УДК 065.012

В.Д. Гогунський,
д.т.н., професор,
Одеський
національний
політехнічний
університет
victor@3g.ua

А.С. Коляда,
аспірант,
Одеський
національний
політехнічний
університет
akolyada@gmail.com

РАЗРАБОТКА ПРОГРАММНОГО ПРОЕКТА ДЛЯ ИЗВЛЕЧЕНИЯ И ОБРАБОТКИ ИНФОРМАЦИИ ИЗ НАУКОМЕТРИЧЕСКИХ БАЗ ДАННЫХ

А.С. Коляда, В.Д. Гогунський. **Разработка програмного проекта для извлечения и обработки информации из наукометрических баз данных.** Рассматриваются требования и архитектура проекта по извлечению публикаций из наиболее известных наукометрических баз данных. Описаны внешние и внутренние интерфейсы системы, приведен пример последовательности работы пользователя с программным продуктом.

A.S. Kolyada, V.D. Gogunsky. Development of software project for information extraction and processing from scientometric databases. *The requirements and architecture of the project for extracting publications from the most prominent scientometric databases are shown. Described the external and internal interfaces of a system, also an example of the user interactiong sequence with the software product is presented.*

Введение. В связи с 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” возникает задача проверки наличия публикаций соискателя в международных наукометрических базах данных. Для решения этой проблемы в [1] разработан способ извлечения данных о публикациях по параметру «Автор» из наиболее известных наукометрических баз данных. Данная статья описывает реализацию этого способа в программном проекте, архитектуру проекта и использованные средства. В связи с требованием индексации публикаций в международных наукометрических базах возникает проблема поиска и идентификация своих публикаций автором в них. С подобной проблемой сталкиваются многие научные сотрудники, аспиранты и преподаватели при подготовке отчетности по научной работе. Задачей данного программного продукта является предоставить список публи-

каций соискателя, которые индексируются в международных наукометрических базах даннях [2].

Материал и результаты исследования.

Одной из первых стадий разработки программного проекта является сбор информации, анализ, спецификация, и проверка требований к программному обеспечению [3]. Программные требования – свойства программного обеспечения, которые должны быть надлежащим образом представлены в нём для решения конкретных практических задач.

К данному проекту представлены следующие требования:

- извлечение информации из веб страниц;
- критерием информации является ФИО автора;
- работа с наиболее распространенными наукометрическими базами данных: Scopus, Web of Science, Base Search, Science Index, Copernicus, WorldCat, ScienceDirect, DOAJ, Springer, Google Scholar;
- обработка результатов с целью определения нерелевантной информации и фильтрации ее;
- предоставление информации пользователю.

На рис. 1 показан высокоуровневый дизайн программного проекта. Здесь видно, что он состоит из двух основных приложений: интерфейс пользователя (frontend) и основного приложения (backend). Основное приложение предоставляет интерфейс, через который интерфейсное приложение получает доступ к основному функционалу для предоставления его наружу либо для отображения в графическом виде пользователю. Таким образом, пользователь или клиентская программа имеет доступ только к интерфейсной части, скрывая детали реализации основного приложения.

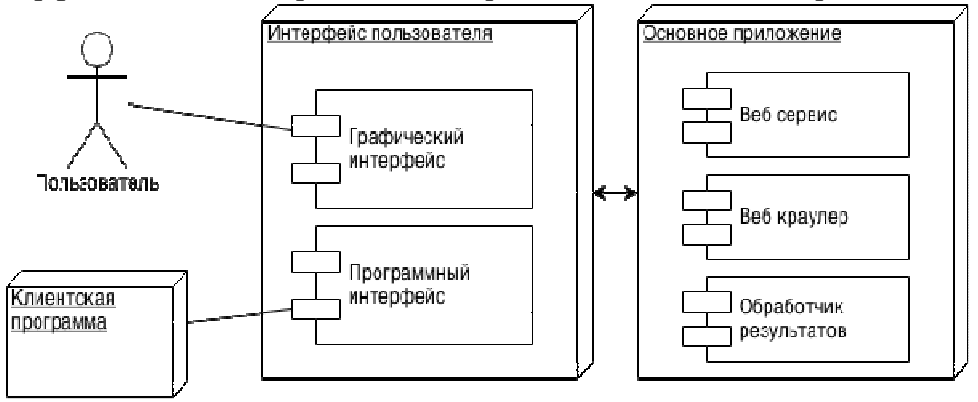


Рис 1. Высокоуровневый дизайн проекта по извлечению информации из наукометрических баз данных

Оба интерфейса (внешний – для клиентских программ и внутренний – для коммуникации с основным приложением) в качестве формата данных используют JSON – открытый стандартный формат обмена данными, который использует читабельный для человека текст. Описание интерфейса основного приложения показано на рис. 2. Также здесь видно, что основное приложение состоит из трех модулей:

- Веб сервис – предоставляет доступ к приложению с помощью протокола HTTP, используя JSON интерфейс;
- Веб краулер – ядро приложения, где и выполняется извлечение публикаций по заданному параметру;
- Обработчик результатов – фильтрация и анализ извлеченных данных с целью отброса нерелевантных результатов и классификации.

Одной из проблем, возникших после извлечения публикаций, является наличие авторов-однофамильцев. Нужен способ отделить публикации только искомого автора. Для этого модуль «обработчик результатов» использует латентно семантический анализ [4]. Цель этого анализа – определение схожих по смыслу публикаций, исходя из их названий.

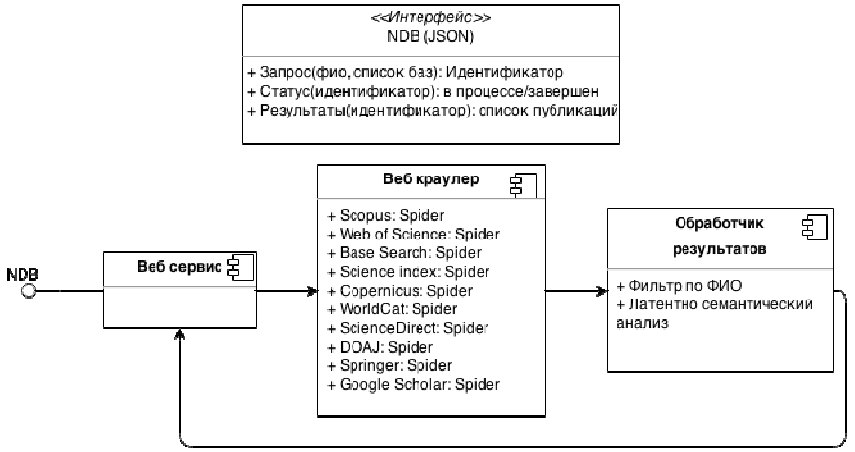


Рис 2. Детализированная архитектура основного приложения

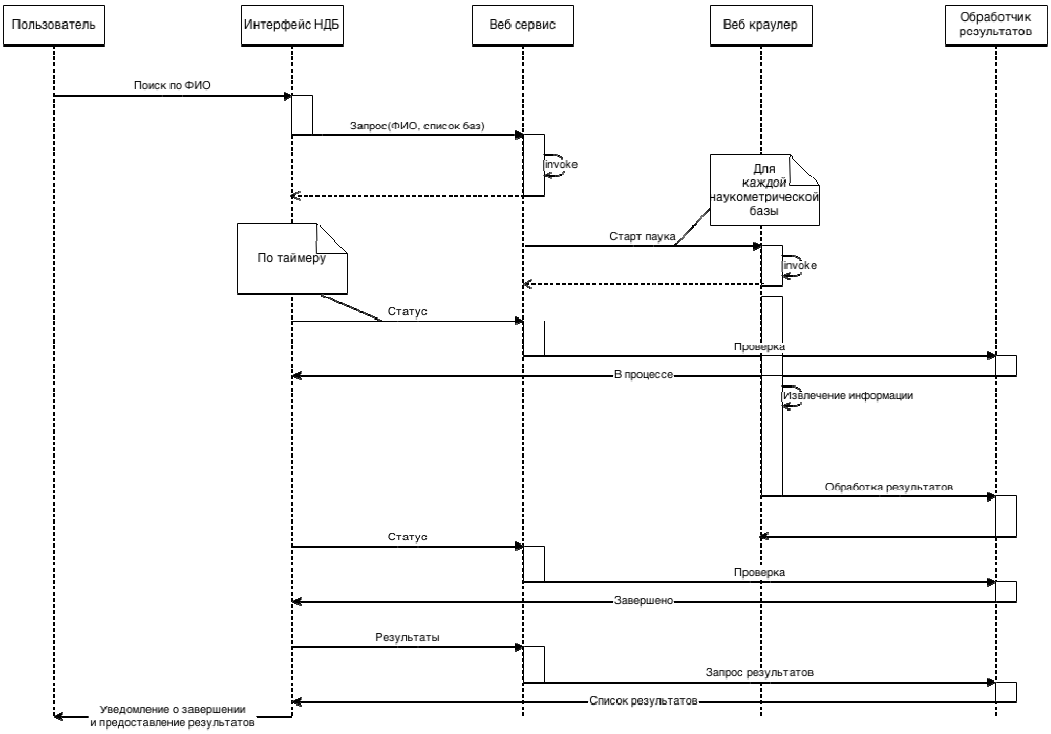
Еще одной значимой проблемой при извлечении данных из веб страниц оказался тот факт, что некоторые части результирующей страницы сгенерированы на стороне клиента, а не получены с сервера после соответствующего запроса. Речь идет о скриптовых программах, встроенных в веб страницу, которые выполняются клиентской программой – веб браузером. Для решения этой проблемы нужно использовать веб браузер для генерации веб страницы в конечном виде и потом только продолжать ра-

боту с ней по извлечению данных. Но использование полноценного браузера слишком накладно с точки зрения ресурсов, из-за того, что он является приложением с графическим интерфейсом, ориентированным на пользователя. Нам же нужно программное взаимодействие с ним.

Решение является в использовании, так называемого, «безголового» браузера (headless browser).

В основном они содержат открытый браузерный «движок» WebKit, который в последних версиях не требует даже наличия графической подсистемы для загрузки и обработки веб страницы. Одна из наиболее распространённых реализаций такого браузера является PhantomJS [5].

На рис. 3 показана диаграмма последовательности работы с программным продуктом. Пользователь (или клиентская программа) посылает запрос на извлечение публикаций по заданному ФИО. На экране пользователя отображается, что процесс запущен и статус, который обновляется с прогрессом поиска. Параллельно отправляется запрос в основное приложение, где запускаются все «пауки», извлекаются полученные результаты и обрабатываются. Интерфейсное приложение с некоторым интервалом опрашивает текущий статус. По окончании работы, обработчик обновляет статус и пользователю предоставляются результаты.



Управління проектами та якістю

Рис 3. Диаграмма последовательности выполнения поиска и извлечения информации из наукометрических баз данных

Выводы. Не смотря на возникшие трудности и проблемы, программный проект по извлечению публикаций из наукометрических баз данных реализован и работает в тестовом режиме. Основной проблемой для поддержки проекта является отсутствие спецификаций на Веб страницы наукометрических баз данных. В любой момент структура и внешний вид могут быть изменены. Дальнейшее развитие проекта предполагает решение этой проблемы с помощью тестовых приложений, которые с некоторой периодичностью будут проверять, не изменилась ли структура и сигнализировать о изменениях для идентификации и исправления.

Литература:

1.Коляда, А. С. Автоматизация извлечения информации из наукометрических баз данных [Текст] / А. С. Коляда, В. Д. Гогунский // Управління розвитком складних систем. – 2013. – № 16. – С. 96 – 99.

2.Бурков, В. Н. Параметры цитируемости научных публикаций в наукометрических базах данных [Текст] / В. Н. Бурков, А. А. Белощицкий, В. Д. Гогунский // Управління розвитком складних систем. – 2013. - № 15. – С. 134 – 139.

3.IEEE Computer Society, edited by Pierre Bourque, Richard E. Fairley (2014). Guide to the Software Engineering Body of Knowledge (SWEBOK®), pp. 346.

4.Коляда, А. С. Латентно семантический подход для анализа информации из наукометрических баз данных [Текст] / А. С. Коляда // Управління розвитком складних систем. – 2014. – Вып. 17. – С. 90 – 94.

5.PhantomJS – a headless WebKit. [Електронный ресурс] // http://phantomjs.org/

Надійшла до редакції 17.04.2014