ЛЕКЦИЯ 1

«Технологии поиска тематической (профессиональной) информации в сети Internet»

План

1. Введение
2. Поисковые каталоги
3. Технология поиска информации:
   Поисковые машины
   Подборки ссылок
   Базы данных адресов
   Поиск медицинской информации

Введение

Сегодня Интернет объединяет множество разных сетей, миллионы компьютеров, около 300 миллионов пользователей всех континентов и, по разным оценкам, число таких пользователей увеличивается на 15-80% ежегодно. Можно выделить два основных направления в использовании Интернет. Это оперативный доступ к поистине необозримым кладезям информации по любой тематике (на сотнях тысяч информационных серверов), поиск и интерактивное общение с другими пользователями, практически любой специализации и географическом расположении. Как сориентироваться в столь масштабном информационном пространстве? Для этого существуют специализированные поисковые серверы. Их можно разделить на тематические каталоги, роботы индексов (поисковые машины), системы мета поиска.

Поисковые каталоги

Основная задача internet – предоставление необходимой информации. Чтобы найти нужную информацию необходимо знать адрес Web-страницы, на которой эта информация находится. Лучше всего искать в Сети необходимую информацию с помощью поисковых систем. Поисковая система представляет собой специализированный Web-узел. Поисковые системы классифицируют по методам поиска.

Поисковые каталоги предназначены для поиска по темам. Обычно они построены по иерархическому принципу, т.е. каждый шаг поиска это выбор подраздела с более конкретной тематикой искомой информации. На нижнем уровне поиска пользователь получает относительно небольшой список ссылок на искомую информацию.

Каталог Интернет-ресурсов – это постоянно обновляющийся и пополняющийся иерархический каталог, содержащий множество категорий и отдельных web-серверов с кратким описанием их содержимого. Способ поиска по каталогу подразумевает «движение вниз по ступенькам», то есть движение от более общих категорий к более конкретным. Одним из преимуществ тематических

каталогов является то, что пояснения к ссылкам дают создатели каталога и полностью отражают его содержание, то есть дает возможность точнее определить, насколько соответствует содержание сервера цели поиска.

Некоторые каталогина главной странице имеют тематический рубрикатор, с помощью которого пользователь попадает в рубрику со ссылками на интересующую его информацию.

Кроме того, некоторые тематические каталоги позволяют искать по ключевым словам. Пользователь вводит необходимое ключевое слово в строку поиска и получает список ссылок с описаниями сайтов, которые наиболее полно соответствуют его запросу. Стоит отметить, что этот поиск происходит не в содержимом WWW-серверов, а в их кратком описании, хранящихся в каталоге.

Информацию можно искать двумя путями:

- можно воспользоваться иерархическим деревом при поиске информации. Т.е. сначала выбирается общая тематика, удовлетворяющая запросу информации, и далее конкретизируете, следуя подсказкам каталога. В конечно результате будет получен список сайтов, содержащих информацию, соответствующую запросу.

- также можно пойти и по другому пути. Проанализировав предполагаемое содержание запрашиваемой информации, выбираются ключевые слова, которые обязательно встретятся в искомых материалах или их заголовках. Набираются эти слова через пробел в строке ввода на главной страницеи нажимаетсяEnter. Система попытается сама подобрать интересующую вас информацию.

Технология поиска информации

Сеть Интернет растет гигантскими темпами и найти информацию, необходимую конкретному пользователю, не очень просто. Но возможно, поскольку в сети есть ресурсы, которые помогут не утонуть в океане информации и новичку, и профессионалу.

Появление всемирной паутины WorldWideWeb стало количественным и качественным скачком в области информационных технологий. Число новых ресурсов и объем информации, которую они содержат, растет лавинообразно, увеличивается количество иголочек в информационном "стоге сена" и, соответственно, размер его самого. Для поиска информации в сети имеются следующие виды ресурсов:

– информационные порталы;

– каталоги интернет-ресурсов;

– поисковые системы.

Сама сеть Интернет постепенно превратилась в Средство Массовой Информации с огромной аудиторией пользователей во всем мире и невероятным объемом информации. Она стала глобальным средством информации, опутавшим каналами связи весь земной шар, но не поглотила привычные нам СМИ, они органически влились в сеть на правах самостоятельных информационных ресурсов. Практически каждая газета, радиостанция или телеканал в любой стране мира имеет свое представительство в сети Интернет.

Электронная версия газеты может и, как правило, сильно отличается от бумажной, значительно превышая ее по объему – формат данных, публикуемых на

интернет-сайтах более гибок, он не ограничен выделенными под материал страницами, газетными и журнальными колонками. Появляется элемент интерактивности – читатели могут оставить свои комментарии и отзывы о прочитанной статье, новости, аналитическом обзоре.

Поисковые инструменты - это особое программное обеспечение, основная цель которого – обеспечить наиболее оптимальный и качественный поиск информации для пользователей Интернета. Поисковые инструменты размещаются на специальных веб-серверах, каждый из которых выполняет определенную функцию:

1. Анализ веб-страниц и занесение результатов анализа на тот или иной уровень базы данных поискового сервера.
2. Поиск информации по запросу пользователя.
3. Обеспечение удобного интерфейса для поиска информации и просмотра результата поиска пользователем.

Приемы работы, используемые при работе с теми или другими поисковыми инструментами, практически одинаковы. Перед тем как перейти к их обсуждению, рассмотрим следующие понятия:

1. Интерфейс поискового инструмента представлен в виде страницы с гиперссылками, строкой подачи запроса (строкой поиска) и инструментами активизации запроса.
2. Индекс поисковой системы – это информационная база, содержащая результат анализа веб-страниц, составленная по определенным правилам.
3. Запрос – это ключевое слово или фраза, которую вводит пользователь в строку поиска. Для формирования различных запросов используются специальные символы ("", |, ~), математические символы (*, +, ?).

Схема поиска информации проста. Пользователь набирает ключевую фразу и активизирует поиск, тем самым получает подборку документов по сформулированному (заданному) запросу. Этот список документов ранжируется по определенным критериям так, чтобы вверху списка оказались те документы, которые наиболее соответствуют запросу пользователя. Каждый из поисковых инструментов использует различные критерии ранжирования документов, как при анализе результатов поиска, так и при формировании индекса (наполнении индексной базы данных web-страниц).

Таким образом, если указать в строке поиска для каждого поискового инструмента одинаковой конструкции запрос, можно получить различные результаты поиска. Для пользователя имеет большое значение, какие документы окажутся в первых двух-трех десятках документов по результатам поиска и на сколько эти документы соответствуют ожиданиям пользователя.

Наиболее популярные технологии поиска информации в Интернет.

Поисковые машины (searchengines)

Машины веб-поиска - это серверы с огромной базой данных URL-адресов, которые автоматически обращаются к страницам WWW по всем этим адресам,

изучают содержимое этих страниц, формируют и прописывают ключевые слова со страниц в свою базу данных (индексирует страницы).

Более того, роботы поисковых систем переходят по встречаемым на страницах ссылкам и переиндексируют их. Так как почти любая страница WWW имеет множество ссылок на другие страницы, то при подобной работе поисковая машина в конечном результате теоретически может обойти все сайты в Интернет.

Именно этот вид поисковых инструментов является наиболее известным и популярным среди всех пользователей сети Интернет. У каждого на слуху названия известных машин веб-поиска (поисковых систем) – Яndex, Rambler, Aport.

Чтобы воспользоваться данным видом поискового инструмента, необходимо зайти на него и набрать в строке поиска интересующее ключевое слово. Далее можно получить выдачу из ссылок, хранящихся в базе поисковой системы, которые наиболее близки запросу. Чтобы поиск был наиболее эффективен, нужно заранее обратить внимание на следующие моменты:

— определиться с темой запроса. Что именно в конечном итоге хочется найти?

— обратить внимание на язык, грамматику, использование различных небуквенных символов, морфологию. Важно также правильно сформулировать и вписать ключевые слова. Каждая поисковая система имеет свою форму составления запроса — принцип один, но могут различаться используемые символы или операторы. Требуемые формы запроса различаются также в зависимости от сложности программного обеспечения поисковых систем и предоставляемых ими услуг. Так или иначе, каждая поисковая система имеет раздел "Help" ("Помощь"), где все синтаксические правила, а также рекомендации и советы по поиску, доступно объясняются (скриншот страничек поисковиков).

— использовать возможности разных поисковых систем. Если не нашли на Яndex, попробовать на Google. Необходимо пользоваться услугами расширенного поиска.

— чтобы исключить документы, содержащие определенные термины, нужно использовать знак "-" перед каждым таким словом. Например, если нужна информация о работах Шекспира, за исключением "Гамлета", то вводится запрос в виде: "Шекспир-Гамлет". И для того, чтобы, наоборот, в результаты поиска обязательно включались определенные ссылки, нужно использовать символ "+". Так, чтобы найти ссылки о продаже именно автомобилей, Вам нужен запрос "продажа+автомобиль". Для увеличения эффективности и точности поиска, используйте комбинации этих символов.

— каждая ссылка в списке результатов поиска содержит сниппет – несколько строчек из найденного документа, среди которых встречаются ключевые слова. Прежде чем переходить по ссылке, необходимо оценить соответствие сниппета теме запроса. Перейдя по ссылке на определенный сайт, нужно внимательно посмотреть взглядом главную страничку. Как правило, первой страницы достаточно, чтобы понять – по адресу Вы пришли или нет. Если да, то дальнейшие поиски нужной информации нужно вести на выбранном сайте (в разделах сайта), если нет – необходимо вернуться к результатам поиска и попробовать очередную ссылку.

— следует помнить, что поисковые системы не производят самостоятельную информацию (за исключением разъяснений о самих себе). Поисковая система – это лишь посредник между обладателем информации (сайтом) и пользователем. Базы данных постоянно обновляются, в них вносятся новые адреса, но отставание от реально существующей в мире информации все равно остается. Просто потому, что поисковые системы не работают со скоростью света.

К наиболее известным машинам веб-поиска относятся Google, Yahoo, AltaVista, Excite, HotBot, Lycos. Среди русскоязычных можно выделить Яndex, Rambler, Апорт.

Поисковые системы являются самыми масштабными и ценными, но далеко не единственными источниками информации в Сети.

Подборки ссылок

Подборки ссылок – это отсортированные по темам ссылки. Они достаточно сильно отличаются друг от друга по наполнению, поэтому чтобы найти подборку, наиболее полно отвечающую Вашим интересам, необходимо ходить по ним самостоятельно, дабы составить собственное мнение.

Базы данных адресов (addressesdatabase)

Базы данных адресов – это специальные поисковые серверы, которые обычно используют классификации по роду деятельности, по выпускаемой продукции и оказываемым услугам, по географическому признаку. Иногда они дополнены поиском по алфавиту. В записях базы данных хранится информация о сайтах, которые предоставляют информацию об электронном адресе, организации и почтовом адресе за определенную плату.

Хочется отметить, что единой оптимальной схемы поиска информации в Интернет не существует. В зависимости от специфики нужной информации, можно использовать соответствующие поисковые инструменты и службы. А от того, как грамотно будут подобраны поисковые службы, зависит качество результатов поиска

Поиск медицинской информации

Описать ресурсы Интернета в какой-либо области - крайне сложная задача. Во-первых, они огромны и практически неисчерпаемы, а во-вторых, быстро обновляются и изменяются. Это базы данных, мультимедийные учебные серверы, виртуальные атласы и учебники, демонстрации клинических случаев, медицинские библиотеки, электронные версии журналов, описания научно-исследовательских проектов, программное обеспечение для обработки изображений и многое другое.

Необходимо отметить основные характеристики ресурсов Интернета, отличающие их от других источников информации. Любой пользователь Интернета может создать свою Web-страницу и наполнить ее любой информацией. При этом отсутствует какая-либо система контроля за содержанием (контентом) сайтов. Медицинский работник, пользующийся Интернетом, должен сам выбрать, каким Web-страницам он доверяет, какую информацию он может использовать для

принятия правильных клинических решений. Большинство сайтов, содержащих достоверную профессиональную информацию, предоставляет доступ к своим архивам за плату. В первую очередь это относится к сайтам медицинских журналов. Вместе с тем, многие Интернет-проекты, в частности существующие за счет грантовой поддержки, являются бесплатными. Для Интернета также характерно то, что наряду с процессом создания новых Web-страниц, происходит закрытие множества сайтов. Например, было установлено, что за 2 года, около 15% ссылок прекратили свое существование. Наконец, серьезную проблему представляет поиск необходимой информации в Интернете. Некоторые пользователи Интернета даже не представляют, что в Сети есть ответы на большинство интересующих их вопросов. Для того чтобы отыскать эти ответы, необходимо знать определенную методологию поиска, механизмы поиска и быть знакомым с основными информационно-поисковыми системами Интернета.

Поскольку Всемирная Паутина не принадлежит никому и развивается стихийно, она представляет собой сложную смесь самой разнообразной информации. Например, университеты и исследовательские институты предоставляют информацию о своих учебных и исследовательских программах и работающих в них специалистах. Специализированные центры дают доступ к поисковым базам данных. Правительственные и международные организации (например, ВОЗ) публикуют информацию о своих программах и проводимой политике. Коммерческие организации обеспечивают поддержку пользователей и публикуют в Сети массу рекламной информации. Профессиональные общества предоставляют информацию для своих членов. Частные лица размещают в Интернете детальное описание своей работы и интересов. И все в большей степени информационная индустрия использует Интернет для продажи доступа к информации, включая такие традиционные источники информации, как библиографические базы данных и журналы.

Поиск профессиональной медицинской информации в Интернете

Во все времена информация являлась одним из ключевых факторов, определяющих развитие медицины. Современные данные необходимы как в научной, так и в практической деятельности медика. Вопрос о поиске медицинских статей, руководств, изображений, программ встает при написании рефератов, дипломных работ при необходимости получения дополнительных знаний, при подготовке кандидатской и докторской диссертации. В настоящее время прекрасные возможности для реализации всего вышеперечисленного предоставляет сеть Интернет, с помощью которой возможно не только подобрать все необходимые материалы, но и сделать это с минимальной затратой времени и средств.

Говоря о поиске информации, в первую очередь подразумевается поиск текстовой информации и рассматриваются основные средства и методы работы с информационно-поисковыми системами (ИПС) и базами данных (БД). Подробные инструкции по работе с той или иной ИПС обычно нетрудно отыскать на конкретном сайте в разделе "Помощь" (Help). ИПС является, по сути, посредником

между пользователем, ищущим информацию, и БД, содержащей гиперссылки, рефераты статей и т.п.

В сети Интернет существуют два основных вида ИПС: классификационные и словарные. В первом случае вся информация, включаемая в базу данных ИПС (гиперссылки на сайты рефераты статей и т.д.), распределяется персоналом Web-сервера (систематизаторами) по заранее определенным категориям (например, "Иммунология", "Руководства", "Конференции" и т.д.). Пользователь подобной ИПС (например, Yahoo) выбирает интересующую его категорию и находит там ссылки на документы данной тематики. В основе словарной ИПС (например, Google) лежит перечень ключевых слов, формируемый компьютерной системой на основе проиндексированных документов, К каждому слову прилагается список документов, в которых это слово встречается, зачастую также с указанием позиции слова в тексте. Основным преимуществом словарных ИПС перед классификационными является возможность поиска ключевых слов не только в заголовках и аннотациях документов, включенных в базу данных, но и в содержании самих документов.

Согласно теоретическим основам поиска информации действия субъекта (т.е. пользователя ИПС) определяются имеющейся у него информационной потребностью, которая зачастую носит достаточно абстрактный характер (мы не всегда точно понимаем, что хотим, до тех пор, пока не выразим наши желания). Выражением информационной потребности является запрос, сформулированный с помощью профессиональных терминов, с использованием специального языка и синтаксиса, с учетом правил работы с конкретной ИПС. Современные системы еще не обладают искусственным интеллектом и не могут задать вам уточняющие вопросы, поэтому удовлетворяющий вашу информационную потребность ответ может быть получен только на очень точно сформулированный запрос. Соответствие ответа ИПС вашей информационной потребности обозначается термином "пертинентность", а соответствие ответа ИПС запросу - термином "релевантность".

Таким образом, релевантность отражает полноту поиска, а пертинентность - его точность. В большинстве случаев далеко не все релевантные документы являются пертинентными. Поэтому, получив ответ ИПС, насчитывающий несколько десятков тысяч ссылок (например, библиографических или гиперссылок), теоретически можно просмотреть их все или первые несколько сотен, являющиеся наиболее релевантными, но более рациональным решением будет уточнение запроса и повторный поиск. Результатом уточнения запроса должен оказаться либо перечень, состоящий из пары десятков наиболее пертинентных документов, либо перечень, насчитывающий около 70-80 документов, выявить среди которых пертинентные вы сможете путем простого просмотра заголовков и аннотаций.

Итак, цель ясна - получение пертинентной информации. Переходя от теории к практике, можно сказать, что для достижения этой цели необходимо ответить на следующие вопросы.

1) Каков предмет поиска? (Что?)

2) Какие существуют системы поиска? (Где?)

3) В чем заключается механизм поиска? (Как?)

При ответе на первый вопрос ключевое значение имеет **правильная формулировка запроса**. Необходимы ключевые слова, наиболее точно отражающие интересы, а также синонимы. Может оказаться полезным подбор слов по категориям, таким как "Заболевание", "Диагностика", "Лечение", "Единицы наблюдения". Нужно быть готовым к тому что количество результатов поиска может оказаться огромным, поэтому лучше заранее продумать такие моменты, как объекты наблюдения (люди и/или животные, мужчины и/или женщины, возраст), временной диапазон публикации статей, тип статей (обзор, клиническое наблюдение и т.д.). Тщательность проработки данного этапа даст 50% успеха и позволит сэкономить время и деньги при поиске.

После того как цель поиска сформулирована, возникает **проблема навигации в сети Интернет**. Для этого пользователь выбирает наиболее удобные для него сайты.

Переходя к третьему вопросу, **технологии поиска**, необходимо еще раз отметить важность предварительной проработки цели поиска и формулировки ключевых слов. Возможности поисковых систем достаточно широки. Умея ими пользоваться, можно значительно сократить затраты времени, составив такой запрос, в результате обработки которого будут найдены именно те статьи, которые вам необходимы.

В основе любого поиска лежит введение ключевых слов в специальную форму поиска. Основным принципом поиска является последовательное **уточнение запроса с помощью различных комбинаций ключевых слов**, операторов, знаков, опций и т.д. Обозначения различных функций в разных поисковых системах различны, но зная общие принципы построения поисковых запросов, легко сориентироваться в каждом конкретном случае. Дополнительную информацию вы можете получить из разделов "Help" представленных сайтов или из руководств по работе с системами поиска Большинство ресурсов, предоставляющих доступ к медицинским статьям, оснащены двумя вариантами поиска: простым (направлен на начинающих пользователей или используется для предварительного ознакомления с предметом поиска и ключевыми словами) и расширенным (профессиональный). В первом случае пользователю предлагается поле для ввода запроса, во втором дополнительные опции для указания даты публикации, имени автора, раздела, в котором будет производиться поиск (название, аннотация, текст статьи), номера первой страницы и т.д.

Из характеристики отдельных опций и операторов, приводимой ниже, становятся понятны различные варианты поиска статей: подбор всех статей по заданной тематике за определенный промежуток времени, поиск конкретной статьи, поиск статей определенного автора или из определенного журнала.

Поиск слов. Простой поиск слов осуществляется по всему тексту статьи или аннотации. Если вводить слово в форму буквами в нижнем регистре, то будут найдены слова и с прописными буквами (например, в результатах поиска «сердце» будут статьи, содержащие и «сердце», и «Сердце»). При написании слов с

прописными буквами будут найдены слова, точно соответствующие введенным. В некоторых системах ввод слов буквами в нижнем регистре без кавычек равнозначен всем словам, начинающимся с введенного (например, cancer даст и cancerous). Использование знаков усечения (* и $) позволяет не вводить часть букв слова с целью учета всех словоформ (например, "womSn" для поиска "woman" и "women") или в случае неуверенности в написании. Символ * заменяет несколько знаков в конце слова, символ $ -один или несколько знаков в определенной позиции. В некоторых системах существуют специальные ключи для орфографической проверки ключевых слов и обозначения акронимов и аббревиатур.

Поиск фразы. Для поиска конкретной фразы вводимые слова должны быть помещены в кавычки (например, "магнитно-резонансная томография").

Расширенный поиск. Существуют специальные формы поиска, с помощью которых можно осуществлять поиск слов по категориям: слова в названии (в результате будут подобраны документы, в которых обсуждается именно заданная тема), слова в тексте статьи (более широкий круг статей, касающихся не только заданной темы), автор (фамилия и инициалы), название журнала, том журнала, номер первой страницы, место работы автора, номер гранта. Используя данные опции, можно найти конкретную статью, зная лишь некоторые отдельные данные, почерпнутые, например, из доклада. Журналы могут идентифицироваться по полному или сокращенному названию.

Отображение результатов поиска. Все статьи, соответствующие критериям поиска, отображаются в виде списка, разбитого на блоки по 10-50 документов. Формат выдачи результатов обычно включает имя автора, название, источник, дату публикации, ссылки на текст аннотации и/или полного текста. Список может быть сортирован по дате публикации, по степени соответствия запросу, по наличию полного текста и т.д.

ВОПРОСЫ ДЛЯ УСТНОГО ОПРОСА

«Технологии поиска тематической (профессиональной) информации в сети Internet»

1. Перечислите известные Вам поисковые ресурсы
2. Назовите наиболее популярные технологии поиска информации
3. Назначение поисковых каталогов
4. Перечислите известные Вам поисковые системы
5. Опишите какую-нибудь поисковую систему