УДК 002.513.5:004.912

И. Р. Ахмадеева, **Ю. А. Загорулько**, **Н. В. Саломатина**
А. С. Серый, **Е. А. Сидорова**, **В. К. Шестаков**

Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

Институт систем информатики им. А. П. Ершова СО РАН
пр. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия

Институт математики им. С. Л. Соболева СО РАН
пр. Акад. Коптюга, 4, Новосибирск, 630090, Россия

E-mail: ah.irishka@gmail.com; zagor@iis.nsk.su; nataly@math.nsc.ru
Alexey.Seryj@iis.nsk.su; lena@iis.nsk.su; shestakov@iis.nsk.su

ПОДХОД К ФОРМИРОВАНИЮ ТЕМАТИЧЕСКИХ КОЛЛЕКЦИЙ ТЕКСТОВ НА ОСНОВЕ ИНТЕРНЕТ-РЕСУРСОВ [*]

Рассматривается задача автоматического формирования коллекций текстов, соответствующих заданным тематикам. Для ее решения предложен подход и разработана система, использующая для сбора текстов механизмы метапоиска и специализированные средства для работы с вики-ресурсами. Проведенные эксперименты с системой показали продуктивность предложенного подхода.

Ключевые слова: текстовые коллекции, интернет-ресурсы, вики-ресурсы, поисковый запрос, метапоиск.

Введение

В современном мире информация стала одним из наиболее важных и ценных ресурсов. Эффективность деятельности отдельных людей, коллективов и организаций все в большей степени зависит от их способности использовать имеющуюся в их распоряжении информацию. В связи с этим все большее внимание уделяется процессам сбора, накопления, систематизации и обработки информации.

Существуют различные виды представления информации: графический, текстовый, аудио, видео, но наибольшей ценностью обладает информация, представленная в текстовом виде. Одним из способов систематизации текстовой информации являются тематические текстовые коллекции, т. е. массивы текстов, относящихся к определенной тематике. Такие коллекции могут играть как чисто информационную роль, так и использоваться при решении широкого класса задач, таких как машинное обучение (алгоритмов), построение и наполнение онтологий, составление различного вида словарей (морфологических, синони-

[*] Работа выполнена при поддержке Министерства образования и науки Российской Федерации (соглашение № 02.G25.31.0054).

Ахмадеева И. Р., Загорулько Ю. А., Саломатина Н. В., Серый А. С., Сидорова Е. А., Шестаков В. К. Подход к формированию тематических коллекций текстов на основе интернет-ресурсов // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2013. Т. 11, вып. 4. С. 5–15.

ISSN 1818-7900. Вестник НГУ. Серия: Информационные технологии. 2013. Том 11, выпуск 4

мов, тезаурусов и т. п.) и рефератов, кластеризация (например, для выявления трендов в заданных областях знаний) и т. д.

В зависимости от дальнейшего использования коллекции к ней предъявляются определенные требования, накладывающие ограничения как на ее количественные, так и качественные характеристики. Такими требованиями могут быть, например, определенный уровень релевантности тематике, отсутствие рекламы и текстов-дубликатов, высокое качество текста, ограничения на объем текста и др.

Текстовые коллекции можно разделить на несколько типов: электронные библиотеки, полнотекстовые базы данных и отдельные корпуса текстов, построенные в соответствии с определенными принципами [1]. У каждой коллекции есть своя специфика, определяемая способом пополнения коллекции и целями, преследуемыми ее создателями. В частности, различается ручной, полуавтоматический и автоматический способы пополнения коллекций. Например, для бесплатных электронных библиотек характерны такие методы комплектования, как ручное копирование с других сайтов или получение текстов от добровольцев (в том числе и от самих авторов), в то время как для коммерческих полнотекстовых баз данных – либо сканирование и распознавание печатных оригиналов, либо покупка электронных копий непосредственно в издательствах. При составлении корпусов текстов также могут использоваться как ручные, так и полуавтоматические и даже полностью автоматические методы, при этом источники текстов выбираются в зависимости от поставленных задач.

Качество текстов, как правило, зависит от типа коллекции. Например, ошибок и опечаток в бесплатных электронных библиотеках будет больше, чем в коммерческих, потому что у последних есть средства на оплату работы корректоров.

Тип коллекции обычно определяет и способ представления ее содержимого пользователям. Например, электронные библиотеки чаще всего предоставляют возможности для просмотра отдельных документов и внутреннего поиска (более развитого в коммерческих версиях), в то время как корпусы текстов обычно предоставляются только целиком единым массивом данных.

В данной работе рассматривается задача автоматического формирования коллекций текстов, соответствующих заданным тематикам. При решении этой задачи в качестве источников данных предлагается использовать интернет-ресурсы двух типов. Ресурсы первого типа представляют собой интернет-страницы (сайты), полученные с помощью поисковых сервисов общего назначения. Для их сбора реализуется механизм метапоиска. Ресурсы второго типа – это сайты, заранее выбранные экспертами. Они характеризуются большей «надежностью», имеют известный формат и, возможно, обладают собственным тематическим классификатором. К последним, в частности, отнесены и вики-ресурсы. Для ресурсов второго типа возможно применение специальных методов обработки, учитывающих их специфику, в то время как для ресурсов первого типа необходимо применять универсальные методы, ориентированные на произвольный формат данных.

В статье описывается архитектура и принципы работы системы, выполняющей задачу формирования тематических текстовых коллекций на основе интернет-ресурсов рассмотренных выше двух типов, рассматривается применение механизмов метапоиска для сбора текстов, а также возможность извлечения текстов заданной тематики из вики-ресурсов. Обсуждаются результаты эксперимента по применению разработанной системы к построению текстовой тематической коллекции.

Назначение, **принципы работы и архитектура системы**

Предлагаемая система предназначена для автоматического формирования тематических текстовых коллекций на основе интернет-ресурсов. При этом формируемые текстовые коллекции должны обладать определенными качественными и количественными характеристиками, которые задаются экспертом перед созданием коллекции.

Качественные характеристики, как правило, определяют следующие свойства коллекций:

- высокий уровень релевантности текстов коллекции заданной теме;
- однородность текстов по стилю или жанру (по заданному набору стилей);
- равномерность распределения по подтемам;

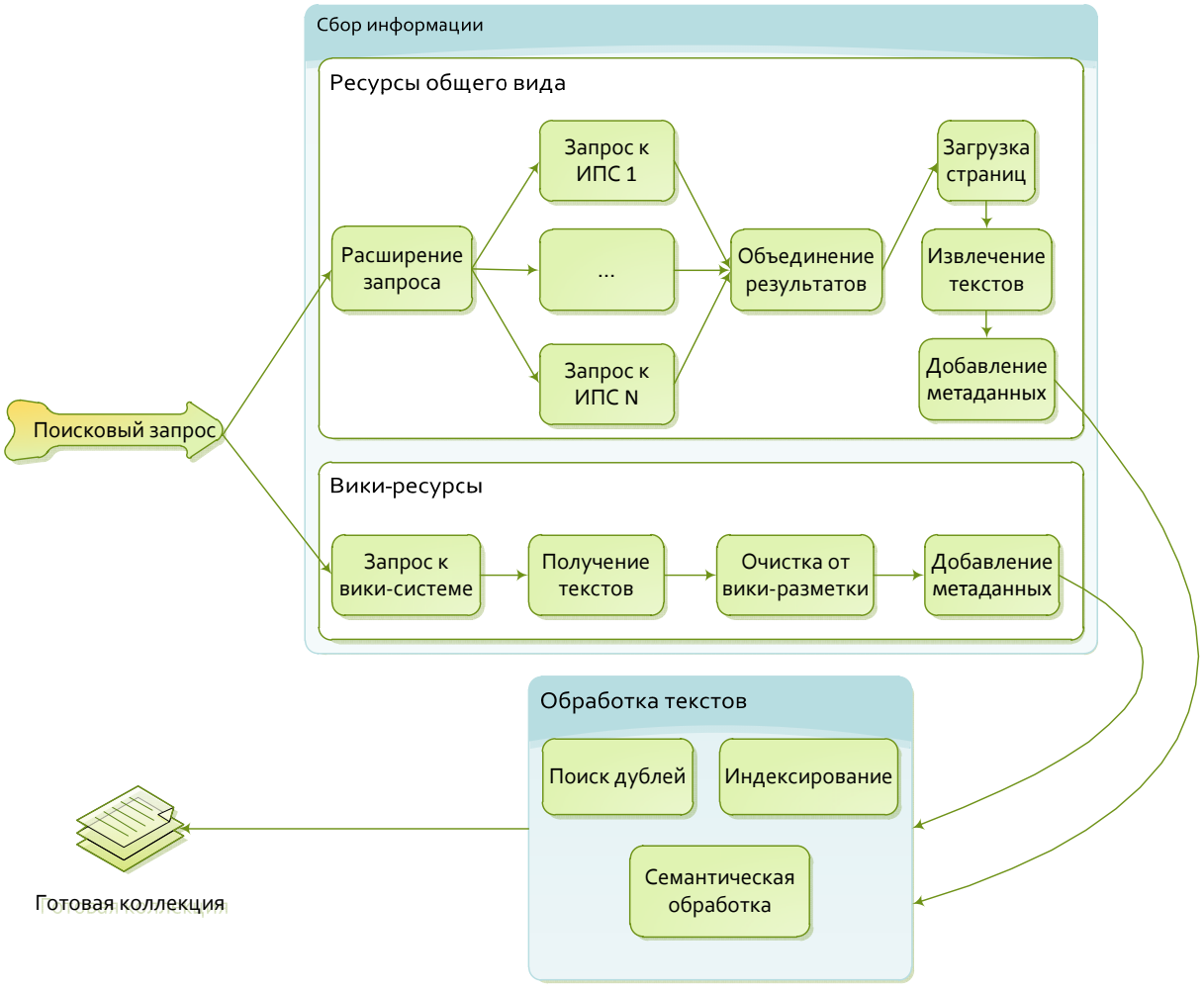
- низкий уровень ошибок (опечаток);
- отсутствие текстов-дубликатов (в том числе семантических);
- отсутствие фрагментов текста, не несущих полезной смысловой нагрузки (реклама, общее меню и т. п.);
- отсутствие избыточных элементов форматирования текста.

Количественные характеристики задают ограничения на общий объем коллекции и объем каждого текста (как сверху, так и снизу).

Рассматриваемая система состоит из нескольких взаимосвязанных компонентов, которые объединяются в два основных блока: блок сбора информации и блок обработки текстов (см. рисунок). Для сбора текстов используются механизмы метапоиска и специализированные средства для работы с вики-ресурсами. В процессе сбора текстов возникают задачи извлечения и очистки данных. На последнем этапе все собранные тексты проходят дополнительную обработку.

Рассмотрим подробнее порядок работы системы. На вход подается поисковый запрос, содержащий набор ключевых слов, характеризующих тематику коллекции. Заметим, что запрос может быть задан непосредственно пользователем или генерироваться другой системой. Далее запрос передается компонентам, отвечающим за сбор информации.

Как было сказано, компонент, работающий с интернет-ресурсами общего вида, для сбора данных использует механизм метапоиска. Сначала он выполняет расширение запроса, затем передает его подключенным к системе информационно-поисковым системам (ИПС).



Архитектура системы

Результаты работы этих систем, содержащие наборы ссылок на релевантные интернет-страницы, объединяются; на основе полученных ссылок выполняется загрузка интернет-страниц, из которых извлекаются тексты. Полученные тексты снабжаются метаданными и передаются следующему компоненту.

Компонент, работающий с вики-ресурсами, осуществляет запрос к вики-системе. Из вики-страниц, находящихся в полученном в ответ списке, извлекаются тексты, выполняется их очистка от вики-разметки, после чего тексты снабжаются метаданными и передаются следующему компоненту.

Последний компонент системы выполняет итоговую обработку собранных текстов, например осуществляет поиск дублей, индексирование, некоторую семантическую обработку. Его подробное описание представляет собой материал для отдельной статьи.

В результате на выходе системы получается готовая текстовая коллекция, в которой тексты хранятся в формате plain text в кодировке UTF-8. Для каждого текста в отдельном файле хранятся его метаданные в формате JSON: исходный поисковый запрос, сниппет, адрес интернет-ресурса, язык, время создания и время последней модификации (какие-то из этих метаданных могут отсутствовать). Результаты обработки коллекции также сохраняются в отдельном файле в формате JSON.

Сбор материала с помощью метапоиска

Основная часть информации в сети Интернет расположена на сайтах, ее объем постоянно растет. Структура и содержание сайтов меняются с течением времени: одни сайты исчезают, появляется огромное количество новых. Среди этого гигантского количества динамически изменяющейся информации очень сложно найти нужную. Сбор текстов, соответствующих определенной тематике, можно осуществить следующим образом: сначала с помощью механизмов поиска найти страницы, релевантные заданной тематике, а затем извлечь из них содержательную текстовую информацию.

Для поиска по всей сети Интернет можно создать собственного поискового агента, который будет обходить страницы Интернета и составлять поисковый индекс, либо использовать результаты поиска существующих информационно-поисковых систем. Создание собственного поискового агента может быть привлекательно тем, что алгоритм его работы был бы открыт и известен, в то время как поисковый алгоритм большинства ИПС является коммерческой тайной и старательно оберегается их создателями. Однако сканирование сети Интернет – достаточно трудоемкий процесс и требует больших временных и аппаратных затрат, к тому же полнота поиска все равно будет хуже, чем в метапоисковой системе, использующей несколько ИПС, покрывающих различные участки Интернета. Помимо этого каждая ИПС имеет свой алгоритм ранжирования результатов поиска, из-за чего поисковая выдача разных ИПС может сильно различаться. Поэтому метапоисковая система, использующая несколько ИПС, может вернуть более точный и полный список результатов на запрос пользователя.

Основная идея метапоиска заключается в формировании поисковой выдачи (результатов поиска) за счет объединения и переупорядочивания поисковых выдач всех задействованных ИПС. Метапоисковые системы в отличие от традиционных ИПС обычно не имеют собственного поискового индекса. Они обрабатывают поисковый запрос пользователя и отправляют его нескольким ИПС, которые либо выбираются из списка ИПС автоматически, либо задаются пользователем вручную. Ответ каждой ИПС обрабатывается, и результаты поиска объединяются в единый список [2].

В работе метапоисковой машины можно выделить три этапа: предобработку запроса, выполнение запроса несколькими ИПС, постобработку результатов. В рамках нашей задачи добавляется еще один этап: загрузка текстов, так как целью системы является именно создание текстовой коллекции.

Предобработка запроса заключается в его расширении с использованием словарей или тезаурусов. В постобработку входит приведение результатов к единому виду, удаление повторяющихся ссылок, фильтрация. Также в метапоисковых системах большое значение имеет объединение и переупорядочивание результатов, полученных от разных ИПС. Однако

для задачи сбора текстов порядок следования результатов не так важен, как для метапоисковых систем, предоставляющих результат пользователю, который хочет получить наиболее релевантные результаты в начале списка.

Расширение поискового запроса. Наличие синонимов и близких понятий в языке делает задачу поиска более сложной. Поиск по ключевым словам позволяет получить только документы, содержащие исходные ключевые слова, а документы, содержащие его синонимы или близкие понятия, при этом не находятся. Для решения этой проблемы и увеличения полноты поиска используется механизм расширения поискового запроса. Его основная идея заключается в дополнении поискового запроса синонимами или квазисинонимами (переформулировки, транслит, гипонимы и даже однокоренные слова) слов, входящих в запрос. Для того чтобы такое расширение было возможным, необходимо знать семантические отношения между словами. Они могут задаваться словарями синонимов и тезаурусами [3].

Типичным ресурсом, применяемым для расширения англоязычных запросов, является WordNet [4–6] – свободно распространяемый тезаурус английского языка, разработанный в Принстонском университете в 1985 г.

Синонимия является основным отношением в словаре WordNet. Слова на основании отношения синонимии объединяются в синонимические ряды (синсеты). Если слово имеет несколько значений, то оно входит в несколько синсетов.

Синсеты связаны между собой небольшим количеством концептуальных отношений: гиперонимией (отношением «общее – частное»), меронимией (отношением «состоит из»), антонимией. Отношения в WordNet связывают только слова, принадлежащие одной части речи. Таким образом, словарь WordNet разделен на четыре части: существительные, глаголы, прилагательные и наречия.

На примере тезауруса WordNet рассмотрим, как происходит расширение запроса. Сначала запрос разбивается на термы, т. е. выделяются максимальные последовательности слов, которые принадлежат какому-либо синсету в WordNet. На основании информации, представленной в WordNet, для каждого терма строятся множества синонимов и гипонимов. После этого каждый терм с помощью логических связок OR объединяется со своими синонимами и гипонимами. Расширенные таким образом термы объединяются с помощью связок AND, в итоге образуя расширенный запрос.

Взаимодействие с поисковыми системами. Для взаимодействия с ИПС используются их поисковые программные интерфейсы (API). Такие ИПС, как Google, Яндекс, Yahoo!, Bing, позволяют отправлять запросы к их поисковым базам и получать ответы в формате XML либо JSON (в бесплатных версиях API существуют ограничения на число запросов). Например, чтобы воспользоваться Яндекс.XML [1] (поисковый API, предоставляемый Яндексом), нужно отправить на специальный адрес запрос (POST либо GET), в котором в качестве параметров указывается поисковый запрос, правила сортировки и фильтрации, число требуемых результатов. В ответ возвращается список ссылок на страницы, которые Яндекс считает релевантными исходному запросу. При этом для каждой страницы выдается заголовок, кодировка, дата и время изменения, тип документа и сниппет – небольшой отрывок текста из найденной страницы, как правило содержащий контекст, в котором встретилось ключевое слово.

После того, как ИПС вернули результаты обработки запроса, последние объединяются, из них удаляются одинаковые ссылки и формируется единый список ссылок на HTML-страницы с некоторой метаинформацией: время получения ссылки, время создания страницы (если известно), заголовок и сниппет. Обычно все ИПС предоставляют эти данные.

Затем происходит загрузка страниц по ссылкам из списка. Основная задача этого этапа состоит в том, чтобы загрузить HTML-страницу по ссылке, используя HTTP-протокол, определить кодировку страницы и преобразовать ее в UTF-8.

Специфика получаемых данных. В число возвращаемых ИПС результатов могут входить не только ссылки на HTML-страницы, но и на документы различных форматов (например, .doc, .pdf). На данном этапе разработки системы рассматриваются только HTML-страницы, а документы остальных типов игнорируются.

[1] Документация Яндекс.XML. URL: http://api.yandex.ru/xml/doc/dg/concepts/about.xml

В настоящее время большая часть информации в сети Интернет представлена в виде HTML-страниц. Основной его недостаток, затрудняющий автоматическое извлечение информации, состоит в том, что в HTML не разделяются данные и их представление. Теги в HTML не имеют семантического значения, один и тот же тег может использоваться как для выделения блока информации, так и для создания навигации по сайту.

В стандарте HTML5 были добавлены новые теги, имеющие семантическое значение [7]. Одним из них является тег <article>, используемый для выделения независимой части документа или сайта (например, статья в блоге). Применение семантических тегов значительно бы облегчило автоматическую обработку сайтов, но далеко не все создатели сайтов применяют новые семантические теги, даже если знают о них.

Интернет – это хаотически развивающаяся система, и нам приходится учитывать все разнообразие составляющих его сайтов при их автоматической обработке и извлечении из них информации.

Извлечение текстов. В типичной HTML-странице можно выделить навигационную часть и содержательную часть. К навигационной части можно отнести меню, «шапку страницы», «низ страницы», элементы оформления. Обычно навигационная часть одинаковая у всех страниц одного сайта. Содержательная часть – основной контент страницы, то, ради чего она была создана.

Много исследований посвящено извлечению содержательной информации из HTML-страницы. Можно выделить два направления разработок:

- анализ структуры отдельной HTML-страницы [8; 9];
- анализ нескольких страниц, выделение повторяющихся элементов [10; 11].

Подход, основанный на анализе множества страниц, базируется на предположении, что у всех страниц с одного сайта элементы оформления одинаковые, но содержательная часть разная. Значит, путем сравнения некоторого количества таких страниц можно выделить повторяющиеся части и удалить их, признав элементами оформления и навигации. То, что останется, и будет содержательной частью. В этом состоит основная идея данного подхода. Например, в работе [11] каждая страница разделяется на блоки, и после сравнения нескольких страниц те блоки, которые оказались уникальными, считаются содержательной частью.

Выделение повторяющихся элементов требует наличия большого количества структурно похожих страниц (с одного сайта), в нашем случае страницы в поисковой выдаче достаточно разнообразные, что делает невозможным применение данного метода в рамках поставленной задачи.

Если нет возможности выполнить разбор нескольких страниц, принадлежащих одному сайту, применяются методы, которые анализируют структуру отдельной страницы и на основе этого анализа делают выводы о ее содержимом. Структуру HTML-документа можно представить в виде DOM-дерева. DOM (Document Object Model – объектная модель документа) является стандартом, регламентирующим способ представления содержимого документа (в частности, HTML-страницы) в виде набора объектов [12]. В подходах, основанных на анализе DOM-дерева одной страницы, исходя из предположений о том, как обычно оформляются HTML-страницы, создаются эвристики, с помощью которых извлекается требуемая информация. Эвристики могут задаваться разработчиком вручную либо автоматически. Например, в системе Lixto информация извлекается из HTML-страниц с использованием шаблонов. Шаблоны задаются при взаимодействии с пользователем: сначала он выделяет на странице интересующий его блок, а система автоматически генерирует шаблон поиска похожих блоков [9].

Для формирования текстовой коллекции нужно уметь находить и извлекать статьи из HTML-страниц. Для статьи характерно большое количество текста и небольшое количество тегов, в отличие от навигационной части, в которой содержится большое количество тегов и небольшое количество текста. Теги, которые могут встречаться в статье:

- <p>, , , <tr>, <td>,
, <hr> (разделение текста на абзацы);
- <a> (ссылки на источники, материалы по теме);
- <img> (иллюстрации);
- <b>, <i>, , , <big>, <font> (форматирование текста).

В ходе анализа DOM-дерева выделяются блоки, которые состоят из допустимых тегов и текста достаточной длины. Статьей считается текстовая информация, содержащаяся в этих блоках. Если страница не содержит таких блоков либо их размер слишком мал, считается, что страница не содержит статьи по интересующей нас теме и удаляется. Это происходит из-за того, что релевантные страницы, возвращаемые ИПС, могут иметь разные типы (необязательно содержат статью): страница обсуждения на форуме, главная страница сайта, страница с описанием товара в интернет-магазине и др.

Релевантным запросу может быть целый сайт, содержащий большое количество статей на интересующую нас тему. В данном случае ИПС может вернуть ссылку на главную страницу. Главная страница сайта обычно содержит небольшое описание сайта, ленту новостей с их краткими аннотациями, ссылки на наиболее популярные и интересные статьи. С помощью вышеизложенного метода мы не обнаружим там статьи и исключим страницу из рассмотрения. В результате нерассмотренными окажутся сразу несколько статей, находящихся на этом сайте. Таким образом, если страница не содержит статьи, это еще не означает, что она не содержит полезной для нас информации. В дальнейшем планируется производить более полный анализ получаемых страниц и на его основе загружать новые.

Использование вики-ресурсов

Вики-ресурс – это веб-сайт, структуру и содержимое которого пользователи могут самостоятельно изменять с помощью инструментов, предоставляемых самим сайтом [2]. Ресурсы такого типа характеризуются значительной структурированностью за счет системы категоризации страниц и наличия большого количества связей между ними, а также присутствием дополнительной метаинформации (вплоть до истории правок).

Вики-ресурсы могут достигать значительных размеров (такие как проекты Фонда Викимедиа), хотя по большей части имеют не очень большой объем и узконаправленную тематику.

Использование вики-ресурсов в современной практике не такая уж редкость. Для примера рассмотрим одну из работ, в которой решается задача, похожая на поставленную в данной статье. Она посвящена составлению подборки текстов по определенной предметной области для построения онтологии [13]. В ней предлагается использовать в качестве источника Википедию. Основная идея заключается в генерации иерархии предметной области на основе связей страниц этой электронной энциклопедии, но при этом, для того чтобы коллекция текстов имела хорошие количественные и качественные характеристики (обеспечивала хорошее покрытие предметной области), предлагается алгоритм оценки и фильтрации, основанный на доступной информации о классификации страниц в виде категорий. Проведенные в рассматриваемой работе эксперименты и оценка их результатов показали, что Википедия является хорошим источником текстов для решения поставленной задачи.

Взаимодействие с вики-ресурсами. В основе любого вики-ресурса лежит вики-движок – комплекс программных средств для преобразования вики-разметки в код, предназначенный для отображения в браузере [14]. Этот же движок обеспечивает индексацию вики-страниц, за счет чего предоставляет возможность внутреннего полнотекстового поиска по содержимому вики-ресурса. Также он предоставляет API, позволяющий программным средствам взаимодействовать с вики-ресурсом. Объединяя все вместе, получаем возможность автоматически получать содержимое определенных статей из вики-ресурса по заданному запросу.

На данный момент в нашей системе поддерживаются вики-системы, работающие на движке MediaWiki [3]. Это один из самых распространенных движков, в частности, на нем работает всем известная Википедия. У него есть собственный язык запросов [4]. Для слов из запроса поддерживается морфология, релевантность выше у тех страниц, у которых искомые слова встречаются в названии страницы. Есть специальный синтаксис для поиска точного совпадения, для простых логических операций и простых шаблонов.

[2] URL: http://ru.wikipedia.org/wiki/Вики
[3] Сайт проекта MediaWiki. URL: http://mediawiki.org
[4] Википедия:Поиск. URL: http://ru.wikipedia.org/wiki/Википедия:Поиск

Специфика получаемых данных. Тексты, получаемые из вики-ресурса, обладают некоторыми особенностями. Прежде всего, в них присутствует вики-разметка – специальные управляющие последовательности символов для оформления текста на ресурсах данного типа. Для формирования текстовой коллекции нужно удалять такую разметку из текстов. С одной стороны, простота синтаксиса позволяет довольно просто выделять ее в тексте, с другой – наличие различных вариантов разметки (в частности такого механизма, как вики-шаблоны, позволяющего включать одни страницы в другие) не дает возможности полностью удалить всю разметку с сохранением всех смысловых частей. Таким образом, необходим баланс между степенью очистки текста от разметки и сохранением его смысла.

Тот же способ форматирования статей, а также стиль написания, присущий вики, задает определенную структуру текста: все абзацы разделены пустой строкой, а для каждого раздела задан заголовок. В некоторой степени это облегчает обработку текстов.

Еще можно отметить ограничение на максимальный размер статей. Для MediaWiki этот параметр явно задается одной из настроек [5]. Следовательно, диапазон изменения размера получаемых текстов будет находиться в ограниченных пределах.

Очистка получаемых данных. Как уже было сказано, очистка текстов в основном заключается в удалении вики-разметки. Для MediaWiki существует довольно большое число разнообразных альтернативных парсеров [6], но эксперименты показали, что ни один из них не позволяет получить хорошее качество очистки. На данный момент в нашей системе используется тот, который показал наиболее приемлемый результат. При этом основным препятствующим фактором является не столько сложность оригинального парсера, сколько ориентация на специфику задачи. Во многих случаях при преобразовании разметки возникает несколько вариантов (смещать баланс в сторону удаления большего количества разметки или в сторону сохранения большего количества текста) и выбор определяется требованиями к результату. Оптимальным решением является разработка собственного средства очистки и его настройка под собственные задачи.

Применение системы для формирования коллекций

Рассматриваемая в данной работе система формирования тематических текстовых коллекций на основе интернет-ресурсов была использована для построения ряда коллекций. Подведем предварительные итоги ее применения.

Первая трудность, которая возникает при создании коллекции в режиме on-line, – это большие временные затраты на ее загрузку. В этом режиме следует ограничиться несколькими десятками, максимум, одной-двумя сотнями текстов. Объемы коллекций, создаваемых для обработки off-line, обычно не имеют ограничений такого сорта. Но ограничения могут возникнуть естественным образом при оценке качества (релевантности) полученного материала.

Следующей проблемой является то, что коллекции, собранные с помощью механизма метапоиска, обладают рядом недостатков, затрудняющих их дальнейшую обработку. Проявляются эти недостатки как на уровне всей совокупности текстов, так и на уровне отдельных текстов. Рассмотрим их на примере коллекции, сформированной по запросу «Ландшафтный дизайн» (838 статей).

Когда составители пытаются отразить в собранной коллекции тему как можно полнее, для запроса обычно используют самые широкие термины интересующей предметной области, допускающие множество разных интерпретаций. Одним из следствий этого являются проблемы, связанные с наличием в коллекции нерелевантных текстов.

В исследуемой коллекции текстов, ранжированных поисковой системой, нерелевантные тексты встречаются как в начале подборки, так и в конце. Экспертная оценка такова: среди текстов с высокой релевантностью содержится примерно 20 % текстов, которые практически не соответствуют запросу. Доля нерелевантных текстов в конце подборки возрастает

[5] Максимальный размер статьи в MediaWiki. URL: http://www.mediawiki.org/wiki/Manual:$wgMaxArticleSize
[6] Alternative MediaWiki parsers. URL: http://www.mediawiki.org/wiki/Alternative_parsers

до 30 %. Этому способствует, в частности, тот факт, что термины запроса встречаются в подзаголовках, предваряющих параграф всего в одну фразу, и учитываются при оценке релевантности. Все остальное содержание статьи бывает посвящено совершенно другой теме, в нашем случае это: 1) автобиографии людей, в интересы которых входил ландшафтный дизайн; 2) реклама компаний, список видов деятельности которых включает ландшафтный дизайн; 3) перечень профессий, которые можно получить в вузе, содержащий словосочетание *ландшафтный дизайн*; 4) статьи Википедии, в списке категорий которых значился ландшафтный дизайн, и т. п.

Среди нерелевантных текстов обнаружены и такие, в которых не было ни слова о ландшафтном дизайне (от 5 до 10 % от доли нерелевантных текстов). В них шла речь о дизайне автомобилей, велосипедов, мобильных телефонов, интерьеров и т. п.

Другой проблемой, вытекающей из использования широкоупотребительных терминов в поисковых запросах, является появление в коллекции дублей (как точных, так и частичных). Для оценки меры сходства документов использовалась одна из модификаций метода шинглов [15]. Метод сравнения текстов выбирался из соображений производительности с учетом экспертной оценки точности и полноты (подробное обоснование выбора выходит за рамки данной статьи). Если значение меры сходства пары документов достаточно велико, более информативный документ (как правило, это больший документ в паре) считается уникальным в коллекции, менее информативный объявляется дубликатом. Если объемы документов отличаются незначительно или совпадают, то уникальным считается тот, который встретился раньше.

В полученной коллекции было обнаружено 49 документов, являющихся точными копиями других, встреченных ранее. Для частичных дубликатов характерны точные повторы одного или нескольких абзацев (иногда они составляют основную часть статьи), а также отдельных фраз в начале и конце текста. Встречаются случаи варьирования внутри всего текста, в том числе «искусственные». Иногда имеет место вложенность текстов: короткий текст погружен в существенно более длинный, объемы текстов при этом могут различаться на порядок. Всего в подборке обнаружено от 63 до 68 дубликатов (в зависимости от выбираемого порогового значения сходства и мнения экспертов). Из них от 14 до 19 являются частичными. В результате оказалось, что из всего количества полученных текстов уникальными оказались чуть менее 92 %.

Один из этапов предобработки отдельного текста включает разбиение его на фразы и слова, что обеспечивает правильное выявление таких значимых структур, как устойчивые повторы, содержащие словосочетания и термины, в которых отражается семантика текстов. Поэтому важно, чтобы в тексте были стандартные разделители между словами и фразами. Для многих интернет-текстов характерно использование символов «_», «|» и др. в качестве таких разделителей, например: оборудование_Apple_Inc, thumb|left|Ёлка с гирляндой. Также «шумом», который должен быть удален, являются и символы разметки заголовков и списков, например: «= =», «*», «#» и др. В этом случае следует учесть, что размеченные структурные элементы представляют собой отдельные фразы.

Следует отметить, что в Википедии принята определенная структура статьи, в которой предусмотрены общие подразделы. Для кластеризации коротких текстов особенно важно, чтобы повторы на уровне фраз (заголовков стандартных разделов) не были причиной «сближения» текстов. Например, такие подзаголовки, как «См. также», «Примечание», «Файл», «Ссылки», «Литература», «Категория» и проч., необходимо предварительно удалить из текстов.

Таким образом, первые эксперименты по испытанию системы показали имеющиеся проблемы и недоработки. Среди них были как ожидаемые (например, проблемы, возникшие из-за недостаточной очистки текстов, полученных из вики-ресурсов, от вики-разметки), так и непредвиденные, в частности вытекающие из особенностей представления текстов в Интернете и методов оценки релевантности ресурсов, используемых различными ИПС.

Заключение

В данной работе рассмотрена задача автоматического формирования тематических текстовых коллекций на основе интернет-ресурсов. Для ее решения предложен подход, на основе которого разработана система, использующая для сбора текстов механизмы метапоиска

и специализированные средства для работы с вики-ресурсами. Проведенные эксперименты с системой показали не только продуктивность предложенного подхода к автоматическому формированию текстовых коллекций, но и выявили ряд недостатков, препятствующих получению коллекций высокого качества. К таким недостаткам, в частности, относятся: низкая релевантность отдельных текстов коллекции поисковому запросу, наличие в текстах «мусора» и др.

Наиболее сложным является устранение первого из отмеченных недостатков, так как оно требует разработки надежного метода оценки релевантности запросу извлеченных и очищенных от «мусора» текстов. Наиболее перспективными и эффективными в настоящее время являются комплексные методы оценки релевантности, основанные как на анализе структуры веб-страниц [16], так и их текстового содержимого (*text content based*), также классифицируемые как методы анализа информации на странице (*on-page*) [17].

Дальнейшее развитие рассматриваемого подхода в части работы с интернет-ресурсами общего вида состоит в реализации механизма распознавания типов найденных ресурсов (форум, блог, новостной сайт и т. д.). Это позволит применять для разных типов ресурсов свои специфические методы извлечения текстов, что должно положительным образом сказаться на качестве результата. Также информация о типе ресурса будет добавляться в метаданные, что может оказаться полезным при дальнейшем использовании коллекции. В плане общего улучшения работы блока сбора информации планируется выполнять анализ ссылок, расположенных на найденных страницах, с целью отбора и использования подходящих по теме. Это даст возможность увеличить объем и качество извлекаемой информации.

Список литературы

1. *Степанов В. К.* Применение Интернета в профессиональной информационной деятельности. М.: ФАИР, 2009. 301 с.

2. *Meng W., Yu C., Liu K. L.* Building Efficient and Effective Metasearch Engines // ACM Computing Surveys (CSUR). 2002. Vol. 34. No. 1. P. 48–89.

3. *Арбатская О. А.* Интеллектуализация тематического поиска в поисковых системах Интернет // Лингвистическое обеспечение информационных ресурсов библиотек, музеев, архивов и других учреждений культуры. СПб.: Сударыня, 2008. С. 173–190.

4. *Voorhees E. M.* Query Expansion Using Lexical-Semantic Relations // SIGIR'94. L.: Springer, 1994. P. 61–69.

5. *Zhang J., Deng B., Li X.* Concept Based Query Expansion Using WordNet // Proc. of the 2009 International e-Conference on Advanced Science and Technology / IEEE Computer Society. 2009. P. 52–55.

6. *Nemrava J.* Using WordNet Glosses to Refine Google Queries // Proc. of the Dateso 2006 Workshop. VSB – Technical University of Ostrava, Dept. of Computer Science, 2006. P. 85–94.

7. *Berjon R., Faulkner S., Leithead T., Navara E. D., O'Connor E., Pfeiffer S., Hickson I.* HTML5: A Vocabulary and Associated APIs for HTML and XHTML // W3C Candidate Recommendation. 2013.

8. *Кузнецов Р. Ф.* Извлечение значимой информации из web-страниц с использованием предложений // RCDL'2006: Сб. тез. постерных докл. VIII Всерос. конф. СПб.: НУ ЦСИ, 2006. 274 с.

9. *Baumgartner R.* Datalog-Related Aspects in Lixto Visual Developer // Datalog Reloaded. Lecture Notes in Computer Science. 2011. Vol. 6702. P. 145–160.

10. *Агеев М. С., Вершинников И. В., Добров Б. В.* Извлечение значимой информации из web-страниц для задач информационного поиска // Интернет-математика 2005. Автоматическая обработка веб-данных. М., 2005. С. 283–301.

11. *Marathe M., Patil S. H., Garje G. V., Bewoor M. S.* Extracting Content Blocks from Web Pages // International Journal of Recent Trends in Engineering, 2009. Vol. 2. No. 4. P. 62–64.

12. *Stenback J., Le Hégaret P., Le Hors A.* Document Object Model (DOM) Level 2 HTML Specification // W3C Recommendation. 2003.

13. *Cui G. Y., Lu Q., Li W. J., Chen Y. R.* Corpus Exploitation from Wikipedia for Ontology Construction // Proc. of the VI International Language Resources and Evaluation (LREC 2008). Marrakech, 2008. P. 2125–2132.

14. *Leuf B., Cunningham W.* The Wiki Way: Quick Collaboration on the Web. Addison-Wesley, 2001. 435 p.

15. *Broder A., Glassman S., Manasse M., Zweig G.* Syntactic Clustering of the Web // Computer Networks and ISDN Systems. 1997. Vol. 29. No. 8. P. 1157–1166.

16. *Lindemann C., Littig L.* Coarse-Grained Classification of Web Sites by Their Structural Properties // Proc. of the VIII Annual ACM International Workshop on Web Information and Data Management. 2006. P. 35–42.

17. *Qi X., Davison B. D.* Web Page Classification: Features and Algorithms // ACM Computing Surveys (CSUR). 2009. Vol. 41. No. 2. P. 1–31.

Материал поступил в редколлегию 21.10.2013

I. R. Akhmadeeva, Yu. A. Zagorulko, N. V. Salomatina, A. S. Sery, E. A. Sidorova, V. K. Shestakov

APPROACH TO FORMING THEMATIC TEXT COLLECTIONS ON THE BASIS OF WEB-RESOURCES

Problem of automatically forming text collections related to given themes on the basis of web-resources are considered. Approach to solution of this problem is suggested and system using metasearch technique and specialized facilities for operation with wiki-resources for collecting texts is developed. Experiments made with the system have proved productivity of the suggested approach.

Keywords: text collections, web-resources, wiki-resources, web search query, metasearch.