Newmann Bauer marketing group

Москва, Певческий переулок, д. 4, стр. 1

www.nbmarketing.ru, info@nbmarketing.ru

Тел. (495) 660-37-04

Обзор и исследование поисковых систем.

24 июня 2010 г.

Аналитическое исследование

Содержание:

История зарождения современных поисковых систем

Первая интернет-страница, при создании которой была применена технология HTTP, появилась сравнительно недавно - в 1990 году. Ее создателем является британский ученный Тим Бернерс-Ли, который также является изобретателем URI, URL, HTTP, World Wide Web. Созданный им сайт info.cern.ch (в данный момент доступный в сети в качестве авторской страницы создателя) является прародителем не только современных информационных ресурсов, но и первым в мире доступным каталогом интернет-сайтов. С этого момента Интернет начал набирать популярность не только среди научных кругов, но и среди простых обладателей персональных компьютеров.

В 1993 году была создана первая в мире поисковая система для Всемирной сети «Wandex». В ее основу был заложен World Wide Web Wanderer бот[1], разработанный Метью Греем из Массачусетского технологического института. Через несколько месяцев после рождения поисковой системы «Wandex» была создана конкурирующая система «Aliweb», которая в отличие от индекса «Wandex» работает до сих пор. В 1994 году была запущена первая полнотекстовая [2] («crawler-based», то есть индексирующая ресурсы при помощи робота) поисковая система «WebCrawler». Основным отличием поисковой системы от своих предшественников является предоставление возможности пользователям осуществлять поиск по любым ключевым словам на любой веб-странице. Сегодня эта технология является стандартом поиска любой поисковой системы. Поисковая система «WebCrawler» стала первой системой, о которой было известно широкому кругу пользователей.

Первой поисковой системой, которая была доступна русскоязычным пользователям Интернета, стала поисковая машина «AltaVista», которая в 1996 году запустила морфологическое расширение для русского языка. В этом же году были запущены первые отечественные поисковые системы – «Rambler.ru» и «Aport.ru». Появление первых отечественных поисковых систем ознаменовало новый этап развития Рунета, позволяя русскоязычным пользователям осуществлять запрос на родном языке, а также оперативно реагировать на изменения, происходящие внутри Сети.

С запуском в 1997 году поисковой системы «Яндекс» отечественные поисковые машины начали конкурировать между собой, улучшая систему поиска и индексации сайтов, выдачи результатов, а также предлагая новые сервисы и услуги.

В западных странах переломный момент в развитии поисковых систем наступил с появлением в 1997 году поисковой системы Google. Компания Google разработала собственную поисковую машину, которая дала пользователям возможность осуществлять качественный поиск с учетом морфологии, ошибок при написании слов, а также повысить релевантность[3] в результатах выдачи запросов. Сегодня компания Google обрабатывает более 40 миллиардов запросов в месяц, что соответствует 62,4 % всех поисковых запросов в мире.

[1] Бот (веб-паук, краулер)- производное от слова «робот». Аппаратно программный комплекс, применяемый в поисковых системах для добавления сайта в электронный каталог. Является составной частью «поисковой машины».

[2] Полнотекстовый поиск — поиск документа в базе данных текстов на основании содержимого этих документов, а также совокупность методов оптимизации этого процесса.

[3] Релевантность - в широком смысле - мера соответствия получаемого результата желаемому результату. Релевантность - в поисковых системах - мера соответствия результатов поиска задаче поставленной в запросе.

Поисковая машина

Поисковая машина - это аппаратно-программный комплекс, осуществляющий быстрый поиск необходимой информации внутри сервера или интернет-ресурса. Основа поисковой машины у всех поисковых систем примерна одинаковая. Как правило, это поисковый бот, необходимый для индексации и поиска сайта, программное обеспечение, отвечающее за составление каталога запроса и ранжирование результатов по релевантности поискового запроса. Но многие крупные поисковые системы держат в секрете содержание своей поисковой машины. Ключевым отличием является база проиндексированных сайтов, релевантность и учет морфологии языка запроса. Все это в совокупности и определяет критерий качества работы поисковых машин.

Классифицируется поисковая машина по области поиска информации:

1. *Локальный поиск.* Предназначен для осуществления поиска информации по какой-либо части всемирной сети, например, по одному или нескольким сайтам, либо по локальной сети. Примером служит поисковый скрипт на сайте или внутренние серверы крупных компаний.

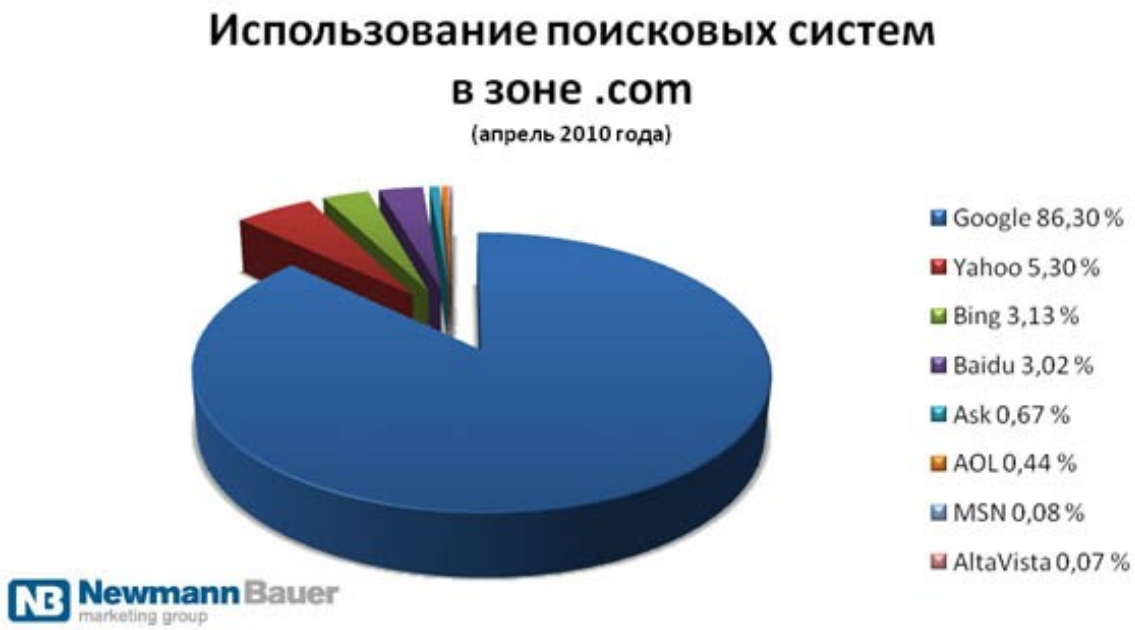
2. *Глобальный поиск*. Предназначен для поиска информации по сети Интернет, либо по региональной части, группе сайтов и т.д. Глобальный поиск используют крупные поисковые системы Яндекс, Google, Yahoo и т.д.

Поисковые машины осуществляют различный поиск информации по сети Интернет. Например, картинки, музыка, географическое положение, личная информация и т.д. Файлы, с которыми работает поисковая машина, могут быть разных форматов (например .html,.htm,.txt,.doc,.rtf, …), графического (.gif, .png, .svg, …) или мультимедийного (видео, звука и другой информации). Но наиболее распространенным является поиск по текстовым документам (web-страницы, документы в формате doc, rtf, txt и др.). Поиск по изображениям, видео, звукам более сложен с технологической точки зрения, поэтому массово не реализован. Такие системы, как, например, Яндекс.Картинки искали не по самим изображениям, а по альтернативным текстам, соответствующим этим изображениям. А каталог поиска картинок в компании Google составляется вручную, что увеличивает релевантность запроса, но тормозит обновление баз изображений.

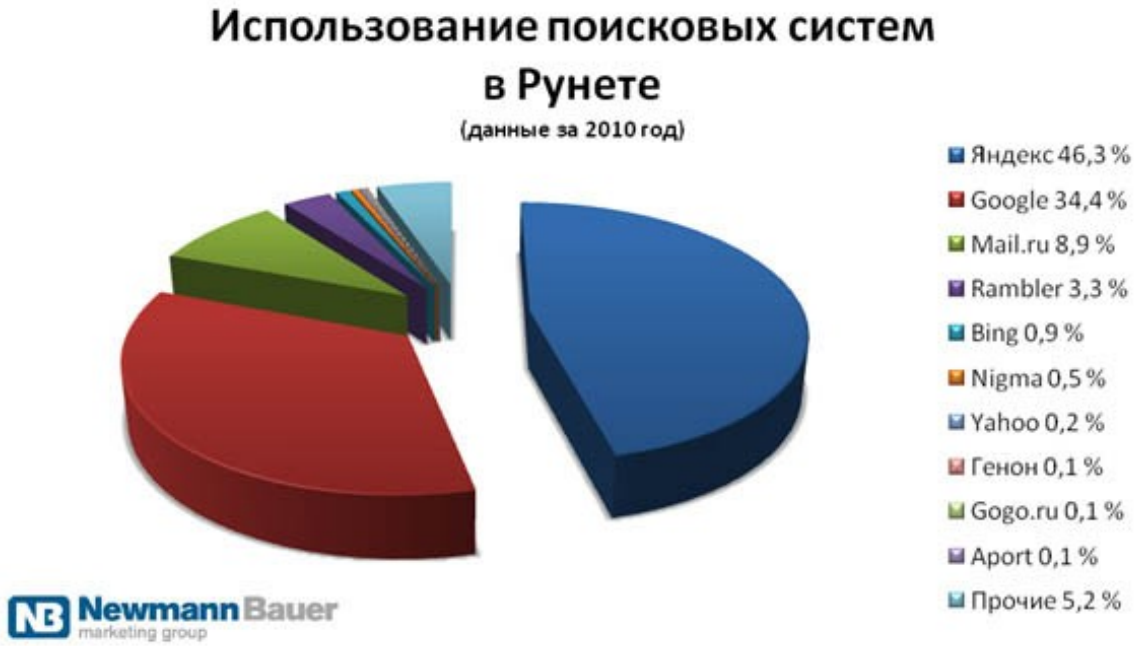
Рейтинг поисковых систем

Характерной ошибкой многих аналитических компаний является попытка сравнения поисковых систем, работающих с западными рынками и рынком Рунета. В широкое применение компьютеры, не говоря уже про Интернет, вошли только в начале 90-х годов прошлого века. И развитие интернет-технологий в постсоветском пространстве происходило благодаря энтузиазму отдельных людей. Крупные IT-компании, работающие сейчас на рынке Рунета, появились в результате консолидации отдельных групп инициативных людей. Примером служит компании «ABBYY», «Студия Артемия Лебедева», «Лаборатория Касперского» и др. На Западе многие компании, работающие в сфере высоких технологий, образовываются в результате грандов и кредитов со стороны государственных и частных структур, а также венчурных фондов. В России только сейчас начинают внедряться технологии поддержки инноваций в сфере высоких технологий.

В зоне .com использование поисковых систем распределилось следующим образом (данные полученные от компании Net Applications):



Российский рынок поисковых систем в подавляющем большинстве представлен одной отечественной (поисковая система Яндекс, доля рынка – 46,3 %) и одной западной (поисковая система Google, доля рынка – 34,4 %) компаниями.



Популярность сразу двух поисковых систем на рынке Рунета обусловлена непохожестью результатов выдачи искомого. Компании Яндекс и Google используют собственные технологии ранжирования и релевантности поисковых запросов, соответственно выдача результатов одной поисковой системы зачастую отличается от другой. Интересен и другой факт, компании Mail.ru и Rambler, растеряв свои позиции в поисковом рейтинге, остаются крупными IT-холдингами благодаря сервисам и сайтам, которые находятся на их платформе. По популярности сервисов и сайтов портал Mail.ru опережает многих отечественных, а также западных конкурентов (доля на рынке 59 %), а в сфере веб-почты компания Mail.ru занимает лидирующие позиции на пространстве СНГ. С отказом компании Mail.ru от технологии поиска Google и Яндекс, и запуском собственного поискового алгоритма go.mail.ru, компания начала наверстывать упущенное в поисковом сегменте рынка и сегодня она занимает третью строчку в поисковом рейтинге.

Статистика переходов

Основным критерием оценки качественной работы поисковых систем служит «статистика переходов» со страницы каталога выдачи запроса. Статистика переходов зависит от времени года, важных событий в мире, праздников, состояния экономики и других немаловажных факторов.

Статистику переходов можно условно разделить на несколько категорий:

Популярные категории. К популярным категориям относятся ежедневные запросы по основным тематикам каталога (фото, видео, картинки и т.п.). Популярными категориями пользуются в основном люди, проводящие в сети Интернет 3-5 часов в сутки, их мнение влияет на составление каталога популярных категорий. Социальный портрет пользователей: пользователи в возрасте 20-30 лет, работающие в частных или госструктурах, социальный статус – служащие, достаток средний или выше среднего, имеющие свободный доступ к сети Интернет. Время посещения сети Интернет – утренние часы, обеденное время, частично вечернее.

Newmann Bauer
marketing group

№	Категории	Статистика
1	Знакомства и общение	17 103 373
2	Интернет	5 992 676
3	Новости и СМИ	3 707 231
4	Кино	3 316 244
5	Игры	3 183 062
6	Развлечения	2 901 185
7	Товары и услуги	2 534 746
8	Дом и семья	2 473 929
9	Авто	2 020 423
10	Города и регионы	1 831 936
11	Софт	1 750 451
12	Спорт	1 565 823
13	Музыка	1 446 285
14	Фото	1 346 065
15	Юмор	1 298 244

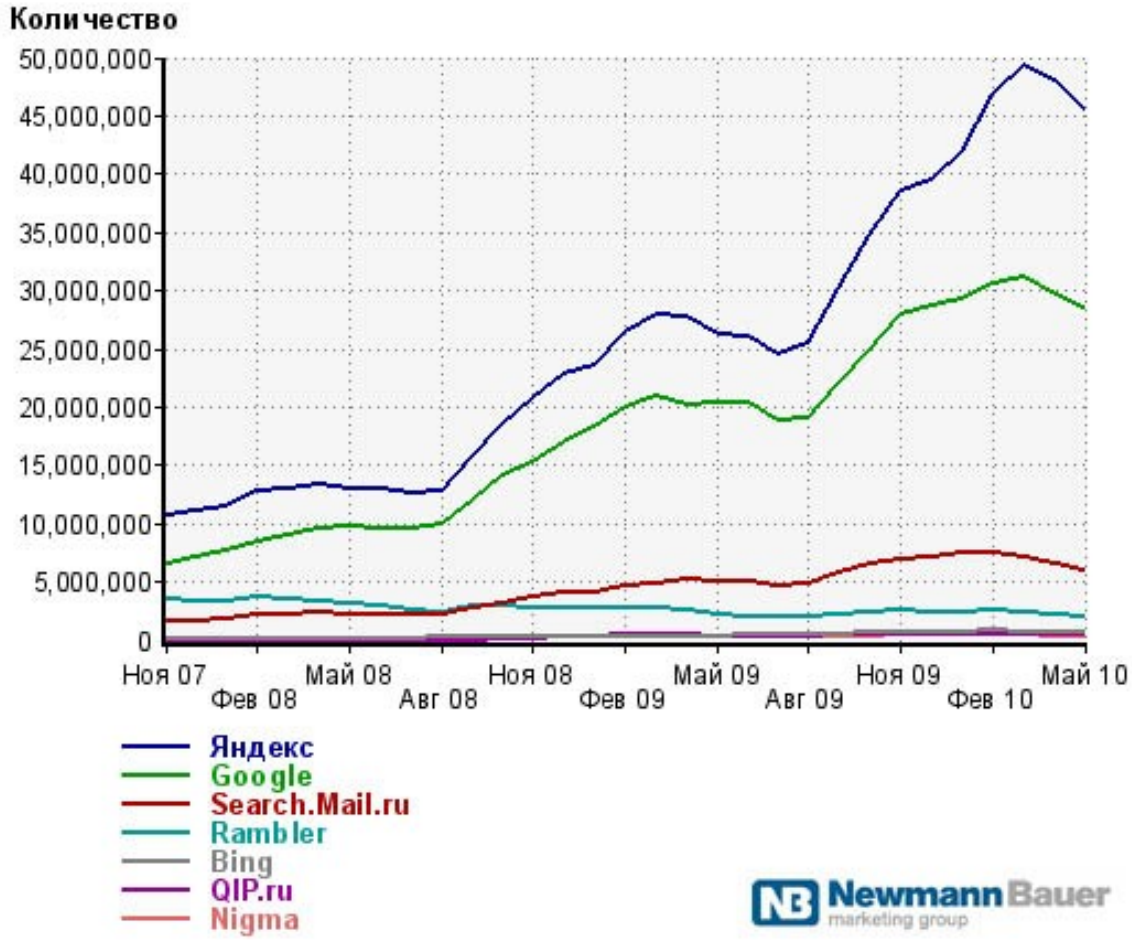
1. Всплеск по интересам. Всплеск запросов по интересам обычно связан со знаменательными событиями в мире или в Рунете, которые так или иначе вызывают интерес у пользователей сети. Существует сезонные всплески интересов, например, Новый Год, Евровидение и т.п. Ниже приведена таблица всплесков по интересам за 2009 год.

Newmann Bauer
marketing group

Newmann Bauer
marketing group

	Москва	Центр	Севера-Запад	Юг	Поволжье	Урал	Сибирь	Дальний Восток	Вся Россия
1.Новый год	6 742 000	2 053 000	2 770 100	1 319 400	3 860 800	1 712 200	2 019 000	422 100	20 907 800
2.Евровидение 2009	2 946 700	781 500	1 071 800	492 300	1 300 700	492 900	581 200	114 400	7 805 800
3.Готовые домашние задания (ГДЗ)	2 921 500	470 700	1 983 200	251 300	1 710 100	1 138 300	503 100	96 800	9 086 700
4.GTA 4	2 000 700	531 700	923 700	369 900	1 003 300	477 600	566 900	151 400	6 039 100
5.ЕГЭ 2009	1 909 000	681 400	744 500	508 100	1 253 400	417 400	625 100	148 900	6 319 700
6.Международный женский день	1 643 000	303 400	500 900	234 100	606 300	330 200	336 600	87 800	4 066 200
7.Ночь музеев	1 532 100	92 600	228 400	33 000	114 800	58 700	66 500	10 300	2 141 600
8.Свиной грипп	1 363 900	402 700	514 500	186 500	647 400	235 600	351 600	78 100	3 784 000
9.The Sims 3	1 339 100	422 200	567 100	252 400	677 300	318 500	389 800	89 100	4 061 700
10.День защитника Отечества	1 174 200	286 900	455 200	239 300	596 300	297 700	380 800	92 400	3 543 700
11.День Святого Валентина	1 027 200	224 700	423 200	167 100	490 200	218 600	257 600	53 600	2 875 200
12.Рождество	911 300	208 100	373 300	149 900	369 100	170 300	206 400	42 100	2 436 200
13.День Победы	893 700	151 800	326 700	108 700	294 600	133 900	157 300	37 500	2 113 900
14.Пасха	760 500	142 300	301 900	120 900	238 900	116 600	151 300	32 700	1 873 700
15.«Ледниковый период 3»	724 400	206 700	324 700	91 000	315 800	137 600	177 800	19 500	2 001 400

По данным компании Liveinternet, на рынке Рунета по количеству переходов лидирует поисковая система Яндекс. Статистика переходов влияет в первую очередь на рейтинг поисковой системы на рынке Рунета. Ниже представлена диаграмма переходов за 2009-2010 год (по данным Liveinternet).



Индекс цитирования

Индекс цитирования (ИЦ) — основной показатель ранжирования сайтов в каталоге выдачи запроса, вычисляемый на основе числа ссылок на данный сайт с других ресурсов сети Интернет. В простейшей разновидности учитывает только количество ссылок на ресурс.

Тематический индекс цитирования (тИЦ) – технология поиска, используемая в поисковой машине Яндекса. ТИЦ определяет степень авторитетности интернет-ресурса с учетом качественной характеристики ссылок на него. Качественные характеристики ссылок в компании Яндекс называют «весом» индекса цитирования. Рассчитывается при помощи алгоритма (сам алгоритм держится в секрете, чтобы уменьшить появления поискового спама). Большую роль играет тематическая близость ресурса и ссылающихся на него сайтов. Количество ссылок на ресурс также влияет на значение его тИЦ, но тИЦ определяется не количеством ссылок, а суммой их «веса».

Первоначально, до того как появились оптимизаторы сайтов, индекс цитирования реально отражал популярность соответствующего ресурса в Интернете. Первой крупной поисковой системой, использовавшей в своем алгоритме индекс цитирования, стала Google (алгоритм PageRank).

Поисковый спам

Поисковый спам – сайты и страницы в сети Интернет, созданные для манипуляции результатами поиска, увеличению тИЦ. Поисковый спам улучшает позиции интернет-ресурса в каталоге выдачи запроса, но негативно влияет на его популярность. Известны случаи, когда интернет-ресурсы с оригинальным контентом теряли популярность и попадали в списки поискового спама, теряя при этом посетителей.

Поисковый спам подразделяется на несколько типов:

- Слова в тегах meta keywords, description, например sex, халява которые не имеют отношение к контенту сайта, но пользующиеся популярностью в поисковых запросах. Для борьбы с этим видом спама поисковые машины анализируют не только теги, но и сам контент.

- Большое количество ключевых слов, искусственное повышение частоты ключевого слова или выражения в тексте, использование элементов разметки HTML (h1-3, strong, b, em, i) для искусственного повышения веса ключевого слова.

- «Невидимый текст» - текст, невидимый посетителями сайта, но индексируемый поисковыми машинами. Применяется цвет текста, соответствующий цвету фона, текст размером в 1 пиксель, блоки текста со стилем «display:none».

Ссылочный спам — ссылки, «накручивающие» link popularity и PageRank сайта. Так как поисковые машины, отвечая на запрос, руководствуются количеством ссылок ведущих на сайт с других ресурсов, можно предложить следующие идеи по увеличению числа этих ссылок:

1. Создавать небольшие сайты с использованием бесплатных хостинг-провейдеров, зарегистрировав их в большом количестве тематических каталогов и с них ссылаться на основной сайт.

2. Принять участие в обмене ссылками.

3. Приобретать ссылки за деньги.

4. Ссылочный спам с гостевых книг, блогов, вики и пр.

Поисковые системы, борясь с появлением поискового спама, создают фильтры, куда добавляются сайты, ссылки с которых не учитываются при ранжировании.

doorway (дорвей, входная дверь)— промежуточные страницы, созданные для накрутки веса страницы при ссылочном ранжировании. Часто страницы-дорвеи перенаправляют посетителя на другую страницу или другой интернет-ресурс. Поисковые машины в ответ удаляют из своей базы данных сайты, в которых есть автоматическое перенаправление.

Маскировка — анализ переменных запроса, при котором поисковой машине отдается содержимое сайта, отличное от того, которое видит пользователь.

Основная проблема, порождаемая поисковым спамом, это появление огромного количества мусорного контента. Все это отрицательно влияет на релевантность результатов выдачи, снижает и искажает эффективность работы поисковых машин. В конечном итоге, Интернет перестает быть источником для получения объективной информации. Также и спам заставил поисковые алгоритмы критически относиться к «добропорядочным» сайтам, на которые не ссылаются другие ресурсы, что уменьшило в итоге релевантность сайтов по менее популярным запросам.

24.06.2010 Newmann Bauer marketing group

Москва, Певческий переулок, д. 4, стр. 1

www.nbmarketing.ru, info@nbmarketing.ru

Тел. (495) 660-37-04