УДК 004.91, 004.514
DOI 10.25205/1818-7900-2018-16-2-49-61

В. В. Исаченко [1], **З. В. Апанович** [1, 2]

[1] *Новосибирский государственный университет*
ул. Пирогова, 1, Новосибирск, 630090, Россия

[2] *Институт систем информатики им. А. П. Ершова СО РАН*
пр. Академика Лаврентьева, 6, Новосибирск, 630090, Россия

vv.isachenko@gmail.com, apanovich@iis.nsk.su

СИСТЕМА АНАЛИЗА И ВИЗУАЛИЗАЦИИ ДЛЯ КРОСС-ЯЗЫКОВОЙ ИДЕНТИФИКАЦИИ АВТОРОВ НАУЧНЫХ ПУБЛИКАЦИЙ

Представлена система разрешения неоднозначности авторства статей на английском языке с использованием русскоязычных источников данных. Система позволяет находить и исправлять ошибки в определении авторства научных публикаций, что может улучшить результаты поиска статей определенного автора и подсчета индекса цитируемости.

В качестве исходного хранилища публикаций использовалась база link.springer.com, для получения достоверной информации об авторах и их статьях использовалась научная электронная библиотека eLIBRARY.ru.

Система предоставляет интерактивную визуализацию результатов и возможность редактирования для повышения качества экспертного анализа. Подходы, используемые в данной системе, применимы для разрешения неоднозначности авторства публикаций из различных библиографических баз данных.

Ключевые слова: разрешение неоднозначности авторства, кросс-языковая идентификация сущностей, обработка естественного языка, интерактивная визуализация, кластеризация.

Введение

Многие научные цифровые библиотеки, такие как DBLP, PubMed, Springer и др., предоставляют функции, которые облегчают исследования целых коллекций документов. Такие системы дают доступ к миллионам библиографических записей, и на данный момент являются важнейшим источником информации для академического сообщества, так как они позволяют производить централизованный поиск публикаций.

Одной из проблем, возникающих при поиске публикаций определенного автора, является то, что такие системы не свободны от ошибок идентификации авторов. Эти ошибки могут быть двух типов: публикации двух разных персон присваиваются одной персоне или публикации одной персоны распределяются по нескольким разным персонам.

От подобного рода ошибок не свободно большинство библиографических систем, в том числе VIAF, SCOPUS и др. Например, на сайте Scopus представлено пять авторов с разными вариантами написания фамилии Непомнящий. При этом публикациям реальной персоны – В. А. Непомнящий, сотрудник ИСИ СО РАН – соответствовали публикации четырех из них, и все они имели различные идентификаторы. Помимо того, что данные ошибки затрудняют поиск статей, относящихся к определенному автору, они могут влиять на такую важную характеристику работы ученых, как индексы цитируемости. Причин возникновения ошибок

Исаченко В. В., *Апанович З. В.* Система анализа и визуализации для кросс-языковой идентификации авторов научных публикаций // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 2. С. 49–61.

© В. В. Исаченко, З. В. Апанович, 2018

достаточно много: множественные варианты транслитерации с русского языка на английский, ошибки автоматических систем по наполнению библиографических баз данных, невнимательность пользователей.

Наиболее точно решает проблему установления авторства публикаций экспертный анализ. Эксперты могут идентифицировать автора неизвестного документа или определить принадлежность произведения другому автору при помощи характерных языковых особенностей, стиля автора. Однако экспертный анализ – трудоемкий процесс, поэтому разрабатываются системы для автоматизации определения авторства документов. В таких системах применяются подходы из теории распознавания образов, математической статистики и теории вероятностей, алгоритмы нейронных сетей, кластерного анализа и др.

К сожалению, автоматическое разрешение неоднозначности не дает стопроцентной точности, и в любом случае требуется вмешательство эксперта. Задача эксперта усложняется тем, что количество документов в коллекции, для которой необходим анализ, может достигать нескольких сотен или даже тысяч. Для упрощения восприятия результатов анализа применяется интерактивная визуализация информации в виде графов, матриц смежности, диаграмм и т. п. Такое представление коллекции документов и полученных результатов значительно ускоряет процесс экспертного анализа.

Задача, в которой все публикации даны на одном языке (например, на английском) достаточно хорошо изучена: существуют решения, которые работают в условия неполноты и разнородности данных и показывают высокую точность результатов [1; 2]. Однако задача кросс-языковой идентификации сущностей (в частности, данных на английском и русском языках) является достаточно новой и требует детального изучения.

В работах [3; 4] описаны эксперименты по кросс-языковой идентификации сущностей при помощи Открытого архива СО РАН на основе исчерпывающей информации о местах работы авторов. Хотя результаты были достаточно обнадеживающими, основной проблемой был локальный характер этого архива, поскольку он касался только сотрудников СО РАН. Таким образом, возник вопрос, с каким более крупным русскоязычным источником можно провести подобные эксперименты. В качестве такого экспериментального источника данных была выбрана научная электронная библиотека eLIBRARY.ru [1], которая содержит большое количество подтвержденных записей о публикациях российских ученых.

В данной статье описана система анализа и визуализации публикаций на естественном языке для автоматизации процесса устранения неоднозначности авторства научных публикаций (рис. 1). Система производит идентификацию авторов коллекции статей на основании извлекаемых метаданных и текста публикации, а также предоставляет интерактивную визуализацию для упрощения интерпретации полученных результатов и анализа коллекции.

Постановка задачи

В качестве исходного хранилища публикаций использовалась база link.springer.com [2]. По фамилии, имени и отчеству (ФИО) на русском языке из данного хранилища извлекается коллекция статей на английском языке. Часть данных о статье, в том числе текст публикации, может отсутствовать. В качестве источника достоверной информации об авторах и их статьях использовалась Научная электронная библиотека (eLIBRARY.ru). Бóльшая часть информации в ней представлена на русском языке.

Для решения проблемы идентификации авторства коллекции документов из хранилища link.springer.com требуется:

- сопоставить статьи из хранилища link.springer.com со статьями из eLIBRARY.ru;
- произвести разделение набора научных статей, соответствующих одному или нескольким авторам, на набор непересекающихся множеств, где каждому множеству соответствует один автор данных научных публикаций;
- произвести визуализацию полученных результатов.

[1] eLIBRARY.RU – Научная электронная библиотека. URL: https://elibrary.ru/.
[2] Springer – International Publisher Science, Technology, Medicine. URL: https://link.springer.com/.

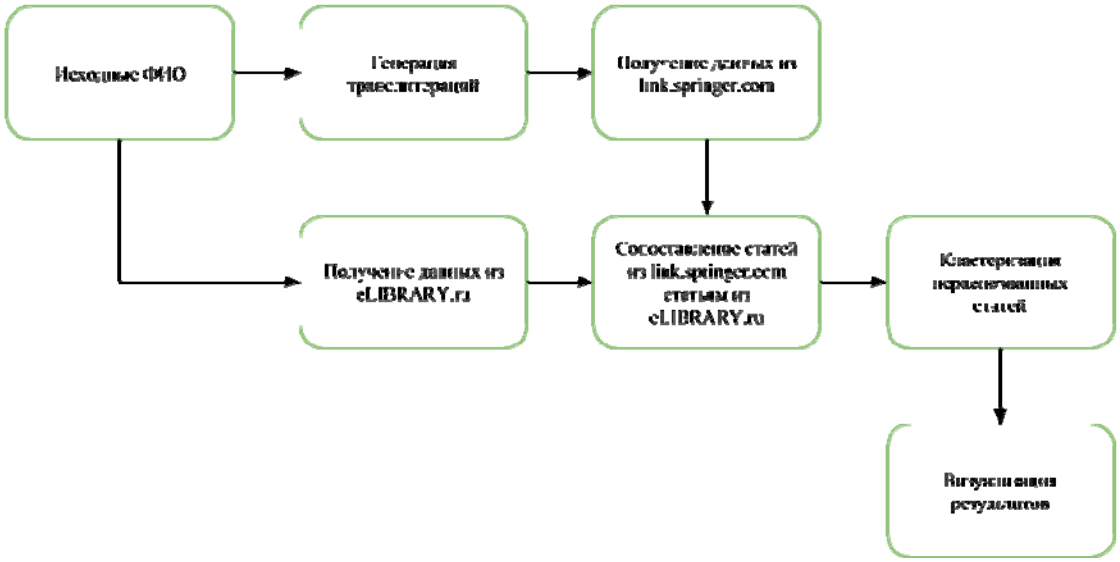

Рис. 1. Схема работы системы

Генерация транслитераций

Как известно, существует проблема неоднозначности транслитерации имен авторов с русского языка на английский язык. Как упоминалось ранее, в системе Scopus хранятся публикации В. А. Непомнящего, отнесенные к разным авторам, имена которых представлены как Nepomniaschy, V.A. Nepomnyashchii, V.A. Nepomnyaschy. Генерация всех возможных транслитераций не представляется возможной, так как реальные данные могут не подчиняться правилам транслитерации букв. Однако чем больше будет покрытие вариантов, тем больше данных будет доступно при поиске. В ранних работах использовались транслитерации, полученные лишь по одному из имеющихся стандартов или по обращению на языковые ресурсы для переводов текста, такие как translate.google.com [3]. Данные подходы покрывают малое количество вариантов транслитерации.

В текущей работе были изучены различные транслитерации букв русского алфавита, используемые в стандартах зарубежных стран и Российской Федерации (ГОСТ 7.79-2000, ГОСТ 16876-71 и пр.) [5], а также транслитерации, используемые в обиходе пользователей сети Интернет [4].

На основании выделенных транслитераций отдельных букв была реализована генерация всех возможных транслитераций имени автора на русском языке. С каждым именем сопоставляются различные варианты сокращений, так как не всегда в хранилище возможно найти статьи по полному имени автора. По всем вариантам транслитерации производится обращение к базе данных Springer. Также осуществляется обращение к eLIBRARY.ru по изначально предложенным экспертом ФИО на русском языке.

Идентификация авторства статей
путем их сопоставления с данными из eLIBRARY.ru

В результате получения входных данных исходными параметрами статей из хранилища link.springer.com являются:

- название статьи;
- список авторов;
- список мест работы для каждого автора;

[3] Google Переводчик. URL: https://translate.google.com/.
[4] ALA-LC Romanization Table for Russian. URL: http://www.loc.gov/catdir/cpso/romanization/russian.pdf

- дата публикации;
- название журнала;
- список тем, затронутых в публикации;
- список ключевых слов;
- текст публикации в формате pdf.

В результате получения входных данных исходными параметрами статей из eLIBRARY.ru являются:

- название статьи;
- список авторов;
- информация об издании, в котором была опубликована статья.

Сопоставление происходит по доступным параметрам следующим образом:

- пусть A – статья из link.springer.com, B – статья из eLIBRARY.ru;
- если название статьи A совпадает с названием статьи B полностью, без учета разделительных символов, регистра и знаков препинания, то считается, что $A = B$;
- иначе производится стемминг названий A и B, и считается коэффициент совпадения названия как доля совпадающих слов в данных названиях;
- коэффициент соавторства данных статей принимается равным доле совпадающих авторов;
- если сумма двух данных коэффициентов превышает пороговое значение, то считается, что $A = B$.

Если название публикации и список авторов указаны в eLIBRARY на русском языке, сравнить их вышеуказанным способом не получится. В таком случае в сравнении использовались результаты машинного перевода названия и списка авторов с русского языка на английский. В качестве инструмента для машинного перевода использовалась система Яндекс.Переводчик.

Иногда среди данных о публикации из link.springer.com доступна информация об издании, в котором был опубликован оригинал статьи на русском языке, включающая в себя номер выпуска, номера страниц и дату публикации. В таком случае производится сравнение этой информации с данными eLIBRARY. В результате такого сопоставления формируются группы статей, которые принадлежат одному автору, найденному в электронной библиотеке eLIBRARY.ru, а также группа статей, которые не были распознаны.

Для оценки качества сопоставления проведены эксперименты на данных сотрудников ИСИ СО РАН. В выборку были включены 25 сотрудников института, чьи публикации содержатся в системе link.springer.com. Средний процент числа публикаций авторов, распознанных системой, составил 79 %, при этом количество публикаций, которые не принадлежат автору, но были отнесены в его группу, близко к нулю. Основной причиной, по которой система не может определить принадлежность статьи ее автору, является неполнота данных. Для улучшения результатов к группе статей, которые не были распознаны, применяется алгоритм подсчета близости и группировки статей, описанный далее.

Алгоритм кластеризации статей

Алгоритм кластеризации статей, не сгруппированных на ранних этапах, заключается в попарном сравнении статей и объединении групп в случае, если коэффициент схожести статей превышает заданный порог. Более формальное описание алгоритма приведено ниже.

Пусть $A = \bigcup_{g_i} A_{g_i}$ – множество статей, полученных после сопоставления публикаций из Springer с публикациями из eLIBRARY.ru, где g_i – номер группы. При этом группа A_{g_i}, где $g_i = -1$ – группа публикаций, для которых не было найдено сопоставление. Тогда применяется следующий алгоритм:

для каждой статьи $s \in A$
    для каждой статьи $t \in A$
        $d := $ *коэффициент сходства* (s,t)
        Если $(d > threshold)$
            Если $(Group(s) = -1 \ \ и \ \ Group(t) = -1)$
                $NewGroup(s,t)$
            Иначе
                $UniteGroups(s,t)$

При объединении групп происходит проверка на то, что обе эти группы не были изначально сформированы на этапе сопоставления со статьями из eLIBRARY.ru. В данном случае объединения не происходит, так как эти группы соответствуют статьям различных авторов, указанным в eLIBRARY.ru.

Асимптотика данного алгоритма $O(N^3)$. Для улучшения данной асимптотики была применена структура данных «Система непересекающихся множеств» [6]. С ее помощью асимптотика операции объединения групп уменьшается до $O(1)$, следовательно, весь алгоритм имеет асимптотику $O(N^2)$.

Подсчет коэффициента сходства статей

Для подсчета близости научных статей из хранилища link.springer.com используются все полученные через API данные, чтобы сократить влияние неполноты данных на результаты идентификации. Сравнение каждого из параметров формирует свой коэффициент, который суммируется в итоговый.

Далее, пусть A и B – различные статьи, полученные из хранилища link.springer.com.

Сравнение названий статей:

- если название статьи A совпадает с названием статьи B полностью, без учета разделительных символов, регистра и знаков препинания, то считается, что коэффициент совпадения названий равен максимальному значению – 1.0;
- иначе производится стемминг названий A и B, и считается коэффициент совпадения названий как доля совпадающих слов в данных названиях.

Сравнение списков авторов публикаций. Коэффициент соавторства статей принимается равным доле совпадающих авторов.

Для *сравнения имен авторов* производятся следующие шаги:

- приведение пары имен к одинаковому формату (например, если одно имя является полным, а во втором отсутствует отчество автора, то из первого удаляется отчество; таким же образом обрабатывается ситуация с сокращениями имен);
- производится сравнение имен с помощью алгоритма сравнения строк.

По результату сравнения двух приведенных к одному формату имен авторов не всегда можно сразу сказать, являются ли эти строки ФИО одного и того же человека. Это обусловлено тем, что транслитерации имени одного человека могут достаточно сильно различаться, либо, наоборот, люди могут являться полными тезками. Для того чтобы уменьшить количество ошибок при сравнении, используется полученная информация о местах работы авторов. В случае если место работы совпадает, коэффициент сравнения имен авторов увеличивается, так как более вероятно, что это один и тот же человек.

Сравнение и формирование коэффициентов схожести тем и ключевых слов статей подсчитывается аналогично коэффициенту соавторства, т. е. они принимаются равными доле совпадающих терминов.

Сравнение даты публикаций. Данный коэффициент является небольшим добавочным коэффициентом и призван улучшить сопоставление документов в соответствии с гипотезой о том, что если между датами публикаций прошло не очень много времени, то вероятность

того, что они принадлежат одному автору выше, чем у тех документов, которые были приняты в печать с довольно продолжительным разрывом во времени:

- если между датами публикаций статьи *A* и статьи *B* разница менее 5 лет, то коэффициент принимается равным 0.1;
- иначе, если между датами публикаций статьи *A* и статьи *B* разница более 25 лет, то коэффициент принимается равным –0.1.

Еще один добавочный коэффициент, основанный на эвристике, – это *сравнение названия журнала*: если названия журналов статей *A* и *B* совпадают, то коэффициент принимается равным 0.1.

Подсчет коэффициента сходства текста публикаций

Для подсчета коэффициента сходства текстов на естественном языке они представляются в виде векторов в многомерном пространстве. Тогда мера близости между ними определяется как косинусное расстояние. Для улучшения качества сравнения текстов на естественном языке, а также уменьшения размерности векторного представления текстов производится их предобработка [7], в которую входят удаление стоп-слов и стемминг.

Для построения векторного представления текстов в ранних работах использовался алгоритм мешка слов (bag of words) с применением TF-IDF меры [8]. TF-IDF – статистическая мера, показывающая важность слова в контексте набора документов. Наибольший показатель будет иметь слово, которое часто встречается в документе, но редко встречается во всей коллекции.

Также были проведены эксперименты по векторизации текстов на естественном языке с применением инструмента word2vec [5]. Это программный инструмент анализа семантики естественных языков, представляющий собой технологию, которая основана на дистрибутивной семантике и векторном представлении слов. Векторное представление слов основывается на контекстной близости: близкие векторы будут иметь слова, имеющие похожий смысл. Векторные репрезентации слов, полученные в результате работы word2vec, обладают следующим свойством: смысл имеют только расстояния между векторами, а не сами векторы. При сложении векторов двух слов получается вектор слова, который показывает нечто общее между исходными. Однако увеличение количества слагаемых быстро приводит к потере какого-либо ценного результата, поэтому нельзя описать основную идею документа простой суммой векторов всех слов, которыми представлен текст.

Одним из вариантов векторного представления текста является представление, в котором каждый элемент соответствует некоторой тематике. Перечислив достаточное количество возможных тематик текста, можно посчитать количество слов в тексте, соответствующих каждой тематике, и получить семантический вектор текста – вектор, каждый элемент которого обозначает отношение данного текста к той или иной тематике.

Таким образом, для построения семантического вектора текста необходимо описать достаточное количество кластеров, отражающих тематику и стиль текста. С помощью алгоритма кластеризации все слова разбиваются на заданное число кластеров, и, если количество кластеров будет достаточно большим, можно ожидать, что каждый кластер будет указывать на достаточно узкую тематику текста, а точнее, на узкий признак тематики или стиля.

Каждое слово имеет отношение ко многим кластерам – к каким-то больше, к каким-то меньше. Поэтому вычисляется семантический вектор слова – вектор, зависящий от расстояния от слова до центра соответствующего кластера в полученном векторном пространстве. После этого, для того чтобы получить семантический вектор текста, необходимо сложить все векторы слов, которые составляют текст. Для улучшения результатов необходимо отбросить все слова-шумы, расстояние от которых до центров кластеров не превышает пороговое значение, а также нормировать полученный семантический вектор текста количеством входящих в него слов.

[5] Word2Vec. URL: https://code.google.com/archive/p/word2vec/.

Сравнение алгоритмов

Для обучения алгоритма word2vec была использована модель, построенная на части дампа сайта wikipedia.org за 2014 г. [6] Данная модель содержит приблизительно двести тысяч векторных представлений слов. На основании этой модели были произведены кластеризация на 100 кластеров с помощью алгоритма k-means, реализованного в библиотеке Accord.Net [7], и подсчет векторного представления текстов по описанному выше алгоритму.

В качестве выборки для тестирования алгоритмов векторного представления текстов были использованы тексты на естественном языке, полученные при обращении в хранилище link.springer.com по имени Быстров Александр Васильевич. В результате получено 10 документов, 2 из которых не содержали текста, поэтому сравнение было произведено по тем 8 документам, которые имели текст публикации.

Ниже представлены матрицы схожести данных текстов, построенные на основании меры TF-IDF и алгоритмов word2vec (табл. 1, 2).

Таблица 1

Матрица смежности,
полученная при сравнении текстовых данных TF-IDF мерой

1.0000	0.0101	0.0073	0.0103	0.1167	0.0084	0.0068	0.0100
0.0101	1.0000	0.0162	0.4791	0.0164	0.1977	0.0327	0.2201
0.0073	0.0162	1.0000	0.0157	0.0206	0.0120	0.0252	0.0204
0.0103	0.4791	0.0157	1.0000	0.0373	0.1679	0.0248	0.2957
0.1167	0.0164	0.0206	0.0373	1.0000	0.0113	0.0344	0.0168
0.0084	0.1977	0.0120	0.1679	0.0113	1.0000	0.0205	0.1296
0.0068	0.0327	0.0252	0.0248	0.0344	0.0205	1.0000	0.0262
0.0100	0.2201	0.0204	0.2957	0.0168	0.1296	0.0262	1.0000

Таблица 2

Матрица смежности,
полученная при сравнении текстовых данных word2vec

1.0000	0.9477	0.9211	0.9426	0.9681	0.9484	0.9448	0.9388
0.9477	1.0000	0.9258	0.9925	0.9613	0.9757	0.9630	0.9768
0.9211	0.9258	1.0000	0.9294	0.9493	0.9216	0.9118	0.9165
0.9426	0.9925	0.9294	1.0000	0.9620	0.9774	0.9633	0.9846
0.9681	0.9613	0.9493	0.9620	1.0000	0.9571	0.9594	0.9534
0.9484	0.9757	0.9216	0.9774	0.9571	1.0000	0.9659	0.9803
0.9448	0.9630	0.9118	0.9633	0.9594	0.9659	1.0000	0.9618
0.9388	0.9768	0.9165	0.9846	0.9534	0.9803	0.9618	1.0000

Как видно из таблиц, результаты, полученные на основании алгоритма word2vec, являются плохо разделимыми. Такое возможно из-за недостаточно точно обученной модели. Для использования более крупных моделей требуется больше вычислительных мощностей и времени, что неприменимо в данной системе, когда эксперту необходимо взаимодействовать с ней и изменять параметры группировки по ходу работы.

[6] Word2vec API. Pretrained models. URL: https://github.com/3Top/word2vec-api/.
[7] Accord.net framework. URL: http://accord-framework.net/.

Результаты тестирования с применением кластеризации

Добавление в систему модуля кластеризации статей, не распознанных на этапе сравнения с публикациями из eLIBRARY, позволило улучшить результат идентификации авторства статей до 92 %. Следует отметить, что получение стопроцентной точности автоматической идентификации представляется маловероятным, при этом экспертный анализ позволяет достичь гораздо более высоких результатов, но отличается высокой трудоемкостью. Таким образом, стоит признать наиболее оптимальным вариант полуавтоматической обработки данных о публикациях с целью установления авторства. При этом необходимо представлять результаты автоматической идентификации в удобном для эксперта формате, чтобы упростить и ускорить процесс экспертного анализа.

Визуализация полученных результатов

Количество документов в коллекции, для которых необходимо произвести атрибуцию, может достигать десятков, а то и сотен. Анализировать полученные результаты в виде текстовых данных затруднительно, эксперт может потратить большое количество времени. Поэтому для упрощения понимания результатов и взаимодействия пользователя с системой используется визуализация информации.

В разработанной системе пользователю предлагается рассмотрение результатов на различных уровнях. Такая методика применяется во многих системах: она позволяет взглянуть на результаты с разных сторон: например, на результаты в целом и на внутреннее представление объектов. Это также позволяет производить более тонкую настройку инструмента пользователем, поскольку он может исключить из рассмотрения ненужные признаки или выделить признаки, вносящие наибольший вклад в целевую функцию.

В главном меню пользователю предлагается ввести ФИО искомого автора и запустить программу. Также есть возможность просмотреть все генерируемые транслитерации и сокращения для данного имени на русском языке. В текстовом поле отображается текущий статус работы системы, ведется логирование всех действий.

Первый уровень – визуализация групп объектов по сущности (автору). Это позволяет сразу взглянуть на итоговые результаты и внести коррективы. В качестве визуализации предлагается круговая диаграмма, в которой каждая доля показывает выделенную алгоритмом анализа группу (рис. 2). Размер долей в круговой диаграмме прямо пропорционален количеству документов из коллекции, которые система определила в данную группу. На этом уровне пользователю предлагается просмотр краткого текстового описания группы документов, которое появляется после нажатия на долю круговой диаграммы. Также доступны тонкие настройки параметров группировки, такие как использование различных параметров в целевой функции и порог целевой функции. При изменении данных параметров система автоматически пересчитывает результаты, что добавляет визуализации интерактивный характер. Также эксперту доступно редактирование полученных результатов. В диалоге (рис. 3) можно изменять группы публикаций с помощью переноса статей из одной группы в другую. При нажатии кнопки «Показать детальнее» открывается следующий уровень визуализации – визуализация отдельной группы статей, а при нажатии кнопки «Сохранить результаты» пользователь может выбрать путь для сохранения данных, а также информации о текущем разбиении.

Следующий уровень представления – внутреннее представление сформированной группы документов (рис. 4). На этом уровне коллекция представлена в виде матрицы смежности документов, попавших в данную группу. Коэффициенты схожести отображены в виде окружностей, радиус которых зависит от веса коэффициента. При нажатии на определенную точку она выделяется красным цветом, появляется текстовое описание пары документов, а также развернутое пояснение полученного коэффициента. В случае если документ был отнесен к данной группе на этапе кросс-языковой идентификации с библиотекой eLIBRARY.ru, окружность изначально имеет зеленый цвет, а информация об авторе, указанном в eLIBRARY.ru, добавляется в краткое текстовое описание.

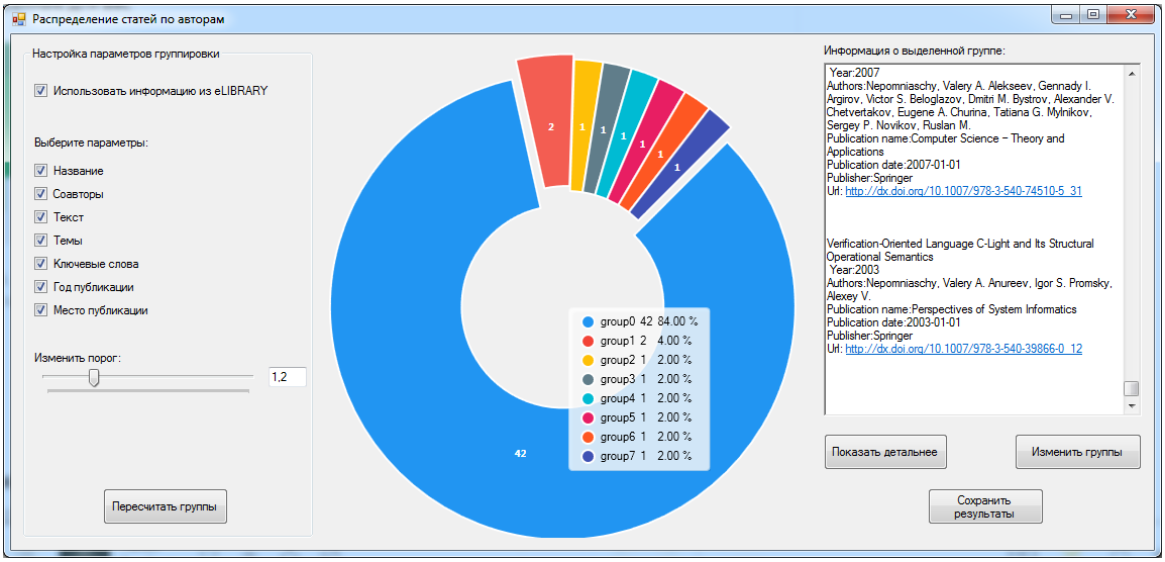

Рис. 2. Представление распределения статей по авторам

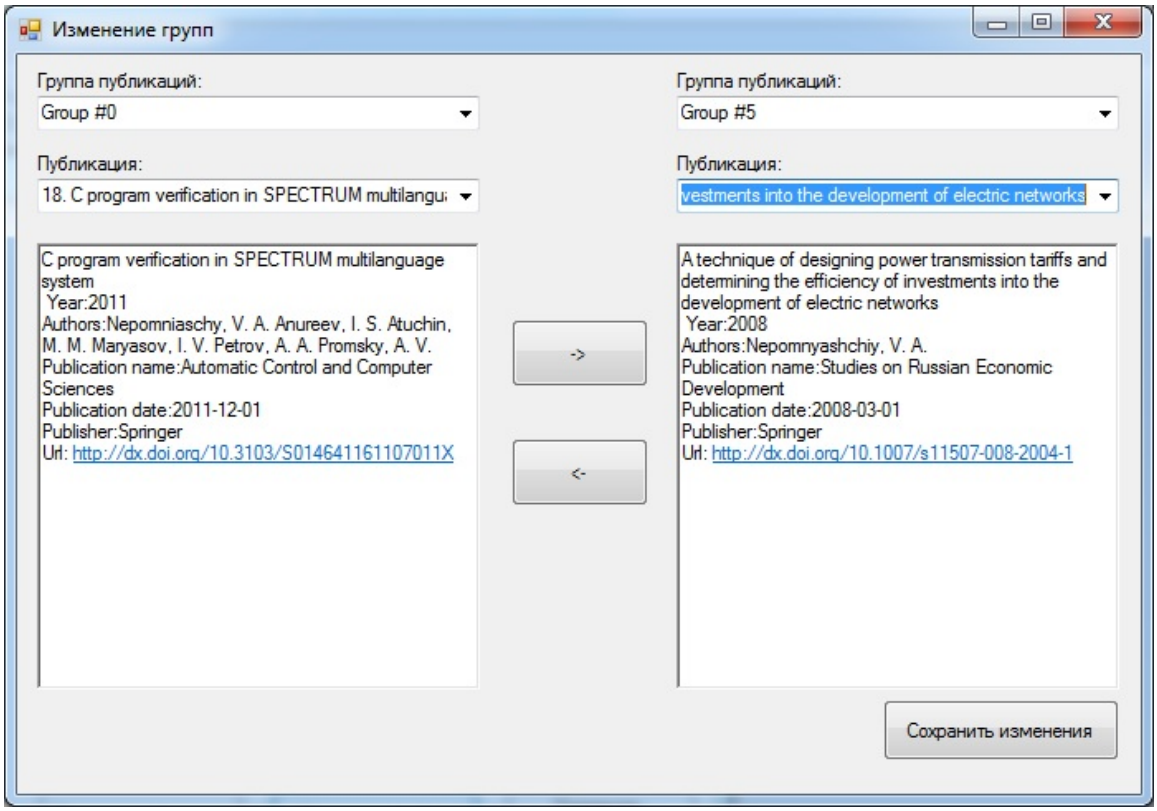

Рис. 3. Диалог для настройки полученных групп

При нажатии кнопки «Соавторство» открывается очередной уровень визуализации, представляющий соавторов научных публикаций в виде матрицы (рис. 5). Данный уровень помогает искать так называемые «выбросы» в группе – такие публикации, которые в действительности не принадлежат данному автору, в отличие от остальных. Например, эксперт точно знает группу ученых, вместе с которыми публиковался данный человек, а значит, может точно определить, что некоторые статьи в этом наборе лишние. Для этого предусмотрено выделение интересующей эксперта статьи, и по нажатию кнопки «Убрать из группы»

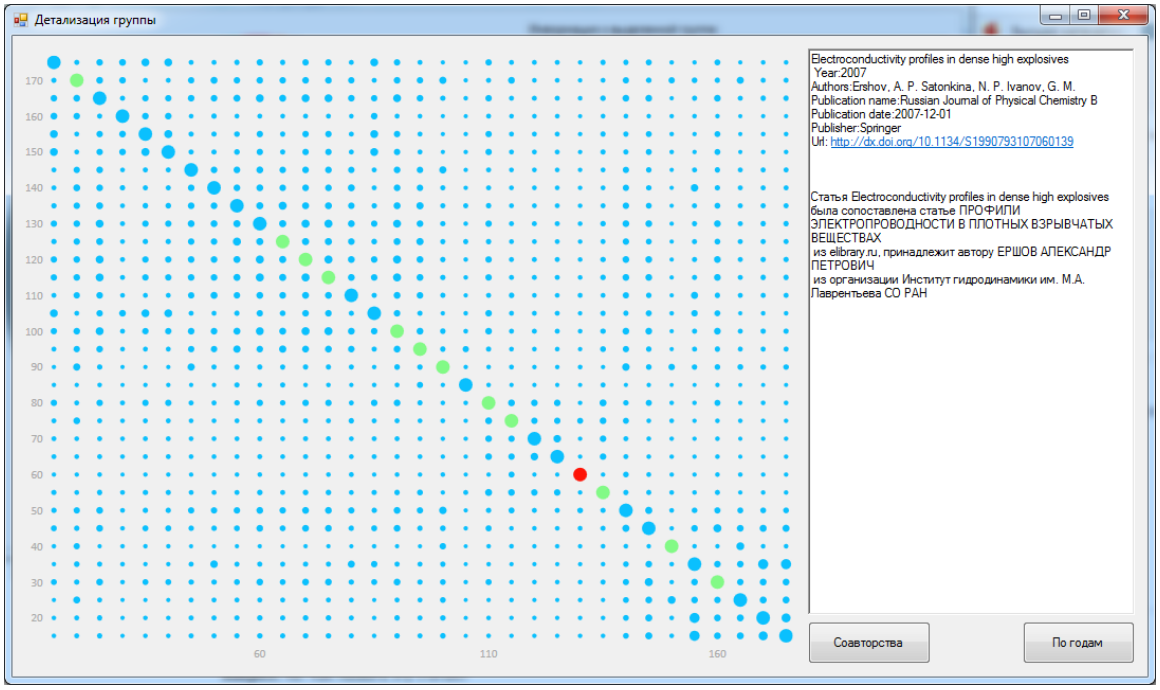

Рис. 4. Результаты в виде матрицы смежности внутри группы публикаций

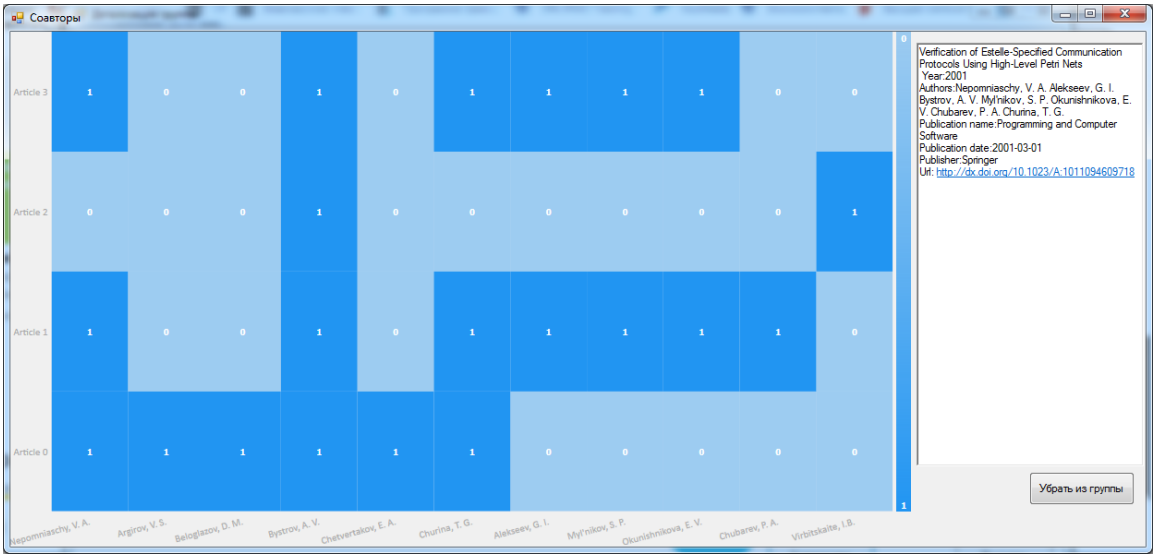

Рис. 5. Таблица соавторства в выделенной группе публикаций

текущая статья будет перемещена из группы. Система либо автоматически распределит эту публикацию в другую группу, либо создаст новую группу, содержащую эту статью.

Помимо перечисленного, пользователю системы предлагается для изучения распределение научных статей автора по году публикации (рис. 6). Оно отображается при нажатии кнопки «По годам» и также может помочь при поиске научных публикаций, не принадлежащих данному автору.

Заключение

В статье представлена система анализа и визуализации для разрешения неоднозначности авторства англоязычных статей хранилища link.springer.com при помощи сопоставления с русскоязычным источником данных elibrary.ru.

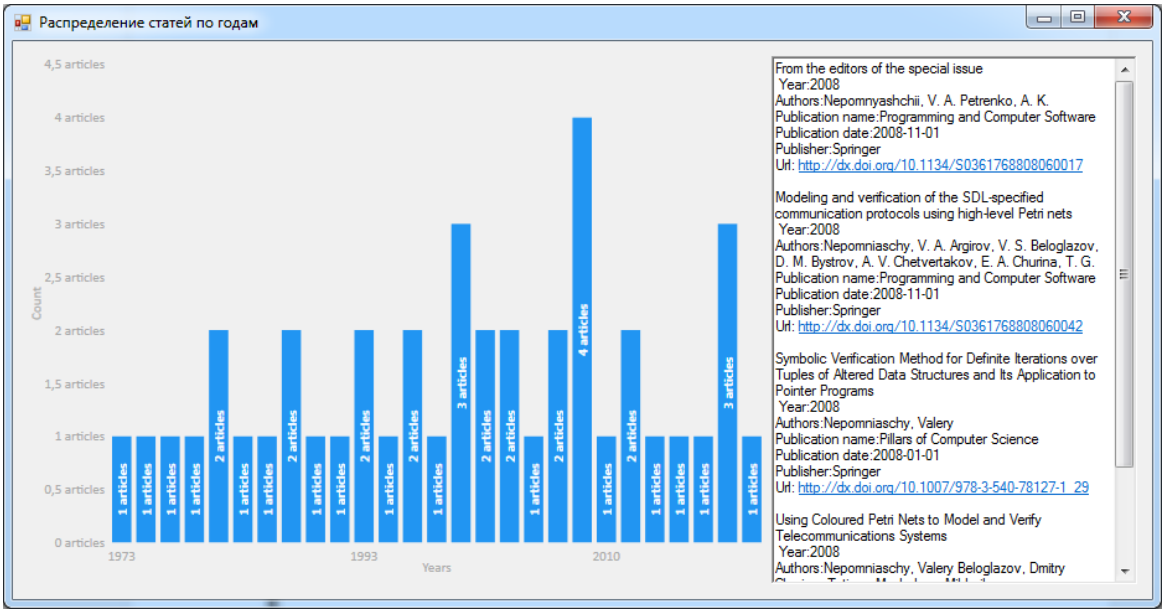

Рис. 6. Распределение статей в группе по году публикации

Реализованная система:

- генерирует множество вариантов транслитераций имен авторов;
- использует в качестве источника достоверных данных на русском и английском языках электронную библиотеку научных публикаций eLIBRARY.ru;
- на основании извлекаемых метаданных и текста публикации идентифицирует авторов исходной коллекции документов;
- показала результат распознавания 92 % (протестирована на выборке авторов из ИСИ СО РАН);
- предоставляет интерактивную визуализацию для упрощения интерпретации полученных результатов и анализа коллекции.

В дальнейшем планируется добавить дополнительные виды визуализации, помогающие не только искать выбросы в полученных группах, но и точнее настраивать алгоритм кластеризации и анализировать полученные группы, например, изменение тематики с течением времени. Также планируется расширить систему для использования различных англо- и русскоязычных баз данных, предоставляющих информацию о публикациях.

Список литературы

1. *Ferreira A. A., Gonçalves M. A., Laender A. H. F.* A brief survey of automatic methods for author name disambiguation // ACM SIGMOD Record. 2012. Vol. 41. No. 2.

2. *Shen Q., Wu T., Yang H., Wu Y., Qu H., Cui W.* Nameclarifier: A visual analytics system for author name disambiguation // IEEE Trans. Vis. Comput. Graph. 2017. Vol. 23. No. 1. P. 141–150.

3. *Apanovich Z. V., Cherepanov D. N., Marchuk A. G.* Cross-language identity resolution and approaches to its solution // Bulletin of the Novosibirsk Computing Center. Series: Computer Science. 2014. P. 41–54.

4. *Apanovich Z. V., Marchuk A. G.* Experiments on Russian-English identity resolution // Proceedings of the ICADL-2015 Conference. Seul, South Korea, LNCS 9469. Springer International Publishing, Switzerland, 2015. P. 12–21.

5. Fifth United Nations Conference on the Standardization of Geographical Names. 1987. Vol. 1: Report of the Conference. P. 40–41.

6. *Cormen T. H., Leiserson C. E., Rivest R. L., Stein C.* Introduction to Algorithms. 3rd ed. MIT Press, 2009.

7. *Большакова Е. И., Клышинский Э. С., Ландэ Д. В., Носков А. А., Пескова О. В., Ягунова Е. В.* Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: Учеб. пособие. М.: МИЭМ, 2011. 272 с.

8. *Salton G., Buckley C.* Term-weighting approaches in automatic text retrieval // Information Processing & Management. 1988. Vol. 24 (5). P. 513–523.

Материал поступил в редколлегию 04.04.2018

V. V. Isachenko [1], **Z. V. Apanovich** [1, 2]

[1] *Novosibirsk State University*
1 Pirogov Str., Novosibirsk, 630090, Russian Federation

[2] *A. P. Ershov Institute of Informatics Systems SB RAS*
6 Academician Lavrentiev Ave., Novosibirsk, 630090, Russian Federation

vv.isachenko@gmail.com, apanovich@iis.nsk.su

SYSTEM OF ANALYSIS AND VISUALIZATION FOR CROSS-LANGUAGE IDENTIFICATION OF THE AUTHORS OF SCIENTIFIC PUBLICATIONS

This paper describes a system for disambiguation of authorship of articles in English using Russian-language data sources. The system allows a user to find and correct mistakes in determining the authorship of scientific publications, which can improve the search results for articles by a certain author and calculation of the citation index.

As a source of publications, the link.springer.com database was used. To obtain reliable information about authors and their articles, the eLIBRARY digital library was used.

The system provides interactive visualization of the analysis results and editing facilities to improve the quality of expert analysis. The approaches used in this system are applicable for disambiguation of the authorship of publications from various bibliographic databases.

Keywords: authorship disambiguation, cross-language identity resolution, natural language processing, interactive visualization, clustering.

References

1. Ferreira A. A., Gonçalves M. A., Laender A. H. F. A brief survey of automatic methods for author name disambiguation. *ACM SIGMOD Record*, 2012, vol. 41, no. 2.

2. Shen Q., Wu T., Yang H., Wu Y., Qu H., Cui W. Nameclarifier: A visual analytics system for author name disambiguation. *IEEE Trans. Vis. Comput. Graph*., 2017, vol. 23, no. 1, p. 141–150.

3. Apanovich Z. V., Cherepanov D. N., Marchuk A. G. Cross-language identity resolution and approaches to its solution. *Bulletin of the Novosibirsk Computing Center. Series: Computer Science*, 2014, p. 41–54.

4. Apanovich Z. V., Marchuk A. G. Experiments on Russian-English identity resolution. *Proceedings of the ICADL-2015 Conference. Seul, South Korea, LNCS 9469*. Springer International Publishing, Switzerland, 2015, p. 12–21.

5. *Fifth United Nations Conference on the Standardization of Geographical Names*, 1987, vol. 1: Report of the Conference, p. 40–41.

6. Cormen T. H., Leiserson C. E., Rivest R. L., Stein C. Introduction to Algorithms. 3rd ed. MIT Press, 2009.

7. Bolshakova E. I., Klyshinskiy E. S., Lande D. V., Noskov A. A., Peskova O. V., Yagunova E. V. Automatic processing of texts in natural language and computer linguistics. Moscow, MIAM Press, 2011, 272 p.

8. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 1988, vol. 24 (5), p. 513–523.

For citation:

Isachenko V. V., Apanovich Z. V. System of Analysis and Visualization for Cross-Language Identification of the Authors of Scientific Publications. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 2, p. 49–61. (in Russ.)

DOI 10.25205/1818-7900-2018-16-2-49-61