З.В. Апанович

**Сопоставление данных
разноязычных ресурсов и
кросс-языковая идентификация
авторов**

Рекомендуемая форма библиографической ссылки

Апанович З.В. Сопоставление данных разноязычных ресурсов и кросс-языковая идентификация авторов // Научный сервис в сети Интернет: труды XVIII Всероссийской научной конференции (19-24 сентября 2016 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2016. — С. 36-45. — doi:[10.20948/abrau-2016-20](https://doi.org/10.20948/abrau-2016-20)

Размещена также [презентация к докладу](#)

Сопоставление данных разноязычных ресурсов и кросс-языковая идентификация авторов¹

З.В. Апанович

*Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск
apanovich@iis.nsk.su*

Аннотация. В данной работе описан алгоритм установления кросс-языковой идентичности авторов научных публикаций. Для их идентификации используется также кросс-языковое сопоставление названий организаций, в которых работают указанные авторы. Алгоритм основан на комбинированном использовании атрибутов и текстовых данных, упрощающих идентификацию персон. Продемонстрированы результаты применения этого алгоритма для сопоставления англоязычных и русскоязычных источников данных.

Ключевые слова: кросс-языковая идентификация авторов, Jaro-Winkler, tf-idf, LDA

Введение

Важным этапом пополнения одной базы знаний при помощи другой является этап установления идентичности сущностей, то есть генерация отношений вида *owl:sameAs*. В нашей работе необходимость решения этой задачи возникла при сопоставлении персонам, описанным в Открытом архиве СО РАН [1], информации об этих персонах, взятой из англоязычных семантических систем [2]. Поскольку в Открытом архиве используются русскоязычные имена персон, а в системах, с которыми мы работали, используются англоязычные имена тех же самых персон, возникла задача кросс-языковой идентификации сущностей. В процессе экспериментов с различными структурированными наборами данных, такими как RKBExplorer.com, а также при работе с полуструктурированными данными, такими как WorldCat и DBLP, обнаружилось значительное количество ошибок идентификации [2, 3]. Достаточно часто публикации нескольких разных персон рассматривались как публикации одной и той же персоны, и наоборот, публикации одной персоны распределялись между несколькими разными персонами.

Типичный пример такой ошибки можно найти по адресу <http://www.worldcat.org/wcidentities/lccn-n80162678>, где находится специальная страничка, посвященная публикациям академика Андрея Петровича Ершова.

¹ Работа выполнена при финансовой поддержке РФФИ (проект № 14-07-00386).

На этой странице публикации приводятся в хронологическом порядке, и выделяется 3 группы публикаций: публикации об А.П. Ершове, публикации А.П. Ершова и публикации А.П. Ершова, опубликованные посмертно. Система позволяет выбрать произвольный год и посмотреть публикации за этот год. Например, в списке публикаций за 2003 год, про которые сказано, что они опубликованы посмертно, обнаруживаются следующие три публикации:

- 1) Parameters of Electrode Ignition Discharges in Supersonic Propane-Air Flows. by A P Ershov; O S Surkont; I B Timofeev; V A Chernikov; V M Shibkov; N V Ardelyan; S N Chuvashov; V L Bychkov; V G Gromov; V A Levin; MOSCOW STATE UNIV (Russia) Department of PHYSICS.²
- 2) Natural Neutron-Fission Wave by A P Ershov; V F Anisichkin³
- 3) Capillary Flow of Cationic Polyelectrolyte Solutions by A P Ershov; Z M Zorin; V D Sobolev; N V Churaev⁴

При внимательном рассмотрении этих публикаций выясняется, что ни одна из них не является посмертной публикацией академика Андрея Петровича Ершова, умершего в 1988 году. В список авторов первой публикации входит Алексей Петрович Ершов из Московского университета имени Ломоносова, одним из авторов второй публикации является Александр Петрович Ершов из Института гидродинамики СО РАН (Новосибирск), а в создании третьей публикации участвовал Альберт Петрович Ершов из Института физической химии РАН (Москва).

Примеры таких неточностей в англоязычных сайтах достаточно распространены. При этом необходимо отметить, что англоязычные сайты являются важным источником информации, поэтому проблему кросс-языковой идентификации сущностей решать надо. В то время как задача идентификации сущностей для одного языка является достаточно хорошо изученной, задача кросс-языковой идентификации сущностей является сравнительно новой. Ее специфической особенностью является то, что при установлении идентичности сущностей необходимо принимать во внимание особенности построения перевода или транслитерации имен собственных с учетом особенностей того или иного языка. В данной работе представляется комбинированный подход к кросс-языковой идентификации авторов научных публикаций, и приводятся результаты работы реализованных инструментов.

² http://www.worldcat.org/title/parameters-of-electrode-ignition-discharges-in-supersonic-propane-air-flows/oclc/64436655&referer=brief_results

³ http://www.worldcat.org/title/natural-neutron-fission-wave/oclc/5649157047&referer=brief_results

⁴ http://www.worldcat.org/title/capillary-flow-of-cationic-polyelectrolyte-solutions/oclc/5649156288&referer=brief_results

1. Работы, связанные с идентификацией авторов

С решаемой проблемой связано несколько групп работ. Во-первых, существуют системы идентификации сущностей на основе сравнения атрибутов в контексте Открытых связанных данных, такие как SILK [4]. Отдельную большую группу составляют работы, ориентированные на идентификацию авторов научных публикаций [5-8]. В этой группе работ лучше всего себя зарекомендовали эвристики, использующие информацию о соавторах и о месте публикации (название конференции, журнала). Чаще всего причиной возникновения ошибок при идентификации сущностей является неполнота данных, что затрудняет сравнение различных записей по атрибутам. Поэтому заслуживают внимания методы анализа текстов на предмет установления их авторства. В настоящее время существуют достаточно продвинутые методы идентификации авторства, включающие анализ на уровне пунктуации, орфографии, синтаксиса, лексико-фразеологическом и стилистическом уровне [10-12]. При сравнении англоязычных текстов русскоязычных авторов эти методы не кажутся самыми подходящими, в силу того, что разные тексты одного и того же автора, скорее всего, переводили разные переводчики с разной манерой перевода. Поэтому мы использовали для идентификации сущностей комбинированный подход, сочетающий сравнение атрибутов публикаций и сравнение текстов публикаций.

Помимо разработки новых алгоритмов идентификации авторов публикаций, многими организациями предпринимаются усилия, направленные на облегчение процесса идентификации авторов научных публикаций. Например, неприбыльный проект ORCID⁵ (Мировой реестр уникальных идентификаторов авторов) позволяет каждому желающему (персоне или организации) получить уникальный идентификатор (Open Researcher and Contributor ID). Затем можно связать с этим идентификатором все возможные способы написания имени, публикации и места работы указанного автора.

Международный библиотечный кооператив OCLC⁶ поддерживает как общий проект нескольких национальных библиотек сервис VIAF⁷ (Виртуальный международный авторитетный файл), который объединяет несколько авторитетных файлов имен.

2. Алгоритм кросс-языковой идентификации авторов публикаций

Следует заметить, что во многих англоязычных ресурсах часто не уделяется должного внимания различным вариантам написания иностранных имен, полученным при помощи транслитерации. Мы использовали для генерации английских транслитераций русскоязычных имен две разных

⁵ <http://orcid.org/>

⁶ <http://www.oclc.org/en-europe/home.html>

⁷ <http://viaf.org/>

программы: нашу собственную программу и транслитерацию, полученную при помощи Google translate⁸.

Группа экспериментов была осуществлена с текстами электронной библиотеки SpringerLink⁹, которая была выбрана для экспериментов по нескольким причинам. Во-первых, в отличие от специализированных библиотек, она является библиотекой широкого профиля. Во-вторых, она содержит полные тексты в формате PDF для многих публикаций. Если же полные тексты публикаций не доступны, SpringerLink содержит подробную полу-структурированную информацию об издании, месте работы авторов (если таковое указано в тексте статьи), списки цитирований и др. В-третьих, каталог этой библиотеки является одним из источников, используемым глобальным ресурсом WorldCat.org.

Общая схема работы алгоритма по идентификации авторов имеет следующий вид:

- По русскоязычному имени автора генерируются все возможные варианты англоязычного написания его имени. Для генерации англоязычных имен доступны два варианта: либо наша программа генерации, либо при помощи переводчика Google translate. По всем вариантам имен осуществляется поиск статей в электронной библиотеке SpringerLink. Результаты поиска визуализируются в виде ориентированного графа.
- Если рассматриваемая персона описана в Открытом Архиве СО РАН, то оттуда извлекается информация о местах работы заданной персоны. При помощи переводчика Google translate осуществляется перевод русскоязычного названия организации на английский. У каждой найденной статьи в библиотеке SpringerLink извлекается место работы заданного автора и осуществляется нечеткое сравнение с местами работы, указанными для данной персоны в архиве СО РАН. Сравнение осуществляется на основе алгоритма Jaro-Winkler [13]. Следует отметить, что процедура сравнения названий организаций имеет достаточно сложную структуру, в силу того что вариантов написания названия одной и той же организации имеется много, включая разные варианты сокращений. В некоторых статьях место работы не указывается вообще, или указывается частично (например, СО РАН).
- Дата публикации статьи сравнивается со временем работы сотрудника в указанной организации, извлекаемым из Архива СО РАН.
- Все найденные статьи разбиваются на группы в соответствии с идентифицированным местом работы. Тексты статей, для которых место работы не указано, сравниваются со всеми статьями, размещенными по другим группам, и если сходство рассматриваемой статьи с публикациями одной из идентифицированных групп превышает пороговое значение,

⁸ translate.google.com

⁹ link.springer.com

статья помещается в эту группу. В настоящий момент для сравнения сходства имеется две возможности: при помощи метода tf-idf [14] и косинусной метрики близости, а также метода LDA (Latent Dirichlet Allocation) [15]. В случае применения метода LDA расстояние между двумя документами вычисляется при помощи дивергенции Кúльбака — Лéйблера [16]. Перед вычислением текстового сходства в текстах удаляются стоп-слова и осуществляется процедура стемминга [17].

- Для тех статей, текст которых оказался не похожим ни на одну из уже существующих групп, создается новая группа под названием NewgroupN, где N-это порядковый номер вновь создаваемой группы.
- Для каждой группы статей создается закладка, названная по одному из известных мест работы заданного автора. Граф сходства между статьями, попавшими в каждую группу, визуализируется в виде неориентированного графа.

Коллекция документов рассматривается как граф, в котором вершинами являются документы, номер вершины соответствует номеру документа в коллекции, а каждая пара документов в коллекции связана ребром, чей вес (W) соответствует сходству между двумя документами. Если величина сходства между двумя документами не превышает установленного порога, ребро между этими вершинами не создается. Пороговое значение зависит от количества вершин. Например, для коллекции из 30 вершин порог равен 0.05. Полученный граф изображается при помощи обычного силового алгоритма, так что похожие документы располагаются ближе друг к другу.

Пользователь имеет возможность получать большое количество необходимой информации, позволяющей с одной стороны, контролировать сам процесс сравнения, с другой стороны, упрощает отладку программы.

3. Некоторые результаты экспериментов

Программа идентификации авторов тестировалась на публикациях сотрудников ИСИ СО РАН, как работающих в данный момент, так и тех, кто когда-либо работал, но впоследствии либо сменил место работы, либо умер. Результаты работы программы для этой группы сотрудников сопоставлялись с данными электронной библиотеки elibrary.ru.

Удивительные результаты были получены при сравнении транслитераций имен собственных персон, сгенерированных при помощи Google translate. Выяснилось, что Google translate не только генерирует транслитерации имен, но и пытается их переводить, поэтому часто результаты транслитерации оказываются довольно неожиданными. Русская фамилия Котов, помимо стандартной транслитерации Kotov получает и дополнительные варианты, такие как Cats, Kotov has и др., а фамилия Непомнящий помимо большого количества разнообразных вариантов написания так же получает и перевод такой как Oblivious. Также, было замечено, что Google translate не порождает полное множество всех возможных транслитераций, и поэтому часть публикаций

выпадает из поля зрения. Например, для такой фамилии как «Непомнящий» Google translate выдает такие переводы как Nepomnyaschiy, Nepomnyaschy и Nepomniaschy, при этом выпадают из поля зрения такие варианты транслитерации как Nepomnyashchii и Nepomnyashchiy, несмотря на то, что на SpringerLink обнаружались такие варианты написания имени «Непомнящий».

Отрицательной стороной этого варианта является то, что варианты сгенерированных транслитераций не являются стабильным множеством и изменяются от одного прогона к другому. Более того, иногда варианты сгенерированных имен не совпадают по кодировке с теми, которые используются на англоязычных сайтах, и по этой причине публикации, которые имеются в электронной библиотеке, не обнаруживаются поиском по ключевым словам.

На рис. 1 показан фрагмент ориентированного графа самоцитирований А.С. Нариньяни, извлеченного из набора данных SpringerLink.com. Примечательно, что только две публикации из двенадцати публикаций, обнаруженных на SpringerLink.com, присутствуют в списке публикаций А.С. Нариньяни, размещенных в elibrary.ru. Полные англоязычные тексты SpringerLink обнаружены для 7 публикаций А.С. Нариньяни. Зеленым цветом выделены публикации А.С. Нариньяни, на которые есть ссылки в англоязычных публикациях, но их тексты отсутствуют в ресурсе SpringerLink.com. Следует отметить, что наша программа транслитерации сумела выделить эти публикации из SpringerLink.com, при этом были обнаружены такие варианты написания фамилии, как Narinyani и Narin'yani с различными вариантами сокращений имени и фамилии, в то время как Google translate нашла только одну публикацию с транслитерацией Narinyani. Также надо отметить, что на таком библиографическом сайте как dblp¹⁰ публикации Александра Семеновича Нариньяни распределены между тремя разными персонками с разными транслитерациями имени и фамилии.

Рис. 1. Фрагмент графа публикаций А.С. Нариньяни, извлеченного из ресурса SpringerLink.com

¹⁰ <http://dblp.uni-trier.de>

Что касается данных в библиотеке SpringerLink, следует отметить значительный разброс в объеме доступной информации о публикациях (от пары абзацев до нескольких десятков страниц), что существенно влияло на точность идентификации. Также, результаты проверки программы на тестовой выборке из 100 персон (около 3000 публикаций) показали, что примерно в 80% случаев, в публикациях не было информации о полном имени персоны, имелись только инициалы. Место работы персон были указаны примерно в 70% случаев.

На рис. 2 показан пример работы программы, ищущей по заданному имени персоны “ВА Непомнящий” публикации в электронной библиотеке SpringerLink.com. Полное имя персоны найдено в Архиве СО РАН и сгенерированы различные англоязычные варианты написания этого имени, они показаны в верхней вкладке слева. В средней вкладке слева показаны англоязычные варианты места работы заданной персоны. В центре показан граф, изображающий публикации, приписанные алгоритмом сравнения Валерию Александровичу Непомнящему, работающему в ИСИ СО РАН. Вершины белого цвета соответствуют публикациям, для которых место работы указано. Вершины желтого цвета соответствуют публикациям, у которых место работы не указано. Всего по запросу “ВА Непомнящий” в SpringerLink.com было найдено 49 публикаций.

Из них автору по имени Valery Nepomniaschy принадлежали 3 статьи, V.A. Nepomnyaschy – 1 статья, V.A. Nepomniaschy– 24 статьи, Valery A. Nepomniaschy– 3 статьи, V. A. Nepomnyashchii – 15 статей, V. A. Nepomnyashchiy – 3 статьи. Все указанные варианты написания имени, отчества и фамилии использовались Валерием Александровичем Непомнящим из ИСИ СО РАН, но помимо него было еще 2 персоны, одна из которых использовала для своего имени вариант написания V. A. Nepomnyashchii, а вторая - V. A. Nepomnyashchiy. На рис. 2 можно видеть, что программа создала две новых группы публикаций для этих персон. Наконец, интересным фактом является то, что elibrary.ru имеется четыре однофамильца Валерия Александровича Непомнящего с похожими инициалами (Виктор Афанасьевич, Виталий Александрович, Владимир Абрамович и Владимир Анатольевич), но ни у кого из них в списках публикаций на elibrary.ru нет ни одной англоязычной публикации.

Рис. 2. Публикации, идентифицированные как принадлежащие В.А. Непомнящему из ИСИ СО РАН. Статьи, в которых место работы указано, показаны более светлым цветом

На всей тестовой выборке, несмотря на неполноту входных данных, обсуждаемую выше, совместное сравнение атрибутов и текстов публикаций показали неплохую точность, близкую к 90%.

Заключение

Первоначально нами ставилась достаточно узкая задача, состоявшая в том, чтобы найти публикации авторов, упоминавшихся в Открытом Архиве СО РАН, в англоязычных электронных ресурсах. Эксперименты показали, что информация о публикациях одного и того же автора может быть рассредоточена по многим различным как англоязычным, так и русскоязычным ресурсам и для создания наиболее полной картины нужна работа по объединению информации из различных источников.

В данной версии программы был реализован подход, когда большой вес приписывался информации о месте работы персоны, и распределение публикаций по группам осуществлялось на основе этой информации. В основном, такое решение диктовалось повышением скорости работы алгоритма. Эксперименты показали, что такое решение было не совсем оправданным, поскольку обнаружились случаи, когда люди меняли место работы, и эта информация не отражалась в архиве СО РАН. В этих случаях, несмотря на то, что текстовый анализ показывал текстовое сходство публикаций из двух разных групп, публикации этих двух групп не объединялись. В настоящее время ведутся эксперименты по выбору

подходящих весовых коэффициентов, позволяющих объединить в одну функцию результаты нечеткого сравнения текстов и атрибутов. Также ведутся эксперименты по повышению качества анализа за счет сравнения научных сообществ, к которым принадлежат разные персоны, в частности, на основе сравнения различных вариантов сетей цитирования и само-цитирования.

Эксперименты также показали, что `elibrary.ru` может быть весьма полезной при идентификации персон и их публикаций за последние 10-15 лет, но при изучении публикаций персон, которые умерли, оказались за границей или сменили место работы, в ней обнаруживается неполнота данных. Стоит отметить, что в данный момент ни одна из существующих библиотек не обладает достаточной полнотой данных и отсутствием ошибок. Поэтому планируется дальнейшее развитие этого алгоритма, поскольку он может быть адаптирован для сопоставления произвольной пары русскоязычных и англоязычных источников данных.

Благодарности

Автор выражает благодарность Александру Гурьевичу Марчуку, предоставившему тестовые данные для экспериментов.

Литература

1. Марчук А.Г., Марчук П.А. Особенности построения цифровых библиотек со связанным контентом // Труды RCDL'2010, Казань, 2010. — 19-23.
2. Apanovich Z.V., Marchuk A.G. Experiments on using the LOD cloud datasets to enrich the content of a scientific knowledge base // KESW 2013, CCIS 394. — Springer Verlag Berlin Heidelberg 2013. — pp. 1-14.
3. Zinaida Apanovich, Alexander Marchuk // Experiments on Russian-English identity resolution//Proceedings of the ICADL-2015 Conference Seul, South Korea, LNCS 9469. — Springer International Publishing Switzerland 2015. — pp. 12-21.
4. Isele R., Jentzsch A., Bizer Ch. Silk Server - Adding missing Links while consuming Linked Data// 1st International Workshop on Consuming Linked Data (COLD 2010), Shanghai, November 2010.
5. Anderson A. Ferreira, Marcos André Gonçalves, Alberto H. F. Laender Disambiguating Author Names in Large Bibliographic Repositories // Internat. Conf. on Digital Libraries, New Delhi, India 2013.
6. Godby C. J., Denenberg R. Common Ground: Exploring Compatibilities Between the Linked Data Models of the Library of Congress and OCLC <http://www.oclc.org/research/publications/2015/oclcresearch-loc-linked-data-2015.html>.
7. Ley M.: DBLP - Some Lessons Learned//PVLDB 2(2), 2009. — pp.1493-1500.

8. Hickey, T. B., Toves J. A. 2014. "Managing Ambiguity // VIAF" D-Lib Magazine 20 (July/August). doi:10.1045/july2014-hickey. <http://www.dlib.org/dlib/july14/hickey/07hickey.html>.
9. Князева А. А. Автоматическое связывание документов / А. А. Князева, И. Ю. Турчановский, О. С. Колобов // Труды XIV Всерос. науч. конф. RCDL'2012, Переславль-Залесский. — Изд-во «Университет города Переславля», 2012. — С. 360-369.
10. Rogov A.A., Sidorov Yu. VI. Statistical and Information-calculating Support of the Authorship Attribution of the Literary Works. Computer Data Analysis and Modeling: Robustness and Computer Intensive Methods // Proc. of the Sixth International Conference. Vol.2: K-S. — Minsk: BSU, 2001. — pp. 187-192.
11. Хмелёв Д.В. Лигвоанализатор: Распознавание автора текста с использованием цепей А.А. Маркова. //Вестник МГУ, сер.9: филология, N 2, 2000. — С. 115-126.
12. Stamatatos, E., A survey of modern authorship attribution methods // Journal of the American Society for Information Science and Technology 60(3) 2009.
13. Cohen W. W., Ravikumar P. D., Fienberg S. E., A Comparison of String Distance Metrics for Name-Matching Tasks // IIWeb 2003. — pp. 73-78.
14. <http://www.codeproject.com/Articles/12098/Term-frequency-Inverse-document-frequency-implemen>
15. Blei D. M., Ng A., Jordan M. Latent Dirichlet allocation // Journal of Machine Learning Research (3) 2003. — pp. 993-1022.
16. Steyvers M., Griffiths T. Probabilistic Topic Models // Handbook of Latent Semantic Analysis, 2007.
17. <http://snowball.tartarus.org/>