Тема "Критерии релевантности веб-страницы запросу в поисковике"

Релевантность (*relevance* от английского *relevant* — относящийся к делу; обозначает соответствие найденного документа запросу, сделанному пользователем поисковой системы.) — мера соответствия получаемого результата желаемому. В терминах поиска — это мера соответствия результатов поиска задаче поставленной в поисковом запросе. Определяет, насколько полно тот или иной документ отвечает критериям, указанным в запросе пользователя. Необходимо учитывать, что в каждой поисковой системе работает собственная программа спайдер, индексирующая веб-страницы, каждая система индексирует страницы своим особым способом и приоритеты при поиске по индексам тоже различны. Поэтому запрос по одним и тем же ключевым словам в каждой из поисковой систем порождает разные результаты.

Концептуально степень релевантности можно измерять любым вещественным числом от 0 до 1, но поскольку человек не способен четко характеризовать степень релевантности, то ее часто оценивают при помощи шкалы.

Базис пространства релевантности: для классификации типов релевантности концептуально пользуются каркасом, который имеет три размерности:

Информационная потребность.

Выделяется 4 представления информационной потребности:

1. реальная потребность: неосознанная истинная информационная потребность пользователя (например, поиск некой новой информации исследователем, про которую он толком ничего не знает);
2. осознанная потребность: то, как пользователь понимает стоящую перед ним неосознанную проблему;
3. выраженная потребность: то, как пользователь описывает свою потребность средствами естественного языка;
4. формализованная потребность: представление средствами языка запросов поисковой системы.

Информационные ресурсы.

Выделяется 4 типа информационных ресурсов, которые могут быть доступны пользователю в процессе поиска: множество документов:

1. набор документов, которые вместе удовлетворяют потребность пользователя;
2. документ: полный информационный ресурс, ссылка на который представляется пользователю в результате поиска;
3. метаинформация: структурированная информация о документе, такая как, например, библиографическая информация, характеристика качества документа или отзывы других пользователей;
4. суррогат: представление документа в виде заголовка, автора, аннотации и т.п.

Контекст использования информации.

Этот контекст концептуализируется при помощи трех компонент:

1. тематика: область интересов пользователя;
2. задача: процесс или задача, для решения которой пользователь инициировал поиск;
3. атрибуты пользователя: описание характеристик пользователя, таких как его знания в этой тематике или время, в течение которого он хочет найти ответ.

Поисковая система состоит из следующих основных компонентов:

Spider (паук) - браузероподобная программа, которая скачивает веб-страницы.

Crawler (краулер, «путешествующий» паук) – программа, которая автоматически проходит по всем ссылкам, найденным на странице.

Indexer (индексатор) - программа, которая анализирует веб-страницы, скаченные пауками.

Database (база данных) – хранилище скачанных и обработанных страниц.

Search engine results engine (система выдачи результатов) – извлекает результаты поиска из базы данных.

Web server (веб-сервер) – веб-сервер, который осуществляет взаимодействие между пользователем и остальными компонентами поисковой системы.

Детальная реализация поисковых механизмов может отличаться друг от друга (например, связка Spider+Crawler+Indexer может быть выполнена в виде единой программы, которая скачивает известные веб-страницы, анализирует их и ищет по ссылкам новые ресурсы), однако всем поисковым системам присущи описанные общие черты.

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 компонент и работает напрямую с html-текстом страницы.

Crawler. 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.

Indexer. 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.

Database. База данных - это хранилище всех данных, которые поисковая система скачивает и анализирует. Иногда базу данных называют *индексом* поисковой системы.

Search Engine Results Engine. Система выдачи результатов занимается *ранжированием* страниц. Она решает, какие страницы удовлетворяют запросу пользователя, и в каком порядке они должны быть отсортированы. Это происходит согласно алгоритмам ранжирования поисковой системы. Эта информация является наиболее ценной и интересной для нас – именно с этим компонентом поисковой системы взаимодействует оптимизатор, пытаясь улучшить позиции сайта в выдаче, поэтому в дальнейшем мы подробно рассмотрим все факторы, влияющие на ранжирование результатов.

Web server. Как правило, на сервере присутствует html-страница с полем ввода, в котором пользователь может задать интересующий его поисковый термин. Веб-сервер также отвечает за выдачу результатов пользователю в виде html-страницы.

Все факторы, влияющие на положение сайта в выдаче поисковой системы, можно разбить на внешние и внутренние. Внутренние факторы ранжирования – это те, которые находятся под контролем владельца веб-сайта (текст, оформление и т.д.).

После наполнения сайта качественным контентом проводится **оптимизация** сайта под поисковики: необходимо уделять особое внимание не только подбору правильных ключевых слов, но и эффективному позиционированию их на страницах сайта. При этом желательно использовать потенциальные возможности всех элементов, каждый из которых вносит свой вклад в повышение релевантности страницы.

Общее правило таково – чем выше релевантность ключевого слова, тем выше позиция страницы в результатах, выдаваемых машиной поиска. Сейчас между поисковиками идет достаточно жесткая конкуренция, поэтому, чтобы выдавать ответы, наиболее релевантные запросам пользователей, они постоянно совершенствуют и усложняют алгоритмы поиска. Кроме того, позиция страницы в результатах поиска зависит от частоты использования ключевых слов.

Цель правильного размещения ключевых слов состоит в выборе метода их включения на страницу, увеличивающего релевантность страницы запросу пользователя.

Рассмотрим места, размещение ключевых слов в которых приведет к повышению эффективности оптимизации страницы сайта под поисковые машины:

1. тег Title заголовка страницы;
2. тег Description описания страницы;
3. тег Keywords ключевых слов;
4. теги заголовков (H1, …);
5. тексты ссылок;
6. содержимое страницы;
7. тег Alt альтернативного текста;
8. комментарии и т. д.

Если ключевые слова написаны на английском языке, то также эффективным будет размещение их в названиях рисунков и именах доменов.

Наиболее «уважаемым» поисковыми системами в HTML-коде страницы является тег Title, в начале которого обязательно должна быть ключевая фраза. Оптимизация содержания страницы без правильного подбора ее заголовка часто становится мало эффективной. Поисковые системы в результатах поиска по запросу используют текст заголовка страницы для оформления ссылки на сайт. Слова в заголовках имеют большой вес у пауков поисковых машин, а также в системах индексации в каталогах, что может серьезно повысить рейтинг страницы. Таким образом, правильно сделанный заголовок страницы - половина успеха. Ключевые слова вписывают в заголовок страницы согласно следующему принципу. В дополнение к ключевому слову или фразе вставляют и другие слова с учетом, что на одно ключевое слово или фразу из двух слов длина заголовка должна быть 3-5 слов.

Тег Description представляет собой краткое описание содержимого страницы, которое получает пользователь в результатах, выдаваемых поисковой системой на его запрос. Именно по этой информации

пользователь чаще всего принимает окончательно решение, переходить по ссылке на выданную страницу или нет. Те поисковики, которые при расчете релевантности учитывают тег Description, уделяют особое внимание содержанию в нем ключевых слов.

Т.о. текст тега Description должен быть осмысленным, кратким и четко характеризующим тематику страницы. В тег надо помещать максимальное количество ключевых слов, причем стараться составить из них предложения, которые будут выглядеть как связный текст.

Некоторые поисковые системы, которые не поддерживают тег Description, выводят в качестве описания страницы краткую аннотацию в виде 150-200 первых символов страницы. Т.о. необходимо тщательно продумывать текст, расположенный вверху страницы с точки зрения размещения в нем ключевых слов.

В тег Keywords вписывают ключевые слова, присутствующие в тексте страницы и имеющие прямое отношение к теме сайта. Для каждой страницы необходимо составлять свой собственный набор ключевых слов, наиболее характерный для описываемого текста. В тег не стоит включать служебные слова (предлоги, союзы и т.д.), повторять слова два и более раз, записывать слова во множественном числе. Очередность слов составляется по степени важности в порядке убывания. Слова достаточно отделять друг от друга пробелами: запятые здесь не нужны. Слова с прописной буквы не надо дублировать словами с заглавной, за исключением аббревиатур, личных имен, названий компаний, торговых марок и т.д. Обычно последовательность ключевых слов включает не более 250 символов. Избыточные символы просто не учитываются поисковыми системами.

На релевантность ключевого слова значительно влияет частота его использования в тэгах <H1>, <H2>..., применяемых для выделения заголовков и подзаголовков текста. Оптимальным вариантом в данном случае будет размещение ключевой фразы, форматированной как h1 или h2, в самом начале страницы. Здесь рекомендуется использовать только ключевое слово или ключевую фразу без добавления других слов.

Многие машины ныне постепенно преобразуются в тематические поисковики. Это значит, что они просматривают сайты целиком, включая содержимое страниц, а также учитывают входящие и исходящие ссылки. Т.е., когда паук исследует сайт, он ищет ссылки и, проползая по ним, анализирует их текст, а также тематику страницы, на которую осуществляется переход. Если в тексте ссылок используются ключевые фразы, то поисковая система считает, что ссылки согласованы с содержимым страницы, и релевантность такого документа повышается.

Следующим местом пристального внимания при позиционировании ключевых слов является непосредственно текст страницы. Начинать страницу необходимо с основного текста, содержащего максимум ключевых слов. Желательно избегать использования графики в самом начале страницы, т.к. она практически не несет никакой нагрузки с точки зрения повышения релевантности страницы.

Использование в коде страницы тэгов Alt, содержащих ключевые слова, применяется для повышения рейтинга страницы в листе ответов в некоторых поисковых системах. Тэг Alt применяется для отображения браузером альтернативного текста, назначаемого для изображений, на случай работы пользователя в режиме с отключенной графикой.

Наличие ключевых слов в тексте комментариев, во всплывающих подсказках (тэг Acronym), а также выделение их полужирным шрифтом может, хотя и незначительно, также повысить рейтинг страницы.

При анализе размещения ключевых слов в тексте сайта обычно используют такие понятия как, вес ключевого слова, плотность и положение на странице.

Вес ключевого слова – отношение частоты использования ключевого слова к общему количеству слов на индексируемой странице, выраженное в процентах. В общем случае, увеличение веса ключевого слова на странице ведет к повышению ее релевантности. Но существует предел, превышение которого расценивается как спам и ведет к исключению страницы из индекса. Идеальной считается плотность ключевых слов в пределах 3-5 %, но небольшие отклонения значительной роли не играют. Но не все поисковые системы воспринимают значение веса как критически важное. Например, всеми любимый Google не слишком придирчив к плотности ключевых слов при условии, что ключевая фраза достаточно часто упомянута на самой странице. Желательно стремиться к равенству веса ключевого слова своей страницы значениям, характерным для наиболее рейтинговых сайтов, использующих такую же ключевую фразу.

Плотность ключевого слова – это показатель, учитывающий, не только, сколько раз встречается ключевое слово на странице, но и, как часто оно используется в определенном объеме текста. “Перебор” по плотности может быть расценен поисковой системой как нарушение правил оптимизации, и страница может быть исключена из индекса.

Положение ключевого слова на странице учитывает, как близко к началу страницы находится заданное слово. Как правило, чем ближе к началу встречается слово запроса, тем выше релевантность страницы данному слову. Обычно размещают ключевые слова в самом начале страницы, упоминают их еще раз в первых абзацах, а также равномерно рассеивают остальные ключевые слова по всей странице.

Источники:

http://www.rupromo.ru/faq/relevantno/
http://rcdl2003.spbu.ru/~igor/papers/exp-survey/node5.html
http://megalib.com/books/797/webm_4.htm