Электронные ресурсы и электронные библиотеки: Ежегодный межведомственный сборник научных трудов. –М.: ГПНТБ России, 2006, с. 37 – 42.

Автоматизация процесса создания цифровых архивов научно-технической документации.

Калафати Юрий Дмитриевич
Моисеев Константин Владимирович
info@controlchaostech.com
Технологии управляемого хаоса, Москва

Необходимость перевода в электронную форму огромного количества бумажных документов, и последующей систематизации электронных документов в архивы делает актуальным автоматизацию этого процесса или хотя бы некоторых его этапов. Потоковые сканеры, интеллектуальные программы для оптического распознавания текста в принципе позволяют решить эту задачу для документов, которые содержат только напечатанную текстовую информацию. Действительно, после сканирования, т.е. перевода документа в цифровую форму, и оптического распознавания текста, документ можно сохранить в одном из текстовых или векторных форматов. Программы семантического анализа текста в некоторых случаях могут выделять из текста, например, заглавие, авторов и другую необходимую для описания документа информацию и затем передавать эту информацию для пополнения базы данных. В результате всех этих процессов может быть создан архив документов с возможностью поиска информации как по электронным каталогам, так и по полным текстам.

Ситуация однако усложняется для научно-технических документов, где содержится большое количество формул, таблиц, иллюстраций и графиков. Во-первых, оптическое распознавание для математических и химических формул не работает. Если сохранять формулы в виде картинок и затем вставлять в документ с текстом, то появляется большое количество ручного труда. Во-вторых, распознанный текст все равно будет содержать символы от неправильно распознанных формул. Как поступать с этим текстом не понятно. Если текст не править, то использовать его можно только в служебных целях (поскольку текст с плохо распознанными формулами выглядит весьма непрезентабельно), например, для полнотекстового поиска. Если текст частично править, то есть выбрасывать все плохо распознанные символы, то все равно текст оказывается неполным и организация архива

возможна только при сохранении оцифрованного (графического) оригинала страницы. Таким образом, частично автоматизированный процесс создания архива научно-технической документации возможен, если за единицу хранения в архиве принять пару документов - оригинальную страницу в графическом формате и соответствующий этой странице распознанный текст.

По этому пути мы пошли при создании архива Соросовского образовательного журнала за 1995 – 2001 гг. Этот журнал посвящен современным аспектам естествознания. В каждом номере журнала публиковались статьи по физике, математике, химии, биологии и геологии. Исходной информацией для нашего проекта послужил полный архив статей, опубликованных в бумажной версии журнала. Объем этого архива составлял чуть менее 10000 страниц. Собранные воедино эти материалы можно рассматривать как энциклопедию современного естествознания. В таком случае одной из основных задач при создании электронного архива становится задача быстрого и интеллектуального поиска в этом архиве необходимой информации с удобным отображением найденных результатов.

На первом этапе был создан электронный авторский и тематический каталог для всех статей журнала. Далее для каждой страницы журнала мы сохраняли два представления – графический образ страницы и соответствующий странице распознанный текст. Графический образ нам нужен был для точного представления материалов статьи – это и иллюстрации, и таблицы, и, конечно же, многочисленные формулы и специальные символы. (Повторимся, что альтернативный вариант для корректного представления только один – полностью переверстать вручную всю имеющуюся информацию. Этот путь является нереальным в силу того, что на это потребовалось бы огромное количество времени, сил и, конечно же, денег). Текстовое представление каждой страницы нам было необходимо для построения полнотекстового индекса, на основе которого реализован механизм полнотекстового поиска. При выполнении поисковой операции результаты поиска представлялись как набор текстовых представлений страниц, для каждой страницы проводилось подчеркивание найденных слов и словосочетаний. После чего от текстового представления страницы можно было перейти к ее оригинальному графическому представлению, на котором, как уже говорилось, можно получить доступ ко всей специфичной информации страницы.

Следует обратить внимание на ряд серьезных проблем и неудобств, с которыми мы столкнулись в процессе работы над архивом Соросовского образовательного журнала. Во-первых, существенной частью информации, с которой работает конечный потребитель,

являются текстовые составляющие каждой страницы. Соответственно, нам пришлось выполнить большой объем ручной работы по приведению текстовых частей каждой страницы к аккуратно смотрящемуся «читаемому» виду. При этом во многих случаях все равно не удалось избежать большого количества ошибок и неточностей, связанных с использованием в тексте специальных символов и, например, химических формул. Следующее неудобство – сопоставление найденных результатов с оригинальной страницей. Увидев подчеркнутые результаты поиска на текстовом представлении, приходилось глазами выискивать соответствующее место на графической картинке. Нам приходилось искать разумный компромисс между качеством изображения текстовой информации на растровых изображениях страниц и размером файлов с этими изображениями. Тем не менее, при достаточно большом увеличении качество отображения текста оказалось неудовлетворительным. Использованный в данном проекте формат графического представления страниц не поддерживал многостраничные документы, что очень сильно замедляло скорость работы файловой подсистемы операционной системы Windows. Таким образом проект по созданию Соросовского образовательного журнала можно назвать автоматизированным лишь с большими оговорками.

Анализируя все вышеперечисленные проблемы, мы обратили внимание на формат хранения файлов DjVu. Этот формат специально оптимизирован для хранения отсканированных документов. В нем применяется механизм «многослойности». В одном графическом слое «нарисована» текстовая информация страницы, ряд графических слоев предназначен для сохранения всей нетекстовой информации из образа страницы – например, иллюстрации, фоновые рисунки и т.д. Такой способ представления

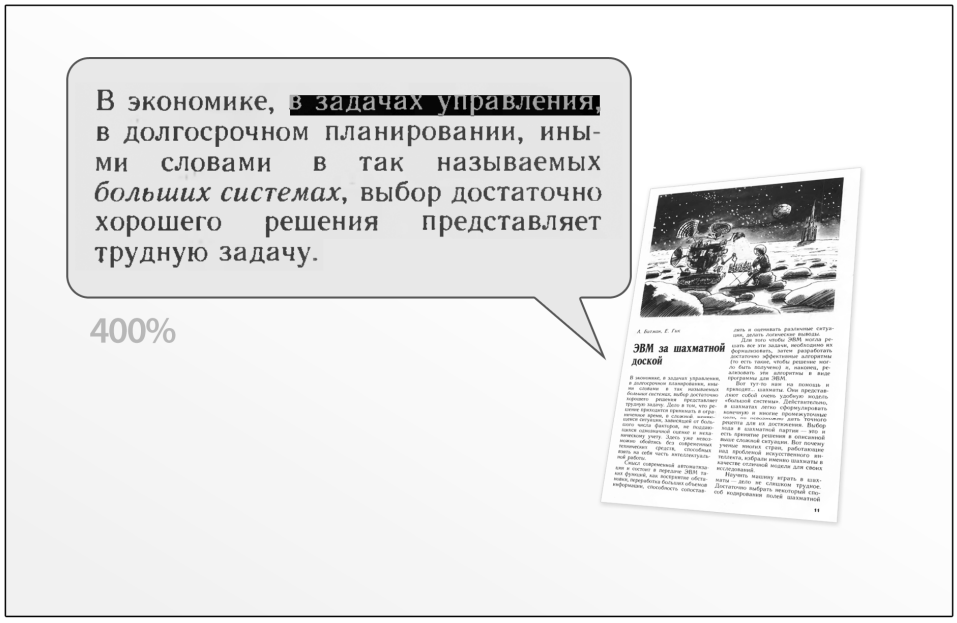

Рис 1. Качество показа растрового текста при увеличении 400% в файле формата DjVu

графической информации позволяет использовать разные алгоритмы сжатия для разных слоев, позволяет достичь отличного качества показа растрового текста при очень небольшом размере файла. На *рис 1* показано качество текста при достаточно большом коэффициенте увеличения. Важнейшей особенностью формата DjVu является возможность подкладывать текстовый слой под графический образ страницы. Пользуясь этой особенностью в процессе распознавания текста на отсканированной странице, мы можем сопоставить каждому распознанному слову прямоугольник на графическом слое, на котором нарисовано соответствующее слово. Перечисленные свойства формата DjVu позволяют автоматизировать процесс подготовки архива. Дополнительным плюсом формата DjVu можно считать возможность создания многостраничных документов.

Перечисленные свойства формата DjVu позволяют автоматизировать весь процесс подготовки архива научно-технической документации.

Издательским проектом, где нами был реализован полностью автоматизированный процесс создания архива, был архив журнала «Химия и жизнь» за 40 лет. Исходным материалом для проекта послужили бумажные журналы суммарным объемом порядка 50000 страниц. В отличии от редакции Соросовского обазовательского журнала в редакции «Химии и Жизни» не было готового электронного каталога. Чтобы не повышать стоимость проекта, было принято решение отказаться от создания авторского, тематического каталога вручную. (В настоящее время ведется работа по частичной автоматизации этого процесса с использованием семантического анализатора.) Поэтому поисковые возможности в архиве журнала ограничились, во-первых, поиском по хронологическому каталогу и, во-вторых, по полным текстам статей. Использование формата DjVu позволило весь этот объем информации при полноцветном сканировании с качеством 300 dpi поместить на 1 DVD диск! Средний размер одной страницы составил 60 килобайт. Весь распознанный в автоматическом режиме текст был подложен под графический образ страницы. Распознанный текст был использован только для проведения полнотекстовой индексации и выполнения поисковых операций. Ввиду того, что текст в явном виде нигде не показывается конечному потребителю издания, у нас появилась возможность не исправлять ошибки в распознанном тексте. Все ошибки, появившиеся при распознавании химических формул, различных специальных символов, а также ошибки, возникающие при автоматической обработке сложной верстки, не мешают нам проводить полнотекстовую индексацию.

Нами было создано специализированное программное обеспечение CCT Publisher, которое помогло объединить в одном издании отсканированные бумажные материалы и

материалы, сверстанные на компьютере. Это было достаточно актуальным потому, что номера журналов «Химия и Жизнь» за последние 4 года были доступны в электронной форме в виде PDF файлов. CCT Publisher позволяет решать следующие задачи: качественное отображение на экране компьютера отсканированных и подготовленных на компьютере материалов. В CCT Publisher реализована функция подчеркивания найденной информации прямо по картинке - графическому образу страницы (за счет имеющейся у нас информации о том, в каком прямоугольнике какое слово нарисовано). В программе существует возможность извлекать текстовую информацию, копируя ее прямо с картинки. Можно копировать образ страницы как графический файл, таким образом можно легко копировать химические и математические формулы при помощи любого стандартного графического редактора как часть картинки. Удобный и интеллектуальный поисковый механизм дает быстрый доступ ко всей необходимой информации.

Рис 2. Схема автоматизированного процесса.

Таким образом, разработанное нами программное обеспечение, а также объединение форматов файлов PDF и DjVu в едином архиве позволило решить практически все проблемы, с которыми мы столкнулись в процессе работы над Соросовским образовательным журналом.

Использование формата DjVu, а также разработанное нами программное обеспечение позволили получить полностью автоматизированный процесс создания электронных архивов с возможностью полнотекстового поиска.

В зависимости от того, какое разработанное нами программное обеспечение будет использовано, созданный архив может быть опубликован в Интернете, на CD или DVD дисках или использован в рамках локальной сети организации. На *рис. 2* представлена схема процесса, который мы активно используем в нашей повседневной работе. Надеемся, что представленный здесь алгоритм работы окажется полезным при обработке больших массивов отсканированных и электронных научно-технических материалов.

В завершение хотелось бы упомянуть наш новый продукт ИРБИС Ретрокон. Это совместное с ГПНТБ решение, ориентированное в первую очередь для создания электронной коллекции на основе отсканированной электронной картотеки. Графические образы библиотечных карточек содержат в основном текстовую информацию. Тем не менее, описанный выше подход для создания электронных архивов оказывается плодотворным для организации работы с электронными карточками. Дело в том, что электронные каталоги создаются в основном вручную. Это дорогостоящее и долгое занятие, тем более неоправданное экономически для тех карточек, которые практически не востребованы. Создание электронной картотеки по предложенной нами схеме позволяет быстро создать архив из всех карточек с возможностью поиска по полным текстам. Уникальные особенности поисковой технологии позволяют получать адекватные результаты поиска даже для карточек с плохим качеством текста.