ИПМ им.М.В.Келдыша РАН

Абрау-2019 • Труды конференции

А.С. Козицын, С.А. Афонин, Д.А. Шачнев

Определение тематической близости научных журналов и конференций с использованием технологий Big Data

Рекомендуемая форма библиографической ссылки

Козицын А.С., Афонин С.А., Шачнев Д.А. Определение тематической близости научных журналов и конференций с использованием технологий Big Data // Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23-28 сентября 2019 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2019. — С. 450-455. — URL: http://keldysh.ru/abrau/2019/theses/15.pdf doi:10.20948/abrau-2019-15

Размещена также презентация к докладу

Определение тематической близости научных журналов и конференций с использованием технологий Big Data

А.С. Козицын, С.А. Афонин, Д.А. Шачнев

НИИ механики МГУ им.М.В. Ломоносова

Аннотация. Количество публикуемых в мире журналов очень велико. В этой связи, необходим программный инструментарий, который позволит анализировать тематические связи журналов. Разработанный авторами и представленный в этой работе алгоритм использует для анализа тематической близости журналов граф соавторства. Алгоритм нечувствителен к языку журнала и подбирает похожие журналы на разных языках, что сложно реализуемо для алгоритмов, основанных на анализе полнотекстовой информации. Апробация алгоритма проводилась в наукометрической системе ИАС ИСТИНА. В разработанном для этих целей интерфейсе пользователь может выбрать один близкий ему по тематике журнал, и система автоматически сформирует подборку журналов, которые могут представлять интерес для пользователя как с точки зрения изучения имеющихся в них материалов, так и с точки зрения публикации собственных статей. В перспективе разработанный алгоритм можно адаптировать для поиска похожих по тематике конференций, сборников публикаций и научных проектов. Наличие такого инструмента увеличит публикационную активность молодых сотрудников, повысит цитируемость статей и цитируемость между журналами. Результаты работы алгоритма определения тематической близости между журналами, сборниками, конференциями и научными проектами также могут использоваться для построения правил в моделях разграничения доступа к данным на основе онтологий предметной области.

Ключевые слова: тематическая классификация, библиографические данные, граф соавторства, информационные системы.

Determining the thematic proximity of scientific journals and conferences using Big Data technologies

A.S. Kozitsin, S.A. Afonin, D.A. Shachnev

Institute of Mechanics Lomonosov Moscow State University

Abstract. The number of scientific journals published in the world is very large. In this regard, it is necessary to create software tools that will allow analyzing

thematic links of journals. The algorithm presented in this paper uses graphs of co-authorship for analyzing the thematic proximity of journals. It is insensitive to the language of the journal and can find similar journals in different languages. This task is difficult for algorithms based on the analysis of full-text information. Approbation of the algorithm was carried out in the scientometric system IAS ISTINA. Using a special interface, a user can select one interesting journal. Then the system will automatically generate a selection of journals that may be of interest to the user. In the future, the developed algorithm can be adapted to search for similar conferences, collections of publications and research projects. The use of such tools will increase the publication activity of young employees, increase the citation of articles and quoting between journals. In addition, the results of the algorithm for determining thematic proximity between journals, collections, conferences and research projects can be used to build rules in the ontology models for access control systems.

Keywords: thematic classification, bibliographic data, graph of co-authorship, Information Systems.

1. Введение

Количество публикуемых в мире журналов очень велико. Например, в информационно-аналитической системе (ИАС) «ИСТИНА» [1] зарегистрировано более 70 тысяч журналов и еще более 200 тысяч сборников научных публикаций. В этой связи, студентам, аспирантам и молодым ученым необходим программный инструментарий, который позволит на основе аккумулированного опыта других научных сотрудников автоматически подбирать наиболее соответствующие по тематике их научным интересам журналы.

Существует несколько возможных способов решения задачи определения тематической близости журналов. Первый способ основан на использовании тематического анализа полнотекстовых описаний журналов, текстов статей, опубликованных в журналах, их аннотаций и ключевых слов. На основе результатов проведенного полнотекстового тематического анализа таких текстов возможно построение оценки смысловой близости интересов пользователей и публикаций в журнале. Для возможности проведения такого анализа пользователь должен описать область своих научных интересов при помощи ключевых слов или загрузить в систему полные тексты своих статей. Кроме того, необходимо иметь достаточно точно описанные тематические профили всех журналов или полные тексты публикуемых в этих журналах статей. Получение достаточно полных полнотекстовых данных является сложной задачей, поскольку во многих журналах открытая публикация полных текстов статей не разрешена.

Использование только ключевых слов для проведения тематического анализа может давать слишком общие результаты. Это связанно с тем, что во многих случаях ключевые слова статьи характеризуют не ее тематику, а связь статьи с одним из приоритетных направлений развития науки, технологий и техники в Российской Федерации. Например, ключевое слово

«Нанотехнология» встречается в статьях совершенно различной тематики: «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 тематического анализа для решения поставленной выше задачи в наукометрических системах является сложно реализуемым.

Альтернативным методом оценки близости журналов по тематическим направлениям является анализ графа соавторства статей, публикуемых в этих журналах. При реализации этого метода предполагается, что большинство авторов публикую свои статьи в тематически близких журналах. Вследствие этого, в близких по тематике журналах часто публикуются одинаковые авторы. В отличии от методов полнотекстового тематического анализа, основанный на использовании графов соавторства подход не требует наличия полнотекстовой информации о статьях, и использует только библиографические данные публикуемых в журналах статей. Такие данные могут быть получены из наукометрических систем (например, ИАС «ИСТИНА»), или систем цитирования (например, WoS).

2. Алгоритм оценки тематической близости журналов

Формально задачу оценки близости журналов можно сформулировать следующим образом. Необходимо построить граф, вершинами которого являются журналы, а веса ребер соответствуют их тематической близости.

Разработанный алгоритм 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. Основным недостатком такого метода является невозможность учитывать значимость авторов для каждой статьи. Во многих случаях статьи пишутся только одним автором, фамилия которого ставится на первом месте в ее библиографическом описании. Остальные соавторы могут участвовать в работе над статьей незначительно, и их основное направление научной деятельности может не совпадать с ее тематикой.

Для проверки гипотезы о значимости порядка авторов при проведении тематического анализа была проведена оценка доли статей, в которых порядок авторов определяется лексикографическим порядком, а не значимостью в работе над статьей. Из наукометрической системы МГУ были отобраны для анализа все статьи в журналах за 2014-2017 гг с количеством авторов от 2 до 7.

Для каждого из указанного количеств авторов были посчитаны процент статей L, для которых правильный набор авторов определяется лексикографическим порядком. Результаты расчета для различного количества авторов приведены в таблице 1.

Количество авторов	L
2	24%
3	16%
4	9%
5	6%
6	6%
7	3%

Таблица 1. Процент статей с лексикографическим порядком.

Из приведенных в таблице данных можно сделать вывод, что в большинстве случаев основным автором является автор, который указан в библиографическом описании первым. Для учета этого факта была разработана формула расчета веса ребер с учетом позиции автора в библиографическом описании статьи. Вес автора для каждой статьи определяется как 1/2+1/(2K) для первого автора и 1/(2K) для остальных соаторов, где K-количество соавторов в статье. Степень связи по заданному автору для двух журналов определяется как минимум из максимумов его весов по подмножествам статей в каждом из журналов. Окончательный вес ребра связи между двумя журналами может быть рассчитан как сумма степеней их связи по всем авторам.

3. Программная реализация и результаты тестирования

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 между журналами производится с заданными интервалами времени и сохраняется в таблицы СУБД.

В разработанном для этих целей интерфейсе [2] пользователь может выбрать один близкий ему по тематике журнал, и система автоматически сформирует подборку журналов, которые могут представлять интерес для

пользователя как с точки зрения изучения имеющихся в них материалов, так и с точки зрения публикации собственных статей. Для удобства работы пользователя предусмотрена возможность добавлять выбранный журнал в заметки, которые впоследствии можно просматривать, редактировать, а также использовать при последующем поиске.

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 как следствие слишком широкой тематической области принимаемых в журнал статей.

4. Заключение

Описанный в настоящей работе алгоритм позволяет автоматически оценивать степень тематической близости научных журналов на основе библиографического описания статей и без использования полнотекстовых версий статей. Следует отметить, что алгоритм нечувствителен к языку журнала и подбирает похожие журналы на разных 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. Наличие такого инструмента увеличит публикационную активность молодых сотрудников, повысит цитируемость статей и цитируемость между журналами.

Результаты работы алгоритма определения тематической близости между журналами, сборниками, конференциями и научными проектами также могут использоваться для построения правил в моделях разграничения доступа к данным на основе онтологий предметной области.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект 18-07-01055.

Литература

1. В.А. Садовничий, В.А. Васенин Интеллектуальная система тематического исследования наукометрических данных: предпосылки создания и методология разработки. Часть 1. // Программная инженерия, 9(2) — М., 2018. — С. 51-58-338.
2. ИАС ИСТИНА. — URL: https://istina.msu.ru.
3. S. Afonin // Ontology models for access control systems. In 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC) — С. 1-6. — doi: 10.1109/RPC.2018.8482178 .

References

1. V.A. Sadovnichii, V.A. Vasenin Intellektualnaia sistema tematicheskogo issledovaniia naukometricheskikh dannykh: predposylki sozdaniia i metodologiia razrabotki. Chast 1. // Programmnaia inzheneriia, 9(2) — M., 2018. — S. 51-58-338.
2. IAS ISTINA. — URL: https://istina.msu.ru.
3. S. Afonin // Ontology models for access control systems. In 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC) — S. 1-6. — doi: 10.1109/RPC.2018.8482178 .