаций и могут участвовать в качестве поисковых при формировании запроса и идентификации авторов.

Трудность возникает, если этой информации недостаточно или ее нет в полном объеме в базе данных или у пользователя. Уточнение осуществляется благодаря экспертным знаниям или за счет семантических связей, которые могут быть реализованы в виде подсказок из базы данных.

Тело публикации, как правило, недоступно для поиска, даже если публикация находится в открытом доступе, но доступно издателям для предварительной лексической, синтагматической, парадигматической, семантической обработки при размещении в библиографических базах данных.

1.2. Набор данных тезауруса адресата

Понятие «адресата в информационной среде», сформулированное для удобства определения пользователей и авторов из баз данных, подразумевает персону – участника информационного процесса, поиска и обмена информацией. Термин «тезаурус адресата (индивидуума)» (ТА) введен в информатику Ю.А. Шрейдером [6] для представления предметной области (ПрО) автора на основе понятийного запаса знаний автора. Термин связан также с представлением «знаний» в информационной системе как «структурированной информации» [7].

Для более подробного знакомства с использованием тезаурусов в поисковых процессах и извлечения знаний можно обратиться к работе [8]. В дальнейшем проявилась важность этого представления, как основы для описания онтологии адресата (ОА) в современных базах данных [9].

Состав данных (информации) тезауруса адресата зависит от понятийного запаса индивидуума. Можно остановиться на следующем наборе данных: частотный словарь индивидуума; варианты сочетаний терминов; контексты частотных терминов; специальные обозначения и формулы; списки цитируемой литературы; списки цитирующих авторов; список публикаций с перекрестными ссылками. Если в информационной системе достаточно данных и публикаций по некоторой предметной области, то на основе множества данных о тезаурусе адресата и метрического анализа можно построить *словарь-тезаурус предметной области автора*. Далее, сравнивая предметные тезаурусы авторов, можно более точно их идентифицировать, а также устанавливать принадлежность текста некоторому автору и его вклад в исследования.

1.3. Инструменты сравнения текстов для идентификации авторов

Рассматриваются методы сравнения текстов для установления авторства, такие как частотные алгоритмы [10], контекстное сравнение [11], тематическая кластеризация и алгоритмы глубокого анализа текстов, связанные с методами машинного обучения [12], [13].

Используя эту совокупность методов, можно сформировать технологию обработки информации для *вновь поступающих данных* в информационную библиографическую систему.

Первый этап предварительной обработки (препроцесса) публикаций для *каждого автора* включает:

- частотную обработку текстов для получения списка терминов с их весом (частотой использования);
- составление списка соавторов;
- формирование множества контекстов для терминов.

В результате накапливаются следующие данные (параметры) *автора*: список (словарь) терминов, ранг (вес) терминов, словоформы терминов, относительная частота терминов (по отношению к другим терминам), абсолютная частота

терминов, конкорданс словарь (словарь с контекстами), рис. 1. На этом этапе также возможно выделить список уникальных терминов, обозначений, формул и других особенностей текста, характерных для некоторых авторов и предметных областей.

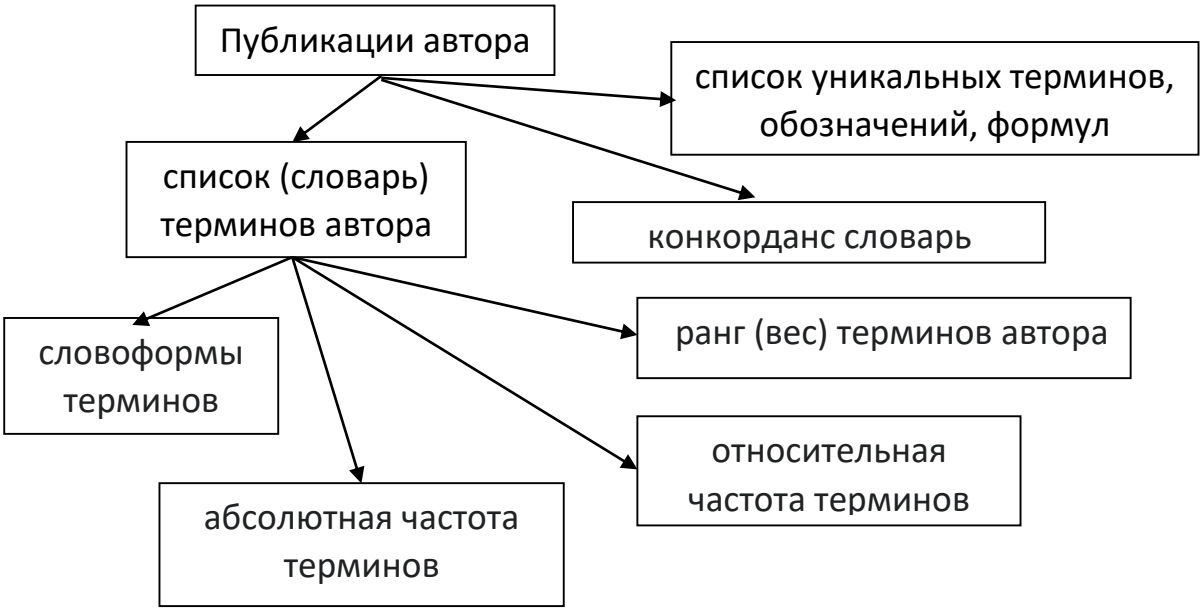

Рис. 1. Схема предварительной обработки публикаций автора.

Второй этап заключается в процедуре сравнения авторов по имеющимся (накопленным) параметрам. Выявляются пересечения множеств терминов, контекстов и уникальных терминов, обозначений и т. д.

После сравнения и выявления множества публикаций, принадлежащих определенному автору, составляются авторский указатель и указатель цитируемых публикаций. При этом можно варьировать строгость принадлежности «спорных» публикаций тому или другому автору, учитывая степень совпадений выявленных параметров (в %, например).

На этом предварительная обработка *вновь поступающих данных об авторе* может быть закончена.

Все это множество связанной полученной информации можно считать тезаурусом адресата.

Замечание 1. Если в систему предполагается загрузить *серию публикаций одного автора* (или авторского коллектива), то можно на предварительном этапе обработки составить тезаурус адресата (адресатов).

Замечание 2. Если поступила единичная работа, то предварительная обработка (по схеме рис. 1) используется для включения в имеющийся авторский указатель или при отсутствии совпадений и спорных свойств публикации (варианты фамилий и других вторичных документов) хранится в статусе подтверждения, но участвует в дальнейшей предметной семантической обработке. Подтверждение можно делать автоматически, если в системе накопится дополнительная информация об авторе или по запросу к автору.

Для дальнейшей семантической обработки публикаций необходимо использовать словари (тезаурусы) профессиональных терминов из предметных областей (например, математических).

Публикации необходимо проиндексировать в соответствии с предметной и тематической направленностью, определяя принадлежность терминов публикаций словарям (тезаурусам) предметных областей. Таким образом можно зафиксировать связи тезауруса адресата (автора) с предметными областями. Эти связи представляют в дальнейшем дополнительные *признаки для предметной идентификации автора*.

Таким образом, публикации, связанные семантически в онтологиях, в результате препроцессорной обработки будут иметь еще ряд признаков идентификации авторов.

2. ПРИМЕРЫ НА НАБОРАХ ДАННЫХ

На примере некоторого множества работ по разделам высшей математики можно рассмотреть варианты идентификации авторов публикаций со схожими наборами вторичных документов.

Для обработки текста используется свободная библиотека для высокопроизводительного полнотекстового поиска Apache Lucene, реализованная на языке Java.

2.1. Установление авторства

Для выделения значимых выражений документа использовался расчет меры tf-idf для терминов документа, извлеченных из индекса, с учетом морфологии [13]. На первом этапе рассматривались только существительные и термины, которые были идентифицированы как имена собственные.

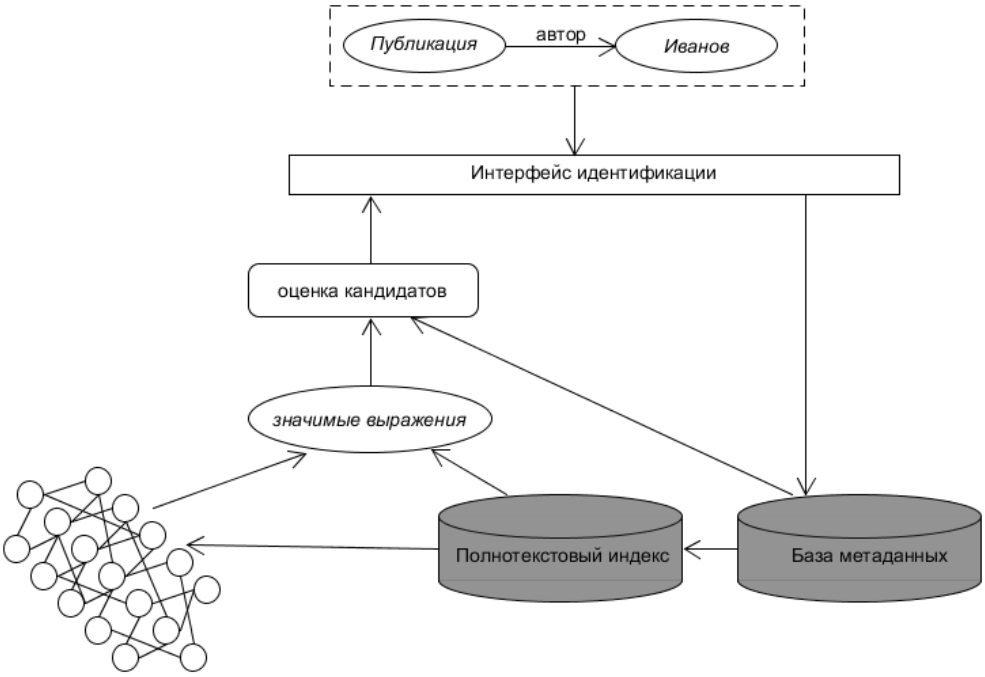

Рис. 2. Общая схема работы с терминами и авторами.

Далее исключались термины, для которых мера tf-idf была меньше порогового значения. Составление комбинаций из двух и трех слов выполнялось на основе использования контекста выделенных слов и правил, учитывающих морфологию. Под контекстом понимаются N слов, находящихся в тексте перед словом, для которого строится вектор, и N слов, находящихся после этого слова. Для выделения контекста используется неглубокая нейросетевая модель word2vec [14]–[16] в режиме «skip-grams». На рис. 2 представлена общая схема работы.

В качестве примера далее на рис. 3 отражен этап формирования тезаурусов предметных областей отдельных авторов, на основе которых можно рассуждать об их (авторов) идентичности.

Из примера видно, что были получены работы авторов с неполным набором вторичной информации. Применение описанного алгоритма позволяет выявить термины, связи и пересечения подмножеств терминов с учетом их контекстов.

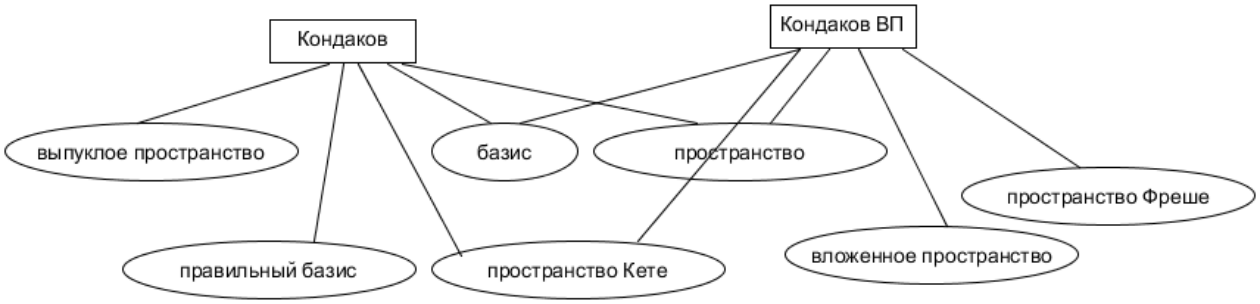

Рис. 3. Общая схема сравнения авторов.

Используем далее дополнительно связи терминов из энциклопедии, классификаторов УДК, MSC и других работ из области аналитических пространств, такие, как представлены на рис. 4.

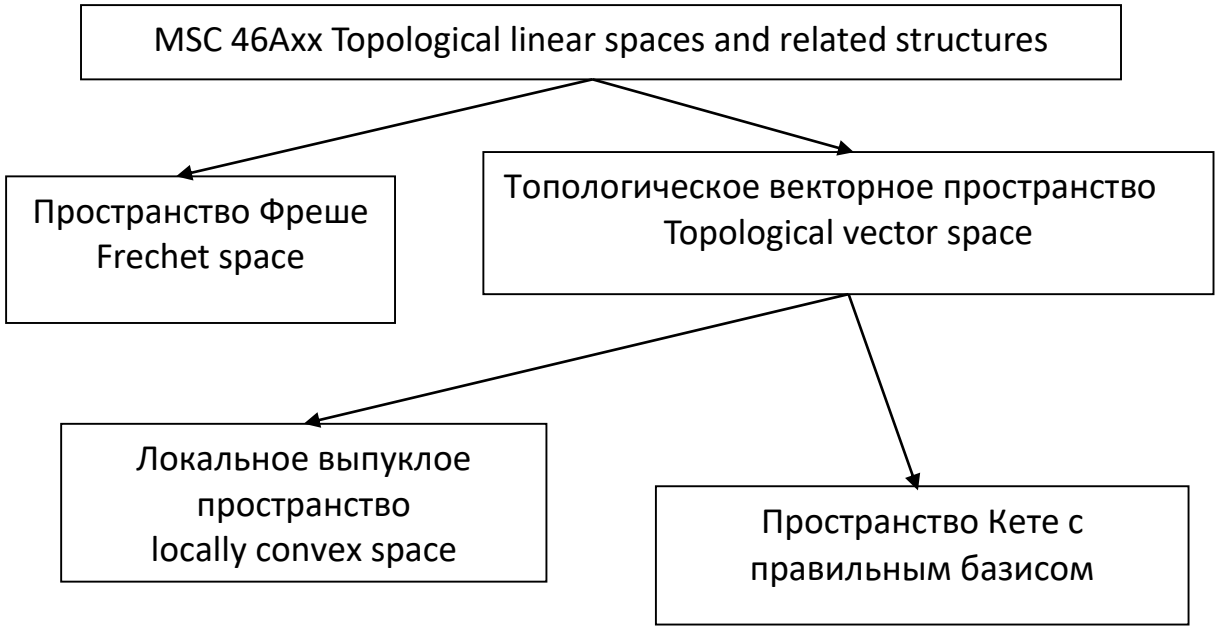

Рис. 4. Связи выявленных терминов авторов.

Было обработано около 5000 авторов публикаций. Отдельно проводится работа по обработке формул и включения их в тезаурус автора. Используется алгоритм сравнения формул на основе векторной модели. Алгоритм условно делится

на две части: первичный отбор формул-кандидатов и последующее их упорядочение по схожести. Описание этого алгоритма выходит за рамки данной статьи.

2.2. Учет авторского вклада

Для учета авторского вклада в публикацию требуется исследовать историю работ автора и его принадлежность научным школам, а также исследования автора в предметных областях. Это имеет особое значение, поскольку соавторство стало носить коммерческий характер и стали возможны платные публикации, «старшее авторство» и цитирование [17].

Совокупность «исторических» данных об авторе и публикации формируется на основе тезауруса адресата следующим образом. Собирается и хранится история публикаций: соавторы, перекрестные ссылки, ключевые слова, внутренние системные индикаторы принадлежности публикаций предметным областям (LibMeta). Практически все современные библиографические коллекции собирают и демонстрируют перечисленные данные. Развитие информационных технологий позволяет привлечь различные методы анализа для установления авторства публикаций. Надо отметить, что для художественных произведений типа «карманных» детективов такая экспертиза проводится давно, поскольку этот процесс изначально построен на коммерческой основе, и необходим учет вклада каждого участника. Для научного сообщества требуется избегать такого подхода, так как это неизбежно ведет к принижению значения исследовательской работы.

Выбраны признаки, по которым распределяются публикации, это история вопроса, новизна, количество публикаций на смежные темы, множества соавторов, экспертное мнение, выраженное в процессе дискуссий и рецензирования.

История публикации. Структура научной статьи предполагает наличие вступления, в котором перечисляются предыдущие исследования. Анализируя этот текст, можно составить списки исследователей и соответствующих библиографических ссылок по выбранной теме, пример – рис. 5.

Далее нужно выбрать пересечения внутри этих множеств и выявить «главных» авторов и их соавторов. Для соавторов выявить частотные характеристики и принадлежность предметной области. Таким образом, получить «карту» публикаций по теме, где будут области пересечения авторских коллективов, где пересекаются {k1, k2, …, kN} авторов (k1>k2> …>kN). В эти области включаются и авторы

из списков цитирования. Отдельно стоящие авторы (A, B, C, …) могут принадлежать множествам «приглашенных» к участию в публикациях, и тогда их роль оценивается экспертами из научного сообщества. Это могут быть авторы публикаций без соавторов, работающие в данной предметной области, и тогда, естественно, их вклад в работу не оспаривается.

Рис. 5. Общая схема сравнения авторов.

Множество авторов k1, которое больше других, может претендовать на множество ведущих ученых, руководителей научных школ и исследовательских проектов (грантов и пр.).

Оценка по ключевым словам. Пересечение ключевых слов в тезаурусах авторов свидетельствует о близости исследований.

Новизна. Анализ коллективов, составляющих множества {k1, k2, …, kN}, позволяет выявить «новых» членов авторского коллектива, за какой-то период времени, «новые» ключевые слова за этот же период времени. Поскольку благодаря тезаурусу адресата можно выяснить, к какому автору относятся «новые» ключевые слова, то можно сделать вывод о том, благодаря кому появился «новый» вклад в публикации и предметную область.

На основе данных тезауруса адресата в системе LibMeta введена метрика оценки авторского участия в публикации по математическим предметным областям. Вычисляются следующие множества:

- ядро ключевых концептов предметной области (Concept Kernel) – {СК = $К_1 \cap К_2 \cap К_3$ }, где $К_1$ – тезаурус ОДУ, К2 – словарь спец. функции и К3 – математической энциклопедии:

$$|KK|=|K_1|+|K_2|+|K_3|, |K_1|=184, |K_2|=151, |K_3|=6263, |KK|=6\,598,$$

- ядро ключевых слов информационных объектов для разных типов ресурсов ресурсов предметной области (Keyword Kernel) – {KK}, |KK|=6810,

- ядро авторских коллективов по годам (Kernel of Copyright Teams) – {КСТ}.

Рассмотрим для примера 2015 год для публикаций, затрагивающих *обыкновенные дифференциальные уравнения Бернулли*[1].

Получаем: $|KK_{2015}|=754$, $КСТ_{2015}$={*'Лазарев'*, *'Неустроева'*, *'Шишкина'*, *'Бочкарев'*, *'Лекомцев'*, *'Сенин'*, *'Янковский'*, *'Кольцун'*}, ядро библиографических ссылок (Bibliographic Reference Kernel) – {BRK} для этих авторов представлено 34 ссылками, |BRK|=34.

Далее оценивается пересечение данных из ТА автора: ключевых слов {KWA}, $|KWA_{Лазарев}|=14$, $|KWA_{Янковский}|=79$, соавторов {CA} $|CA_{2015}|=163$, библиографических списков {RL} $|RL_{Лазарев}|=3$, $|RL_{Янковский}|=16$, с множествами $\{KK_{2015}\}$, $\{КСТ_{2015}\}$, $\{BRK_{2015}\}$ к общим характеристикам предметной области:

$$\{KWA\} \cap \{KK\}, \{CA\} \cap КСТ\}, \{RL\} \cap \{BRK\}.$$

На основании этого вводятся оценки вклада автора в предметную область $KWA_{Лазарев}/KK_{2015}=14/754$, «средний» вклад автора в предметную область в этом году $CA_{2015}/КСТ_{2015}=163/8$, «средний» вклад $|RL_{Лазарев}|/|BRK|=3/34$, $|RL_{Лазарев}|/|BRK|=16/34$.

Эти оценки показывают вклад автора в предметную область и конкретные исследования (публикации) «во времени». Подчеркнем, что эти оценки не отражают картину реального мира, но справедливы для характеристики того множе-

[1] http://libmeta.ru/concept/showRelatedValues/404?attribute=119

ства объектов, которые загружены в систему. Авторы, у которых наибольший процент «пересечений» с онтологией ПрО, могут считаться «ключевыми» исследователями в предметной области.

Рассмотрим матрицу (Таблица 1) признаков публикации по новизне, где критерий – это новые ключевые слова.

Таблица 1: Соответствие *Ключевых слов* предметной области публикациям и авторам

	публикации (art)	сравнение {art} и {artПрО}	ПрО цифровой библиотеки	содержит art%
	авторы (au)	сравнение {au} и {auПрО}		содержит au%
Ключевые слова	тезаурусы автора (auths)	сравнение {auths} и {thsПрО}		содержит auths%
	UDC (udc)	сравнение {udc} и {udcПрО}		содержит udc%
	MSC (msc)	сравнение {msc} и {mscПрО}		содержит msc%
	Формулы (form)	сравнение {form} и {formПрО}		содержит form%

Таблица 1 многомерная и содержит наибольшее количество возможных связей ключевых слов. В ней присутствует связь *ключевых слов* с авторским предметным тезаурусом (если он есть) и тезаурусом ПрО, который заложен в основу онтологии ПрО в семантической библиотеке. На основе сравнения множеств из столбцов получаем значения (например, в процентном отношении) в последнем столбце и принимаем решение о принадлежности ключевых слов ПрО семантической библиотеке или о новом множестве для этой библиотеки и ПрО, т. е. можно принять решение о том, насколько новая публикация соответствует ПрО.

Замечание 3: В данном исследовании не дается никакой оценки обоснования исследований авторов и качества научных работ.

Замечание 4: Все оценки делаются только на основе публикаций, вторичной информации или полных текстов (если они доступны) и авторских методов, отслеживания связей в цифровой библиотеке.

Замечание 5: Реальный вклад автора в публикацию и исследования может оценить только научное сообщество. В цифровой библиотеке можно установить количество и тип связей по выбранным признакам и на основе массива данных, который есть в библиотеке. Этот анализ дает картину вклада публикации и рейтинг автора в масштабах имеющихся данных, но не качества публикации и знаний автора в целом.

Замечание 6: В библиотеке LibMeta используется технология создания предметного авторского тезауруса, и на его основе можно получить представление о тезаурусе адресата как участника обмена информацией в информационной среде. Эта технология позволяет рассматривать *значение и вклад* публикаций автора применительно к различным предметным областям, которые составляют пересечение множеств в рамках предметного авторского тезауруса.

В работе представлена идеальная схема оценки роли автора и установления авторства, конечно, в ней есть спорные факторы, но схема может быть использована как первое приближение, если авторство статьи вызывает сомнения по причине неточности вторичных данных в цифровой библиотеке.

В реальности провести границу между претендентами на авторство публикации может быть непросто, что иногда является предметом спора научных школ. Известны случаи, когда идея и ее реализация в исследованиях принадлежат различным людям, которые могут знать или не знать о работах друг друга. Здесь поднимаются вопросы плагиата и приоритетов в науке. Пример тому являет история разногласий Ньютона и Лейбница по вопросу вклада каждого в развитие математического анализа [18]. Именно в цифровых библиотеках можно учесть, если не все, то многие признаки авторства, что показано на примерах математических статей в LibMeta.

ЗАКЛЮЧЕНИЕ

Предложена технология предварительной обработки публикаций для дальнейшего размещений в семантической библиотеке. Использование данных теза-

уруса адресата позволяет накапливать структурированную информацию об авторах и публикациях, что способствует на предварительном этапе идентифицировать авторов и оценить их вклад в исследования.

Использование персональной среды для научного исследования на базе индивидуальных библиографических коллекций и результатов, собранных автором в процессе исследований, позволяет рассматривать задачи идентификации и определения авторского вклада как часть функционирования семантической библиотеки.

Благодарности

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект № 20-07-00324, и в рамках темы Министерства науки и высшего образования РФ «Математические методы анализа данных и прогнозирования».

СПИСОК ЛИТЕРАТУРЫ

1. *Krämer T., Momeni F., Mayr P.* Coverage of Author Identifiers in Web of Science and Scopus. – arXiv preprint arXiv:1703.01319, 2017 – arxiv.org.

Clement T.P. Authorship Matrix: A Rational Approach to Quantify Individual Contributions and Responsibilities in Multi-Author Scientific Articles // Science and Engineering Ethics. 2014. V. 20. P. 345–361. URL: https://doi.org/10.1007/s11948-013-9454-3.

2. *Frische S.* It is time for full disclosure of author contributions// Nature. 2012. P. 489.

URL: http://www.nature.com/news/it-is-time-for-full-disclosure-of-author-contributions-1.11475.3.

3. *Cozzarelli N.R.* Responsible authorship of papers in PNAS // Proceedings of the National Academy of Sciences of the United States of America. 2004. V. 101, No. 29. P. 10495.

4. MARC 21 Formats. URL: http://www.loc.gov/marc/marcdocz.html.

5. *Шрейдер Ю.А.* Тезаурусы в информатике и теоретической семантике // Научно-техническая информация. Сер. 2. 1971. № 3. С. 21–24.

6. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. СПб.: Питер, 2000. 384 с.

7. *Лукашевич Н.В.* Тезаурусы в задачах информационного поиска. М.: Изд-во МГУ, 2011. 495 с.

8. *Муромский А.А., Тучкова Н.П.* Об онтологии адресата в математической предметной области // Электронные библиотеки. 2018. Т. 21, № 6. С. 506–533.

9. *Борисов Л.А., Орлов Ю.Н., Осминин К.П.* Идентификация автора текста по распределению частот буквосочетаний // Препринты ИПМ им. М.В. Келдыша. 2013. № 27. 26 с.
URL: http://library.keldysh.ru/preprint.asp?id=2013-27.

10. TextSTAT - Simples Text Analyse Tool. URL: http://neon.niederlandistik.fu-berlin.de/textstat/.

11. *Mohsen A.M., El-Makky N.M., Ghanem N.* Author Identification Using Deep Learning, 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, 2016. P. 898–903.
URL: https://doi.org/10.1109/ICMLA.2016.0161.

12. *Маннинг К.Д., Рагхаван П., Шютце Х.* Введение в информационный поиск. Cambridge University Press, 2018. 482 p.

13. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR, 2013.

14. *Mikolov T., Yih W.T., Zweig C.* Linguistic Regularities in Continuous Space Word Representations // Proceedings of NAACL HLT, 2013.

15. *Le Q., Mikolov T.* Distributed Representations of Sentences and Documents // International Conference on Machine Learning, 2014. P. 1188–1196.

16. *Strange K.* Authorship: Why not just toss a coin? // American Journal of Physiology-Cell Physiology. 2008. V. 295, No. 3. P. 567–575.
URL: https://doi.org/10.1152/ajpcell.00208.2008.

17. *Meli D.B.* Equivalence and Priority: Newton versus Leibniz: Including Leibniz's Unpublished Manuscripts on the Principia. Clarendon Press, 1993. P. 318.

AUTHORS IDENTIFICATION WITHIN THE SUBJECT AREA IN THE SEMANTIC LIBRARY

O. M. Ataeva[1, [0000-0003-0367-5575]], V. A. Serebriakov[2, [0000-0003-1423-621X]], N. P. Tuchkova[3, [0000-0001-6518-5817]]

[1,2,3]*Dorodnicyn Computing Centre FRC CSC RAS, Moscow*

[1]oli@ultimeta.ru, [2]serebr@ultimeta.ru, [3]natalia_tuchkova@mail.ru

Abstract

The peculiarities of the task of authors identifying and determining author's contribution to publications in digital bibliographic codes are considered. The features of the problem of insufficient identification are manifested in the repetition of information, doubling, the presence of authors with completely coincidental names, self-quotation, autoplagiate and plagiarism itself. It is proposed to use publication information that has already been accumulated in the digital library in the form of related object area data and a variety of target thesaurus data, as the author and user of the library. This information contains links whereby keyword contexts, multiple co-authors, and term associations in dictionaries and thesauruses can be used to identify authorship. It is important that an array of scientific publications is considered, since they have an established traditional structure, which allows comparing fixed text elements (annotations, keywords, classifier codes, etc.). Thus, even if the names in the publications are fully matched, the question of authorship can be raised if the publications in the digital library correspond to different subject areas. Resolution of such contradictions is accomplished by evaluating a plurality of links of all elements of secondary publication information. The result of the comparison could be the addition of the author to a specific area, i.e. the extension of the addressee's thesaurus and the author's personal thesaurus, or the appearance of full namesakes in the library, but from different areas of knowledge. It has been shown that modern data analysis tools allow you to evaluate the author's contribution to publication, despite the fact that of course, only the scientific community can evaluate the real contribution to scientific research.

Keywords: *comparison of scientific texts, semantic search, thesaurus for the ontology of knowledge, information query using the thesaurus, methods of authors identification*, *addressee thesaurus, secondary information, individual frequency dictionary, LibMeta*.

REFERENCES

1. *Krämer T., Momeni F., Mayr P.* Coverage of Author Identifiers in Web of Science and Scopus. – arXiv preprint arXiv:1703.01319, 2017 – arxiv.org.

2. *Frische S.* It is time for full disclosure of author contributions// Nature. 2012. P. 489.
URL: http://www.nature.com/news/it-is-time-for-full-disclosure-of-author-contributions-1.11475.3.

3. *Cozzarelli N.R.* Responsible authorship of papers in PNAS // Proceedings of the National Academy of Sciences of the United States of America. 2004. V. 101, No. 29. P. 10495.

4. MARC 21 Formats. URL: http://www.loc.gov/marc/marcdocz.html.

5. *Shrejder Yu.A.* Tezaurusy v informatike i teoreticheskoj semantike // Nauchno-tekhnicheskaya informaciya. Ser. 2. 1971. № Z. S. 21–24.

6. *Gavrilova T.A., Horoshevskij V.F.* Bazy znanij intellektual'nyh si-stem. SPb.: Piter, 2000. 384 s.

7. *Lukashevich N.V.* Tezaurusy v zadachah informacionnogo poiska. M.: Izd-vo MGU, 2011. 495 s.

8. *Muromskij A.A., Tuchkova N.P.* About ontology of the addressee in mathematical subject domain // Russian Digital Library Journal. 2018. V. 21, № 6. P. 506–533.

9. *Borisov L.A., Orlov Yu.N., Osminin K.P.* Identifikaciya avtora teksta po raspredeleniyu chastot bukvosochetanij // Preprinty IPM im. M.V. Keldysha. 2013. № 27. 26 s. URL: http://library.keldysh.ru/preprint.asp?id=2013-27.

10. TextSTAT - Simples Text Analyse Tool. URL: http://neon.niederlandistik.fu-berlin.de/textstat/.

11. *Mohsen A.M., El-Makky N.M., Ghanem N.* Author Identification Using Deep Learning, 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, 2016. P. 898–903.
URL: https://doi.org/10.1109/ICMLA.2016.0161.

12. *Manning K. D., Raghavan P., Schutze H.* Introduction to Information Retrieval. Cambridge University Press, 2018. 482 p.

13. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR, 2013.

14. *Mikolov T., Yih W.T., Zweig C.* Linguistic Regularities in Continuous Space Word Representations // Proceedings of NAACL HLT, 2013.

15. *Le Q., Mikolov T.* Distributed Representations of Sentences and Documents // International Conference on Machine Learning, 2014. P. 1188–1196.

16. *Strange K.* Authorship: Why not just toss a coin? // American Journal of Physiology-Cell Physiology. 2008. V. 295, No. 3. P. 567–575.
URL: https://doi.org/10.1152/ajpcell.00208.2008.

17. *Meli D.B.* Equivalence and Priority: Newton versus Leibniz: Including Leibniz's Unpublished Manuscripts on the Principia. Clarendon Press, 1993. 318 p.

СВЕДЕНИЯ ОБ АВТОРАХ

АТАЕВА Ольга Муратовна – научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, кандидат техн. наук, специалист в области системного программирования и баз данных.

Olga Muratovna ATAEVA – researcher of the of Dorodnicyn computing center FRC SCS RAS, PhD, expert in the field of system programming and databases.

email: oli@ultimeta.ru

СЕРЕБРЯКОВ Владимир Алексеевич – специалист в области теории формальных языков и её приложений, доктор физ.-мат. наук, профессор, зав. отделом Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН. Руководитель и участник разработки ряда известных программных проектов, в частности, ИСИР и ИСИР РАН, «Научный портал РАН».

Vladimir Alekseevich SEREBRIAKOV – expert in the field of theory of formal languages and its applications, doctor of sciences, professor, head of Dorodnicyn computing center FRC SCS RAS department. Head and participant in the development of a number of well-known program projects, in particular, ISIR and ISIR RAS, Scientific portal RAS.

email: serebr@ultimeta.ru

ТУЧКОВА Наталия Павловна – старший научный сотрудник Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН, кандидат физ.-матем. наук, окончила ВМиК МГУ им. М.В. Ломоносова. Специалист в области алгоритмических языков и информационных технологий.

Natalia Pavlovna TUCHKOVA – senior researcher of Dorodnicyn computing center FRC SCS RAS, PhD in physics with a math degree, graduated from CS Faculty of Lomonosov MSU. The expert in the field of algorithmic languages and information technologies.

email: natalia_tuchkova@mail.ru

Материал поступил в редакцию 25 ноября 2020 года