МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное
учреждение высшего профессионального образования
«Юго-Западный государственный университет»
(ЮЗГУ)

Кафедра информационных систем и технологий

УТВЕРЖДАЮ
Проректор по учебной работе
_____О.Г.Локтионова
14 мая _____2013г.

Инструментарий поиска информационных ресурсов

Методические указания
по выполнению практических работ
по курсу «Социальные проблемы информатизации»
для студентов специальностей 230400

Курск 2013

УДК 681.3(075)

Составитель: Л.А. Лисицин

Рецензент
Кандидат технических наук, доцент *Мельник Е.В.*

Инструментарий поиска информационных ресурсов

[Текст]: методические указания по выполнению практических работ / Юго-Зап. гос. ун-т; сост.: Л.А. Лисицин, Курск, 2013. 16 с.: таб. 1. Библиогр. с. 16.

Содержат сведения по приемам работы с широко распространенными поисковыми системами в сети интернет. Добавлены необходимые инструкции по работе в HTML формате и других средствах представления ресурсов. Материал ориентирован на практическую работу студентов в компьютерной среде.

Отражен порядок выполнения практических работ и правила оформления отчетов.

Методические указания соответствуют требованиям программы, утвержденной учебно-методическим объединением по специальностям «Информационные системы».

.

Текст печатается в авторской редакции

Подписано в печать Формат 60x84 1/16.
Усл.печ. л. __. Уч.-изд. л. __. Тираж 50 экз. Заказ . Бесплатно.
Юго-Западный государственный университет.
305040, г. Курск, ул. 50 лет Октября, 94.

Содержание

ИНСТРУМЕНТАРИЙ ПОИСКА ИНФОРМАЦИОННЫХ РЕСУРСОВ

Цель работы: Изучение принципов построения и функционирования поисковых средств мировых информационных ресурсов и практическое освоение инструментария поиска.

Поисковые машины

Все поисковые машины, предназначенные для сети Интернет, имеют более или менее схожие принципы работы. Компактные копии документов, известных серверам поисковых систем, хранятся на локальном диске. Каждый из поисковиков опрашивает свой внутренний каталог по ключевым словам или фразам, которые пользователь указывает при определении сценария поиска. Различие состоит лишь в объеме просматриваемой информации и алгоритме поиска, плюс, в наличии дополнительных сервисов (например, встроенных тематических каталогов).

Поиск ведется в базе локальной машины, а в ответ на запрос выдаются подходящие адреса во всех концах интернета. Безусловно, поисковая машина ведет постоянный опрос узловых адресов в сети, пополняя собственную базу данных. В значительной степени, доступность документа для поисковой системы зависит от его автора. В его власти использовать в гипертексте наиболее запрашиваемые ключевые слова и поместить документ на доступном для основных поисковых машин сервере.

Поскольку поисковые машины существуют в Интернете, в основном, за счет публикуемой рекламы, как правило, самые популярные системы поиска могут предоставить наилучшие возможности. Для рядового пользователя услуги поисковых серверов предоставляются бесплатно. Достаточно лишь указать адрес поисковой системы в рабочей строке браузера или обратиться к ней через каталог закладок.

В данной работе сообщается о некоторых наиболее известных международных поисковых системах, а также о ряде российских поисковых машин.

Поисковый сервер Yahoo!

Американский поисковый сервер [http://www.yahoo.com].

(Первая публикация в Сети: апрель 1994 года. Разработчики Дэвид Фило (David Filo) и Джерри Янг (Jerry Yang), Стенфордский Университет (США).

Имя "Yahoo!" можно перевести как "ура!" или как аббревиатуру "Yet Another Hierarchical Officious Oracle" (в свою очередь переводимую как "иная иерархия официальной истины").

До сегодняшнего дня Yahoo остается лидером по популярности среди поисковых систем Интернета в мире. Одним из главных достоинств является наличие встроенного многоступенчатого тематического каталога, опрашивающего крупнейшую в сети базу данных. Недаром среди персональных закладок многих пользователей можно обнаружить ссылки на подразделы каталога Yahoo. При опросе поисковая машина Yahoo обращается не только к собственному списку сетевых ресурсов, но и к серверам поисковой машины Alta Vista. Среди существенных недостатков Yahoo можно отметить игнорирование многих российских и израильских серверов, плюс, обилие устаревших ссылок.

Интерфейс поисковой системы Yahoo постоянно модифицируется и усовершенствуется, зона поиска все время расширяется. Возникают новые версии поисковой машины для людей различных возрастов. Создано множество национальных Yahoo-серверов. Печатается даже специальный журнал как в виртуальной, так и в глянцевобумажной версиях. Однако основные методы поиска остаются неизменными: пользователь или шаг за шагом уточняет область поиска, следуя указателям тематического каталога, или вписывает ключевые слова по стандартной схеме, указанной ниже.

Для полноценного поиска по ключевым словам необходимо выбрать меню Options (Возможности). После клика в данном меню перед пользователем возникает поисковое окно, в котором он может выбрать ареал поиска: по ресурсам Web (Yahoo!), среди групп новостей (Usenet) или по электронному адресу (E-mailadreses).

Пользователь может определить и само исследуемое пространство: внутренний каталог Yahoo (Yahoo Categories) или мировую паутину (Web Sites). Поиск по внутреннему каталогу будет сильно

ограничен, вероятно, для того чтобы не утомлять неквалифицированного пользователя лишней информацией.

Кроме всего прочего пользователь может запросить отображать среди результатов поиска информацию за определенный промежуток времени и установить порционность выводимых сообщений.

Ему также предоставляется возможность выбрать метод поиска:

1) "разумный" поиск (Intelligent default),
2) по точному соответствию фразы (An exact phrase match),
3) по соответствию всех слов (Matches on all words (AND)),
4) по соответствию одного из слов (Matches on any word (OR)),
5) по имени человека (A person's name).

Наибольшие затруднения, как правило, представляет "разумный" поиск. Поэтому уделим этому методу особое внимание. Необходимо уяснить лишь десяток несложных правил:

1) для начала поиска, после указания ключевого слова (на английском языке) необходимо нажать на меню Search (Поиск) или на клавишу Enter (Ввести) на клавиатуре,
2) если поиск ведется по одному ключевому слову, пробел после слова ставится лишь в том случае, если Вы желаете исключить из вероятного списка те документы, в которых к ключевому слову примыкают дополнительные знаки (например, знаки препинания),
3) при поиске по соответствию хотя бы одного из перечисленных слов достаточно отделить слова пробелами (например, best provider),
4) при поиске по соответствию всех слов необходимо перед вторым, третьим и т.д. словами поставить знак "" (например, best provider),
5) при необходимости исключить из общего списка документы содержащие некое слово, нужно использовать знак "-" (например, best -provider),
6) при поиске фразы рекомендуется использовать кавычки,
7) если Вы ищете слово, начинающееся с заглавной буквы, - возьмите его в кавычки (например, "Provider"),

8) для поиска по известному заголовку можно использовать дополнительный ключ "t:" (например, t:best),

9) для поиска слова среди доменных имен (URL) желательно указать дополнительный ключ "u:" (например, u:best),

10) если Вы сомневаетесь в правильности написания того или иного слова, - используйте значок "*" (например, pr*v*der).

Разобравшись со спецификой поиска в одной системе, пользователь без труда освоит любую другую поисковую машину.

Поисковая машина AltaVista

Лидер 1995-96 годов. Была создана в лабораториях одной из крупнейших компьютерных компаний Digital Equipment Corporation (DEC). В вольном переводе с итальяно-американского сленга имя AltaVista звучит, как "Там-за-горизонтом". С первых дней своего существования эта поисковая система была заявлена как безусловно наилучшая: использующая все безграничные ресурсы Web и позволяющая достичь максимальных скоростей поиска.

AltaVista представляет настоящий интерес для высокопродуктивного поиска (www.altavista.com) на 25 языках, среди которых иврит и русский. Поиск может вестись как на просторах Web, так и среди Usenet.

Существуют простой и усложненный методы поиска. Данная поисковая машина не предлагает пользователю поработать по тематическому каталогу. Он может использовать стандартные процедуры поиска, уже описанные для системы Yahoo, или изучить дополнительные команды (в AltaVista самая длинная командная строка).

По сути, к уже знакомым операциям добавляются несколько логических и синтаксических операций. Некоторые из них дублируют более простые операции. Полный список операций поиска в AltaVista содержится во вкладке Help (Помощь) в основном окне поисковой системы.

Название этого поискового механизма имеет неоднозначный перевод с английского: "экс-сайт" может быть воспринято как "terra incognita" (неведомое пространство) Интернета. За время существования с октября 1995 года (разработчик Стенфордский Универси-

тет), завоевал немалую популярность за счет совершенно нового подхода к алгоритму поиска.

Поисковая машина Excite

Поисковая машина Excite сама разбирается с путаницей слов: синонимов и омонимов, контекстов и скрытых смыслов. При выдаче результатов поиска Excite сопровождает их комментариями о точности совпадения с начальным запросом (до 100%). Однако, если подобная концепция поиска пользователя не удовлетворяет, можно воспользоваться обычной схемой поиска по ключевым словам.

Показательно, что за последний год Excite очень сильно изменил пользовательский интерфейс: появился прекрасный тематический каталог, предоставлена возможность обращения к локальным серверам Excite в странах Европы, при обращении к ссылке Power Search (Усиленный Поиск) Вы с удивлением обнаруживаете, что по умолчанию теперь предлагается поиск по ключевым словам, а не по фразам. Вероятно, алгоритм концептуального поиска, долгое время скрываемый от пользователей и конкурентов, не до конца оправдал себя. Тем не менее, при поиске научных статей по заранее известному названию или заголовку, люди чаще всего прибегают именно к этой поисковой машине.

Поисковая система HotBot

Разработка программисткой компании Inktomi и мультимедийного интерактивного журнала HotWired. Основной идеей системы HotBot является достижение максимального удобства при поиске информации за счет изначального определения ареала и метода поиска. На этапе подготовки к поиску пользователь может определить временной промежуток для искомой информации (от недели до двух лет со дня опубликования в Сети), континент и тип домена, установить режим вывода результатов поиска и многое другое.

Однако, эти достоинства могли бы остаться незамеченными, если бы на сервере HotBot не был размещен лучший на сегодняшний день тематический каталог сетевых ресурсов, позволяющий

пользователю воспользоваться услугами представленных в нем компаний.

Подробные комментарии и объяснения по работе с сервером HotBot можно найти по адресу help.hotbot.com.

Поисковый сервер Infoseek

Надежная система как для любительского, так и для профессионального поиска.

Поисковый сервер Infoseek (можно перевести как "ищущий информацию") существует с 1994 года. На сегодняшний день используются две версии: для глобальных и для локальных сетей. Поиск осуществляется поключевым словам (фразам) или по тематическому каталогу. Инициализация системы производится нажатием клавиши "Seek" (Найти). Основные достоинства: самая крупная база данных, собирающая информацию с локальных серверов от Бразилии до Голландии, плюс удачно реализованная возможность уточнять ареалпоиска после получения результата добавлением новых ключевых слов.

Обычно, используют Infoseek как последнее средство поиска, в случае, если другие поисковые системы не обнаружили нужной информации по интересующему вопросу. Почему? Потому что по стандартным запросам Infoseek выдает на несколько порядков больше информации, чем любая другая поисковая система

Поисковая машина Lycos

Даёт пользователю возможность без труда находить не только документы с упоминанием ключевых слов, но и графические и звуковые файлы по фрагменту имени файла. Позволяет также предельно локализовать область поиска и обладает хорошо структурированным каталогом.

Для поиска того или иного файла достаточно ввести его имя (с указанием типа файла или без оного) и нажать на кнопку Find (Найти) в окне браузера или на клавишу Enter (Ввести) на клавиатуре. При получении результатов поиска Вы видите не только имена

искомых файлов, но и адрес FTP-сервера, на котором данный файл хранится, с указанием конкретной папки. Это позволяет Вам воспользоваться для перекачки файла специально предназначенной для этого программой (например, CuteFTP).

Всего в Мире Интернета существует свыше 200 поисковых систем. Невозможно (да и не имеет смысла) изучить каждую из них до мелочей. Уже прочитанной информации достаточно для начального поиска.

Российские поисковые системы

Выделяют среди них 5: Rambler, Яndex [http://www.yandex.ru] , Следопыт, Апорт и Ау [http://www.au.ru] . Всех их отличает молодость, оригинальность решений (зачастую, - следствие бедности) и стремление во что бы то ни стало помочь русскоговорящему пользователю, не владеющему английским языком или просто желающему искать информацию на родном языке. Кроме того, в базах данных этих поисковых машин можно обнаружить документы, не доступные поисковым гигантам всемирной паутины.

Поисковая машина Rambler

Сам разработчик, Дмитрий Крюков, переводит название своей системы как "праздно шатающийся человек". На сегодняшний день Rambler является не только наиболее популярным, но и наиболее мощным поисковым механизмом в Русской Сети. Существуют две версии поисковой машины: русская и английская. Опрос проводится по более чем 2 миллионам документам и каждый день база пополняется тысячами новых материалов. Осенью 1997 года эта поисковая система была официально включена компанией Microsoft в русскую версию Internet Explorer 4.

Поиск осуществляется по стандартно-упрощенной схеме с возможностью использования логических операторов "+" и "-" для увеличения или уменьшения веса данного ключевого слова. Полное описание алгоритмов поиска можно найти по адресу www.rambler.ru/query.html.ru [http://www.rambler.ru/query.html.ru]

или после клика в строке "Запросы". Популярность системы Rambler объясняется публикуемыми результатами рейтингов различных узлов российской сети. Было введено тематическое ранжирование сайтов, что сделало результаты опросов (по частоте посещений данного узла) более реальными. Для ознакомления с ними поэкспериментируйте с кнопками "TOP100" и "Рейтинг сайтов" в левой части основной рабочей страницы поисковой системы.

Поисковая машина Яndex

Это разработка компании CompTek International по декларируемым задачам более всего напоминает американскую машину Excite. Та же забота об удобстве поиска. Клиент может просто вписывать целые фразы и доверять поиск системе после нажатия на кнопку "Найти!". Основным достоинством поисковой системы является учет русской морфологии и синтаксических связей. Предусмотрена

возможность уточнять запрос. Все это привело к включению Яndex в список поисковых систем под шапкой Microsoft Internet Explorer 4. Для более подробного ознакомления с особенностями этой поисковой машины достаточно нажать на кнопку "Помощь".

Система «Следопыт»

На примере системы "Следопыт" (любимого детища компании МедиаЛингва) разберемся с тем, что такое "метапоисковая" машина. Говоря кратко, это - машина-паразит. В лучшем смысле этого слова (для наглядности найдите в Глоссарии слово "Хост"). Такая машина исследует чужие базы данных. Так же, как Yahoo может искать внутри каталогов AltaVista, "Следопыт" просматривает каталоги той же AltaVista, плюс, европейской "искалки" EuroSeek, а также уже знакомых Excite, HotBot, Rambler и WebCrawler. Уже только этого было бы достаточно для упоминания "Следопыта" в книге. Но, кроме этого, данная система - еще и переводчик с русского на английский и обратно. При этом переводится лишь сам запрос, результат поиска выдается на языке оригинала.

Более подробную информацию можно найти на головной странице поисковой системы при условии, если выбранный Вами для работы браузер поддерживает нужные Java-скрипты. Упрощенную версию программы можно загрузить на свой компьютер с сайта www.medialingua.ru.

Система "Апорт!"

Разработка компании "Агама" при поддержке российского отделения одного из лидеров компьютерного рынка - "Intel", и Артемия Лебедева. Эта поисковая система, опрашивая свыше миллиона документов, позволяет не только переводить запросы с русского на английский и обратно, но и переводить результат поиска с английского на русский. Безусловно, переводится не весь документ, а лишь аннотация к документу. В противном случае процесс обработки результатов поиска мог бы безмерно затянуться. Кроме того, в поисковой системе "Апорт!" предусмотрено автоматическое исправление ошибок при составлении запроса.

Посиковая система "Ау!"

В сотрудничестве с системой "Апорт!" развивается с лета 1997 года поисковая машина фирмы "Роцит" - "Ау!". Эту поисковую машину отличает наличие хорошо структурированного, хотя и крохотного (всего несколько тысяч документов) тематического каталога.

Дополнительно следует почитать коллекцию поисковых систем и статью о внешних метапоисковых системах. URL-адреса основных средств поиска в Интернет. Формирование запроса, копирование содержимого www-страниц на свой компьютер.

Релевантность

От английского "Relevant" - относящийся к делу; означает соответствие найденного документа запросу пользователя поисковой системы

Релевантность (от англ. relevancy) - это степень соответствия документа запросу. Релевантность не является чем-то, что живет в документе само по себе. **Каждая поисковая система определяет релевантность документа запросу пользователя в соответствии с заложенным в нее алгоритмом**. И, хотя алгоритмы у всех разные, ищут поисковые машины примерно одинаково, так как алгоритмы построены на общих принципах. Основные отличия поисковых машин заключаются не в алгоритмах определения релевантности, а в способах их реализации.

Связывание данных

В упрощенной форме связывание представляет собой отношения между данными, поставляемыми объектом источника данных, и HTML-потребителем данных. Данное отношение называется **связью**, потому что значение элемента данных (который называется datem, что является сокращением от data item - элемента данных) синхронизировано между клиентом и сервером. Когда HTML-потребитель данных (например, текстовое окно HTML) модифицирует элемент данных, то модифицированный элемент данных сохраняется в объекте источника данных. Напротив, если объект источника данных изменяет значение данных, то модифицированный элемент данных отправляется потребителю данных. Путем дальнейшего обобщения многочисленные потребители могут быть связаны с одним элементом данных, и все значения всех потребителей будут синхронизированы со значением, указанным объектом источника данных. Значения в объекте источника данных связаны со значениями в одном или большем количестве потребителей данных.

Доступны два различных стиля связывания: связывание текущей записи (current record binding) и связывание таблицы с повторением (repeated table binding). Связывание текущей записи использует HTML-элементы для отображения данных из текущей записи в наборе записей. В качестве текущей могут быть установлены различные записи. В таком случае элементы обновляются динамически для отображения данных в записи. Связывание таблицы с повторением позволяет определить набор связанных элементов, называемых шаблоном, который повторяется один раз для каждой записи в наборе записей. Разработчики Web-страниц также имеют возмож-

ность ограничения числа записей, повторяющихся в таблице. Этот элемент называется разбиением таблицы.

Связывание данных

Таблица 1.

Запрос	Кол-во найденных документов	Оценка релевантности
Широта охвата поисковых систем	14300 – Google, 1543070 - Яндекс	Горазда релевантнее поисковая система Google
Релевантность	2810000 - Google, 1 095 487 - Яндекс	Из-за неопределенности запроса трудно определить релевантность
Запрос 1		
…		
Запрос n		

Задание 1. Определить широту охвата и релевантность по запросам (не менее 5 запросов по 10 поисковым системам), занести данные в таблицу (согласно примерам), провести оценку релевантности.

Задание 2. Произвести приемы простого поиска информации, использование знаков + и -, применение джокера, контекстного поиска (для поисковых систем, где такие команды доступны).

При выполнении приемов простого поиска информации показать роль прописных букв, поиск по заголовкам, поиск web-узлов, поиск URL-адресов, поиск ссылок.

Осуществить поиск средствами расширенного поиска: OR, AND, NOT, NEAR, вложением команд.

Контрольные вопросы.

1. Сколько поисковых систем существует в интернете?
2. Какие алгоритмы поиска вам известны?
3. Какие дополнительные сервисы предлагают поисковые системы?
4. Назовите российские поисковые системы.
5. Назовите характерные сервисы российских поисковых систем.
6. Назовите лидеров среди российских поисковых систем.
7. Что такое расширенный поиск?
8. Какие команды используют для точного поиска?
9. Как правильно формировать запрос?
10. Что такое релевантность?
11. Как произвести оценку релевантности?
12. Что такое охват?
13. Что такое связывание данных?
14. Как воспользоваться услугой переводчика при поиске информации?

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Информатика. Базовый курс / под ред. С.В. Симоновича. 2-е изд. СПб.: Питер, 2011. 640 с.

2. Финкель, Е. Интернет для профессионалов / Р. Аллен, П. Гралла. СПб.: Питер, 2010. 153 с.