УДК 004.62

А.С. Коляда
Аспирант кафедры управления системами безопасности жизнедеятельности
В.Д. Гогунский
Доктор технических наук, профессор, заведующий кафедрой управления системами безопасности жизнедеятельности

Одесский национальный политехнический университет, Одесса

АВТОМАТИЗАЦИЯ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ НАУКОМЕТРИЧЕСКИХ БАЗ ДАННЫХ

Рассмотрен процесс автоматического извлечения информации из международных наукометрических баз данных. Разработана система извлечения информации о публикациях по параметру "Автор".

Ключевые слова: наукометрическая база данных, веб-скрапинг, «паук», «краулер»

Розглянуто процес автоматичного добування інформації з міжнародних наукометричних баз даних. Розроблено систему вилучення інформації про публікації за параметром "Автор".

Ключові слова: наукометрична база даних, веб-скрапінг, «павук», «краулер»

The process of automatic extraction of information from the international science-metric database. A system for extracting information on the publications by "Author".

Keywords: science-metric database, web scraping, «spider», «crawler»

Постановка проблемы

Наукометрическая база данных – библиографическая и реферативная база данных, а также инструмент для отслеживания цитируемости научных статей. В основном она появилась с развитием интернета, но история начинается еще с 70-х годов XIX века, когда впервые появились два индекса научного цитирования – индекс юридических документов Shepard's Citations в 1873 году и индекс научных публикаций по медицине Index Medicus в 1879.

Приказом Министерства образования, науки, молодежи и спорта Украины от 17.10.2012 № 1112 "Про опублікування результатів дисертацій на здобуття наукових ступенів доктора і кандидата наук" определены требования на публикацию статей в изданиях, которые включены в международные наукометрические базы данных. По мнению некоторых авторов, этот приказ призван заменить перечень профильных изданий ВАК Украины, который периодически обновлялся в ручном режиме, а значит, требовал постоянного внимания официальных инстанций на более прозрачный по принципу формирования, автоматически обновляющийся список журналов, включенных в международные наукометрические базы [1; 2].

В связи с этим возникает проблема поиска и идентификации автором в этих базах своих публикаций. С подобной проблемой сталкиваются многие научные сотрудники, аспиранты и преподаватели при подготовке отчетности по научной работе. Наиболее простой способ определения публикаций в различных изданиях, которые включены в наукометрические базы, состоит в том, чтобы вести реестр регистрации журналов и включения их в эти базы.

Множество данных в слабо структурированной системе всемирной паутины образует сложную структуру организации информационных взаимодействий, изменяющихся во времени. При этом некоторые издания могут быть включены в одну и более наукометрических баз. Число публикаций постоянно увеличивается. Форматы представления библиографических данных и в публикациях, и в наукометрических базах существенно отличаются. Поиск публикаций в такой разнородной неформализованной среде часто удается осуществить только в ручном режиме.

Процесс поиска в данной среде является скорее искусством, нежели информационной технологией, и зависит от умений и навыков пользователя. Проблема состоит в том, чтобы максимально формализовать и автоматизировать этот процесс.

Для решения этой проблемы необходим способ извлечения данных из наукометрических баз в структурированном виде для возможной их дальнейшей обработки.

Анализ последних исследований и публикаций

Исследованиями в направлении извлечения информации из глобальной сети Интернет занимаются крупные компании Google, Yandex, Microsoft. Они используют результаты исследований в реализации поисковых машин, которые являются главным компонентом поисковых систем. Поисковая машина представляет собой комплекс программ, предназначенный для поиска информации.

Одной из главных функций поисковых машин является извлечение информации из сети. Далее происходит обработка результатов, их индексация для ускорения выдачи результатов поиска и повышения его релевантности.

Основными компонентами подсистемы сбора и извлечения информации являются:

– «Паук» (Spider) – программа для загрузки веб-страниц;

– «Краулер» (Crawler) – программа для автоматического прохода по всем ссылкам, найденным на странице.

Паук скачивает веб-страницы тем же способом, что и веб-браузер, т. е. имитируется действие пользователя. Но веб-браузер отображает эту информацию в графическом виде, а паук сохраняет ее для дальнейшей обработки. Краулер выделяет все ссылки, присутствующие на странице и переходит по всем или определенным ссылкам, исходя из заданных заранее условий. Следуя по найденным ссылкам, он перенаправляет страницы пауку для их загрузки.

Робот Googlebot – это разработанная Google программа сканирования Интернета («Паук»). Сканирование является процессом, в ходе которого робот Googlebot обнаруживает новые и обновленные страницы для добавления в индекс. Google использует огромную сеть компьютеров, чтобы извлечь содержание миллиардов веб-страниц. Робот Googlebot функционирует автономно и применяет алгоритмический процесс: компьютерные программы определяют сайты, которые нужно сканировать, а также частоту сканирования и количество извлекаемых страниц на каждом сайте.

Процедура сканирования начинается с получения списка URL веб-страниц, который создается на основе результатов предыдущих сеансов сканирования. Его дополняют данные из файлов Sitemap, предоставленных веб-мастером. Просматривая эти сайты, робот Googlebot находит на каждой странице ссылки и добавляет их в список страниц, подлежащих сканированию. Все новые и обновившиеся сайты, а также неработающие ссылки помечаются для обновления в индексе.

Цель статьи

Основная цель настоящей статьи – разработать способ извлечения данных о публикациях по параметру *Автор* из наиболее известных наукометрических баз данных с возможностью расширения поддерживаемых источников. Второстепенной задачей является знакомство с наиболее известными наукометрическими базами данных.

Основной материал исследования

На сегодня насчитывается значительное количество международных наукометрических баз данных, которые различаются структурой и способом хранения информации. Программный интерфейс для доступа к базе если и существует, то зачастую не афишируется. Не существует единого, универсального интерфейса, который подходил бы ко всем базам. Но есть один интерфейс, который имеют многие наукометрические базы данных, и который ориентирован больше на пользователя, нежели на программное обеспечение.

Доступ к содержимому (в ограниченном виде) предоставляет веб-интерфейс. Пользователь с помощью веб-браузера загружает веб-страницу определенной наукометрической базы данных и, используя поиск по заданным параметрам, получает необходимую информацию на странице.

В данном исследовании предлагается извлекать программным способом информацию, ориентированную на пользователя (человека). Таким образом имитируется работа пользователя, который загрузил бы тысячи веб-страниц и собрал информацию определенной структуры в локальное место хранения.

Для автоматического поиска и извлечения данных используется подход, основанный на процессе, применяемом в поисковых машинах, – веб-скрапинг.

Веб-скрапинг – это процесс извлечения информации из веб-страниц, который фокусируется на преобразовании неструктурированных данных в сети (например, в формате HTML) в структурированный формат данных, который может быть проанализирован и сохранен (рис. 1).

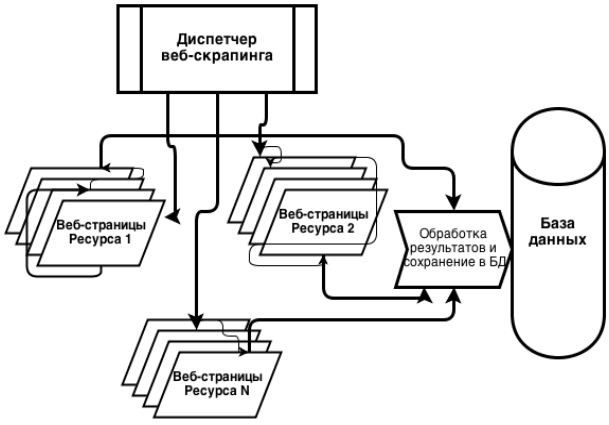

Рис. 1. Процесс веб-скрапинга

Веб-скрапинг относится к автоматизации работы во всемирной паутине, также использует программы типа паук и краулер для обхода и загрузки веб-страниц. В отличии от поисковых машин, сканируется узкий круг веб-страниц, заданный начальными условиями и извлекается только полезная информация.

После извлечения информации в структурированном виде возможна дальнейшая ее обработка, которая может включать в себя фильтрацию результатов по некоторым критериям, подсчет различных коэффициентов и показателей, а также наиболее важную и сложную задачу – определение однофамильцев и, соответственно, повышение точности результатов.

Исходя из содержимого извлеченной информации, можно решать проблему однофамильцев несколькими способами или их комбинацией:

– семантический анализ темы или направления публикации;

– анализ ключевых слов публикации;

– анализ источника публикации.

В результате данного исследования спроектирована система извлечения информации о научных публикациях по параметру поиска “Автор”. Используя это свойство, программа выполняет поиск по известным ей наукометрическим базам данных, загружает результаты и извлекает информацию определенной структуры.

На данный момент поддерживаются следующие широко известные международные наукометрические базы данных.

– *Scopus* – библиографическая и реферативная база данных и инструмент для отслеживания цитируемости статей, опубликованных в научных изданиях. Позиционируется издательской корпорацией Elsevier как крупнейшая в мире универсальная реферативная база данных с возможностями отслеживания научной цитируемости публикаций [3];

– *Российский индекс научного цитирования (РИНЦ)* – библиографическая база данных научных публикаций проекта eLIBRARY. РИНЦ выполняет функцию не только инструмента для оценки учёных или научных организаций на основе цитирования, но и авторитетного источника библиографической информации по научной периодике [4];

– *BASE (Bielefeld Academic Search Engine)* – многопрофильная поисковая система для научных интернет-ресурсов, созданная библиотекой университета Билефельд, Германия. Является одной из самых обширных поисковых систем публикаций в мире, особенно для открытого академического доступа к веб-ресурсам [5];

– *Index Copernicus* – интерактивная база данных из внесенной пользователем информации об ученом профиле, научных учреждений, публикаций и т.д. База данных имеет несколько инструментов оценки производительности, которые позволяют отслеживать влияние научных работ и публикаций, отдельных ученых или научно-исследовательских учреждений. Также Index Copernicus предлагает традиционное реферирование и индексирование научных публикаций [6];

– *Springer* – международная издательская компания, специализирующаяся на издании академических журналов и книг по естественно-научным направлениям (теоретическая наука, медицина, экономика, инженерное дело, архитектура, строительство и транспорт). Является вторым по величине издательством в мире после Elsevier в области «STM» (science, technology, medicine — *англ.* наука, технологии, медицина) [7].

Ниже рассмотрены технологии и средства, использованные при реализации системы извлечения неформализованной информации из веб-страниц.

Используемый формат извлеченных данных – текстовый формат обмена данными (JSON). Структура состоит из нескольких полей таких, как *Автор*, *Название* (публикации), *Источник*, *Дата и Наукометрическая база* и др. Структура не жесткая, может различаться набором полей для разных результатов, но такие поля, как *Автор* и *Наукометрическая БД* являются обязательными.

Структура данных имеет следующий вид:

{"title" : "Informational Model of Natural Language Processing",
"url" : "http://hdl.handle.net/10525/263",
"author" : ["Palagin, Aleksandr",
"Gladun, Viktor",
"Petrenko, Nikolay",
"Velychko, Vitalii",
"Sevruk, Aleksey",
"Mikhailyuk, Andrey"],
"spider": "base-search",
"source": "Institute of Information Theories and Applications FOI ITHEA",
"date": "2008",
"desc": "The formal model of natural language processing in knowledge-based information systems is considered. The components realizing functions of offered formal model are described." }.

Реализация загрузки веб-страниц, навигация по ссылкам и извлечение данных из веб ресурсов производится при помощи веб-скрапинг фреймворка Scrapy [8], схема работы которого показана на рис. 2.

Извлеченные данные сохраняются в NoSQL базе данных MongoDB [9], потому что они не имеют жестких связей, как в реляционных базах данных.

Фреймворк Scrapy предоставляет удобный способ расширения числа поддерживаемых наукометрических баз данных путем добавления новой программы-паука, ориентированного на работу с веб-ресурсом конкретной базы данных.

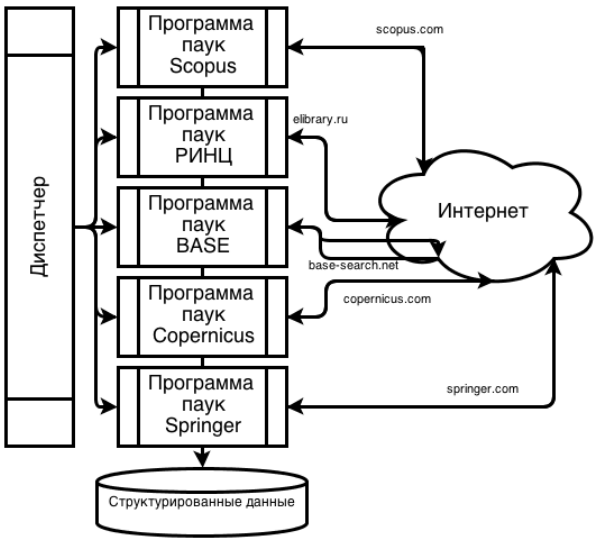

Рис. 2. Архитектура веб-скрапинг фреймворка Scrapy

Используемый набор технологий и программного обеспечения позволяют создать программный продукт по извлечению информации из неоднородных и неформализованных источников (таких как наукометрические базы) и преобразованию ее в структурированный вид с возможной дальнейшей обработкой. Эти данные необходимы в первую очередь аспирантам и соискателям при подготовке к защите диссертаций [10 – 13]. Кроме того предлагаемая система может быть полезна при оценке деятельности ВУЗов [14].

Выводы и перспективы дальнейших исследований

Размещение публикаций в международных наукометрических базах может иметь позитивные последствия для науки Украины. На примере базы Scopus на сайте Национальной библиотеки Украины имени В. И. Вернадского показано, какую полезную информацию можно извлечь из нее: рейтинг ученых Украины, рейтинг организаций Национальной академии наук Украины, рейтинг высших учебных заведений Украины и т.д.

Представленный способ извлечения информации из международных наукометрических баз данных является своего рода универсальным интерфейсом для программного доступа к их содержимому (хоть и ограниченному) [10]. Используется процесс веб-скрапинга для обрабоки и извлечения неформализованных данных с дальнейшим приведением их к нормальному виду. Для поиска своих публикаций автору требуется ввести свою фамилию и/или имя и запустить работу программы. Далее результаты в структурированном виде сохраняются в локальную (относительно наукометрических баз) базу данных и готовы к дальнейшей обработке или просмотру.

Дальнейшее развитие заключается в исследовании и решении проблемы определения однофамильцев и отброса нерелевантных результатов, что повысит качество результатов поиска данных и удобство использования программного продукта на практике. Также планируется расширение поддерживаемых наукометрических баз для охвата как можно большего числа научных изданий.

Список литературы

1. Формализация проблемы извлечения знаний из естественно языковых текстов. [Текст] / [Палагин А., Кривый С., Петренко Н., Бибиков Д.]. – Sofia: Information technologies & knowledge, 2012. – 100 с.

2. Флегантов Л. Для чего нам нужны международные наукометрические базы данных [Электронный ресурс]. - http://web-in-learning.blogspot.com/2012/11/blog-post_24.html

3. Scopus (Elsevier): Elsevier receives millionth response to Editor, Author and Reviewer Satisfaction Survey. - http://www.scopus.com/search/form/authorFreeLookup.url.

4. РИНЦ – Научная электронная библиотека. - http://www.elibrary.ru.

5. Bielefeld Academic Search Engine. - http://www.base-search.net.

6. Index Copernicus. Indeksacja czasopisma. - http://www.journals.indexcopernicus.com/search_article.php.

7. Springer Science + Business Media. - http:// www.springer.com.

8. Scrapy – a fast high-level screen scraping and web crawling framework. - http://scrapy.org.

9. MongoDB – an open-source document database. - http://ru.wikipedia.org/wiki/MongoDB.

10. Білощицький А.О. Наукометричні бази та індикатори цитування наукових публікацій / Білощицький А.О., Гогунський В.Д. // Інформаційні технології в освіті, науці та виробництві. – Вип. 4 (5). – О.: Бахва А.О., 2013.

11. Коляда А.С. Разработка проекта информационно-аналитической системы извлечения и обработки информации из наукометрических баз данных / Коляда А.С., Негри А.А., Колесникова Е.В. // Управління проектами: стан та перспективи. Матеріали IX Міжнар. наук.-практ. конф. – Миколаїв: НУК, 2013. – 348 с.

12. Про опублікування результатів дисертацій на здобуття наукових ступенів доктора і кандидата наук». – Наказ МОНмолодьспорту від 17 жовтня 2012 року № 1112 України.

13. Оборський Г.О. Нові тенденції і завдання щодо підготовки науковців вищої кваліфікації [Текст] / Оборський Г.О., Гогунський В.Д. // Інформаційні технології в освіті, науці та виробництві. – Вип. 2 (5). – О.: Бахва А.О., 2013. – С. 15 – 22. – [http://sbornik.college.ks.ua].

14. Про затвердження орієнтовних критеріїв оцінювання діяльності вищих навчальних закладів. - Наказ МОН України від 20.06.1013 р. № 809.

Статья поступила в редколлегию 21.10.2013

Рецензент: д-р техн. наук, проф. С.В. Руденко, Одесский национальный морской университет, Одесса.