Электронный архив газет: Web-публикация, ассоциация информации с базой данных, создание полнотекстового поиска

© А. Г. Марчук © С. В. Лештаев
Институт систем информатики СО РАН,
Новосибирск
mag@iis.nsk.su svles@iis.nsk.su

Аннотация

В докладе описана разработанная система представления (публикации) информации архива сканированных газет на сайте. В практическом плане, в электронный архив газет входит ряд нетривиальных программных частей: отображение изображений сканов высокого разрешения на веб-сайте; распознавание текста страниц, создание полнотекстового индекса для поиска по ключевым словам, организация связи опубликованных выпусков газет и базы данных фотоархива СО РАН.

Благодарности

Авторы выражают благодарность принимающим участие в работе над проектом «Открытый архив СО РАН» сотрудникам Института систем информатики СО РАН: А.А.Фурсенко, И.Ю.Павловской, И.А.Крайневой, В.Э.Филиппову, П.А.Марчуку. Работа выполнена при поддержке гранта РФФИ 14-07-00386А.

1 Введение

Исследуется задача представления (публикации) набора данных. Исходным набором данных для проекта является множество из 24532 сканированных страниц и разворотов газеты «За науку в Сибири» (современное название «Наука в Сибири»). Необходимое требование решения - максимально удобный доступ к данным через браузер.

Аналогичную, хотя и более масштабную задачу решала команда Google в проекте Google newspapers https://news.google.com/newspapers. В другом проекте, на сайте https://Issuu.com отображаются журналы, оцифрованные и опубликованные в формате PDF.

Первично, в таких системах, как и в нашей, решается задача доступа через Интернет и Web-браузер к сканам высокого разрешения. Мы остановили свой выбор на технологии Deep Zoom [10] фирмы Microsoft. Привлекательным свойством технологии является возможность лёгкой подстройки визуального качества изображения пользователем в процессе чтения, вплоть до предельных характеристик его детальности.

Другими задачами, решавшимися в проекте, являлись: обеспечение связи сканированных изображений с базой данных и реализация текстового поиска по набору задаваемых слов.

Связь между сканированными образами и базой данных осуществляется в обе стороны, т.е. из базы данных мы имеем ссылки на конкретные места в изображениях и, наоборот, некоторые образы статей размечены дополнительной информацией и ссылками, ведущими в базу данных. Созданный в рамках проекта технологический комплекс позволяет в удобном виде выделять информационные фрагменты из изображений страниц и связывать фрагменты с элементами базы данных. Авторы не нашли подобных решений в других разработках, связанными с публикацией сканов газет и журналов.

Для того, чтобы сделать поиск по ключевым словам в текстах выпусков, необходимо распознать текст и сделать полнотекстовый индекс. В какой-то мере, Google в своей системе решает эту задачу. А Issuu.com не распознаёт текст с изображений, если предоставленный пользователем PDF состоит из них, но если PDF содержит текст, то поиск работает.

Частично результаты исследования этой задачи были опубликованы в статье [7] и магистерской диссертации [6]. В этой статье приводится краткое описание новых решённых частей продолжающегося проекта.

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

Рисунок 1 Изображения страниц одного выпуска состыкованы по порядку в горизонтальный ряд и выровнены по высоте.

2 Отображение выпуска газеты

В разных системах Web-публикации сканированных копий газет и журналов, используется небольшой набор вариантов отображения страниц на графику браузера. Обычно используются традиционные имиджи и HTML. Некоторые используют публикацию через PDF формат файла (сайт Issue.com). Проблемой таких подходов является необходимость соблюдать компромисс между качеством изображений и объёмом JPEG или PDF-файлов. Для более качественной публикации «картинок» страниц, используют специализированные технологии и системы, в частности, применяют технологию Deep Zoom, для браузера доступную в рамках Silverlight. Google в этой задаче, использует свою технологию отображения карт.

В нашем решении используется Deep Zoom, для некоторых целей планируется добавить варианты PDF с текстом и технологию Open Seadragon. Последнее связано с тем, что Silverlight не набрал популярность настолько, чтобы использоваться повсеместно, в некоторых браузерах требуется разблокировать в настройках запуск Silverlight-компонента на сайтах, так как по умолчанию он блокируется для безопасности. Open Seadragon реализует ту же технологию средствами HTML-5, хотя и с рядом ограничений.

Deep Zoom и Open Seadragon в качестве источника используют «пирамидальную» группу сжатых копий изображения (DZI), каждая копия разделена на части по 256x256 пикселей в формате JPG. Это решение создаёт много служебных файлов: в 24532 сканах набора газет за 37 лет всего более четырёх миллионов JPG файлов.

Это, в свою очередь, породило проблему, которая выражается напр. в том, что копирование полного набора файлов кассеты может длиться часами. Для хранения такого объёма маленьких файлов и, что важно – оперативной выдачи, был разработан и применён формат быстрого архива без сжатия Sarc (Simple Archive). Он состоит из трёх блоков: 1) тело архива содержит без изменений

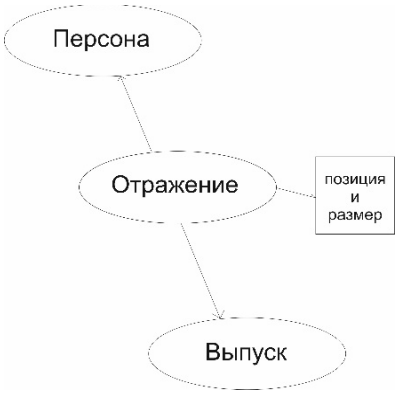

Рисунок 2 Орграф показывающий связывание в RDF данных представителей классов (название класса указано в овале) выпуска, персоны и отражения. Название текстового свойства позиции и размера области отражения указано в квадрате. Допускается множество отражений в одном выпуске, но у отражения может быть только один персонаж.

байты файлов в архиве подряд по порядку добавления в архив; 2) содержание архива в XML формате, описывающее пары: имя файла и позиция начала его байт в теле архива; 3) размер содержания в байтах, этот блок всегда фиксированного размера 8 байт (длинное целое). Объектное представление архива состоит из содержания в виде словаря пар.

Отображение выпуска осуществляется в виде состыкованных изображений страниц подходящего размера. Deep Zoom предоставляет также функциональность отображения коллекции изображений, принимая в качестве источника файлы Deep Zoom Collection, состоящие из XML описания каждого изображения коллекции: относительные размеры, позиции и путь к DZI.

Для публикации всего множества выпусков необходимо описать их в базе данных. Авторы использовали RDF СУБД, созданную на основе технологии Polar [8]. Каждая страница выпуска газеты является отдельным документом и описывается в данных как сущность класса документа. Весь выпуск описывается сущностью класса многостраничный документ и класса коллекции.

3 Взаимосвязь публикуемых сканов газет с базой данных

Привязка сканированного материала к базе данных осуществляется через отождествление области, на которой имеется изображение статьи с

Рисунок 3 Орграф примера RDF данных, показывающий связи между представителями классов выпуска газеты; отражения статьи; персонажей. Допускается множество отражений персонажей в одной статье.

документом. Далее, следует использование ассоциированных с документом отношений, таких как отражение, авторство, место.

Например, если в статье (на изображении газетного разворота) отражён или описан человек (организация или событие), информация о котором есть или может быть внесена в базу данных, информационным специалистом делается такая привязка. Такая технология опробована и используется в проекте «Фотоархив СО РАН», редактору предоставлена функциональность отметить область отражения или описания и указать персонажа. В RDF данных это записывается узлом класса отражение. Для каждого отражения указаны дуга к узлу персонажа, дуга к многостраничному документу, позиция и размер области отражения. На сайте фотоархива СО РАН на странице описания человека (организации или события) автоматически размещается ссылка на отражение, которая приводит на страницу выпуска и показывает область отражения.

Статья в выпуске рассматривается как самостоятельный информационный объект, она имеет название и прочие атрибуты описания. В RDF она записывается сущностью класса article, а множество персонажей, описанных в ней записано сущностью класса отражения. Сама статья отражена в выпуске так же, сущностью класса отражение с указанием позиции и размера.

4 Распознавание текстов

Для распознавания текста, имеющихся на сканированных изображениях авторы выбрали OCR движок tesseract, он бесплатный и формирует результирующую информацию в виде удобного формата HOCR. Это HTML со специальной идентификацией и классификацией элементов, в формате фиксируется позиция и размер абзацев, строк и каждого слова. Для этого формата можно создать конвертер в PDF. Из платных приложений ABBY Finereader распознаёт более точно, чем tesseract. Чтобы с помощью него определить позицию слова, можно распознавать изображение в формат PDF с параметром «точная копия».

4.1 Кодирование слов русского языка

Для создания базы данных всех распознанных слов, создаётся база широкого набора, в идеале – всех слов русского языка и применяется кодирование встреченных при обработке текстов слов и имён собственных. Используется список из 4 159 394 словоформ для 142 792 лемм [12] опубликованный на сайте http://www.speakrus.ru/dict/. Для кодирования с помощью технологии Polar авторами реализована таблица имён, осуществляющая биективное отображение строк в числа и обратно [4]. При этом каждое слово получает числовой код, и в каждом падеже или времени это слово имеет такой же уникальный код, т.е. слова переводятся в нормальную форму. Неинформативные слова: предлоги, союзы и т.п. называются «стоп-словами», получают специальный код.

4.2 Преобразование HOCR в поток слов

В HOCR текст уже разделён на слова, но запятые, точки и другие знаки препинания оставляются вместе со словом. Разделённые на две части слова для переноса на новую строку с помощью тире остаются двумя словами, при этом тире остаётся в конце первой части. И из-за неточности распознавания слова могут быть разделены на части. Поэтому, если слово не найдено в базе всех слов, выполняется попытка найти его конкатенацию со следующим или предыдущим словом. Кавычки могут играть роль выделения наименования, оно может состоять из нескольких слов. Для решения перечисленных проблем подходит применение синтаксического и семантического анализатора, созданного по специальной грамматике. Это исследование продолжается. Сокращения слов можно расшифровать с помощью онлайн сервиса http://www.sokr.ru/. И необходим список имён,

Рисунок 4 Изображение первой страницы первого выпуска газеты “За науку в Сибири”. На изображении выделена цветом (при печати изображение преобразовано в чёрно белое) фотография президента АН СССР Келдыш Мстислава. Под выделением размещена ссылка на страницу сайта фотоархива с его информационным портретом.

фамилий, названий. Результаты исследования и разработок применимы не только для распознанных газетных сканов, но и к другим вариантам работы с текстами.

4.3 Полнотекстовый индекс с координатами каждого слова

Всего на сканированной странице ~1500 слов, следовательно, всего на 24532 страницах менее 100 миллионов слов, что позволяет рассматривать решения размещения базы данных в оперативной памяти. После распознавания с помощью технологии Polar создаётся таблица слов, строки в ней соответствуют словам (всем, кроме стоп-слов) в страницах газет и содержат код слова (4 байта), целочисленный идентификатор страницы (4 байта) из базы данных страниц и позицию слова (4 байта x, 4 байта y). Всего не более 100 миллионов строк, не более 1,6 гигабайт на таблицу, она помещается в оперативную память. Это является достаточным условием быстрой работы с ней.

5 Пользовательский поиск

Пользователю предоставляется функция поиска [5], он указывает несколько ключевых слов, или наименований, возможно с опечатками или в каком-то падеже или времени. Для определения искомых ключевых слов используется нечёткий поиск [3], [11]. Пока есть ограничения на работу с именами людей: фамилия, имя, отчество должны совпадать полностью (кроме падежа) и в образце, и в тексте, в качестве частного послабления, в тексте возможно указано только фамилии и имени или фамилии с инициалом имени. Названия организаций часто состоят из нескольких слов или аббревиатуры, при поиске требуется совпадение этих слов в такой же последовательности.

В качестве результата поиска пользователю предоставляется множество выпусков в виде списка ссылок, отсортированных по убыванию релевантности найденных в них страниц. В одном выпуске может быть найдено несколько страниц, для каждой в списке содержится отдельная ссылка. При переходе по ссылке отображается соответствующая ей страница, и во всём выпуске подсвечиваются точками позиции всех искомых слов.

Оказалось, что предоставить пользователю найденные сканированные страницы недостаточно – ему трудно увидеть эти слова в образе страницы и появляется ощущение, что эта страница предоставлена ошибочно. В настоящее время делается попытка решить эту проблему в двух направлениях. Если пользователю по ссылке показать PDF вариант страницы, то он может самостоятельно найти на ней искомое слово с помощью поиска браузера. Другой вариант – найти искомые слова внутри демонстрируемого образа показать пользователю копию, в которой найденные слова окрашены.

5.1 Нечёткий поиск

Для того, чтобы найти множество похожих слов на данное слово не используя перебор всех слов применяется нечёткий поиск. Один из вариантов нечёткого поиска выполняется с помощью триграмм

[1], то есть троек символов, был реализован авторами. Предварительно, каждое слово в базе слов преобразуется в неупорядоченное множество триграмм, т.е. буквенных троек, «вырезанных» из слова с помощью скользящего окна, например: слово → { __с, _сл, сло, лов, ово, во_, о__ }. Каждому трёхбуквенному коду сопоставляется множество слов, в которых она содержится. При поиске искомое слово тоже преобразуется в множество триграмм.

Нечёткость поиска означает, что данное для поиска слово может отличаться от искомого заменой одного символа, пропущенным, лишним символом, перестановкой символов или соседние символы переставлены местами.

Сравним множество триграмм двух похожих слов: 1) если в одном слове символ заменён другим, то множества отличаются ровно тремя триграммами; 2) если в одном из слов вставлен лишний символ, то его множество содержит триграмм на одну больше и множества отличаются двумя триграммами; 3) если в одном из слов перестановка соседних символов, то множества отличаются на 4 триграммы.

Следовательно, чтобы найти похожие слова с точностью до одной из перечисленных ошибок, для каждой триграммы определяется список содержащих её слов. Можно перебрать все варианты выборок 4х из списков (без повторов). Для каждого варианта вычислить пересечение остальных списков (без выбранных четырёх). Объединение этих пересечений является результатом. При объединении подсчитывается количество копий каждого слова, чем больше, тем больше триграмм совпало, тем больше точность совпадения.

Если введённое слово не найдено, то нечёткий поиск найдёт список возможных вариантов искомого слова, отсортированный по убыванию точности совпадения. Для Яндекса, Google и т.п., возможно, лучше выбрать самое популярное из вариантов, используя статистику поисковых запросов и частоту появления слов в текстах [12]. В реализованном авторами поиске используются все варианты. К каждому найденному искомому слову добавляется множество его синонимов. Список синонимов русского языка можно найти в интернете [2]. Синонимы далее считаются одинаковыми словами.

5.2 Расчёт релевантности страниц

Чем больше в тексте страницы найдено различных искомых слов, тем больше её релевантность. Когда на нескольких страницах одинаковое число различных искомых слов, то релевантность можно различить по длине наименьшему из расстояний между различными словами.

6 Генерация отражений информации фотоархива

Связывание массива обработанных страниц газеты «За науку в Сибири» производилось с базой данных, вручную сформированной при обработке фотографий и списков из истории Сибирского отделения РАН, так называемым фотоархивом СО РАН [9].

Полнотекстовый поиск позволяет выполнить автоматическое выявление страниц, на которых есть упоминание имён персон, организаций, событий и др. (далее объекта) из базы данных фотоархива и зафиксировать это соответствие. Как результат, на изображениях страниц некоторым точкам или областям устанавливается гиперссылка на описание, имеющееся на сайте фотоархива.

Для того, чтобы информационный оператор проверил правильность ссылок, идентификаторы всех страниц, на которых найдено хотя бы одно наименование сохраняются в специальный список. После проверки и возможной коррекции, список уничтожается.

7 Исправление ошибок распознавания

Полиграфия газет обычно не слишком высокого качества. Поэтому большое количество символов распознаются неправильно. Планируется создать приложение для их исправления. Для этого используется база всех слов. Множество всех слов из всех распознанных страниц, которых нет в базе слов объединяются в один список. Далее для каждого ошибочного слова из списка автоматически подбирается список правильных вариантов из базы слов, это те, которые отличаются одним символом (или вставленным/удалённым символом). Вероятность того, что при распознавании символы поменяются местами мала. Более вероятны различия в 6 триграммах, когда два символа в слове неправильные. Для приложения планируется создать пользовательский интерфейс, позволяющий просматривать контекст и в один клик заменять слово на его верный вариант, или записать слово в наименования.

Заключение

Основными тезисами доклада являются: публикация сканов выпуска с помощью технологии Deep Zoom и формата архива Sarc; взаимосвязь отражённых на страницах выпуска объектов и объектов базы данных; распознавание текста с изображений и организация пользовательского поиска.

Описанные результаты в основном доведены до программного решения, применённого при обработке газеты «Наука в Сибири». В силу тематической близости, удобным оказалось

устанавливать связь с базой данных фотоархива СО РАН, результат размещён сайте фотоархива http://soran1957.ru. Опубликовано ~1700 выпусков газеты «За науку в Сибири» с 1965 по 1997. Редакторы разметили области, отражающие значимые персонажи и организации. Ранее проведено распознавание текста из сканов этого архива с помощью OCR движка Cunei, но позиции слов не вычислены. По распознаванию с помощью движка tesseract проведены успешные эксперименты на единичных страницах, требуется применение для массового распознавания всех сканов архива. Проведены эксперименты создания полнотекстового индексирования (без позиций), создана база слов, и проверен нечёткий поиск. Завершены успешные эксперименты отображения страниц (отдельно от выпуска) с помощью Deep Zoom, Open Seadragon, создания текстовых PDF (которые могут индексировать поисковики). Остальные исследования и разработки продолжаются.

Литература

[1] Ukkoncn, E. Approximate string-matching with q-grams and maximal matches. Theoretical Computer Science 92 (1992) 191-211. [Электронный ресурс] – Режим доступа: https://www.cs.helsinki.fi/u/ukkonen/TCS92.pdf (дата обращения: 29.05.2016).

[2] Абрамов. Н. Словарь синонимов. Полная парадигма. Морфология. Перевод в текст Александр Ильин, 2003. http://www.speakrus.ru/dict/

[3] Васильева О. В. Методы поиска и представления информации о расписании вуза. диплом. работа. Новосибирск. НГУ, 2015.

[4] Карасюк П.К. Технологии создания и использования больших таблиц имён. диплом. работа. Новосибирск. НГУ, 2015.

[5] Кристофер Д. Маннинг, Введение в информационный поиск / Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце, перевод Д. Клюшин, 2014.

[6] Лештаев С.В. Архитектура и программное обеспечение архивных фактографических систем: работа с многостраничными растровыми изображениями. дис. … магистра. Новосибирск. НГУ, 2012.

[7] Лештаев С.В., Марчук А.Г. Система создания электронных архивов газет с поиском по ключевым словам, // Системная информатика. — 2014. — № 3. — С. 1-11.

[8] Марчук А.Г. "PolarDB – система создания специализированных NoSQL баз данных и СУБД" // Моделирование и анализ информационных систем. Т. 21, № 6 (2014), с.169–175.

[9] Научная статья про фотоархив [Электронный ресурс] – Режим доступа: http://soran1957.ru/ (дата обращения: 29.05.2016).

[10] Официальная страница Deep Zoom [Электронный ресурс] – Режим доступа: https://www.microsoft.com/SilverLight/deep-zoom/ (дата обращения: 29.05.2016).

[11] Сметанин Н. Нечёткий поиск в тексте и словаре. 9 марта 2011 [Электронный ресурс] – Режим доступа: https://habrahabr.ru/post/114997/ (дата обращения: 25.05.2016).

[12] Хаген М. Частотный словарь. Полная парадигма. Морфология. Совмещённый словарь. 2014. http://www.speakrus.ru/dict/

Digital newspaper archive: Web-publication, linking with database and creating full-text search

Alexander G. Marchuk, Sergey V. Leshtaev

The paper describes the system for creating, maintaining, and Web publishing of scanned set of newspaper pages. The properties of this system include: the use of Deep Zoom technology for visualization of images of pages, text recognition and building the search index, providing of links between scanned pages and database objects.

Technology described was used to publish newspaper “Za nauku v Sibiri” and integration of newspaper articles with the database of the archive soran1957.ru