Нахождение скрытых зависимостей между объектами на основе анализа больших массивов библиографических данных

Козицын Александр Сергеевич
К.ф.-м.н.
Московский Государственный Университет им. М.В. Ломоносова.
119192, Москва, Мичуринский пр., д.1, лаб. 404
alexanderkz@mail.ru

Афонин Сергей Александрович
К.ф.-м.н.
Московский Государственный Университет им. М.В. Ломоносова
119192, Москва, Мичуринский пр., д.1, лаб. 404
serg@msu.ru

Аннотация: В докладе рассматривается несколько алгоритмов выявления скрытых связей между научными публикациями, журналами и конференциями с использованием статистического анализа для обработки библиографических данных. В работе представлены результаты исследований по автоматическому нахождению переводов статей, разрешения неоднозначностей при определении авторов статей по библиографическим данным, определения степени тематической близости журналов и конференций. Представленные результаты могут использоваться для верификации собираемых наукометрическими системами данных, улучшения качества аналитической обработки наукометрических данных, для построения эргономичных интерфейсов систем сбора данных, а также для определения правил политик безопасности.

Ключевые слова: библиографические данные, скрытые зависимости, статистика.

Discovering hidden dependencies between objects based on the analysis of large arrays of bibliographic data.

Kozitsyn Aleksandr Sergeevich
Ph.D.
Московский Государственный Университет им. М.В. Ломоносова.
119192, Moscow, Michurinskii pr., 1, lab. 404
alexanderkz@mail.ru

Afonin Sergei Aleksandrovich
Ph.D.
Московский Государственный Университет им. М.В. Ломоносова
119192, Moscow, Michurinskii pr., 1, lab. 404
serg@msu.ru

Annotation: The paper discusses several algorithms for identifying hidden links between scientific publications, journals, and conferences using statistical analysis to process bibliographic data. The paper presents the results of research on automatic finding translations of articles, ambiguity resolution when identifying authors of articles on bibliographic data, determining the degree of thematic proximity of journals and conferences. The presented results can be used for verification of data collected by scientometric systems, improving the quality of analytical processing of scientometric data, for constructing ergonomic interfaces of information systems, and for defining rules for security policies.

Keywords: bibliographic data, hidden dependencies, statistics.

1 Введение

Алгоритмы автоматического выявления связей и взаимных зависимостей внутри совокупности информационных объектов широко используются для обработки и анализа больших массивов данных. Они применяются для автоматического или автоматизированного построения онтологий предметных областей, для выявления связей между структурированными данными в социальной сфере и в бизнесе, для построения прогнозов и в других областях. Несмотря на наличие общих принципов и подходов к решению таких задачи, в каждом конкретном случае требуется учитывать специфику обрабатываемых данных, описывающих семантическую зависимость между объектами.

При построении алгоритмов обработки научных публикаций необходимо учитывать, что библиографические описания статей являются частично структурированными данными. В них можно выделить как элементы структурного описания, например, год и журнал, так и полнотекстовые данные, например, название статьи. В этом случае появляется необходимость использовать алгоритмы, совмещающие лингвистические и статические методы обработки информации. Выбор конкретного алгоритма зависит от характера данных, их объемов и поставленных целей.

2 Цели выявления зависимостей

Выявление скрытых зависимостей между объектами при анализе больших объемов наукометрических данных позволяет решать задачи автоматического или автоматизированного построения онтологий, уточнения наукометрических показателей, построения дополнительных правил при реализации политик безопасности доступа к данным, поиска данных и другие подобные задачи.

2.1 Построение онтологий предметных областей и реализации политик безопасности доступа к данным

Методы автоматического и автоматизированного построения онтологий [1], [2] используются для создания семантического описания предметной области. В основе таких методов находятся алгоритмы выявления скрытых семантических зависимостей между объектами, являющимися элементами онтологий. В случае библиографических данных такими элементами могут являться публикации, авторы, доклады, журналы, сборники, конференции, издательства, тематические рубрикаторы и другие объекты. Для решения задач обеспечения безопасного доступа к данным для этих элементов должны быть определены отношения, которые могут использоваться для определения правил политики безопасности в рамках модели логического разграничения доступа ABAC (Attribute-Based Access Control) [3], которая на настоящий момент является одним из наиболее перспективных и активно развивающихся направлений исследований в области обеспечения информационной безопасности.

Традиционные модели разграничения доступа, в частности, ролевая модель RBAC, мандатная модель MAC или дискреционная модель DAC, ориентированы на применение в информационных системах с разделением обязанностей пользователей и фиксированными уровнями доступа. Преимуществом таких моделей является их простота, но они не позволяют реализовать многие естественные правила доступа, которые востребованы в современных приложениях. В частности, ролевая модель, в ее классическом виде, не позволяет выразить правило "пользователь имеет доступ к объектам, которые он создал". В дискреционной модели возможность доступа определяются на основе матрицы доступа субъект-объект, что требует определения прав для каждой пары субъект-объект.

В модели ABAC решение о предоставлении доступа зависит от значений атрибутов объектов и субъектов доступа, то есть объектов информационной системы и пользователя, запрашивающего выполнение операции. Политика безопасности представляется в виде набора правил доступа, учитывающих значения атрибутов объектов и исполь-

зующих связи между ними. В рамках модели ABAC разработано множество конкретных решений [4], как узкоспециализированных, в которых вычисление значений атрибутов информационных объектов производится на языке высокого уровня и изменение политики требует изменения исходного кода приложения, так и общих, с возможностью гибкой настройки правил и возможностью автоматической проверки их непротиворечивости [5].

Использование такого подхода требует определения атрибутов или связей между объектами. Часть отношений может задаваться явно при вводе данных в систему пользователями. Например, при регистрации в наукометрической системе статьи пользователи обычно указывают авторов статьи и журнал. Однако, значительная часть отношений не вводится в систему пользователем. Например, зависимости между публикациями, зависимости между журналами, связи между авторами, которые существуют в реальном мире, но не описываются пользователями в явном виде при вводе данных. Выявление скрытых отношений и их использование при описании политик безопасности позволяет расширить возможности для описания правил доступа к данным.

2.2 Наукометрия

В настоящее время во многих научных организациях внедряются системы автоматизации сбора данных о научной и педагогической деятельности сотрудников. Это обусловлено тем обстоятельством, что управление большими организациями науки и образования невозможно без внедрения методов оценки эффективности деятельности отдельных сотрудников и коллективов. Для проведения такой оценки необходимо использовать современные автоматизированные средства сбора, верификации, хранения и интеллектуального анализа больших объемов библиографической информации, которая описывает результаты научной деятельности его сотрудников организации, и предоставляет данные для оценки эффективности ее деятельности.

Одним из ключевых показателей в наукометрии традиционно является цитируемость публикаций. Предполагается, что ссылаясь на какую-либо публикацию, авторы подтверждают ее авторитетность и восстребованность в научном сообществе. Существует несколько общепринятых показателей, основанных на цитировании.

Импакт-фактор – один из основных общепринятых показателей качества журнала. Для расчета используются две величины: количество сделанных в текущем году ссылок на статьи в журнале за предыдущие два года и общее количество статей в журнале за два предыдущих года. Отношение этих двух величин показывает среднюю цитируемость статей в журнале и называется импакт-фактором журнала. Для определения количества ссылок анализируются статьи из фиксированного списка (около 8.5 тысяч) журналов. Ссылки из других источников не рассматриваются. Импакт-фактор журнала используется в наукометрических системах для оценки публикаций автора, и, как следствие, является важным критерием для авторов при выборе журнала для размещения публикации. Журналы с высоким импакт-фактором, как правило, имеют большее предложений от авторов, и могут предъявлять более жесткие требования к качеству статьи при рецензировании.

Индекс Хирша – оценивает публикационную активность автора. Для расчета индекса Хирша все работы автора сортируются по убыванию количеств ссылок на них. Величина индекса Хирша равна максимальному номеру статьи, для которой ее номер меньше либо равен количества ссылок на эту статью. Этот индекс имеет важное значение для авторов, поскольку указывается в различных формах при подаче заявок на получение различных грантов и прохождении конкурсных процедур. Высокий индекс Хирша руководителя и участников проекта увеличивает вероятность положительного решения по заявке и получения финансирования на проведение исследований.

Цитируемость - показатель, оценивающий качество отдельной статьи. Величина этого показателя зависит от базы данных публикаций, которая используется для расчета (Web Of Science, Scopus, Google Scolar, РИНЦ), а также от накладываемых ограничений: учет самоцитирования; учета взаимных ссылок; нормировка по количеству ссылок и других. В наукометрических системах цитируемость статей является нормирующим коэффициентом, который в значительной степени влияет на итоговый показатель по сотруднику.

Расчет всех приведенных выше показателей невозможен без правильного определения автора статьи. В библиографических данных статьи указывается ФИО автора и, в большинстве случаев, место работы. Однако, этих данных недостаточно для однозначной идентификации автора. В крупных организациях встречается большое количество совпадений ФИО. Например, в МГУ им. М.В. Ломоносова полное совпадение фамилии, имени и отчества встречается более чем у 150 Ивановых. В этой связи требуется использование алгоритмов, позволяющих правильно определять авторов работы по дополнительным признакам.

Вторым важным аспектом при подсчете цитируемости является определение переводных версий статей. Перевод статьи не является отдельной научной публикацией, поскольку не содержит никакой новой научной информации по сравнению с оригиналом, и, в ряде случаев, делается без участия автора статьи. Однако, его необходимо учитывать при оценке работы автора, поскольку, например, цитируемость перевода статьи на английский язык по базе данных публикаций Web Of Science может значительно превышать цитируемость оригинала на русском языке. Таким образом, автоматизация процесса построения связей между оригинальными статьями и их переводами позволит более объективно оценивать перечисленные выше показатели оценки научной продукции.

2.3 Выполнение поисковых запросов

Одной из основных функций любой информационной системы является обеспечение возможности выполнения поисковых запросов пользователей. В случае работы с библиографическими данными поиск должен, в первую оче-

редь, обеспечиваться по публикациям и авторам. При этом система может предоставлять пользователям разные возможности по проведению такого поиска. Наиболее простой в реализации поиск по именам, названиям и ключевым словам используется во многих информационных системах, поскольку требует только построения эффективного поискового индекса по используемым словам. Однако, такой вариант поиска не позволяет найти статью, если пользователь только приблизительно знает ее название или делает тематическую выборку статей по своему направлению. В этом случае, он не знает точных слов, которые используются в названии статьи и в наборе ключевых слов. Для реализации подобного поиска необходимо использование более сложных механизмов обработки данных, позволяющих пользователю описывать свою информационную потребность.

В первую очередь, для систем обработки библиографических данных необходимо обеспечивать возможность тематического поиска[6]. Определение тематических направлений и определение тематической близости между объектами информационного поиска позволит значительно улучшить его точность, а также повысить качество ранжирования найденных результатов при показе пользователю.

3 Определение тематической близости журналов и конференций

Задача подбора журналов, похожих на заданный журнал полезна для молодых сотрудников, которые еще не очень хорошо знакомы с библиографией предметной области. На основе одного заданного пользователем журнала автоматически формируется подборка похожих по тематическому направлению журналов, которые могут представлять интерес для пользователя как с точки зрения изучения имеющихся в них материалов, так и с точки зрения публикации собственных статей. Существует несколько подходов к решению этой задачи.

Первый подход основан на использовании тематического анализа текстовой информации: полнотекстовых описаний журналов; текстов статей, опубликованных в журналах; аннотаций и ключевых слов. На основе результатов проведения такого тематического анализа с использованием различных методов кластеризации возможно построение оценки тематической близости журналов. Однако, такой подход требует наличия в информационной системе достаточно точно описания тематических профилей всех журналов и полные тексты статей, которые часто являются недоступными из-за правовых ограничений, которые накладывают редакции журналов. Использование только ключевых слов для проведения тематического анализа не позволяет получить достаточно точное описание предметной области журнала, поскольку дает слишком общие результаты. Это обусловлено тем фактом, что во многих случаях подбор ключевых слов характеризует не тематику статьи или проекта, а их связь с одним из приоритетных направлений развития науки, технологий и техники в Российской Федерации. Например, ключевое слово «Нанотехнология» встречается в статьях совершенно различных тематических направлений: «Разработка и производство новых наноструктурированных алмазоподобных углеродных покрытий трибологического назначения»; «Разработка новой медицинской нанотехнологии для поражения раковых клеток при детских острых лимфобластных лейкозах»; «Использование радионуклидов и источников ионизирующего излучения в нанохимии, ядерной медицине и для исследования процессов, происходящих в окружающей среде»; «Разработка и создание сверхчувствительных полевых и зарядовых наноструктур для считывающих и сенсорных устройств наноэлектроники». В этой связи, использование полнотекстового подхода для проведения тематического анализа данных в наукометрических системах затруднено или невозможно.

Альтернативным подходом для проведения оценки близости журналов по тематическим направлениям является использование графа соавторства статей в анализируемых журналах. Предполагается, что многие авторы имеют определенную тематику проведения исследований и публикуют свои материалы в нескольких тематически близких журналах. Вследствие этого факта, в близких по тематике журналах часто публикуются одинаковые авторы. В отличии от методов тематического анализа, этот подход, основанный на использовании графов соавторства, не требует наличия полнотекстовой информации об опубликованных в журнале статьях, и использует только данные об авторстве статей. Такие данные могут быть получены из наукометрических или библиографических информационных систем, или систем цитирования (например, WoS).

Формально задачу оценки близости журналов можно сформулировать следующим образом. Необходимо создать функцию для отображения множества всех пар журналов на заданное множество действительных чисел. При этом значение функции для тематически близких, согласно экспертной оценке, пар журналов должно быть выше, чем у пар журналов с различной тематикой.

Разработанный авторами доклада алгоритм для решения поставленной выше задачи на первом шаге для каждой пары журналов вычисляет все пары статей, опубликованных в этих журналах одним автором. Если паре журналов соответствует только одна пара статей, то такие пары считаются не связанными. Если паре журналов соответствует несколько пар статей, то журналы считаются связанные ребром с определенным весом. В рамках настоящей работы рассматривалось несколько методов определения веса ребра. Наиболее простым методом является определение веса ребра равным количеству уникальных авторов среди соответствующих пар статей. Однако, этот метод не учитывает значимость авторов в статье. Во многих случаях основным автором статьи является один автор, фамилия которого ставится на первом месте. Остальные соавторы могут участвовать в работе над статьей незначительно, и их основное направление научной деятельности может не совпадать или не полностью совпадать с тематикой статьи.

Для проверки этого утверждения и оценки доли статей, в которых порядок авторов определяется лексикографическим порядком, а не значимостью в работе над статьей при разработке алгоритма использовался следующий метод.

Из наукометрической системы МГУ им. М.В. Ломоносова [7], [8] были отобраны для анализа все статьи в журналах за 2014-2017 гг с количеством авторов от 2 до 7. Для каждого количества авторов от 2 до 7 были посчитаны две величины: количество статей, в которых первый автор стоит в правильном лексикографическом порядке, и общее количество статей. На основе этих данных определен процент статей, для которых правильный набор авторов не случаен. Результаты расчета приведены в таблице 1.

Таблица 1 – Доля статей с неслучайным распределением авторов в правильном лексикографическом порядке

Количество авторов	Доля статей
2	0.24
3	0.16
4	0.09
5	0.06
6	0.06
7	0.03

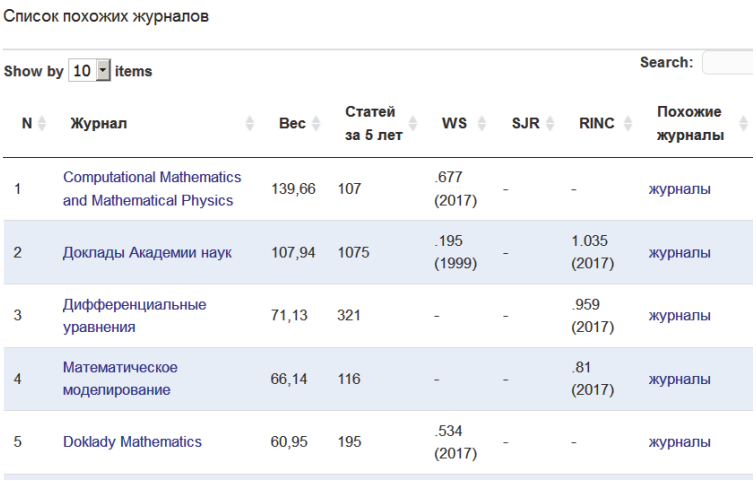
Из приведенных в таблице данных можно сделать вывод, что в большинстве случаев основным автором является автор, который в статье указан в списке первым. Для учета этого факта была разработана формула расчета веса ребер с учетом позиции автора в библиографическом описании статьи. Вес автора в каждой статье определяется следующим образом. Половина веса отдается первому автору, а оставшаяся половина веса распределяется равномерно по всем авторам (в том числе, первому).

Степень связи по заданному автору для двух журналов определяется как минимум из максимумов его весов по подмножествам статей в каждом из журналов. Таким образом, связь по автору для журнала получает большой вес только если автор имеет значимые для него статьи в обоих журналах. Величина связи пары журналов определяется как сумма связей по всем авторам.

При выборе языка для программной реализации алгоритма учитывались такие особенности алгоритма как большой объем обрабатываемых данных, необходимость быстрого доступа к хранящимся в СУБД данным, небольшие требования к объемам памяти для создания временных структур данных и отсутствие необходимости вести диалог с пользователем. Учитывая эти требования, для реализации был выбран язык PL/SQL.

Расчет тематической близости журналов производится с заданными интервалами времени и сохраняется в таблицы СУБД. Результаты промежуточных расчетов для обеспечения лучшего быстродействия сохраняются в кэширующих таблицах.

Апробации алгоритма проводилась на данных о публикациях сотрудников МГУ им. М.В. Ломоносова в наукометрической системе МГУ. Для просмотра результатов расчета реализован веб-интерфейс (Рис. 1) с использованием открытой библиотеки DataTables [9]. В информационную карточку каждого журнала добавлена ссылка для перехода к таблице со списком тематически похожих журналов. В этой таблице указываются названия близких по тематике журналов и меры сходства. Кроме того, для возможности быстрой оценки авторитетности каждого журнала из списка, в таблице приводятся данные о количестве публикаций в этом журнале за 5 лет (зарегистрированных в системе «ИСТИНА»), а также данные Web of Science и РИНЦ. С целью удобной навигации по графу близости журналов в разработанном интерфейсе также реализована возможность перехода по ссылкам на список похожих журналов непосредственно из каждого элемента списка. Средствами библиотеки DataTables для быстрого поиска по названиям журналов реализован механизм быстрой фильтрации по части названия журнала.


Список похожих журналов

Show by 10 items Search:

N	Журнал	Вес	Статей за 5 лет	WS	SJR	RINC	Похожие журналы
1	Computational Mathematics and Mathematical Physics	139,66	107	.677 (2017)	-	-	журналы
2	Доклады Академии наук	107,94	1075	.195 (1999)	-	1.035 (2017)	журналы
3	Дифференциальные уравнения	71,13	321	-	-	.959 (2017)	журналы
4	Математическое моделирование	66,14	116	-	-	.81 (2017)	журналы
5	Doklady Mathematics	60,95	195	.534 (2017)	-	-	журналы

Рисунок 1 – Интерфейс поиска тематически близких журналов

Тестирование разработанной программной реализации алгоритма проводилось по следующей методике. Из полученных результатов случайным образом было выбрано 200 пар связей журналов. Экспертами была проведена ручная оценка совпадения тематик журналов с простановкой баллов (2 – точная; 1- не совсем точная; 0 – ошибочная). Общая сумма баллов делилась на удвоенное количество анализируемых связей. Оценка точности по этой методике составила 78\%.

В качестве примера ошибок алгоритма можно привести ,например, список журналов, которые определены как близкие по тематике к изданию «Труды Высшей школы Министерства внутренних дел СССР»: «Философские науки»; «Логические исследования»; «Известия МГТУ "МАМИ"»; «Логико-философские исследования»; «Вестник Московского университета. Серия 7: Философия». Такие ошибки могут возникать в следствии слишком широкой тематической области принимаемых в журнал статей или большого количества авторов, интересующихся сразу двумя тематическими направлениями в науке.

Следует отметить, что алгоритм нечувствителен к языку журнала и подбирает похожие журналы на других языках, что сложно реализуемо для алгоритмов, основанных на анализе полнотекстовой информации. В перспективе разработанный алгоритм можно адаптировать для поиска похожих по тематике конференций, сборников публикаций и научных проектов.

4 Поиск переводов статей

Сложность определения переводных версий статей обусловлена тем фактом, что ввод в систему информации о статье и ее переводе может осуществляться не только в разное время по мере выхода изданий, но и разными авторами. В этой связи, актуальной становится задача автоматизации поиска и сопоставления переводных версий статей в процессе сбора подобной информации, поскольку ручная обработка таких объемов данных невозможна.

Задача поиска переводов статей на основе автоматического перевода их названий является очень сложной, поскольку в названиях используются многозначные слова, и необходимо при переводе учитывать специфику предметной области статьи. В таблице 2 приводится пример автоматического перевода названия статьи двумя популярными переводчиками Гугл и Промт.

Таблица 2 – Пример автоматического перевода названий

Английское название статьи	Перевод названия Промт	Перевод названия Гугл	Русское название статьи
Methods for estimating the energy of extensive air showers	Методы для оценки энергии обширных атмосферных ливней	Методы оценки энергии обширных атмосферных ливней	Методы получения оценок энергии широких атмосферных ливней
Rayleigh and Love surface waves in isotropic media with negative Poisson's ratio	Рэлей и Любовные волны поверхности в изотропических СМИ с отношением отрицательного Пуассона	Поверхностные волны Рэлея и Лайва в изотропных средах с отрицательным коэффициентом Пуассона	Поверхностные волны Релея и Лява при отрицательном коэффициенте Пуассона изотропных сред
Cubic auxetics	Кубический auxetics	Кубические аксетики	Кубические ауксетики
Soil wedge structures in the southern coast of the finland gulf	Структуры клина почвы в южном побережье залива финляндии	Почвенные клиновые сооружения на южном побережье Финского залива	Клиновидные структуры на южном берегу финского залива

Как видно из приведенной таблицы, в большинстве случаев имеется большое смысловое сходство автоматического перевода и перевода, сделанного автором, но набор слов существенно различается. Это объясняется, в первую очередь, неоднозначностью терминов в любом языке. В одних случаях в языке перевода отсутствуют полностью эквивалентные термины языка оригинала, в других - автоматическая система выбирает не совсем верные термины.

В настоящее время, в связи усилением борьбы с плагиатом, активно развивается направление поиска эквивалентных текстов на разных языках, обсуждаемых, в том числе на конференции "Обнаружение заимствований" [10]. Например, в системе "Антиплагиат" создан модуль "Переводные заимствования", который способен определять степень эквивалентности текстов, написанных на разных языках. Используемый в системе метод анализа основывается на понятии n-грамм. Элементами n-грамм являются классы эквивалентных слов, что позволяет учитывать наличие эквивалентных терминов в разных языках [11]. Такой подход эффективен для поиска переводов полных текстов, но имеет ряд существенных недостатков, которые затрудняют его использование для поиска переводных статей по названиям. Во-первых, построение классов эквивалентных слов требует настройки под каждую пару языков. В системе "Антиплагиат" используется только русско-английский перевод, а в случае перевода статей необходимо учитывать все возможные языки. Во-вторых, использование n-грамм возможно только для достаточно длинных частей текста, и плохо применимо к названиям статей.

Альтернативным подходом к автоматизации процесса поиска переводных версий статей является использование статистических данных о распределении статей по журналам. Основой разработанного авторами алгоритма является предположение, что оригинальная статья и ее перевод должны быть опубликованы одним и тем же авторским коллективом с разницей не более года в журналах на разных языках. Разработанный в рамках решения поставленной выше задачи алгоритм включат три этапа.

На первом этапе производится поиск пар журналов, которые печатают переводные статьи. Поиск производится на основе сравнения количества похожих статей, имеющих одинаковой список авторов. Для быстрого сравнения списка авторов статей используется хэш-функция, позволяющая построить индекс по всему массиву статей, загруженных в систему. В рамках принятого алгоритма для сравнения журналов были опробованы разные метрики, использующие количество статей в каждом из журналов и количество статей, имеющих одинаковый набор авторов и отличающихся по дате публикации не более чем на год. Результатом работы первого этапа алгоритма является построение двудольного графа журналов, которые печатают переводные статьи.

На втором этапе на основе построенного множества журналов производится поиск возможного перевода статьи. Поиск осуществляется среди статей, которые могут являться переводами (имеют совпадающее множество авторов, и дата публикации отличается не более чем на год) и опубликованы в журналах, связанных ребром в построенном ранее графе журналов.

На заключительном третьем этапе для англоязычных переводов производится проверка степени соответствия между названиями статей на основе сравнения переводов. Поскольку автоматический перевод названия статьи может значительно отличаться от авторского варианта перевода, для определения степени соответствия названий используется сравнение английского названия с множеством всех возможных переводов слов русскоязычного названия.

Следует отметить, что третий этап алгоритма используется только для уточнения результата работы первых двух этапов, и, в целом, алгоритм может использоваться для поиска переводных версий статей между любыми парами языков без использования словарей.

Для апробации алгоритма использовались данные о публикациях сотрудников МГУ им. М.В. Ломоносова. Авторами статьи разработан модуль, добавленный в функционал нукометрической системы организации [8].

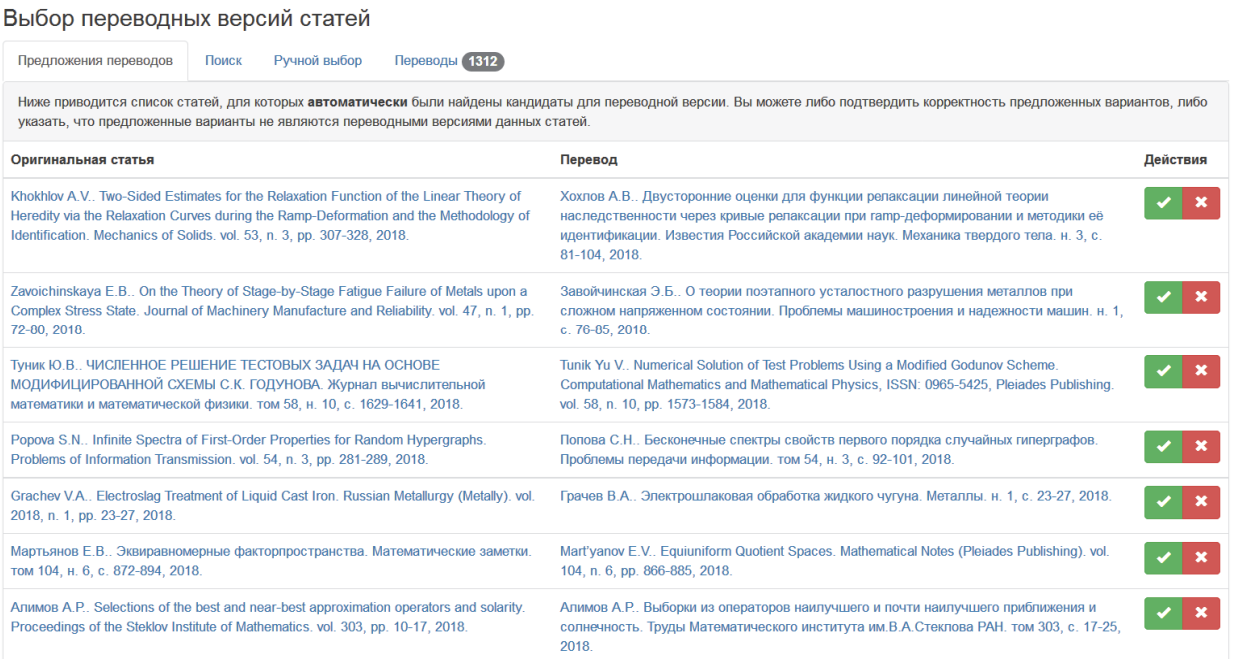


Рисунок 2 – Интерфейс подтверждения переводов статей

Разработанный для этих целей интерфейс (Рис. 2) позволяет экспертам проводить оценку результатов работы модуля и отмечать в системе правильные и ошибочные варианты предлагаемых переводов.

5 Разрешение неоднозначностей для авторов с одинаковыми фамилиями

Одной из задач, решаемых в рамках автоматизации сбора библиографических данных о публикациях, является определение автора по указанным в публикации фамилии и инициалам. Если для редких фамилий поиск соответствия между указанным в статье автором и зарегистрированным пользователем в системе этот вопрос решается простым поиском совпадения строк, то для распространенных фамилий необходимо проводить более сложный анализ. Например, среди 90 тысяч пользователей большой наукометрической системы МГУ им. М.В. Ломоносова около тысячи имеет фамилии "Иванов", "Кузнецов", "Смирнов", "Петров", "Попов". И 50 тысяч сотрудников имеют более 10 однофамильцев. Кроме того, следует учитывать большое количество соавторов, которые не зарегистрированы, но тоже должны правильно распознаваться системой. Доля зарегистрированных пользователей среди всего списка соавторов составляет менее 30%.

Точность решения задачи распознавания можно увеличивать на основе использования дополнительной информации об устойчивых группах соавторов. Один из таких методов, основанный на поиске максимально связанных подграфов в графе соавторства, описан в работах [12], [13]. Метод достаточно эффективен, однако имеет ряд недостатков. Во-первых, он имеет достаточно большую вычислительную сложность, во-вторых, не использует информацию об авторизации пользователя. Последний аспект особенно важен, поскольку 93% публикаций вносится одним из соавторов работ. Этот факт необходимо использовать для уточнения распознавания авторов.

Разработанный алгоритм на первом шаге выделяет список возможных авторов для каждой фамилии, упоминающейся в библиографическом описании статьи. Поиск происходит по совпадению фамилии и инициалов. При этом для каждого автора учитываются все варианты написания его фамилии и инициалов, встречавшихся ранее. Такой анализ необходим для работы со статьями, изданными на других языках. Если среди возможных авторов встречается авторизованный в настоящий момент пользователь, то он считается определенным, и остальные возможные авторы из списка для этой фамилии удаляются.

Далее производится сортировка по количеству вариантов для каждой фамилии и, начиная с наименее частотных фамилий, для пар фамилий осуществляется оценка вероятности соавторства для каждой пары авторов, соответствующих этим фамилиям. Для каждого последующего ФИО выбирается автор с наилучшим ребром связи с предыдущими. Если остаются нераспознанные, то определяется лучшее ребро для каждой пары вариантов.

Тестирование алгоритма проводилось на графе соавторства, имеющего около 226 тысяч вершин (авторов) и 5 миллионов ребер. Для построения графа соавторства использовалась информация из статей, тезисов, книг и проектов.

В ходе тестирования, было обработано 540 тысяч статей. Совпадение результатов расчета алгоритма с реальными данными составило 520 тысяч записей.

Дальнейшее улучшение результатов возможно за счет использования двудольного графа (пользователь-автор), позволяющего учесть факт регулярного ввода информации о публикации нескольких авторов одним пользователем. Например, в случае ввода информации о всех сотрудниках кафедры ученым секретарем кафедры. Также, перспективным является использование весовой функции при определении возможных авторов на первом шаге алгоритма. Однако, часть ошибок являются неустранимыми. Например, в случае смены фамилии автором, или неправильного ее указания.

Недостатком алгоритма является невозможность его использования для статей, написанных одним автором. Однако, этот недостаток частично компенсируется тем обстоятельством, что такие статьи авторы обычно вводят самостоятельно и авторство однозначно определяется по авторизованному пользователю.

Разработанные алгоритмы позволяют упрощать ввод данных в систему, подсказывая сотруднику правильный вариант, повышают точность вносимых данных, а также позволяют анализировать внесенные данные на наличие потенциальных ошибок.

6 Заключение

Описанные в настоящей работе алгоритмы оценки тематической близости журналов, определения переводов статей и идентификации авторов статьи по библиографическим данным позволяют в автоматическом режиме строить дополнительные связи между объектами в наукометрических или библиографических информационных системах для решения задач уточнения наукометрических данных, улучшения результатов работы информационного поиска и построения более широкого набора правил определения прав доступа к данным в логических моделях разграничения доступа ABAC. Программные реализации алгоритмов были апробированы на данных наукометрической системы МГУ им. М.В. Ломоносова.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта №18-07-01055

Список использованной литературы

[1] Платонов А.В., Полещук Е.А., МЕТОДЫ АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ ОНТОЛОГИЙ//Программные продукты и системы. 2016. № 2. С. 47-52.

[2] Бубарева О.А., Исследование механизмов автоматического построения онтологий над множеством неструктурированных данных//измерения, автоматизация и моделирование в промышленности и научных исследованиях (ИАМП-2018) Материалы XIII Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых с международным участием. 2018, С.477-481

[3] Jin, X., Krishnan, R., and Sandhu, R. A untied attribute-based access control model covering DAC, MAC and RBAC. In Data and Applications Security and Privacy XXVI, Lecture Notes in Computer Science, vol 7371, 2012, pp. 41-55.

[4] Servos D., Osborn S. L.. Current Research and Open Problems in Attribute-Based Access Control // ACM Comput. Surv., 49(4), 2017, pp. 65:1-65:45.

[5] Afonin S. Ontology models for access control systems. In 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC), pages 1-6, 2018.

[6] Васенин В. А., Афонин С. А., Козицын А. С. Автоматизированная система тематического анализа информации // Информационные технологии - приложение. 2009. № 4. С. 1-32.

[7] В. А. Садовничий, В. А. Васенин. Интеллектуальная система тематического исследования наукометрических данных: предпосылки создания и методология разработки. Часть 1. Программная инженерия, 9(2):51-58, 2018.

[8] Наукометрическая система МГУ им. М.В. Ломоносова. - URL: http://istina.msu.ru.

[9] Библиотека datatables. – URL: https://datatables.net/

[10] Научная Конференция "Обнаружение заимствований - 2017". - URL: http://www.oz2017.ru .

[11] Плагиат в научных статьях: трудности обнаружения перевода. - URL: http://ai-news.ru/2018/01/plagiat_v_nauchnyh_statyah_trudnosti_obnaruzheniya_perevoda.html.

[12] Афонин С.А., Гаспарянц А.Э., Автоматическое построение функции оценки качества в задаче разрешения неоднозначности имен авторов научных публикаций. Программная инженерия. 2015. № 10. С. 31-37.

[13] Афонин С.А., Гаспарянц А.Э., Разрешение неоднозначности авторства публикаций при автоматической обработке библиографических данных. Программная инженерия. 2014. № 1. С. 25-29.