**Михаил Абрамзон**

# Яндекс для всех

Санкт-Петербург
«БХВ-Петербург»
2007

УДК 681.3.06
ББК 32.973.26-018.2
А16

**Абрамзон М. Г.**

А16 Яндекс для всех. — СПб.: БХВ-Петербург, 2007. — 544 с.: ил. + CD-ROM

ISBN 978-5-9775-0144-6

Рассказывается о ведущем поисковом ресурсе российской части Интернета — Яндексе. Описаны его сервисы для поиска документов и новостей, блогов и адресов, товаров и музыкальных произведений. Рассмотрены почтовый сервис и сервис для создания и размещения сайтов на Народ.ру. Читатель узнает о том, что такое "электронные деньги" и как с их помощью оплатить товар. Большое внимание уделяется общедоступным поисковым программам, которые Яндекс предлагает своим посетителям для поиска информации не только на сайтах, но и на своем компьютере. На компакт-диске находятся программы Яндекса, описанные в книге, и дополнительные материалы.

*Для начинающих пользователей Интернета*

УДК 681.3.06
ББК 32.973.26-018.2

**Группа подготовки издания:**

| | |
|---|---|
| Главный редактор | *Екатерина Кондукова* |
| Зам. главного редактора | *Игорь Шишигин* |
| Зав. редакцией | *Григорий Добин* |
| Компьютерная верстка | *Натальи Смирновой* |
| Корректор | *Татьяна Кошелева* |
| Дизайн обложки | *Инны Тачиной* |
| Зав. производством | *Николай Тверских* |

Лицензия ИД № 02429 от 24.07.00. Подписано в печать 31.08.07.
Формат 70×100 1/16. Печать офсетная. Усл. печ. л. 43,86.
Тираж 2500 экз. Заказ №
"БХВ-Петербург", 194354, Санкт-Петербург, ул. Есенина, 5Б.

Санитарно-эпидемиологическое заключение на продукцию № 77.99.02.953.Д.006421.11.04
от 11.11.2004 г. выдано Федеральной службой по надзору
в сфере защиты прав потребителей и благополучия человека.

Отпечатано с готовых диапозитивов
в ГУП "Типография "Наука"
199034, Санкт-Петербург, 9 линия, 12

ISBN 978-5-9775-0144-6

# Оглавление

**Аркадий Волож —**
основатель и генеральный директор компании "Яндекс".

Является соучредителем и членом Совета директоров компании Infinet Wireless (производитель оборудования WiMAX в России). Был соучредителем компании CompTek International, одного из крупнейших дистрибьюторов сетевого и телекоммуникационного оборудования в России, и руководил этой компанией с 1989 по 2000 год. Принимал активное участие в процессе дерегулирования частот для беспроводных сетей, легализации IP-телефонии. У Аркадия высшее образование в области прикладной математики.

# Введение

## Так начинался "Яндекс"

В 2000 году акционерами CompTek — компании, создавшей и в течение долгого времени развивавшей проект Яndex, была учреждена компания "Яндекс". Компания ru-Net Holdings инвестировала 5 млн 280 тыс. долларов и получила в новой компании долю в 35,72%. В число акционеров вошли также менеджмент и ведущие разработчики поисковой системы. Генеральным директором стал Аркадий Волож. Но история Яндекса началась задолго до этого события.

Десятью годами ранее в компании "Аркадия" начались разработки поискового программного обеспечения. Через два года были созданы две информационно-поисковые системы — Международная классификация изобретений, а также Классификатор товаров и услуг. Системы работали под DOS и позволяли проводить поиск, выбирая слова из заданного словаря с использованием стандартных логических операторов.

Еще через год "Аркадия" стала одним из подразделений компании CompTek и в течение двух последующих лет выполняла работы по усовершенствованию поисковых технологий. В сотрудничестве с лабораторией Ю. Д. Апресяна (Институт проблем передачи информации РАН) был разработан словарь, обеспечивающий поиск с учетом морфологии русского языка. Теперь пользователи могли задавать в запросе любые формы слов.

Эти разработки позволили создать "Библейский компьютерный справочник", справочник стандартов "Информ — Норматив", электронные научные издания "А. С. Грибоедов", "Пушкин. Электронный фонд русской классической литературы", словарь языка Грибоедова.

Следующим шагом стала разработка алгоритма построения гипотез, после чего морфологический разбор перестал быть привязанным к словарю — если какого-либо слова в словаре нет, то находятся наиболее похожие на него словарные слова, и по ним строится модель словоизменения.

Летом 1996 года руководство компании CompTek и разработчики поисковой системы пришли к выводу, что развитие самой технологии важнее и интереснее, чем создание прикладных продуктов на базе поиска. Исследования рынка показали своевременность и большие перспективы поисковых технологий.

Первая демонстрация продуктов серии Яndex (Яndex.Site, Яndex.Dict) была проведена на выставке Netcom'96 18 октября 1996 года. Первый из них обеспечивает поиск по своему сайту и установлен на сотнях серверов Рунета. Второй продукт, морфологическое расширение запроса, до сих пор используется для передачи запросов на AltaVista.

А уже 21 ноября была выполнена первая установка системы Яndex.Site на веб-сервере Издательского дома "Открытые системы". Это дало возможность искать информацию с учетом морфологии русского языка. И в том же ноябре всем пользователям была предоставлена бесплатная возможность русифицированного поиска необходимой информации с учетом морфологии русского языка с помощью поискового сервера AltaVista.

Еще через полгода появился Яndex.CD — поиск документов на CD-ROM, а затем Яndex.Lib — полнофункциональная библиотека Яndex для встраивания в различные приложения и базы данных. И наконец осенью 1997 года был открыт Yandex.Ru.

Основными нововведениями поисковой системы Yandex.Ru были проверка уникальности документа — этим достигалось исключение копий в различных кодировках, и отличительные свойства поискового алгоритма Yandex: поиск с учетом морфологии русского языка, с учетом расстояния и тщательно разработанный алгоритм оценки релевантности.

Оптимизация поискового алгоритма позволила реализовать проблему поиска по разным зонам текста, ограничение поиска на группу сайтов, поиск по ссылкам и изображениям. Также, впервые в Рунете, было введено понятие *индекс цитирования* — количество сайтов, ссылающихся на данный ресурс. Затем был открыт "Семейный Яндекс" с фильтрацией результатов поиска от мата и порнографии.

Название Яndex появилось в то время, когда будущий генеральный директор будущей компании Аркадий Волож и будущий директор по технологиям компании Илья Сегалович разрабатывали технологию поиска неструктурированной информации с учетом морфологии русского языка. Требовалось слово, отражающее суть новой технологии, которое бы хорошо звучало, легко писалось и запоминалось. Тогда на основе английского слова index был предложен вариант — yet another indexer ("еще один индексатор" или Языковой иНдекс). Кроме этого, Аркадий предложил букву "Я" в названии — специфически русскую — русской и оставить, для наглядности. Так появилось слово "Яndex".

# Поиск, почта и все остальное

Сегодня Яндекс — это не только поиск. И поиск не только среди документов. Яндекс теперь вполне может быть назван порталом, предоставляющим посетителям разнообразные услуги.

## Поиск в Интернете


**Рис. В.1.** Слоган, который знают все!

В год, когда была образована компания "Яндекс", на канале НТВ прошла рекламная компания поисковой системы, во время которой был впервые озвучен слоган "Найдется все!" (рис. В.1). Сегодня этот лозунг знает каждый, кто использует Яндекс для поиска информации. В канун 2007 года на домашней странице портала выводилось сообщение — "Поиск по 1 372 783 513 веб-страницам". Это, конечно, далеко не весь Рунет, но его значительная часть.

Домашних страниц поискового сервера у Яндекса несколько. Это главная страница портала **http://www.yandex.ru**, через которую можно выполнить обычный и расширенный поиск. Аскетичный поиск **Ya.ru**, где на домашней

странице нет ничего кроме поисковой строки. Есть также семейный поиск, поиск для слабовидящих и медиативный.

Кроме того, пользователь может настроить формат представления результатов поиска, а также вид домашней страницы Яндекса.

Сам поиск может вестись по нескольким направлениям:

- по веб-страницам;
- по новостям;
- картинок;
- товаров и услуг;
- в блогах;
- контактов фирм и организаций;
- легальной музыки;
- на картах.

## Словари и энциклопедии

Поиск по словарям — тоже поиск, но более конкретный. Его результаты основываются на статьях словарей и энциклопедий, которые включены в состав электронной справочной литературы Яндекса. На конец 2006 года поиск мог проводиться среди 29 словарей. А кроме того, здесь же можно выполнить перевод слов с/на английский, немецкий, французский, испанский, итальянский и, конечно, русский язык.

## Почта

У вас еще нет почтового ящика на Яндексе? Подумайте, не настала ли пора его получить. Размер ящика не ограничен, ограничен лишь размер одного письма — он не может превышать 10 Мбайт. Почта проверяется на спам и на вирусы. Для борьбы со спамом (а это страшная вещь — бывают дни, когда из сотен пришедших писем действительно нужных остается не более десятка) применяется разработанная Яндексом программа фильтрации спама и массовых рассылок "Спамооборона".

Почтовая система предоставляет и множество других "вкусностей". Это и импорт адресных книг из почтовых клиентов (кому хочется каждый раз заполнять адресную книгу заново?), и сбор почты с других серверов, и доступ к

своему почтовому ящику по защищенному каналу. Все это и многое другое мы с вами рассмотрим в *главе 2*, посвященной этому сервису.

## Яндекс-каталог

Каталог Яндекса появился позже поисковой системы. С одной стороны, это еще один сервис, привлекающий посетителей. С другой — дополнительная возможность организации поиска по отобранным модераторами каталога ресурсам. В дополнение к обычной рубрикации по темам (Бизнес, Дом, Развлечения, Отдых и пр.) Яндекс предлагает классификацию сайтов по типу содержащейся в них информации (Справки, Товары и услуги, Публикации и пр.). Несколько позже появилась рубрикация и по регионам.

Если большинство российских, да и не только российских, каталогов отталкивались в своем развитии от Yahoo!, то в Яндексе была разработана собственная система рубрикации. И хотя количество рубрик в каталоге относительно небольшое, дополнительные признаки, которые проставляются для каждого сайта, позволяют перейти к нужной группе ссылок за минимальное количество щелчков. А для ранжирования ссылок в рубриках используется тематический индекс цитирования (тИЦ).

## Яндекс и Народ

Одним из пользующихся популярностью сервисов, предоставляемых Яндексом, стал сервис бесплатного размещения сайтов. Причем это не просто хостинг, где пользователи могли разместить собственные сайты. Свое название "Народ" сервис подтвердил еще и тем, что помимо хостинга предложил воспользоваться набором подготовленных шаблонов, позволяющих создать страницу пользователю, даже ничего не понимающему в вопросах программирования веб-страниц. После регистрации, пользователь мог выбрать один из шаблонов (сейчас их свыше ста), наполнить его своей информацией и пустить "в плавание" по широким просторам Интернета.

Этим сервисом воспользовались многие, а в некоторых учебных заведениях его применяют даже в процессе обучения. Создаются здесь персональные страницы и визитки предприятий, фотоальбомы и резюме, сайты увлеченных людей и интернет-магазины. Многие сайты были включены в каталог Яндекса, а это не такая простая задача. По данным тематического индекса цитирования составляется выборка ТОП100 народных сайтов.

## Яндекс.Деньги

Яндекс.Деньги — это платежная система, с помощью которой вы можете:

- совершать платежи в Интернете;
- совершенно безопасно хранить информацию о зачислениях и платежах;
- управлять своими средствами через Интернет.

Эта система — не банк, в ней не открывается счет пользователя системы. Только кошелек — и пополнив его любым способом, можно оплачивать свои покупки в интернет-магазинах, передавать свои средства другим пользователям этой системы или получать переводы от них, через специальные обменные системы переводить или получать электронные деньги из других подобных систем, например, WebMoney. Средства, находящиеся в вашем кошельке, могут быть перечислены на ваш счет в любом банке, находящемся на территории России.

Система поддерживает два типа кошельков — Яндекс.Кошелек, доступ к которому осуществляется через сайт Яндекс.Денег, и Интернет.Кошелек, для работы с которым на компьютер пользователя устанавливается специальная программа. Кошельки между собой несовместимы и действуют полностью самостоятельно. Поэтому каждый может завести себе два различных кошелька и пользоваться ими независимо.

Несмотря на то что Яндекс.Деньги — не банк, все средства, находящиеся в этой системе, обеспечены реальными банковскими счетами компании-оператора, размещенными в следующих банках:

- ИМПЭКСБАНК;
- Банк "ТАВРИЧЕСКИЙ";
- Внешторгбанк;
- Росбанк;
- Сбербанк.

## Ленты новостей

Уже давно новости можно читать не только на сайтах, где они публикуются, но и подключившись к RSS-потокам. RSS — формат представления данных (международный стандарт для синдикации веб-контента). Аналогичные функции выполняет и формат Atom, но он имеет расширенные по сравнению с RSS характеристики.

Многие блоги (сетевые дневники) и многие новостные источники предоставляют информацию в формате RSS. Эти потоки состоят из сообщений, где каж-

дое сообщение является записью в дневнике или новостью. Яндекс.Лента — специальный сервис для сбора таких информационных RSS-потоков в одном удобном для использования месте.

Из огромного списка возможных источников вы выбираете блоги (сетевые дневники) или новости, которые хотите читать, собираете из них ленту и читаете. В процессе чтения вы можете отмечать понравившиеся сообщения, чтобы потом просмотреть их отдельно. Нужные сообщения вы также сможете найти и с помощью поиска по вашей ленте.

Каждая лента в сервисе представляет собой набор RSS-потоков, сообщения из которых сортируются по времени поступления. Вы можете создать себе несколько лент (например, по тематикам) и наполнить их интересными лично вам потоками с помощью формы добавления потока.

## Решения для корпоративных пользователей

По-настоящему богаты те, кто может себе позволить делиться с другими. Если исходить из этого, Яндекс — богатая компания. То, что было разработано для себя и является основой бизнеса, предлагается всем желающим. Как на платной основе, так и на бесплатной.

Корпоративным клиентам предлагаются два продукта — Спамооборона и Яндекс.Сервер.

Корпоративный продукт "Спамооборона" — это серверное решение для фильтрации спама. Основные свойства системы: полнота и высокая точность фильтрации, актуальная база знаний о спаме, наличие гибких настроек. Установив ее на корпоративном почтовом сервере, вы резко снизите количество спама, доходящего до почтовых ящиков ваших сотрудников.

Хотите, чтобы на вашем портале было легко найти любую информацию — установите Яндекс.Сервер. Большинство возможностей этого продукта теперь доступно в бесплатной версии, более чем достаточной для большинства интернет-проектов.

## Персональные продукты

Персональный поиск Яндекса — это программа на вашем компьютере, осуществляющая поиск по файлам и письмам с учетом морфологии русского

языка. Совершенно бесплатная, обладающая прекрасными поисковыми возможностями. Позволяет во много раз быстрее искать, к примеру, в базах почтовой программы The Bat! письма, чем выполнять поиск стандартными средствами почтовика.

Яндекс.Бар — это уже совсем иной продукт. Удобство его использования почувствует в первую очередь тот, кто много и часто пользуется Яндексом. А все потому, что в этот плагин, работающий и на MS IE, и на FireFox, включено большинство служб Яндекса, а также обеспечен доступ к вашим личным ресурсам (почте, ленте, денежным средствам).

Есть на Яндексе и другие сервисы и службы — игры и общение, соревнования по поиску и фотоальбомы. Сервисы постоянно развиваются, а количество их увеличивается. Но "нельзя объять необъятное", говорил незабвенный Козьма Прутков. И не отвлекаясь на новинки, разберем, чем же является Яндекс сегодня.

**Илья Сегалович** —
директор "Яндекса" по технологиям и разработке,
один из основателей компании.

Поисковыми технологиями Илья начал заниматься в 1990 году — в компании "Аркадия", где руководил группой программного обеспечения. В период с 1993 по 2000 год, Илья работал в компании CompTek International, где возглавлял отдел поисковых систем. При непосредственной поддержке Ильи Сегаловича созданы Национальный корпус русского языка (Ruscorpora) и Российский семинар по оценке методов информационного поиска (РОМИП). Илья Сегалович имеет высшее образование в области геофизики. Вместе со своей женой Марией Илья поддерживает благотворительную студию "Дети Марии" (социальная помощь детям-сиротам и детям-инвалидам).

---

## Глава 1

# Поиск (Найдется всё!)

*Главная задача информационно-поисковой системы — это поиск информации, релевантной информационным потребностям пользователя. Слово релевантность означает соответствие между желаемой и действительно получаемой информацией. Релевантность можно еще представить как меру близости между реально полученными документами и тем, что следовало бы получить из системы.*

*"CITForum: Поисковые системы"*

## 1.1. Что такое "поисковая машина"

Каждому из нас в определенный момент времени бывает необходима информация, отсутствующая среди записей, заметок и данных на нашем компьютере. Где в таком случае вы будете ее искать? Одним из наиболее простых и удобных способов поиска является Интернет (далее также "Сеть"). Здесь есть все — техническая и экономическая информация, справочники и научные издания, расписания транспорта и онлайновые магазины, книги и курсы валют. Все можно найти, не отрываясь от стула. Но у этой хорошей стороны

Интернета есть и обратная сторона — количество информации в Сети растет даже не по часам, а по минутам и секундам. Найти нужную информацию обычным *серфингом* уже невозможно. Простой и удобный протокол HTTP, используемый для серфинга, удобен для навигации и просмотра страниц, но совершенно не предназначен для поиска.

Первым шагом на пути систематизации информации, размещаемой в Интернете, стало создание *каталогов* сайтов, в которых ссылки на ресурсы группировались по тематическому признаку. Так построено большинство современных каталогов, но началом всему стал проект Yahoo!, открытый в 1994 году. Вторым шагом после создания каталога стал поиск по размещенным в нем ссылкам. Понятно, что это был поиск не по всем ресурсам Интернета, а лишь по тем, которые присутствовали в каталоге. Даже сегодня, спустя десятилетия после появления первых каталогов, в них присутствует лишь малая толика интернет-ресурсов. В одном из самых крупных каталогов — DMOZ (Open Directory Project) находятся ссылки на 4 миллиона сайтов, распределенных по 590 000 категорий, а в базе Яндекса размещена информация свыше чем о 2 278 900 000 документов. Показатели для поиска несравнимые.

Поэтому не удивительно, что почти одновременно с появлением каталогов, появились и *поисковые машины*. Первой из них стал проект WebCrawler, появившийся в 1994 году. Следом за ним открылись поисковые системы Lycos и AltaVista, а в 1997 году Сергей Брин и Ларри Пейдж создали Google. В том же году была официально анонсирована и поисковая система Яндекс, ставшая самой популярной в русскоязычной части Интернета.

## 1.1.1. Компоненты поисковых машин

Информация в Сети не только пополняется, но и постоянно изменяется, но об этих изменениях никто никому не сообщает. Отсутствует единая система занесения информации, одновременно доступная для всех пользователей Интернета. Поэтому с целью структурирования информации, предоставления пользователям удобных средств поиска данных и были созданы поисковые машины.

Поисковые системы бывают разных видов. Одни из них выполняют поиск информации на основе того, что в них заложили люди. Это могут быть каталоги, куда сведения о сайтах, их краткое описание либо обзоры заносят редакторы. Поиск в них ведется среди этих описаний.

Вторые собирают информацию в Сети, используя специальные программы. Это поисковые машины, состоящие, как правило, из трех основных компонентов:

- Агента;
- Индекса;
- Поискового механизма.

**Агент**, или более привычно — паук, робот (в англоязычной литературе — spider, crawler), в поисках информации обходит сеть или ее определенную часть. Этот робот хранит список адресов (URL), которые он может посетить и проиндексировать, с определенной для каждой поисковой машины периодичностью скачивает соответствующие ссылкам документы и анализирует их. Полученное содержимое страниц сохраняется роботом в более компактном виде и передается в Индекс. Если при анализе страницы (документа) будет обнаружена новая ссылка, робот добавит ее в свой список. Поэтому любой документ или сайт, на который есть ссылки, может быть найден роботом. И наоборот, если на сайт или любую его часть нет никаких внешних ссылок, робот может его не найти.

Робот — это не просто сборщик информации. Он обладает довольно развитым "интеллектом". Роботы могут искать сайты определенной тематики, формировать списки сайтов, отсортированных по посещаемости, извлекать и обрабатывать информацию из существующих баз данных, могут выполнять переходы по ссылкам различной глубины вложенности. Но в любом случае, всю найденную информацию они передают базе данных (Индексу) поисковой машины.

Поисковые роботы бывают различных типов:

- *Spider* (паук) — это программа, которая скачивает веб-страницы тем же способом, что и браузер пользователя. Отличие состоит в том, что браузер отображает информацию, содержащуюся на странице (текстовую, графическую и т. д.), паук же не имеет никаких визуальных компонентов и работает напрямую с HTML-текстом страницы (аналогично тому, что вы увидите, если включите просмотр HTML-кода в вашем браузере).
- *Crawler* (краулер, "путешествующий" паук) — выделяет все ссылки, присутствующие на странице. Его задача — определить, куда дальше должен идти паук, основываясь на ссылках или исходя из заранее заданного списка адресов. Краулер, следуя по найденным ссылкам, осуществляет поиск новых документов, еще неизвестных поисковой системе.

- *Индексатор* разбирает страницу на составные части и анализирует их. Выделяются и анализируются различные элементы страницы, такие как текст, заголовки, структурные и стилевые особенности, специальные служебные HTML-теги и т. д.

**Индекс** — это та часть поисковой машины, в которой осуществляется поиск информации. Индекс содержит все данные, которые были переданы ему роботами, поэтому размер индекса может достигать сотен гигабайт. Практически, в индексе находятся копии всех посещенных роботами страниц. В случае если робот обнаружил изменение на уже проиндексированной им странице, он передает в Индекс обновленную информацию. Она должна замещать имеющуюся, но в ряде случаев в Индексе появляется не только новая, но остается и старая страница.

**Поисковый механизм** — это тот самый интерфейс, с помощью которого посетитель взаимодействует с Индексом. Через интерфейс пользователи вводят свои запросы и получают ответы, а владельцы сайтов регистрируют их (и эта регистрация — еще один способ донести до робота адрес своего сайта). При обработке запроса поисковый механизм выполняет отбор соответствующих ему страниц и документов среди многих миллионов проиндексированных ресурсов и выстраивает их в порядке важности или соответствия запросу.

Названные выше компоненты не обязательно входят в состав поисковой машины так, как они здесь описаны. У разных поисковиков реализация может отличаться друг от друга. К примеру, связка Spider+Crawler+Индексатор может быть выполнена в виде единой программы, которая скачивает известные веб-страницы, анализирует их и ищет по ссылкам новые ресурсы.

## 1.1.2. Характеристики поисковых машин

В статье, посвященной поисковой машине Rambler (**http://www.rambler.ru/doc/architecture.shtml**), называются основные характеристики, которые могут быть применены к любым поисковикам:

- полнота;
- точность;
- актуальность;
- скорость;
- наглядность.

*Полнота* поиска характеризуется отношением количества найденных по запросу документов к общему количеству документов в Интернете, соответствующих данному запросу. Если по запросу "кристаллическая решетка" будет найдено 150 документов, а общее количество документов в Интернете, соответствующее этому запросу, составляет 1000, то полнота поиска составит 0,15. (Эта величина приблизительная, поскольку неизвестно точно, сколько же на самом деле существует в Интернете страниц, отвечающих условию поиска.) Чем более полно проанализированы и занесены в Индекс документы, тем выше будет показатель полноты поиска.

*Точность* поиска определяется как степень соответствия найденных документов запросу пользователя. Допустим, мы хотим найти документы, в которых встречается выражение "сын знахаря". В результатах поиска мы увидим документы, в которых встречается точно такое выражение. Но присутствуют и документы, содержащие искомые слова, но не выражения, например: "родители привозят *сына* в небольшой городок на Адриатическом побережье, к местному *знахарю*". И если всего найдено 200 документов, из которых только в 80 встречается именно искомое словосочетание, то точность поиска будет оценена как 80/200 (0,4). Чем точнее поиск, тем выше вероятность, что пользователь найдет нужные документы, тем меньше будет избыточной, лишней информации.

Для повышения точности результата в различных поисковых системах применяются различные способы. Каждый поисковик использует свои решения, в целом предназначенные для выполнения близких по сути задач. К примеру, вот что по этому поводу сказано на сайте Рамблера:

> Повышение точности в поисковой машине Рамблер достигается за счет использования различных технологий на всех этапах обработки и поиска информации. Одним из наиболее интересных процессов является распознавание грамматических омонимов. *Омонимы* — это слова, которые имеют одинаковое написание, но различный смысл. Различают лексические и грамматические омонимы. Лексические омонимы относятся к одной части речи, как, например, существительное "бор": хвойный лес, стальное сверло и химический элемент. Грамматические омонимы относятся к разным частям речи, поэтому по написанию у них обычно совпадают только отдельные формы. Примерами грамматических омонимов могут служить слова "печь" (существительное русская *печь* и глагол *печь* пирожки) и "рядовой" (прилагательное *рядовой* сотрудник и существительное *рядовой* Иванов).
>
> Омонимы не только увеличивают размер индексной базы (так как для каждого такого слова приходится хранить все его возможные значения), но и отрицательно сказываются на точности поиска. Если пользователь ищет слово "данные", ему неинтересно получить в найденном все документы, которые содержат слово "дать". Для того чтобы результаты поиска были точ-

нее, модуль синтаксического анализа проводит разбор окружения слов-омонимов с целью установления их наиболее вероятных значений. Например, если рядом со словом "печь" стоит существительное ("пирожки", "картошка"), то с высокой вероятностью "печь" в данном контексте является глаголом. На сегодняшний день анализатор способен распознавать значения только грамматических омонимов.

Синтаксический анализ позволяет также с определенной вероятностью распознавать некоторые имена собственные. Например, если в тексте несколько слов подряд написано с большой буквы, они чаще всего представляют собой имя собственное (Петр Петрович, Московский Государственный Университет). Данные о таких конструкциях учитываются при индексации и обработке запроса.

Еще один способ повышения точности поиска — это выделение устойчивых обозначений и поиск их как отдельных лексических единиц. На сегодняшний день в Рамблере реализована система распознавания таких конструкций, как, например C++, б/у, п/п-к. Если по запросу C++ поднимать все тексты, в которых присутствуют латинская буква C, а также знак +, то получится огромное количество документов, далеко не все из которых соответствуют запросу; кроме того, это большая работа, значительно увеличивающая время поиска.

Источник: "Принципы работы поисковой машины Рамблер" (**http://www.rambler.ru/doc/architecture.shtml**).

А вот что на эту же тему пару лет назад сказал И. Сегалович, директор Яндекса по технологиям и разработке:

Алгоритм поиска учитывает социальную структуру Интернета. Он умеет отличать мнение людей от технической, вспомогательной и рекламной информации, то есть лучше распознавать, какой ресурс является авторитетным в своей области. Также введена дополнительная очистка результатов поиска от дубликатов. Теперь пользователь избавлен от повторения в списке найденного почти одинаковой информации. Поиск в Интернете — это серьезная наука, поэтому для повышения качества сервиса в Яндексе проводятся регулярные исследования. В прошлом году мы организовали отдел *асессоров* — пользователей, которые систематически по заданной методике оценивают релевантность результатов. Обратная связь от асессоров дает нам возможность настраивать параметры алгоритма ранжирования и увеличивать точность поиска. Стало удобнее работать с региональной информацией. Теперь Яндекс автоматически определяет, в каком городе находится компьютер, с которого поступил запрос, и, если уточнение по региону имеет смысл, предлагает повторить поиск, ограничив его сайтами данного региона. Поиск поддерживает шесть языков — к русскому и английскому добавились украинский, белорусский, французский и немецкий. Язык документов и сайтов определяется автоматически, а ограничить об-

ласть поиска нужным языком можно в настройках или расширенном поиске. Расширенный поиск стал проще и функциональней, заданные с его помощью ограничения теперь видны на странице найденных результатов. Благодаря "умной подсказке" пользователи расширенного поиска смогут увидеть сформированный запрос, как если бы он был задан на русском языке.

Какова психология того, кто ищет информацию? Считается, что наиболее подходящие (релевантные) документы должны быть на первой-второй страницах результатов поиска. Если количество полученных результатов больше, человек вряд ли будет просматривать остальные страницы. И даже если в числе найденных есть документ, полностью отвечающий заданным условиям, но находится он на странице из второго десятка, ищущий этот документ не увидит — он просто не дойдет до этой страницы. Поэтому громадное значение приобретает и *ранжирование* документов в результатах поиска по их релевантности запросу.

По поводу релевантности Яндекс говорит, что это "соответствие ответа вопросу", но при этом важны две составляющие — полнота (ничто не забыто) и точность (отсутствие лишнего).

Релевантность различают как содержательную и формальную. Воспользовавшись словарями, представленными в Яндексе, предложу определения этих понятий:

- *содержательная* релевантность — соответствие документа информационному запросу, определяемое неформальным путем;
- *формальная* релевантность — соответствие, определяемое алгоритмически путем сравнения поискового предписания и поискового образа документа на основании применяемого в информационно-поисковой системе критерия выдачи.

В простейшем случае, релевантность текста определенному запросу — это процент вхождения запроса к общему объему текста. Для поисковых систем высокорелевантным текстом считается такой, где вхождение запроса в текст примерно равно 4–7% — меньшего может не хватить, большее чревато тем, что система сочтет текст за поисковый спам и наложит на страницу некий понижающий фильтр или может вообще убрать страницу из результатов выдачи по искомому запросу.

Конечно, каждая поисковая система использует гораздо более сложные способы вычисления релевантности документов запросу пользователя. Тем не менее, несмотря на то что алгоритмы у всех поисковых машин разные, они построены на общих принципах — основные отличия результатов выдачи

заключаются не в алгоритмах определения релевантности, а в конкретных способах реализации этих алгоритмов.

Какие же факторы, помимо вхождения слов запроса в текст документа, оказывают дополнительное влияние на его место среди других документов? Каждая поисковая машина, стремясь привлечь качеством выдачи запрашиваемой информации, разрабатывает собственные критерии подсчета релевантности. Это и плотность ключевых слов на странице, и разделы страниц, где находятся эти слова, объем содержания, тексты заголовков и ссылок и многое другое. Учитываются и такие рассчитываемые показатели сайтов, как индекс цитирования, тематический индекс цитирования, Page Rank. И при этом происходит постоянное изменение степени влияния на результаты тех или иных параметров, их состав и принцип расчета.

## 1.2. Как устроена поисковая машина Яndex

Поисковая машина Яndex относится ко второму рассмотренному ранее типу поисковых машин. У Яndexa есть свои пауки-агенты, есть свой Индекс и поисковый механизм. Эта поисковая машина ориентирована в первую очередь на российскую часть всемирного Интернета, т. е. индексируются в ней русскоязычные сайты, расположенные в доменах **ru** и **su**. Сделаны небольшие исключения для наиболее авторитетных зарубежных сайтов. Сложнее с русскоязычными сайтами, которые зарегистрированы в международных или региональных (государственных) доменах других стран — **com**, **org**, **de**, **us** и других, но они все же попадают в Индекс и учитываются при поиске.

Большинство значимых зарубежных нерусскоязычных сайтов может быть найдено по ссылке, при этом, в отличие от русскоязычных сайтов, в Индекс они не попадают. Упрощается ситуация в том случае, когда у компаний, таких как BMW, IBM и многих других, появляются русскоязычные версии сайтов, без проблем индексируемые Яндексом.

Поисковая машина — самый востребованный ресурс Яндекса. Ежедневно его посещают около четырех с половиной миллионов посетителей, при этом количество просмотренных поисковых страниц приближается к сорока миллионам. При этом пользователи, выполняющие на нем поиск, этого не замечают — складывается впечатление, что Яндекс работает индивидуально для каждого из них.

Так, при запросе средней "тяжести", то есть при поиске не очень частотного слова, время отклика системы (без учета времени передачи данных по каналу от поисковой системы к пользовательскому компьютеру) исчисляется десятыми долями секунды.

В условиях постоянного роста количества пользователей и их запросов главной задачей поисковой машины является сохранение приемлемых с точки зрения пользователей скорости и полноты выполнения запросов. Эта задача решается несколькими способами, каждый из которых является необходимым, но не достаточным в отрыве от других. Способы достижения высоких результатов на сегодня применяются следующие:

- оптимизация базовых поисковых алгоритмов и архитектуры поиска;
- регулярное увеличение мощностей вычислительных ресурсов поисковой системы;
- использование архитектурной возможности масштабирования системы.

Оптимизация поисковых алгоритмов проводится постоянно. Результаты таких работ вводятся в действие до двух раз в год. Сказывается их внедрение на уменьшении нагрузки на поисковую машину (в год эта величина составляет 20–30%), а также на уменьшении времени отклика.

Увеличение мощности — это постоянное обновление используемого оборудования. Сюда входит и переход на более мощные процессоры, увеличение оперативной памяти, увеличение объемов дисковых хранилищ. Способ хотя и необходимый, но крайне затратный. Результативность выполненного апгрейда можно косвенно оценить увеличением объема поисковой базы, находящейся в его распоряжении.

Третий способ — использование мастшабируемости системы. В двух словах суть его заключается в том, что каждый уровень системы распараллеливается на несколько одинаковых узлов. Например, при наличии десяти поисковых серверов, обрабатывающих поступающие запросы, очередной запрос будет направляться на тот из них, которых в данный момент времени будет свободен.

Аналогично обстоит дело и со сбором информации. Этим занимается робот-паук, который обходит страницы с заданными URL и скачивает их в базу данных, а затем архивирует и перекладывает в хранилище суточными порциями. Робот размещается на нескольких машинах, и каждая из них выполняет свое задание. Так, робот на одной машине может качать новые страницы, которые еще не были известны поисковой системе, а на другой — страницы,

которые ранее уже были скачаны не менее месяца, но и не более года назад. Хранилище у всех машин едино.

При необходимости работу можно распределить другим способом, например, просто распределив между роботами всю работу, учитывая лишь ее объемные показатели. Параллельная работа программы позволяет легко выдерживать дополнительную нагрузку — при увеличении количества страниц, которые нужно обойти роботу, достаточно просто распределить задачу на большее число машин.

В хранилище информация в сжатом виде собирается и разбивается на части. Эти части постепенно распределяются между множеством машин, на которых запущена программа-индексатор. Как только индексатор на одной из машин заканчивает обработку очередной части страниц, он обращается за следующей порцией. В результате на первом этапе формируется много маленьких индексных баз, каждая из которых содержит информацию о некоторой части Интернета. При увеличении нагрузки на машины, занимающиеся индексированием, проблема может быть решена простым добавлением машин в систему.

После того как все части информации обработаны, начинается объединение (слияние) результатов. Основная база участвует в анализе как одна из частей нового индекса. Так, если объединяются 70 новых частей, то в анализе участвует 71 фрагмент (70 новых + основная база предыдущей редакции). Специальная программа ("сливатор") составляет таблицы перенумерации документов базы. Содержимое всех частей объединяется. Среди страниц с одинаковыми адресами выбирается наиболее свежая версия; если при скачивании URL последней информацией была ошибка 404 (запрашиваемая страница не существует), она временно удаляется из индексной базы. Параллельно осуществляется склейка дублей — страницы, которые имеют одинаковое содержимое, но различные URL, объединяются в один документ.

Сборка единой базы из частичных индексных баз представляет собой простой и быстрый процесс. Сопоставление страниц не требует никакой интеллектуальной обработки и происходит со скоростью чтения данных с диска. Если информации, которая генерируется на машинах-индексаторах, получается слишком много, то процедура "сливания" частей проходит в несколько этапов. Вначале частичные индексы объединяются в несколько промежуточных баз, а затем промежуточные базы и основная база предыдущей редакции пересекаются. Таких этапов может быть сколько угодно. Промежуточные базы могут сливаться в другие промежуточные базы, а уже потом объединяться окончательно. Поэтапная работа незначительно замедляет формирование единого индекса и не отражается на качестве результатов.

Источник: "Принципы работы поисковой машины Рамблер" (**http://www.rambler.ru/doc/architecture.shtml**).