ИПМ им.М.В.Келдыша РАН

Абрау-2019 • Труды конференции

Б.А. Низомутдинов, А.С. Тропников

Автоматизированный сбор данных для наукометрического анализа

Рекомендуемая форма библиографической ссылки

Низомутдинов Б.А., Тропников А.С. Автоматизированный сбор данных для наукометрического анализа // 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. 523-531. — URL: http://keldysh.ru/abrau/2019/theses/76.pdf doi:10.20948/abrau-2019-76

Размещена также презентация к докладу

Автоматизированный сбор данных для наукометрического анализа

Б.А. Низомутдинов, А.С. Тропников

Университет ИТМО

Аннотация. Представлены алгоритмы отбора и последующей обработки библиографической информации с целью последующего наукометрического анализа публикаций, проиндексированных в системе РИНЦ. Информационная база формировалась теме «электронное правительство» и конвертировалась в формат системы VOSviewer для дельнейшей обработки и формирования наукометрических карт с кластеризацией предметных областей. В результате применения представленной технологии был сформирован массив из 23 тыс. описаний публикаций, пригодный для загрузки в систему VOSviewer.

Ключевые слова: парсер, библиометрия, наукометрия, электронное участие

Automated data collection for scientometric analysis

B.A. Nizomutdinov, A.S. Tropnikov

ITMO University

Abstract. This article presents the algorithms for a selection and subsequent processing of bibliographic information for the subsequent scientometric analysis of publications indexed in the RINC system. The information base was formed on the subject of "e-government" and was converted into a format of the VOSviewer system for further processing and formation of scientometric maps with clustering of subject areas. As a result, an array of 23 thousand descriptions of publications was formed, suitable for uploading to the VOSviewer system.

Keywords: parser, bibliometrics, scientometrics, e-participation

1. Введение

Для хранения данных о научных публикациях используются цифровые библиографические и реферативные базы данных. Такие базы представляют собой инструмент для отслеживания цитируемости статей, опубликованных в научных изданиях. Также они являются одним из главных источников получения наукометрических данных для проведения оценочных исследований.

Наиболее авторитетными цифровыми международными реферативными базами являются Web of Science и Scopus.

В Российской федерации максимально полной цифровой библиографической и реферативной базой является «Научная электронная библиотека (ELIBRARY.RU). Это крупнейшая в России электронная библиотека научных публикаций, обладающая богатыми возможностями поиска и получения информации. Библиотека интегрирована с Российским индексом научного цитирования (РИНЦ) – бесплатным общедоступным инструментом измерения и анализа публикационной активности ученых и организаций созданным по заказу Минобрнауки РФ.

В мировой практике имеется богатый опыт применения наукометрического анализа, с данными получаемых из международных реферативных баз данных Web of Science и Scopus [1-3]. Так, для получения основной метаинформации публикаций по заданной теме, в системе Web of Science можно выгрузить файлы формата .txt, содержащий краткую информацию (автор, название, источник, аннотация) о работе и пристатейные ссылки. Полученный текстовый файл может быть использован в различных прикладных приложениях, способных проводить сложный анализ на основе скаченных данных.

Для наукометрического анализа разрабатывается и модернизируется специализированное программное обеспечение, предусмотренное для автоматизированной обработки массивов информации, выгруженных из библиографических баз данных.

Однако, осуществить такую выгрузку для отечественных публикаций в машиночитаемом формате нет возможности. Так, к примеру, система ELIBRARY не предоставляет доступа по API, инструментарию выгрузки подборок публикаций в формате XML и др. Тем самым, процесс получения метаданных по интересующим публикациям для дальнейшего анализа усложняется и требует значительных трудозатрат. Для оценки возможностей использования данных из РИНЦ для загрузки в системы, традиционно работающих с данными, полученными из международных баз (Web of Sciense и Scopus), в 2018 году был проведен пилотный проект с ручной выгрузкой из базы РИНЦ и конвертацией в международный обменный формат [4]. Результаты этого кейса показали, что кириллическая информация вполне может использоваться для проведения сравнительных библиометрических исследований, однако необходимо решить вопросы получения данных автоматизированными методами.

Целью данной работы является получение метаданных публикаций по заданной теме при помощи автоматизированных средств и дальнейшая конвертация в требуемый формат, для последующего наукометрического анализа.

Для данного исследования была определена тема «электронное правительство» и смежные научные направления.

Анализ электронного правительства как научной дисциплины является предметом обсуждений и дискуссий. Предпринимаются различные попытки исследования электронного правительства и электронного участия, в которых проводится систематический обзор литературы, используется систематизация и структурирование метаданных [5-7]. Однако, несмотря на растущий интерес к данным темами, имеются некоторые ограничения: а именно недостаток данных о российском сегменте научных публикаций и отсутствие сравнительных исследований международного и российского контекстов. Полученные, в ходе данной работы, данные позволят расширить исследовательский контекст, охват и количество доступных данных, для выявления новых моделей и взаимосвязей.

В данной работе не будет представлена подробная интерпретация собранных метаданных, основная цель материала - апробация методики по сбору и обработке данных для наукометрического анализа из базы ELIBRARY автоматизированным способом. Собранный массив данных изучается междисциплинарной командой и будет интерпретирован на следующем этапе работ.

2. Методология и методы исследования

За последние годы в наукометрии наблюдается устойчивая тенденция к использованию автоматизированных средств сбора, обработки и анализа получаемых данных. Данный тезис касается практически всех областей современной науки.

Для изучения российского контекста исследований по теме «электронное правительство», мы выбрали библиотеку ELIBRARY, которая индексирует документы, опубликованные в российских журналах, конференциях, а также диссертации. Данная электронная библиотека является наиболее полным источником библиографических данных для России, однако, анализ данных из этой системы является трудозатратой задачей, так как в ELIBRARY нет автоматического инструментария для извлечения библиографических данных, выгрузки результатов в машиночитаемом формате или по API.

В данном исследовании, при формировании массива данных были применены возможности системы парсинга для извлечения метаданных публикаций по заданной тематике, с дальнейшей обработкой и кластеризацией.

Разработанная система парсинга позволяет осуществлять процесс сбора необходимой информации из заданных источников, а также проводить предварительную конвертацию их результатов в требуемый формат.

В качестве инструмента для анализа полученных данных был использован комплекс VOSviewer.

VOSviewer является специальной программой, при помощи которой исследователи могут визуализировать в виде наукометрических карт. При построении карты, VOSviewer извлекает термины из аннотаций и ключевых слов статей и добавляет их в единый текстовой корпус. Далее, VOSviewer

выделяет кластеры терминов на основе их совместной встречаемости в используемых текстах. При визуализации VOSViewer окрашивает отдельные кластеры в разные цвета.

2.1. Сбор ссылок и формирование базы

Поиск и отбор статей был осуществлен при помощи портала ELIBRARY, который позволил собрать ссылки на статьи, касающихся заданной тематики. В качестве поисковых запросов были использованы такие термины, как «электронное правительство», «Electronic Governance», «электронное участие», «электронная демократия», «Electronic Participation» и др.

При помощи существующего инструментария ELIBRARY был задан ряд сложносоставных запросов с набором различных параметров. По результатам запросов был получен список из более чем 23 тысяч научных работ. В электронной библиотеке можно искать публикации по заданному ключевому слову, однако, как уже было выше отмечено, нет возможности выгрузить найденные данные в обменный формат.

Один из инструментов, который был задействован для формирования базы ссылок, это «Подборки публикаций». Все найденные публикации в ELIBRARY добавлялись в подборки. В итоге была сформированы подборки из 23 тыс. статей (или 230 страниц - каждая страница содержит 100 ссылок).

В российской научной электронной библиотеке не существует возможностей для агрегации и анализа такого объема данных. С помощью существующих инструментов возможно выводить результаты запросов по 100 записей на странице, без способа выгрузки данных в отдельных файл. Данные ограничения не позволяют эффективно обрабатывать большие объемы данных.

Для решения данной проблемы в среде разработке Embarcadero Delphi 7 был реализован инструмент для обработки исходного кода страницы, выгрузки прямых ссылок на научные работы и сохранения их в отдельном файле. При помощи отдельного скрипта был запущен цикл по загрузке результатов поиска, выгрузке исходного кода в программу и открытии новой страницы в электронной библиотеке для продолжения цикла.

В исходном коде страницы, полученном в результате работы скрипта, хранятся ID-адреса научных материалов. При помощи набора тэгов, программа осуществляла обработку исходного кода и сохраняла набор ID-адресов, находящихся в коде. Используя данные ID-адреса, есть возможность для создания прямых ссылок на научные материалы: ссылки создаются по одинаковому шаблону. На основе ID формировался итоговый адрес публикации вида "https://elibrary.ru/item.asp?id=" + "ID публикации".

По завершению работы программы был получен файл с более чем 23 тысячами прямых ссылок на научные работы. Данные ссылки были использованы в парсере, чтобы получить подробную информацию о материале: авторство, аннотация, ключевые слова, издание и т.п. На основе полученных ссылок была сформирована база источников, для дальнейшего парсинга.

2.2. Парсинг

Парсинг сайтов является наилучшим и эффективным решением автоматизации сбора и обработки информации. Парсинг - это принятое в информатике определение последовательного синтаксического анализа информации, размещённой на интернет - страницах. Парсер - это программа или скрипт, позволяющая выполнить такой анализ и представить результат в нужном для пользователя виде.

Процесс парсинга html - страницы, который проводился в данной работе, можно разделить на 4 основных этапа:

- Первый этап - получение исходного кода html-страницы. На этом шаге выполняется копирование исходного кода страницы с дальнейшим извлечением из неё информации.
- Второй этап - извлечение из полученного кода нужной информации. Получив исходный код html-страницы, необходимо выполнить над ним обработку, т.е. отделить искомый текст от гипертекстовой разметки, выстроить иерархическое дерево элементов документа (DOM) и извлечь из страницы искомую информацию. По заданным критериям выделить только основную информацию, которая представляет интерес. Для сбора важных параметров публикации были заданы границы парсинга каждого поля, на примере одной страницы статьи в ELIBRARY. Анализировался исходных код страницы и выделялись теги HTML в которых содержались параметры. К примеру, название публикации на странице в ELIBRARY содержится в метатеге: <title>Название публикации</title>.
- Третий этап - конвертация собранных данных в формат Web of science, в итоге данные с каждой страницы сразу конвертировались в формате WOS. Подробнее о конвертации изложено в пункте 2.2. Конвертация.
- Четвертый этап - сохранение результата. После успешного извлечения данных страницы их необходимо сохранить в требуемом виде для дальнейшей обработки.

В итоге, с помощью парсинга с заданных URL происходил сбор основных метаданных публикаций. Общая схема работы представлена на рисунке 1.

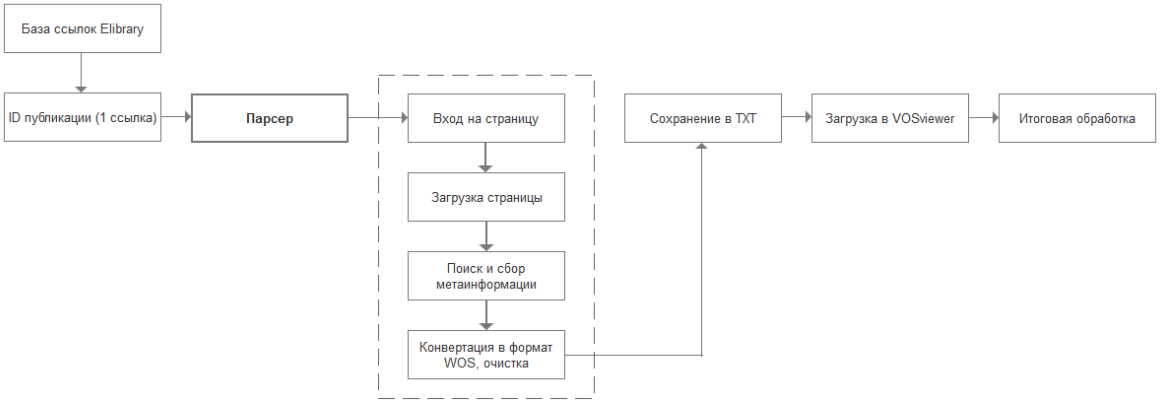

Рис. 1. Модель сбора данных с Elibrary для обработки в VOSviewer

При большом количестве страниц обхода появилась проблема блокировки парсера сервером ELIBRARY. После обхода 100 страниц происходила блокировка IP адреса, с которого работал парсер. Для решения этой задачи потребовалось оптимизировать скорость и частоту запросов таким образом, чтобы избежать блокировок.

Экспериментальным путем было подобрано оптимальное время цикла запроса, дополнительно была задана динамическая пауза между обходами страниц, и в итоге составляла от 3 до 30 секунд, т.е. имитировалась работа обычного пользователя.

Также, во избежание блокировки использовался динамический подход для формирования данных User Agent. User agent - это клиентское приложение, использующее определённый сетевой протокол. При посещении веб-сайта клиентское приложение обычно посылает веб-серверу информацию о себе. Это текстовая строка, являющаяся частью HTTP-запроса, начинающаяся с User-agent: или User-Agent:, и обычно включающая такую информацию, как название и версию приложения, операционную систему компьютера и язык.

Пример User-agent, которые использовались:

- msnbot/1.1 (+http://search.msn.com/msnbot.htm)
- Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)
- Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Iron/28.0.1550.1 Chrome/28.0.1550.1
- Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)
- Opera/9.80 (Windows NT 6.1; WOW64) Presto/2.12.388 Version/12.16

Для увеличения количества потоков и скорости обхода также планировалось задействовать прокси-сервера (proxy servers). Но, найти бесплатные и качественные proxy - на данном этапе работы не удалось, поскольку большая их часть заблокирована ELIBRARY. На следующем этапе исследование планируется платная аренда proxy.

2.3. Конвертация данных

Одна из задач данного исследования - обработка полученной информации в системе VOSviewer. Данная программа может работать с метаданными определенного типа и формата. Для этого, собранная информация с Elibrary конвертировалась в формат Web of Science, так как данный формат поддерживает VOSviewer.

Формат WOS предполагает хранение информации о публикации в текстовом файле, где данные о публикации хранятся в одном массиве, при этом, каждая публикация размечена особым образом, в начале каждого элемента стоит двух символьное обозначение, есть также обозначение, которое указывает на окончания отдельной публикации. Эти обозначения полей (метатеги) из двух символов определяют поля в записях. Так, формат WOS

предусматривает более 70 метатегов [2]. После анализа структуры страницы публикаций в ELIBRARY было выявлено, что для большей части публикаций можно собрать следующие данные:

Таблица 1. Модель сбора данных для обработки в VOSviewer

Тег	Расшифровка тега
SO	название публикации
AU	авторы
PY	год
DE	ключевые слова
AB	аннотация
LA	язык
SE	издание
ER	тег, который закрывает публикацию

После того, как парсер «обходил» 1 страницу, собранные метаданные конвертировались в обменный формат Web of Science и сохранялись в локальную базу данных.

Таблица 2. Пример заполненной таблицы

Тег	Расшифровка тега
SO	"Открытое правительство" как способ борьбы с дефицитом демократии в условиях господства информационно-коммуникационных технологий
AU	Шингирей А.
PY	2018
DE	Электронная демократия; информационно-коммуникационные технологию; открытое правительство
AB	Статья посвящена исследованию динамики событий, меняющих формат взаимоотношений правительства и граждан: мы сегодня живём в обществе, где постоянно растущее число людей может получить необходимую информацию по различным, интересующим их, политическим и социальным вопросам посредством Интернета, что, в свою очередь, формирует их собственное мнение в данных сферах общественной жизни. Это приводит к необходимости поиска путей совершенствования и трансформации существующей демократической модели в новую, более надежную и отлаженную. Одним из возможных способов борьбы с кризисом репрезентативной демократии является интегрированная система «Открытое правительство» (система «ОП»), которая основана на принципах открытости власти, вовлеченности граждан и общественном контроле.
LA	русский
SE	Этносоциум и межнациональная культура
ER	

Далее была произведена пробная загрузка данных в систему VOSviewer и построена модель библиометрической сети, отображающая частоту использованных в публикациях ключевых слов и смысловую связь между ними.

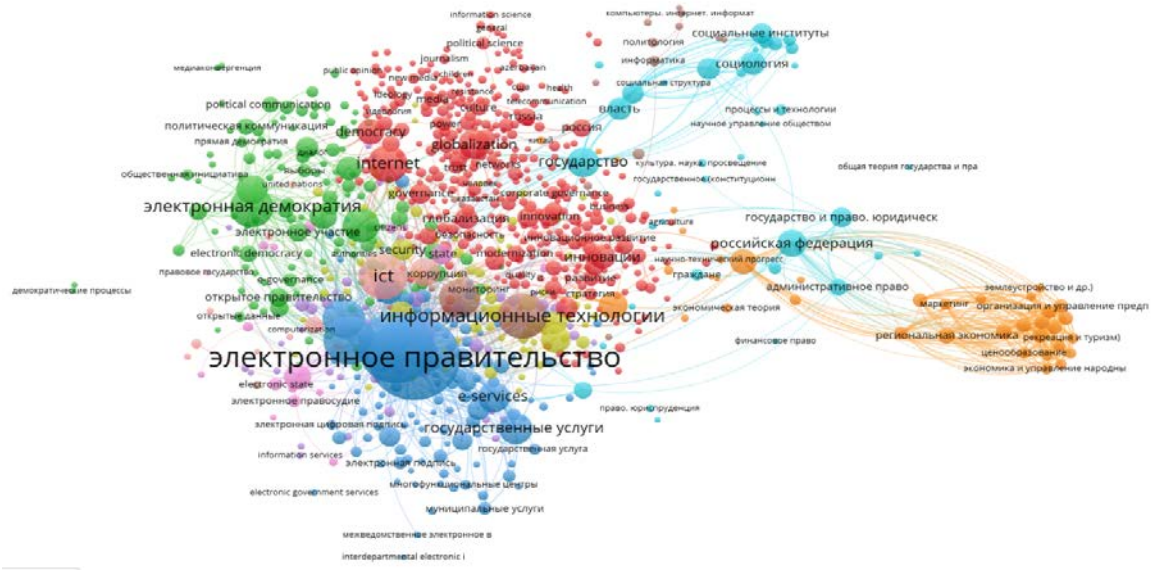

Рис. 2. Пример обработки полученных данных в системе VOSviewer

3. Результаты исследования

Произведенный анализ показал, что открытые данные, полученные путем парсинга Elibrary, можно использовать для наукометрического анализа. При этом, данный подход можно использовать и для других тем исследований, для поиска концептуальных основ исследовательской области. Представленные результаты могут служить руководством для дальнейших междисциплинарных работ.

Исследовательский интерес на последующих этапах работы следует сконцентрировать на качественном анализе собранных данных с целью изучения междисциплинарных направлений российской науки, а также более подробной интерпретации. Кроме того, необходимо расширить набор метаинформации, который собирался в данном исследовании. Также, одним из будущих направлений исследований может стать сетевой анализ публикаций (журналы и конференции) для изучения их значимости и специализации в контексте исследования, а также выявление моделей взаимосвязей. Это может улучшить наше понимание предметной области с институциональной точки зрения. Кроме того, поскольку количество наукометрических методов и программных средств увеличивается, это может быть эффективным способом сравнения и перекрестной проверки полученных результатов.

Работа выполнена при поддержке Российского научного фонда (РНФ) в рамках проекта №18-18-00360 «Электронное участие как фактор динамики политического процесса и процесса принятия государственных решений»

Литература

1. Digital Libraries and Information Access: Research Perspectives. Edited by. G.G. Chowdhury and Schubert Foo. — Chicago, Illinois: Neal-Schuman/ALA, 2012.
2. Hossfeld U., Levit G.S., Prokudin D. Selection Methods of Digital Information Resources for Scientific Heritage Studies: A Case Study of Georgy F. Gause // ACM International Conference Proceeding Series. 2017. Vol. Part F133135. P. 69-74. DOI: 10.1007/978-3-030-02846-6_11
3. Van Eck N.J., Waltman L. Citation-Based Clustering of Publications using CitNetExplorer and VOSviewer // Scientometrics. 2017. Vol. 111(2). P. 1053-1070. DOI: 10.1007/s11192-017-2300-7
4. Прокудин Д.Е., Панфилов Г.О. Концепция «электронного участия» в современной российской науке: наукометрический анализ // Интернет и современное общество: сборник тезисов докладов. Труды XXI Международной объединенной научной конференции «Интернет и современное общество» (IMS-2018), Санкт-Петербург, 31 мая – 2 июня 2018 г. — Электрон, дан. — СПб: Университет ИТМО, 2018. С. 92-94.
5. Qi T. A scientometric analysis of eparticipation research // International Journal of Crowd Science. 2018. Vol. 2 (2). P. 136-148. DOI: 10.1108/IJCS-08-2018-0015
6. Rana N.P. Dwivedi Y.K., Williams M.D. A meta-analysis of existing research on citizen adoption of e-government // Information Systems Frontiers. 2015. Vol. 17(3). P. 547-563. DOI: 10.1007/s10796-013-9431-z
7. Reece B. E-government literature review // Journal of E-government. 2006. Vol. 3(1). P. 69-110. DOI: 10.1300/J399v03n01_05