УДК 004.62, 004.623, 930.25

Шевченко В.А., Васецкий А.М.

АВТОМАТИЗАЦИЯ СБОРА БОЛЬШИХ ОБЪЁМОВ ДАННЫХ С ИНТЕРНЕТ САЙТА

Шевченко Владислав Андреевич, студент 4 курса бакалавриата факультета информационных технологий и управления;
Васецкий Алексей Михайлович, старший преподаватель кафедры информационных компьютерных технологий, e-mail: amvas@muctr.ru;
Российский химико-технологический университет им. Д.И. Менделеева, Москва, Россия
125480, Москва, ул. Героев Панфиловцев, д. 20

Рассматриваются вопросы автоматического скачивания материалов сайта по заданным критериям поиска. На первом этапе двухступенчатого алгоритма формируется подборка документов, удовлетворяющих заданному пользователем запросу к внутренней поисковой системе сайта на базе ElasticSearch. На втором этапе происходит автоматическое скачивание всех страниц, относящихся к выбранным документам. Для этого используется разработанное нами программное обеспечение на основе .NET фреймворка и загрузчика Wget.

Ключевые слова: *ElasticSearch; поиск; скрипт; сайт; электронный архив; Wget.*

AUTOMATION OF GATHERING OF LARGE VOLUMES OF DATA FROM THE INTERNET SITE

Shevtchenko V.A., Vasetskiy A.M.
D. Mendeleev University of Chemical Technology of Russia, Moscow, Russia

The questions of automatic downloading of the site materials according to the specified search criteria are considered. At the first stage of the two-step algorithm, a collection of documents is created that satisfy the user-specified request to the internal ElasticSearch search engine of the site. At the second stage, all pages referring to the selected documents are downloaded automatically. To do this, we use the software developed by us based on the .NET framework and the Wget downloader.

Keywords*: ElasticSearch; script; site; archive; Wget; collection; request.*

В последние годы большое развитие получили сайты, содержащие большой и сверхбольшой объём данных. Однако разработчики зачастую намеренно не предоставляют достаточно удобных сервисов пользователям, заставляя тех производить большое количество рутинных ручных операций для сохранения результатов своей работы. В частности, к подобным ограничениям относится постраничное сохранение документов, а также сокрытие разработчиками информации о структуре записей, хранящихся у них на сайте. Например, такой подход характерен для ряда электронных архивов и библиотек. Причины такого отношения к исследователям и простым пользователям лежат вне предметной области данной статьи.

Нами была предпринята попытка облегчить работу с данными для пользователей, используя легальные механизмы доступа к информации. В качестве примера рассматривался сайт *«Память народа»* [1], разработанный корпорацией ЭЛАР. Данный исторический ресурс содержит большое количество отсканированных архивных документов в графическом формате. В частности, среди них находятся документы по боевым действиям в ходе Великой Отечественной войны 1941-45 гг. из хранилища Центрального архива министерства обороны РФ (ЦАМО).

На основании работы [2] количество оперативных документов, хранящихся на данный момент в базе данных сайта [1], составляет по уточнённым нами данным 3430708 записей, и их объём оценивается примерно в 60 терабайт. Более подробно распределение данных по годам приведено на рис.1.

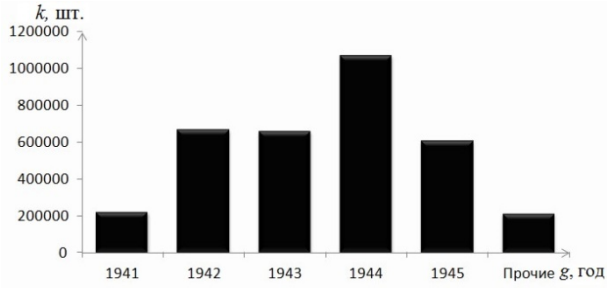

Рис.1. Распределение записей по годам в базе данных сайта [1]; *k* – количество записей; *g* – год записи

Следует отметить, что в ходе работы над этой темой нами был выявлен ряд явных ошибок при датировке. Встречались записи, атрибутированные совершенно некорректными годами, как результаты описок или опечаток. К сожалению, работники ЭЛАР никак не отреагировали на наши рекомендации по исправлению атрибутов документов.

Обращение к базе данных рассматриваемого сайта [1] заключается в использовании пользовательских запросов на базе поисковой системы ElasticSearch [3]. Следует отметить, что аналогичная система используется и в ряде других сайтов [4,5].

Интерфейс поискового механизма включает в себя краткую и расширенную формы. По умолчанию

мы будем рассматривать его наиболее полную версию, поскольку с точки зрения синтаксиса запросов они идентичны. В формировании запроса участвуют поля, получаемые из пользовательского интерфейса, а также технические поля, заполняемые автоматически. Следует также отметить наличие скрытых полей, недоступных через интерфейс сайта. Структурно все они могут быть разбиты на текстовые, числовые, смешанные, логические, списки и поля дат.

Наличие и количество полей каждого типа регулируются разработчиками в зависимости от решаемых задач и могут меняться с течением времени. Также периодически отмечаются изменения, вносимые в формат запросов и ответов разработчиками. Поэтому следует ожидать и дальнейших изменений *JSON*-структур, примеры которых приведены далее в статье. Здесь *JSON* (*JavaScript Object Notation*) – текстовый формат обмена данными, основанный на *JavaScript*.

Общий вид структуры запроса к разделу «Документы частей» несколько отличается от ранее рассмотренного в работе [2] и выглядит теперь следующим образом (приводится с сокращениями):

{"query":{"bool":{"should":[{"bool":{"should":[{"match_phrase":{"document_type":"Боевые донесения, оперсводки"}},
{"match_phrase":{"document_type":"Боевые приказы и распоряжения"}},
{"match_phrase":{"document_type":"Отчеты о боевых действиях"}},
{"match_phrase":{"document_type":"Переговоры"}},
{"match_phrase":{"document_type":"Журналы боевых действий"}},
{"match_phrase":{"document_type":"Директивы и указания"}} ,
{"match_phrase":{"document_type":"Приказы"}}, …},
{"bool":{"should":[{"bool":{"must":[{"range":{"date_from":{"lte":"1941-6-30"}}},
{"range":{"date_to":{"gte":"1941-1-1"}}}],"boost":3}},
{"bool":{"must":[{"range":{"document_date_b":{"lte":"1941-6-30"}}},
{"range":{"document_date_f":{"gte":"1941-1-1"}}}],"boost":1}}]}} ,
{"bool":{"should":[{"match":{"document_name":{"query":"100 сд","type":"phrase"}}},
{"match":{"authors":{"query":"100 сд","type":"phrase"}}},
{"match":{"army_unit_label.division":{"query":"100 сд","type":"phrase"}}},
{"nested":{"path":"page_magazine","query":{"bool":{"should":[{"match":
{"page_magazine.podrs":{"query":"100 сд","type":"phrase"}}}]}}}}}],
"minimum_number_should_match":3}},"_source":["id","document_type","document_number","document_date_b","document_date_f","document_name","archive","fond","opis","delo","date_from","date_to","authors","geo_names","operation_name","secr","image_path","delo_id","deal_type","operation_name"],"size":10,"from":0}.

Пользователь может заполнять с помощью интерфейса сайта только некоторые поля, например, *document_name*, *document_type*, *document_date_b*, *document_date_f*, *authors*, *operation_name.* Следует также отметить, что ряд полей, таких как *army_unit_label.division, page_magazine.podrs*, и, вероятно, им подобные, заполняются автоматически копиями поля *document_name*. Это связано со структурой атрибутирования записей, находящихся на сайте, и в данной статье не рассматривается.

Следует отдельно отметить весьма громоздкую конструкцию полей вида *document_type,* которая занимает примерно половину от общего объёма текста запроса.

В то же время поля *fond*, *opis*, *delo*, отвечающие за наиболее важные при цитировании архивных материалов ссылки на номера фонда, описи и дела, в данный момент пользователям не предоставляются. Поле *size* отвечает за количество выдаваемых документов и по умолчанию равно 10.

Ответ на запрос в виде *JSON*-структуры выглядит следующим образом (приводится только первая запись из 27):

{"took":67,"timed_out":false,"_shards":{"total":1,"successful":1,"failed":0},"hits":{"total":27,"max_score":26.933163,"hits":[{"_index":"pamyat_2017_05_03","_type":"document","_id":"10104371","_score":26.933163,
"_source":{"document_number":"3","deal_type":null,"archive":"ЦАМО","delo":"4","date_to":null,"fond":"807",
"operation_name":null,"document_name":"Разведсводка штаба 100 сд",
"geo_names":null,"document_date_b":"1941-06-27","secr":"ns","image_path":"Передача_008_КП097 P-C28/807-0000001-0004/00000015.jpg","delo_id":0,"id":10104371,"document_date_f":"1941-06-27",
"document_type":"Разведывательные бюллетени и донесения","opis":"1","date_from":null,"authors":"100 сд, Коган, Ященко"}},…

Для исследователей представляют интерес далеко не все поля. В частности, полезную информацию можно почерпнуть из следующих полей:

– *total* – общее количество найденных записей;

– *id* – индивидуальный номер записи;

– *fond*, *opis*, *delo* – фонд, опись, дело – реквизиты архивной записи;

– *document_name* – название документа;

– *document_date_b* – начальная дата документа;

– *authors* – авторы документа.

Остальные поля являются либо техническими, либо не столь информативными.

В ходе выполнения работы [2] было разработано программное обеспечение на языке C++ с использованием фреймворка *.NET*, позволяющее напрямую работать с поисковым механизмом сайта *«Память народа»* без применения интернет-браузеров. Его дальнейшее развитие позволяет реализовать следующий расширенный набор функций:

– формирование запросов различного вида;

– кодирование запроса в формате *UTF*-8;

– отсылка запроса в поисковый механизм сайта методом *POST*;

– получение ответа в виде *JSON*-структуры;

– обработка (декомпозиция) ответа;

– формирование файла задания для менеджера закачек *Wget*.

Таким образом, используя имеющийся в нашем распоряжении инструментарий, можно реализовать двухступенчатый алгоритм работы, представленный ниже:

1) получение от пользователя исходных данных, включая те поля, которые недоступны с официального интерфейса сайта;

2) формирование и отсылка запроса к внутренней поисковой системе сайта;

3) получение ответа в виде *JSON*-структуры;

4) декомпозиция ответа и вывод результатов пользователю в табличном виде;

5) получение от пользователя задания на интересующие его документы;

6) последовательное формирование запросов на индивидуальные страницы выбранных документов;

7) обработка ответов поисковой системы и получение прямых ссылок на страницы;

8) формирование файла-задания для скачивания выбранных страниц менеджером закачек *Wget* с учётом формируемой структуры хранения страниц документов;

9) запуск *Wget* с заданными параметрами;

10) ожидание завершения работы *Wget*.

Пример запроса на получение страниц выбранного пользователем документа:

{"query":{"bool":{"must":[{"match":{"document_id":10104371}}]}},"sort":[{"id":{"order":"asc"}}],"size":10000}

Пример ответа на него (для одной страницы записи):

{"took":2,"timed_out":false,"_shards":{"total":1,"successful":1,"failed":0},"hits":{"total":1,"max_score":null,"hits":[{"_index":"pamyat_2017_05_03","_type":"page_document","_id":"10117333","_score":null,
"_source":{"operation":"I","source_id":10117333,"id":10117333,"delo_id":0,"document_id":10104371,
"position":14,"list":null,"secr":"ns","censor":null,"image_path":"Передача_008_КП097Р-С28/807-0000001-0004/00000015.jpg",
"action":"U"},"sort":[10117333]}]}}

Таким образом, обрабатывая постранично каждый документ, можно извлечь из него путь к каждому изображению. В вышеприведённом фрагменте это *Передача_008_КП097Р-С28/807-0000001-0004/00000015.jpg*. Далее, эта ссылка преобразуется в полный путь к файлу изображения.

На последнем этапе формируется файл-задание для менеджера закачек *Wget*. Помимо этого, в данном файле находится необходимый набор настроек менеджера закачек, необходимых для скачивания файлов, поскольку прямые ссылки в чистом виде сайтом не поддерживаются.

После передачи файла-задания в менеджер закачек пользователю остаётся только контролировать стабильность связи с сайтом, которая до сих пор остаётся довольно слабым местом его работы. Чтобы не создавать дополнительной нагрузки на него наша программа ведёт скачивание файлов последовательно, а не параллельно. Тем не менее, данная схема работы позволяет достаточно быстро получить любой многостраничный документ, даже если он содержит несколько десятков или даже сотен страниц.

Список литературы

1. Архивный сайт «Память народа» [Электронный ресурс]. Режим доступа: https://pamyat-naroda.ru (дата обращения 27.05.2017).

2. Васецкий А.М., Лисовский А.А., Филиппова Е.Б. Расширение функционала поискового механизма сайта программными средствами // Успехи в химии и химической технологии. 2016. Т. 30, № 4 (173). С. 103-105.

3. Сайт «Elastic» [Электронный ресурс]. Режим доступа: https://www.elastic.co/ (дата обращения 27.05.2017).

4. Обобщённый электронный банк данных «Мемориал» [Электронный ресурс]. Режим доступа: https://www.obd-memorial.ru (дата обращения 27.05.2017).

5. Архивный сайт «Подвиг народа» [Электронный ресурс]. Режим доступа: http://www.podvignaroda.mil.ru (дата обращения 27.05.2017).