УДК 681.3.01

Р.Ю. Вишняков

ДОКУМЕНТНЫЙ МУЛЬТИМЕДИА КАТАЛОГИЗАТОР

Учреждения, оперирующие большими объемами документов, для ускорения документооборота и автоматизации его обработки нуждаются в комплексе автоматизированного ввода документов с бумажных носителей. Ключевым звеном такого комплекса является быстрый документный сканер с соответствующим аппаратно-программным обеспечением. Поскольку сканирование организовывается в потоковом режиме, то одной из задач такого комплекса также является распределение документов в различные каталоги по некоторым признакам. В дальнейшем такую процедуру будем называть каталогизацией, а ее реализующую программную систему – каталогизатором.

Рассмотрим возможные способы реализации архива электронных документов.

В виде базы данных *(например, на основе SQL Server).* Все данные архива хранятся в базе данных, доступ к которой осуществляется с помощью SQL-сервера. Такой способ наиболее желателен, так как язык SQL и дополнительные средства для работы с БД покрывают интересы по быстрому поиску нужной информации, но он не приемлем к мультимедийным типам файлов в том виде, в каком он прекрасно работает с документными файлами. В хранении мультимедиа файлов, которые требуют добавочных описаний, возникают дополнительные трудности по накоплению ключевых данных для их архивации в БД.

В виде иерархической файловой структуры. Архив документов в виде иерархической файловой структуры представляет собой некое пространство, где по определенным условиям и правилам выполняется архивация всех данных. Недостатком данного способа организации архива является постоптимизация данных, а также недостаточно оптимизированное управление данными.

Разбиение группы файлов по темам и их сортировка по каталогам. Виртуальная файловая структура. Этот способ объединил два выше приведенных способа хранения и каталогизации информации. В данном случае объединяются гибкость запросов SQL-сервера и удобство управления данными с простотой иерархической файловой структуры. Смысл создания виртуальной файловой структуры состоит в том, чтобы не создавать физических каталогов сортировки, а объединить БД и привычное представление информации внутри каталогов архива.

Предлагаемый мультимедийный каталогизатор построен по данному принципу и функционально обеспечивает несколько режимов работы. Рассмотрим эти режимы.

Ручной режим. Данный режим позволяет работать индивидуально с каждым файлом и фактически является связующим звеном между программами, предназначенными для непосредственной работы с каталогизируемыми файлами типов документов и мультимедиа. Ручной режим облегчает функции каталогизации путем накопления всей рабочей информации, поступающей от пользователя к муль-

тимедиа каталогизатору. Функционально структура ручного режима каталогизации представлена на рис.1.

Здесь ведется журнал работы с потоками команд, осуществляется присвоение аддитивных фильтров сортировки для дальнейшего полуавтоматического или автоматического режима сортировки и т.д. Итак, ручной режим можно охарактеризовать, как режим работы без логических условий распределения документов по каталогам. Также можно отметить, что работа в ручном режиме для **Пользователя** практически ни чем не отличается от работы с EXPLORER в Windows.

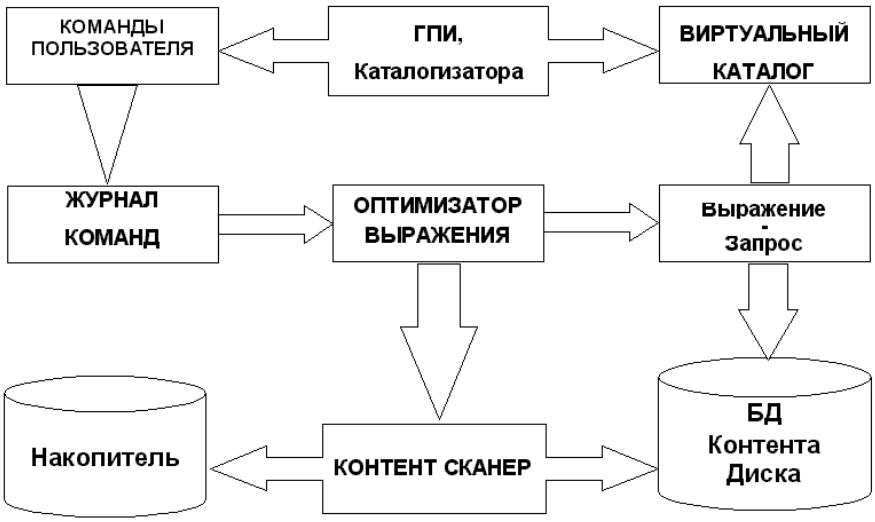

Рис.1. Функциональная структура ручного режима каталогизации

Полуавтоматический режим. В данном режиме первый этап работы осуществляется на уровне ручного режима, но при этом Пользователь вправе задавать логику сортировки файлов, а также условия аддитивных фильтров, что представлено на рис. 2 функциональной структурой каталогизации. Но сортировка в полуавтоматическом режиме при заданных логических выражениях и аддитивных фильтрах осуществляется только в случае 100% удовлетворения всем условиям, иначе **Пользователю** предлагается отреагировать на возможные варианты критических ситуаций, которые могут при этом являться дополнением к логическому выражению.

Например, рассмотрим логическое выражение, формируемое **Пользователем** в процессе работы с документным мультимедиа каталогизатором:

ВЫБОР(“C:\”;*.doc; сканеры<и>компьютеры<и>бумага;
ЕСЛИ(SOF>24K;ЕСЛИ(SOF<120K;ВЕРНУТЬ(#00001))))

После поступления действий от **Пользователя** составляется выше приведенное выражение для запроса в базу данных контента, по исполнении которого получается выборка *всех документных файлов размером более 24кб и менее 120 кб с диска C:\, содержащая слова СКАНЕРЫ, КОМПЬЮТЕРЫ, БУМАГА.*

Пример аддитивного фильтра:

ВЫБОР(WIDTH;HEIGHT;COLOR; OPTION: *.pcx; *.bmp; *.gif; *.jpg)

Данный фильтр позволяет задавать выборку по внутренним характеристикам файла, таким как глубина (бит), ширина, высота, ярлык (принадлежность).

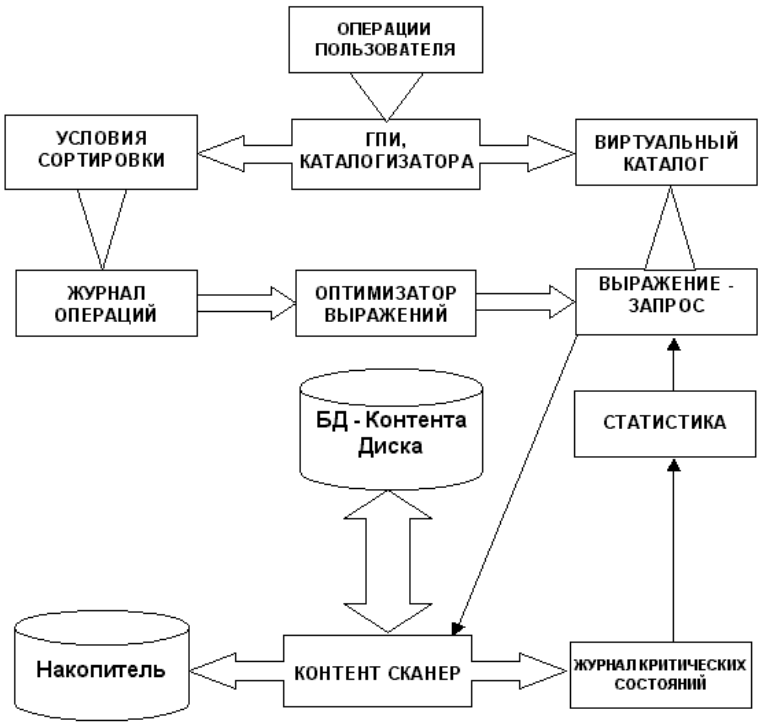

Рис.2. Функциональная структура полуавтоматического режима каталогизации

Автоматический режим. В данном режиме работа **Пользователя** заключается в том, чтобы запустить ПРОЦЕСС каталогизации документов по заранее заданным условиям, фильтрам и задать припуск логических выражений. В процессе автоматической каталогизации так же как в ручном и полуавтоматическом режиме формируется журнал действий, по которым можно совершать «откаты» до ключевых состояний. В данный режим функция «отката» была введена для профилактических и отладочных целей, но она также может пригодиться в случае неправильно заданных фильтров. Единственным недостатком «отката», является невозможность восстановления удаленной информации (например, если были использованы условия по удалению файлов или директориев – физическое удаление). В случае кризисных ситуаций в автоматическом режиме, каталогизатор принимает наиболее близкое условие к переменной каталогизации, но при этом формирует высокоприоритетную аннотацию (Report), относящуюся к данному моменту. По окончании каталогизации **Пользователь** может оценить обстановку по сгенерированной аннотации с выводом логических выражений и принятыми решениями.

В виртуальной структуре архива файлы физически находятся в одном каталоге, а логически представляются в виде независимой структуры, которую можно редактировать в процессе работы. Такая структура архива представлена на рис. 3.

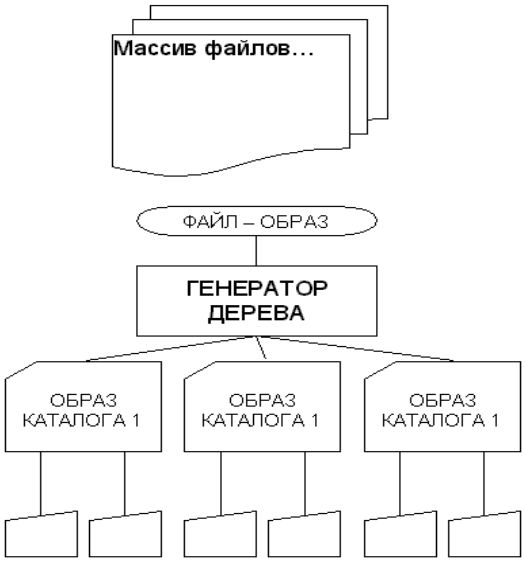

Рис.2. Древообразная структура архива

Предлагаемый каталогизатор ориентирован на создание и обработку «виртуальной» файловой структуры архива. Он реализует три режима каталогизации: ручной, полуавтоматический и автоматический.

После создания структуры архива (см. рис.1) пользователь должен получать файл с записанной в него структурой архива (Образ архива). Каталогизатор имеет два режима работы: режим ожидания (режим готовности) и режим активности.

Режим ожидания (Stand-by). В этом режиме (готовности) организован сервис, который приведен в состояние активности по умолчанию. Сервис – автономная часть программы, предназначенная для выполнения поставленных перед нею задач в автономном режиме на уровне операционной системы, не занимающая при этом значительных ресурсов ПК.

В состоянии ожидания сервис осуществляет прослушивание каталогов, дисков, файлов, заданных пользователем или внесенных в информацию образа наиболее часто используемых путей. Смысл автономного прослушивания заключен в оптимизации оперативности информации. Сервис позволяет мгновенно реагировать на изменения, происходящие в отслеживаемых ресурсах, а это, в свою очередь, – своевременно обновлять БД-ОБРАЗ о всевозможных изменениях, происходящих на прослушиваемых ресурсах. Включая такую возможность, мы отказываемся от форсированного набора информации о необходимых ресурсах в БД-ОБРАЗ.

Созданные структуры можно модифицировать. Для этого, используя файл структуры, **Пользователь** сможет просматривать архив, выбирая нужные документы, точно так же как при работе с физическими каталогами. Фактически каталогизатор создает дополнительный описательный «раздел», который позволяет работать с файлами на уровне операционной системы. Единственное отличие от работы с реальными каталогами состоит в том, что создается промежуточный сценарий выполнения каталогизации, а затем интерпретатор каталогизации начинает

выполнения скрипта. При этом происходящие за каждый опорный шаг изменения фиксируются. Благодаря ведению журнала выполнение каждого шага есть надежная транзакция.

Электронные документы имеют сквозную нумерацию (как физическую, так и виртуальную). Это позволяет при необходимости переходить на следующий или предыдущий документ в пачке. Если документ распознан плохо, то должна приводиться схема сверки по принципу 1:1 или попросту нужно просмотреть отсканированный документ-оригинал.

С целью повышения интеллектуальности системы и качества каталогизации предполагается использование подключаемых модулей, реализующих дополнительные критерии каталогизации.

Каталогизация: в качестве универсального критерия, применимого практически во всех областях, может использовать критерий ключевых слов. Выбор ключевых слов из документов представляет собой задачу с множеством вариантов решений. Необходимо рассмотреть и экспериментально опробовать некоторые варианты.

Теоретически можно добиться довольно неплохого результата при использовании следующего алгоритма полуавтоматической каталогизации:

1. Создание кланов для всей группы документов (исключение слов с использованием клана, не прошедших порога индексирования и эвристики).
2. Частотный фильтр. Из практики известно, что слова, не являющиеся названиями клана, но часто встречающиеся, можно рассматривать как ключевые. Здесь **Пользователю** необходимо только задавать порог, после которого клан будет профильтрован (остаток – ключевые слова).
3. Задание тем путем указания заголовка темы из входящих в нее понятий (возможно с заданной частотой).
4. Автоматическое формирование тем (расширенное сканирование).
5. Ручная корректировка дерева каталога (например, увеличение глубины или расширения дерева).

Другим подходом к каталогизации может быть введение направленной каталогизации, когда предмет сканирования приблизительно ясен, и мы можем определить его в какую-либо тему, которая, в свою очередь, включают в себя темы, подтемы, разделы, подразделы и понятия. Но такой метод требует постоянного обновления базы, и может выполнять абсолютно бесполезную работу при отсутствии темы направленности.

Рассматриваемый в настоящей работе каталогизатор разработан в рамках работ по созданию систем безбумажной обработки информации, проводимых в международной немецко-российской лаборатории ELDIC.

ЛИТЕРАТУРА

1. *Вишняков Ю.М. Родзин С.И.* Проблемы поиска и распознавания в презентационной графике // Международный межвузовский сборник «Проблемы и перспективы развития устройств автоматики, связи и ВТ». Ростов-на-Дону: РГУПС, 2000.