Федеральное агентство по образованию

Государственное образовательное учреждение
высшего профессионального образования

«МАТИ» – Российский государственный
технологический университет им. К.Э. Циолковского

Кафедра «Высшая математика»

ПРОГРАММЫ ПОИСКА ИНФОРМАЦИИ В ПОЛНОТЕКСТОВЫХ БАЗАХ ДАННЫХ

Аналитический обзор

Составитель: Захарченко В.М.

Москва 2005

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

Тенденции развития.

Поиск информации - в полнотекстовых базах данных одно из самых быстро развивающихся направлений в информационных технологиях. Причин такого роста несколько:

В глобальной сети – Интернете стали доступны огромные массивы текстовой информации. По-видимому, скоро вся информация, накопленная человечеством (текстовая, графическая и пр.) будет оцифрована и складирована на серверах глобальной сети. Это значит, что любая информация в принципе доступна и может быть выведена на любой компьютер, подключенный к сети. Точно также как доступна иголка в стоге сена, ее очень легко взять, если знать где она лежит.

Другая причина в том, что существующие технологии поиска информации в интернете недостаточно эффективны при поиске в массивах информации такого объема. В результате поиска по запросу поисковая система обычно выдает огромное количество ссылок, большинство которых не отвечают запросу, и являются для вас информационным мусором. Это конечно не стог сена, но и не чисто нужная информация. И чем дальше развивается Интернет, чем больше накапливается в нем информации, тем труднее найти в нем то что нужно. В какой-то степени можно даже говорить о кризисе поиска информации в Интернете.

Также за последние годы резко выросли технические возможности современных компьютеров по хранению больших объемов информации и по скорости ее обработки. В результате стало возможным создавать полнотекстовые базы, включающие миллионы страниц тематически отобранных материалов: из интернета, из баз данных на серверах локальных сетей предприятий, с тиражируемых носителей, из любых других источников. На персональных компьютерах индивидуальных или корпоративных пользователей, на отдельных серверах в интернете стало возможно накапливать электронные архивы, эквивалентные десяткам тысяч томов и проводить в них поиск в реальном времени за 1-2 сек.

В результате оказалось, что реальную возможность высокоэффективного поиска информации дает двухэтапный процесс поиска, где первый этап, это предварительный поиск и отбор информации в тематические базы данных, а второй этап, это поиск нужной информации конечным пользователем в сетевых или локальных полнотекстовых базах. Эффективность поиска в полнотекстовых базах данных, в предварительно отобранном по каким то признакам материале значительно выше, чем поиск в массиве разнородной информации.

Подтверждением тому является бурный рост количества баз данных представленных в сети и на CD. Огромное количество юридических систем, энциклопедий, справочных изданий, электронных библиотек. Параллельно также развивается рынок программ для поиска информации в больших текстовых массивах или полнотекстовых базах данных объемом от единиц до десятков и сотен гигабайт.

Ранее такие программы использовались в основном на серверах поисковых систем типа Яндекс, Рамблер, Jahoo, Google и пр. Работа с большими объемами данных и высокая скорость обеспечиваются в таких системах во многом за счет аппаратной базы, использующей множество мощных серверов, объединенных в единую сеть поисковой системы. Ясно, что эти системы создавались как штучные, не предназначенные для массового тиражирования.

Возросшие возможности современных ПК позволяют иметь на сервере локальной сети или локальном компьютере поисковые системы, по своим техническим характеристикам (скорости поиска, объему базы данных) не уступающие большим системам, а по эффективности поиска, превосходящие их, за счет поиска не во всей глобальной информационной куче, а в массиве информации, прошедшем предварительную интеллектуальную обработку.

Следует ожидать, что спрос на тематические базы данных будет стремительно расти и дальше. Появление поисковых систем нового поколения, использующих смысловую оценку содержания текстов и документов, скорее всего приведет к еще большему спросу на такие базы. Дело в том, что интеллектуальный поиск требует большой вычислительной мощности.

Какие бы суперкомпьютеры не использовались в качестве серверов поисковых систем в интернете, все равно, на всех пользователей мощности не хватит. Удел глобального поиска

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

в интеренете – это быстрый поверхностный поиск, использующий сравнительно простые алгоритмы, с выдачей множества ссылок. Индивидуальный пользователь ПК имеет ресурсовооруженность в тысячи раз большую. Он может позволить себе запустить любую программу поиска и анализа информации в базе на сотню гигабайт и на несколько минут и на несколько часов, например на ночь. В результате он получит исчерпывающую информацию по интересующему его вопросу. Эффективность такого поиска не идет ни в какое сравнение с существующим поиском в Интернете.

В будущем возможно появление поисковых систем, выполняющих поиск и смысловой анализ информации в масштабах глобальной сети. Такой поиск, по алгоритмам близкий к реализуемым в мозгу человека, могут обеспечить только полностью параллельные вычислительные системы с многосвязной структурой, подобной нейронной сети. Эквивалентная вычислительеая мощность таких систем в 10^8-10^9 раз превышает мощность современных компьютеров. Несмотря на огромную разницу в производительности, начало производства таких систем реально в ближайшее десятилетие. Экспериментальный вариант подобной поисковой системы на основе оптической нейронной сети был разработан автором ещё в начале восьмидесятых годов.

Сейчас, в силу вышеперечисленных причин, конкуренция на рынке ПО для массовых поисковых систем быстро возрастает. Ряд фирм занимается созданием программ для персональных компьютеров работающих с локальными или сетевыми полнотекстовыми базами данных большого объема. В игру включились такие серьёзные фирмы как Oracle и Microsoft. Однако успех может прийти и к мелкой фирме, и даже программисту одиночке. Пример тому поисковая система Google, когда за несколько лет был пройден путь от созданной студентом удачной программы поиска, до компании с капитализацией в миллиарды долларов. Правда, массовых программ поиска информации, предназначенных для конечного пользователя, на рынке пока очень мало. Купить серьёзную программу отечественного или зарубежного производства в «коробочном варианте» достаточно сложно. Обычно их устанавливают или разрабатывают под заказ, а значит рынок массовых поисковых систем еще не устоялся. Это хороший шанс для российских программистов, которым достаточно сложно вклиниться в старые, десятилетиями отрабатываемые западными фирмами сегменты рынка ПО, такие как СУБД или EPR. Поисковые системы находятся в состоянии интенсивного развития и требуют привлечения новых идей из математики, лингвистики, семантики, искусственного интеллекта, теории нейронных сетей - областей, которые вполне соответствуют специфике российского научного и творческого потенциала.

Я начинал заниматься информационно-поисковыми системами всех видов еще в семидесятых годах прошлого столетия. Вернувшись после некоторого перерыва к этому увлекательнейшему занятию, я попробовал разобраться в том, что творится в области разработок ПО для поиска информации на российском рынке. Даже при поверхностном анализе оказалось, что у нас существует несколько десятков фирм и коллективов, ведущих серьёзную работу в области разработки систем поиска информации. Причем не только на уровне программирования, а на серьёзном научном уровне, включая исследования в области лингвистики, семантики, анализа смысла текста и прочих вещей, без которых серьёзная информационно-поисковая система работать не может. Имеющиеся на российском рынке зарубежные программы, как правило не используют лингвистическое и программное обеспечение для анализа текстов на русском языке.

Основные свойства и характеристики поисковых систем.

Область применения программы.

Самое простое применение поисковой системы, это поиск текстовых файлов на собственном компьютере. Нечто вроде дополнения к файловому менеджеру. Вещь очень нужная и полезная. Вы вводите слова, содержащиеся в тексте документа, вам выдается список файлов. Такие системы, пригодны для поиска в массиве из нескольких тысяч небольших документов, расположенных на ПК пользователя. Эти системы ищут только документы, для поиска информации, например справочного характера, они непригодны.

Поисковые системы для корпоративных пользователей. Такие программы предназначены для работы с массивами текстовых документов предприятия имеющих

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

объемы от нескольких гигабайт до нескольких десятков гигабайт. Кроме того, такие программы обязательно реализованы в сетевом варианте, при котором доступ к базе данных на сервере локальной сети, осуществляется с рабочих станций сотрудников.

Поисковые системы для интернет проектов. Предназначены для поиска html документов в Интернете. Рассчитаны на упрощенный поиск в большом количестве небольших документов. Результат поиска в таких системах – список ссылок на html файлы в сети плюс короткие цитаты из контекста, обычно по одной. Из-за больших объемов информации в сети эти программы должны иметь очень высокую скорость поиска. Из-за большого количества мусора в сети, необходима сортировка выдачи по степени релевантности или другим критериям (например рейтингу сайта).

Самое сложная задача, это поиск информации в больших полнотекстовых массивах. В базы данных таких систем могут закачиваться любые текстовые источники информации, в том числе большого объема: энциклопедии, справочники, архивы периодических изданий, целые библиотеки специальной литературы, архивы документов корпораций, специализированные архивы типа исторических, патентных, судебных, расшифровки разговоров, протоколы и многое другое. Если в ответ на Ваш конкретный запрос система выдаст ссылку на энциклопедию, то это Вряд ли Вас обрадует. Если в этой энциклопедии сто ответов на запрос, то система должна обработать каждый и выдать отдельно все соответствующие тексты. Такая система должна искать не просто документы, а информацию, содержащуюся в них.

Если поисковая система предназначена для индексации и поиска информации в глобальной сети или для доступа к большим хранилищам текстовой информации, объёмом до десятков терабайт, то программное обеспечение системы разрабатывается специально для комплекса серверов, в качестве которых используются мощные специализированные компьютеры типа кластерных систем, имеющих десятки параллельно работающих процесоров и большой объем оперативной памяти. Например, поисковая система Google в качестве аппаратной базы использует сеть из нескольких тысяч таких суперкомпьютеров, размещенных по всему миру.

Программы для различных категорий пользователей

Программа для реализации собственного проекта. Обычно создается для поисковой системы в интеренете, интранет сети большой организации, крупного банка текстовых данных с доступом через сеть (например национальная библиотека). Для реализации проекта создается команда проектировщиков, программистов и т.п., которая самостоятельно или с посторонней помощью создает, сопровождает и развивает систему.

В случае успешно созданного проекта, комплекс программ может быть доработан до необходимой степени универсальности и использован для разработки поисковых систем на заказ. Самостоятельно такой программный комплекс не поставляется, так как требует конфигурации и настройки программных средств под требования заказчика, частичной доработки программ, постоянного сопровождения на случай сбоев системы.

Если программный комплекс доработан настолько, что -покрывает потребности большого круга пользователей, не требует постоянного сопровождения разработчиков, имеет программный интерфейс, доступный программистам среднего уровня, сопровождается качественной документацией, не использует чужих компонентов без лицензии, то он может поставляться на рынок как инструментарий разработчика. В этом случае фирма-поставщик имеет более-менее определенные цены на свою продукцию. Обычно имеется несколько стандартных версий, представляющих урезанные варианты полной конфигурации.

Программы для конечного пользователя. Представляют собой готовый коммерческий продукт. Имеют хорошо отработанный пользовательский интерфейс, позволяющий обычному пользователю управлять всеми функциями системы. Обычно разработаны «с нуля», без использования «чужих» программных средств. Такие программы распространяются в «коробочном» варианте по определенной цене.

Состав программы

Программа работы с полнотекстовыми базами данных обычно состоит из нескольких функциональных блоков:

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

Программа сканирования файловой структуры исходного массива документов. Исходный массив документов может размещаться в интернете, локальной сети предприятия, на дисках персонального компьютера. Задача программы, достучаться до каждой директории и каждого файла и передать файл соответствующей программе обработчику.

Комплекс программ извлечения текстовых данных из файлов различных форматов. Такие программы называются парсерами от английского parcer - программа выполнения грамматического разбора. Например парсер для файлов формата .doc, созданных редактором MS Word. Парсер для HTML, pdf и других типов файлов. На вход парсера поступает документ в формате исходного редактора, на выходе чистый текст для индексирования.

Программа создания индекса. Индекс полнотекстовой базы это файл, в котором записана информация о каждом слове исходного массива документов: к какому документу оно принадлежит, в какой части документа находится, относится оно к заголовку, основному тексту и т.д. Структура индекса зависит от используемого метода доступа к текстовым данным. Например, при использовании инвертированных списков создается словарь, где к каждому слову приписывается список номеров всех документов или текстовых фрагментов, в которых оно содержится.

Программа леммитизации, работающая с морфологическим словарем. Морфологический словарь содержит так называемые парадигмы (конструкции) слов русского языка, в виде базового слова и соответствующих ему форм. Для существительных это именительная форма единственного числа плюс все падежные формы, формы единственного и множественного числа. Обычно пользователю при поиске нужен текст с любой формой слов запроса. Чтобы не заставлять его перечислять все возможные формы этого слова, слово при индексации и поиске заменяется на базовое. Эта процедура называется леммитизацией. На входе слово в любой форме, например «иду» или «шел», на выходе исходная форма «идти».

Программа обрезки окончаний слов. Часто используется вместо программы леммитизации из-за своей простоты. Дает худшие результаты поиска, так как не учитывает родственные связи слов имеющих близкий смысл, но разное написание. Например: идти-шел, лет-годов, лев-львов и т.д.

Программа поиска слов в базе данных. Существует три основных варианта организации поиска в текстовом массиве:

а) Контекстный поиск, при котором весь текст последовательно просматривается программой поиска, слова сравниваются с запросом, выполняются логические операции и дополнительные условия поиска. Контекстный поиск позволяет просто реализовать самые сложные виды поиска, но имеет существенный недостаток - он очень медленный. Скорость просмотра порядка нескольких мегабайт в секунду, поэтому время поиска в базе объемом 10Гб займет более часа, что совершенно неприемлемо.

б) Подокументно контекстный поиск. Этот поиск использует предварительно созданный индекс, в котором есть списки слов каждого документа. В результате поиска по индексу определяется документы, содержащих слова запроса. Например на запрос «Александр Васильевич» будет выдан список всех документов, содержащих слово «Александр» и слово «Васильевич». Как эти слова расположены в тексте документа не учитывается. Например в начале документа есть Александр Петрович, а в конце Семен Васильевич. Этот документ будет выдан индексом. Поэтому для уточнения результатов, программа производит контекстный поиск в найденных документах. При этом учитывается расстояние между ключевыми словами, их взаимное расположение. Будут отфильтрованы документы содержащие только нужное сочетание «Александр Васильевич» Такая подокументно-контекстная схема пригодна для поиска небольших документов, типа писем, приказов и пр. Во первых, в маленьких текстах меньше вероятность случайного совпадения всех слов запроса. Во вторых, контекстный просмотр, таких документов занимает меньше времени. При поиске в больших базах данных, содержащих тексты книг, справочников и т.п. поиск сильно замедляется.

в) Индексный поиск по всему содержанию документов. Это самый сложный и быстрый вид поиска. Индекс содержит полную информацию о всех словах текстов базы данных, включая взаимное расположение слов. Содержание запроса сравнивается одновременно со всем полем информации, содержащейся в базе данных. При этом поиске ищутся не документы, а нужная информация. Затем по найденным фрагментам текста

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

выдаются тексты самих документов. Скорость такого поиска на ПК до десятков гигабайт в секунду.

Лингвистическое обеспечение

При анализе текста в серьезных поисковых системах используются комплекс словарей отображающих экспертные знания о словообразовании данного языка. В этот комплекс могут входить:

- словарь словоизменения (морфологического анализа)
- словарь моделей управления предикатов русского языка
- тезаурус общей лексики (прежде всего синонимы и обобщающие понятия)
- специальные словари и правила, например словари служебных идиоматических единиц (многословные предлоги, союзы, наречия, вводные), часи составных наименований организаций и др. В основе морфологических словарей русскоязычных поисковых систем как правило лежит Грамматический словарь Зализняка.

Некоторые системы имеют модули работы со смыслом текста, используя семантические сети, ассоциативные связи и прочие высоконаучные вещи. Такая интеллектуальная система может найти в тексте фрагменты по смыслу отвечающие запросу, хотя они вообще не содержат слов из запроса.

Сетевые версии поисковых программ.

В описаниях многих программ заявлена возможность работы в локальной сети. Сетевые версии программ сложнее локальных и стоят значительно дороже. Существует несколько вариантов сетевого исполнения поисковой программы:

Первый и самый простой, это возможность поиска в сетевом окружении. Такая программа может индексировать файлы, расположенные не только на своем компьютере, но и на дисках других ПК, объединенных в локальную сеть. При этом поиск может осуществляться только с ПК, на котором установлена система и расположена база данных, включая поисковый индекс. Многопользовательский режим, при котором пользователи со своих ПК могут обратиться в базу данных за информацией, такая система не реализует.

Второй вариант это поисковые системы работающие по интернет протоколу. В этом случае база данных и основная программа установлены на центральном сервере локальной сети, а все пользователи имеют доступ к информации со своих компьютеров через стандартный интернет броузер, такой как Internet Explorer. То есть все происходит точно также, как и при поиске в глобальном интернете. Пользователь работая в интранет сети, для доступа к базе данных вводит специфические адреса интранетовских баз данных и дальше ищет информацию по стандартной схеме, со стандартным интерфейсом поисковой системы. Естественно, что поисковые программы, созданные на основе поисковыхъ интернет систем, в основном используют интернет протоколы для организации многопользовательской работы, например Яндекс.

Следующий уровень, это программные системы имеющие клиент-серверную архитектуру с собственной клиентской частью программы. Программа клиент устанавливается на всех рабочих станциях сети, а программа сервер обеспечивает индексирование информации всей сети, создание базы данных на сервере и доступ к ней всех пользователей. Такие системы сложнее в разработке, но их функциональные возможности больше, чем у использующих стандартный броузер. Например для разграничения доступа пользователей к различным видам корпоративной информации используются мощные системные средства, интерфейс пользовательской части можно сделать намного функциональнее и дружелюбнее.

Требуемые аппаратные ресурсы.

Если программа предназначена для использования в качестве вспомогательного средства, помогающего ориентироваться в содержании дисков вашего компьютера, то чем меньше ресурсов она требует, при приемлемых результатах работы, тем лучше.

Если качество работы поисковой системы критично для результатов вашей профессиональной деятельности, то подход должен быть иным. Современная программная система, предназначенная для решения сложной интеллектуальной задачи информационного поиска должна полностью использовать все гигагерцы и гигабайты компьютера пользователя. Назначение такой системы – экономить время, силы и деньги

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

пользователя, а не быстро дешевеющую память и процессорную мощность. Поэтому скромные требования к аппаратным ресурсам, это не обязательно хорошо.

Операционные системы. (Программная платформа).

Основная операционная система, используемая на большинстве персональных компьютеров, и в локальных сетях небольших предприятий это конечно ОС Windows. В основном версии 2000 и XP, с ядром NT. Современные сетевые приложения работающие в среде этих ОС могут быть рассчитаны на интеграцию в локальные и клиент-серверные приложения разработчика через 32-битный COM (Component Object Model) API (Application Programming Interface). (COM+) и не всегда совместимы с предыдущими версиями Windows, выпуск которых уже практически прекращен. Поэтому программное обеспечение системы поиска должно работать под управлением Windows 2000, XP,NT. Желательно и под Windows-98, по крайней мере для клиентской части программы.

Многие корпоративные сети и интернет серверы работают под управлением различных версий ОС Unix и Linux. Поэтому наличие версий ПО под эти платформы является необходимым для потребителей этого круга. Другим же пользователям это говорит о солидности разработки и о серьёзных намерениях фирмы разработчика.

Базовая СУБД

Любой поиска информации, простой или сложный, начинается с доступа к текстовым данным в полнотекстовой базе. Поэтому метод доступа к данным (деревья, хэш таблицы, инвертированные списки и др.) это тот фундамент, на котором строится вся поисковая система. Он относится к ядру СУБД – комплекса программ, предназначенного для работы с базами данных. Так как базы данных могут быть разные: реляционные, полнотекстовые, иерархические, то и методы доступа должны соответствовать специфике данных и соответствующим алгоритмам обработки. Например, наиболее распространенные реляционные СУБД, предназначенные для табличных данных, не лучший выбор для работы с массивами текстовой информации. Поэтому для разработчика ПО системы поиска информации, первостепенным является вопрос выбора СУБД. Для потребителя это также важнейшая характеристика системы, во многом определяющая все ее возможности.

Для разработчиков существуют разные варианты реализации доступа к данным в системах поиска:

Первый, это самостоятельная разработка с нуля. Имеет большие преимущества, так как наиболее полно реализует выбранную математическую модель поиска и поисковые алгоритмы. Позволяет разработчикам полностью владеть всеми компонентами системы, что необходимо для процесса её развития и адаптации под новые задачи. Также важно, что разработчик имеет авторские права на все части системы. Это во многом упрощает маркетинг и проблемы с лицензированием. Самостоятельная разработка дает возможности производителю продавать ПО как продукт для конечного пользователя, как инструмент для разработчиков, разрабатывать специализированные заказные системы, для сетевого и локального доступа, поисковые оболочки для тиражируемых баз данных на CD.

С точки зрения пользователя такая система также предпочтительней: разработчик отвечает за систему в целом, предоставляет широкий спектр возможных доработок, система имеет большой потенциал перспективного развития, более гибкая ценовая политика фирмы разработчика, порядок с лицензированием.

Наиболее успешные фирмы на российском рынке, например ABBY, 1С, Яндекс, разрабатывают свои программные системы самостоятельно.

Сложности самостоятельной разработки также очевидны. Самостоятельная разработка должна основываться на математической модели и алгоритмах в чем то провосходящих имеющиеся на рынке системы, иначе она не имеет смысла. Для реализации программы уровня ядра СУБД или оригинального метода доступа нужны системные программисты очень высокого уровня квалификации.

Второй вариант, использование готовых программных модулей производства других фирм в качестве основы, на которую наращиваются специализированные программные средства прикладного уровня. Для этого могут использоваться СУБД Oracle, имеющая модуль полнотекстового индексирования Oracle Text Mining. Некоторые разработки

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

ориентированы на СУБД MS SQL Server 2000, также имеющую функции работы с базами текстовых данных.

Ситуация может быть и обратной, когда разработчик программ доступа к текстовым данным, использует разработки других фирм, в области лингвистического обеспечения. Например фирма 1С при создании программных оболочек для различного рода текстовых баз на CD использует морфологический словарь фирмы «Орфо».

Для реализации достаточно крупного проекта может быть закуплен инструментальный пакет программных средств, таких как Yandex Professional, Yandex Enterprise, предназначенный именно для разработчиков интернет проектов.

История фирмы разработчика

История коллектива разработчиков – это важный показатель качества программного продукта. Потребительские качества программного продукта во многом определяются опытом создания поисковых систем. Многие коллективы разработчиков вышли еще из советских времен, из команд разработчиков отраслевых информационных систем.

Маркетинговая политика (Цены)

Так как в задачу этого материала входит потребительский анализ программ, то их цена одна из самых важных характеристик поисковой системы.

Многие фирмы выпускают свой программный продукт в нескольких версиях с разным функциональным набором и разной ценой. Самые простые версии могут распространяться бесплатно.

Сетевые версии содержат цену на серверную часть программы и клиентские программы для рабочих станций. Обычно цена устанавливается на пакет из сервера и нескольких (5-10) программ клиентов.

В цену программы может включаться стоимость обслуживания и поддержки: консультации по телефону и e-mail, выезд специалиста, замена программ на новые версии. Чаще цены на обслуживание устанавливаются отдельно.

Перечень основных характеристик поисковой системы, используемый для анализа.

Страна производитель, название фирмы
Аппаратные требования.
Программная платформа.
Максимальный объем базы данных (Гб)
Максимальное количество баз данных.
Максимальное количество файлов в БД.
Типы индексируемых файлов, например: doc, txt, rtf, html, pdf, xls, zip
Наличие пробной версии и её ограничения.
Скорость индексирования Мб/мн.
Максимальный объем одного индексируемого документа
Тип поиска: по документам, по страницам, по всему текстовому полю БД.
Организация БД и выдачи: подокументная, постраничная, пофрагментная.
Средн. время поиска по запросу в БД определенного объема.
Возможный объём поискового запроса (количество слов)
Возможность установки расстояния между словами при поиске.
Использование логических операторов в запросе.
Поиск по атрибутам документов (дата, автор, размер, название и пр.)
Сортировка результатов поиска.
Наличие морфологического словаря.
Наличие дополнительных лингвистического обеспечения. Поиск по смыслу.
Возможность поиска в сетевом окружении.
Сетевая версия с многопользовательским режимом.
Диапазон цен: от цены за минимальную версию до цены за полную версию.

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

Данные о поисковых системах, с результатами тестирования.

Поисковая система Windows XP

Название системы: Служба индексирования и поиска Windows XP. Поставляется в составе операционной системы Windows XP Professional.

Фирма-разработчик: Microsoft Corporation в сотрудничестве с американской фирмой Inso (Бостон). Корпорация Inso известна разработками программ-просмотрщиков файлов различного формата (программа Quick View Plus) и доступа к данным.

Назначение системы: Система индексирования, текстов, содержащихся на компьютере в файлах почти всех известных форматов, и быстрого индексного поиска по текстовым запросам различного вида по всему проиндексированному тексту.

Отличительные черты программы: Встроена в операционную систему. Процесс индексирования осуществляется в промежутках бездействия компьютера.

Аппаратные требования: Компьютер, обеспечивающий нормальную работу Windows XP. Это значит, процессор с частотой от 1000 МГц и оперативная память от 256Мб. Объём свободного дискового пространства около половины объёма индексируемого текста.

Программные требования: ОС Windows XP Professional

Пользовательский интерфейс. Позволяет проводить два вида поиска. Один для простого контекстного поиска по текстовому запросу, другой для поиска по индексированному массиву файлов. Индексированный поиск также имеет два уровня сложности: стандартный поиск и расширенный. Расширенный поиск включает множество различных функций поиска и язык запросов, позволяющий управлять этими функциями. Язык запросов достаточно сложен для неподготовленного пользователя. Видно, что это упрощенная модификация профессионально ориентированного языка администратора баз данных, предназначенная для управления созданием, индексацией, актуализацией и настройкой параметров работы с текстовой базой данных.

Чтобы добраться до программ индексирования надо последовательно открыть окна: Панель управления / Администрирование / Управление компьютером / Службы и приложения / Служба индексирования / System Раздел “Папки” показывает каталоги на диске, которые содержат индексируемые файлы. Эти каталоги можно удалить, добавить, изменить.

Управление поиском осуществляется в разделе «Опрос каталога».

Общее впечатление:

Интерфейс громоздок и неудобен, не рассчитан на конечного пользователя. Работает неустойчиво. Часто возникают ошибки выполнения программы. Язык запросов очень неудобен и трудоемок. Например, для ввода запроса нужно напечатать:

внесен* ~ изменен* ~ утвержден* ~ материал* ~ кадастр* ~ делен* ~ округ* ~ осуществл*

где звездочка означает произвольное окончание, а знак тильды ограничение расстояния между словами.

При просмотре информации нет быстрого перехода к найденным текстам или хотя бы вывода фрагментов текстов с подсвеченными словами запроса. Для навигации по документам требуется ручная загрузка документов и поиск внутри документов.

Основные характеристики:Форматы исходных файлов: html, txt, все форматы документов Microsoft Office, а также форматы, для которых пользователем установлены дополнительные программы-фильтры. Кроме того индексируются почта и новостные сообщения.

В индекс включаются данные о текстовом содержании документов и атрибуты, типа имени автора, и данные о файле (размер, дата создания).

При поиске используются несколько типов запросов, в том числе текстовые запросы в свободной форме; запросы фраз; запросы по шаблонам, векторные запросы с использованием весовых коэффициентов.

Некоторые формы запросов рассчитаны на подготовленного пользователя, типа: {prop name=DocTitle} Contains {phrase}рыжая собака{/phrase}{/prop}

@DocTitle"рыжая собака"

Результаты тестирования: Тестирование велось на массиве документов .html формата объемом 10,0 Гб. с числом документов – 268456. .

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

Скорость индексирования и объем индекса: Время индексирования - 480 мин. что составляет скорость индексирования в 21Мб/мин. Объем индекса 2474Мб, т.е. 25% объема исходных файлов. Для меньшей базы в 500Мб объем индекса составил 40%.

Время поиска:Использовались запросы от пяти до десяти слов. Время поиска по всем видам запросов: произвольное расположение слов в документе или в виде фразы из текста, где слова расположены подряд, или с ограничением расстояния дали одинаковое время поиска по всему массиву от 3 до 5 сек. Так как все словоформы русских слов не учитываются, то количество слов, по которым проводится поиск в 2-3 раза меньше необходимого, используемого при поиске другими системами. Поэтому реальная скорость поиска раза в два меньше.

Релевантность:При поиске в небольших по размерам документах, когда все слова соединены логическим оператором «И» результаты поиска приемлемые. При поиске в больших документах необходимо устанавливать предельное расстояние между словами. В поисковой системе XP устанавливается расстояние только в 50 слов, что недостаточно и ведет к слишком большой выдаче. Наибольшая точность выдачи достигается при поиске на полное совпадение фразы. При обычном текстовом запросе на естественном языке релевантность выдачи низкая из-за отсутствия морфологического словаря. Часто выданные документы вообще не соответствуют запросу, ни по содержанию слов, ни по смыслу.

Базовое ПО (СУБД): По всей вероятности используются компоненты СУБД Microsoft SQL Server для индексирования и поиска текстовой информации.

Работа в сети: Программа может индексировать сетевые диски,. но индекс создается на локальном компьютере, на котором только и можно проводить поиск. Распределенного доступа к центральной базе данных программа не обеспечивает.

Лингвистическое обеспечение: Словоформы русского языка учитываются за счет отбрасывания окончаний при контекстном поиске. . Использование ручной обрезки слов увеличивает трудоемкость составления запроса и не всегда эффективно.Морфологический словарь только для английского языка.

Вывод: Чтобы программу могли использовать русскоязычные пользователи необходима доработка системы, прежде всего в части использование морфологического словаря русского языка и интерфейса. В существующем виде программа совершенно неудобоварима.

Cros

Название системы: «Cros"

Фирма-разработчик:

Компания "Кронос-Информ" образована в 1992 году. Основная часть сотрудников компании это коллектив бывших разработчиков информационных систем ФСБ (КГБ СССР). Основные разработки - программы для работы с полнотекстовыми базами данных и СУБД для интегрирования разнородных данных.

Назначение системы:

Классическая информационно-поисковая система с поиском по ключевым словам и стандартным дополнительным критериям. Поиск подокументно-контекстный. Работает в в персональном варианте и в локальной сети. Для работы в Интернете или Интранет сети поставляется отдельная программа. Есть версии под DOS и под Windows.

Отличительные черты программы:

Система кроме индекса хранит исходные текстовые документы в сжатом виде. Сжатие с использованием ключей позволяет защитить информацию от несанкционированного доступа.

Аппаратные требования

Процессор с частотой не ниже 300 МГц. Оперативная память от 64-до 256Мб в зависимости от ОС.

Программные требования

Windows NT, Windows 2000, Windows XP, Windows 9x, ME

Основные характеристики

Форматы исходных файлов: doc, html, rtf, lex, xls, pdf.

Многопользовательский режим. Работа в локальной сети .

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

Поиск по ключевым словам и атрибутам документов.

Максимальный размер базы данных: нет данных (из-за ограничений на пробную версию удалось создать базу только в 550Мб исходного текста).

Скорость поиска: в базе 200Мб – 1сек. (по данным журнала Chip) Скорость индексирования: объем текстов 200Мб за 750сек.. (по данным журнала Chip)

Базовое ПО (СУБД):

Самостоятельная разработка.

Лингвистическое обеспечение:

словоформы русского языка учитываются за счет отбрасывания окончаний при контекстном поиске. Морфологический словарь, не используется.

Результаты нашего тестирования:

Из-за ограничений пробного варианта на количество файлов, тестирование проводилось на базе 550Мб, содержащей 1000 файлов формата html и txt. Для совместимости с тестом Chip, тестирование проводилось на такой же машине: CPU AMD Duron 1000, 256Мб, HDD-IDE,

Время индексирования текста 550Мб составило 35 мин. Это дает 16Мб/мин, что совпадает с ркезультатами тестов украинских товарищей.

При поиске по запросу «Александр Васильевич», система нашла 40 документов меньше чем за 1сек.. Из них 14 содержали нужное сочетание. Остальные содержали «Александр» и «Васильевич», но в других сочетаниях имени и отчества. При вводе ограничения на расстояние между словами «Александр Васильевич :1» система нашла 14 документов, поиск длился 15сек. Это значит, что индексный поиск осуществляется только по отдельным словам запроса. В результате поиска выдается список документов, содержащих слова запроса, независимо от их расположения в тексте документа. При вводе дополнительного критерия поиска, учитывающего расстояние между ключевыми словами, программа переходит на контекстный поиск в найденном. Такая подокументно-контекстная схема вполне пригодна для поиска небольших документов, типа писем, приказов и пр. При поиске информации в больших базах данных, содержащих тексты книг, справочников и т.п. такой поиск слишком медленный.

Также последовательная загрузка и контекстный просмотр больших текстов затрудняют быструю навигация по всем найденным фрагментам текста.

Маркетинговая политика:

Цены на продукт фиксированы. Цена от $140 до $100 за одного пользователя. Цена на интернет – компонент, от $150.до $900. Пприобретение только по безналичному расчету, с заключением договора. Распространяется пробная версия с ограничением по числу документов в базе и функциям. Оплаченная программа привязывается к ПК пользователя регистрационным ключом.

Вывод:

Полнофункциональная информационно-поисковая система для офисной коллекции документов. Для поиска в больших полнотекстовых базах данных не предназначена. Предъявляет низкие требования к аппаратуре. Самостоятельная разработка. Коллектив с большим опытом работы. Внятная ценовая политика. Хорошие результаты сравнительных тестирований (Chip). Реальный коммерческий продукт с ценой соответствующей характеристикам.

diskMETA

Разработка украинской компании ЗАО “МЕТА” (Харьков). Фирма известна как разработчик наиболее популярной украинской поисковой системы «МЕТА». На основе поискового движка создан ряд программных продуктов: для организации поиска на сайте – siteMETA, программа для создания CD архивов с полнотекстовым поиском, CDMETA, программная библиотека разработчика МЕТА РХ.

В 2003-2004гг. фирма выпустила серию программ для создания архивов с полнотекстовым поиском на персональном компьютере. В основе всех версий лежит одно поисковое ядро, используемое во всех поисковых системах. Версии программ называются:

diskMETA-Lite бесплатная тестовая версия с ограничениями по объему базы: 2 индекса по 1000 документов. Формат исходных файлов .txt, .html, .doc. Морфологического

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

словаря нет. Словоформы учитываются "вероятностной морфологией", т.е. анализом схожести слов путем отбрасывания окончаний и пр.

diskMETA-Personal отличается только количеством индексируемых документов – 100 000, с которым уже можно реально работать. Цена $19,50.

diskMETA-Pro - Количество документов в индексе увеличено до 500 000. Добавлена возможность индексирования файлов формата .pdf, .xls и поиск в архивах chm, .zip, .rar. Цена $48,50 Может быть загружена пробная версия сроком на 1 месяц.

diskMETA-Workgroup - Количество документов в индексе увеличено до 1000 000. При поиске используется морфологический словарь русского и английского языков. В этой версии можно индексировать сетевые диски, то есть указать при индексации каталоги не только на ПК с установленной системой, но и каталоги на дисках других компьютеров локальной сети. Цена $97,50

Основные характеристики:

По данным производителя с сайта и из документации к системе.

Скорость индексирования в среднем 1Гб в час, т.е. около 17Гб/мин., что не очень быстро.

Скорость поиска приводится 1сек. По какому объему, по какому запросу, на какой машине не указывается.

Результаты тестов:

Тестировалась версия diskMETA-Pro, которая работоспособна 30 дней после скачивания.

Результаты нашего тестирования в базе 10Гб- 256000 документов на ПК с CPU AMD XP 2500:

База данных в 10 Гб была проиндексирована за 5 часов, что составляет 2Гб/час или свыше 30 Гб/мин. , хотя заявленная разработчиком скорость составляет 1Гб/час. В ходе дальнейшего тестирования неожиданно выяснилось, что система при индексировании больших файлов, учитывает только 50Кб от начала файла, а остальной текст просто игнорирует. Так как приблизительно половину объема тестовой базы данных составляли текстовые документы размером от нескольких сот килобайт до нескольких мегабайт, то diskMETA проиндексировала только половину текстового массива и скорость индексирования действительно не превышает 1Гб/час.

Тестирование на объем запроса:

длина запроса ограничивается программой до 100 символов, в которые в среднем помещается около 10 слов.

Время поиска по списку стандартных тестовых запросов №1-5 составляет: 2,4/ 7,8/ 1,7/ 2,0/ 1,1 Среднее время поиска 3.0 сек. Так как объем реально проиндексированного массива почти в два раза меньше, то цифры эти должны быть увеличены приблизительно в полтора раза. Поэтому считаем среднее время поиска равным 4.5 сек.

Релевантность выданных документов по запросам:

2 из 19 / 1 из 2000 / 2 из 84 / 0 из 8 / 2 из 14 При установке соответствующего расстояния между словами в соответствии с языком запросов: [2, Александр Васильевич] точность поиска резко возрастает: 2 из 2 / 1 из 1 / 2 из 2 / 0 из 8 / 2 из 14 /. Расстояние нужно подбирать по величине запроса.

При поиске по запросу №4 «русский полководец Александр Васильевич Суворов» документ содержащий это словосочетание вообще не находился. При проверке оказалось, что это сочетание слов находится на расстоянии 60Кб от начала файла, а система не индексирует всё что дальше 50Кб от начала файла.

Общее впечатление:

Система очевидно представляет самостоятельную разработку поискового ядра плюс программное окружение. Использованы некоторые чужие программы, например для сжатия текстов.

Интерфейс пользователя типичен для интернет-поисковика, переделанного под настольную систему. Дизайн интерфейса мог бы быть и поизящнее.

Набор функций также стандартен для интернет поисковика. При поиске по стандартной схеме с установкой расстояния и логики, выдача также стандартная. Ранжирование и сортировка выдачи по степени релевантности не впечатляют, часто находится огромное количество мусора, не относящегося к запросу.

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

Скорость индексирования невысокая, в четыре раза ниже чем у Яндекса, являющегося де-факто эталоном поисковой системы на нашем рынке.

Скорость поиска, приведенная к объему текста, достаточно высокая, приблизительно такая же как Яндекса, хотя по результатам тестов и ниже в два раза чем заявленная производителем (1 сек.).

Недостатки:

Первый и самый главный: diskMETA не является в нынешнем варианте системой полнотекстового поиска, так как она не индексирует весь текст исходного массива документов. Только при доработке программ индексирования и поиска система сможет стать серьёзным конкурентом лучшим поисковым системам.

Система имеет общий для всех интернет поисковиков недостаток: она не ищет внутри документов и предназначена для поиска документов, а не информации.

Отсутствует сетевая версия, обеспечивающая многопользовательский доступ к серверной базе данных. Это не позволяет использовать diskMETA для большинства корпоративных приложений.

Однако, учитывая смешные цены, программа вполне может быть использована для поиска в коллекциях небольших по размеру документов, размещенных на персональных компьютерах.

http://meta.ua/

http://diskmeta.com

http://sitemeta.com

ODB-Text версии 3.4

Информационно-поисковая система для корпоративного архива документов.

Производство НПЦ “ИНТЕЛТЕК ПЛЮС”. Фирма существует с 1992 года. Основу коллектива составляют выпускники МГТУ им. Н.Э. Баумана.

В основу деятельности положена разработка одной из первых в России объектной СУБД "ODB-Jupiter", которая в отличие от реляционных, была пригодна для работы со слабоструктурированной и неструктурированной информацией.

Как видно из описания системы, она развивалась в направлении автоматизации документооборота предприятия и полнотекстовый поиск по запросам на естественном языке составляет только часть её функций.

Система представляет собой полноценную сетевую программу, построенную по архитектуре клиент-сервер и состоящую из двух основных программ: клиента и сервера. В системе реализован единый сервер для делающий возможным доступ к базам данных как из локальных сетей организаций, так и через Интернет.

Программная платформа: Windows 95/98/NT/2000

Типы индексируемых файлов: .txt, doc, xls, .htm, .ppt

Офисная направленность системы видна и из набора функций:

- редактирование всех документов базы данных;
- поддержка коллективной обработки документов в локальной сети, когда несколько пользователей могут в один момент работать с одним и тем же документом.
- систематизация документов с помощью рубрикатора;
- разделение документов базы данных по типам;
- назначение набора реквизитов каждому типу документов и соответсвенно каждому экземпляру документа, что характерно для табличных баз данных.

Существует демонстрационная версия программы, правда с ограниченимя по объему хранимой информации. К программе прилагаются три тестовые базы данных, демонстрирующие основные функции системы.

База "Офис" используется для ведения архива офисных документов (нормативные акты, договора, платежные поручения, письма и др.).

База "Библиотека" поволяет создавать архив электронных изданий. Карточка реквизитов документов базы содержит основную служебную информацию об издании.

База "Кадры" позволяет вести картотеку сотрудников.

Характеристики:

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

К сожалению никаких количественных характеристик системы на сайте фирмы не представлено, скорее всего это говорит о том, что ничего выдающегося в этих характеристиках нет. Направленность системы на автоматизацию офисного документооборота также не предполагает больших объемов текстовых баз данных, высокой скорости поиска и индексирования.

Тестовые испытания демонстрационной версии не проводились из-за ограничений на объем базы данных.

Цены.

Базовая поставка, включающая сервер и 10 клиентских мест 19425 р.

Интернет-расширение + лицензия на 10 одновременных подключений 14925 р.

Услуги по обслуживанию системы стоят от 10 до 20% полной стоимости в год.

Вывод:

Несмотря на явную офисную ориентацию, ODB-Text интересна как одна из немногих полнотекстовых поисковых систем представленных на российском рынке в виде законченного коммерческого продукта. Важно также, что основой системы служит собственная объектная СУБД, с серъёзным программным окружением, и то что фирма ведет серьёзную исследовательскую работу в области поиска информации.

Адрес в интеренете: http://www.inteltec.ru/

Ищейка.

Разработка российской компании iSleuthHound Technologies.

Как сказано в рекламных материалах, предназначена для мгновенного интеллектуального поиска в текстовых массивах информации.

Программа имеет три версии:

Бесплатная ознакомительная версия, поддерживает поиск всего 500 документов трех типов.

Версия «Ищейка Проф» и «Проф DeLuxe» поддерживают свыше 10 форматов файлов. Неограниченное число документов и баз данных (зон поиска). Цена $15-$29.

Версия Ищейка Сервер 1.3. Сетевая версия. Позволяет пользователям работать с системой как с обычным браузером для доступа к документам базы данных. Кроме того эта старшая версия поддерживает 17 типов документов. Цена около $300 с лицензией на 10 пользователей..

По данным с сайта производителя: Объем базы данных - неограничен. Количество файлов – неограничено. Основные типы файлов: doc, txt, rtf, html, pdf, xls, zip, всего около 17.

Никаких численных данных о характеристиках системы не приводится.

Результаты нашего тестирования

Для тестирования использовалась демонстрационная версия программы Версия Ищейка Сервер 1.3. , имеющая ограничения на один индекс в размере 1000 документов. Тестирование производилось на массиве из 914 документов общим объемом 501Мб. В качестве тестовых документов использовались правовые документы, электронные издания специальной и художественной литературы. Использовать стандартный тестовый массив в 10Гб не удалось из-за ограничений демонстрационной версии программы.

Программа проиндексировала 501Мб текста за 18 мин. Скорость индексирования 28Мб/мин или 1.7Гб/час, что является вполне приемлемым средним показателем.

Объем индексных файлов почему-то составил 590Мб, вместо 125 предсказанных системой в начале процесса.

Скорость поиска:

При поиске по общеупотребительным словам русского языка, программа тратит от 1 до 2 сек на слово. Например: “русский язык” -2.9сек. “Варшавский конкурс фантастов” – 5.3сек.

При поиске по словам типа имен и фамилий результаты очень разные. Напроимер, слово “Ботвинник” программа искала 4сек. Сочетание “Ботвинник Михаил Моисеевич” искалось 7 сек. Было выдано 5 документов, и ни одного с нужным сочетанием, хотя в базе содержалось 6 документов с этой ссылкой. По отдельным словам эти ссылки находятся.

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

Вообще система нормально ищет по одному слову. Поиск по нескольким словам дает большинство нерелевантных документов.

Скорость поиска в несколько секунд на базе 500Мб нельзя назвать мгновенной. Она раз в десять ниже, чем например у Яндекса. Релевантность и полнота поиска также очень низкие, хотя для других типов запросов они может быть и выше. Скорее всего дело в том, что программа предназначена только для поиска в персональных и офисных коллекциях документов, а поиск в больших текстах с разнообразной лексикой для нее слишком трудное испытание.

Другие характеристики:

Объем документов – в принципе неограничен, но как показали тесты, с большими документами программа работает плохо.

Поиск ведется по документам, отдельные фрагменты текста программа на релевантность не оценивает.

Нет установки расстояния между словами, при увеличении числа слов в запросе резко падает релевантность выдачи. Поэтому программа рассчитана на запросы от одного до нескольких слов, лучше всего выполняются запросы по одному слову.

Время поиска увеличивается линейно с ростом числа слов запроса и логарифмически с ростом объема базы. Исходя из этого среднее время поиска в 10Гб по запросу из 4-5 слов будет около 1мин.

Программа работает с логическими операторами и осуществляет дополнительный поиск по атрибутам документов: дата создания, название документа, название папки.

Морфологический словарь отсутствует, но поиск по вероятностной морфологии дает неплохие результаты.

Сетевая версия Ищейки должна обеспечивать работу корпоративной интранет сети предприятия (до 10 пользователей). Пользователи имеют доступ к базе данных по типу обычного поисковика (Яндекс, Google и т.п.).

Программная платформа: ОС Windows всех версий.

Аппаратные требования минимальные: несколько мегабайт оперативной и дисковой памяти..

Вывод:

Программа предназначена прежде всего для использования как персональная поисковая система. Может использоваться для построения интранет сети небольшого предприятия. Для поиска в произвольных текстовых массивах не годится. Очень скромные требования к аппаратуре и очень низкая цена.

Адрес сайта компании: http://www.isleuthhound.com/ru/

Yndex (Яндекс, Yandex)

Название поисковой системы: Компания производит не одну поисковую систему, а целый спектр программных продуктов. Их названия состоят из названия компании + дополнительное определение типа Standart, Professional и т.п. Само название Yndex происходит от скрещивания слов (Языковый Index).

История компании «Yndex » восходит к 1990 году, когда в компании «Аркадия», возглавляемой Аркадием Борковским и Аркадием Воложем, начались разработки поискового ПО. Затем в альянсе с фирмой CompTek был разработан ряд тематических информационно-поисковых систем, затем программы поиска для интернета. Собственный поисковик «Yndex.ru» появился в 1997 году. Сейчас Yndex самая большая и популярная поисковая система русскоязычного Интернета.

Назначение системы: Программные продукты Yndex покрывают почти весь спектр по категории пользователей и по области применения. Это:

- инструмент разработчика корпоративных информационно-поисковых систем;
- поисковые системы для WEB – сайтов;
- поисковая система для всего русскоязычного Интернета;
- информационно-поисковые системы для локальных баз на CD;

В этом перечне отсутствует только версия для конечного пользователя. Все продукты, даже самые простые предназначены для профессиональной установки и настройки.

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

Основные характеристики:

Форматы исходных файлов: Зависит от версии. От только html для самой дешевой версии Standart, до XML, RTF, PDF, DOC, MP3 и форматов баз данных для версии Enterprise.

Многопользовательский режим. Клиентская часть – стандартный интеренет броузер, с которого доступен поиск в базах данных и доступ администратора баз данных.

Максимальный размер базы данных: ограничений нет.

Поиск с учетом морфологии русского языка. Поиск с учетом расстояния между словами, в том числе в пределах абзаца и на точное совпадение фразы. Возможно использование логических ореаторов. Ранжирование найденных документов по релевантности.

Результаты тестов:

Тестовые испытания проводились на пробной версии Яndex.Server Standard на машине CPU Athlon XP 2500, ОЗУ 512Мб, HDD-IDE 120Gb/8Мб.

Скорость индексирования по данным тестовых испытаний. Тестовый массив html файлов обьемом текста 10Гб был проиндексирован за 2ч.40мин, что дает скорость 70Мб/мин или 3.7Гб/час

Скорость поиска: На объеме 10 Гб при объеме запроса 5-10 слов среднее время поиска 6сек. Время поиска менялось от 0.5сек до 14сек. Для тестовых запросов №1-5, соответственно 5,5/14/4,5/0,5/1,1 сек. Среднее время 5,12 сек. Время поиска увеличивается при поиске по словам с высокой встречаемостью, имеющим множество словоформ.

Объем запроса ограничен. Максимальная длина вводимого запроса 210 символов. Но при поиске по запросу №5 из 11слов с использованием 4 логических операторов ИЛИ система ничего не нашла. При сокращении запроса на 1 слово поиск возобновился.

Базовое ПО (СУБД): Собственная разработка. Стандартное программное ядро поисковой системы работает во всех вариантах системы.

Лингвистическое обеспечение: Морфологический словарь русского языка

Варианты поисковых систем Яndex и их цена:

Наиболее простая версия Яndex.Server Standard (работает только с файлами формата HTML) распространяется как Shareware. Имеет стандартный интерфейс. При работе выдает надписи «незарегистрированная копия». Стоимость регистрации $390. Расчет на ограниченность возможностей версии, рискованность создания серьёзной системы без технической поддержки и несолидность работы WEB сервера с незарегистрированной копией.

Более продвинутая версия Яndex.Server Professional Имеет парсеры, извлекающие текстовую информацию, только для форматов html и txt. В дополнение к возможностям стандартной редакции, позволяет полностью настроить дизайн страницы с результатами поиска с использованием скриптов, написанных на Perl, C++ или XSLT, или представить эти результаты в виде XML-документа с определенной схемой. Имеет возможность реализовать "расширенный поиск" для пользователей, не знакомых с языком запросов, организовать поиск по тематическим разделам, сгруппировать найденные документы по различным признакам. Позволяет сделать тонкую настройку индексируемых зон и атрибутов в HTML-документе. цена $3900

Яndex.Server Enterprise. Основное отличие от предыдущей версии – возможность индексирования текстов из документов в формате XML, RTF, PDF, MSDOC, MP3 и различных баз данных. Имеет дополнительные возможности по поиску в нескольких базах данных и объединению результата. Представляет собой полнофункциональный инструмент разработчика крупных корпоративных хранилищ документальной информации, объединяющей текстовые данные из файлов различных форматов. Цена $15900

Яndex.Server In-The-box $21900 Этот вариант системы есть в прайс-листе, но описание её отсутствует.

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

Прайс-лист	
Яndex.Server Standard+тех. Поддержка	$390
Яndex.Server Professional	$3900
Яndex.Server Enterprise	$15900
Яndex.Server In-The-box	$21900
Яndex.Publisher	$1/CD

Яndex Publisher – программа поиска для полнотекстовых баз данных, распространяемых на CD. На CD записываются коллекция документов, готовый индекс, словари и программное обеспечение. Учитывая, что большинство CD у нас продается по цене ~100руб, треть стоимости диска за поисковую систему – это многовато, тем более, что для объемов в несколько сот мегабайт подойдут многие, не столь мощные программы.

Выводы:

По-видимому, лучшая на сегодняшний день российская программа поиска информации для полнотекстовых баз данных в Интернете и Интранет сетях. Использует технологию поиска, отработанную для поисковой машины в Рунете. Очень быстрая индексация, серьёзное лингвистическое обеспечение. Пробная версия с минимальными ограничениями.

Предназначена для разработчиков, а не для конечного пользователя. Версия Enterprise, позволяющая индексировать файлы нескольких распространенных форматов стоит свыше $15000. Отсутствие персональной поисковой системы в линейке продуктов Яндекс, можно объяснить только маркетинговой политикой. Но так как конкуренты ждать не будут, то следует ожидать появления поисковой программы Яндекс и в этой нише.

Поисковая машина МБД (MBD Search Engine)

Название поисковой системы: Поисковая машина МБД (MBD Search Engine). Новая разработка на основе экспериментальной системы, созданной ещё в 1990г. в в министерстве электронной промышленности СССР. Сокращение МБД от названия проекта (Машина Баз Данных) В 1991г. в ООО «МБД» была разработана программа для работы с базами неструктурированных текстовых и числовых данных. В 1996г. фирма прекратила свою работу и возобновила только в 2003г. под названием МБДСофт .

Назначение системы: Система предназначена для очень быстрого поиска и просмотра информации в больших базах (до сотен гигабайт), с многократной корректировкой запроса. Работает на отдельном компьютере или в локальной сети. Система ориентирована на.конечного пользователя.

Отличительные черты программы:

Система ищет не документы, а информацию, соответствующую запросу, в общем текстовом поле всей базы данных. В ответ на запрос выдаются все текстовые страницы, содержащие найденную информацию, независимо от количества найденных страниц на документ. Поэтому система может использоваться для быстрого индексного поиска в коллекциях больших документов, типа энциклопедий, справочников, книжных архивах.

Система может работать с поисковыми запросами большого объема, до нескольких сотен слов, как на естественном языке, так и с логическими операторами.

Система кроме индекса хранит исходные текстовые документы в сжатом виде. База данных имеет постраничную (книжную) организацию, что психологически привычно для пользователя и позволяет осуществлять быстрый просмотр найденных фрагментов в страничном контексте.

При создании базы данных создается полный словарь системы, доступный пользователю в режиме поиска.

Аппаратные требования: Процессор с частотой от 1000 МГц. Оперативная память от 256Мб.

Программные требования: Windows 2000, Windows XP.

Форматы исходных файлов: doc, html, rtf, txt.

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

Работа на отдельном ПК и в локальной сети. Клиент – серверная архитектура. Многопользовательский режим.

Поиск только индексный по всем текстам базы с учетом расстояния между словами и логичекими операторами. Удобная установка расстояния с помощью движка.

Большой объем запроса до нескольких сот слов, с логическими операторами и установкой расстояния.

Программа ищет внутри документов, оценивая на релевантность запросу каждый фрагмент текста, а не весь документ в целом. В результате поиска выводятся страницы документов с фрагментами текста, релевантными запросу. Слова входящие в запрос выделены другим цветом.

Быстрое перелистывание найденных страниц, в пределах документа и переход между документами.

Сортировки результатов поиска по релевантности нет. По мнению разработчиков, система рассчитана на быстрый поиск с интерактивной корректировкой запроса по результатам последовательных поисков и полный просмотр релевантной выдачи без мусора.

Возможен вывод оригиналов документов в формате родительской программы.

Максимальный размер базы данных: (неограничен). Для версии Standard - 100Гб.

Число баз данных: неограничено.

Максимальный и минимальный размер индексируемых документов: неограничен.

Наличие пробной версии: Версия «MBD Search Engine Standard» может быть скачана с сайта фирмы как пробная. Специальных ограничений (кроме работы в сети) версия не имеет.

Базовое ПО (СУБД): Самостоятельная разработка.

Лингвистическое обеспечение: используется базовый морфологический словарь русского языка объемом около 1 миллиона слов.

Результаты тестов:

Тестовые испытания версии Standard проводились на машине с CPU Athlon XP 2500, ОЗУ 512Мб, HDD-IDE 120Gb/8Мб.

Скорость индексирования по данным тестовых испытаний. Тестовый массив html файлов обьемом текста 10Гб был проиндексирован за 4ч.10мин, что дает скорость 40Мб/мин или 2.4 Гб/час

Скорость поиска: На тестовой БД объемом 10Гб, время поиска для тестовых запросов №1-5, соответственно 0,86/1,53/1,41/0,53/0,9 сек. Время поиска менялось от 0.5сек до 1.53 сек. Среднее время 1.05 сек.

При вводе в окно поиска целой страницы текста, содержащей около 200 слов, программа успешно переварила её и нашла страницу за 35 сек.

Время поиска увеличивается при поиске по словам с высокой встречаемостью..

Цены на продукт фиксированы. Цена $100 за одно рабочее место, независимо сервер или клиент. Сетевая версия поставляется с лицензией на 5 или более рабочих мест. Пользователи с лицензией получают техническую поддержку и новые версии системы бесплатно.

Вывод: Система с очень быстрыми поиском и просмотром. Ищет информацию внутри документов. Работает с большими базами данных. Имеет постраничную (книжную) организацию базы данных. Позволяет создавать полный словарь всех индексируемых документов и использовать его при поиске. Имеет простой, дружественный интерфейс, предназначенный для конечного пользователя. Предназначена для поиска в произвольных текстовых массивах, в том числе для персонального и корпоративного применения.

Недостатки: достаточно высокие требования к аппаратному и программному обеспечению. Для поиска документов в офисных коллекциях желательно дополнить систему поиском по атрибутам и сортировкой результатов. Мал набор типов индексируемых файлов, необходимо добавить хотя бы .xls и .pdf.

Адрес в интернете: http://www.mbdsoft.ru

Russian Context Server (Россия). RCO - Гарант-парк

Название поисковой системы: Russian Context Server.

Фирма-разработчик: компания "Гарант-Парк-Интернет"

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

Торговая марка: RCO – Russian Context Optimizers. Основные продукты: системы поиска информации, программы для лингвистической и аналитической обработки документов в уже существующих базах данных и информационно-поисковых системах, в основном на базе СУБД Oracle и MS SQL. Собственная разработка информационно-поисковой системы.

Назначение системы: Поисковая система Russian Context Server это инструмент разработчика, поисковая машина, предназначенная для поиска на WWW-сервере, в полнотекстовых базах данных, файловых архивах. Основной заказчик система информационно-правовая система “Гарант”.

Отличительные черты программы:

Система кроме индекса хранит исходные текстовые документы во внутреннем HTML формате. Для разных типов данных используются разные виды индексов. Для поиска по атрибутам табличная структура, для поиска по тексту инвертированный индекс. Мощное лингвистическое обеспечение.

Программные требования

Платформа Windows NT.

Основные характеристики:

- Форматы исходных файлов: html. Для, импорта данных из документов других форматов требуются дополнительные программы (поставщики).
- Многопользовательский режим. Архитектура клиент-сервер.
- Поиск по ключевым словам и атрибутам документов.
- Поиск с ранжированием документов по релевантности.
- Максимальный размер базы данных: нет данных
- Скорость поиска: (По данным взятым с сайта RCO) На объеме 10 Гб при запросе 1атрибут и 1 слово время поиска 5-10сек. Даннае приведены для старой машины с CPU-PII266, RAM-128Мб. При выполнении нашего теста уменьшение времени поиска за счет большей производительности ПК примерно компенсируется увеличением времени за счет увеличения объема запроса до 8-10 слов. Скорость поиска зависит не только от производительности процессора, но и от скорости считывания данных с диска. А этот показатель растет медленнее чем производительность CPU. Поэтому тестовый показатель должен быть порядка 10 сек.
- Скорость индексирования для той же машины 100Мбайт в час = 2 МБайт/мин. В пересчете на производительность современного ПК она составит не более 10Мб/мин, что очень мало и объясняется по-видимому сложностью используемых аогоритмов анализа данных при индексировании.

Базовое ПО (СУБД): Разработка на основе языков программирования реляционных баз данных.

Лингвистическое обеспечение:

- словарь словоизменения (более 100 тыс. слов русского языка)
- словарь моделей управления предикатов русского языка
- тезаурус общей лексики (прежде всего синонимы и обобщающие понятия)
- специальные словари и правила.

Используется морфологический анализ при поиске английских слов.

Пробная версия: как указано на сайте, в пробном варианте доступна только версия для Oracle. Остальные только по специальному договору с фирмой.

Ценовая политика: Цены на продукт определена в зависимости от версии системы. Версия определяется установленными ограничениями, в основном на число документов в базе.

Наименование версии.	Цена за 1 лиценз.	Цена годовой поддержки	**Ограничения**
Standard	$600	$200	поиск по 1 серверу до 5000 документов
Advance	$1500	$200	число серверов не ограничено, до 5000

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

			документов на сервере
Enterprise	$3000	$200	число серверов и документов не ограничено

Вывод: Архитектура системы характерна для реляционных СУБД. Применительно к текстовому поиску это несколько утяжеляет систему и ухудшает характеристики. Интерфейс требует профессиональной настройки. Система предназначена для разработчиков корпоративных систем поиска информации, прежде всего полнотекстовых баз данных в Интернете. Для использования в качестве персональной системы поиска информации конечным пользователем не предназначена.

Основное направление деятельности компании это теория и практика компьютерного распознавания смысла текста – наиболее сложное и перспективное направление развития информационно-поисковых систем. В этой облпсти RCO можно считать лидером. Специализация на программах интеллектуальной обработки текста определила ориентацию на использование для поиска информации промышленных СУБД других фирм, прежде всего СУБД Oracle и MS SQL Server.

Следопыт

Название системы: “Следопыт 3.0"

Фирма-разработчик: ЗАО “Медиа Лингва”

Компания **ЗАО "МедиаЛингва** - российский разработчик программного обеспечения, в области лингвистических, поисковых и мультимедиа технологий. Основана в августе 1995 г.

Назначение системы:

Информационно-поисковая система для полнотекстового поиска на персональном компьютере или в локальной сети предприятия. Запрос на естественном языке или с использованием логических операторы. с поиском по ключевым словам и стандартным дополнительным критериям.

Отличительные черты программы:

В качестве базовой СУБД программа использует внешнюю СУБД - Microsoft SQL Server 7.0. или Microsoft SQL Server 2000. Эта СУБД должна быть установлена на компьютере пользователя, или устанавливается одновременно с установкой всей поисковой системы. Поэтому основные технические характеристики системы ограничены возможностями универсальной СУБД фирмы Microsoft.

Программная система поставляется в трёх вариантах: “корпоративном”, “профессиональном” и “персональном”. Варианты отличаются набором программного окружения базовой СУБД, которое и является разработкой ЗАО “Медиа Лингва”.

Минимальный “персональный вариант” позволяет проводить индексирование файлов на жестком диске ПК, компакт-дисках и других съемных носителях. Индексатор запускается по команде пользователя или работает в фоновом режиме при помощи автоиндексатора .

Форматы индексируемых исходных файлов: doc, html, rtf, txt,xls,ppt. Профессиональная и корпоративная версии также индексируют файлы pdf, архивные файлы и почтовые сообщения. Индексируется текстовое содержание и некоторые атрибуты документов.

Лингвистическое обеспечение: в персональной версии используется программа отбрасывания окончаний, а в старших версиях используется морфологический словарь русского языка.

Поисковый запрос может задаваться в виде фразы на естественном языке, допускается формулирование запроса с одновременным использованием русских и английских слов. Есть также возможность поиска по ключевому слову, без учета словоформ.

Возможен поиск с использованием формального языка запросов с применением логических операторов "И", "ИЛИ", "НЕ". Логический поиск доступен , только в профессиональной и корпоративной версии. Поиск ведется по ключевым словам и некоторым атрибутам документов.

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

Корпоративная версия системы может работать в локальной сети, обеспечивая многопользовательский режим доступа к данным. Поиск в сетевом окружении или клиент-серверная архитектура обеспечивающая многопользовательский доступ.

Результаты тестирования этой системы приведены по данным журнала Chip: Сравнительный тест восьми персональных поисковых систем Автор: Олег Пилипенко. Журнал "ЧИП"(Украина) №3, 2003г.)

Адрес: http://www.it2b.ru/print2.view4.page1.html

Тестирование проводилось на базе объемом в 200Мб текста, что соответствует объему офисной коллекции документов небольшой фирмы. Максимальный размер базы данных нигде не указан.:

Скорость индексирования: 200Мб текста было проиндексировано за 240сек., что составляет 50Мб/мин. и является очень неплохим показателем. С какой скоростью система сможет проиндексировать 10Гб текста и сможет ли вообще, сказать трудно, так как никаких данных на этот счет найти не удалось..

Скорость поиска: в базе 200Мб поиск длился около 4сек., что совпадает с данными приведенными на сайте разработчика – время поиска 1-5сек., правда не уточнено в каком объеме и по какому запросу. В любом случае это очень низкая скорость, в десятки раз уступающая таким системам как Яндекс. Очевидно, что это плата за использование универсальной СУБД, имеющей не самые лучшие показатели при полнотекстовом поиске информации.

В качестве системных требований на сайте разработчика указаны:

Установленные на компьютере пользователя Microsoft SQL Server v.7.0 или 2000 Standard или Enterprise Edition с установленной системой полнотекстового поиска (Full-Text Search).

Техническая поддержка включает консультации по программе по телефону, e-mail или ICQ.

Цены:

Следопыт 3.0 - Персональный (коробка) – цена 10 у.е. Символическая цена за символическую программу. Наши пользователи привыкли работать с нелицензинными копиями самых лучших программ, поэтому вряд ли кто станет работать со столь слабой версией, да ещё платить за это деньги.

Следопыт 3.0 - Профессиональный (коробка) - 40 у.е.. Эта версия включает практически все функции системы, кроме работы в локальной сети. Цена вполне соответствует набору предлагаемых функций и характеристикам системы.

Следопыт 3.0 - Корпоративный (коробка) с лицензией на 6 рабочих мест – 750 у.е.

Следопыт 3.0 - Корпоративный ПЛЮС с лицензией на неограниченное количество рабочих мест. -1250 у.е. Реальные покупатели программных разработок на российском рынке – это большие и маленькие фирмы, эксплуатирующие программы для поиска документов в многопользовательском режиме в локальной сети. На этих покупателей и рассчитана основная версия поисковой системы Следопыт.

Выводы

Система имеет базовый джентльменский набор функций информационно-поисковой системы, предназначенной для поиска в офисной коллекции документов, включая поиск по атрибутам, морфологический словарь, запрос с логическими операторами и пр. Использование стандартной СУБД (MS SQL Server) ухудшает скоростные характеристики системы. Для быстрого поиска информации в произвольных текстовых массивах система не предназначена. Для поиска документов на ПК или в сети небольшой фирмы, вполне подходит. Цена нормальная. Немаловажное достоинство то, что система существует и продается как законченный коммерческий продукт: в коробке и по определенной цене.

Адрес в интернете: http://www.sledopyt.ru/

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

Данные о некоторых поисковых системах по опубликованным материалам.

К этому разделу относятся прежде всего поисковые системы, не представленные на российском рынке в виде готового коммерческого продукта, а также не имеющие демонстрационной версии, пригодной для тестирования.

В дальнейшем, по мере возможности, описание этих систем будет дополняться, в том числе и результатами тестирования.

AltaVista Search Engine 3.0.

Продукция компании AltaVista (Compaq). Использует ту же технологию, что и глобальная поисковая система AltaVista. В России представлена компанией РБК (РосБизнесКонсалтинг) СОФТ, разработавшей лингвистическое обеспечение для русской версии.. Поставляется в комплекте SDK (набор программ для разработчиков). Стоимость лицензиии от $31500 до $220000.

Адрес: http://www.rbcinfosystems.ru

Информационно поисковая система Convera (Excalibur) Retrieval Ware.

Информационно поисковая система Convera (Excalibur) Retrieval Ware появилась на рынке в 1998 году. Использует нечеткий поиск. Продукция компании Convera Technologies Corp (США). (ранее Excalibur Technologies). В России представлена компаней «Весть-МетаТехнология»

Адрес: http://www.vest-meta.ru

MnoGoSearch.

Популярная, свободно распространяемая (FreeWare) поисковая система. Последняя версия mnoGoSearch 3.2.7. Её анализ представляет интерес для понимания того, чем отличается бесплатно распространяемая программа, от программы за $31000 ?

Адрес: http://www. mnogosearch.org

ИПС Артефакт.

Разработка компании Интегрум-Техно. Отечественная разработка, начало которой восходит к 70-м годам. Индексный поиск, морфологический анализ, логические операторы при поиске. Известна DOS версия начала 90-х годов. Сейчас ИПС используется для онлайнового доступа к крупнейшему в стране банку текстовой ирформации объемом до 800Гб. На сайте фирмы есть анонс ИПС как коммерческого продукта пятилетней давности. Для конечного пользователя система очевидно не предназначена. Цены только договорные.

Адрес сайта: http://www.integrum.ru

Галактика ZOOM.

Информационно поисковая система с возможностью анализа текстовых данных. Продукт компании Галактика, специализирующейся на программных системах управления бизнесом. Сведения об ИПС даны только самые общие, без характеристик. Пробной версии нет. Заказная система по договорным ценам.

Адрес в интернете: http://www.galaktika.ru

Электронный архив Евфрат.

Производства российской компании Cognitive Technologies. Евфрат – серия программ для организации электронного документооборота на предприятии. Имеет функции ирформационно-поисковой системы. Предназначен для офисных коллекций документов.

Адрес в интернете: http://www.evfrat.ru

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

CCT Archive

Программа полнотекстового поиска. Программы используют оригинальную технологию поиска информации "НЕЗАБУДКА", разработанную в Институте радиотехники и электроники РАН, основанную на “записи информации на траектории нелинейной динамической системы”. Предназначена для электронных архивов и библиотечных ИПС.

Адрес в интернете: http://www.controlchaostech.com.

1С

Разработки фирмы 1С (Россия). Программные оболочки для информационных правовых систем. Отдельно не поставляется, но используется для юридических баз данных, распространяемых на CD.

Адрес в интернете: http:/www.1c.ru

СУБД MS SQL Server 2000.

Для разработчиков баз данных. Начиная с версии 7.0 компания Microsoft включила в комплект поставки MS SQL Server специальную компоненту - систему полнотекстового поиска по базе данных, позволяющую пользователю находить нужные записи по разнообразным условиям. В стандартную поставку MS SQL Server 2000 входит комплект лингвистических модулей для многих языков за исключением русского.

Oracle Text.

Система является компонентой последних версий известной СУБД Oracle. Компонента обеспечивает полнотекстовый поиск по всем документам, включенным в базу данных. Предназначена для разработчиков. Поддержки русского языка система не имеет, что вызвало появление российских разработок лингвистического обеспечения для этой СУБД (например RCO).

ABBY Retrieval & Morphology 4.0 Engine.

Фирма Abby (Россия). Средство разработки ИПС в виде набора библиотек (DLL) программных.. Инструментарий разработчика предназначен для встраивания функций индексирования, полнотекстового поиска и морфологического анализа текстовых данных на 34 языках во внешние локальные и клиент-серверные приложения через 32-битный COM API интерфейс. Инструментарий поддерживает языки программирования C++, Visual Basic и Delphi. Цена 1580 у.е. (3 лицензии разработчика).

Адрес: http://www.abbyy.ru

Megaputer

Российская компания Мегапьютер (Megaputer) предлагает системы поиска для локальных и корпоративных хранилищ на основе СУБД Oracle и программ анализа числовых и текстовых данных собственного производства.

Адрес в интернете: http://www.megaputer.ru

ИРБИС

Система автоматизации библиотек ИРБИС. Разработана в ГПНТБ России. В системе реализованы все типовые библиотечные технологии, включая технологии комплектования, систематизации, каталогизации, читательского поиска, книговыдачи и администрирования. Поисковые функции входят в АРМ читателя. Цена - за полный комплект несколько тысяч долларов.

Адрес в интернете: http://www.elnit.ru/

Google Desktop Search (GDE)

Бесплатная локальная версия известной поисковой системы Google, одноименной американской компании. Бета версия персональной поисковой системы появилась в октябре 2004г. Так как поисковые возможности Google широко разрекламированы, то характеристики этой системы представляют интерес для российских пользователей.

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com

К сожалению, как сам Google Desktop Search, так и ряд других бесплатных зарубежных поисковых систем пока малопригодны для текстовых массивов на русском языке. Они не работают с русской морфологией, плохо индексируют русскоязычные текстовые массивы.

Предельный объем текста в поисковом индексе также на два порядка ниже, чем у поисковых систем высшего класса, которым в основном и посвящен этот обзор. Например, для GDE максимальный объем индексируемого текста 2Гб. Для других систем он и того ниже. Поэтому тестирование этих систем ничего не даст для сравнения с

Ознакомиться с общими характеристиками этих систем можно по нескольким обзорам, напечатанным в журнале Компьютерра и доступным в Интернете.

1. «Кто не спрятался, я не виноват» Краткий обзор зарубежных бесплатных персональных поисковых систем. Автор Александр Прокудин. Опубликовано: 17.11.2004 Еженедельник «Компьютерра» http://www.computerra.ru/offline/2004/567/36689/

2. «Четыре Геркулеса». Обзор характеристик четырех поисковых бесплатных систем: Google, BinkX, Copernic, FileHand Search 2.0. Автор: Владимир Гуриев Опубликовано в журнале "Компьютерра" №41 от 2 ноября 2004 года. http://www..computerra.ru/ offline/2004/565/36526/

Выводы:

Накопление больших массивов информации в интернете, локальных сетях, автономных компьютерах, вызвало вызвало бурный рост нового класса программных средств – программ поиска информации в полнотекстовых базах данных.

Существующие технологии поиска недостаточно эффективны, чтобы найти в большом количестве разнородной информации глобальной сети, сведения отвечающие запросу. Поэтому развивается процесс двухэтапного поиска: первый - тематический поиск и отбор данных в полнотекстовою базу данных, второй - поиск в полнотекстовой базе данных. Можно прогнозировать рост спроса на такие базы, от небольших тематических объемом с один CD, до больших, объемом сотни гигабайт и даже терабайт, вмещающих содержание крупных библиотек.

На рынке сейчас представлено множество разношерстных и разнокалиберных программ поиска. Их функциональные возможности находятся в диапазоне от простого контекстного поиска до интернетовских поисковых машин и систем поиска для крупных библиотек . Диапазон цен: от бесплатных программ до систем стоимостью в сотни тысяч долларов.

В основном программы делятся на три основные категории: программы для поиска в персональной коллекции документов, программы для создания корпоративных хранилищ документов, программы поиска информации в произвольных текстовых массивах.

На основе анализа характеристик существующих поисковых систем и динамики их развития можно сделать вывод, что доминирующее место на рынке постепенно занимают поисковые системы, имеющие мощное поисковое ядро, обеспечивающее очень высокую скорость обработки неограниченных по размеру текстовых массивов. Производители ПО, имеющие в своем распоряжении такое программное ядро, могут разработать, и разрабатывают, для него программное окружениие, перекрывающее все функции персональных, корпоративных и универсальных поисковых систем. Вопрос ценовой конкуренции с менее производительными системами решается путем выпуска ряда различных по цене версий, на основе одного программного ядра, но с ограничениями по функциям.

P.S.

К сожалению я не смог на данный момент охватить все имеющиеся поисковые системы, достойные упоминания. По многим системам даны только самые краткие сведения, возможна и ошибочная информация.

Все замечания и дополнения просьба присылать по адресу: zakhar@mbdsoft.ru Валерию Захарченко. Последние версии материала можно также прочитать на сайте http://www.mbdsoft.ru.

PDF created with FinePrint pdfFactory trial version http://www.fineprint.com