ИНТЕРНЕТ

Посов Илья Александрович

ЭФФЕКТИВНЫЙ ПОИСК В ИНТЕРНЕТЕ

По материалам тренинга Ю.М. Лифшица «Эффективный поиск в Интернете» [1]

Говорят, что Аристотель был последним человеком, который знал современную ему культуру в полном объеме. Сейчас время изменилось, и единственное, что может претендовать на подобную осведомленность, это всемирная сеть Интернет. Интернет содержит немыслимое количество знаний и, что не менее важно, умеет искать среди них нужные. Чтобы задать Интернету вопрос, используются поисковые системы. Современные поисковые системы хорошо справляются со своими задачами, но, к сожалению, сейчас они недостаточно интеллектуальны, чтобы говорить с человеком на его языке. На современном этапе человек должен понимать образ мыслей поисковых систем, чтобы уметь ими пользоваться. Далее мы расскажем о поиске в Интернете, об устройстве поисковых систем, о приемах поиска и о стандартных ошибках, которые совершают пользователи. Одну стандартную ошибку можно назвать прямо сейчас – пользователь никогда не найдет ничего в Интернете, если не примет решение начать поиск. К Интернету стоит чаще обращаться с вопросами: круг возможных вопросов практически неограничен.

...только пользователь может определить, нашел ли он ту информацию, которую хотел.

РЕЛЕВАНТНОСТЬ

Это слово означает соответствие найденной информации введенному запросу. Релевантность, безусловно, субъективна, только пользователь может определить, нашел ли он ту информацию, которую хотел. И даже для пользователя эта задача может быть сложной, если он ищет что-то, сам не зная что.

Поисковая система ищет страницы, содержащие введенные в запросе слова. «Смысл» фразы в запросе никаким образом не анализируется. Запрос «хочу почитать про ТРИЗ» будет искать не те страницы, на которых можно почитать про ТРИЗ, а страницы, на которых написано «хочу почитать про ТРИЗ». Это упрощает задачу поисковой системы – не надо анализировать смысл запроса. К тому же, при таком понимании поиска становится понятно, как придать релевантности конкретное числовое значение. Для вычисления релевантно-

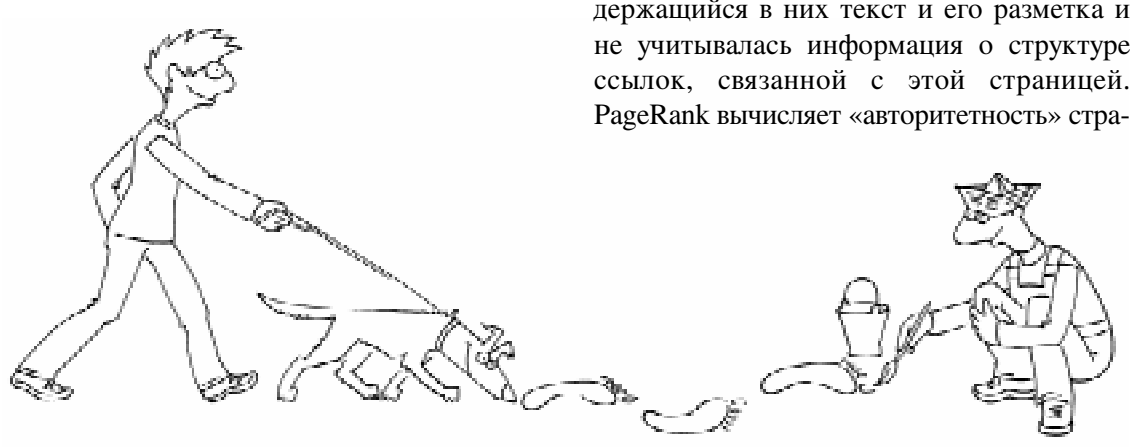
сти страницы обычная поисковая система использует следующие характеристики:

- *Наличие слов на странице.* Чем больше слов запроса присутствует на странице, тем больше релевантность. В идеале все слова должны присутствовать на странице.
- *Частота слов.* Релевантность тем больше, чем чаще слова запроса встречаются на странице. Частоту можно комбинировать с обратной частотой – насколько редко слово встречается во всем Интернете. Наиболее ценны те слова, которые много встречаются в тексте страницы, но в общем случае используются редко.
- *Форматирование слов запроса на странице.* Особенно хорошо, если искомые слова оказываются в заголовке или просто выделены по отношению к остальному тексту (жирным, курсивом).
- *Близость слов запроса друг к другу.*
- *Соответствие тематик сайта и запроса.* Тематику текста можно определять, например, таким методом. Выбрать нескольких рубрик: «автомобили», «компьютеры», «погода» и т. п. Каждой рубрике приписать список слов, которые ей соответствуют. Для рубрики «автомобиль» это будут слова: покрышки, аккумулятор и т. п. Теперь, если в документе преобладают слова некоторой рубрики, можно считать, что он к ней относится. Ключевые слова для рубрик можно определять автоматически. Достаточно взять набор документов, про которые достоверно известно, что они относятся к некоторой рубрике, и выделить в этом наборе часто повторяющиеся слова.

Особенно просто определить рубрику сайта, если он находится в каталоге, связанном с поисковой системой.

- *Регистрация в каталоге, связанном с поисковой системой.* Yahoo, Yandex, Google и другие поисковые системы имеют свои каталоги сайтов.
- *Количество ссылок, ведущих на страницу.* Этот показатель говорит о ее популярности. Можно анализировать и текст входящих ссылок, но поисковая система должна делать это аккуратно. Автор страницы не отвечает за все ведущие на нее ссылки, и недоброжелатель может специально давать ссылкам текст, не относящийся к содержанию страницы.
- *Качество ссылок, ведущих на страницу.* Это один из ключевых критериев, определяющих релевантность. Страница не просто должна содержать слова из запроса – она должна быть «авторитетным» источником информации. Авторитет страницы тем больше, чем больше других авторитетных страниц на нее ссылается. Впервые этот критерий качества страниц появился в поисковой системе Google, авторы назвали его PageRank, и дальше мы посмотрим на него подробнее.

В ноябре 1997 при запросе собственного названия только одна из четырех ведущих поисковых систем выдавала себя в первой десятке [2]. Такое низкое качество результатов поиска авторы Google Сергей Брин и Лэрри Пэйдж в своей статье [2] объясняли тем, что для определения релевантности страниц учитывался только содержащийся в них текст и его разметка и не учитывалась информация о структуре ссылок, связанной с этой страницей. PageRank вычисляет «авторитетность» стра-



Знание этого алгоритма позволит обманывать поисковую систему...

ницы. Он не зависит от введенного поискового запроса, и, таким образом, каждой странице Интернета соответствует свой PageRank. Его можно посмотреть, например, с помощью находящихся в Интернете сервисов, которые отображают PageRank по введенному адресу страницы.

Несколько слов о вычислении PageRank. PageRank страницы соответствует вероятности, что случайным образом блуждающий по сети пользователь в заданный момент времени окажется на этой странице. Случайное блуждание по сети означает следующее поведение пользователя. Когда он находится на очередной странице, то с вероятностью d (параметр алгоритма, выбирается близким к 1) он переходит по случайной ссылке из этой страницы, а с вероятностью $1-d$ ему становится «скучно» и он переходит на любую случайную страницу сети. Как вариант, он переходит не на случайную страницу сети, а на один из нескольких заранее заданных авторитетных сайтов, откуда и продолжает случайное блуждание.

Все тонкости вычисления PageRank, а также подробности, как именно описанные выше критерии релевантности собираются в одно число, являются коммерческой тайной поисковой системы. Знание этого алгоритма позволит обманывать поисковую систему, так что она будет выдавать неправильные результаты поиска. Правда, обманывать поисковую систему можно и без этих знаний – например, еще недавно запрос в Google «Враг Народа» первым результатом выдавал страницу действующего президента России. Вообще, поисковая система борется с теми, кто занимается нечестным продвижением сайтов. Если она видит, что кто-то искусственно завысил для своей страницы, например, количество ведущих на нее ссылок (или любой другой описанный выше критерий), то поисковая система может просто выбросить страницу и не совершать по ней поиск.

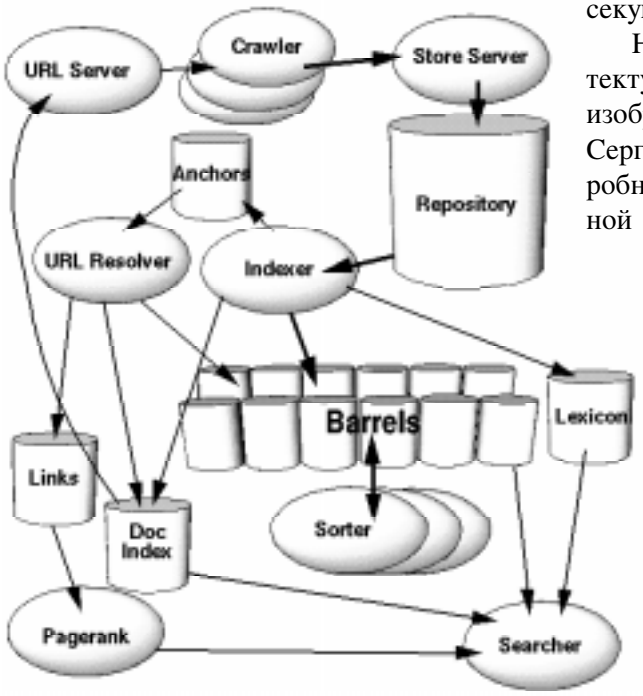


Рис. 1

АНАТОМИЯ ПОИСКОВОЙ СИСТЕМЫ

Для эффективного поиска полезно понимать, как устроена поисковая система и как именно определяются страницы, соответствующие введенному пользователем запросу. Вообще-то, примерное устройство поисковых систем достаточно хорошо известно пользователям Интернета, и сейчас никого не удивляет, что поиск страницы во всей огромной сети занимает десятые доли секунды.

На рис. 1 приведено изображение архитектуры поисковой системы Google. Это изображение взято из статьи [2], в которой Сергей Брин и Лэрри Пэйдж впервые подробно описали устройство крупномасштабной поисковой системы.

Поисковая система состоит из трех основных компонентов – это краулеры, база данных и обработчик запросов. Краулеры (другое название «пауки») – программы, которые блуждают по Интернету, сохраняют все найденные страницы в базу данных и заполняют индексы, которые потом будут использоваться для поиска. Реальный поиск совершается не по страницам Интернета, а по сохраненной информации. Это означает, например, что вы можете не обнаружить нужные Вам слова, на

найденной поисковой системой странице. За время, прошедшее от последнего посещения ее краулером, она могла измениться. Такое может происходить, например, с форумами, где страницы часто меняются из-за посылаемых сообщений. Но в общем случае подобная ситуация редкая. Можно добавить: чтобы обойти весь Интернет, краулерам необходимо всего (или «целых») 2–3 недели.

Обработчик запросов совершает поиск, он ищет страницы, содержание которых соответствует запросу. Для этого он смотрит в «обратный индекс» и для каждого слова запроса находит список страниц, которые это слово содержат. Далее отбираются страницы, содержащиеся во всех списках, то есть те страницы, которые содержат все слова запроса. Отобранные страницы обрабатываются дальше, например, чтобы проверить, как близко слова запроса находятся внутри каждой страницы (см. критерии релевантности). Окончательно страницы выводятся пользователю в порядке уменьшения их релевантности.

ТЕХНИЧЕСКИЕ АСПЕКТЫ ПОИСКА

Начать можно с технических аспектов поиска, небольших советов про то, как и чем пользоваться. Первый совет – использование возможностей браузера по загрузке сразу нескольких страниц. Если после нажатия на кнопку «поиск» вы выберете и откроете какую-то одну ссылку, то, скорее всего, и найдете всего один результат. Пробуйте сразу открыть все ссылки, которые, как кажется, соответствуют цели поиска, и пусть они все вместе загружаются. Впоследствии вы по ним пройдетесь и изучите внимательно.

Каждую ссылку удобно открывать не в отдельном окне браузера, а на отдельных вкладках (например, для браузеров Opera и Mozilla новая вкладка создается сочетанием клавиш Ctrl+T). Старые версии Internet Explorer не умеют работать с вкладками, но можно поставить расширение, которое это позволит.

Найденные страницы полезно сохранять на случай, если их впоследствии захочется изучить подробнее. Стандартный метод для этого – использование функции «Избранное», которая есть во всех браузерах. Если избранное почему-либо неудобно, можно пользоваться сервисами в Интернете, например, http://del.icio.us. Это сайт, который дает возможность хранить и обмениваться ссылками. В браузеры можно встроить кнопки, при нажатии на которые ссылка на текущую страницу сразу сохраняется в вашей учетной записи на сайте del.icio.us. Вам необходимо предоставить только несколько слов (тегов), описывающих сохраняемую ссылку. При удачном использовании технология присвоения ссылкам тегов может оказаться удобнее распределения ссылок по каталогам избранного.

Следующий технический аспект – использование поисковых форм в браузере. Это некоторое окно, похожее на окно для ввода адреса страницы, но в него вводится запрос к поисковой системе. Поиск совершается сразу же, нет необходимости загружать главную страницу поисковой системы. В браузерах Opera, Mozilla, Internet Explorer 7.0 окно поиска уже встроено. В других случаях его можно установить отдельно как расширение. Пример такого расширения – поисковая панель от Google. Помимо окна поиска она содержит полезные функции, как например отображение PageRank просматриваемой страницы.

Поисковое окно, встроенное в браузер, можно настраивать, меняя указания, в каких поисковых системах оно способно совершать поиск. Добавьте в

Функция «расширенный поиск».

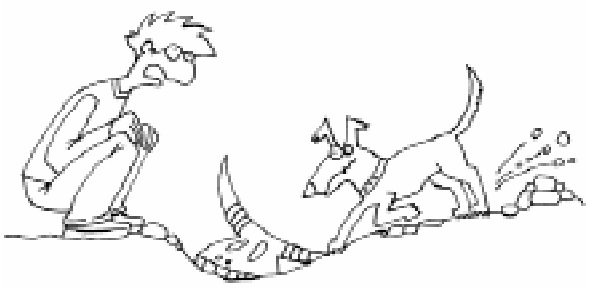
ваш браузер возможность искать во всех поисковых системах, которыми вы привыкли пользоваться.

Чем еще можно пользоваться при поиске? Функцией «расширенный поиск». Действительно, возможности искать страницы, содержащие заданный набор слов, может быть недостаточно. Иногда нужен другой поиск. Самый простой пример, добавить в поиск условие – найти страницы, *не* содержащие заданного слова. В Google и Yandex подобный запрос выглядит так: «Windows–95», (перед 95 стоит знак «–»). Будут искаться страницы, которые содержат слово Windows, но не содержат слово 95. Оформление расширенных запросов меняется в каждой поисковой системе, так же как меняется сам набор возможностей расширенного поиска. Удивительно, что совсем недавно Yandex не понимал запроса «Windows–95». Работу расширенного поиска в каждой поисковой системе нужно изучать отдельно.

Помимо того, что условия расширенного поиска можно вводить в строку запроса, есть возможность просто перейти на страницу расширенного поиска. Переход находится обычно рядом с полем для ввода текста запроса. Там уже с помощью ряда полей и переключателей можно выбрать, что и как мы хотим найти.

Некоторый набор возможностей следует выделить отдельно. Правда, они есть не во всех поисковых системах.

- *Поиск страниц, ссылающихся на данную.*
- *Поиск страниц, похожих на данную.* Имеется в виду, похожих по содержанию.
- *Использование двойных кавычек.* Запрос ««белеет парус одинокий»» ищет страницы, содержащие ровно эту фразу. Более того, слова в кавычках при поиске не будут склоняться, как это происходит в обычном случае. Например, запрос «идет» находит страницы, содержащие слова «идем», «иди» и т. п. На это способны не все поисковые системы. И особенно, если они иностранные. Иностранные наверняка не знают морфологию русского языка. Даже Google научился ей только недавно.
- *Региональный поиск* (например, только петербургские сайты). Аналогично можно сузить область поиска до страниц какого-то определенного сайта.
- *Поиск по каталогу.* Поисковые системы имеют свои каталоги сайтов, и помимо этого в Интернете есть много самостоятельных каталогов с сайтами.
- *Воскрешение «умерших» страниц.* Как видно из описания процесса поиска, поисковой системе совершенно необязательно хранить страницы в том виде, в котором она нашла их в Интернете. Достаточно хранить слова, которые содержатся на странице. Но некоторые поисковые системы хранят страницы целиком. Если, например, страница временно недоступна, а ее все равно хочется просмотреть, мы можем попросить поисковую систему показать сохраненную копию. Ссылки для этого располагаются рядом с результатами поиска.



Воскрешение «умерших» страниц.

Сохраненные копии Интернет страниц с 1996 года можно найти также в архиве Интернета, сайт http://www.archive.org/. Вводите любую страницу, выбирайте год, месяц, и смотрите, как она выглядела в то время.

- *Логический запрос:* использование ключевых слов AND, OR, NOT. Например, можно искать «Windows AND (98 OR 95)». Это выдаст страницы, которые содержат либо слова Windows 98, либо Windows 95.
- *В Google встроен калькулятор.* Попробуйте набрать в строке поиска 6*7, и вы узнаете, чему равно это произведение. Калькулятор имеет много возможностей, в частности умеет производить действия с комплексными числами, что иногда может оказаться полезным. (Это, конечно, не имеет отношения к поиску, но об этом все равно хочется упомянуть).

Последний аспект поиска – язык ресурсов, которые мы ищем. Обычно на английском языке содержится более полная информация по разным вопросам. При поиске неплохо искать информацию и на русском, и на английском. В случае трудностей с языком можно пользоваться сервисами для перевода: http://lingvo.yandex.ru, http://www.lingvo.ru, http://translate.google.com. Последний позволяет переводить сайты целиком. Первые полезны для правильного перевода терминов. Например, как перевести на английский язык термин «искусственный интеллект»? Вроде, ничего не мешает сказать что-то наподобие «simulated intellect», но вы ничего не найдете про искусственный интеллект, если введете такой запрос. Правильно термин звучит «artificial intelligence», и именно он используется в Интернете. Впрочем, про выбор ключевых слов при поиске будет написано дальше.

ПРОЦЕСС ПОИСКА

Первый шаг в процессе поиска – это принятие решения о поиске. Без этого шага найти ничего невозможно, и, как уже говорилось вначале, принимать решение о поиске стоит чаще.

Следующий шаг – сформулировать тему и определить тип ресурсов, которые надо найти. Ресурсами для поиска могут быть:

– тематические ресурсы (например, http://improvement.ru) – сайты, посвященные одной конкретной теме;

– сообщество – например, сообщества на http://www.livejournal.com;

– популярная статья по интересующему вопросу или их коллекция;

– форум;

– каталог или коллекция ссылок;

– файл;

– энциклопедическая статья;

– электронная книга или электронная библиотека;

– сайты-Сервисы;

– контактная информация – адрес домашней страницы, e-mail.

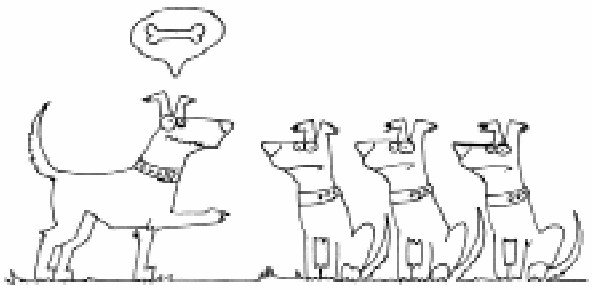
Определенный заранее тип ресурсов, во-первых, делает задачу поиска более конкретной, а во-вторых, позволяет выбрать поисковую систему.

Помимо поисковых систем общего назначения, таких как Google, Yandex, Yahoo, Rambler, существуют специализированные поисковые системы. Например, энциклопедические статьи хорошо искать на сайте http://Wikipedia.org или http://slovari.yandex.ru. Для поиска научных статей удобен сайт http://scholar.google.com, сайты http://ebdb.ru и http://www.poiskknig.ru позволяют искать электронные книги; http://filesearch.ru – поиск файлов; http://market.yandex.ru – поиск товаров и т. д.

Помимо этого существуют нестандартные, метапоисковые системы, их суть в том, что они осуществляют поиск не сами, а с помощью других поисковых систем. Результат поиска обрабатываются дальше, в частности, чтобы представить его в некотором более удобном виде. Например, системы http://www.nigma.ru и http://www.clusty.com совершают обычный поиск, а потом выделяют среди найденных документов группы похожих по содержанию. Результат можно просматривать отдельно по каждой группе. Система http://www.quintura.ru – визуальная поисковая система, на нее лучше один раз посмотреть, чем прочитать.

Другие полезные для поиска сайты – каталоги http://list.mail.ru, http://www.dmoz.org и http://yahoo.com.

Из всего многообразия сайтов более всего выделяется http://Wikipedia.org. Это онлайн энциклопедия, одной которой часто достаточно, чтобы найти нужную информацию. Даже если вы начнете искать в Google нечто, что есть в Wikipedia, ссылка



...метапоисковые системы ... осуществляют поиск не сами, а с помощью других поисковых систем...

на статью окажется одной из первой в результатах поиска Google.

После того как мы определились с тем, что и каким методом мы хотим найти, нужно подобрать ключевые слова. На этом шаге мы пытаемся предугадать, какие слова используются на нужном нам сайте. Пример неправильного выбора ключевых слов уже был приведен, термина «simulated intellect» не существует, и эти ключевые слова не позволят найти информацию об искусственном интеллекте. Ключевые слова можно искать в энциклопедических статьях. Например, если начать поиск информации в Wikipedia, то по статье можно узнать, какие основные слова используются в данной области. Например, в статье про искусственный интеллект вы обнаружите, что ИИ – стандартное сокращение для этого термина (соответственно, AI по-английски), и по одному этому сокращению можно находить в Интернете интересные результаты, не те же, что при поиске по ключевым словам «искусственный интеллект». Wikipedia можно также использовать для перевода терминов с русского на английский и наоборот. Если вы читаете статью, например, про «фразеологизм», то можно найти слева на странице список языков, на которых есть эта же статья. Если перейти к английскому варианту, окажется, что статья написана про слово Idiom, что, следовательно, и является переводом термина на английский. Обычные словари не дают перевода слова «фразеологизм».

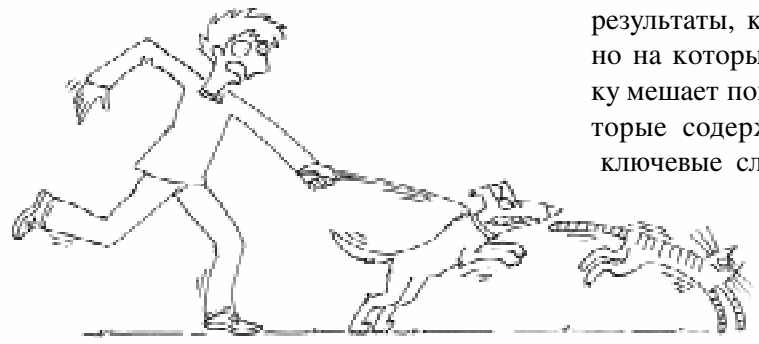
Кроме того, искать ключевые слова можно в описаниях сайтов в каталогах или в заголовках на уже найденных сайтах, в аннотациях к научным статьям и иногда в разделе «ключевые слова», который есть в некоторых статьях как раз для поддержки поиска.

Несколько примеров частых и полезных ключевых слов: *Скачать*, *download*, *free*. *Pdf*, *ppt*, *doc*, *zip*, *txt*, *mp3* – расширения файлов, которые вы хотите найти. *Форум*, *каталог*, *ссылки*, *forum*, *directory*, *links*. *Для начинающих*, *первые шаги*, *руководство*, *советы*, *правила*, *faq*, *for newbies*, *for beginners*, *guide*, *rules*, *checklist*. *Конспект лекций*, *обзор*, *lecture notes*, *survey*. *Как*, *где*, *почем*, *хорошо*, *правильно*. Последний набор ключевых слов соответствует тому, что вы хотите найти обсуждение на форуме. Если кто-то задал интересующий вас вопрос со словами *как*, *где*, *почем*, наверняка ему уже ответили, и вы сможете прочитать ответ. Аналогично пример запроса из самого начала статьи «хочу почитать про ТРИЗ» может оказаться не таким уж и бессмысленным, вы найдете не столько книги о ТРИЗ, сколько их обсуждения.

Последнее ключевое слово: « – это». Попробуйте ввести «Неокортекс – это», одной из ссылок должно появиться определение этого термина. Хотя искать энциклопедические определения лучше либо сразу в энциклопедиях, либо с помощью сервисов подобных http://slovari.yandex.ru.

Какие трудности могут возникать при поиске?

Мы можем знать, какую надо решить проблему, но не знать, что для этого нужно искать. Часто поиску мешают интересные результаты, которые, не относятся к делу, но на которые хочется отвлекаться. Поиску мешает поисковый спам – страницы, которые содержат популярные при поиске ключевые слова, но без всякой полезной информации. Служат они, например, для раскрутки других сайтов. Последняя проблема при поиске: можно не найти нужные материалы, если делать только поверхностный поиск.



Часто поиску мешают интересные результаты, которые, не относятся к делу...

Перед заключением перечислим основные ошибки пользователей, совершающих поиск. Использование лишних слов в строке запроса. Лишние слова могут помешать найти интересный сайт, на которых эти слова не используются. Поиск не даст хороших результатов при неудачно подобранных ключевых словах, неправильно переведенных терминах, недостаточной гибкости в смене ключевых слов. При полномасштабном поиске меняйте ключевые слова 3 или 4 раза.

Иногда стоит пробовать искать так, как вы уже пытались. Если какой-то метод вы «уже пробовали», это не значит, что во второй раз вы не обнаружите новых результатов. Неудобно совершать поиск, если не разделять во времени поиск и чтение. Как уже говорилось, полезно сначала отобрать потенциально интересные документы и только потом их изучить.

Наконец, последнее правило: если поисковая система не выдала результатов, значит, ей был задан плохой запрос. Трудно сказать, насколько серьезный надо проводить поиск информации, чтобы уверенно заявить – в Интернете об этом не написано.

ЗАКЛЮЧЕНИЕ

В заключение перечислим основные моменты, упомянутые ранее в статье. Помните о возможностях Интернета и чаще принимайте решение о поиске. Используйте поиск не только на русском языке, но и на английском. Используйте для поиска специализированные сайты, каталоги и особенно энциклопедию http://wikipedia.org. Основная техника поиска – выбор правильных ключевых слов, их нужно уметь определять, читая уже найденные документы.

С помощью поиска в Интернете можно развлекаться. Попробуйте игру Uno Google. Следует подобрать два слова таким образом, чтобы при поиске по этим словам получался ровно один результат. Попробуйте найти другие способы развлечения с помощью поисковых систем.

Несколько других практических заданий на поиск: найдите технику конькового хода на лыжах (с картинками). Найдите список тиров вашего родного города. Найдите материалы на тему «как делать хорошие презентации». Найдите в Интернете практические задания на тренировку поиска в Интернете. С помощью поиска в Интернете определите, в каком литературном произведении и какого автора была написана фраза про Аристотеля, приведенная в начале статьи.

Последнее, что хочется сказать. Интернет содержит немыслимое количество информации, но далеко не вся она достоверна. Всегда следите за тем, что и на каком сайте вы читаете, и критически относитесь к тому, что узнаете.

Источники

1. Тренинг Ю.Лифшица «Эффективный поиск в Интернете», http://yury.name/search.html

2. *Sergey Brin and Lawrence Page*. The Anatomy of a Large-Scale Hypertextual Web Search Engine, http://infolab.stanford.edu/~backrub/google.html

Посов Илья Александрович, аспирант математико-механического факультета СПбГУ.