Розглядається проблема обробки даних, представлених у публічному доступі глобальної мережі інтернет. Визначено завдання, рішення яких дозволяє вирішити проблему. Представлений спосіб вилучення інформації із слабоструктурованих веб сторінок на прикладі наукометричних баз даних. Розроблено програмне забезпечення для автоматизації процесу вилучення інформації із наукометричних баз даних і збереження їх з можливістю подальшої обробки

Ключові слова: веб сторінка, інтернет, інформація, слабоструктурований, вилучення

Рассматривается проблема обработки данных, представленных в публичном доступе глобальной сети интернет. Определены задачи, решение которых позволяет разрешить проблему. Представлен способ извлечения информации из слабоструктурированных веб страниц на примере наукометрических баз данных. Разработано программное обеспечение для автоматизации процесса извлечения информации из наукометрических баз данных и сохранения их с возможностью дальнейшей обработки

Ключевые слова: веб страница, интернет, информация, слабоструктурированный, извлечение

УДК 004.62

ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ ИЗ СЛАБО-СТРУКТУРИРОВАННЫХ ВЕБ СТРАНИЦ

А. С. Коляда
Аспирант*
E-mail: akolyada@gmail.com
В. Д. Гогунский
Доктор технических наук, профессор*
E-mail: vgog@i.ua
*Кафедра управления системами безопасности жизнедеятельности
Одесский национальный политехнический университет
пр. Шевченко, 1, г. Одесса, Украина, 65044

1. Введение

В области компьютерных наук информация занимает ключевое место. В глобальной сети Интернет информация предоставляется пользователю в виде веб страниц, которые не имеют строго определенной структуры. Возникает актуальная проблема извлечения данных из таких источников для дальнейшей работы с ними [1]. Применение способа веб скрапинга порождает задачу анализа и идентификации слабо структурированных данных. Слабоструктурированные представления данных отличаются отсутствием строгих структур таблиц и отношений в моделях реляционных баз данных, тем не менее, эта форма данных содержит теги и другие маркеры для отделения семантических элементов, а также для обеспечения иерархической структуры записей и полей в наборах данных [2].

2. Постановка проблемы и цель исследования

Развитие интернет-технологий в области организации хранилищ данных, репозиториев и электронных библиотек с предоставлением доступа к базам данных научных публикаций создает условия для развития исследований в разных областях знаний, которые в определенной мере отображаются в научных публикациях. Именно множество публикаций составляет основу формирования новых знаний. Разработка интеллектуального интерфейса для взаимодействия с различными наукометрическими базами данных позволит существенно упростить поиск информации [3].

Проблема заключается в анализе информации, содержащейся на веб-странице. Цель данного исследования – разработать способ извлечения информации из слабо структурированных веб страниц на примере наукометрических баз данных (НМБД).

3. Анализ предыдущих исследований

Глобальная сеть Интернет является самым большим источником данных, большая часть которых представляется в виде веб-страниц, не имеющих строго формализованной структуры. Извлечением информации из таких источников занимаются такие крупные корпорации как Google и Microsoft. Для качественного поиска используются сложные математические модели, семантический анализ и другие способы анализа информации. Поэтому данные, которые поступают на вход этих систем, должны быть структурированы определенным образом. Однако большинство наукометрических баз данных представлены в форме уникальных структур, что усложняет получение структурированных данных из подобных слабоструктурированных веб страниц.

Извлечение структурированных данных из веб страниц сводится к решению следующих задач [4 – 7]:

- поиску и получению целевых страниц для извлечения данных (проблема навигации);
- распознаванию участков, содержащих нужные данные (проблема распознавания данных);

- поиску структуры найденных данных (проблема поиска общей структуры данных);
- обеспечению однородности извлекаемых данных (проблема сопоставления атрибутов извлекаемых данных);
- объединению данных из разных источников (проблема объединения данных).

4. Модель программного обеспечения для извлечения информации из слабоструктурированных данных

Для решения задачи извлечения данных на примере наукометрических баз данных [8], предлагается модель программного обеспечения (рис. 1), которая состоит из следующих компонентов:

- программа для извлечения данных из конкретных НМДБ;
- блок фильтров извлеченных результатов;
- база данных для конечных результатов.

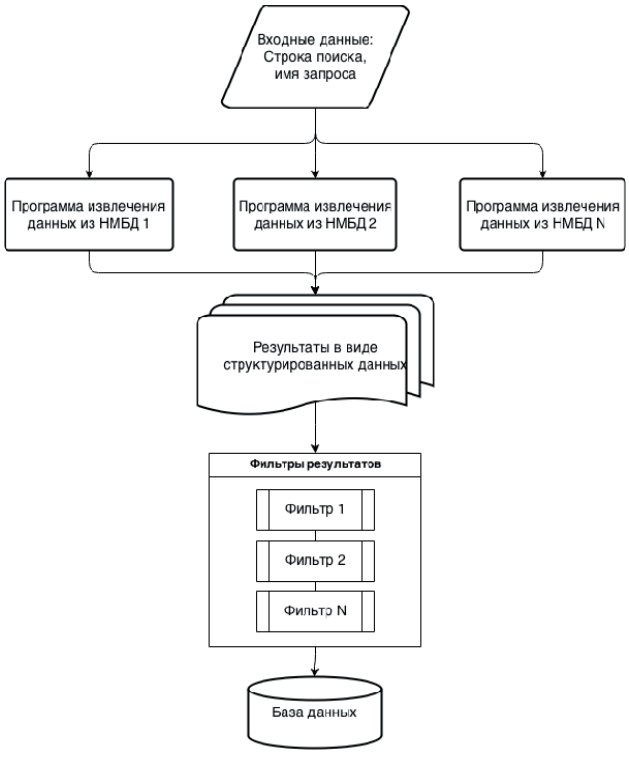

Рис. 1. Схема работы программного комплекса извлечения информации из наукометрических баз данных

Для каждой НМБД создается отдельная программа извлечения данных, так как все базы имеют разный интерфейс и структуру. Эти программы содержат в себе логику работы с конкретной НМБД, а также необходимые параметры, константные данные для выполнения этой работы.

После завершения работы программ извлечения данных выходные результаты каждой из них собираются в общий массив, который далее передается в блок фильтров.

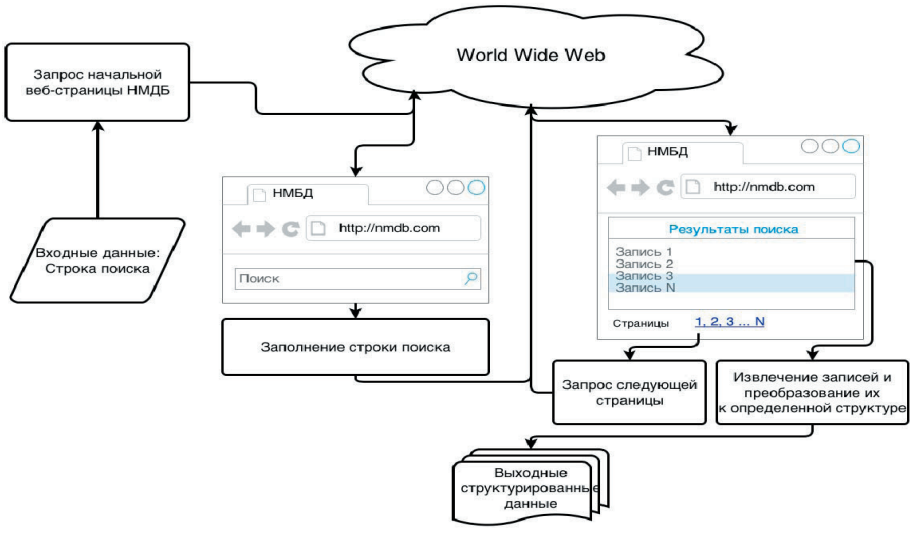

Рис. 2. Общая схема работы программы извлечения данных из НМДБ

Блок состоит из одного или нескольких фильтров, которые отбрасывают нерелевантные результаты, согласно некоторым параметрам, специфичным для этого фильтра.

Например, результаты программ извлечения данных могут содержать записи не соответствующие запросу поиска. Для этого можно использовать фильтр, который будет оставлять только результаты, соответствующие поисковой строке. Также можно использовать фильтр для отброса результатов однофамильцев, удаление дубликатов и т.д.

После обработки блоком фильтров, оставшийся набор результатов записывается в базу данных для дальнейшего представления и анализа.

4. 1. Процесс извлечения данных из веб страниц

Рассмотрим подробнее работу программы извлечения данных. На рис. 2 показана общая схема ее работы. Каждая программа может иметь различия в деталях из-за слабо структурированных данных.

На вход программы подается строка поиска. Далее выполняется запрос на загрузку начальной страницы конкретной НМБД, где программным путем имитируется работа пользователя браузера - вводится строка поиска в поле поиска и выполняется запрос на выдачу результатов. Из веб-страницы результатов, с помощью подпрограммы, извлекаются все поля каждой записи и преобразовываются в структуру строго определенную программным комплексом. Отсутствующие поля остаются пустыми. Таким же образом извлекаются ссылки на другие страницы, так как большое количество результатов может быть разделено на страницы. Далее выполняются запросы на эти ссылки, и процесс повторяется сначала, пока не будут обработаны все результаты, либо сработает ограничение на количество, которое для каждой НМБД устанавливается отдельно (один из параметров). В конце работы программы из-

влечения данных получаем набор структурированных данных, готовых для дальнейшей обработки.

Рассмотрим сам процесс извлечения информации со страницы НМБД. Веб страница, которую возвращает сервер, отформатирована с использованием языка разметки (в основном HTML), для дальнейшего отображения в том или ином виде с помощью специальной программы (веб-браузер). На рис. 3 показан пример визуализации веб-браузером некоторой области данных и исходный код этих данных. Здесь, например, название статьи “Features of Digital Devices Design of Modern PLD of the Xilinx Incorporation” заключено в следующие специальные последовательности символов, называемые тегами: <h1 class=”Title”>Здесь название статьи </h1>. Для извлечения этой информации, выполняется поиск этих тегов и извлекается их содержимое. Таким образом, заполняется одно из полей результатов. Для автоматизации этого процесса, программы извлечения данных используют язык запросов к элементам языка разметки (Xpath).

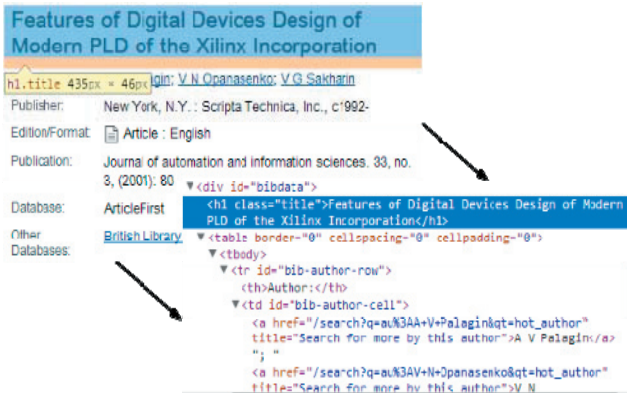

Рис. 3. Данные с веб-страницы и их исходный код языка разметки HTML

4. 2. Особенности извлечения данных на примере наукометрических баз данных

Исследуем особенности извлечения данных из наукометрических баз данных, поддерживаемых разработанным программным обеспечением. На текущий момент определена следующая структура данных для каждой публикации (табл. 1).

Таблица 1

Структура извлеченной информации

Поле	Описание
Наукометрическая база	Название базы источника публикации
Авторы	Авторы публикации
Название	Название публикации
Дата	Дата публикации
Источник	Источник публикации или издательство
Описание	Аннотация или краткое описание публикации
URL	Веб-ссылка на публикацию

BASE (base-search.net)

Наукометрическая база данных BASE позволяет выполнять поиск на разных языках и не задает строгих правил задания строки поиска (например, инициалы автора могут быть с точками или без них, а также слитно).

Результаты поиска подаются в визуально структурированном виде, имеются следующие поля: название публикации, автор(ы), предмет, издательство, год и URL. Но исходный код на языке разметки HTML имеет сложную структуру и вдобавок имена тегов зависят от языка интерфейса сайта.

Поэтому перед началом работы с этой базой, мы устанавливаем язык интерфейса – английский. Результаты поиска находятся внутри тегов-контейнеров с именем класса “ResultsContent”. Для каждого результата мы анализируем его содержимое: теги с именем класса “ItemLeft_en” содержат имя поля, а теги с именем класса “ItemRight_en” – значение. Далее адаптируем эту информацию под структуру данных (табл. 1) и получаем извлечённую запись.

Scopus (scopus.com)

Поиск по наукометрической базе данных Scopus выполняется только на латинице. При этом для фамилии и инициалов имеются два разных поля ввода. Инициалы следует указывать с точкой. Работа с этой базой данной имеет особенности, в основном, из-за того, что результаты поиска – это информация об авторе. Поэтому для Scopus окончательная структура данных расширена на 2 поля: количество документов и h-индекс.

Результаты поиска представляются в виде таблицы со следующими полями: автор(ы), количество документов, предмет и т. д. Если автор имеет ссылку на расширенную информацию, мы переходим по этой ссылке, а также записываем ее с соответствующее поле (URL).

На странице расширенной информации данные представлены в виде таблицы, состоящей из трех колонок: имя поля, разделитель, значение поля. Перебирая строки таблицы, мы заполняем выходную структуру данных.

Science Index (elibrary.ru)

Поиск поддерживается на многих языках. Для более эффективного поиска по автору, используется расширенный поиск, где указывается фамилия автора и инициалы, которые разделены пробелом.

Результаты поиска подаются в таблице, каждая строка которой содержит неструктурированную информацию: название статьи, автор(ы), источник, URL и дата публикации. Для адаптации этой информации под общую структуру, используются следующие манипуляции со строкой результата (рис. 4):

- название и URL публикации извлекается и тега <b>, который находится внутри тега <a>;
- авторы публикации извлекаются из тега <i>, который находится внутри первого тега <font>;
- из второго тега <font> извлекается дата и источник публикации.

Mlibrary(lib.umich.edu)

Наукометрическая база данных Мичиганского университета Mlibrary предоставляет поиск на латинице и имеет расширенный режим поиска для задания интересующих параметров.

Мы используем параметр “Автор” для поиска. Замечено, что запись инициалов через пробел выдает больше результатов.

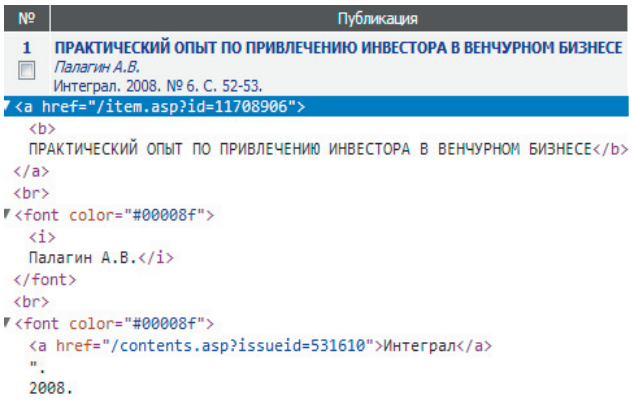

Рис. 4. Пример строки результата поиска в наукометрической базе данных Science Index

Результаты поиска выдаются в виде списка с названием публикации и ссылкой на полное описание. Мы переходим по этим ссылкам и извлекаем информацию из содержимого. Структура исходной информации представлена в следующем виде: теги с именем класса “article-field-label” содержат имя поля, теги с именем класса “article-field-value” – значение. Проходом по всем полям мы извлекаем информацию только интересующих нас (табл. 1).

WorldCat(worldcat.org)

Поиск по базе WorlCat также выполняется на латинице с использованием расширенного режима, где указываем параметр “Автор” и “Формат публикации – статья”. Как и с базой Mlibrary, инициалы автора в строке поиска следует указывать разделенные пробелом.

Результаты поиска – список публикаций с кратким описанием и ссылкой на полное описание. Снова мы переходим по всем ссылкам и работаем с информацией на этих страницах. Содержимое страниц этой наукометрической базы данных имеет хорошо выраженную структуру, что очень редко для веб-страниц. Здесь каждое поле имеет свой идентификатор, по которому мы извлекаем значение. Например, идентификатор “bib-author-cell” содержит значение поля “Авторы”, а “bib-publisher-cell” – значение поля “Издательство”. Таким образом, мы легко заполняем свою структуру данных (табл. 1).

5. Выводы

Самым распространенным примером слабоструктурированных данных являются веб страницы глобальной сети Интернет.

Информация на веб страницах слабо структурирована и не формализована, что затрудняет автоматизацию извлечения данных. На примере наукометрических баз данных, доступ к содержимому которых доступен только через веб интерфейс, предложен способ извлечения информации из таких слабо структурированных данных. Для каждой наукометрической базы данных описывается набор правил, с помощью которых извлекается информация и приводится к определенной типовой структуре.

Дальнейшие исследования состоят в разработке алгоритма усовершенствованной фильтрации нерелевантных данных, так как результаты, которые возвращают наукометрические базы данных, могут содержать сторонние записи. Для анализа и обработки извлеченных данных планируется исследовать применение семантического анализа к ним, что позволит классифицировать знания, содержащиеся в этих данных [9, 10].

Литература

1. Коляда, А. С. Автоматизация извлечения информации из наукометрических баз данных [Текст] / А. С. Коляда, В. Д. Тогунский // Управління розвитком складних систем. 2013. – № 16.
2. Buneman, Peter Semistructured data, Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems [Text] / Peter Buneman. – Tucson, Arizona, United States. – May 11 15, 1997. – P.117 121.
3. Бурков, В. Н. Параметры цитируемости научных публикаций в наукометрических базах данных [Текст] / В. Н. Бурков, А. А. Белощицкий, В. Д. Тогунский // Управління розвитком складних систем. – 2013. – № 15. – С. 134 – 139.
4. Arens, Yigal. Retrieving and integrating data from multiple information sources [Text] / Yigal Arens, Chin Y. Chee, Chun-Nan Hsu, Craig A. Knoblock // International Journal of Intelligent and Cooperative Information Systems. – 1993. – Issue 02
5. Yung-Jen Hsu, Jane. Template-based information mining from HTML documents [Text] / Jane Yung-Jen Hsu, Wen-tau Yih // Proceedings of the fourteenth national conference on artificial intelligence and ninth conference on Innovative applications of artificial intelligence. – 1997. – P. 256 – 262.
6. Smith, Dan. Information extraction for semi-structured documents [Text] / Dan Smith, Mauricio Lopez // In Proceedings of the Workshop on Management of Semistructured Data. – 1997.
7. Li, Zhao. Web data extraction based on structural similarity [Text] / Zhao Li, Wee Keong Ng, Aixin Sun // Journal Knowledge and Information Systems archive. – November 2005. – Vol. 8, Issue 4. – P. 438 – 461.
8. Коляда, А. С. Разработка проекта информационно-аналитической системы извлечения и обработки информации из наукометрических баз данных [Текст] : матеріали IX Міжнар. наук.-практ. конф / А. С. Коляда, А. А. Негри, Е. В. Колесникова // Управління проектами: стан та перспективи. — Миколаїв : НУК, 2013. — 348 с.
9. Палагин, А. Формализация проблемы извлечения знаний из естественно языковых текстов [Текст] / А. Палагин, С. Кривый, Н. Петренко, Д. Бибиков. — Sofia : Information technologies & knowledge, 2012. — 100 с.
10. Baumgartner, Robert The Personal Publication Reader: Illustrating Web Data Extraction, Personalization and Reasoning for the Semantic Web [Text] / Robert Baumgartner, Nicola Henze, Marcus Herzog // Lecture Notes in Computer Science 2005. – Vol. 3532. – P 515–530.