Поиск переводов статей с использованием статистических данных

А.С Козицын, С.А. Афонин, А.А. Зензинов

НИИ механики МГУ, Москва

Аннотация. В настоящее время происходит активное внедрение наукометрических систем для автоматизации процесса анализа эффективности деятельности научных организаций с целью применения различных методов стимулирования научной деятельности. Одним из наиболее важных индикаторов является количество публикаций и их цитируемость. Для оценки этого показателя необходимы средства автоматизированного построении связей между оригинальными статьями и их переводами. В настоящей работе анализируются существующие методы оценки близости оригинального текста и его возможного перевода, показывается их недостаточная эффективность для построения связей между статьями и описывается разработанный авторами метод автоматического поиска переводов статей в больших коллекциях библиографических данных. Особенностью разработанного алгоритма является использование статистических данных о публикации статей в различных журналах и информации о соавторах анализируемых статей. Представленный в настоящей работе алгоритм позволяет осуществлять поиск переводов статей без предварительной настройки на заданные пары языков оригинала и перевода статьи, а также не требует использования больших коллекций обучающих выборок. Апробация программной реализации алгоритма проводилась в наукометрической системе МГУ им. Ломоносова. Результаты тестирования показали ее достаточную эффективность и возможность использования разработанного алгоритма для автоматического построения рекомендаций пользователям для отметки в системе переводных версий статей.

Ключевые слова: библиографические данные, автоматический перевод, статья, граф соавторства.

Linking translated articles using authorship statistics

A.S Kozitsyn, S.A. Afonin, A.A. Zenzinov

Institute of mechanics Lomonosov Moscow state university, Moscow

Abstract. During the last decades scientometric techniques have been used for research activity stimulation. Number of published articles and number of their citation counts are among the most important scientometric parameters. In an automated environment, when the publications metadata is gathered from various

sources, correct linking of original papers with their translations into different languages is extremely important. In the paper we show that the known text similarity measures are inefficient in the context of article linkage problem. We propose a method for semi-automatic article linkage using statistical data on authors publication activities only. This approach may be used for linking articles without training for the language of translation. The method was evaluated on real-world collection of publications metadata of ISTINA information system.

Keywords: bibliographic data, automatic translation, article, co-author graph.

Использование наукометрических систем для управления большими научно-образовательными организациями является необходимым условием для обеспечения возможности эффективного управления[1]. Такие системы позволяют строить агрегированные оценки по различным показателям эффективности научной и педагогической деятельности сотрудников организации для принятия управленческих решений. Наборы используемых для анализа показателей в различных системах могут отличаться и зависят от сферы деятельности организации. Однако, вне зависимости от конкретного вида обрабатываемых данных, необходимыми элементами любой системы обработки наукометрических данных являются механизмы верификации собираемой системой информации, которые включают в себя проверку полноты и точности предоставляемых данных.

Одним из важнейших показателей, который, как правило, описывает эффективность научной деятельности сотрудников организации, является количество публикаций и их цитируемость. На основе этого показателя оценивается как индивидуальная деятельность сотрудников организации, так и эффективность деятельности отдельных научных коллективов и организации в целом. Например, при подаче заявок на различные конкурсы требуется предоставление информации по имеющимся у заявителя публикациям для оценки квалификации заявителя, агрегированные данные по публикациям должны предоставляться в отчетах в вышестоящие инстанции, проведение внутренних конкурсов при замещении вакантных должностей также требует оценки квалификации сотрудников с использованием этого показателя. Для построения более объективных оценок при анализе публикаций необходимо учитывать, что авторы вводят как оригинальные статьи, так и их переводы в иностранных журналах. Переводы статей позволяют собирать дополнительную информацию о цитируемости автора, в том числе в метриках Web Of Science и Scopus, однако не могут учитываться как самостоятельные статьи при подсчете общего количества статей автора за период.

Наиболее простым техническим решением является предоставление возможности пользователю указать наличие перевода статьи при регистрации ее в системе. Однако, опыт эксплуатации подобных систем показывает, что пользователи забывают вносить подобную информацию, если интерфейс добавления данных не дает соответствующих подсказок или указаний.

Поскольку ввод данных о статье и ее переводе в наукометрическую систему может осуществляться в разное время и разными пользователями, необходима разработка алгоритмов, которые на этапе предварительной верификации данных производили поиск возможных связей статей и показывали рекомендации пользователю, а также могли производить автоматический поиск возможных переводов в уже сформированном массиве статей.

Задача автоматического перевода названий статей является очень трудоемкой, поскольку в названиях используются многозначные слова, и необходимо при переводе учитывать специфику предметной области статьи. В таблице 1 приводится пример автоматического перевода названия статьи двумя популярными переводчиками Гугл [2] и Промт [3].

Английское название статьи	Перевод названия Промт	Перевод названия Гугл	Русское название статьи
Self-Purification of Agrosoddy-Podzolic Sandy Loamy Soils Fertilized with Sewage Sludge	Самоочищение песчаных песчаных суглинковых почв, оплодотворенных осадками сточных вод	Самоочищение Агрозодди-Подзолик Сэнди глинистые почвы, оплодотворенные с отстоем сточных вод	Степень самоочищения агродерново-подзолистых супесчаных почв, удобренных осадком сточных вод
Poynting’s effect of cylindrically anisotropic nano/microtubes	Эффект Пойнтинга цилиндрически анизотропного нано/микротруб	Влияние Пойнтинга на цилиндрические анизотропные нано / микротрубки	Эффект Пойнтинга для цилиндрически-анизотропных нано/микротрубок
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

Calculating lateral distribution functions of the Cherenkov light from extensive atmospheric showers in terms of a multilevel scheme	Вычисление боковых функций распределения Излучения Черенкова от обширных атмосферных душей с точки зрения многоуровневой схемы	Вычисление боковых функций распределения черенковского света из обширных атмосферных ливней в терминах многоуровневой схемы	Расчет функций пространственного распределения черенковского света ШАЛ в рамках многоуровневой схемы
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» [4]. 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 [5]. 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. Такой подход позволяет находить возможные переводы, основываясь только на структуре связей в графе соавторства статей, не требуя использования статистической информации о языке оригинала и перевода статьи.

Основой разработанного авторами доклада 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.

После построения пар статей, которые могут являться переводами, производится построения двудольного графа журналов, которые печатают переводные статьи. Метрика для оценки степени связи журналов в графе строится на основе мощности множеств статей в каждом из журналов и мощности множества пар статей в этих журналах, которые могут являться переводами. Результатом работы этого этапа алгоритма является множество пар журналов, в одном из которых часто печатаются переводные статьи из второго журнала. В процессе работы системы граф связей журналов уточняется на основе вносимых пользователями данных о своих статьях. Для этого используется как явное указание пользователями связей между оригинальной и переводной статьей, так и информация о DOI статьи, задаваемых авторами. Многие авторы указывают библиографические данные оригинала статьи в русскоязычном журнале, внося DOI переводной версии для учета ссылок из Web Of Science. Таким образом, собрав из внешних источников информацию о статье по DOI можно точно определить название переводного журнала для указанного в статье русскоязычного журнала.

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.

Следует отметить, что алгоритм может использоваться как для обработки полной коллекции статей, так и для обработки статей, вносимых в наукометрическую информационную систему авторами непосредственно в момент их добавления. В последнем случае, одним из требований является достаточная производительность реализации алгоритма, позволяющая давать рекомендации пользователю непосредственно при редактировании информации о статье в интерфейсе системы. Использование хэшфункции для множества авторов статьи позволяет производить поиск возможных вариантов перевода и давать рекомендации менее чем за 0.1 сек.

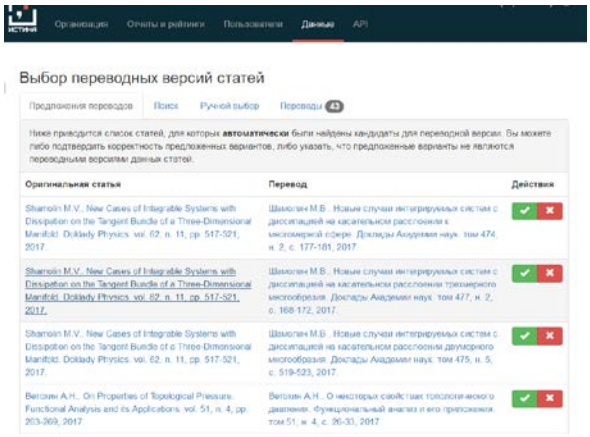

Рис 1. Интерфейс подтверждения найденных вариантов перевода.

Для апробации алгоритма использовались данные о публикациях сотрудников МГУ им. М.В. Ломоносова. Авторами статьи разработан модуль, добавленный в функционал нукометрической системы организации [6]. Разработанный для этих целей интерфейс (рис. 1) позволяет экспертам проводить оценку результатов работы модуля и отмечать в системе правильные и ошибочные варианты предлагаемых переводов. На настоящий момент из 675 оцененных экспертами вариантов 625 вариантов признаны правильными и 50 ошибочными. Таким образом, точность алгоритма составляет 92%. Ошибки определения обусловлены тем, что один и тот же коллектив авторов может публиковать в течение года несколько работ по схожей тематике. В некоторых случаях названия статей бывают настолько схожими, что даже по названиям статьи трудно выбрать правильный вариант (рис. 1).

Литература

1. Садовничий В. А., Васенин В. А., Афонин С. А.и др. Информационная система "ИСТИНА" как big data - инструментарий в области управления на основе анализа наукометрических данных. Материалы Всероссийской конференции с международным участием "Знания-Онтологии-Теории" (ЗОНТ-2015), 6-8 октября. Т. 1, Институт математики им. С.Л.Соболева СО РАН Новосибирск, 2015. С. 115-123.
2. Переводчик «Гугл». — URL: http://translate.google.ru
3. Автоматический переводчик «Промпт». — URL: http://www.translate.ru.
4. Научная Конференция «Обнаружение заимствований – 2017». — URL: http://www.oz2017.ru .
5. Плагиат в научных статьях: трудности обнаружения перевода. — URL: http://ai-news.ru/2018/01/plagiat_v_nauchnyh_statyah_trudnosti_obnaruzheniya_perevoda.html.
6. Васенин В. А., Афонин С. А., Голомазов Д. Д., Козицын А. С. Интеллектуальная Система Тематического Исследования НАучно-

технической информации (ИСТИНА). Информационное общество. № 1-2. С. 21-36. 2013.

References

1. Sadovnichii V. A., Vasenin V. A., Afonin S. A.i dr. Informatsionnaia sistema "ISTINA" kak big data - instrumentarii v oblasti upravleniia na osnove analiza naukometricheskikh dannykh. Materialy Vserossiiskoi konferentsii s mezhdunarodnym uchastiem "Znaniia-Ontologii-Teorii" (ZONT-2015), 6-8 oktiabria. T. 1, Institut matematiki im. S.L.Soboleva SO RAN Novosibirsk, 2015. S. 115-123.
2. Perevodchik «Gugl». — URL: http://translate.google.ru
3. Avtomaticheskii perevodchik «Prompt». — URL: http://www.translate.ru.
4. Nauchnaia Konferentsiia «Obnaruzhenie zaimstvovanii – 2017». — URL: http://www.oz2017.ru .
5. Plagiat v nauchnykh statiakh: trudnosti obnaruzheniia perevoda. — URL: http://ai-news.ru/2018/01/plagiat_v_nauchnyh_statyah_trudnosti_obnaruzheniya_perevoda.html.
6. Vasenin V. A., Afonin S. A., Golomazov D. D., Kozitsyn A. S. Intellektualnaia Sistema Tematicheskogo Issledovaniia NAuchno-tekhnicheskoi informatsii (ISTINA). Informatsionnoe obshchestvo. № 1-2. S. 21-36. 2013.