ПОИСК ИНФОРМАЦИИ в ИНТЕРНЕТ

Информатика для СПО

Проблема поиска информации в Интернет

Активные пользователи Интернет тратят на поиск необходимой информации несколько часов в день, часто результаты этого поиска отзываются тщетными а более половины найденной информации признается бесполезной. Проблема заключается в том что информация разрознена. На статических сайтах большого размера требуется более эффективный поиск информации. Для динамических сайтов и порталов необходим быстрый поиск в большой коллекции документов, находящихся не только на различных сайтах но и на различных серверах.

Задача поиска информации

Задача поиска информации сплетается в сложный клубок задач, требуя выяснения: что представляет собой интересующая информация; как ее идентифицировать в запросе к системе, осуществляющей поиск; как его организовать; что делать с найденными результатами при различных механизмах поиска. Сегодня в ответ на большинство запросов информация, выдаваемая поисковыми системами, является неполной, несистематизированной, неверифицированной, загрязненной большим количеством спама. Релевантность поиска не растет или даже падает, пользовательский интерфейс не всегда удобен - бесконечная лента результатов поиска, где в одну кучу свалены форумы, сайты, товарные предложения, новости, бесконечные входные страницы, липовые сайты, фальшивые каталоги, платные ссылки. Поэтому ответом на запрос «океан» реально является ответ на запрос «названия фирм, магазинов со словом океан», «Мировой океан и другие океаны, как географические объекты». Такая же ситуация характерна для большинства популярных однословных и двухсловных нечетких запросов.

Потребность в решении задачи поиска информации

Необходим инструмент, позволяющий быстро и просто связывать работников с релевантной информацией, а также решить задачи быстрого получения точного результата при минимальном административном участии, с интеллектуальным пользовательским интерфейсом.

Поиск информации заключается в получении ответов на запросы:

- навигационные - показать место, где лежит информация;
- информационные - показать саму информацию;
- транзакционные показать, где можно купить товар.

Знания необходимые для эффективного полиска

Для эффективного поиска в Интернет необходимо знать:

- какие существуют машины поиска;
- как добываются ими сведения о ресурсах сети;
- какие свойства искомых ресурсов нужно сообщить машине;
- что будет результатом поиска;
- от чего зависит результат поиска и как его можно улучшить;
- какой должна быть технология эффективного поиска.

Средства для поиска информации в интернете

Поиск в Интернет осуществляется с помощью каталогов, поисковых машин и порталов, метапоисковых систем и коллекций ссылок. Объектами поиска являются отдельные страницы в любом формате, структурированные энциклопедии и словари, содержание БД, файлы (программ, изображений, музыки), рассылки, группы новостей, сообщения на форумах и собственно новости.

Типы поисковых машин

Можно выделить два типа поисковых машин. Первый тип образуют машины с **классифицированными списками ресурсов - Каталоги**. Пользователю предоставляется набор информации о ресурсах в сети в форме систематически организованных и связанных наборов сведении, каждый из которых может иметь достаточно сложную структуру. Создать каталог с минимальным первичным НАПОЛНЕНИЕМ просто, но раскрутить его до известности и постоянной аудитории очень сложно. Поэтому естественным источником пользователей Для таких «каталогов» опять же являются ПОИСковые системы.

Второй тип поисковых машин

ВТОРОЙ тип поисковых систем составляют машины, которые используют алгоритмы Поиска ресурсов выполняемого на основе ключевых слов. Запросные машины выдают пользователям информацию преимущественно В виде текста.

Различия между поисковыми машинами разных типов

Различие между машинами этих типов не является очень заметным, так как машины с Классифицированными списками сайтов допускают ПОИСК ПО запросам определенных видов. В свою очередь, запросные машины часто содержат классифицированные списки ресурсов, но более бедные по содержанию.

Акцент на пути, по которым пользователи получают информацию

Создатели Google первыми поняли, что изучать стоит не только тексты, но и пути, по которым пользователи приходят к этим текстам Важные документы быстро становятся известными в Интернет, у них высокий авторитет. На сегодняшний день крупные поисковые системы используют этот подход.

Механизм регистрации страниц в поисковых БД

Каждая машина поиска содержит БД, отображающую Содержание web-страниц. Для представления в этой БД страница или сайт должны быть зарегистрированы в поисковой машине. Процедура регистрации предоставляется любой поисковой службой, но большинство из них автоматически пополняют свои БД с помощью специальных агентов-роботов, которые регулярно посещают узлы и страницы, прослеживают ссылки, отыскивают новые или изменившиеся ресурсы и направляют полученные сведения в БД.

Механизм поиска – функция механизма обработки запросов

Разнообразие применяемых механизмов поиска основано на предлагаемом машиной многообразии способов обработки запросов. Используются ключевые слова, спецификации терминов, которые должны или не должны присутствовать в искомом материале, усечение запроса (внешнее и внутреннее), автоматическое порождение запроса по ключевым словам, поиск по точному или приблизительному совпадениям, поиск на основе выделения специализированных полей, поиск на основе ограничений значений.

Ранжирование результатов

Результаты поиска, полученные машиной, обычно ранжированы по степени связанности с содержанием и формой запроса. Для начальной части итогового списка иногда предоставляются дополнительные возможности сортировки по дате, узлам и др.

Методы поиска

Основное направление развития систем поиска идет в направлении создания средств для формирования того информационного пространства в Интернет, которое пользователь может обозреть и эффективно использовать, в противовес созданию средств поиска, которые дадут пользователю точный ответ на запрос. Реализация этого направления может быть достигнута использованием трех идей: персонализации поиска, обобщенного представления его результатов и создания метаданных для эффективной реализации поиска.

Персонализация поиска

Персонализация поиска по содержанию состоит в выявлении преимущественного интереса пользователя, формировании содержания предметной области, соответствующей этому интересу, и ,проведении поиска на множестве тех источников информации, которые наилучшим способом соответствуют этой области. Персонализация такого вида проявляется в современных машинах в виде ориентации на определенную предметную область или категорию пользователей а также в возможности задания при поиске некоторых предпочтений. Персонализация поиска по методам работы означает предоставление механизма, позволяющего изменять стратегию поиска в процессе его осуществления. Многие популярные поисковые машины уже сейчас формируют результаты, содержашие ссылку на страницу, предлагающую способы их улучшения. Эти рекомендации имеют общий характер и не учитывают содержание запроса. .

Обобщенное представление результатов поиска

Обобщенное представление результатов поиска подразумевает передачу пользователю укрупненного набора найденных ресурсов, полезность которого была бы ясна без проверки каждого ресурса по отдельности. В существующих машинах эти возможности проявляются, в частности, в сортировке результатов поиска по узлам или в поиске, осуществляемом в классифицированных коллекциях web-ресурсов, предварительно собранных машиной.

Создание базы метаданных

Создание базы метаданных означает индексацию web-ресурсов на основе управляемого предметного анализа и организацию доступа к индексированным ресурсам. Практически это означает создание шлюзов для различных предметных областей, наполнение их ссылками на тщательно отобранные ресурсы и составление описаний ресурсов (например, используя ключевые слова, классификационные обозначения) для Эффективного поиска и просмотра.

Структура поисковых систем

Поисковые системы состоят из пяти отдельных программных компонент:

- Spider (паук) это программа, которая скачивает Web-страницы. Она соединяется с Web-сайтом и загружает страницу. Паук не имеет никаких визуальных компонент. То же действие (скачивание) вы можете наблюдать. когда просматриваете некоторую страницу и когда выбираете в меню браузера раздел «просмотр HTML-Кода».

Структура поисковых систем

- Crewler (путешествующий паук) -он скачивает страницы, может идти по странице и находить все ссылки. Это его задача -определять, куда дальше должен идти паук, основываясь на ссылках или исходя из заранее заданного списка адресов. Spider и Crewler вместе составляют робот. Кроме роботов, например, у Яндекса есть несколько агентов, которые определяют, доступен ли в данный момент сайт или документ, на который стоит ссылка в соответствующем сервисе.

Структура поисковых систем

- Indexer (Индексатор) - разбирает страницу на различные ее части, анализирует их и составляют поисковый образ документов. Элементы типа заголовков страниц, подзаголовков, ссылок, текста, структурных элементов, элементов редактирования и других стилевых частей страницы вычленяются и анализируются.

Структура поисковых систем

- Database (БД) это хранилище поисковых образов всех скаченных, обработанных и заиндексированных документов.
- Search Engine Result Engine (поисковая машина) содержит словарь словоформ. Именно система выдачи результатов решает, Какие страницы удовлетворяют запросу пользователя. Это та часть поисковой системы, с которой осуществляется поиск.

Соответствие результата поиска ожиданием пользователя – главный критерий хорошей поисковой машины

Типы и объемы хранимых данных могут различаться (например, Google сохраняет части или целиком страницы, а Altavista - каждое слово каждой страницы). Все машины просматривает всю Сеть. Критически важно то, насколько ответ на запрос соответствует пользовательским ожиданиям, насколько достоверны предоставленные поисковой машиной ссылки; собственно, этим и отличается хорошая машина от плохой. Причина успеха Google - в ранжировании страниц; для этого используется не только PageRank, но и еще более полутора сотен критериев.

Критерии отбора результатов поиска

Несмотря на то, что поисковые системы сильно изменились, большинство до сих пор отбирают результаты поиска на основании примерно следующих критериев:

- Title (заголовок). Присутствует ли ключевое слово в заголовке?
- Domain/ url (Домен/адрес). Присутствует ли ключевое слово в имени домена или в адресе страницы?
- Stile (стиль) жирный (Strong или B), курсив (EM или !), заголовки HEAD. Есть ли место на странице, Где ключевое слово использовано в жирных, курсивных или текстовых заголовках?

Критерии отбора результатов поиска

- Density (плотность). Как часто ключевое слово употреблено на странице? Количество ключевых слов относительно текста страницы называется плотностью ключевого слова.
- MetaInformation (метаданные). Некоторые поисковые системы до сих пор читают метаключевые слова (meta keywords) и метаописания (meta description).
- Outbound Links (ссылки наружу). На кого есть ссылки на странице и встречается ли ключевое слово в тексте ссылки?

Критерии отбора результатов поиска

- Inbound Links (внешние ссылки). Кто еще В Интернет имеет ссылку на сайт‘? Каков текст ссылки? Это называется внестраничный критерий, потому ЧТО автор страницы не всегда может им управлять.
- Insite Links (ссылки внутри страницы). На какие еще страницы сайта содержит ссылки эта страница?

Механизм индексации страниц

Кода пользователь вводит запрос в поисковую систему, браузер обращается к DNS - серверу, а тот отправляет его на один из кластеров, наиболее близкий и наименее загруженным. После этого все операции проходят внутри этого кластера. Прежде чем страница будет проиндексирована, она проходит через ряд преобразователей. Определяется, какие слова встречаются в документе с какой частотой, на каких позициях, с какой гарнитурой и т.д. Прямой индекс содержит идентификаторы документов, каждому из которых соответствует список идентификаторов слов. При поиске же используется инвертированный индекс. В нем, наоборот. каждому слову соответствует идентификатор документа. Параллельно с индексом, например, google хранит полные текстовые копии страниц.

Объем индексной информации

Пройдя через аппаратный распределитель нагрузки, запрос пользователя попадает на один из web-серверов поисковой машины, который координирует действия всех других машин. Слова, введенные пользователем, перекодируются в идентификатор слова, и индексные серверы с помощью инвертированного индекса формируют список ID документа, отсортированный по релевантности. Или список слов через web-сервер передается архивным серверам. Они извлекают заголовки страницы и куски текста, которые содержат наибольшее количество ключевых слов и кажутся поисковику наиболее релевантными. Вся эта информация возвращается web-серверу, и он формирует выдачу - HTML-страницу с результатами, которая предъявляется пользователю, Несмотря на то что в результате индексирования размер сайтов уменьшается в разы, все равно в сумме получаются терабайты информации на индексных серверах.

Словарная и нечеткая морфология поиска

Существуют два типа морфологии поиска словарная и нечеткая. Типом морфологии поиска определяется алгоритм, по которому будет составляться индекс и выполняться поиск файлов в указанной области. От значения этого параметра зависят результаты поиска. Если выбрана словарная морфология, при поиске учитываются все грамматические варианты слов. Будут найдены документы, которые содержат фразы и слова запроса во всех грамматических формах («человека», «человеку», «люди», «людьми» и т. д.). Результаты поиска окажутся более точными, однако составление или обновление индекса займет довольно много времени. Если же выбрана нечеткая морфология, слова будут приведены к наиболее вероятной основе без учета грамматических форм. Использование нечеткой морфологии значительно уменьшает время индексации и поиска.

Особенности языков запросов машин поиска

Языки запросов различных машин поиска в основном являются сочетанием следующих функций: операторов булевой алгебры AND (И),OR (ИЛИ), NOT (НЕ); операторов расстояния ограничивают порядок следования и расстояния между словами, например: NEAR - второй термин должен находиться на расстоянии от первого, не превышающем определенного числа слов; FOLLOWED BY - термины следуют в заданном порядке; ADJ -термины, соединенные оператором, являются смежными.

Возможность усечения терминов

В языках запросов появилась возможность усечения терминов -использование символа «*» вместо окончания термина позволяет включить в искомый список все слова, производные от его начальной части (шаблона); учитывается морфология языка машина автоматически учитывает все формы данного термина, возможные в языке, на котором ведется поиск. Возможность поиска по словосочетанию, фразе; поиск ограничивается элементом документа (слова запроса должны находиться именно в заголовке, первом абзаце, ссылках и т. д.); можно ограничить выдачу по дате опубликования документа, на количество совпадений терминов; можно искать графические изображения; язык становится чувствительным к строчным и прописным буквам.

Этапы обработки результатов запроса

Результат запроса (список ссылок) обрабатывается в два этапа. На первом этапе производится отсечение очевидно нерелевантных источников, попавших в выборку в силу несовершенства поисковой машины или недостаточной «интеллектуальности» запроса.

Что такое метапоисковая система?

Метапоисковые системы – это машины, каталоги которых имеют список поисковых машин, ориентированных на обслуживание определенных потребностей. Пользователю метамашины либо предоставляют списки поисковых систем. либо позволяют направить запрос конкретной машине, либо дают возможность указать область поиска и свойства искомого ресурса. Результаты работы метамашин - это списки, которые являются либо смесью результатов от всех использованных машин, либо отдельными друг от друга рсзультатами, полученными каждом машиной.

Что такое метапоисковая система?

Метапоисковые системы не имеют собственных поисковых БД, не содержат никаких индексов и при поиске используют, ресурсы множества поисковых систем. За счет этого полнота поиска в таких системах максимальна и вероятность нахождения нужной информации очень высока.

Создатели метапоисковых систем не совсем оправдано надеются, что поисковые системы, которые они используют, возвращают релевантные результаты поиска, и слишком полагаются на позицию, на которой в данной поисковой системе находится документ.

Поисковые машины и поисковые системы

В таких системах анализ полученных описаний Документов не производится, что может поставить нерелевантные Документы, идущие первыми в одной поисковой системе, выше релевантных в другой, чем существенно понизить качество самого поиска. Этот принцип оказался хорошим ДЛЯ создания анализатора позиции сайта в поисковых системах, но в целом ДЛЯ систем метапоиска оказался неудовлетворительным. Созданы системы с возможностью выбора тех поисковых машин, в которых, по мнению пользователя, он с большей вероятностью может найти то, что ему нужно. Например, у метапоисковой машины Nigma есть восемь поисковых систем).

Пропускная способность канала связи – главное ограничение на работу системы метапоиска

Такой подход позволяет уменьшить используемые вычислительные ресурсы метапоискового сервера, не перегружая его большим объемом ненужной информации, и сэкономить трафик. В любой системе метапоиска наиболее узким местом является пропускная способность канала передачи данных, потому что затраты времени на обработку информации на порядки меньше времени доставки страниц, запрошенных у поисковых серверов.

Метапоисковые агенты

Для передачи запроса к поисковой системе используется специальный метапоисковый агент, который отвечает не только за процесс ретрансляции запроса и приема страниц, но и за то, чтобы запрос был передан в правильной кодировке, принятой в каждой из выбранных поисковых систем, иначе будет получен совершено другой набор описаний документов или не будет получен вовсе, что негативно скажется на качестве поиска. После обработки полученного запроса каждая система возвращает метапоисковому агенту множество описаний и ссылок на документы, которые считает релевантными данному запросу.