

УДК 621:622:665:681

ЦИФРОВАЯ, ПОЛНОТЕКСТОВАЯ БИБЛИОТЕКА «НЕФТЬ И ГАЗ»

Руднев Н.А., Абызгильдин А.Ю.

Уфимский государственный нефтяной технический университет (УГНТУ), г.Уфа

Основными источниками информации всегда являлись библиотеки. За несколько тысяч лет эти источники превратились из собрания глиняных табличек в архивы огромного количества бумажных книг. При наличии электронных средств переработки информации поиск данных по книгам стал таким же архаичным, как и применение клинописи. Никто не возражает против приятного проведения времени с книгой в руках, однако при современной динамике обучения, научных исследований, развития производства и скорости изменения ситуации в производственных процессах, эта роскошь становится непозволительной.

Базой большинства современных экспертных систем является материал в электронном виде, и таким образом, предметом работы экспертов или инженеров знаний или библиотекарей будет электронная библиотека или информация, организованная в систему, выполняющую аналогичные функции.

Электронная библиотека (www.elibrary.sitcity.ru, www.ogbus.ru/library) созданная в УГНТУ объединяет тему «Нефть и газ» по технологии переработки углеводородного сырья, машинам и аппаратам химических производств, бурению и разработке нефтяных и газовых скважин, охране труда и промышленной безопасности, а также по общим дисциплинам технического вуза. В настоящее время идет работа по разделению электронной библиотеки для пользования студентами при обучении по специальностям 090800 Бурение нефтяных и газовых скважин (ГБ), 090600 Разработка и эксплуатация нефтяных и газовых месторождений (ГР), 170503 Техника переработки твердого топлива, нефти и газа (МА), 250400 Химическая технология природных энергоносителей и углеродных материалов (ТП), 250100 Химическая технология органических веществ (ТС), 251800 Основные процессы химических производств и химическая кибернетика (ТК), 330500 Безопасность технологических процессов и производств в нефтегазовой отрасли (БП), 320700 Охрана окружающей среды и рациональное использование природных ресурсов (ОС), 550800 Химическая технология и биотехнология (ХБ).

Библиотека «Нефть и газ» по объему превышает любую русскоязычную электронную библиотеку содержит книги с учебного и научного абонемента УГНТУ. Количество книг в библиотеке – более 5000 (более 1 500 000 стр.).

Исследование по созданию библиотеки проводилось на инициативной основе, использовались компьютеры с обычной, для учебных заведений конфигурацией, собранные из комплектующих различных фирм. Исследование проводилось по двум направлениям – методическому, включающему вопросы, связанные с определением тематики, необходимого количества книг, их выбора, заказа и техническому, включающему вопросы, связанные с компьютерным оснащением, вводом информации, ее обработкой, хранением и выводом. Оба направления сопровождалась разработкой программного обеспечения.

В результате исследований по методическому направлению, анализ рабочих программ вуза по темам основных специальностей показал количество книг, обеспечивающих обучение по общеобразовательным, общинженерным и специальным дисциплинам – около 600. Дополнительная литература по тем же специальностям составляет около 1000 книг. Кроме того, количество в 200 книг опре-

делено по методикам обучения, психологии и педагогике высшей школы. Таким образом, примерное количество учебной литературы для технического вуза составляет около двух тысяч. Следует учесть, что для учебных заведений различного профиля, институтов и университетов, эта цифра может изменяться, к тому же сюда не вошли лекции, методические и учебные пособия.

Исследования по техническому направлению составили большую часть работы. Сюда вошли вопросы, связанные со сканированием – наиболее трудоемкий этап, и менее трудоемкие этапы с точки зрения использования труда человека, но занимающие значительно больше машинного времени этапы: передача информации по электронным сетям, каталогизация данных, несколько этапов обработки растровых изображений, распознавание текстов, совмещение растровых и векторных изображений, индексация текстового материала.

Разработка системы поиска, индексация текстовой информации и разработка сервисных программ для реализации каждого этапа составили программное сопровождение технологии переработки информации.

Сканирование производилось при помощи планшетного сканера в 40% случаев, т.е. около 1500 книг отсканировано вручную. Ручной перевод требует значительного количества времени и технических ресурсов – для сканирования 300 страниц – около часа рабочего времени человека и около 4 часов машинного времени для обработки изображений. Ручной перевод неизбежен при переводе уникальных книг, однако часть книг может быть разрезана, что во многих случаях безболезненно для фондов, так как многие издания (особенно учебные) имеются в нескольких экземплярах. Несмотря на «кошунственность» уничтожения книг, преимущества компенсируют материальную утрату: книга обретает другую форму существования, неподвластную времени; пространственно книга не занимает практически никакого места; книга становится мгновенно доступной неограниченному числу абонентов. Автоматический перевод требует в два раза меньше времени и обработка изображений ограничивается конвертированием исходных файлов для сжатия и переименованием файлов, однако соотношение цена/скорость не всегда решается в пользу дорогостоящих бесконтактных, вакуумных или роторных (барабанных) сканеров.

В результате сканирования 1000 книг с разрешением 300 dpi (минимально необходимое разрешение для последующего распознавания текста), черно-белого изображения, получено около 300 000 файлов формата *.tif., общим объемом – 20 GB. Полученные файлы преобразованы в формат *.djvu, общим объемом – 5 GB, а также распознаны в файлы *.txt, общим объемом – 700 MB. В сумме общее количество файлов – около миллиона. Пользовательский вариант (1000 книг), включая поисковую систему и файл индексации, занимает 7,88 GB.

При обработке данных использовалось свободное от учебных занятий время двух компьютерных классов (по 15 компьютеров с тактовой частотой 500 и 900 МГц), три сервера и четыре сканера (один из них сетевой), объединенных высокоскоростной сетью.

Система, использованная при разработке методики показана на рисунке 1. Система обеспечивает автоматический перевод информации с бумажных носителей в электронный вид (сканирование), возможность быстрой перекачки информации между компьютерами системы (сети), работу с большим количеством файлов (файл-сервер), хранение значительных объемов информации на постоянных носителях (CD, DVD); весь объем информации (1000 книг) может быть переработан за

несколько недель. Производительность системы ограничивается только самым медленным этапом – сканированием. Максимальная производительность подобной машинной системы – 60 тыс. книг/год (18 млн. стр.) при 24-х часовой загрузке машин, может быть достигнута увеличением количества устройств ввода информации до 12 сканеров, с 8-ми часовым рабочим днем, при условии достаточного дебита «месторождения» информации, т.е. источников книг или исходной библиотеки.

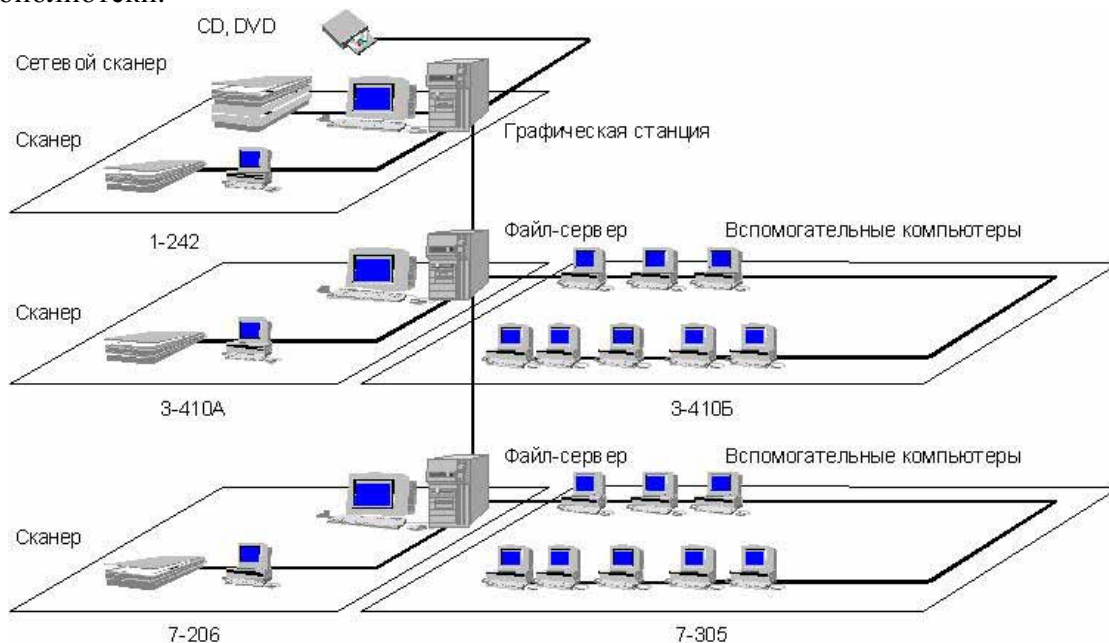


Рис. 1. Структура информационной системы

Высокая производительность системы обеспечивается организацией схемы движения информации, условно изображенной на рисунке 2, подобной для технологических систем, применяемых при непрерывных процессах переработки нефтяных фракций и химической промышленности. Непревращенные компоненты сырья после отделения от готового продукта возвращаются (рецикл) для соединения с исходным сырьем (исходной информацией).

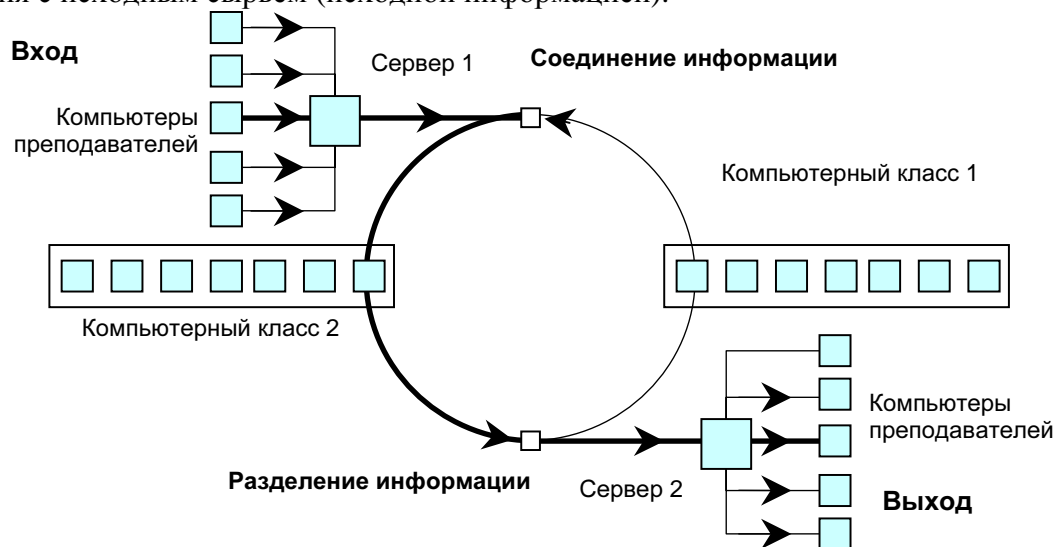


Рис. 2. Схема движения информации

Программное обеспечение разрабатывалось для сопровождения всех этапов технологического процесса перевода информации с бумажных носителей в электронный вид. Написана программа, для переименования файлов: эта задача не представляет проблем при небольшом количестве файлов, однако когда количество папок – несколько сотен, а файлов, содержащихся в них – несколько сотен тысяч, при этом каждый файл должен иметь уникальное имя, процесс занимает несколько часов. Для автоматизации процесса «работы над ошибками» программа, в числе других, содержит также модуль, для сравнения файлов и выявления различного рода ошибок.

На этапе распределения файлов по вспомогательным компьютерам системы, использовалась уникальная программа для разделения информации, контроля состояния процесса, сигнализации об окончании обработки информации и автоматического сбора готовой информации на сервер. При разработке систем просмотра, интерфейс адаптировался к стандартному виду, аналогичному web-сайтам, поэтому необходимым стандартным элементом является интернет-браузер и плагин для браузера.

Разработана также уникальная программа для пользователей библиотеки, совмещающая функции просмотра и поиска, представляющая собой систему, разработанную на основе совмещения текстовых файлов и файлов, полученных конвертацией. Для пользователей предлагается растровая картинка страницы, а также ее векторизованный текстовый аналог. В полной версии программы, используемой разработчиками, возможен выход на исходные файлы сканированных изображений, для их редактирования.

При разработке системы поиска на языке Object Pascal реализовано 11 различных алгоритмов точного поиска подстроки в строке и составлена программа для их анализа. На основе наиболее эффективного алгоритма составлена программа для определения страниц содержания и предметного указателя с последующим быстрым поиском, в противном случае осуществляется глобальный поиск в источнике.

Продолжительность поиска в настоящее время составляет 0.1-1мин/1000 источников, без потери полноты и качества. Дальнейшая разработка сервисной программы для составления файла-спецификации (индексации текста), несмотря на значительный размер такого файла (около 1,5 GB) позволила увеличить скорость – время для вывода результата по запросу в последней версии программы не превышает 5 с. при обычном режиме поиска, что практически не замедляет работу с библиотекой. При повторном запросе, по уже использованным параметрам, результат выводится без задержки.

Первоначальная версия программы содержала в составе интерфейса доступ ко всем сервисным программам, использованным при составлении библиотеки, однако обращение к ним пользователей не планировалось, поэтому в дальнейшем эта функция была исключена.

Интерфейс последней версии программы для пользователей библиотеки разрабатывался из принципа максимальной простоты, и в режиме просмотра (рис. 3) содержит два поля – список книг, содержащихся в библиотеке, расположенный в алфавитном порядке по первым авторам, и окно просмотра растровой картинки страницы. При необходимости использования части текста или цитат нажатием кнопки «Текст» страница в распознанном виде выводится в отдельном окне.

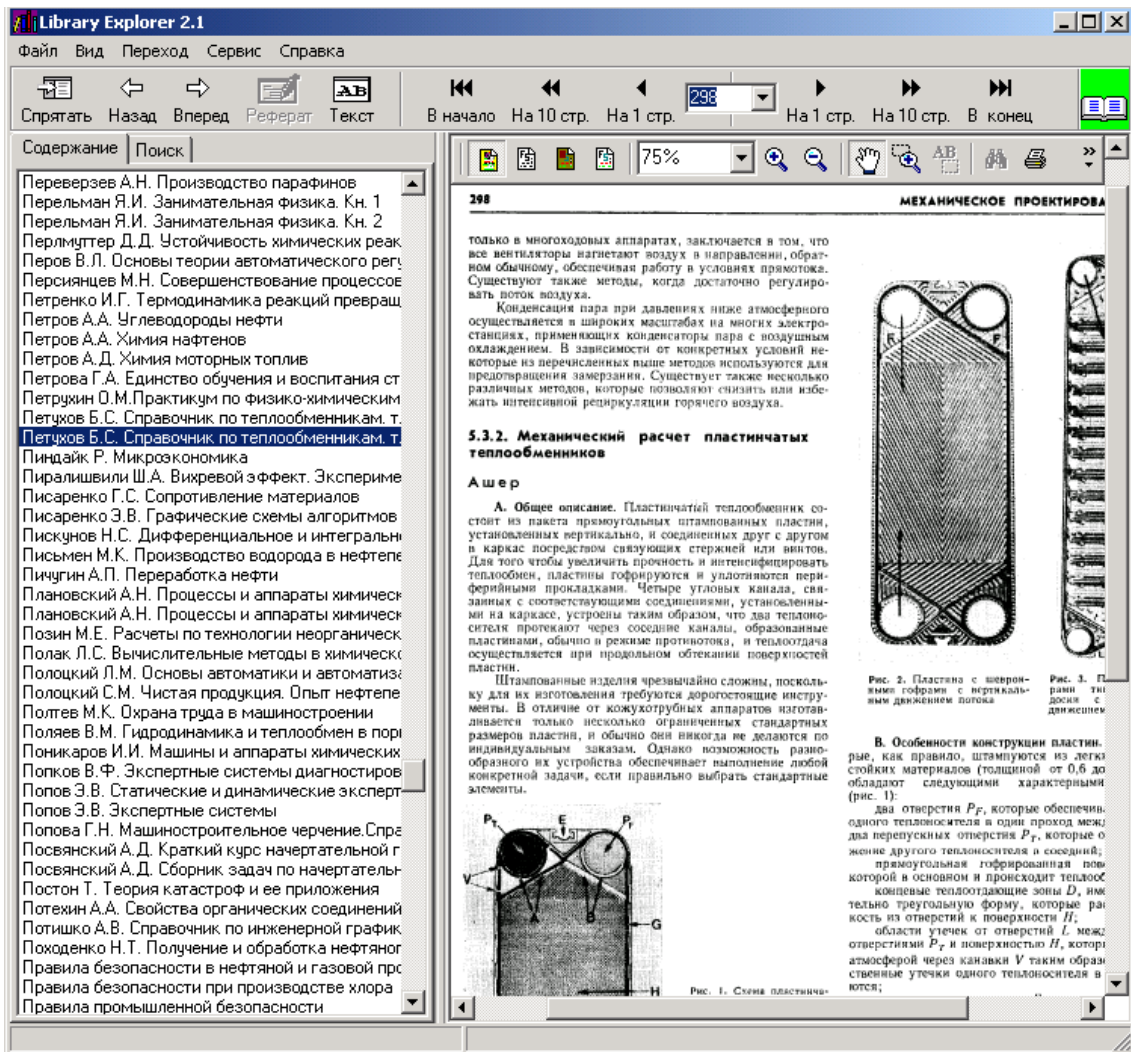


Рис. 3. Закладка «Содержание» программы просмотра и поиска

Закладка «поиск» интерфейса программы (рис. 4) содержит окно для ввода запроса, вывода результатов поиска, в котором можно выбрать необходимую страницу в растровом и текстовом (распознанном) виде, аналогично закладке «содержание». При необходимости сохранения результатов поиска нажатием кнопки «Реферат» текст всех страниц, содержащих фразу или слово по запросу будет сохранен с указанием номера страниц и источника для дальнейшей работы в стандартном редакторе «Блокнот».

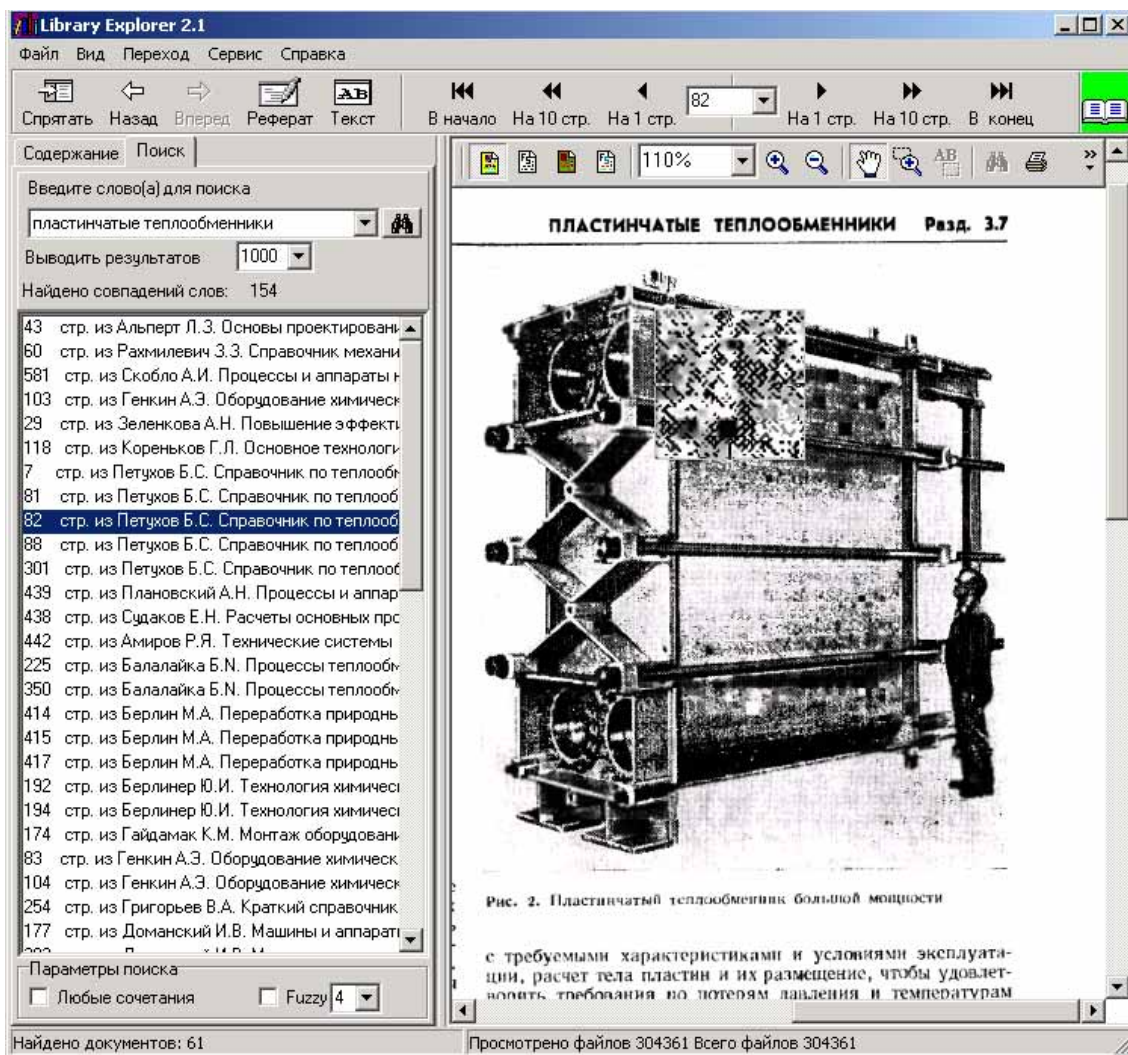


Рис. 4. Закладка «Поиск» программы просмотра и поиска

Опции поиска ограничены выбором количества найденных страниц, «любые сочетания» или вся фраза и «fuzzy». Первоначальная версия программы также содержала опции поиска по авторам, заглавиям, году издания, частоте упоминания, темам, регистрам, использовала стандартные булевы операции (и, или, и-не), характерные для всех поисковых машин (search engine). Тестирование программы показало, что увлечение максимальным набором функций, с предельным расширением спектра опций, не повышает качества поиска и достаточно оставить выбор «любые словосочетания» или поиск полной фразы. Возможно, это объясняется спецификой тематики библиотеки, так как информация ограничена довольно широкой, но одной областью науки и запрос к поиску оформляется сочетанием терминов не допускающим иного толкования (например, запрос «ректификация коллоидных растворов» не имеет смысла и дает результат или 0 или несколько сотен тысяч). Кроме того, работа с библиотекой показывает, что необходимость даже в этой функции возникает довольно редко, хотя функция «поиск в найденном» была бы полезна и в дальнейшем будет добавлена.

В программе поиска реализован также алгоритм нечеткой логики (fuzzy) для поиска слов, отличающихся приставками, окончаниями падежами и т.п., а также

при поиске или введении запроса слов с ошибками. Так, например, при поиске слова «бублуотека» (библиотека) программа дает сходжение результатов поиска 97%, при поиске слова «риктефукция» (ректификация) – 96%. В обычном режиме, несмотря на ошибки, полученные при распознавании в текстовых файлах, поиск по электронной библиотеке показывает 99,9% информации. Процент снижается при количестве букв в слове менее 3-х, например, при поиске слова «да» (найдено 12 217 совпадений) – 60%, за счет выдачи слов, содержащих искомое сочетание («тогда»), при поиске слова «нет» (14 170 совпадений) - 90%; и слов в искомой фразе больше 5-ти, однако такие параметры поиска задаются редко.

Изучение процесса и результатов поиска, конечно, является предметом дальнейших исследований, так как в некоторых случаях получены неожиданные результаты и, будучи разработчиками, мы сами не знаем всех возможностей системы. Тем не менее, уже на данном этапе библиотека является полезным информационным продуктом, при работе с литературой для экспертов, профессионалов и «новичков» - аспирантов и студентов, так как позволяет практически мгновенно выводить на экран компьютера любую страницу из 300 000, любой книги из 1000, с искомой информацией. При работе с обычной библиотекой этот процесс занял бы не менее недели, при условии, что все книги на месте. Это иллюстрируют такие образные примеры: если все книги, имеющиеся в электронной библиотеке поставить друг на друга, высота стопки будет сравнима с высотой 13-ти этажного дома, вес этих бумажных книг – более 1 тонны, в то же время вся готовая электронная библиотека помещается на 10-ти компакт-дисках.

В числе уже имеющихся возможностей библиотеки – кроме описанного выше автоматического составления подборки текста по запросу, подсчет индекса цитируемости, проверка перекрестных ссылок, проверка новизны (в пределах заложенной информации). Практически полностью отпадает необходимость в каталогах, картотеках, кодификаторах, классификации и классификаторах, словарях терминов и определений. Производится автоматическое объединение справочников, иностранных словарей и словарей иностранных слов.

Несмотря на то, что специальных опций для указанных функций в программе не заложено, при запросе на любую фамилию, количество найденных страниц, за исключением книг, автором которой имярек является и упоминается не в связи с его публикациями, является индексом цитируемости. Например, при запросе «Губкин» (один из основоположников развития учения о нефти в России) найдено 276 ссылок – (45- собств. труды, 116- историч. контекст) = 115, т.е. ИЦ = 0,115 при расчете на 1000 книг.

По ссылке на автора или публикацию, материал, при его наличии в библиотеке, можно мгновенно найти и сравнить соответствие между ссылкой и источником. При запросе «Список литературы» поиск выводит все книги, на которые производились ссылки. Анализ количества повторений, а также перекрестных ссылок показывает авторов «злоупотребляющих» использованием «чужого» материала.

Проверка новизны при запросе по ключевым словам, входящим в формулы изобретений, показывает всю информацию, сравнение года издания которой, с датой изобретения может показать его действительность. К сожалению, в составе библиотеки недостаточное количество патентов и авторских свидетельств, для получения достоверных результатов, однако даже имеющегося небольшого количества достаточно, для анализа в узкой области специализации.

Что касается каталогов и картотек, необходимость при наличии электронной библиотеки в них отпадает автоматически, так как книгу можно найти как по автору (заодно поиск показывает все ссылки на работы указанного автора), так и по названию (заодно поиск показывает и все ссылки на книгу). Алфавитный указатель по тому или иному параметру, при наличии компьютера, как известно, может быть составлен без проблем. Универсальная десятичная классификация, введенная с 1895 г., составляемая по нескольким словам, теряет свою эффективность при наличии системы, определяющей содержание книги по всем словам, в нее входящим.

Так как в состав библиотеки входит 5 словарей и словари, прилагаемые к некоторым книгам, при запросе на поиск слова в большинстве случаев результаты поиска показывают несколько переводов с русского на различные языки и наоборот. Аналогичная ситуация и со справочниками (более 70 в библиотеке): приводятся данные из всех справочников по запросу. Например, при запросе «золото» поиск, в числе других страниц, в которых это слово встречается, показывает все страницы из 5-ти справочников со всеми физическими и химическими константами. То же касается и таких терминов, например, как транс-4,5-диметил-1,3-диоксан (4 справочника).

В числе предполагаемых возможностей библиотеки – автоматическое составление рефератов, обзоров и сборников, проверка новизны изобретений (при условии перевода всех патентов и авторских свидетельств, что уже на данном этапе технически осуществимо), составление сборников стандартов и нормативной документации, архивирование проектных чертежей большого формата. Большая область для развития – внедрение в электронную библиотеку компонентов интеллектуальной обработки информации и использование в качестве базы для экспертных систем.

Конечно, при переходе на более высокий уровень развития электронных библиотек – технических библиотек промышленных предприятий и научных учреждений, а также центров научно-технической информации, увеличение объема единиц приведет к возникновению новых проблем. Однако, опыт разработки, полученный нами, показывает, что они решаемы как на методическом, так и на техническом уровне. Несомненно, также, что переход на более высокий уровень развития технологии переработки информации покажет несовершенство разработанного нами метода, однако, как нам кажется, на данный момент он является приемлемым. Есть опыт успешного внедрения в информационные системы научно-исследовательских учреждений и промышленных предприятий. Дальнейшее развитие данного направления позволит не только повысить качество учебного процесса, но и повысить квалификацию персонала промышленных предприятий, что в конечном итоге направлено на повышение эффективности и интенсификацию промышленного производства.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Feigenbaum, 1980; Boose, 1989; Gaines, 1993; Chervinskaya, Wasserman, 2000. <http://www.bekhterev.org/psy1.htm>
2. «Библиотекосведение» 2000, №6, с.3.
3. <http://www.computerra.ru/offline/2000/333/2901/>