Об идентификации авторов научных работ в цифровых коллекциях

О.М. Атаева, В.А. Серебряков, Н.П. Тучкова

Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН

Аннотация. Рассматриваются особенности задачи идентификации авторов в цифровых библиографических коллекциях. Особенности проблемы недостаточной идентификации заключаются в повторах информации, двойниковании и наличии авторов с полностью совпадающими именами, самоцитировании, автоплагиате и собственно плагиате. Предлагается использовать технологию создания тезауруса адресата, как механизма накопления научной информации об авторе, который наряду с авторским указателем и другими признаками будет способствовать идентификации автора и научных работ.

Ключевые слова: способы идентификации авторов, тезаурус адресата, вторичная информация, частотный словарь индивидуума.

On Authors Identification in Digital Collections of Scientific Works

O.M. Ataeva[*][0000-0003-0367-5575], V.A. Serebryakov[**][0000-0003-1423-621X], N.P. Tuchkova[***][0000-0001-6518-5817]

Dorodnicyn Computing Center FRC CSC of RAS
[*]oli@ultimeta.ru, [**]serebr@ultimeta.ru, [***]natalia_tuchkova@mail.ru

Abstract. The peculiarities of the task of identifying authors in digital bibliographic collections are considered. The features of the problem of insufficient identification are repeats of information, twins with fully matching names, self-citation, autoplagiarism and plagiarism itself. It is proposed to use the technology of creation of the thesaurus of the addressee as a mechanism of accumulation of scientific information about the author, which along with the author's index will contribute to the identification of the author and scientific works.

Keywords: methods of identification of authors, thesaurus of the addressee, secondary information, frequency dictionary of the individual

1. Введение

С проблемами идентификации авторов в библиографических системах сталкиваются практически все известные на сегодняшний день цифровые ресурсы. Если в текущий момент в базе данных «все хорошо», то из этого не следует, что при обновлении информации не появится «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. И в результате приходится вручную формировать нужный список, несмотря 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 в статьях при передаче издательствам, тем точнее он идентифицируется, естественно. Некоторые издательства делают обязательными ссылки на индексы авторов соответствующих баз данных, с которыми эти издательства сотрудничают. Этот факт, что идентификаторы авторов сопровождают публикации, говорит о том, что другие способы идентификации авторов оказываются недостаточно надежными. Рассмотрим, что еще есть в арсенале современных информационных технологий, чтобы выяснить какому именно автору принадлежит научная работа, которую надо проиндексировать в базе библиографических данных.

2. Средства идентификации автора

2.1. Множества данных для идентификации автора

Структура научной публикации – это вполне устоявшаяся для многих отечественных и международных журналов особенность научных статей. 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 70-х годах прошлого века появилось семейство стандартов для машиночитаемой каталогизации (*Machine-Readable Cataloging*, MARC) [1] 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 и 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 структуры научных статей.

Таким образом, был сформирован список обязательных полей вторичной информации о документе «научная статья»: автор(ы), аффилиация 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 публикации 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 может быть осуществлено за счет семантических связей, которые могут быть реализованы в виде подсказок из базы данных или благодаря экспертным знаниям.

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.

2.2. Набор данных тезауруса адресата

Понятие адресата, сформулированное для удобства определения пользователей и авторов базы данных, подразумевает персону – участника информационного процесса поиска и обмена информацией. Термин «тезаурус адресата (индивидуума)» (ТА) введен в информатику Ю.А. Шрейдером [2] для представления предметной области (ПО) автора на основе понятийного запаса знаний автора. Термин связан также с представлением «знаний» в информационной системе, как «структурированной информации» [3]. Для более подробного знакомства с использованием тезаурусов в поисковых процессах и извлечении знаний можно обратиться к работе [4].

В дальнейшем проявилась важность этого представления, как основы для описания онтологии адресата (ОА) в современных базах данных [5].

Состав данных для тезауруса адресата зависит от индивидуума, можно, в частности, остановиться на следующем: 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 можно на основании метрического анализа построить *словарь-тезаурус предметной области автора*. Далее сравнивая предметные тезаурусы авторов можно более точно их идентифицировать, а также устанавливать принадлежность текста некоторому автору.

2.3. Инструменты сравнения текстов для идентификации авторов

Методы сравнения текстов для установления авторства включают частотные алгоритмы [6], контекстное сравнение [7], тематическую кластеризацию и алгоритмы глубокого анализа текстов, связанные с методами машинного обучения [8].

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.

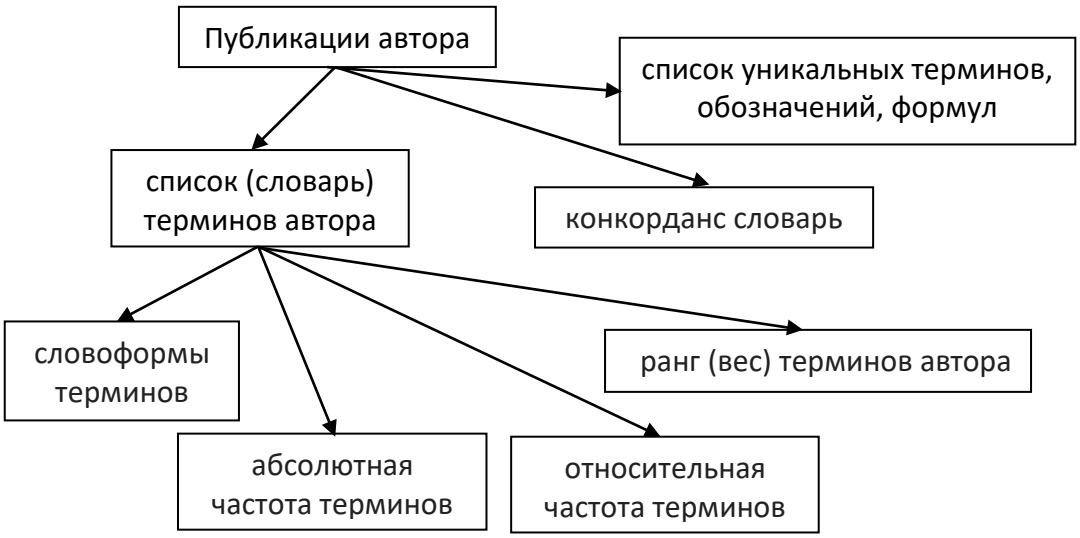

Рис. 1. Схема предварительной обработки публикаций автора

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 и выделить список уникальных терминов, обозначений, формул и других особенностей текста, характерных для некоторых авторов и предметных областей.

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, составляется 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 свойствах 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 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.

3. Пример 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.

Для выделения значимых выражений документа использовался расчет меры tf-idf для терминов документа извлеченных из индекса, с учетом морфологии [9]. На первом этапе рассматривались только существительные и термины, которые были идентифицированы как имена собственные. Далее исключались термины, для которых мера tf-idf была меньше порогового. Составление комбинаций из двух и трех слов выполнялись 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 [10, 11, 12], в режиме «skip-grams». Ниже на рисунке 2 представлена общая схема работы.

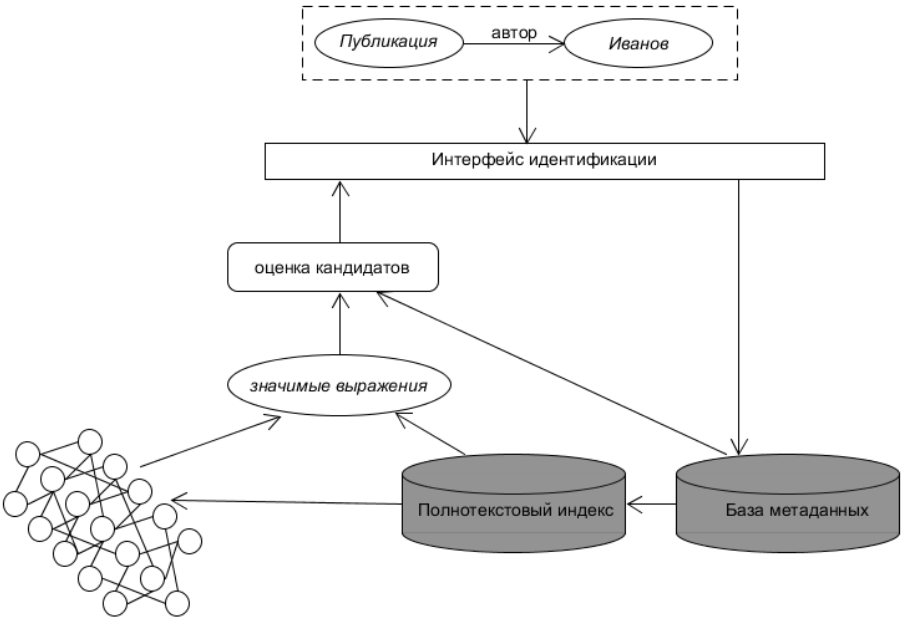

Рис. 2. Общая схема работы с терминами и авторами

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.

В качестве примера ниже приведен вариант построения тезаурусов предметных областей отдельных авторов, на основе которых можно рассуждать об их идентичности.

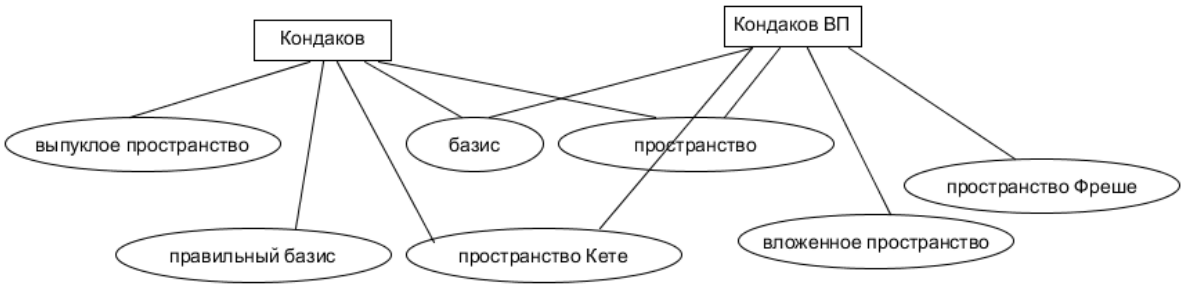

Рис. 3. Общая схема сравнения авторов

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.

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

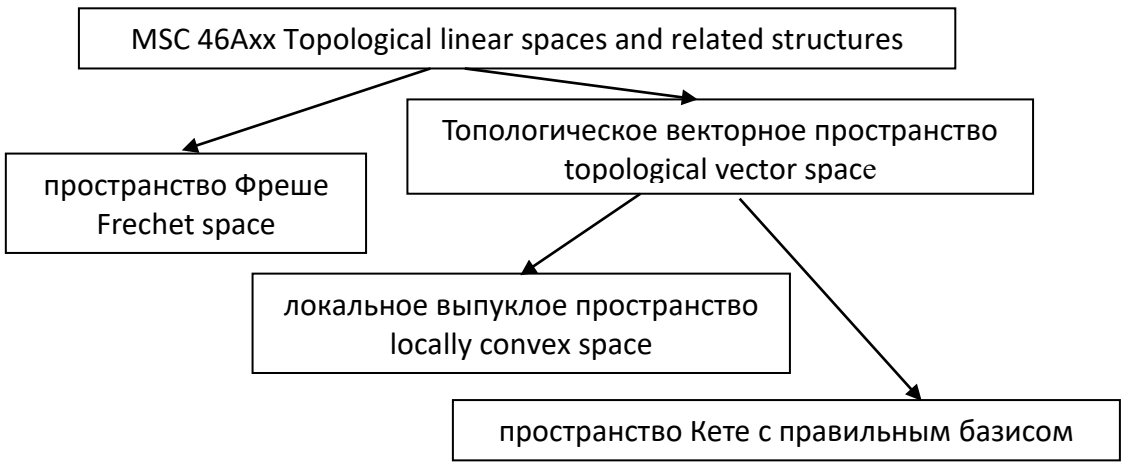

Рис. 4. Связи выявленных терминов авторов

Получаем полное пересечение множеств терминов, авторов, что позволяет идентифицировать авторов как одну персону в системе и составить тезаурус предметной области автора, причем связать термины этого тезауруса с классификаторами в системе.

4. Заключение

Предложена технология предварительной обработки публикаций для дальнейшего размещений в цифровой библиотеке. Использование данных тезауруса адресата позволяет накапливать структурированную информацию об авторах и публикациях, что способствует на предварительном этапе идентифицировать авторов.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проекты 18-00-00297комфи, 20-07-00324.

Литература

1. http://www.loc.gov/marc/marcdocz.html.
2. Шрейдер Ю.А. Тезаурусы в информатике и теоретической семантике // Научно-техническая информация. Сер. 2. 1971. № 3. С. 21–24.
3. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. СПб.: Питер, 2000. 384 с.
4. Лукашевич, Н.В. Тезаурусы в задачах информационного поиска. – М.: Издво МГУ, 2011. 495 с.
5. Муромский А.А., Тучкова Н.П. Об онтологии адресата в математической предметной области. // Электронные библиотеки, 2018, 21(6) . С. 506-533.
6. 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.
7. http://neon.niederlandistik.fu-berlin.de/textstat/.
8. Mohsen A.M., El-Makky N. M. and Ghanem N. Author Identification Using Deep Learning, 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, 2016, pp. 898-903, doi: 10.1109/ICMLA.2016.0161.
9. Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск. – 2011.
10. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR, 2013.
11. Mikolov T., Yih W.T., Zweig C. Linguistic Regularities in Continuous Space Word Representations // Proceedings of NAACL HLT, 2013.
12. Le Q., Mikolov T. Distributed Representations of Sentences and Documents // International Conference on Machine Learning, 2014. pp. 1188-1196.

References

1. http://www.loc.gov/marc/marcdocz.html.
2. Шрейдер Ю.А. Тезаурусы в информатике и теоретической семантике // Научно-техническая информация. Сер. 2. 1971. № 3. С. 21–24.
3. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. СПб.: Питер, 2000. 384 с.
4. Лукашевич, Н.В. Тезаурусы в задачах информационного поиска. – М.: Издво МГУ, 2011. 495 с.
5. Муромский А.А., Тучкова Н.П. Об онтологии адресата в математической предметной области. // Электронные библиотеки, 2018, 21(6) . С. 506-533.
6. 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.
7. http://neon.niederlandistik.fu-berlin.de/textstat/.
8. Mohsen A.M., El-Makky N. M. and Ghanem N. Author Identification Using Deep Learning, 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, 2016, pp. 898-903, doi: 10.1109/ICMLA.2016.0161.
9. Christopher D. Manning Prabhakar Raghavan Hinrich Schütze. An Introduction to Information Retrieval – 2011.
10. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR, 2013.
11. Mikolov T., Yih W.T., Zweig C. Linguistic Regularities in Continuous Space Word Representations // Proceedings of NAACL HLT, 2013.
12. Le Q., Mikolov T. Distributed Representations of Sentences and Documents // International Conference on Machine Learning, 2014. pp. 1188-1196.