ИПМ им.М.В.Келдыша РАН

Абрау-2017 • Труды конференции

А.С Козицын, С.А. Афонин,
А.А. Зенизинов

Метод уточнения библиографических данных научных статей на основе статистики в больших коллекциях библиографических данных

Рекомендуемая форма библиографической ссылки

Козицын А.С, Афонин С.А., Зенизинов А.А. Метод уточнения библиографических данных научных статей на основе статистики в больших коллекциях библиографических данных // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18-23 сентября 2017 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2017. — С. 277-280. — URL: http://keldysh.ru/abrau/2017/10.pdf doi:10.20948/abrau-2017-10

Размещена также презентация к докладу

Метод уточнения библиографических данных научных статей на основе статистики в больших коллекциях библиографических данных

А.С Козицын, С.А. Афонин, А.А. Зенизинов

НИИ механики МГУ, Москва

Аннотация. В докладе описывается два метода уточнения библиографических данных статей. Первый метод позволяет автоматизировать процесс поиска переводных версий статей. Второй – уточнять автора статьи по списку соавторов.

Ключевые слова: библиографические данные, автор, статья, статистика.

Одной из актуальных задач управления большими научно-образовательными организациями является внедрение методов оценки эффективности деятельности ее сотрудников и подразделений в целом. Если на уровне отдельных лабораторий, кафедр, отделов и небольших коллективов руководитель может проводить такую оценку на интуитивном уровне по результатам встреч, семинаров и отчетов, то на уровне крупной организации подобная оценка невозможна без использования автоматизированных систем сбора и агрегации наукометрических показателей. Набор показателей должен быть достаточно большим, чтобы охватывать все основные сферы деятельности организации и учитывать особенности отдельных подразделений. Вместе с тем, необходимо учитывать, что любые наукометрические показатели не могут являться абсолютным критерием оценки каждого сотрудника и дают только приблизительную оценку.

В это связи, для решения задач управления любым крупным ВУЗом требуется создание системы сбора, хранения и анализа больших объемов библиографической информации, которая описывает результаты научной деятельности его сотрудников, и предоставляет данные для оценки эффективности деятельности организации.

Часть данных может собираться из внешних систем, например, Web of Science. Однако, в силу того, что большая часть данных во внешних системах отсутствует или крайне разрознена, необходимо разрабатывать собственные механизмы сбора данных от сотрудников организации. Важным элементом любой системы сбора, обработки и анализа данных является создание механизмов верификации собираемой системой информации, которые

включают в себя проверку полноты и точности предоставляемых данных. Верификация данных осуществляется с использованием трех механизмов проверки: проверка конечными пользователями; проверка ответственными сотрудниками подразделений и механизмами автоматического получения данных из сторонних систем.

Одним из важнейших показателей, которые необходимо собирать для оценки научной деятельности сотрудников организации, является количество публикаций, а также их цитируемость, распределение по журналам и темам. При анализе этого показателя необходимо учитывать, что авторы вводят как оригинальные статьи, так и их переводы в иностранных журналах. Переводы статей также имеют большое значение для анализа, по ним может производиться дополнительный сбор информации о цитируемости автора, в том числе в метриках Web Of Science и Scopus. Однако, было бы некорректным учитывать переводы как самостоятельные статьи при подсчете общего количества статей автора за период.

Сложность определения переводных версий статей обусловлена тем, 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. Поскольку перевод статьи печатается позже, на момент вноса в систему информации о статье ее перевода может еще не существовать. В этой связи, становится актуальной задача автоматизации поиска и сопоставления переводных версий статей в процессе сбора подобной информации, поскольку ручная обработка таких объемов данных невозможна. Некоторые журналы на своих страницах в Интернет размещают информацию о наличии у них переводных изданий, однако такая информация плохо структурирована и ее автоматическая обработка очень сложна. Кроме того, значительная часть переводов размещается в иностранных изданиях, не являющихся полными переводными версиями русскоязычного издания.

Одним и подходов к автоматизации процесса поиска переводных версий статей является использование статистических данных о распределении статей по журналам.

Разработанный в рамках данной работы алгоритм состоит из двух этапов. На первом этапе производится поиск пар журналов, которые печатают переводные статьи. Предлагаемый в данной работе поиск производится на основе сравнения количества похожих статей, имеющих одинаковый список авторов. Для быстрого сравнения списка авторов статей используется хэш-функция, позволяющая построить индекс по всему массиву статей, загруженных в систему. В рамках данной работы 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 множество пар журналов, в одном из которых часто печатаются переводные статьи из второго журнала.

Кроме статистической оценки на основе пар похожих статей для пополнения и уточнения списка переводных журналов используется вводимый авторами статей DOI. 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 заключительном этапе работы алгоритма, на основе полученного списка пар журналов проводится поиск пар статей, которые опубликованы в этих журналах и удовлетворяют приведенным выше критериям: опубликованы одними и теми же авторами и дата публикации отличается не более чем на год. Полученные пары статей считаются возможными вариантами перевода.

Тестирование алгоритма проводилось на данных о публикациях сотрудников МГУ им. М.В. Ломоносова. Массив данных содержал описание 78 тысяч статей, в качестве возможных переводов было выделено 2 тысячи статей, точность оценки составила 65%. Основная сложность автоматического поиска переводов заключается в том, что многие авторы публикуют похожие результаты по нескольку раз, немного меняя название статьи. В таких случаях даже эксперту в предметной области трудно дать однозначную оценку, является статья не очень точным переводом или самостоятельной статьей.

Результаты работы алгоритма не позволяют точно выявлять переводы, однако позволяют уведомлять пользователей системы о наличии возможных переводов статей с возможностью поставить соответствующую отметку в данных о статье.

Еще одной важной задачей является определение авторов статей по ее библиографическому описанию. В большинстве систем хранения библиографических систем авторы идентифицируются уникальными ключами (ResearcherID, IRID, ORCID и другие). Вместе с тем в библиографических ссылках на сегодня указывается только фамилия автора и его инициалы или имя без указания ResearcherID или каких-либо уникальных ключей. Вследствие этого возникает задача определения уникального ключа автора по фамилии в конкретном библиографическом описании. Среди сотрудников любой большой организации, в том числе среди сотрудников МГУ им. М.В. Ломоносова, имеется большое количество однофамильцев. Кроме того, в список потенциально возможных авторов пополняется соавторами, работающих в других организациях. Как следствие, по указанным в библиографических данных фамилии и инициалам находится от двух до восьми возможных авторов, зарегистрированных в системе, с различными уникальными идентификаторами.

В этой связи возникает задача автоматического определения правильных авторов по списку соавторов статьи. Один из подходов к решению такой задачи на основе анализа графа соавторства изложен в статье [1]. Метод обладает достаточно хорошей точностью, однако его ресурсоемкость и большое время работы затрудняет его использование при непосредственно вводе авторами статей в систему.

Для осуществления анализа списка соавторов и выделения наиболее вероятного списка авторов статьи в рамках данной работы был реализован алгоритм, использующий анализ частотностей упоминаний двух авторов в одной работе. Кроме того, используется информация об авторизации пользователя, осуществляющего ввод информации в систему, поскольку по статистики более 80% статей вносится в систему одним из соавторов.

На первом шаге алгоритм выделяет список возможных авторов для каждой фамилии, упоминающейся в библиографическом описании статьи. 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.

В качестве данных для обучения использовался массив статей, книг, отчетов, патентов и проектов, содержащий около 900 тысяч записей об авторстве. Тестирование проводилось на массиве авторства статей, содержащих 540 тысяч записей. Совпадение результатов расчета алгоритма с реальными данными составило 520 тысяч записей.

Недостатком алгоритма является невозможность его использования для статей, написанных одним автором. Но этот недостаток частично компенсируется тем, 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.

Литература

1. 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.