«Яндекс» — поисковый движок, принадлежащий российской корпорации «Яндекс», основной продукт компании.

Доля «Яндекс.Поиска» составляет 57,5 % на рынке Рунета (октябрь 2015 года) и 7 % на рынке Турции (по данным на октябрь 2015 года).

Основная информация

Поисковая машина состоит из трёх основных компонентов:

Агент — это поисковый робот. Он обходит сеть, скачивает и анализирует документы. В случае обнаружения новой ссылки при анализе сайта она попадает в список веб-адресов робота. Поисковые роботы бывают следующих типов: *пауки* (англ. *Spider*) — загружают сайты подобно браузерам пользователя; *«путешествующие» пауки* (англ.*Crawler*) — обнаруживают новые, ещё неизвестные ссылки на основе анализа уже известных документов; *индексаторы* — занимаются анализом обнаруженных веб-страниц и добавляют данные в *индекс*. Множество выкачанных документов разбиваются на непересекающиеся части и очищаются от разметки.

Индекс — база данных, собранная роботами-индексаторами поисковых машин. По индексу и осуществляется поиск документов.

Поисковый механизм

Поисковый запрос от пользователя после анализа загруженности поисковой системы отправляется на наименее загруженный сервер. Для обеспечения такой возможности сервера «Яндекса» объединены в кластеры и даже кластеры кластеров. Затем пользовательский запрос обрабатывается программой под названием «Метапоиск». Метапоиск осуществляет анализ запроса в реальном времени: определяет географическое положение пользователя, проводит лингвистический анализ и т. д. Также программа определяет, относится ли запрос к категории наиболее популярных или недавно заданных. Выдача на такие запросы некоторое время хранится в памяти (кэше) метапоиска, и в случае совпадения показываются заранее сохранённые результаты. Если запрос является редким и совпадений в кэше не найдено, система перенаправляет его на программу

«Базового поиска». Тот анализирует индекс системы, также разбитый по разным дублирующимся серверам (это ускоряет процедуру). Затем полученная информация снова попадает на метапоиск, данные ранжируются и показываются пользователю в готовом виде.

Индексирование

В целом «Яндекс» индексирует следующие типы файлов: html, pdf, rtf, doc, xls, ppt, docx, odt, odp, ods, odg, xlsx, pptx.

Поисковая система способна также индексировать текст внутри объектов Shockwave Flash (если текст не помещен на само изображение), если эти элементы передаются отдельной страницей, имеющей MIME-тип application/x-shockwave-flash, и файлы с расширением .swf.

В «Яндексе» работают 2 сканирующих робота — «основной» и «быстрый». Первый отвечает за интернет в целом, второй индексирует сайты с часто меняющейся и обновляемой информацией (новостные сайты и информационные агентства). В 2010 году «быстрый» робот получил новую технологию под названием «Orange», разработанную совместно калифорнийским и московским подразделениями «Яндекса».

В логах сервера роботы «Яндекса» представляются следующим образом:

Mozilla/5.0 (compatible; YandexBot/3.0) — основной индексирующий робот.

Mozilla/5.0 (compatible; YandexBot/3.0; MirrorDetector) — робот, определяющий зеркала сайтов. Если найдутся несколько сайтов с одинаковым содержимым, в результатах поиска будет показан только один.

Mozilla/5.0 (compatible; YandexImages/3.0) — индексатор «Яндекс.Картинок».

Mozilla/5.0 (compatible; YandexVideo/3.0) — индексатор «Яндекс.Видео».

Mozilla/5.0 (compatible; YandexMedia/3.0) — робот, индексирующий мультимедийные данные.

Mozilla/5.0 (compatible; YandexBlogs/0.99; robot) — робот поиска по блогам, индексирующий комментарии постов.

Mozilla/5.0 (compatible; YandexAddurl/2.0) — робот, обращающийся к странице при добавлении её через форму «Добавить URL».

Mozilla/5.0 (compatible; YandexFavicons/1.0) — робот, индексирующий иконки сайтов (favicons).

Mozilla/5.0 (compatible; YandexDirect/3.0) — робот, индексирующий страницы сайтов, участвующих в «Рекламной сети „Яндекса“» (РСЯ).

Mozilla/5.0 (compatible; YandexDirect/2.0;
Dyatel) — «простукивалка» «Яндекс.Директа».

Mozilla/5.0 (compatible; YandexMetrika/2.0) — робот «Яндекс.Метрики».

Mozilla/5.0 (compatible; YandexCatalog/3.0; Dyatel) — «простукивалка» «Яндекс.Каталога».

Mozilla/5.0 (compatible; YandexNews/3.0) — индексатор «Яндекс.Новостей».

Mozilla/5.0 (compatible; YandexAntivirus/2.0) — антивирусный робот «Яндекса».

Поисковые запросы

Интерфейс «Яндекса» располагает довольно сложным языком запросов, позволяющим ограничить область поиска отдельными доменами, языками, типами файлов и т. д.

Для настройки используются следующие символы-джокеры:

"" — точная цитата

| — ставится между словами, если нужно найти одно из них

* — ставится между словами, если между ними пропущено какое-то слово

site: — поиск на определённом сайте

date: — поиск документов по дате, например, date:2007

+ — ставится перед словом, которое обязательно должно присутствовать в документе

& — ставится между словами, которые должны встречаться в пределах одного предложения

и др.

«Яндекс» автоматически, наряду с оригинальной «точной формой» запроса, ищет его различные вариации и формулировки.

Поиск «Яндекса» учитывает морфологию русского языка, поэтому вне зависимости от формы слова в поисковом запросе выдача будет производиться по всем словоформам. Если морфологический анализ нежелателен, можно перед словом поставить восклицательный знак (!) — поиск в этом случае покажет только конкретную форму слова. Кроме того, при поисковом запросе практически не учитываются так называемые стоп-слова, то есть предлоги, знаки препинания, местоимения и т. д., ввиду их большого распространения.

Защита от спама и вирусов

По состоянию на 2013 год «Яндекс» является самой безопасной поисковой машиной на планете и третьим по степени защищённости среди всех веб-ресурсов.

Проверка веб-страниц и предупреждение пользователей появились на «Яндексе» в 2009 году: с тех пор на странице результатов поиска рядом с опасным сайтом появляется пометка «Этот сайт может угрожать безопасности вашего компьютера». Для обнаружения угроз используются сразу две технологии. Первая куплена у американского антивируса «Sophos» и основана на сигнатурном подходе: то есть при обращении к веб-странице антивирусная система обращается к базе данных уже известных вирусов и вредоносных программ. Такой подход отличается высокой скоростью, но практически бессилен перед новыми вирусами, ещё не попавшими в базы данных. Поэтому «Яндекс» использует наряду с сигнатурным ещё и свой собственный антивирусный комплекс, основанный на анализе поведенческого фактора. Программа «Яндекса» при обращении к сайту проверяет, запрашивал ли последний у браузера дополнительные файлы, перенаправлял ли на посторонний ресурс и т. д. Таким образом, если получены данные, что сайт начинает выполнение неких действий (запускаются каскадные таблицы стилей, модули Java Script и полноценные программы) без ведома пользователя, он помещается в «чёрный список» и базу вирусных сигнатур. Информация о заражении сайта появляется в результатах поиска, и через сервис «Яндекс.Вебмастер» соответствующее уведомление получает владелец сайта.

После первой проверки «Яндекс» делает вторую, и если информация о заражении во второй раз подтвердится, проверки будут проходить чаще, пока угроза не будет устранена. Общее число заражённых сайтов в базе «Яндекса» не превышает 1 %.

Ежедневно в 2013 году «Яндекс» проверяет 23 млн веб-страниц (обнаруживая при этом 4300 опасных сайтов) и показывает пользователям 8 млн предупреждений. Ежемесячно проверяется примерно миллиард сайтов.