**Аналитика и полезная информация**

**Андрей Иванов**

**Подготовка сайта для индексирования в поисковых системах**

**Индексирующие поисковые системы** - сегодня основные поставщики новых посетителей на сайт. В то же время алгоритмы работы поисковиков никогда не публикуются, на что у разработчиков есть все основания - борьба с замусориванием поисковиков (поисковым спамом или спамдексингом) отнимает много сил и средств у коллектива любого поискового проекта.

Вопрос о том, как попасть на первую страницу результатов поиска по профильным, то есть точно относящимся к тематике сайта, запросам, волнует многих. Явный дефицит популярной информации на эту тему породил большое количество домыслов - о "чудодейственных" методах, о "коррумпированности" и "протекционизме" администраций поисковых систем, "продающих" и раздающих выгодные места своим людям.

Опыт анализа сайтов, владельцы которых испытывают проблемы с посещаемостью из поисковых систем, показывает, что причины, как правило, гораздо проще. Они заключаются в элементарной неинформированности разработчиков и владельцев сайта о самых простых правилах, которые необходимо учитывать, чтобы сайт был хорошо представлен в поисковых системах.

**Поисковые запросы и релевантность**

Разумеется, для того, чтобы на сайт из поисковой машины пришли посетители, ему недостаточно просто присутствовать в результатах поиска. Необходимо попасть в начало списка результатов поиска.

Критерии, от которых зависит положение ссылки на сайт на страницах результатов поиска, можно разделить на две группы - текстовые и нетекстовые. Из названий ясно, что текстовые критерии - это характеристики текстового содержимого страницы. Нетекстовые критерияъи оценивают связи страницы, а ее текст при этом роли не играет.

Учет текстовых критериев осуществляется на этапе создания сайта и текста. Работа с нетекстовыми параметрами возможна уже после того, как сайт выложен в Сеть и подан на индексацию.

В работе с индексирующими поисковыми системами всегда следует помнить, что командой на поиск в поисковике является поисковый запрос, введенный пользователем - слово, несколько слов, словосочетание, фраза.

Здесь возникает довольно существенный смысловой (семантический) разрыв. За то, что происходит в голове пользователя, поисковик не отвечает, а для пользователя не существуют понятия "правильного" и "неправильного" запросов.

Поэтому всегда перед тем, как планировать сайт, стоит ознакомиться с тем, как часто и в какой форме пользователи **на самом деле** запрашивают информацию, предполагаемую к размещению в Интернете.

В Рунете единственным источником достоверных данных о поисковых запросах является сервисная служба системы рекламных объявлений "Яндекс-Директ".

**Текстовые критерии релевантности**

Сегодня можно полностью забыть такое устаревшее понятие, как "ключевые слова". Первоначально ключевыми назывались слова, внесенные в особое поле страницы, помеченное тегом <meta name="keywords" ...>. Но сегодня поисковые системы игнорируют содержание этого поля. Чтобы избежать разночтений, мы вообще откажемся от использования этого термина.

Результаты поиска определяются прежде всего поисковым запросом; поэтому, когда речь далее пойдет о ключевых словах и "соответствии-несоответствии" текста страницы запросу или отдельным словам, всегда будут иметься в виду слова, из которых состоит поисковый запрос.

**1. Титульная фраза**

Титульная фраза (титул) - это текст, который содержится на странице в дескрипторе title, в коде страницы он оформляется как <title>Титульная фраза</title>. Слова титульной фразы имеют очень большое значение при упорядочивании (ранжировании) результатов поиска в поисковой системе. А именно, если формулировка запроса совпадает с титульной фразой, либо титульная фраза содержит несколько слов запроса, то страница с таким титулом при прочих равных условиях окажется выше.

Стоит отметить противоречивый статус титульной фразы, ставший причиной наиболее грубых ошибок многих вебмастеров.

Титульная фраза не отображается на странице сайта в браузере. Точнее, отображается, но на рамке окна браузера, то есть в месте, куда обычно никто не смотрит. Именно поэтому настолько распространена ошибка, когда вебмастер ставит для всех страниц сайта одинаковый титул. Обычно это бывает само название сайта, что весьма корректно выглядит, когда сайт, например, показывается заказчику.

Но когда сайт подается на индексацию в поисковые системы, картина меняется, так как в результатах поисковых систем содержание тега <title> отображается в самом интересном месте - как заголовок найденной страницы! Именно титульная фраза страницы является наиболее ярким элементом списка результатов поиска по поисковому запросу. Здесь она играет свою главную роль - заголовка, "лица" страницы. И выбор пользователя более чем наполовину определяется точностью, понятностью и привлекательностью титула.

Таким образом, можно вывести первые правила подготовки сайта к индексации поисковыми машинами:

1. Все страницы сайта должны иметь разные титулы.

2. Титульная фраза каждой страницы должна быть точной и ясной.

3. Содержание страницы должно соответствовать титулу.

4. В титульной фразе страницы должны быть слова, часто встречающиеся в поисковых запросах пользователей, ответом на которые может являться данная страница.

5. Не стоит использовать титульные фразы длиной свыше 80 символов.

**2. Описание страницы**

Описание страницы - это содержание поля, помеченного тегом meta description, оформляется как <meta name="description" content="Описание страницы, максимальная длина - 200 символов">. Слова описания обычно не влияют на ранжирование результатов поиска, и среди русских поисковых систем данный тег используется на практике лишь "Яндексом". Однако "Яндекс" сейчас - ведущая система, и грамотно составлять описание даже в расчете только на "Яндекс" вполне целесообразно.

Текст описания показывается в результатах поиска "Яндекса" под текстом титульной фразы. Таким образом, правильно оформив титул и описание, можно достичь максимальной привлекательности ссылки на вашу страницу в "Яндексе". Это срабатывает, когда пользователь, получив ответ, выбирает из нескольких альтернатив.

К сожалению, большинство вебмастеров пренебрегают такой возможностью улучшить внешний вид (понятность) своих ссылок в поисковике, ставя в описание всех страниц сайта одну и ту же фразу, либо вообще игнорируя это поле.

Но в идеале нужно... Сформулируем очередные правила:

6. Описания всех страниц сайта в теге <meta name="description" ...> должны быть разными.

7. Описание каждой страницы должно соответствовать ее содержанию.

**3. Прочие текстовые критерии**

Поисковые машины используют также следующие текстовые критерии релевантности найденных страниц запросу:

позиция найденных слов в титульной фразе;

"вес" найденного слова и близость его к началу документа;

наличие найденных слов в заголовках (<h1>, <h2> и т. д.) и фрагментах, выделенных жирным шрифтом (<bold>, <strong>);

близость найденных слов друг к другу (компактность цитаты).

Объясним, что подразумевается под "найдеными словами". Индексирующие поисковые системы основаны на идее поиска "цитаты". Подразумевается, что пользователь хочет увидеть документ, в котором есть слова, введенные им в виде запроса. Поисковик находит в своем индексе такие документы и выдает их адреса. В вышеперечисленных критериях "найденное слово" - это слово, имеющееся как на веб-странице, так и в поисковом запросе, по которому должна быть показана ссылка на данную веб-страницу.

С ростом количества документов в Сети все острее становилась проблема ранжирования ссылок, выдаваемых поисковиками в ответ на запросы пользователей. Их стало слишком много. Поэтому разработчики стали один за другим вводить критерии, позволяющие как бы "уточнить" релевантность страницы - степень соответствия документа запросу.

Если в документе речь идет о каком-то предмете, по идее, этот предмет должен часто упоминаться (называться), т.е. частота слова - названия предмета будет повышенной. Этот параметр называется "весом" слова.

Авторы текстов знают, что главную идею материала лучше всего выносить в начало текста, поэтому некоторые поисковые системы положение найденного слова по отношению к началу документа стали также принимать за критерий релевантности.

Положение слов относительно друг друга, особенно в титуле, тоже является критерием. В случае поисковых запросов, состоящих более чем из одного слова, справедливо предположение (вспомним основную идею индексирующих поисковиков), что пользователь ищет цитату, то есть именно вхождение введенного словосочетания. Возможно, в запросе и пропущены некоторые слова, но в любом случае порядок слов и близость их друг к другу важны. Например, в запросе стропила плотники легко узнается название повести Сэлинджера "Выше стропила, плотники!", а плотники стропила напоминают уже о Бродском ("поднять не звали плотников стропила"). Подобная компактность цитаты хорошо распознается поисковыми системами.

Почему в качестве критерия релевантности используется наличие слов запроса в выделенных фрагментах и заголовках документа, очевидно: именно таким образом создатели текстов выделяют наиболее значимые слова, на которые хотят обратить внимание.

Как видим, все текстовые критерии ранжирования результатов в поисковых системах имеют простые логические объяснения. Разработчики поисковых систем ранжируют страницы, исходя из предположения, что вебмастера создавали и структурировали их в расчете на "живых" посетителей.

Распределение частот поисковых запросов. Вебмастера и поисковые системы. Спамдексинг

Поисковые запросы повторяются. Разные люди, в разных городах, в разное время и в разные поисковики вводят одни и те же поисковые запросы: работа, чат, секс, mp3, windows, линукс и др. Частоту использования каждого запроса можно подсчитать. Некоторые встречаются в среднем раз в неделю, а есть и такие, что ежедневно повторяются сотни раз. Чем длиннее запрос, тем реже он встречается. И наоборот, наиболее частотным запросом из группы, относящейся к какой-либо теме, всегда является наиболее общий запрос - одно слово или словосочетание, определяющее тему.

Поисковые системы сегодня стали основным поставщиком новых пользователей на любой интернет-ресурс, и вебмастера это прекрасно знают. Знают и закономерности спроса. Вебмастер, поддерживающий сайт по трудоустройству, например, в Санкт-Петербурге, безусловно осведомлен, что запрос работа встречается гораздо чаще, чем работа в Питере. Но нечеткий однословный запрос работа используют не только питерцы, но и москвичи, новосибирцы, казанцы... Поэтому вебмастерам сайтов по трудоустройству каждого из этих городов хочется попасть на первую страницу поиска именно по "главному" однословному запросу. Но городов много, а на первой странице результатов может разместиться всего лишь 15-20 ссылок.

С целью улучшить свою позицию в результатах поиска вебмастера применяли и применяют ряд технических приемов, предназначенных именно для обмана, "накручивания" поискового робота. Явление это называется спамом поисковых систем или спамдексингом (от spam + indexing) и появилось практически одновременно с появлением поисковиков.

Так как последние раньше использовали только текстовые критерии для ранжирования результатов, то типовым приемом спамдексинга стала так называемая "накачка" - искусственное завышение частот (весов) нужных слов на странице.

Варианты "накачивания" использовались самые различные: от примитивного "скрытого", невидимого обычному пользователю текста, до создания специальных текстов, где веса слов тщательно рассчитывались в соответствии со значениями, снимаемыми со страниц, занимающих первые позиции в ссылках по необходимым запросам.

Накачиваются сами страницы, титульные фразы, теги noframes и keywords. Специально создаются страницы с подобными "оптимизированными для поисковиков" текстами, попадая на которые, пользователь автоматически перебрасывается на другой сайт. Либо просто видит крупную надпись "Вход". Они так и называются - входные страницы. Более того, разработана специальная техника (клоакинг), когда поисковому роботу показывается одна страница, а пользователям - совершенно другая.

Подобные действия являются бичом поисковиков, потому что сильно перегружают индексы, уродуют внешний вид страниц с результатами поиска и резко снижают релевантность системы, то есть ее качество и ее конкурентное преимущество в борьбе за популярность с другими поисковиками.

История поисковиков есть история их постоянной борьбы со спамом. Совсем недавно в поисках оружия против текстовой "накачки" поисковики начали дополнять текстовые алгоритмы ранжирования результатов поиска нетекстовыми критериями - а именно, ссылочными.

**Нетекстовые критерии релевантности**

Нетекстовые критерии можно разделить на три основных типа. Мы даем условные названия, так как устоявшейся терминологии пока нет:

ссылочно-расчетные,

ссылочно-текстовые,

каталожные.

Как мы уже говорили, в случаях нетекстового критерия на ранжирование страницы влияет не ее содержание, а другие факторы.

**Ссылочно-расчетные критерии.** К ссылочно-расчетным критериям относится алгоритм поисковой системы Google - так называемый взвешенный индекс цитирования PageRank (PR). Подробная статья о нем Криса Райдингса (перевод и комментарии А. Садовского) имеется на сайте Александра Садовского http://www.digits.ru/

Индекс цитирования учитывает, как много ссылок имеется в Сети на ваш сайт, и насколько авторитетны ссылающиеся на вас сайты. Авторитетность "цитирования" определяется также по количеству ссылок на "цитирующего". Для расчета индекса цитирования Google и подобные ему системы регулярно "перетряхивают" невообразимо огромные матрицы связей между сайтами Интернета, пересчитывая вес ссылок и авторитетность ресурсов (заметим, что Google использует для этого "ферму" более чем из 10 000 серверов!). Подобный алгоритм используется также "Яндексом".

**Каталожные критерии.** К каталожным критериям можно отнести индекс цитирования "Яндекса" (citation index, CY) и алгоритм выдачи результатов поиска "Рамблера". В обоих случаях принцип один - наверху результатов поиска по запросу в индексах выдается несколько ссылок на сайты, зарегистрированные в принадлежащих порталам тематических каталогах. В случае "Яндекса" показывается до трех ссылок на сайты из каталога "Яндекса", если в их описании имеются слова запроса. Ранжирование производится по убыванию индекса цитирования "Яндекса". Каталожные ссылки в результатах поиска "Яндекса" четко отличаются от обычных счетных: они не нумеруются, а отмечаются точками. Описание ресурса в каталоге "Яндекса" проверяется, а частенько и составляется вручную составителями каталога (так называемыми модераторами). Индекс цитирования, естественно, рассчитывается автоматически.

В свою очередь, "Рамблер" "замешивает" на первую страницу результатов поиска по запросу до пяти ссылок на сайты, зарегистрированные в каталоге-рейтинге Rambler's Top100, если их посещаемость (по "хостам") выше некоторого минимального порога. Критерием для внесения данных ссылок в список результатов поиска служит наличие слов запроса в описании сайта, сделанном для Rambler's Top100. В отличие от "Яндекса", где размер описания ограничен двумя сотнями символов, на индексацию в Rambler's Top100 может быть подано описание размером до 4000 символов.

При внесении в рейтинг описание сайта также может проверяться сотрудником "Рамблера" - модератором рейтинга.

**Ссылочно-текстовые критерии.** К ссылочно-текстовым критериям может быть отнесен алгоритм, не так давно появившийся в "Яндексе". В данном случае "Яндекс" индексирует текст ссылки (то есть текст, выделенный как гиперссылка, на странице сайта). Другими словами, индексируется содержимое тега <a href=...>, и если запрос совпадает с этим содержимым, в результатах поиска выдается адрес, указанный в теге.

"Яндекс", очевидно, запоминает только внешние ссылки с сайтов, то есть страница, на которой указывает ссылка, не должна принадлежать сайту, где эта ссылка размещена.

При этом при показе результатов поиска подобная страница не имеет титула, и "Яндекс" сопровождает ссылку на нее явным указанием - "адрес найден по ссылке".

**Спрос и предложение - точки сближения в поисковых системах**

Поисковые запросы являются реальным выражением спроса пользователей на информацию в Интернете. Предложение информации осуществляется на страницах ресурсов Интернета. Индексирующие поисковые системы сегодня являются бесспорным лидером трафикогенерации для большинства сайтов и являются наиболее естественным связующим звеном, соединяющим спрос и предложение.

Как мы уже показали, в основе всех текстовых критериев поисковых систем лежит представление, что пользователь ищет цитату из какого-то документа. На самом же деле это не так. Пользователи ищут не цитаты из документов, а решение своих проблем - ответы на свои вопросы, которые они даже не всегда могут точно сформулировать. Обсуждение этого вопроса выходит за рамки предмета данной статьи, приведем лишь одно из доказательств.

Вот оно: язык спроса радикально отличается от языка предложения. Характеристиками языка предложения (текстов сайта) в массе являются грамотность, литературная корректность, четкость изложения мыслей. Этого совершенно нельзя сказать о языке спроса (языке запросов) - основная масса поисковых запросов относится к нечетким, т. е. допускающим более одного варианта понимания; запросы часто слишком лаконичны, полны опечаток, смешивают кириллицу и латиницу и т.п.

Поисковикам неоткуда брать информацию для своих индексов, кроме как со страниц, подготовленных вебмастерами. И они берут ее в рафинированном, литературном виде.

С другой стороны, вебмастерам никогда не удастся обучить пользователей формулировать поисковые запросы "как надо", длинными гладкими фразами на хорошем русском языке.

Поэтому первым шагом даже не подготовки сайта к индексации, а его разработки должно быть изучение того, как пользователи ищут в Сети информацию. Ту информацию, которую владелец сайта собирается предлагать на своем ресурсе.

Технически этот анализ осуществляется с помощью получения реальных формулировок запросов и информации об их частотах, выделении устойчивых направлений спроса, базовых формулировок и типовых конструкций поисковых запросов (подробнее об этом рассказано в статье "Анализ спроса и повышение видимости в поисковых машинах").

Далее разработчикам предстоит решить, как же соединить требования, предъявляемые к печатному тексту, с необходимостью включения в них слов и конструкций, информация о которых была получена путем анализа поисковых запросов. Процесс этот, увы, алгоритмизировать невозможно. Здесь разработчикам сайта придется решить не очень сложную, но важную лингвистическую задачу - попытаться описать свой бизнес не словами генерального директора, вебмастера или отдела маркетинга компании, а словами публики.

**С чего начать**

Наверное, эту статью логичнее было бы начать с конца. Сначала - изучение спроса, потом создание текстов, потом - оформление текстов с учетом текстовых критериев ранжирования, потом оформление описаний страниц для большей привлекательности ссылок на них в поисковых системах, и лишь потом - старт проекта и начало работ по использованию нетекстовых критериев (расстановке ссылок и пр.).

Действительно, это - самый правильный путь. Но на следование им не всегда хватает времени, ресурсов и терпения.

Однако мы еще раз повторим, что основная масса ошибок вебмастеров, снижающих видимость сайта в поисковых системах, касается не "коренных" задач планирования, и не неосведомленности в тонкостях построения "иерархических", "циклических" и "обширных" систем связывания страниц для оптимизации PageRank. И дело не в неспособности писать хитрые "входные страницы", чтобы не быть "забанеными" модераторами или фильтрами поисковых машин.

Основная ошибка банальна - это неумение работать с титульными фразами и описаниями. Исправив грубые ошибки в титулах и описаниях, можно, во-первых, значительно расширить круг поисковых запросов, по которым ссылки на страницы сайта будут показаны на лучших позициях. И, во-вторых, придать своим ссылкам в поисковиках достойный, привлекательный внешний вид, вызывающий желание щелкнуть по ним.

Начинать нужно именно с этого.