Занятие 5.
Теория поиска информации

Большинство современных молодых людей имеют мобильные телефоны; подавляющее большинство из них не знает, как работает сеть мобильной связи.

Поисковыми системами можно пользоваться так же, как мобильными телефонами — освоив нехитрые элементы управления ими, научившись нажимать нужные кнопочки. Для того чтобы уметь пользоваться поисковой машиной (или поисковой системой), не обязательно понимать, как она работает, и тем более знать ее внутреннее устройство.

Однако ситуация здесь примерно такая же, как с автомобилем: на нем можно ездить, умея лишь заправлять его, управлять им и зная правила дорожного движения. Но тот, кто хочет чувствовать себя за рулем уверенно и при необходимости самостоятельно устранять мелкие неисправности, должен знать устройство автомобиля, названия и принципы функционирования хотя бы важнейших узлов. Поэтому далее мы будем говорить об устройстве поисковых машин и основных терминах, которые приступающий к поиску информации в интернет пользователь должен знать так же хорошо, как автомобилист термины «карбюратор» и «зажигание».

Как работают поисковые машины

Поисковая машина (для краткости ее часто называют просто поисковик) представляет собой комплект программ, в основе которого **лежат следующие пять:**

- **Spider** («паук») — программа, которая загружает в поисковую машину Web-страницы. Работает аналогично браузеру, установленному на компьютере пользователя, но ничего не отображает ни на каком экране. Если вы хотите иметь представление о том, что именно загружает в поисковую систему «паук», откройте какую-нибудь Web-страницу и выберите в меню Вид браузера пункт Просмотр HTML (или «исходного») кода.
- **Crawler** («червяк», или «путешествующий паук») — программа, способная найти на Web-странице все ссылки на другие страницы. Ее задача — определить, куда дальше должен ползти «паук», руководствуясь ссылками или заранее заданным списком адресов.
- **Indexer** (индексатор) — программа, которая «разбирает» страницу на составные части и анализирует их. Вычленяются и анализируются заголовки Web-страниц, заголовки документов, ссылки, текст документов, отдельно — текст, выделенный полужирным шрифтом, курсивом и т.д.
- **Database** (база данных) — хранилище всех данных, которые поисковая система загружает и анализирует. Требует огромных ресурсов как для хранения, так и для последующей обработки.
- **Search Engine Results Engine** (система выдачи результатов поиска) решает, какие страницы удовлетворяют запросу пользователя и в какой степени. Именно с этой частью поисковой системы «общается» пользователь.

Первые две программы, работающие «в связке», часто называют *поисковый робот* (а иногда — *HTTP-робот).*

Как видите, поисковая машина, получив запрос на поиск, не отправляется в длительное путешествие по «Всемирной паутине, как полагают некоторые пользователи, а анализирует лишь ту информацию, которую собрала ранее. С одной стороны, это позволяет резко повысить скорость обработки запроса на поиск. С другой, ограничивает область поиска внутренними ресурсами поисковой системы, которые, во-первых, ограниченны (ни одна поисковая машина не в состоянии загрузить в свою базу данных информацию со всех узлов Сети), во-вторых, уже в какой-то степени устарели. Ситуация в интернет изменяется очень быстро. Если «паук» с целью обновления информации об уже проиндексированных однажды Web-страницах «заползает» на них раз в два месяца, пользователь рискует получить в результатах запроса ссылку на уже несуществующую Web-страницу.

Процесс загрузки из Сети информации и предварительного анализа ее поисковой машиной называется *индексация,* а сама база данных поисковой машины, в которой хранится собранная информация, — *индекс.*

Глубина индексации может быть разной. Полные тексты документов, размещенных на странице, в базу данных копируют не все поисковые роботы — некоторые ограничиваются лишь заголовками. Когда пользователь формирует запрос на поиск, поисковая машина просматривает свою базу данных и выдает перечень Web-страниц, содержащих слова, введенные пользователем в поле ввода (их часто называют *ключевые слова).* Таких страниц может быть очень много. Задача поисковой машины — отобрать те из них, которые в наибольшей степени отвечают запросу пользователя (т.е. *релевантны* ему) и указать ссылки на них в числе первых.

Дисциплина «Поиск информации в интернет» появилась совсем недавно (первые поисковые машины — около десяти лет назад), терминология еще не устоялась, поэтому не удивляйтесь, если обнаружите, что в какой-то статье или книге *автоматическим индексом* (или даже просто *индексом)* называют саму поисковую машину, а состоит она только из двух частей: поискового робота и базы данных. Используемая нами терминология также не является общепринятой и в ближайшем будущем может быть частично заменена другой.

Высокая скорость поиска обеспечивается не только за счет того, что поисковая машина обращается к уже собранной и хранящейся тут же, у нее «под рукой», информации. Анализируя собранные данные, поисковая машина выполняет индексацию базы Данных, в процессе которой каждому слову ставятся в соответствие его «координаты» — номер документа, в котором имеется данное слово, а зачастую и позиция слова в документе (номер предложения и номер слова в нем).

Аггоритмом поиска можно назвать метод, руководствуясь которым поисковая машина принимает решение, включать или не включать ссылку на страницу либо документ в результаты поиска.

Почти каждая поисковая машина использует свой собственный алгоритм поиска, и его детали представляют собой ноу-хау разработчиков поисковика. Но большинство из них отбирают документы, отвечая сами себе на вопросы:

- Присутствует ли ключевое слово в заголовке документа?
- Присутствует ли ключевое слово в имени домена или в адресе страницы?
- Встречается ли ключевое слово в подзаголовках документа либо в элементах текста, выделенных полужирным, курсивом либо как-то иначе?
- Как часто ключевое слово встречается на странице? (Долю ключевых слов в тексте страницы иногда называют *плотностью ключевого слова.)*
- Встречаются ли ключевые слова в описаниях страниц, выполненных их разработчиками, и среди ключевых слов, указанных разработчиками страниц? (Поскольку очень часто разработчики Web-страниц с целью привлечения к ним внимания лукавят при их описании и выборе ключевых слов, данным критерием пользуются не все поисковики.)
- На какие Web-узлы имеются ссылки на анализируемой странице и встречается ли ключевое слово в тексте ссылки?
- Какие Web-узлы имеют ссылку на анализируемый сайт? Каков текст ссылки? (Это так называемый *внестраничный критерий,* потому что автор страницы не всегда может им управлять.)
- На какие еще страницы данного сайта содержит ссылки анализируемая страница?

Как видите, поисковая система должна провести довольно детальный анализ каждой страницы, информацию о которой она заносит в свою базу данных.Мы привели лишь очень краткое описание того, как работает поисковая система, но для нашей книги этого более чем достаточно. В следующем разделе мы поговорим о возможных алгоритмах поиска более подробно.

Алгоритмы поиска

Как уже говорилось, применяемые поисковиками алгоритмы являются нх ноу-хау. Тем не менее о некоторых закономерностях. которые используются при разработке алгоритмов и предшествующему их применению анализу текста, поговорить стоит.

Некоторые из этих закономерностей быт» подмечены Джорджем Зипфом (George К. Zipf); он опубликовал свои законы в 1949 году. Пять лет спустя знаменитый математик Беноит Мандлеброт (Benoit Mandlebrot) внес небольшие изменения в формулы Зипфа. добившись более точного соответствия теории практике. Хотя некоторые исследователи и подвергают исследования Зипфа острой критике, без учета подмеченных им закономерностей сегодня не способна работать ни одна система автоматического поиска информации.

Зипф заметил, что длинные слова встречаются в тексте реже, чем короткие (по-видимому, это как-то связано с природной ленью человека и вообще

любого живого существа). На основе этой закономерности Зипф вывел два закона.

Первый из них связывает частот) появления того или иного слова в каком-то тексте (она называется частота вхождения слова) с рангом этой частоты.

Если к какому-либо достаточно большому тексту составить список всех используемых в нем слов, а затем проранжировать эти слова — расположить их в порядке убывания частоты вхождения в данном тексте и пронумеровать в возрастающем порядке, — то для любого слова произведение его порядкового номера в этом списке (ранга) и частоты его вхождения в тексте будет величиной постоянной

В математике такая зависимость отображается гиперболой. Отсюда, в частности, следует, что, если наиболее распространенное слово встречается в тексте 100 раз, то следующее по распространенности встретится не 99 и не 90, а примерно 50 раз (статистика не гарантирует точных цифр).

Это также означает, что самое популярное слово в английском языке (the) употребляется в 10 раз чаще, чем слово, стоящее на десятом месте, в 100 раз чаще, чем сотое, и в 1000 раз чаще, чем тысячное.

Значение вышеупомянутой постоянной в разных языках различно, но внутри одной языковой группы она остается неизменной. Так, например, для английских текстов постоянная Зипфа равна приблизительно 0,1. Для русского языка постоянная Зипфа равна *примерно 0,06-0,07.*

Второй закон Зипфа констатирует, что частота и количество слов, входящих в текст с этой частотой, связаны между собой. Если построить график, отложив по одной оси (оси X) частоту вхождения слова, а по другой (оси Y) — количество слов, входящих в текст с данной частотой, то получившаяся кривая будет сохранять свои параметры для всех без исключения созданных человеком текстов.

Зипф считал, что его законы универсальны. Они применимы не только к текстам. В аналогичную форму выливается, например, зависимость между количеством городов и чистом проживающих в них жителей. Характеристики популярности узлов интернет также отвечают законам Зипфа.

Многие исследования показывают, что законам Зипфа подчинены также и запросы работников различных организаций к Web-пространству. Следовательно, работники чаще всего посещают небольшое количество сайтов, при этом достаточно большое количество остальных Web-ресурсов посещается лишь один-два раза.

С другой стороны, каждый Web-сайт получает большую часть посетителей, пришедших по гиперссылкам из небольшого количества сайтов, а из всего остального Web-пространства на него приходит лишь небольшая часть посетителей. Таким образом, объем входящего трафика от ссылающихся Web-сайтов также подчиняется распределению Зипфа.

Не исключено, что в законах отражается «человеческое» происхождение объекта.

Джон Клайнберг из Корнеллского университета первым предложил способ фильтрации информации, позволяющий выявлять наиболее актуальные для каждого конкретного момента времени проблемы, обозначенные в текстах. Этот способ базируется на анализе больших объемов текстовой информации. Когда происходит какое-либо важное событие, о нем начинают активно писать, что приводит к своеобразным «скачкам» в частоте употребления тех или иных слов.

Клайнберг разработал алгоритм, позволяющий анализировать частоту использования того или иного слова, т.е. выполнять ранжирование слов по частоте вхождения. На выходе алгоритм представляет собой рейтинг слов, на основании которого можно делать выводы о популярности той или иной темы и производить сортировку информации.

Чтобы испытать свою разработку, ученый решил проанализировать тексты всех президентских докладов о положении в США (State of the Union addresses) начиная с 1790 года. В итоге получилось, что в период Войны за независимость американских колоний часто употреблялись слова militia («ополчение») и British («британский»), а в период с 1947 по 1959 годы наблюдался «скачок» в использовании слова atomic («атомный»). Таким образом, ученому удалось доказать работоспособность системы.

Как поисковые машины могут использовать законы Зипфа?

Для того чтобы ответить на этот вопрос, воспользуемся первым законом Зипфа и построим *график* зависимости ранга от частоты. Как уже упоминалось, его форма всегда *примерно* одинакова.

Можно предположить, что наиболее значимые для текста слова лежат в средней части представленного графика. Оно и понятно: слова, которые *встречаются слишком час*то, — это предлоги, местоимения и т.д. (в английском, немецком и некоторых других языках — еще и артикли). Редко встречающиеся слова также в большинстве случаев не несут особого смыслового значения, хотя иногда, наоборот, весьма важны для текста (об этом будет сказано чуть ниже). Каждая поисковая система решает, какие слова отнести к наиболее значимым, по-своему, руководствуясь общим объемом текста, частотными словарями и т.п. Если к числу значимых слов будут отнесены слишком многие, важные термины будут забиты «шумом» случайных слов. Если диапазон значимых слов будет установлен слишком узким, за его пределами окажутся термины, несущие основную смысловую нагрузку.

Для того чтобы безошибочно сузить диапазон значимых слов, создается словарь «бесполезных» слов, так называемых *стоп-слов* (а словарь, соответственно, называется *стоп-лист).* Например, для английского текста стоп-словами станут артикли и предлоги the, a, an, in, to, of, and, that... и др. Для русского текста в стоп-лист могли бы быть включены все предлоги, частицы и личные местоимения: на, не, для, это, я, ты, он, она и др.

Исключение стоп-слов из индекса ведет к его существенному сокращению и повышению эффективности работы. Однако некоторые запросы, состоящие только из стоп-слов (типа «to be or not to be»), в этих случаях уже не

пройдут. Неудобство вызывают и некоторые случаи полисемии (многозначности слова в зависимости от контекста). Например, в одних случаях английское слово «can» как вспомогательный глагол должно быть включено в список стоп-слов, однако как существительное оно часто несет большую содержательную нагрузку.

Но поисковая машина оперирует не с одним документом, а с их огромным количеством. Допустим, нас интересуют статьи академика Вернадского. Если бы поисковая машина оценивала частоту вхождения слова «Вернадский» по вышеописанному алгоритму, эта частота была бы близка к нулю, названное слово не вошло бы в число значимых и документы, содержащие это слово, упоминались бы в конце результатов поиска (а документы-аутсайдеры ни один нормальный пользователь не просматривает). Чтобы такого не произошло, поисковые машины используют параметр, который называется *инверсная частота термина*. Значение этого параметра тем меньше, чем чаще слово встречается в документах базы данных. На основе этого параметра вычисляют весовой коэффициент, отражающий значимость того или иного термина. Часто встречающееся слово (например, слово *иногда)* имеет близкий к нулевому весовой коэффициент, слово же Вернадский — напротив, весьма высокий.

Современная поисковая машина может вычислять весовые коэффициенты слов с учетом местоположения термина внутри документа, взаимного расположения терминов, морфологических особенностей термина и т.п. В качестве терминов могут выступать не только отдельные слова, но и словосочетания. Такого рода «математический анализ» позволяет поисковой машине с высокой точностью распознать *суть* текста.

Пространственно-векторная модель ПС

Базы данных поисковых машин могут быть устроены по-разному. Один из вариантов — пространственно-векторная модель. Она позволяет получить результат, хорошо согласующийся с запросом даже в том случае, если в найденном документе не оказывается одного или нескольких введенных пользователем ключе-вых слон, но при этом его (документа) смысл все же соответствует запросу. Такой результат достигается благодаря тому, что все документы базы данных размещаются в воображаемом многомерном пространстве (с размерностью выше трех, представить которое весьма трудно). Координаты каждого документа в этом пространстве зависят от содержащихся в нем терминов (от их весовых коэффициентов, положения внутри документа, от «расстояния» между терминами и т.п.). В результате оказывается, что документы с похожим набором терминов располагаются в этом пространстве поблизости. Получив запрос, поисковая система удаляет лишние слова, выделяет значимые термины, вычисляет вектор запроса в пространстве документов и выдает ссылки на документы, попавшие в определенную область пространства.

В пространственно-векторной модели термины «взаимодействуют» друг с другом, что повышает релевантность найденных документов запросу пользователя. Поисковая машина, работающая в соответствии с такой моделью, лучше воспринимает запросы на естественном языке, чем машина,

воспринимает запросы на естественном языке, чем машина, использующая более привычную «матричную» модель (в которой просто составляется матрица «термины-документы»; если в документе упоминается какой-то термин, в матрице проставляется число, учитывающее его весовой коэффициент, не упоминается — ставится ноль).

Схема работы каждой поисковой системы держится в секрете. Выше мы в весьма упрощенной форме изложили лишь основы алгоритма работы поисковой системы. В реальности механизм индексации и структура базы данных ПС значительно сложнее. Но и сказанного вполне достаточно для того, чтобы при формулировке запросов вы старались выбирать слова, наиболее точно характеризующие предмет поиска. Впрочем, о точности и полноте поиска мы более подробно поговорим в следующем разделе.

Полнота и точность поиска

Если бы интеллект поисковой машины был сравним с человеческим, в результате поиска мы получали бы несколько документов, содержащих исчерпывающую информацию о предмете поиска. К сожалению, это (пока) не так, и в результатах запроса обычно фигурируют сотни документов, не имеющих отношения к тому, что мы на самом деле хотели получить. Называются такие документы *нерелевантными.*

Релевантность

Итак, *релевантным* (от англ. relevant подходящий, относящийся к делу) называется документ, имеющий отношение к сделанному вами запросу, т.е. содержащий нужную нам информацию.

Следует отмстить, что обсуждение понятия релевантности в контексте информационно-поисковых систем ведется уже около полувека, но его конкретного общепринятого определения все еще нет.

По-разному дают определение релевантности и словари. Так, «Экономический словарь», расположенный на сайте www.km.ru, считает, *что релевантность* — это смысловое соответствие между информационным запросом и полученным сообщением. Поисковый узел Яндекс (www.yandex.ru) трактует этот термин как меру соответствия результатов поиска задаче, поставленной в запросе *(что,* в общем-то, эквивалентно определению «Экономической) словаря»).

Но иногда этому термину дают несколько расширенное толкование. Так, в одной из статей на том же поисковике Яндекс творится, что «при поиске в интернет важны две составляющие— полнота (ничего не потеряно) и точность (не найдено ничего лишнего). Обычно это все называют одним словом — релевантность». Другими словами, релевантность — это опять-таки соответствие отпета вопросу, но с учетом таких понятий, как полнота и точность поиска.

Коэффициенты полноты и точности

Коэффициентом полноты поиска (или просто *полнотой поиска)* называют отношение количества полученных релевантных результатов к общему ко-

личеству существующих в поисковом массиве документов, релевантных данному поисковому запросу.

Коэффициент точности поиска (или просто *точность поиска)* — это отношение количества релевантных результатов к общему количеству документов, ссылки на которые содержатся в ответе ПС.

В реальных ПС коэффициент полноты поиска может достигать значений 0,7-0,9, а коэффициент точности обычно находится в пределах 0,1-1,0.

Иногда при оценке эффективности ПС используют и другие критерии — *коэффициент потерь информации* и *коэффициент поискового шума.*

В идеальной ПС *коэффициент потерь информации* = 0, а *коэффициент поискового шума* = 1. В реальности эти коэффициенты совсем другие.

Нередко количество размещенных в Сети документов, релевантных запросу пользователя, достигает десятков и сотен тысяч. Вместе с тем содержащаяся во многих из них релевантная информация совпадает, и пользователю достаточно изучить лишь несколько документов из числа найденных. Таким образом, при непрофессиональном поиске не требуется высокое значение **коэффициента полноты**, который даже при успешном поиске вполне может быть близок к нулю. Следовательно, этот коэффициент в данном случае является второстепенным критерием качества информационного поиска.

Рассмотренные выше параметры интуитивно понятны и наглядны. Хотя с научной точки зрения для объективной оценки возможностей ПС их недостаточно (точнее, эти параметры следует заменить другими), для целей нашей книги большего и не нужно.

Морфологический анализ

Если вы захотите найти документы, содержащие слова «мальчик вошел в лес» и введете их в поле запроса, многие ПС включат в результаты поиска и документы, содержащие, допустим, слова **«Мальчик вошел** в тень **леса», «Мальчик,** найденный в лесу охотниками», и т.д. Поисковая машина способна сама решить весьма непростую проблему словоизменения.

Поиск, при котором учитываются словоизменения, называется *морфологическим.* Его способны осуществлять все русскоязычные и многие зарубежные ПС. Когда мы вводим в поле запроса слона «мальчик пошел в лес», мы, скорее всего, хотим найти документы, содержащие *все четыре* слова. Однако поисковая машина, представив вначале документы со всеми словами, начнет затем давать ссылки на документы, в которых есть *хотя бы одно* из указанных нами ключевых слов. Существуют способы точно указать поисковой машине, как должны быть связаны между собой введенные пользователем ключевые слова. Для этого используются логические операторы, а сам поиск называется *булев.* Более подробно об операторах мы поговорим далее.

Этапы поисковой процедуры

Процедура поиска имеет вполне определенную этапность — от определения информационной потребности и области поиска до анализа результатов и выбора пертинентных объектов.

Приведем еще одну аналогию, которая относится к шахматному искусству. Начало шахматной партии — дебют — обеспечивает развитие фигур на доске и определяет стратегическую канву будущей партии. Несмотря на то что шахматы допускают миллиарды последовательностей ходов, количество дебютов, на самом деле, ограничено несколькими сотнями. Точно так же, как в шахматном искусстве, в искусстве поиска можно определить первый этап — дебют. На этой фазе определяется цель поиска, его стратегия и область проведения (поисковые серверы, каталоги, тематические порталы).

Информационные потребности пользователя могут относиться к разным областям, которые могут быть как узкоспециализированными, так и достаточно типовыми. На практике основная часть информационных потребностей приходится именно на типовые области применения:

- поиск отдельных Web-страниц;
- поиск новостей;
- поиск людей и организаций;
- поиск литературных произведений;
- поиск программного обеспечения;
- поиск музыкальных произведений;
- поиск графических изображений;
- поиск видеоинформации;
- поиск коммерческой информации.

Вторым этапом в шахматах является миттельшпиль. При хорошо разыгранном дебюте и определенной стратегической направленности партии, наибольшее значение на этом этапе уделяется многовариантному анализу и тактическим решениям. В этом случае шахматист-профессионал просчитывает в уме несколько десятков вариантов (из миллионов возможных). Лишние неэффективные варианты он просто не рассматривает, руководствуясь логическими образами, заложенными на уровне подсознания.

Точно так же вторая, оперативная, часть поисковой процедуры предполагает многовариантность подходов и решений при формализации запросов в процессе их отработки. В этом случае также аналитик-профессионал приходит к необходимости использования весьма ограниченного числа поисковых серверов, каталогов и отдельных web-ресурсов для решения своей задачи.Основной задачей второго этапа является формирование эффективных запросов к ИПС. Наибольшую проблему при формировании запросов представляет то, что на каждом поисковом сервере используется свой информационно-поисковый язык (ИПЯ), несмотря на то что у различных языков этого типа много общего, — например, схожий набор булевых операций. В настоящее время не существует

единого стандарта, подобного стандарту языка SQL для СУБД, хотя на протяжении многих лет ведутся попытки такой стандартизации.

Последняя часть шахматной партии — эндшпиль — заключается в поиске вариантов при очень ограниченном количестве ресурсов (фигур). В этом случае количество вариантов, как правило, значительно более скромное, чем на втором этапе, и их правильный выбор определяет результат всей партии.

Точно так же третий этап поиска в сети Internet является определяющим, — от его реализации зависит, будет ли найденное решение пертинентно. На этом этапе пользователь работает с конечными документами, полученными в виде отклика ИПС. От правильного выбора набора документов-первоисточников зависит результат работы всех трех этапов поисковой процедуры.

Советы по поиску в интернете

Советы по поиску в интернете взяты с сайта ПС Яндекс, поэтому все перечисленные советы напрямую относятся к этой ПС. В других ПС некоторые советы могут не работать.

Проверяйте орфографию

Если поиск не нашел ни одного документа, то вы, возможно, допустили орфографическую ошибку в написании слова. Проверьте правильность написания.

Если вы использовали при поиске несколько слов, то посмотрите на количество каждого из слов в найденных документах (перед их списком после фразы «Результат поиска»).

Какое-то из слов не встречается ни разу? Скорее всего, его вы и написали неверно.

Используйте синонимы

Если список найденных страниц слишком мал или не содержит полезных страниц, попробуйте изменить слово. Например, вместо «рефераты» возможно больше подойдет «курсовые работы» или «сочинения».

Попробуйте задать для поиска три-четыре слова-синонима сразу. Для этого перечислите их через вертикальную черту (|). Тогда будут найдены страницы, где встречается хотя бы одно из них. Например, вместо «фотографии» попробуйте «фотографии | фото | фотоснимки».

Ищите больше, чем по одному слову

Слово «психология» или «продукты» дадут при поиске поодиночке большое число бессмысленных ссылок. Добавьте одно или два ключевых слова, связанных с искомой темой. Например, «психология Юнга» или «продажа и покупка продовольствия». Рекомендуем также сужать область вашего вопроса. Если вы интересуетесь автомобилями ГАЗа, то запросы «автомобиль Волга» или «автомобиль ГАЗ» выдадут более подходящие документы, чем «легковые автомобили».

Не пишите большими буквами

Начиная слово с большой буквы, вы не найдете слов, написанных с маленькой буквы, если это слово не первое в предложении. Поэтому не набирайте обычные слова с Большой Буквы, даже если с них начинается ваш вопрос Яндексу.

Заглавные буквы в запросе рекомендуется использовать только в именах собственных. Например, «группа Черный кофе», «телепередача Здоровье».

Ищите без морфологии

Вы можете заставить Яндекс не учитывать формы слов из запроса при поиске. Например, запрос !иванов найдет только страницы с упоминанием этой фамилии, а не города «Иваново».

Ищите похожие документы

Если один из найденных документов ближе к искомой теме, чем остальные, нажмите на ссылку «найти похожие документы». Ссылка расположена под краткими описаниями найденных документов. Яndex проанализирует страницу и найдет документы, похожие на тот, что вы указали. Но если эта страница была стерта с сервера, а Яндекс еще не успел удалить ее из базы, то вы получите сообщение «Запрошенный документ не найден».

Используйте знаки «+» и «-»

Чтобы исключить документы, где встречается определенное слово, поставьте перед ним знак минуса. И наоборот, чтобы определенное слово обязательно присутствовало в документе, поставьте перед ним плюс. Обратите внимание, что между словом и знаком плюс-минус не должно быть пробела. Например, если вам нужно описание Парижа, а не предложения многочисленных турагентств, имеет смысл задать такой запрос «путеводитель по парижу -агентство -тур».

Плюс стоит использовать в том случае, когда нужно найти так называемые стоп-слова (наиболее частотные слова русского языка, в основном это местоимения, предлоги, частицы). Чтобы найти цитату из Гамлета, надо задать запрос «+быть или +не быть».

Попробуйте использовать язык запросов

С помощью специальных знаков вы сможете сделать запрос более точным. Например, укажите, каких слов не должно быть в документе, или что два слова должны идти подряд, а не просто оба встречаться в документе.

О языке запросов мы поговорим подробнее на следующем занятии.

Сохранение информации из интернета

Самая главная операция любого пользователя интернета – сохранение найденной информации.

Итак, сохранение документа с помощью меню броузера.

Имеют значение два обстоятельства: тип броузера, в каком виде вы хотите сохранить документ.

Microsoft Internet Explorer позволяет сохранить документ как:

- web-страницу полностью (со всеми иллюстрациями, которые разместятся в отдельной папке, что довольно удобно);
- web-архив (с включенными иллюстрациями);
- web-страницу, один файл (без иллюстраций, только HTML);
- текстовый файл (только текст документа).

Вы можете также указать кодировку страницы.

1. В меню Файл выберите Сохранить как.
2. Дважды щелкните папку, в которую хотите поместить страницу.
3. В поле Имя файла введите соответствующее имя.
4. В поле Тип файла выберите тип файла.

Чтобы сохранить все файлы, необходимые для отображения данной страницы, включая рисунки, кадры и таблицы стилей, выберите вариант *Веб-страница, полностью.* В этом случае сохранятся все файлы в соответствующих форматах.

Чтобы сохранить всю информацию, необходимую для отображения данной страницы, в виде одного файла в кодировке MIME, выберите вариант *Веб-архив*. Выбор этого типа файла приведет к сохранению всей активной web-страницы.

Чтобы сохранить только активную HTML-страницу, выберите *Веб-страница, только HTML.* Выбор этого типа файла приведет к сохранению информации, содержащейся на web-странице, но при этом не сохранятся рисунки, звуковые эффекты и прочие файлы.

Чтобы сохранить только текст, содержащийся на активной веб-странице, выберите *Только текст.* Выбор этого типа файла приведет к сохранению информации, содержащейся на web-странице, в обычном текстовом формате.

Opera позволяет сохранить документ как:

- HTML-файлы (без иллюстраций, только HTML);
- HTML-файлы с рисунками (со всеми иллюстрациями, которые разместятся в той же папке, что и документ);
- текстовый файл (только текст документа).

В случае сохранения файлов других типов (doc, ppt, pdf и т.д.) броузер автоматически начнет «скачивание» файла после Вашего подтверждения.

Существуют и специальные утилиты для «скачивания» из интернета (ReGet).

Они могут решать, например, такую проблему как восстановление перекачки после обрыва связи.

В случае, если Вы ищете информацию в разных документах, будет оптимально использовать любой текстовый редактор (MS Word, например) для копирования информации из web-страниц.

Принцип работы: найденную информацию на web-странице Вы выделяете в броузере, копируете в буфер обмена, открываете текстовый редактор, вставляете из буфера текст.

Использованная литература:

1. Гусев В.С. Google: эффективный поиск. Краткое руководство. – М.: «Вильямс», 2006.
2. Ландэ Д.В. Поиск знаний в INTERNET. Профессиональная работа.: Пер. с англ. – М.: «Вильямс», 2005.
3. Язык запросов. Как искать? Помощь Яндекса. http://www.yandex.ru/search/?id=481939