КРОНОС ИНФОРМ

Научно-производственная компания
«КРОНОС-ИНФОРМ»

Адрес: 125130, г. Москва, ул. Приорова, д. 30
телефон/факс: +7 (495) 276-1211
www.cronos.ru, e-mail: info@cronos.ru

ОСНОВНЫЕ ПРОБЛЕМЫ ОБРАБОТКИ И ИСПОЛЬЗОВАНИЯ ТЕКСТОВЫХ МАССИВОВ

Цифровизация бизнеса в настоящее время привела к тому, что в любой компании накоплены значительные объемы текстовой информации в электронном виде. Причем, эта информация представлена в различных форматах и кодировках, а хранится как на серверах компании, так и на рабочих станциях сотрудников. Найти необходимую информацию возможно только в рамках контекстного поиска в интерактивном режиме в многомерном мире различных папок. Поиск интересующих данных выливается в трудоемкую задачу, анализ при такой структуризации данных возможен только в «ручном режиме».

С ростом объема накопленной информации подобный поиск требуемых данных становится практически бесполезен. Например, даже небольшая компания за несколько лет может накопить десятки тысяч документов, а архивы больших корпораций легко переваливают за миллионы. Чтобы ознакомиться с ними, аналитику не хватит всей жизни, и зачастую он даже не может точно судить о наличии в массиве данных той или иной информации. Встает проблема правильной организации архива с возможностью быстрого поиска по нему.

В Windows есть встроенная функция поиска файлов, но ее возможности ограничены. Условием поиска может служить только одно или несколько слов, использование более сложных условий не поддерживается. И скорость поиска оставляет желать лучшего, небольшой массив из тысячи документов может обрабатываться несколько минут. В результате на выходе только список файлов, в которых встречается искомое слово. А если документ содержит несколько страниц, то, чтобы найти нужный контекст, Вам придется повторить поиск, открыв файл в соответствующей программе.

Расположенные в разных папках документы могут дублировать друг друга, перегружая результаты поиска информации многократными повторами. Причем, документы с идентичным текстом могут иметь разные названия, а документы с одинаковым именем – разное содержимое. В таких условиях выбрать вручную документы, подлежащие удалению, будет довольно сложно, да и поиск таких дубликатов может затянуться надолго.

Кроме того, документы поступают в различных форматах и кодировках, для работы с которыми требуется установить соответствующие программы. И если для просмотра современных форматов, вроде текстовых файлов и документов Word, есть средства почти на всех компьютерах, то при обработке текстов в иных кодировках и форматах могут возникнуть сложности.

Следует рассмотреть и вопросы безопасности данных, складывающихся из контроля несанкционированного доступа и защите данных от утери. Разбросанные по локальной сети папки с документами могут быть доступны сотрудникам, которым они не требуются для работы, что представляет собой потенциальную угрозу утечки или уничтожения информации. Отсутствие какого-либо контроля доступа к этим папкам не позволит выявить лиц, виновных в утечках, и даже сам факт несанкционированного доступа к информации может оказаться сокрыт. Произвести тонкую настройку доступа к папкам, в принципе, может и системный администратор организации, но это является грубейшим нарушением политики информационной безопасности, которая требует разграничения полномочий между IT –

сотрудниками и сотрудниками безопасности, и в первую очередь, по вопросам контроля доступа.

В целях защиты информации необходимо периодически создавать резервные копии массивов данных. Ручное копирование папок с документами отнимает много времени и требует определенной самодисциплины, чтобы не забывать проводить данную операцию с заданной периодичностью. Создавать резервные копии можно и автоматически, но для этого придется найти, освоить, настроить и использовать соответствующий инструмент, что само по себе достаточно трудоемко. А хранение и обработка резервных копий потребует подготовки отдельных регламентов взаимодействия. Зачастую, для этого опять же нужно будет призвать на помощь IT – сотрудника, что также идет в разрез с политикой информационной безопасности.

С учетом вышеизложенного, работа с массивами документов без использования специального инструмента окажется малопродуктивной, будет отнимать много времени и потребует расширять штат сотрудников для своевременной обработки. Конечно, во многих компаниях существуют системы документооборота, которые частично решают отдельные проблемы, но эти системы настроены для работы только с определенными типами данных, предназначены для контроля движения текущих документов между сотрудниками и не способны эффективно работать с большими массивами архивной информации.

РЕШЕНИЕ ПРОБЛЕМ, ПРЕДЛАГАЕМОЕ ДОКУМЕНТАЛЬНОЙ СИСТЕМОЙ ПОИСКА ИНФОРМАЦИИ «CROS»

В целях эффективной работы с массивами текстовой информации Научно-производственной компанией «Кронос-Информ» была разработана Документальная система поиска информации «Cros» (далее – ДСПИ «CROS»), имеющая Свидетельство об официальной регистрации программы для ЭВМ № 990498 от 14 июля 1999 г., выданное Российским агентством по патентам и товарным знакам. С ее помощью можно значительно упростить работу, сократить время поиска, а также использовать дополнительные возможности для работы с текстовыми массивами.

В ДСПИ «CROS» хранение документов осуществляется в специальном банке документов. При первоначальном заполнении банка документы копируются из папок внутрь банка, и дальнейшая работа с ними производится по созданному таким образом массиву. При поступлении новых порций данных их можно добавить в банк документов в ручном или в автоматическом режиме. Если исходные документы хранятся в архивных файлах, при загрузке производится их извлечение напрямую из архива без предварительной распаковки. В банке документов файлы хранятся в сжатом виде, что позволяет значительно снизить занимаемое ими место, при этом использование механизмов сжатия не влияет на скорость и удобство работы. Так, тестовый массив документов в форматах *.DOC, *.DOCX, *.RTF, исходным объемом 10 гигабайт, в банке данных ДСПИ «CROS» занимает 3,7 гигабайт.

Для контроля дублирования документов при загрузке доступен специальный режим, который будет проверять, вносился ли аналогичный файл в банк ранее, или нет. Условия добавления настраиваются при каждой загрузке, путем активизации опций контроля дублей. Проверку можно производить по имени файла, по дате создания, по его размещению и по содержимому, используя формируемый автоматически уникальный идентификатор «сигнатуру», а также по любой задаваемой комбинации этих параметров. Удаление дублей доступно и в режиме массовой коррекции, после загрузки документов в банк, как по полному совпадению контролируемых параметров, так и используя опцию «похожести документов».

Необходимо отметить, что ДСПИ «CROS» понимает текст в различных форматах и кодировках, что позволяет не заботиться о конвертации документов в определенный вид, а загружать их «как есть». Система автоматически распознает формат документа и для поиска соответствий в каждом из них переводит поисковые условия в подходящий этому документу вид. Для обработки файлов различных форматов не требуется обязательная установка

соответствующих им программ, необходимые для распознавания текстов средства встроены непосредственно в нашу программу.

Программные средства системы позволяют создать структуру «областей поиска» (например, можно разбить документы по их тематике, по структурным подразделениям, к которым они относятся, по годам и т.д.), и приписывать загружаемые файлы к одной или нескольким из них. В дальнейшем поиск можно производить не по всему массиву, а только по выбранным областям поиска в случае, если требуется отобрать файлы, относящиеся только к каким-то определенным разделам. Также предусмотрен сквозной поиск по всем, или предварительно выбранным, банкам данных по заданному критерию отбора.

Внутри банка документов каждый файл имеет набор атрибутов: дата и время создания файла, дата и время добавления файла в банк, имя файла, имя папки, из которой был взят файл, и содержимое файла. Все эти атрибуты тоже могут участвовать в поиске, когда требуется, например, выбрать файлы, добавленные в определенный период времени, или файлы определенного формата. Помимо стандартных атрибутов пользователь может создавать свои собственные, размещая в них дополнительную информацию, относящуюся к этим файлам, которые могут заполняться как вручную, так и автоматически на основании заданных условий. Например, если известно, что в загружаемых документах в первой строке указан автор, можно создать правило, по которому в атрибут «автор» будет добавляться содержимое первой строки. Или в файле название документа указано после контекста «Название:», то атрибут можно заполнять содержимым, идущим после указанного слова. В дальнейшем, все пользовательские атрибуты также можно будет использовать при поиске и сортировке. Это крайне актуально при обработке однотипных для компании текстовых документов, типа актов, счетов и т.д.

Даже при создании гигантских банков в десятки гигабайт скорость поиска информации в ДСПИ «CROS» остается высокой. Там, где стандартный поисковик Windows будет тратить десятки минут для самых простых видов поиска, ДСПИ «CROS» справится за секунды. Необходимо отметить, что время обработки запросов в системе далеко не пропорционально объему банка документов и, фактически, не выходит за диапазон режима on-line. Тесты показали, что запрос из одного слова по содержимому файлов в папке, содержащей более 40`000 файлов, встроенный поисковик Windows отработал за 16 минут, а аналогичный запрос по банку ДСПИ «CROS», в который была загружена эта папка, выдал результат через секунду. И, наконец, запрос по одному слову в банке данных из 100`000`000 документов (больший мы создавать не стали) составил также одну секунду.

ДСПИ «CROS» позволяет делать запросы не только по одному или нескольким словам, но и добавлять сложные логические условия: например, найти документы, в которых одновременно присутствует одно заданное слово и отсутствует другое; документы, где несколько слов расположены рядом друг с другом или в заданном диапазоне промежуточных слов; документы, где поисковый контекст может встречаться с опечаткой в произвольном месте, и т.д.

В отличие от других поисковиков, которые предоставляют пользователю только список найденных документов, ДСПИ «CROS» позволяет эффективно работать с полученной выборкой. Встроенный просмоторщик прямо в окне программы показывает текст из отобранных документов и подсвечивает в нем поисковые слова. Текст не требуется листать вручную, система позволяет быстро переходить от одного подсвеченного слова к другому. Если при поиске была включена опция «подсчет числа вхождений», то для каждого из отобранных файлов будет указано, сколько раз в нем встречаются искомые слова, что позволяет по параметру «число вхождений» производить сортировку, выделяя наиболее информативные документы из списка.

ДСПИ «CROS» имеет широкие возможности в организации и настройке системы доступа. Есть возможность создать отдельный логин и пароль каждому пользователю и указать, какие действия он может производить. Например, можно ограничить возможности пользователя только поиском информации и подготовкой отчета, закрыв режимы редактирования и

добавления записей. Можно скрыть от пользователя часть областей поиска или даже настроить видимость отдельных документов на основании определенной иерархии: например, чтобы руководитель видел документы, добавляемые всеми подчиненными, а подчиненные видели только свои документы, или документы своей рабочей группы. Предусмотрена возможность активизировать журнал учета действий пользователей, автоматически сохраняя в отдельном банке, какие запросы он производил, какие документы добавлял, удалял или редактировал.

Для защиты данных от потери и повреждения следует периодически создавать резервные копии банков, для чего создан режим, позволяющий переписать весь банк в один компактный файл резервной копии. Чтобы не выполнять эту операцию вручную, можно настроить автоматическое создание копий в заданный момент времени (например, ночью, когда никто с банком не работает). Автоматизация работы производится с помощью режима «Планировщик», который, помимо создания копий банков, позволяет выполнять и другие работы: добавление документов в банк, проверку банка на наличие ошибок, оптимизацию банка и т.д.

Если функция «Копирования банка» закрыта средствами ДСПИ «CROS», то восстановить банк для работы на версии программы с иным серийным номером путем его физического копирования невозможно.

ЗАКЛЮЧЕНИЕ

Использование ДСПИ «CROS» значительно облегчит работу с документами, а рутинные операции предоставит выполнять программному обеспечению по расписанию. Это высвободит рабочее время сотрудников для других задач. Использование системы доступа защитит данные от утечки, а автоматически создаваемые резервные копии позволят не волноваться об их порче и утрате.

Стоимость одной лицензии ДСПИ «CROS» составляет 5`000 рублей, а невысокие системные требования позволяют использовать ее на обычном офисном компьютере. Доступна и сетевая версия программы, обеспечивающая одновременный доступ к банкам данных с различных станций локальной сети.

Для ознакомления с ДСПИ «CROS» с официального сайта нашей компании можно скачать Руководство пользователя и пробную версию программы (http://www.cronos.ru/download-demo.html), которая позволяет протестировать все режимы работы, и ограничена только максимальным объемом создаваемого массива в пять тысяч документов.

Отвечу на все Ваши вопросы.

Начальник отдела по работе с клиентами
Андрей Жуковский

Andrey@cronos.ru
+7 (495) 276-12-11 доб. 105