Комбинированный подход к кросс-языковой идентификации сущностей

© З.В. Апанович © А.Г. Марчук
Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук,
Новосибирск
apanovich@iis.nsk.su mag@iis.nsk.su

Аннотация

В данной работе описаны эксперименты по установлению идентичности сущностей, происходящем в процессе использования англоязычных данных из облака LOD для пополнения контента русскоязычных научных баз данных и знаний. Один из возможных подходов состоит в комбинированном использовании структурированных и текстовых данных, которые содержат дополнительную информацию, упрощающую идентификацию персон. В качестве тестовых данных использовались данные электронной библиотеки SpringerLink и данные открытого Архива СО РАН.

1 Введение

Один из проектов, осуществляемых в Институте систем информатики Сибирского отделения Российской академии наук (ИСИ СО РАН) направлен на пополнение Открытого архива СО РАН [8] данными облака Открытых Связанных Данных (Linked Open Data, LOD) [10]. Для интеграции Связанных Данных в приложения требуется решить проблему доступа к связанным данным (1), выравнивания словарей или онтологий (2), установления идентичности сущностей (3) и фильтрации данных (4). В работе [1] подробно рассмотрены методы установления соответствия между онтологиями при помощи различных методов визуализации, а также продемонстрированы проблемы, возникающие при установлении идентичности сущностей на примере RDF-данных портала RKBExplorer.com. Было показано, что при сопоставлении русскоязычных и англоязычных баз знаний, часто одной реальной персоне ставится в соответствие нескольких виртуальных персон, и наоборот, публикации нескольких разных персон приписываются одной персоне. Один из возможных подходов к устранению этой проблемы состоит в совместном использовании структурированных и текстовых данных, которые содержат дополнительную информацию, упрощающую идентификацию персон. В данной работе будут представлены эксперименты по идентификации сущностей на примере электронной библиотеки SpringerLink (http://link.springer.com/) и Открытого архива СО РАН.

Труды XVII Международной конференции DAMDID/RCDL'2015 «Аналитика и управление данными в областях с интенсивным использованием данных», Обнинск, 13-16 октября 2015

2 Входные данные и алгоритм идентификации сущностей

Важным этапом пополнения одной базы знаний при помощи другой является этап установления идентичности сущностей, то есть, генерация отношений вида *owl:sameAs*. В нашем случае необходимо правильно сопоставить персонам, описанным в Открытом архиве СО РАН информацию про эти персоны, взятую из других семантических систем. Проблема осложняется тем, что в случае Открытого архива используются русскоязычные имена персон, а в большинстве систем, с которыми мы работали, используются англоязычные имена тех же самых персон. Конечно, может возникнуть вопрос, почему бы не воспользоваться русскоязычным источником данных, например данными научной электронной библиотеки elibrary.ru? Эта библиотека представляет персонифицированную информацию по российским исследователям, но, к сожалению, срок развития этой электронной библиотеки значительно уступает периоду времени, охватываемому Открытым архивом СО РАН. Поэтому elibrary.ru может быть весьма полезной при идентификации персон и их публикаций за последние 10-15 лет, но она становится мало полезной при изучении публикаций таких персон, как академик Андрей Петрович Ершов. Нам не удалось обнаружить в elibrary.ru информации про таких людей, как А.П. Ершов, Б.А. Трахтенброт, В.Е. Котов и многих других исследователей, заложивших основы советской и российской информатики.

Электронная библиотека SpringerLink была выбрана для экспериментов по нескольким причинам. Во-первых, в отличие от специализированных библиотек, она является библиотекой широкого профиля, что больше

соответствует содержанию Архива СО РАН. 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 в WorldCat.org [4], – проекте по созданию глобального многоязычного каталога, объединяющего как каталоги реальных библиотек, таких как Библиотека Конгресса США, так и нескольких электронных библиотек. WorldCat.org входит в облако Открытых Связанных Данных. Данные WorldCat.org доступны в RDF-формате, в качестве базовой онтологии используется расширение schema.org под названием BiblioGraph (http://BiblioGraph.net).

Эксперименты с WordlCat.org показали, что этот ресурс, также как и рассмотренные нами ранее, не свободен от ошибок идентификации русскоязычных авторов. Например, в списке сущностей WorldCat Identities, (http://www.worldcat.org/wcidentities/lccn-n80162678), содержащем описания особо выдающихся личностей, была найдена запись, посвященная академику Андрею Петровичу Ершову. Она содержала информацию о книгах и статьях академика А.П. Ершова вперемешку с публикациями другого А.П. Ершова (Александра Петровича) из Новосибирска. При этом статьи Александра Петровича, появившиеся с 1989 по 2012 год, были описаны как «посмертные произведения академика А.П. Ершова» (академик А.П. Ершов умер в 1988 году).

В случае обычных библиотек, основу идентификации сущностей составляют так называемые «авторитетные файлы», создаваемые вручную экспертами. В настоящее время ведется интеграция авторитетных файлов различных библиотек во главе с лабораторией OCLC и библиотекой Конгресса США (viaf.org), в котором, в частности, интегрируется разноязычная информация о персонах и их произведениях, организациях, и географических названиях. Несмотря на то, что для интеграции используются очень качественные данные, результаты не всегда идеальны. Так, например, персона, по имени Андрей Петрович Ершов, идентифицированная как http://viaf.org/viaf/5347110, имеет в списке работ, ему приписываемых, как книги, у которых он действительно был либо автором, либо они были выпущены под его редакцией, так и книги, к которым он вряд ли имел отношение. С другой стороны, там есть описание персоны по имени Александр Петрович Ершов (http://viaf.org/viaf/196995053), которому так же приписаны как работы под редакцией Андрея Петровича Ершова, так и работы по экономике («Регулирование доходов населения…»), которые вряд ли были написаны одним и тем же человеком. То есть, возникают ситуации, когда одной и той же персоне приписываются работы нескольких разных персон, и с другой стороны работы одного человека распределяются по разным персонам.

Известной системой для идентификации сущностей на основе сравнения значений атрибутов в контексте Открытых связанных данных является SILK[6]. Эвристики, используемые в VIAF и DBLP, описаны в работах [5, 7]. Заслуживают внимания также работы по сравнению атрибутов русскоязычных библиографических записей. [12, 13].

Чаще всего причиной возникновения ошибок при идентификации сущностей является неполнота данных, что затрудняет сравнение различных записей по атрибутам. Во многих англоязычных библиотеках часто не уделяется должного внимания различным вариантам написания иностранных имен, обусловленным транслитерацией. Частично в решении этой проблемы может помочь VIAF, хотя наши эксперименты показывают, что в случае наличия полного Имени, Отчества и Фамилии, проще сгенерировать все возможные варианты транслитерации. С другой стороны, с возрастанием уровня интеграции различных ресурсов, количество различных персон с одинаковыми или похожими фамилиями тоже возрастает, и возникает необходимость в автоматизации методов идентификации. При этом для решения проблемы идентификации (особенно когда это касается статей, а не книг) одних атрибутов, хранящихся в структурированных данных, оказывается недостаточно. Наиболее полными источниками информации являются собственно тексты публикаций, из которых возможно, не вся информация была трансформирована в полезные атрибуты. Часто такие однофамильцы работают в разных областях науки и их можно различить при помощи достаточно грубого сравнения текстов. В настоящее время существуют достаточно продвинутые методы идентификации авторства, включающие анализ на уровне пунктуации, орфографии, синтаксиса, лексико-фразеологическом и стилистическом уровне [9, 14]. 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. Группа таких экспериментов была осуществлена с текстами электронной библиотеки SpringerLink. Рассматривались только публикации, в которых заданная персона указывалась в качестве автора (не рассматривались книги, изданные «под редакцией» заданной персоны).

Общая схема работы алгоритма по идентификации сущностей имеет следующий вид:

- По русскоязычному имени автора (например, АП Ершов) выбирается элемент Архива СО РАН (Ершов, Андрей Петрович), после чего генерируются все возможные варианты англоязычного написания его имени и по всем вариантам имен осуществляется поиск статей в электронной библиотеке SpringerLink.
- Поскольку в Архиве СО РАН используются русскоязычные названия организаций, а в электронной библиотеке SpringerLink извлекаются англоязычные названия, осуществляется перевод названия организации при помощи переводчика Google (Google translator.com). 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 [3]. 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. В настоящий момент для сравнения сходства имеется две возможности: при помощи метода tf-idf [16] и косинусной метрики близости, а также метода LDA (Latent Dirichlet Allocation) [2]. В случае применение метода LDA, расстояние между двумя документами вычисляется при помощи дивергенции Кульбака — Лейблера. Перед вычислением текстового сходства, в текстах удаляются стоп слова и осуществляется процедура стемминга [15].
- 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.

Коллекция документов рассматривается как граф, в котором вершинами являются документы, а 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. Сила притяжения и сила отталкивания зависят от веса ребра и вычисляются по следующим формулам.

Сила притяжения $= Temperature * SpringForce(d) * W\ SpringForceK$;

Сила отталкивания $= Temperature * ElectricForce(d) / W * ElectricForceK$;

$SpringForce(D) = 2 * \mathbf{log}(D)$;

$ElectricForce(D) = 1 / d^2$, где - это расстояние, а W – сходство между двумя вершинами.

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 сущностей тестировалась на работах академика А.П. Ершова, список публикаций которого имеется в электронном архиве А.П. Ершова, и на публикациях сотрудников ИСИ СО РАН, работающих в данный момент. Результаты работы программы для этой группы сотрудников сопоставлялись с электронной библиотекой elibrary.ru.

Что касается данных в библиотеке SpringerLink, во-первых 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. 1 показан пример работы программы, ищущей по заданному имени персоны “А.П. Ершов” 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 академику Андрею Петровичу Ершову. Вершины белого цвета соответствуют публикациям, для которых место работы указано. Вершины желтого цвета соответствуют публикациям, у которых место работы не указано. Всего по запросу “А.П. Ершов” в SpringerLink.com было найдено 91 публикации. Из них 5 статей принадлежали автору по имени Andrei P. Ershov, одна статья – автору по имени Andrei Ershov, 84 - авторам по имени A.P. Ershov, и одна статья – автору по имени A.P. Yershov. На закладках над рисунком приведены названия мест работы авторов всех найденных статей. Из этих 91 публикаций академику А. П. Ершову реально принадлежали 21 публикации, остальные статьи принадлежали еще нескольким разным А.П. Ершовым.

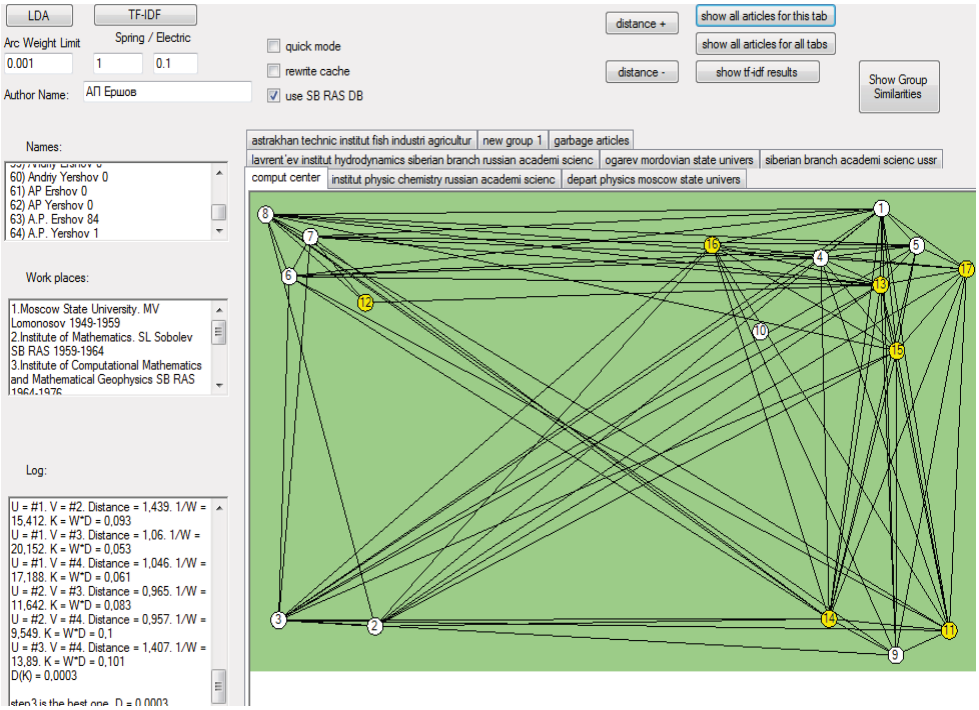

Рис. 1. Публикации, идентифицированные как принадлежащие академику А.П. Ершову. Статьи, в которых место работы указано, показаны более светлым цветом.

Программой правильно идентифицировала 19 публикаций академика Ершова, и 66 публикаций остальных А.П. Ершовых. В результате проведенных экспериментов удалось не только правильно расклассифицировать большую часть публикаций А. П. Ершова, но и обнаружить в библиотеке SpringerLink несколько публикаций академика А. П. Ершова, не отраженных в электронном архиве А.П. Ершова.

На всей тестовой выборке, в условиях существенной неполноты данных, упоминаемой выше, совместное сравнение атрибутов и текстов публикаций показали неплохую точность, близкую к 93%.

4 Заключение

В данной версии программы был реализован подход, когда большой вес приписывался информации о месте работы персоны и распределение публикаций по группам существенно зависело от 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 которому принадлежит та или иная публикация, с использованием различных вариантов сетей цитирования и само-цитирования.

Благодарности

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

Литература

[1] Apanovich Z.V., Marchuk A.G. Experiments on using the LOD cloud datasets to enrich the content of a scientific knowledge base, *P.Klinov and D.Mouromtsev (Eds.) KESW 2013, CCIS 394,* Springer Verlag Berlin Heidelberg 2013, pp. 1-14.

[2] Blei D. M., Ng A., Jordan M. Latent Dirichlet allocation *Journal of Machine Learning Research* (3) 2003 pp. 993-1022.

[3] Cohen W. W., Ravikumar P. D., Fienberg S. E.: A Comparison of String Distance Metrics for Name-Matching Tasks. IIWeb 2003, pp. 73-78.

[4] 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.

[5] Hickey, T. B., Toves J. A.. 2014. "Managing Ambiguity In VIAF" *D-Lib Magazine* 20 (July/August). doi:10.1045/july2014-hickey.http://www.dlib.org/dlib/july14/hickey/07hickey.html.

[6] 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.

[7] Ley M.: DBLP - Some Lessons Learned. *PVLDB* 2(2), 2009, pp. 1493-1500.

[8] Marchuk A.G., Marchuk P.A. Specific features of digital libraries construction with linked content. *Proc. of the RCDL'2010 Conf.* – 2010. – P. 19–23. (In Russian).

[9] Rogov A.A., Sidorov Yu. Vl. 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* (September 10-14, 2001, Minsk). Vol.2: K-S/ Edited by Prof. Dr. S. Aivazian, Prof. Dr. Yu. Kharin and Prof. Dr. H. Rieder. Minsk: BSU, 2001. – P. 187-192.

[10] Schultz A. et al. How to integrate LINKED DATA into your application //Semantic technology & Business Conference, San Francisco, June 5, 2012. http://mes-semantics.com/wp-content/uploads/2012/09/Becker-etal-LDIF SemTechSanFrancisco.pdf.

[11] Steyvers M., Griffiths T. Probabilistic Topic Models Handbook of Latent Semantic Analysis. 2007.

[12] Барахнин В. Б., Нехаева В. А., Федотов А. М. О задании меры сходства для кластеризации текстовых документов / // Вестн. НГУ. Сер.: Информ. технологии. – 2008. – Т. 6, вып. 1. – С. 3–9.

[13] Князева А. А. Автоматическое связывание документов / А. А. Князева, И. Ю.Турчановский, О. С. Колобов // Электронные библиотеки :перспективные методы и технологии, электронные коллекции (RCDL'2012) : тр. XIV Всерос. науч. конф., Переславль-Залесский, 15–18 окт. 2012 г. – Переславль-Залесский : Изд-во «Университет города Переславля», 2012. – С. 360–369.

[14] Хмелёв Д.В. Лигвоанализатор: Распознавание автора текста с использованием цепей А.А. Маркова. Вестник МГУ, сер.9: филология, N2, 2000, с.115-126.

[15] http://snowball.tartarus.org/

[16] http://www.codeproject.com/Articles/12098/Term-frequency-Inverse-document-frequency-implemen

A Combined Approach to Cross-Language Identity Resolution

Zinaida V. Apanovich, Alexander G. Marchuk

This paper describes experiments on the cross – language identity resolution problem that arises when the English-language LOD datasets are used to populate the content of a Russian scholarly knowledge base. One possible approach is the combined use of structured and text data, containing additional information and facilitating the identity resolution. The dataset of the Open Archive of the Russian Academy of Sciences and SpringerLink e-library are used as test examples.