

Каждый из нас не раз в своей жизни сталкивался с проблемой поиска информации. Как правило, используются самые разные источники. Интернет, файлы на диске, базы данных или информационные системы.

Агентство «Поиск»

Условно программы и системы поиска можно разделить на три группы. Это глобальные поисковые интернет-системы (Yahoo, Google, «Яндекс»), готовые решения для бизнеса (корпоратив-

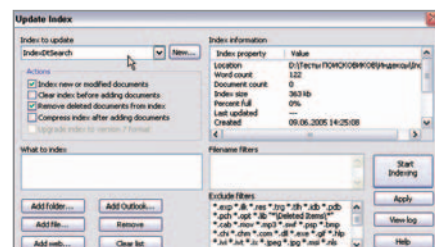
ные технологии поиска и обработки информации) плюс их локальные (десктопные) версии и простой файловый поиск на компьютере. Различные направления подразумевают различное исполнение.



Технологии поиска

С файловым поиском на персональном компьютере все просто. Он не отличается особыми возможностями и какими-либо изысками и осуществляется обычным перебором файлов на предмет соответствия, а скорость и результат напрямую зависят от того, что мы вводим в строке запроса.

Совсем иначе дело обстоит с поисковыми системами, работающими в Глобальной сети. Тут уже не до перебора. Гигантские объемы (тот же «Яндекс» может похвастаться индексацией более чем 11 Тбайт данных), специфика организации информации и глобальная хаотичность сделают простой поиск не только неэффективным, но также длительным и трудоемким. Без знания скрытых воз-



↑ Создание нового индекса

GOOGLE

Персонализация поиска

Google, вслед за революционными технологиями, перевернувшими Интернет с ног на голову, бесплатной электронной почтой с размером ящика 2,5 Гбайт и прочими успехами, готовит пользователям новый сюрприз. Уже не первый год ведется работа по сбору и анализу информации, поступающей от пользователей. Это личные и официальные данные, интересы и увлечения, хобби и история поисковых запросов, наконец. Все это делается для того, чтобы как можно больше персонализировать поиск. То есть научить машину работать таким образом, чтобы она максимально точно удовлетворяла запросы каждого конкретного пользователя. Перспектива вполне заманчивая — после того как каждого пользователя обмерят и взвесят, он сможет получить в распоряжение «личный Интернет» от Google.

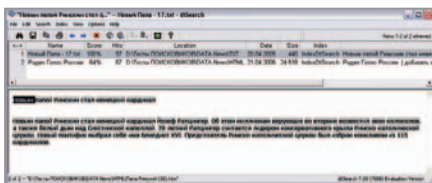
Именно с теми результатами запросов, которые наиболее точно подойдут для этого пользователя. Конечно, скрытых камней не избежать. Сбор информации о пользователях, например, чтение почты, — это нарушения тайны переписки. Плюс ко всему после продажи акций Google вся информация, на которой строится персонализированная система результатов, может попасть «не в те руки». Да и просто осознавать, что кто-то думает за тебя, не так уж и хорошо. Да, я никогда не набирал в строке поиска Google фразу «рыбий жир». Но вдруг мне захочется найти полную информацию об этой полезной и противной субстанции? С учетом моих личных данных и истории запросов, пока я смогу найти с помощью Google нужную мне страницу, интерес к рыбьему жиру рискует пропасть напрочь.

с поиском по сети в пару сотен компьютеров, так и среди десятка мегабайт данных на одном домашнем ПК, — одни и те же.

Первым делом любая программа создает область поиска (проводится индексация выбранных пользователем файлов на дис-

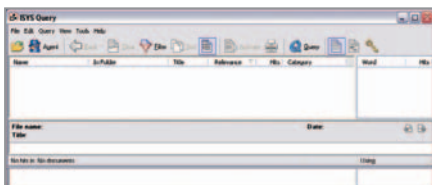


↑ «Ищайка» споткнулась: по ключевому слову «папа» первый найденный документ содержит вхождение «папок»



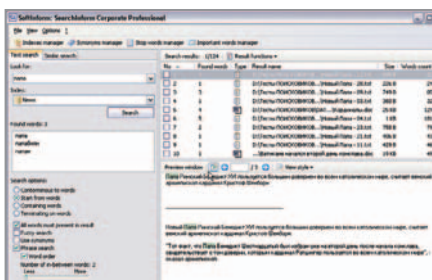
↑ dtSearch: использование в качестве запроса целого фрагмента текста

можностей данных систем найти нужный документ по запросу — дело не одной минуты, а иногда и не одного часа. Проблема в том,



↑ Основное меню ISYS: Desktop

что такая реализация поиска очень зависит от фразы или формулировки запроса пользователем. Чем запрос менее конкрети-



↑ Результаты SearchInform. Одно ключевое слово запроса. Все в порядке. А кто бы сомневался?

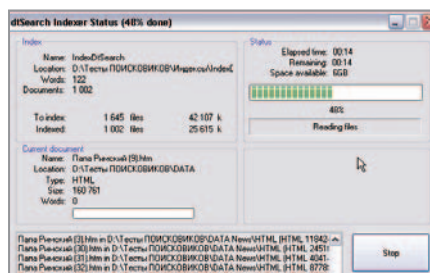
зирован, тем хуже результаты поиска. Но технология работает. Пусть не идеально и не всегда так, как нам хочется.

Третьими значатся готовые решения на базе поисковых технологий и их домашние аналоги. Схема работы таких систем проста: все тот же поиск по фразе (с учетом всех словоформ, синонимов и так далее). Ключевым моментом является скорость индексации информации и обработки запроса, а для корпоративных систем — удобство работы с сетевыми индексами и возможность подключения различных источников данных.



Принцип поиска

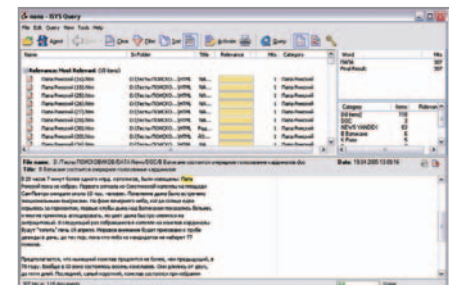
Понятно, что на данный момент в мире поиска наблюдается доминирование именно



↑ Индексация файлов на диске

«фразовых» поисковиков. Практически все «ищайки» могут похвастаться наличием десктоп-варианта своих систем. Технологии, которые применяются как в случае

ке, электронной почты, полей баз данных и так далее — в зависимости от возможностей той или иной системы). В дальнейшем именно созданный индекс (организованная структура, в которой содержится информация об обработанных данных) используется для работы — быстрого получения списка нужных документов согласно запросу. Принципы индексирования доку-



↑ С релевантностью у ISYS: Desktop все в порядке. Правда, почему-то только для документов html. А где же doc и txt?

ментов и хранения проиндексированных данных в разных системах, как и способы их «извлечения» и последующий вывод в результирующем списке, понятно, разные.



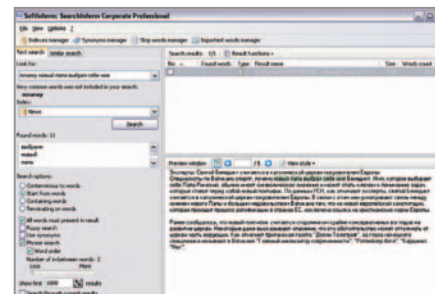
Диспозиция

Чтобы выяснить, какая же система поиска информации на данный момент самая быстрая и толковая, были выбраны несколько популярных программ от российских и западных разработчиков и с ними прове-

ли несколько тестов на скорость индексации документов, быстроту и качество поиска. Заметим, что в тесте участвовали «домашние» версии этих программ.

Нами была собрана обширная библиотека литературы разной направленности, начиная с художественной и заканчивая новостными лентами (11 Гбайт текстов в таких форматах, как **txt**, **html** и **doc**). Первым заданием была индексация всего этого многообразия документов, далее мы провели тесты на скорость поиска и

на релевантность результатов путем сравнения полученных списков по одинаковому запросу. В качестве запроса использовались фразы и отрывки из новостей, содержащихся в подборке документов. Тестовой площадкой был выбран компьютер со следующими характеристиками: процессор **Athlon 2,2 MHz**, объем оперативной памяти **1 Гбайт**, IDE-винчестер **Seagate** на **160 Гбайт** со скоростью вращения **7 200 оборотов в минуту**, операционная система **Windows XP**.



↑ Программа SearchInform нашла документ по фразе, состоящей из шести слов, без каких-либо проблем

Перечисленные ниже программы объединяют следующие преимущества: достаточно быстрый полнотекстовый поиск информации как на английском, так и русском (в некоторых случаях с учетом морфологии языка), привычный и удобный интерфейс и относительно небольшой объем индекса. Представленные версии программ обеспечивают приблизительно одинаковый уровень сервиса.

■ **Ищайка Проф Deluxe 4.5** — полноценная поисковая система от **iSleuthHound Technologies** для быстрого поиска документов и файлов на жестких дисках.

■ **ISYS: Desktop 7.0** — быстрый и удобный инструмент для поиска информации от компании **ISYS Search Software** (как на персональном компьютере, так и среди тысяч документов в локальной сети).

■ **dtSearch 7.0** — программы поиска документов различных форматов и электронных сообщений. Один из мировых лидеров быстрого и качественного поиска.

■ **SearchInform Desktop Edition 1.0.20** — программа полнотекстового поиска документов на персональном компьютере. Совмещает в себе возможности фразового поиска с учетом морфологии и словаря синонимов и новой технологии поиска похожих документов.

Ищайка Проф Deluxe

Версия ▶ 4.5

Разработчик ▶ iSleuthHound Technologies

Сайт ▶ www.isleuthhound.com

Объем дистрибутива ▶ 5,68 Мбайт

ОС ▶ Windows

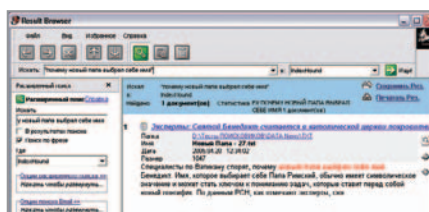
Интерфейс ▶ русскоязычный

Условия распространения ▶ shareware

Цена ▶ \$29,95

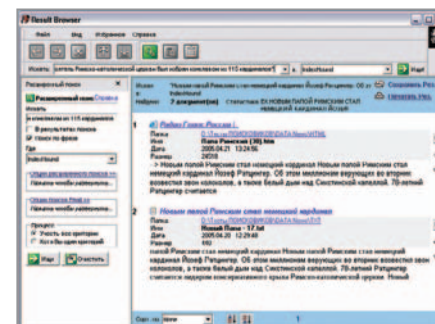
Программа, разработанная компанией **iSleuthHound Technologies**, справилась с заданием по перевариванию 11 Гбайт текстов ни много ни мало за 21 час 12 минут. Размер индекса при этом составил те

Аутсайдер



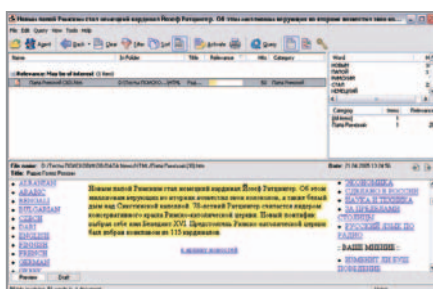
↑ Шесть слов в запросе не проблема

самые 11 Гбайт, полученные на входе. Должны заметить, что у **Ищайки** оказались самые незавидные показатели по сравнению с остальными программами, подвергнутыми испытанию.



↑ Ищем фрагмент текста. После полуминуты работы «Ищайка» выдала результат, отвечающий требованиям запроса

Продукт от знатоков



↑ У программы явно нелады с русским языком. Но с поиском документов по целому абзацу текста она все же справилась

Следующей в списке значится программа **ISYS: Desktop 7.0**. Это мощный инструмент поиска данных, разработанный компанией, которая занимается системами поиска без малого 17 лет, поэтому полученный результат не удивляет. С обработкой 11 Гбайт документов **ISYS: Desktop** управилась за 2 часа 53 минуты, создав свой индекс размером чуть более 4 Гбайт. Неплохой результат и приемлемый размер, правда, лучше всего программа работает с форматом **html**, а вот от **doc** и **txt** почему-то

ISYS: Desktop

Версия ▶ 7.0

Разработчик ▶ ISYS Search Software

Сайт ▶ www.isys-search.com

Объем дистрибутива ▶ 39,2 Мбайт

ОС ▶ Windows

Интерфейс ▶ англоязычный

Условия распространения ▶ shareware

Цена ▶ неизвестна

забывает. Во время обработки компьютер не только не зависал намертво, но и оставлял небольшую часть системных ресурсов для более-менее удобной работы.

dtSearch

Версия ► 7.0

Разработчик ► dtSearch Corp.

Сайт ► www.dtsearch.com

Объем дистрибутива ► 23,1 Мбайт

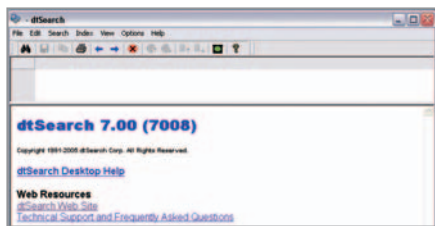
ОС ► Windows

Интерфейс ► англоязычный

Условия распространения ► shareware

Цена ► \$199

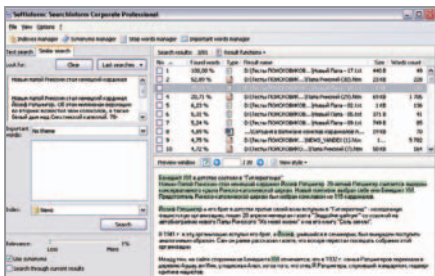
Скорость индексирования и размер созданного индекса очередной версии программы от известнейшей компании



Основной экран программы dtSearch

Победитель

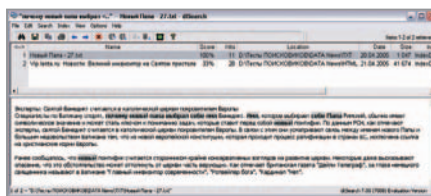
Последней в списке, но отнюдь не последней по производительности, значится система **SearchInform Desktop Edition** от компании **СофтИнформ**. Лидер



Уникальная функция поиска документов, похожих по содержанию. Полезно и позволяет сразу же найти все документы, схожие по смыслу с текстом запроса

гонки обработал тестовые документы за 1 час 51 минуту, ограничившись всего лишь 2 Гбайт при создании индекса. Если учесть небольшую, но очень полезную функцию **SearchInform — Smart Indexing**, которая позволяет в режиме реального времени отслеживать нагрузку на процессор компьюте-

Качественно



dtSearch находит файлы и по фразе с несколькими ключевыми словами

dtSearch Corporation тоже оказались на должном уровне. 11 Гбайт материала **dtSearch 7.0** «пробежала» за 2 часа 57 минут, создав для личных целей поиска индекс размером 4,15 Гбайт.

К несомненному достоинству этой поисковой машины можно смело отнести ее интерфейс. Он настолько удобен и прост, что при работе с этим мощным инструментом у вас не должно возникнуть никаких проблем.

SearchInform Desktop Edition

Версия ► 1.0.20

Разработчик ► SoftInform Search Technology

Сайт ► www.searchinform.com/site/ru/index.htm

Объем дистрибутива ► 3,5 Мбайт

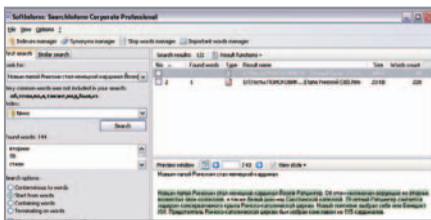
ОС ► Windows

Интерфейс ► русско-, англоязычный

Условия распространения ► shareware

Цена ► 1 500 руб.

ра и регулировать уровень потребления программой системных ресурсов при индексации, то пальма первенства разработке **СофтИнформ** досталась более чем по праву.



Никаких проблем с поиском документов по целому абзацу. И с форматами, в отличие от ISYS, программа дружит

СПЕЦИМВОЛЫ «ЯНДЕКС»

Упрощаем поиск

Технологическая начинка и рекламное слово маркетологов «Яндекс» вывели эту систему на первое место среди русских интернет-поисковиков. Опрос по умолчанию русскоязычных серверов и ориентировка в первую очередь на русскоязычные запросы (реализована поддержка морфологии и шести кириллических кодировок) делают «Яндекс» серьезным конкурентом Google в секторе Рунета.

Но «Яндекс», как и любой другой поисковый сервер, конечно, далек от совершенства. Обычно при задании запроса (при простом поиске) системой отбираются документы, содержащие все набранные в окне поиска слова, причем все они должны находиться в пределах одного предложения, иначе документ не соответствует запросу.

Ряд простейших приемов ввода специальных символов и ключевых слов в поисковую строку значительно оптимизирует поиск. Например, если между словами поставить запятые, «Яндекс» выдаст лишь те документы, которые содержат хотя бы одно из введенных слов. Тильда (~) позволяет искать документы, содержащие в абзаце первое слово, но не содержащие второго. Если же требуется осуществить поиск не в первом абзаце, а по всему документу, необходимо просто удвоить соответствующий знак. Одинарные знаки-операторы «&» и «~» ищут в пределах абзаца, двойные «&&» или «~~», соответственно, во всем документе.

Описывать все операторы, поддерживаемые поисковой машиной «Яндекс», не имеет смысла (притом что получить описание языка поисковых запросов или руководство по поиску в любой системе можно, как правило, выбрав гиперссылку, например, «Помощь» или «Справка», на главной странице системы). Однако стоит заметить, что, овладев нехитрыми приемами поиска, можно значительно сократить время работы в Интернете. Необходимо лишь помнить, что разные поисковые системы используют различные друг от друга инструменты оптимизации поиска.

Подводя итог, можно с уверенностью говорить о том, что будущее за технологиями полнотекстового поиска похожих по содержанию и по смыслу документов — как в сфере Интернета, так и в корпоративном секторе. Неограниченные возможности развития, адекватность результатов и скорость об-

работки любого по объему запроса делают данные технологии намного более удобными и востребованными.

Единственное, что может остановить пользователя, — высокая цена на подобные продукты. Но хорошая вещь всегда стоит своих денег.

Андрей Белоокий