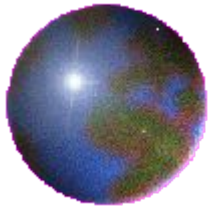
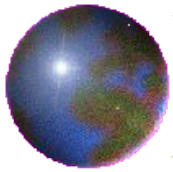


Поиск информации в интернете

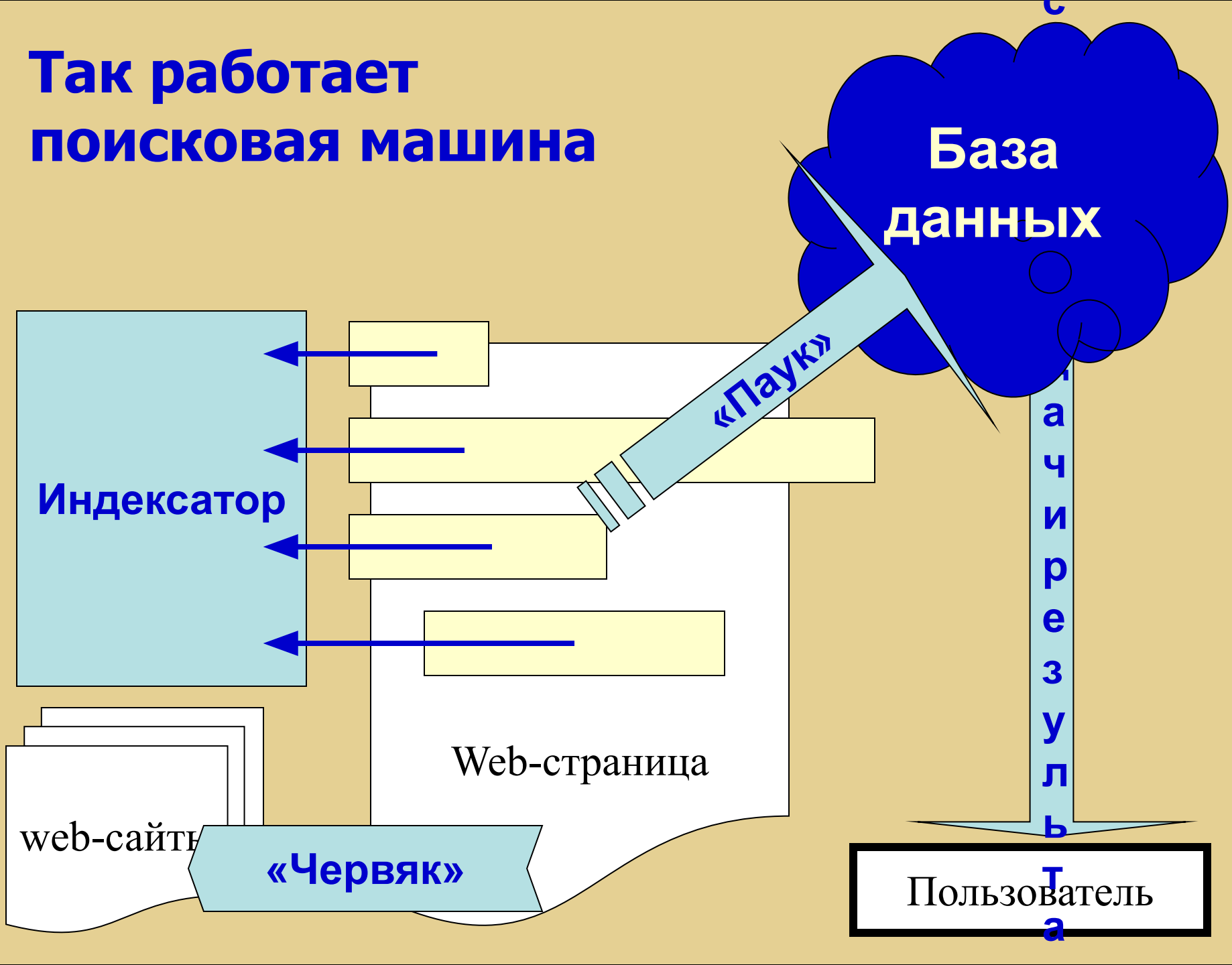
Теория поиска информации

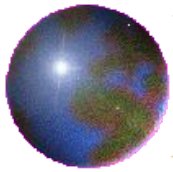


Теория поиска информации

- Прежде чем мы обратимся к ближайшему рассмотрению ПС, необходимо рассмотреть процесс поиска информации в теории.
- Начнем с устройства поисковой машины:

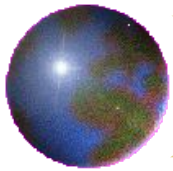
Так работает поисковая машина





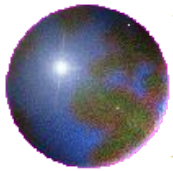
«Паук» (spider)

- Программа, которая загружает в поисковую машину web-страницы.
- Работает аналогично браузеру, установленному на компьютере пользователя, но ничего не отображает ни на каком экране.
- Передает в поисковую систему HTML-код документа.



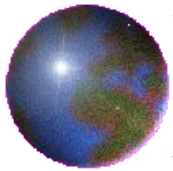
«Червяк» (crawler)

- 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.



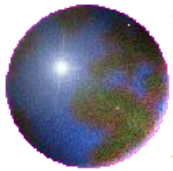
Индексатор (Indexer)

- Программа, которая «разбирает» web-страницу на составные части и анализирует их.
- Вычленяются и анализируются заголовки, ссылки, текст документов.
- Отдельно анализируется текст, набранный полужирным шрифтом, курсивом и т.п.



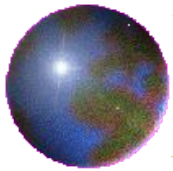
База данных (database)

- Хранилище всех данных, которые поисковая система загружает и анализирует.
- Требует огромных ресурсов как для хранения, так и для последующей обработки.



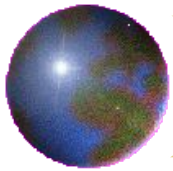
Система выдачи результатов поиска (Search Engine Results Engine)

- Программа, которая решает, какие страницы удовлетворяют запросу пользователя и в какой степени.
- Именно с этой частью поисковой машины «общается» пользователь.



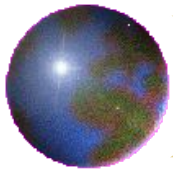
«Паук» и «червяк»

- Первые две программы, работающие «в связке», часто называют поисковый робот или HTTP-робот.



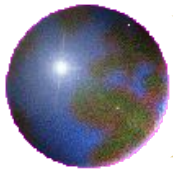
Работа ПС

- Таким образом, после получения запроса ПС анализирует ту информацию, которую собрала ранее.
- Плюсы: многократно повышается скорость обработки запроса.
- Минусы: область поиска ограничена внутренними ресурсами ПС, информация в базе данных быстро устаревает.



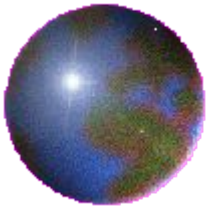
Индексация и индекс

- Процесс загрузки информации из интернета и предварительного анализа ее поисковой машиной называют индексацией.
- Саму базу данных ПС, в которой храниться вся информация – индекс.

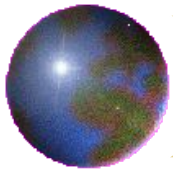


Индексация

- Глубина индексации может быть разной.
- Полные тексты документов, хранящихся на сайте, в базу данных копируются не всегда, иногда поисковые роботы ограничиваются урезанными версиями или вообще только заголовками.

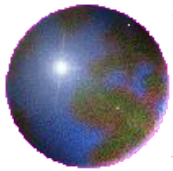


Механизмы и алгоритмы поиска



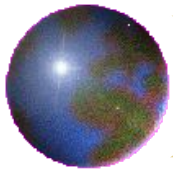
Алгоритмы поиска

- Каждая ПС использует свой алгоритм поиска и его детали представляют собой ноу-хау разработчиков поисковика.
- Алгоритм поиска – метод, руководствуясь которым ПС принимает решение, включать или не включать ссылку на web-страницу в результаты поиска.



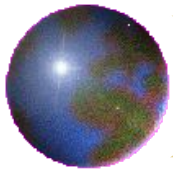
Закономерности поиска

- Некоторые из закономерностей поиска информации были описаны профессором филологии из Гарварда Джорджем Зипфом в 1949 году.
- Без учета собранных им закономерностей сегодня не способна работать ни одна система автоматического поиска информации.



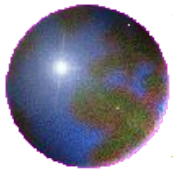
Законы Зипфа

- Зипф заметил, что длинные слова встречаются в текстах любого языка реже, чем короткие.
- Это по всей видимости связано с природой человека и вообще любого живого существа.
- На основе этого наблюдения Зипф вывел два закона.



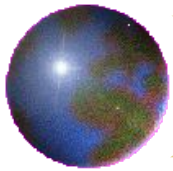
Первый закон Зипфа

- Первый закон связывает частоту появления (вхождения) того или иного слова с рангом этой частоты.
- Наиболее часто встречающимся словам присваивается ранг, равный единице.
- Тем словам, что встречаются реже – ранг, равный двойке и т.п.



Первый закон Зипфа

- Зипф обнаружил, что произведение частоты вхождения слова и его ранга является постоянной величиной.
- Такая зависимость обычно отображается гиперболой.
- Значение константы Зипфа для разных языков различно, но внутри одной языковой группы оно остается неизменным.

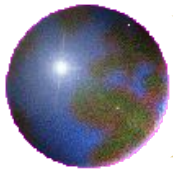


Первый закон Зипфа

- Частота появления (вхождения) того или иного слова является отношением количества появления слова к общему количеству слов в тексте.
- Таким образом, частота слова не может быть больше единицы и составляет в реальности сотые и тысячные доли.

Первый закон Зипфа

- Для русского языка константа равна примерно 0,06-0,07.

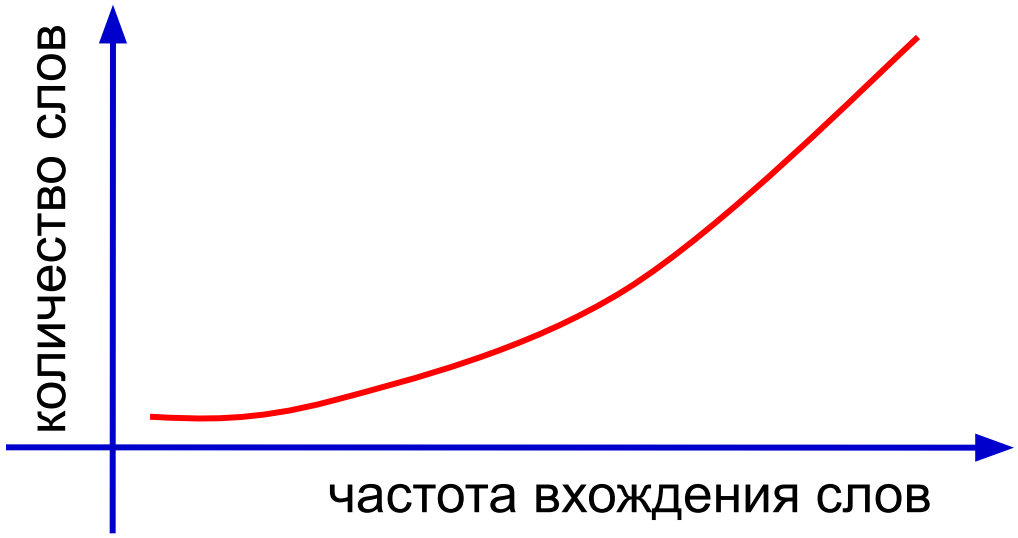


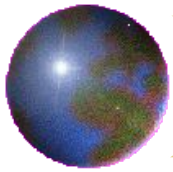
Первый закон Зипфа

- Примеры работы закона:
  - Если наиболее распространенное слово встречается в тексте 100 раз, то следующее по распространенности встретится не 99 и не 90 раз, а примерно 50!
  - Самое часто встречаемое слово в английском языке the употребляется в 10 раз чаще, чем слово, имеющее ранг, равный 10. В 100 раз чаще, чем слово, имеющее ранг 100 и т.д.

Второй закон Зипфа

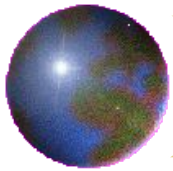
- Зипф определил, что частота вхождения слов и количество слов, входящих в текст с данной частотой, тоже взаимосвязаны.





Второй закон Зипфа

- Получившая кривая будет сохранять свои параметры для всех текстов в пределах одного языка.
- С другой стороны, на каком бы языке текст ни был написан, форма кривой Зипфа останется неизменной. Отличаться будут лишь коэффициенты.

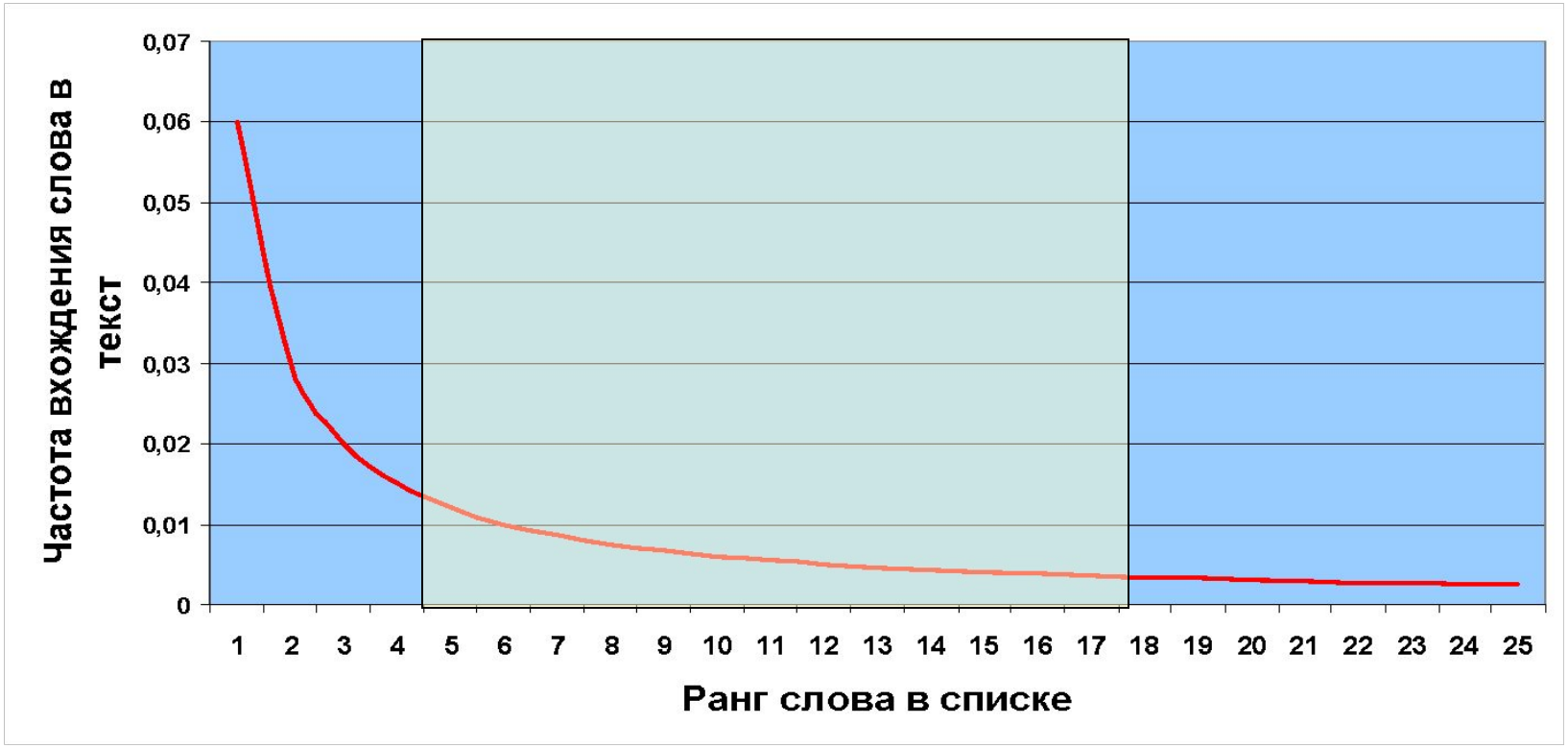


Следствия законов Зипфа

- Законы Зипфа универсальны. Они применимы не только к текстам.
- В аналогичную форму выливается, например, зависимость между количеством городов и числом проживающих в них жителей.
- Характеристики популярности ресурсов интернета отвечают законам Зипфа.
- В законах Зипфа отражается «человеческое» происхождение объектов.

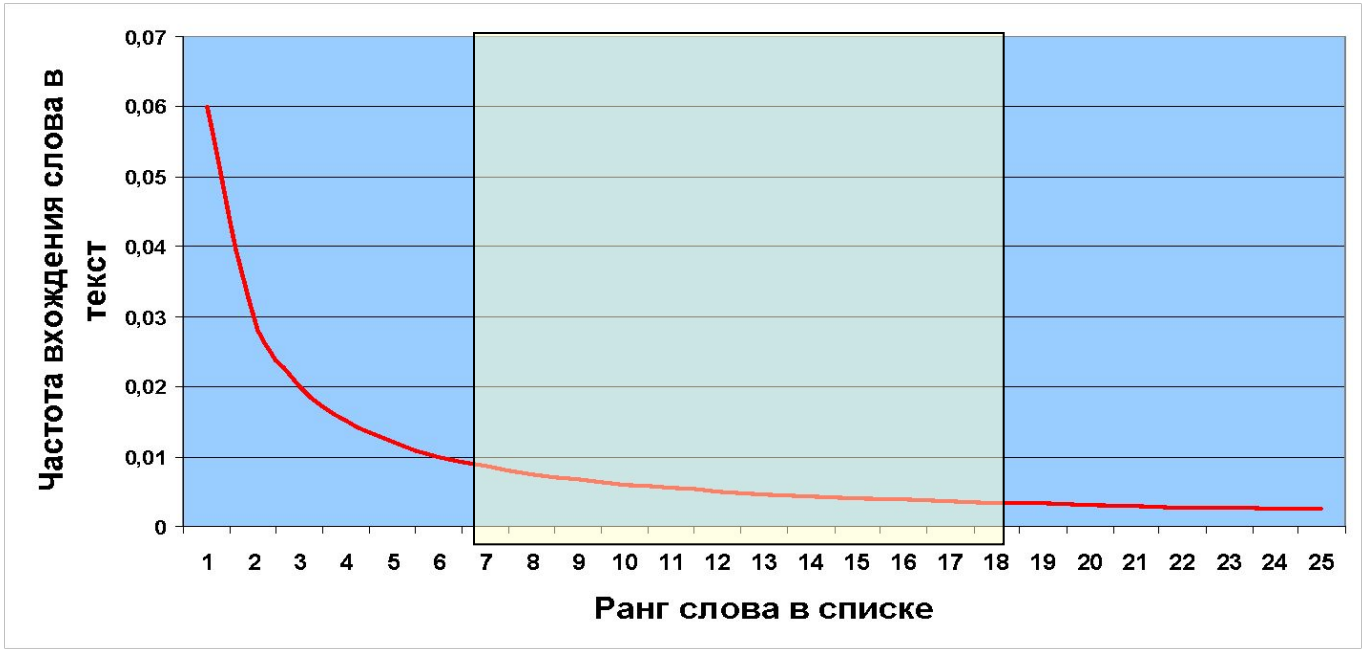
Как ПС используют законы Зипфа

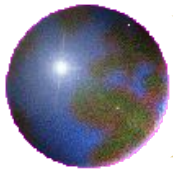
- Рассмотрим график первого закона:



Как ПС используют законы Зипфа

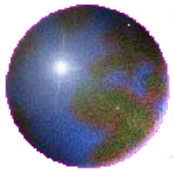
- Из анализа графика можно предположить, что наиболее значимые для текста слова лежат в средней части графика.





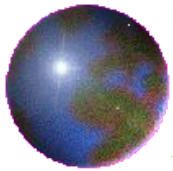
Центральная часть графика

- Центральная зона графика содержит термины, наиболее характерные для данного текста.
- Они в совокупности выражают специфичность текста, отличие его от других, охватывают его основное содержание.



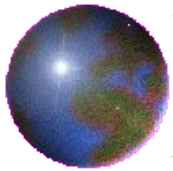
Левая и правая часть графика

- Действительно, наиболее часто встречаемые слова – слева – это предлоги, местоимения, артикли и т.д.
- Справа – редко встречаемые слова. Они не несут в большинстве случаев особого смыслового значения.
- Хотя иногда, они, наоборот, бывают весьма важны (об этом чуть позже).



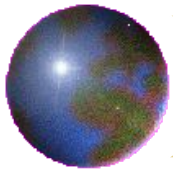
Значимые слова

- Каждая ПС по-своему решает, какие слова отнести к наиболее значимым.
- Однако, если к числу значимых будет отнесены слишком много слов, то важные термины будут забиты «шумом» случайных слов.
- Если значимых слов будет слишком мало, то есть риск потерять главное.



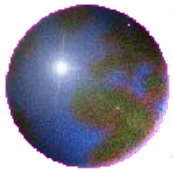
Стоп-слова

- Для того, чтобы безошибочно сузить диапазон значимых слов, создается словарь «бесполезных» слов или «стоп-слов».
- Словарь этих слов («стоп-лист») содержит, например, артикли и предлоги, частицы и личные местоимения.



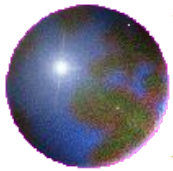
Весовой коэффициент

- При определении значимых слов применяется и т.н. «весовой коэффициент».
- Часто встречаемое слово имеет весовой коэффициент, близкий к нулю.
- Слово, встречаемое редко, - весьма высокий коэффициент.



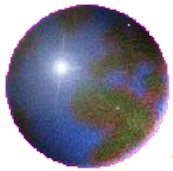
Весовой коэффициент

- Параметр, определяющий «весовой коэффициент», называется инверсная частота термина.
- ПС может вычислять «весовой коэффициент» с учетом местоположения слова внутри документа, взаимного расположения разных слов, морфологических особенностей и т.п.



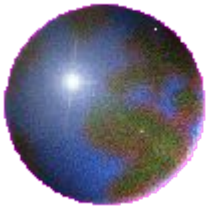
Принцип работы современной ПС

- Современные ПС имеют пространственно-векторную модель построения базы данных.
- Она позволяет получить результат, отвечающий запросу даже в том случае, когда в найденном документе не окажется ни одного ключевого слова!

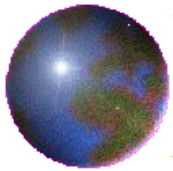


Принцип работы современной ПС

- Это достигается благодаря тому, что все документы базы располагаются в виртуальном многомерном пространстве.
- Координаты каждого документа зависят от содержащихся в тем терминов, их весовых коэффициентов, положения терминов внутри документа и т.п.
- Таким образом, документы с похожим набором терминов оказываются в этом пространстве поблизости и ПС их выдает в ответ на запрос.

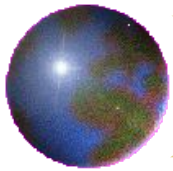


Полнота и точность поиска



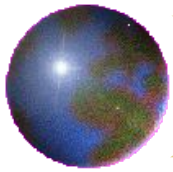
Релевантность

- Релевантным называется документ, имеющий отношение к сделанному Вами запросу, т.е. формально содержащий запрашиваемую Вами информацию.
- Англ. relevant – «подходящий, относящийся к делу».



Релевантность

- Конкретное общепринятое определение релевантности еще не сложилось.
- «Экономический словарь» (www.km.ru) толкует релевантность как «смысловое соответствие между информационным запросом и полученным сообщением».
- Яндекс: «мера соответствия результатов поиска задаче, поставленной в запросе».



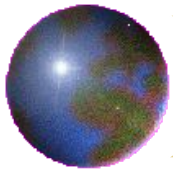
Релевантность

- В то же время, на Яндексе говорится:
  - «При поиске в интернете важны две составляющие – полнота (ничего не потеряно) и точность (не найдено ничего лишнего). Обычно все это называют одним словом – релевантность».

Полнота поиска

- Коэффициентом полноты поиска называют отношение количества полученных релевантных документов к общему количеству существующих в базе данных релевантных документов:

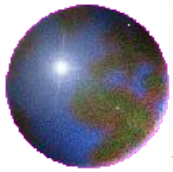
$$\textbf{Коэф. полноты поиска} = \frac{\textbf{Полученные релевантные документы}}{\text{Общее количество релевантных документов в базе данных ПС}}$$



Полнота поиска

- В идеальной ПС коэффициент полноты поиска = 1.
- А противоположный ему коэффициент потерь информации = 0.

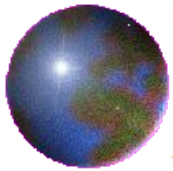
- В реальности коэффициент полноты поиска = 0,7-0,9



Точность поиска

- Коэффициентом точности поиска называют отношение количества релевантных результатов к общему количеству документов, содержащихся в ответе ПС на запрос:

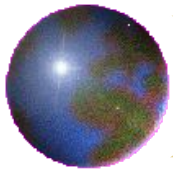
$$\textbf{Коэф. точности поиска} = \frac{\textbf{Количество релевантных документов}}{\text{Общее количество документов в ответе ПС на запрос}}$$



Точность поиска

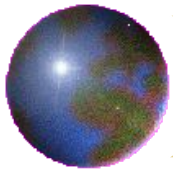
- В идеальной ПС коэффициент точности поиска = 1.
- А противоположный ему коэффициент поискового шума= 0.

- В реальности коэффициент точности поиска = 0,1-1



Полнота и точность

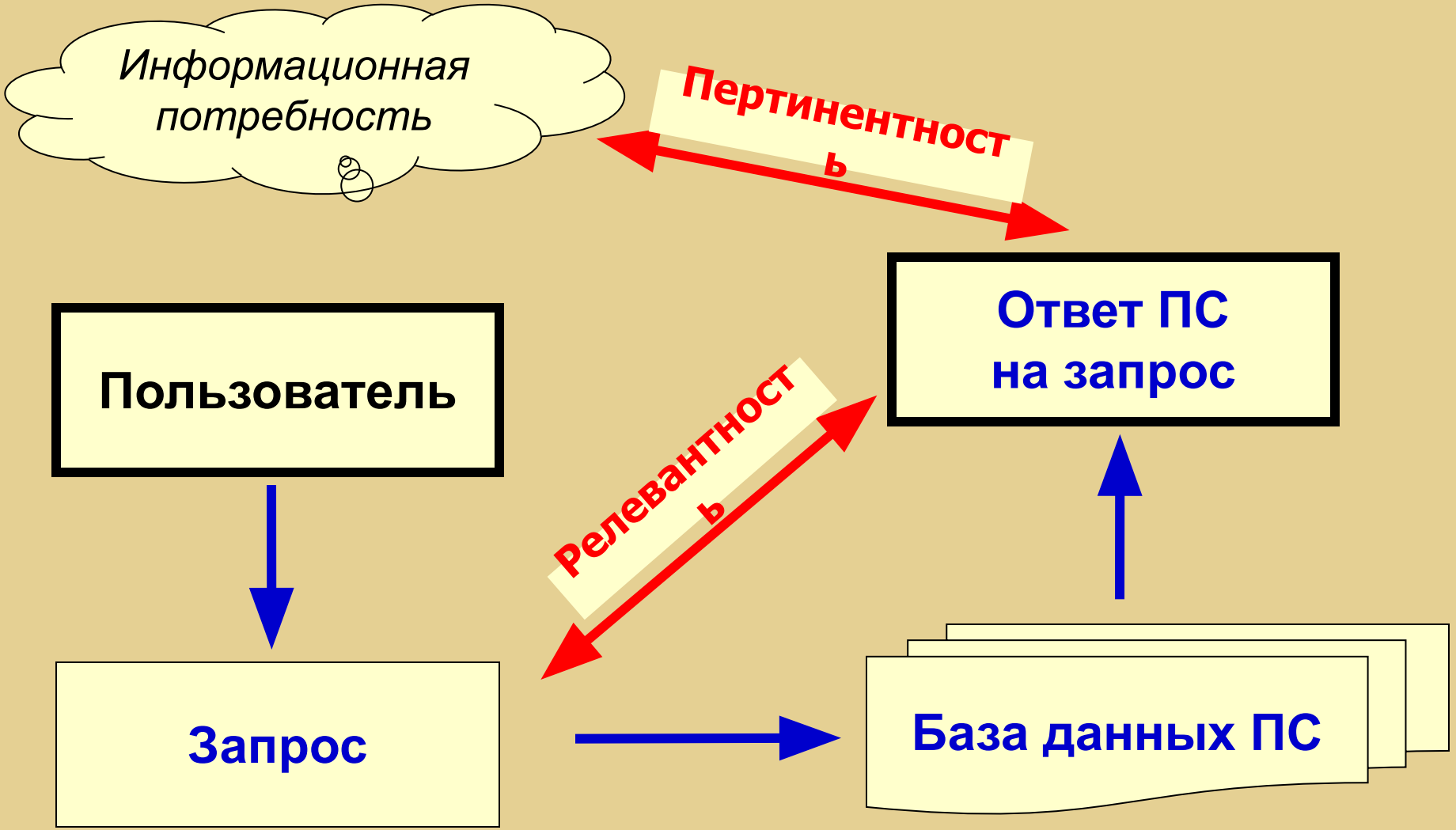
- Нередко количество размещенных в интернете релевантных пользователю документов может составлять десятки тысяч.
- В то же время релевантная информация в них совпадает, и пользователю достаточно изучить лишь несколько документов из числа найденных.
- Таким образом, полнота в сравнении с точностью является второстепенным критерием качества информационного поиска.

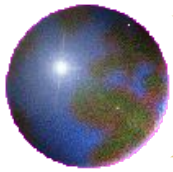


Пертинентность

- На практике используется еще и неформальное понятие – пертинентность.
- Это соотношение объема полезной для пользователя информации к объему полученной.
- Зачастую это соотношение имеет решающее значение.

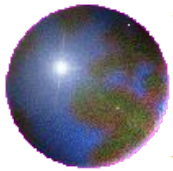
Релевантность и пертинентность





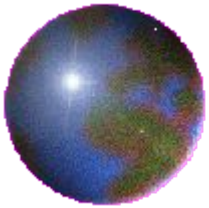
Повышение пертинентности

- Средства повышения пертинентности:
  - уточнение формулировок запросов,
  - ранжирование по весовым критериям,
  - ограничение числа выданных в результате поиска документов.

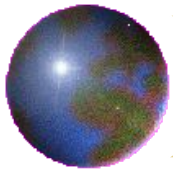


Пертинентность

- Проблеме пертинентности уделяется большое внимание в современных ПС.
- Так, ПС Google реализовала алгоритмы достижения неформальной релевантности (пертинентности) и благодаря этому стала самой популярной ПС в интернете.

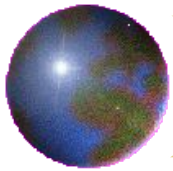


Морфологический анализ



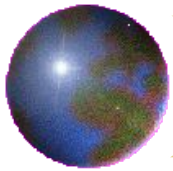
Морфологический анализ

- Почти все современные ПС учитывают изменения слова в поиске документов.
- Указывая в строке поиска слово, мы увидим в результате поиска документы, содержащие варианты этого слова, измененные по падежам, числу, спряжению и т.д.



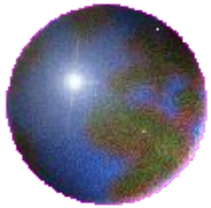
Морфологический анализ

- Для непрофессионалов морфологический анализ – это удобная функция.
- Она позволяет производить поиск по всем вариантам слов сразу и находить даже документы, где слово используется в другой форме.

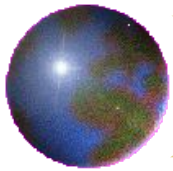


Морфологический анализ

- Для профессионального поиска морфологический анализ не всегда пригоден. Он лишает поиск гибкости.
- Морфологический анализ может увеличить количество документов, выдаваемых по запросу, но количество релевантной информации уменьшится.

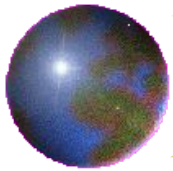


Эффективный поиск



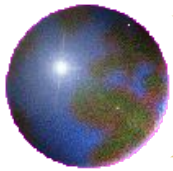
Эффективный поиск

- Будем считать, что эффективность поиска информации тем выше, чем больше коэффициенты полноты и точности,
- в то же время – меньше время и другие ресурсы, затрачиваемые на проведение поиска.



Расширенный поиск

- Многие современные ПС с целью повышения эффективности поиска позволяют вместо простого поиска производить т.н. «расширенный».
- Он доступен по ссылке на странице поиска и представляет собой форму, которую нужно заполнить, ответив на дополнительные вопросы.



Сложный поиск

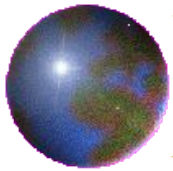
- Кроме этого возможен и т.н. «сложный» поиск с использованием булевых операторов, то есть поиск с помощью логических операторов.
- Булевый поиск станет темой нашего следующего занятия.

Этапы поисковой процедуры

Формирование потребности в информации

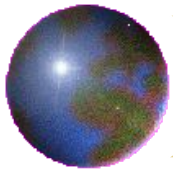
Формирование эффективного запроса к ПС

Поиск нужной информации в ответе ПС



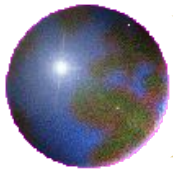
Формирование потребности

- На этой фазе определяется цель поиска, его стратегия и область проведения поиска.
- Информационные потребности могут относиться к разным областям, но на практике они сводятся к общим шаблонам поиска:



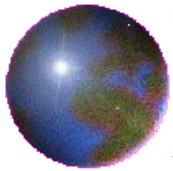
Шаблоны поиска

- Поиск новостей,
- поиск людей,
- поиск предприятий и организаций,
- поиск документов,
- поиск музыки, видео и графики,
- поиск программного обеспечения,
- и т.д.



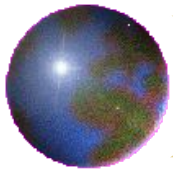
Формирование запроса

- Вторая часть поисковой процедуры предусматривает многовариантность подходов и решений при формализации запроса.
- Здесь же решается вопрос о выборе конкретной ПС или каталога.



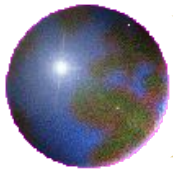
Формирование запроса

- Основная задача при этом – формирование эффективного запроса.
- Основная проблема заключается в том, что в каждой ПС используется свой информационно-поисковый язык.
- Хотя у различных языков этого типа много общего, например, схожий набор булевых операций.



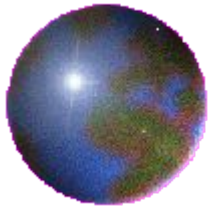
Формирование запроса

- В настоящее время не существует единого стандарта языка запросов к ПС, хотя попытки стандартизации ведутся.
- Таким образом, в наших лекциях мы обратимся только к двум ПС: Google и Яндекс для иллюстрирования работы языка запросов.

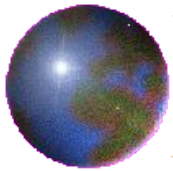


Поиск нужной информации

- Третий этап является определяющим: от его реализации зависит, будет ли найденная информация пертинентной.
- На этом этапе пользователь работает с конечным результатом поиска – откликом ПС на запрос.

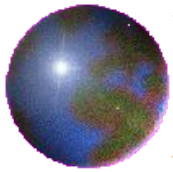


Советы по поиску в интернете



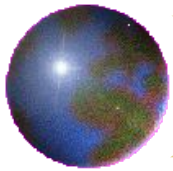
Необходимое замечание

- 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.



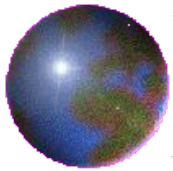
Проверяйте орфографию

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



Используйте синонимы

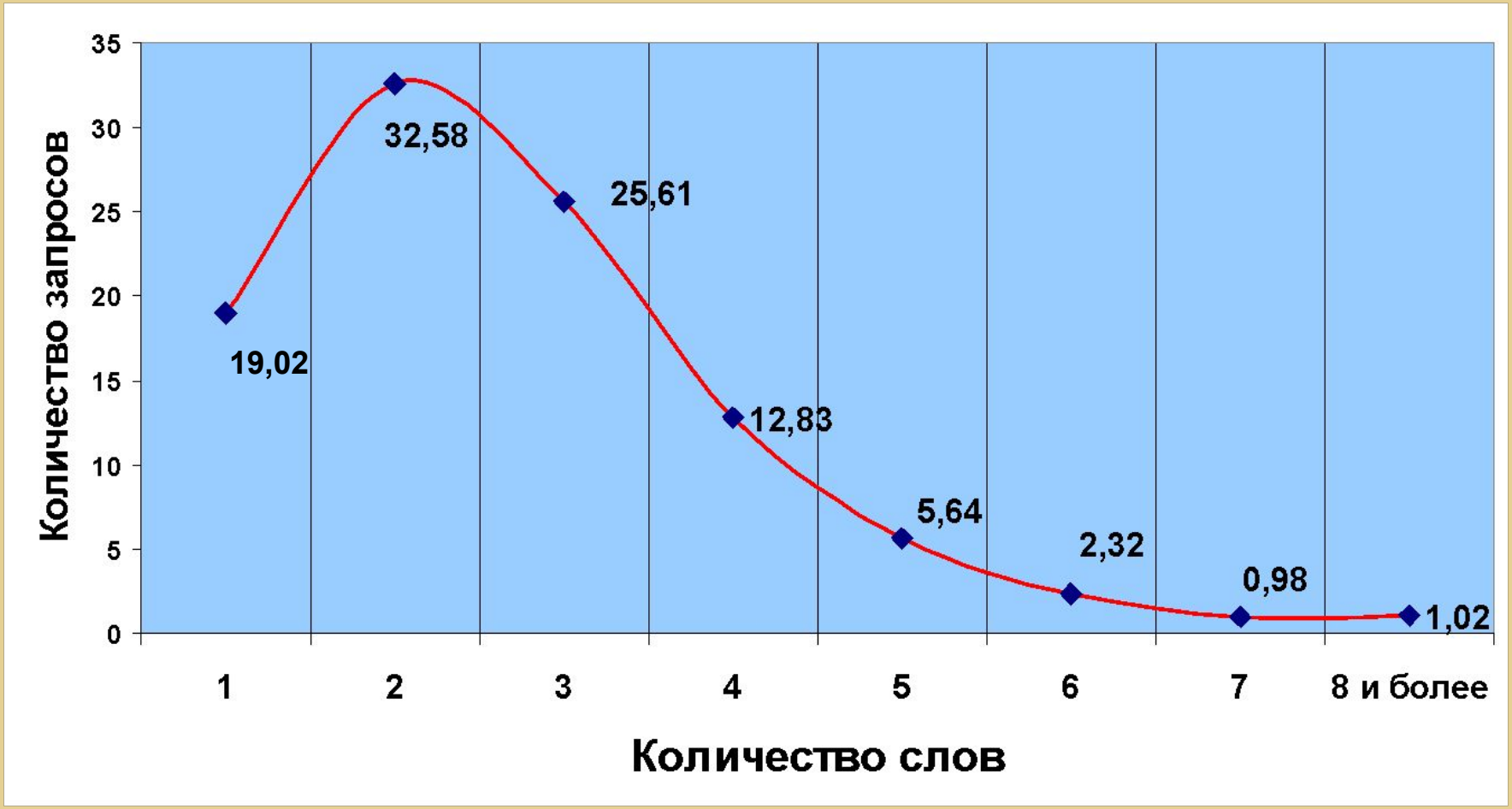
- Если список найденных страниц слишком мал или не содержит полезных страниц, попробуйте изменить слово.
- Попробуйте задать для поиска три-четыре слова-синонима сразу.
- Для этого перечислите их через вертикальную черту (|). Тогда будут найдены страницы, где встречается хотя бы одно из них.



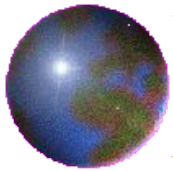
Ищите больше, чем по одному слову

- Многие слова при поиске поодиночке дадут большое число бессмысленных ссылок.
- Добавьте одно или два ключевых слова, связанных с искомой темой. Например, «психология Юнга».
- Рекомендуем также сужать область вашего вопроса. Запрос «автомобиль Волга» выдаст более подходящие Вам документы, чем «легковые автомобили».

Распределение запросов по количеству слов

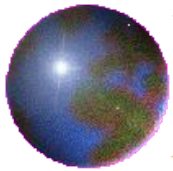


OneStat.com, 2004 г.



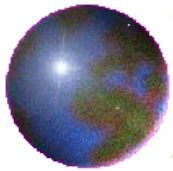
Не пишите большими буквами

- 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 названиях и именах собственных. Например, *министр Иванов, телепередача Здоровье.*



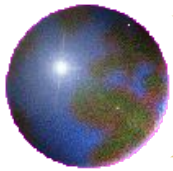
Ищите без морфологии

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



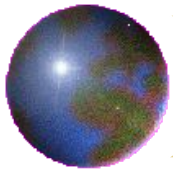
Ищите похожие документы

- Если один из найденных документов ближе к искомой теме, чем остальные, нажмите на ссылку «найти похожие документы».
- ПС проанализирует страницу и найдет документы, похожие на тот, что вы указали.



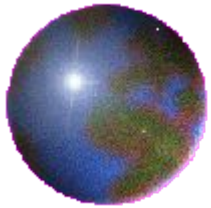
Используйте знаки «+» и «-»

- 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.

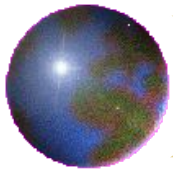


Используйте язык запросов

- С помощью специальных операторов вы сможете сделать запрос более точным.
- Например, укажите, каких слов не должно быть в документе, или что два слова должны идти подряд одно за другим, а не просто встречаться в документе.
- О языке запросов мы поговорим подробнее на следующем занятии.

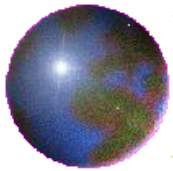


Сохранение информации из интернета



Сохранение web-страниц

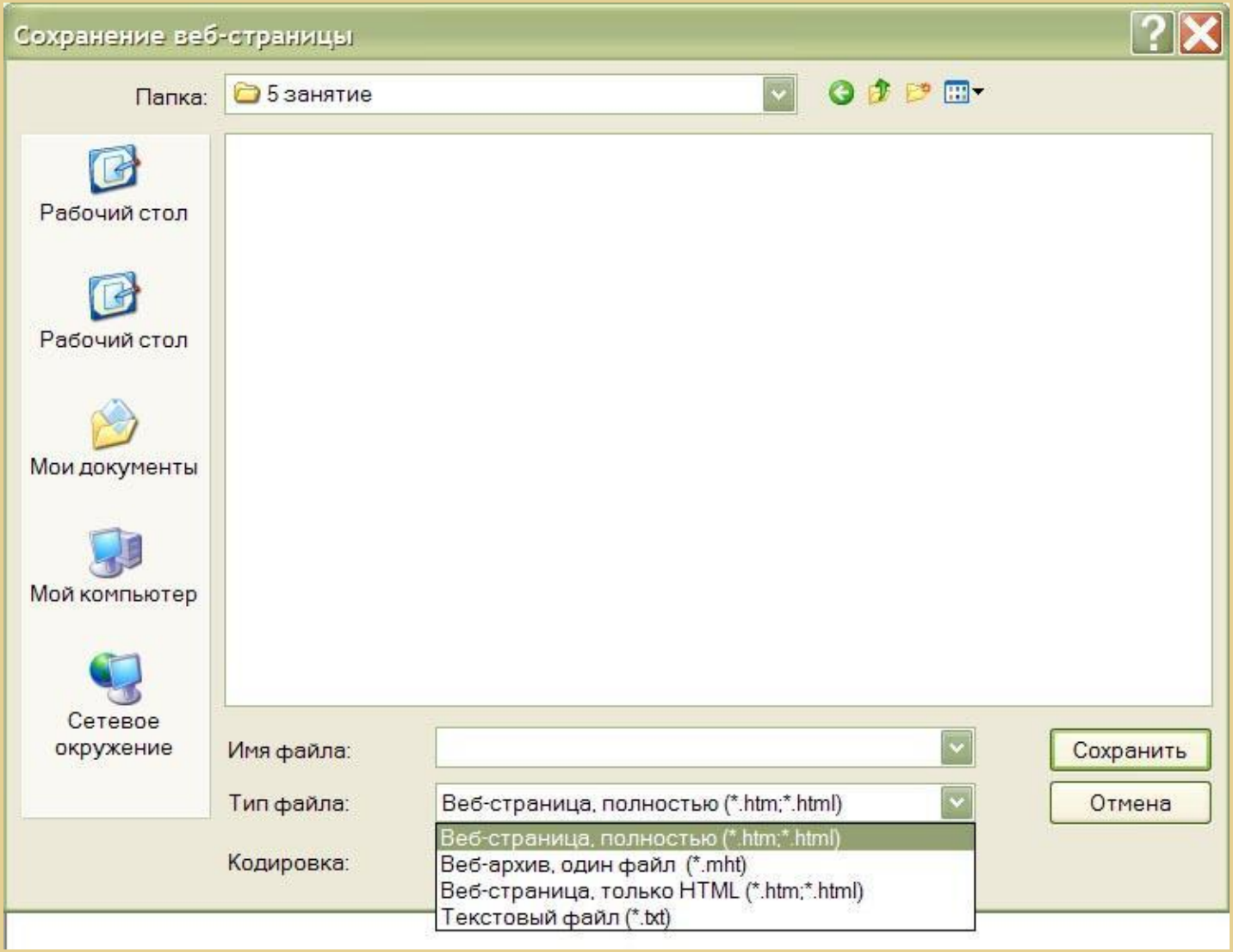
- Самая главная операция любого пользователя интернета – сохранение найденной информации.
- Итак, сохранение документа с помощью меню броузера.
- Имеют значение два обстоятельства:
  - тип броузера,
  - в каком виде вы хотите сохранить документ.

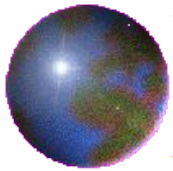


Сохранение web-страниц

- 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.

Сохранение в Microsoft Internet Explorer

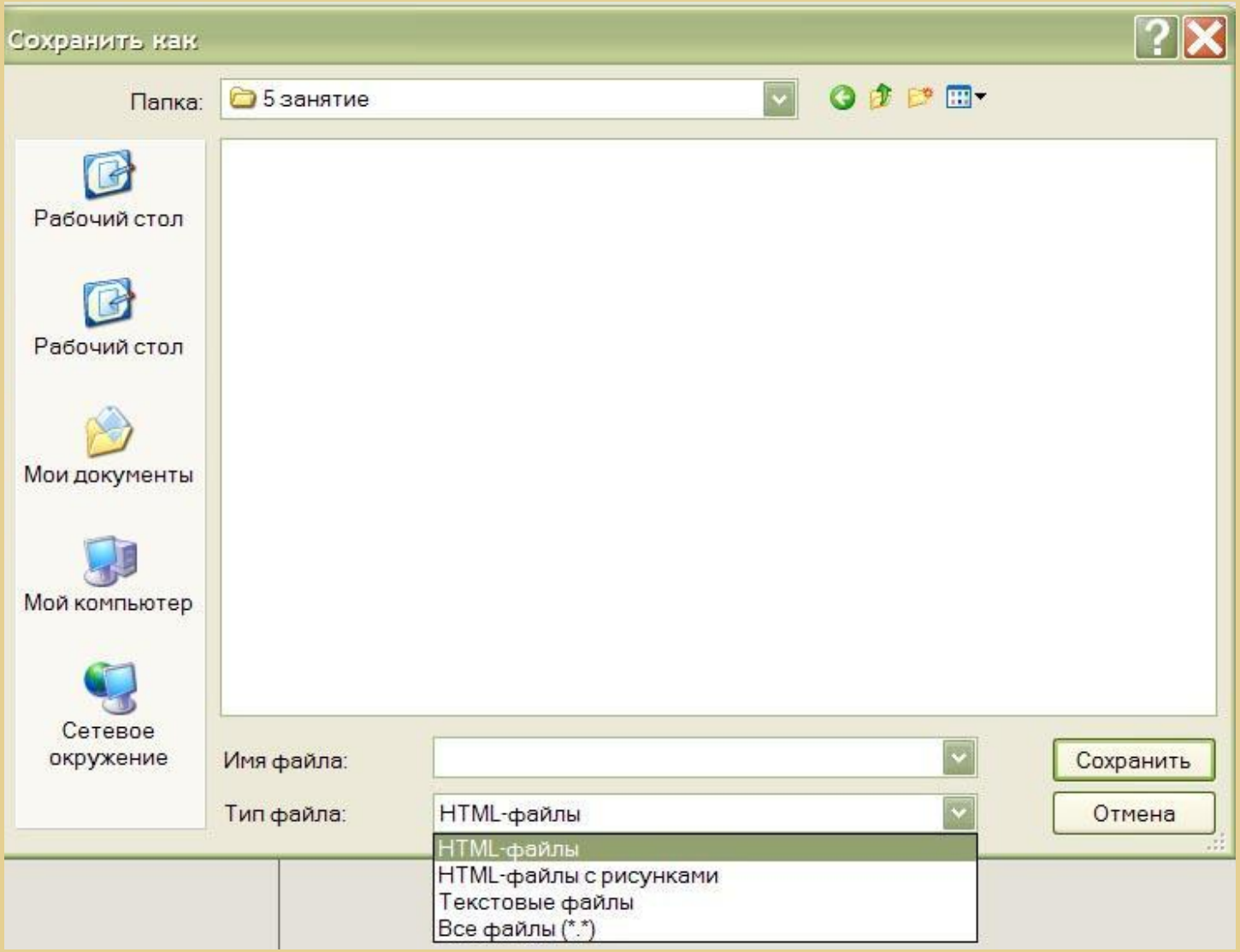


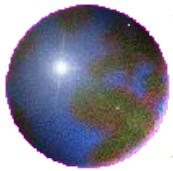


Сохранение web-страниц

- 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).

Сохранение в Opera



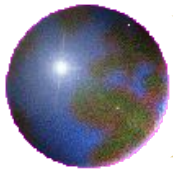


Сохранение файлов других типов

- 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.

Совет по сохранению информации

- 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.



Таким образом,

- Мы изучили устройство поисковой системы,
- разобрали теоретические подходы к поиску информации,
- рассмотрели советы по эффективному поиску в интернете,
- изучили способы сохранения информации из интернета.

Источники информации

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