БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Химический факультет
Кафедра общей химии и методики преподавания химии

А.А.Рагойша

ПОИСК ХИМИЧЕСКОЙ ИНФОРМАЦИИ В ИНТЕРНЕТЕ
ч. I. ПОИСКОВЫЕ СИСТЕМЫ И ТЕМАТИЧЕСКИЕ КАТАЛОГИ

Онлайновое издание (препринт)
http://www.abc.chemistry.bsu.by/lit/Rahoisha_2003.pdf
Размещено на сайте ABC Chemistry
http://www.abc.chemistry.bsu.by/

Минск
2003

УДК 54:004.738.52
ББК

Автор*:* *А.А.Рагойша,* доцент кафедры общей химии и методики преподавания химии БГУ

Рецензенты: *Ю.И.Воротницкий*, директор ЦИТ БГУ, канд. физ.-мат. наук;
Е.А.Стрельцов, доцент кафедры неорганической химии БГУ, доктор хим. наук

Утверждено Советом химического факультета 22.04.03, протокол № 5.

Рагойша А.А.

Поиск химической информации в Интернете. Ч. I. Поисковые системы и тематические каталоги: Учеб. пособие для студ. хим. фак.— Мн.: БГУ, 2003.— 87 с.

В учебном пособии рассматриваются принципы проведения информационного поиска в текстовых электронных базах данных. Даны подробные указания по методике работы с основными универсальными и специализированными поисковыми системами, метапоисковыми системами и тематическими каталогами. Пособие может использоваться студентами, аспирантами и научными работниками естественнонаучных специальностей.

УДК 54:004.738.52
ББК

ISBN
ISSN

ВВЕДЕНИЕ

Научные ресурсы Интернета огромны и разнообразны. Здесь можно найти электронные аналоги печатных изданий (журнальные публикации, патенты, диссертации, отчеты НИР, препринты, материалы конференций, справочники, словари, монографии, учебники и учебно-методические разработки), а также специфичные для компьютерной сети формы (электронные журналы, онлайновые конференции, компьютерные программы, интерактивные учебники, видеоматериалы, презентации, трехмерные модели молекул и кристаллов).

Интернет — динамическая система: одни документы появляются, другие по желанию авторов исчезают бесследно. Сколько их — точно не знает никто. Известно только, что счет идет на миллиарды. Хотя в Интернете отсутствует единый каталог ресурсов, его функции частично выполняют поисковые системы и иные средства навигации, которые подробно обсуждаются в данном пособии.

На нынешней стадии развития Интернета бесплатные научные ресурсы по степени их доступности условно можно разделить на две части. Одна часть — документы, открытые для всех и *известные поисковым системам*. Другая — так называемый *скрытый Интернет* (*hidden Internet*): базы данных, имеющие свои собственные поисковые программы, и сайты, работающие только с зарегистрированными пользователями. Поисковые системы не имеют сведений о каждом из документов скрытого Интернета, но содержат адреса пунктов доступа к ним.

Вопросы, общие для обеих частей Интернета, рассматриваются в главе 1 (терминология) и главе 2 (принципы работы с текстовыми базами данных). В остальных главах обсуждаются наиболее эффективные средства информационного поиска, использующиеся в открытом Интернете.

Книга является пособием к курсу «Поиск химической информации в электронных базах данных и в Интернете». Материал, ***обязательный для изучения***, размещен в главах **1, 2, 3, 12, 15** и в **первых параграфах** остальных.

Обязательными для ознакомления являются главы **4—11, 13, 14, 16—18**. Их основная функция — служить инструкциями при поисковой работе с конкретными информационными источниками.

1. ТЕРМИНОЛОГИЯ ИНТЕРНЕТА

Официальная история Интернета начинается в 1962 г. — именно тогда впервые прозвучала мысль, что компьютеризованная коммуникационная система должна представлять собою не цепочку приемопередающих устройств, а двухмерную самоуправляемую сеть, причем состоящую из равноправных компьютеров. Эта система не содержала бы командных пунктов и сохраняла бы свою работоспособность даже при утере отдельных ее частей. Сообщения по сети должны были бы пересылаться небольшими фрагментами («*пакетами*»); каждый пакет добирался бы к месту назначения самостоятельно и своим путем, наиболее удобным в конкретный момент.

Второй поворотный пункт в истории датируется концом 60-х гг., когда несколько университетов США приступили к реализации этих идей, объединив свою мощную технику в общую компьютерную сеть. Исследователи, подавая команды со своего терминала, управляли вычислительными процессами на удаленных компьютерах — так возник первый метод работы в сети, получивший название **телнет** (**telnet**). Для переноса файлов с одного компьютера на другой был разработан протокол **FTP** (***F**ile **T**ransfer **P**rotocol*), который до сих пор используется при извлечении информации из многочисленных общедоступных *ftp*-архивов. (**Протокол** — свод правил, в соответствии с которыми передаются и принимаются данные). Обмен индивидуальной корреспонденцией неожиданно быстро вырос в целую систему **электронной почты** (*e-mail*). Успех первой компьютерной сети был настолько очевиден, что, поначалу в США, потом в других странах пошли процессы формирования новых сетей, их объединения — наступила эпоха Интернет.

Интернет *(inter — меж- + net — сеть) — сеть, объединяющая много компьютерных сетей.*

Компьютеры совершенствовались, научились обрабатывать не только числа, а и тексты, в результате Интернет стал приобретать черты огромной библиотеки. Но он бы оставался исключительно научным инструментом, если бы в начале 90-х гг. английский ученый Т. Бернерс-Ли не разработал новый алгоритм обмена информацией — протокол **HTTP** (***H**yper**T**ext **T**ransfer **P**rotocol*) — и, на этой основе, новую — *гипертекстовую* — форму функционирования Сети — **World Wide Web** («Всемирную Паутину»), или **WWW**, или просто **Web**.

Гипертекстовая структура документа не является изобретением компьютерной эры. Конечно же, абсолютное большинство нашей печат-

ной литературы построено по линейному принципу: прочитав первое предложение, мы последовательно переходим ко второму предложению, затем к третьему и т. д. Но, например, в энциклопедии мы встречаемся и с иной методикой объединения тематически зависимых текстов: здесь термины, набранные *курсивом*, логически связывают статьи, расположенные в разных местах тома или в нескольких томах. Такое слово, которое является органической частью одного информационного блока и направляет читателя к иному информационному блоку, дает начало **гиперсвязи** (**hyperlink**, или просто **link**) между документами. Документ, который содержит в своем теле *гиперсвязи*, или **ссылки**, называется **гипертекстовым** (**hypertext**). Таким образом, энциклопедия — это печатный гипертекст. *World Wide Web* — высшая форма существования гипертекста: в *WWW* многие миллиарды текстовых, графических, аудио-, видео- и иных файлов, находящихся на разных компьютерах, объединены гиперсвязями в единое информационное поле.

В настоящее время Интернет охватывает практически весь земной шар; информационные потоки в нем не знают государственных границ и лимитируются лишь пропускной способностью коммуникационных линий. С точки зрения методов переноса данных, Интернет — это совокупность нескольких систем; предметом нашего рассмотрения будет одна из них, самая интересная — *World Wide Web*.

Если оставить за рамками обсуждения аппаратуру, координирующую работу сети, остальная часть *WWW* имеет двухуровневую структуру: в узлах находятся **веб-серверы** — компьютеры, на которых хранится информация,— а к ним тем или иным способом присоединены персональные компьютеры пользователей. Любой пользователь может войти в контакт с любым веб-сервером — если, конечно, умеет ориентироваться в сети.

Каждый сервер имеет свой числовой **IP-адрес** (например, 80.94.164.106), по которому компьютеры находят друг друга. Человеку же более удобна буквенная запись, поэтому серверу, кроме *IP*-адреса, приписывается уникальное название — **доменное имя**.

Доменное имя состоит из нескольких частей (минимум — двух), объединенных точками. Веб-сервер БГУ, например, имеет имя *www.bsu.by*, компании *Microsoft* — *www.microsoft.com*, а химического факультета Кембриджского университета — *www.ch.cam.ac.uk*.

Обычно — но не обязательно — имя веб-сервера начинается с букв *www*, затем указывается (полностью или сокращенно) владелец и/или название информационного массива, содержащегося на сервере.

Последняя часть имени (в наших примерах *.by*, *.com* и *.uk*) называется

суффиксом, или **доменом высшего уровня**. Двухбуквенный суффикс — это код страны, в которой зарегистрирован сервер (*.by* — Беларусь, *.uk* — Великобритания, *.de* — Германия, *.ru* — Россия и т. д.). Соединенные Штаты почти не используют свой географический суффикс *.us*, но вместо него записывают трехбуквенный функциональный: *.com* означает коммерческое учреждение, *.gov* — правительственное, *.edu* — образовательное, *.org* — иное некоммерческое, *.net* — связанное с координацией работы всей сети. (Отметим, что сервер, зарегистрированный в какой-либо стране, реально может находиться в совершенно иной точке земного шара. Причины бывают разные, обсуждать их здесь не будем, только приведем, хотя и далекую, но аналогию: немалая часть мирового торгового флота ходит не под своим, а под панамским флагом). Менее распространены, но тоже существуют, международные домены *.int*, *.biz*, *.info*, *.name*, *.museum*.

Весь информационный массив, находящийся на сервере и доступный внешним пользователям, называется **сайтом** (**site**). Сайтами же называют и автономные тематические разделы; так, например, на типичном университетском сайте обычно размещаются сайты факультетов, лабораторий, научных коллективов и даже отдельных сотрудников.

Веб-сервер хранит информацию в файлах и базах данных; по запросу требуемые сведения копируются и пересылаются на компьютер пользователя. Документ, который можно получить в ходе одного обращения к серверу, называется **веб-страницей** (**Web page**), или просто **страницей**. Обычно размер страницы соответствует такому объему материала, который способен разместиться на 1—3 экранах монитора. Страница не обязательно состоит только из текста; она может включать в себя графические, аудио-, видеофрагменты и исполняемые программные модули.

Каждая веб-страница имеет свой идентификатор — **URL** (***U**niform **R**esource **L**ocator*). Это ее адрес в сети; в некоторой степени *URL* выполняет функции библиографического описания, применяющегося в печатной литературе.

Какова может быть структура *URL* в простейших случаях, разберем на следующем примере:

http://www.chemistry.bsu.by/abc/intro/default.htm

Здесь первая группа символов, отделенная двоеточием и двумя косыми чертами (*http://*), означает, что с данным документом компьютер должен работать по протоколу *HTTP*. Это стандартный протокол, чаще всего встречающийся в *World Wide Web*.

Вторая группа (*www.chemistry.bsu.by*) — доменное имя веб-сервера, располагающего данной порцией информации.

Третья группа (*/abc/intro/default.htm*) указывает, в каком именно каталоге, подкаталоге, файле находится документ.

Подобным же образом построены *URL* файлов, хранящихся в *ftp*-архивах и доступных по протоколу *FTP*, например:

ftp://ftp.bsu.by/Educational_Resources/chem.doc

В тексте *URL* разрешено использование букв латинского алфавита, цифр и некоторых математических символов; пробелы запрещены, и при необходимости вместо пробела применяют знак подчеркивания (как в последнем примере в фрагменте *Educational_Resources*). Доменное имя записывается только строчными буквами, а вот в названиях каталогов и файлов заглавные буквы допускаются.

Для получения требуемой информации пользователь в идеальном случае должен знать *URL* конкретной страницы. *World Wide Web* — система огромная и к тому же очень динамичная, поэтому полного перечня абсолютно всех ресурсов *WWW* нет и быть не может в принципе. Тем не менее, существуют способы ориентации и в этом океане материала.

Проблема решается сравнительно просто, если пользователю известно имя сервера, на котором находятся нужные сведения. Дело в том, что каждый правильно организованный сайт имеет так называемую **Главную страницу** (**Home Page**), от которой разрастается система гиперсвязей. Если сравнить сумму информационных ресурсов сервера с книгой, то Главная страница была бы аналогом титульному листу, оглавлению и, нередко, аннотации. Как правило, *URL* Главной страницы имеет вид **http://доменное_имя_сервера/**. Например, поиск сведений о химическом факультете Белгосуниверситета логично было бы начинать с Главной страницы сервера БГУ, находящейся по адресу *http://www.bsu.by/*.

В тех случаях, когда место хранения документа неизвестно, обращаются к *поисковым системам, тематическим каталогам и метасайтам*.

Поисковые системы (**search engine**) постоянно сканируют веб-ресурсы специальными программами («*роботами*», или «*пауками*») и в результате создают базы данных — списки обнаруженных веб-страниц. Любой пользователь может обратиться к такой базе данных с запросом. Если в поисковом бланке указать слова, которые должны присутствовать в нужном документе, то *поисковая система* проанализирует имеющийся у нее материал и сообщит адреса подходящих сайтов и страниц.

Существует много поисковых систем — и универсальных, и специализированных; каждая характеризуется своими плюсами и минусами, связанными с объемом, тематикой, глубиной индексирования и т. д. В настоящее время наибольшую базу данных имеет поисковая система **Google** (*http://www.google.com/*) — более трех миллиардов ссылок на до-

кументы, написанные на разных языках. *WWW* Беларуси неплохо проиндексирован в *Open.BY* (*http://poisk.open.by/*), а русскоязычные ресурсы *World Wide Web* — в российских *Рамблере* (*http://www.rambler.ru/*) и *Яндексе* (*http://www.yandex.ru/*). Примером поисковой системы, специализирующейся исключительно на научной информации, является *Scirus* (*http://www.scirus.com/*).

Метапоисковые системы (**meta search engine**) тоже могут сообщить информацию об адресах сайтов и веб-страниц, но берут они сведения не из собственных баз данных, а заимствуют у нескольких обычных поисковых систем.

Базы данных поисковых систем создаются, в основном, автоматически, а вот **тематические каталоги** (**directory**) составляются человеком. Самый популярный универсальный тематический каталог — это *Yahoo!* (*http://www.yahoo.com/*). Веб-страницы и сайты в *Yahoo!* не просто рассортированы по тематическим разделам и подразделам, но и снабжены краткими содержательными аннотациями. Подобным образом построены и другие каталоги.

Следует отметить, что у больших каталогов есть свои собственные поисковые программы, а часть материала поисковой системы может быть скомпонована в форме тематического каталога.

Небольшие по размеру сайты, содержащие только ссылки на внешние веб-страницы, называются **метасайтами** (**metasite**). Те из них, которые имеют узкую специализацию, оказываются особенно полезными при поиске информации по определенной тематике.

Многопрофильные сайты, предлагающие широкий спектр информационных услуг, называются **порталами** (**portal**). Самый лучший химический портал — это **Chemweb** (*http://www.chemweb.com/*).

Прикладные программы, которые используются для просмотра веб-страниц, называются **браузерами** (**browser**). Самые популярные браузеры — *Microsoft Internet Explorer, Netscape Navigator, Opera* — функционально близки; *MS Internet Explorer* у нас более известен, поскольку входит в состав стандартного инсталляционного пакета операционной системы *Windows*.

Работа с веб-ресурсами может проходить в режимах **Browse** и **Search**.

В режиме *Browse* пользователь, зная *URL* нужной страницы, вызывает ее на экран и при необходимости по гиперсвязям переходит к следующим документам — «*листает*» (*browse*) их.

В режиме *Search* пользователь, не зная *URL*, ищет документ по некоему набору признаков. К режиму *Search* приходится прибегать при извлечении любой информации из любой базы данных.

2. ПОИСК ИНФОРМАЦИИ В ТЕКСТОВОЙ БАЗЕ ДАННЫХ

2.1. Структура текстовой базы данных

Информационный массив электронной базы данных состоит из логически целостных автономных частей, которые называются **записями** (***record***). В библиографической, реферативной, патентной базе данных записью является материал отдельной публикации (статьи, патента); в базе данных поисковой системы — материал одной веб-страницы. Кроме того, некоторые базы данных содержат и другие — вспомогательные — блоки различной степени сложности (например, оглавления, алфавитные, предметные указатели и т. п.).

Запись, в свою очередь, подразделяется на смысловые фрагменты — **поля** (***field***): текстовые (например, имена авторов, название статьи, реферат) либо числовые (например, дата публикации). В пределах конкретной базы данных перечень полей стандартен для всех записей; для разных баз данных списки используемых полей могут различаться.

2.2. Принципиальная схема поиска и извлечения информации

Каждая база данных обслуживается **поисковой программой** (***search engine***; ***search and retrieval software*** и т.д.), с помощью которой пользователь проводит целенаправленное извлечение нужной информации. Поисковые программы могут различаться структурой и своими возможностями, но принципы их действия во многом схожи. Простейший алгоритм работы с любой текстовой базой данных выглядит следующим образом: пользователь сообщает задание — некий термин; программа же отбирает и передает пользователю (**извлекает**) все записи, в которых этот термин упоминается.

Безусловно, обычно на практике используется более сложная схема. Задание чаще содержит не одно, а несколько связанных друг с другом слов. Поиск может проводиться не по всему объему записей, а только по отдельным полям (например, извлечение публикаций, содержащих термины **white phosphorus** в названии и **Black** в списке авторов). Программы способны обнаруживать однокоренные слова в разных грамматических формах (*phosphorus, phosphorous, phosphoric, polyphosphate, polyphosphates* и т. п.) и даже ранжировать записи по степени соответствия заданию. Тем не менее при планировании работы следует учиты-

вать, что **поиск не является смысловым**; в его основе лежит посимвольное сравнение текста задания с текстами, хранящимися в базе данных.

2.3. Запрос. Логические операторы

Успех поисковой работы во многом определяется тем, насколько правильно составлено **поисковое задание**, или **запрос**, (**search phrase**; **search terms**; **query**). В запросе пользователь перечисляет термины, которые должны быть (либо должны отсутствовать) в извлекаемых записях, и указывает взаимосвязь между этими терминами. Поисковое задание записывается в специальной графе (или графах) *поискового бланка* (см. п. 2.7).

В настоящее время наиболее распространен способ формирования запроса с помощью **логических** (*булевых*) **операторов**; соответствующая методика поиска называется **Boolean Search**.

Три логических оператора — «*и*», «*или*», «*и не*» — используются практически в любой базе данных. Их функции и правила применения стандартны для всех поисковых программ.

Логическое «*и*» обычно обозначается словом ***and***. В искомой записи обязательно *должны присутствовать оба термина*, связанные этим оператором.

Пример: По запросу **ion and atom** поисковая программа должна извлечь записи, в которых обязательно присутствуют оба слова: и *ion*, и *atom*.

Логическое «*или*» обычно обозначается словом ***or***. В искомой записи обязательно *должен присутствовать хоть один* из терминов, связанных этим оператором.

Пример: По запросу **ion or atom** поисковая программа должна извлечь записи, в которых обязательно присутствует либо слово *ion*, либо *atom*, либо оба эти слова.

Логическое «*и не*» обычно обозначается словом ***not***. В искомой записи обязательно *должен отсутствовать* термин, перед которым стоит этот оператор.

Пример: По запросу **ion not atom** поисковая программа должна извлечь записи, в которых обязательно присутствует слово *ion* и обязательно отсутствует слово *atom*.

Примечания:

1. Известны программы, в которых такие же операторы могут отображаться не словами, а символами (**&**, **|** , ~ и др.).

2. Некоторые поисковые программы используют иные варианты написания оператора ***not*** (***andnot***; ***and not***; ***but not***).

Поисковое выражение может содержать большое количество терминов, объединенных разными операторами. Следует учитывать, что логические операторы иерархически неравноценны. В частности, операция ***or*** всегда выполняется после операций ***not*** и ***and***, и задание

chromium and phosphate or sodium and silicate

для поисковой системы имеет смысл «Найти записи, содержащие слова *chromium* и *phosphate*, а также записи, содержащие слова *sodium* и *silicate*».

Весь список приоритетности операторов запоминать совершенно не обязательно — очередность выполнения действий достаточно просто и наглядно устанавливается с помощью скобок. Для запроса, упомянутого в предыдущем абзаце, допустима равноценная форма:

(chromium and phosphate) or (sodium and silicate)

Как в любой иной математической формуле, в логическом выражении действия в скобках выполняются в первую очередь; количество пар скобок не ограничивается; скобки могут быть вложенными.

Пример. Найти информацию о бессиликатных покрытиях или связующих на основе фосфатов магния или хрома. Результатом поиска должны быть записи, обязательно содержащие несколько групп терминов: во-первых, слова покрытие (*coating*) или связующее (*binder*); во-вторых, слова магний (*magnesium*) или хром (*chromium*); в-третьих, слово фосфат (*phosphate*); и, в-четвертых, обязательно не содержащие слово силикат (*silicate*). Сумма требований может быть выражена следующим способом:

(coating or binder) and (magnesium or chromium) and phosphate not silicate

(*Внимание*! В таком виде запрос еще не полностью готов к использованию; необходимые уточнения обсуждаются ниже.)

В последнее время во многих базах данных поисковую фразу предлагается записывать без операторов. Прежде чем делать это, пользователю следует детально разобраться, как конкретная программа интерпретирует такой текст. Дело в том, что на свободное место между терминами перед началом поиска автоматически подставляется **оператор по умолчанию** (**default operator**) — в одних программах ***or***, в других — ***and***. Понятно, что при разных операторах результаты поиска могут получиться принципиально различными.

2.4. Операторы расстояния

Сам факт одновременного присутствия неких терминов в записи совсем не обязательно означает, что эти термины связаны между собой по смыслу. Например, в записи, извлеченной по заданию **chromium and phosphate**, может идти речь не о фосфате хрома, а о хромированном корпусе анализатора бесфосфатных материалов.

Информационный шум в значительной степени ослабляется, если при построении поискового задания прибегнуть к **операторам расстояния** (**proximity operator**), устанавливающим допустимое удаление терминов друг от друга и их порядок расположения в извлекаемых записях.

Для обозначения неизменяемой **фразы**, состоящей из нескольких слов, и неизменяемой **строки символов** почти все поисковые программы используют **кавычки** (чаще всего, двойные).

Пример: по заданию **"chromium phosphate"** ведется поиск таких записей, в которых слова *chromium* и *phosphate* находятся в непосредственном соседстве, причем слово *phosphate* через пробел следует за словом *chromium*.

Пример: по заданию **"1,4-дибромбутен-2"** ведется поиск записей, в которых имеется именно такая, как указано внутри кавычек, последовательность цифр, букв и знаков препинания.

То, что написано внутри кавычек, программа считает единым поисковым термином, который можно объединять с другими терминами в логическое выражение.

Пример. По заданию **"ion-selective electrode" and chloride** извлекаются записи, содержащие одновременно: а) текстовый фрагмент *ion-selective electrode*; б) слово *chloride*. Записи, содержащие фрагмент *ion selective electrode* без дефиса между первыми двумя словами, извлекаться не должны — в этом случае нет 100 %-ного соответствия поисковому термину.

Примечание. В некоторых старых программах вместо кавычек все еще используется оператор ***adj*** (т. е. запрос **НИИ adj ФХП** означает то же, что **"НИИ ФХП"**).

Для обозначения максимально допустимой удаленности терминов друг от друга достаточно большое количество программ использует операторы ***near*** и ***with***. К сожалению, пока что отсутствует стандартизация в их написании и значении: одинаковые по виду операторы могут выполнять разные функции в разных базах данных, а для выполнения однотипных действий могут использоваться разные операторы.

Наиболее распространенный формат **термин1 near/n термин2** или

термин1 with/n термин2 (где ***n*** — целое положительное число) означает, что между первым и вторым терминами в тексте записи должно быть менее ***n*** иных слов.

Пример. По запросу **chromium near/2 coating** извлекаются записи, содержащие фрагменты *chromium coating* и *chromium containing coating*, но не извлекаются с фрагментом *chromium forms a hard coating* (термины излишне удалены друг от друга).

Операторы ***near*** и ***with*** могут устанавливать и последовательность расположения терминов в искомом документе (например, один из них разрешает любой порядок, а второй — только такой же, как в задании).

В некоторых базах данных с помощью операторов ***near*** и ***with*** указывают, что связанные ими термины должны находиться в одном поле записи; в одном абзаце; в одном предложении.

При формулировании запроса следует учитывать, что:

- в *WWW* используются разные варианты написания операторов расстояния: ***w/n , with/n, with; near/n, near***;
- операторы расстояния применяются для поиска только тех групп слов, которые находятся в одном и том же поле записи;
- операции ***with/n***, ***near/n*** и т. п. выполняются до операций ***not***, ***and*** и ***or***;
- при выборе оптимальной величины ***n*** следует помнить, что, подсчитывая расстояние между терминами, программа обычно пропускает служебные слова (артикли, предлоги).

2.5. Термины в поисковом задании

а) Шаблон

Для полного извлечения полезной информации необходимо, чтобы поисковое задание содержало в себе многочисленные варианты (грамматические формы, а нередко и однокоренные слова) терминов, на основе которых базируется поиск. Разные методики применяются для того, чтобы задание оставалось компактным, но учитывало многовариантность слов; в одной из них этим целям служат *шаблоны*.

Шаблон (**wildcard**) условным символом заменяет переменную часть термина.

Для обозначения *одного переменного символа* чаще всего используется вопросительный знак (**?**). Так, например, два слова *leucocyte* и *leukocyte* в поисковую фразу могут быть внесены как один термин **leu?ocyte**. Такой метод особенно полезен в тех случаях, когда требуется учесть отличия в английском и американском написании.

Для обозначения *любого количества символов* (в том числе, нулевого) обычно используется *звездочка (asterisk)* *****. Этот шаблон — **усечение** (**truncation**) — чаще применяется в конце слова (*right hand truncation*), но некоторые программы разрешают использовать его и в середине, и в начале слова. Шаблон позволяет включить в один поисковый термин все грамматические формы слов (например, не перечислять *phosphate* и *phosphates*, а записать **phosphate***) или однокоренные слова (не *phosphorus*, *phosphorous*, *phosphoric*, *phosphates*, а только **phosph***).

Если правая часть поискового термина заменяется шаблоном, в левой части должно оставаться не менее трех букв.

При выборе шаблона следует быть осмотрительным, так как существует вероятность неожиданного увеличения информационного шума. К примеру, по заданию **chlor*** можно получить не только материалы о *chlorine*, *chloride*, *chlorate*, но и о далеком по смыслу, но созвучном *chlorophyll*.

Пример. Найти информацию о бессиликатных покрытиях или связующих на основе фосфатов магния или хрома. На этот раз приведем окончательный вариант запроса, пригодный для использования в реальном поиске:

— со всеми операторами

(coating* or binder*) and (magnesium or chromium) and phosphate* not silicate*

— для программ, в которых ***and*** является оператором по умолчанию

(coating* or binder*) (magnesium or chromium) phosphate* not silicate*

Обе формы учитывают возможность присутствия в записях имен существительных в единственном и множественном числах (катион в названии соли может быть только в единственном числе).

Шаблон ***** в середине и начале слова сильно увеличивает длительность поиска, поэтому при работе в Интернет по возможности следует избегать такого его применения.

Замена букв на шаблон внутри кавычек не допускается. (Редкие исключения из этого правила будут обсуждаться при рассмотрении конкретных баз данных).

б) Учет словоформ (stemming)

Многие из современных программ способны самостоятельно расширять задание, варьируя окончания (иногда и суффиксы) слов запроса; такой режим **учета словоформ** терминов называется **stemming** (от *stem* — основа). Например, при использовании в качестве поискового термина слова **boiling** по этому алгоритму извлекаются записи со словами *boil*,

boils, *boiled*, *boiler*, *boilers*.

Программы, по умолчанию работающие в режиме *stemming*, обычно содержат средства его отключения — либо для отдельного термина поискового задания (в этом случае термин необходимо отметить условным знаком), либо для всего запроса (с помощью переключателя, кнопки и т. п., имеющихся на поисковом бланке). Такая необходимость может возникнуть, если читатель проводит узконаправленный поиск по четко выбранному набору слов.

Пример. В запросе о температуре кипения вещества следует учесть, что в искомой статье слово *температура* может быть в любом падеже, но слово *кипения* должно присутствовать только в одной грамматической форме, а именно в родительном падеже единственного числа.

в) Регистр букв (Case Sensitivity)

Обычно для поисковой программы строчные и заглавные буквы абсолютно эквивалентны (**case insensitive**). Например, в задании можно записать либо *bell*, либо *BELL*, либо даже *beLL* — итог будет один и тот же: пользователь получит информацию об авторе по фамилии *Bell*, о компании *Bell Communications Research, Inc.*, о колоколах, куполах, конусах (*bell*). (Более избирательного результата добиваются, указывая поля, в которых планируется вести поиск — об этом см. п. 2.7).

Программы, *чувствительные к регистру, т. е. различающие строчные и заглавные* буквы, (*case sensitive*), все же существуют, поэтому пользователь должен узнать заранее, как проводится поиск по имени собственном в конкретной базе данных.

г) Стоп-слова (Stopwords)

В каждом языке есть служебные слова, которые используются часто, но сами не несут большой смысловой нагрузки: артикли, предлоги и т. д. Для того чтобы ускорить процесс обнаружения требуемой информации, поисковая программа может игнорировать присутствие таких слов и в запросе, и в анализируемых документах. Термины, которые исключаются из поиска, называются **стоп-словами** (**stopword**).

Примеры стоп-слов: *a, the, of* (англ.); *в, из, над* (рус.).

Реакция на стоп-слово внутри кавычек бывает разной: одни программы его отбрасывают, другие учитывают при поиске.

Достаточно большое количество программ, особенно из числа обслуживающих научные базы данных, не разделяет слова на полноценные и неполноценные и допускает поиск по любому термину, в том числе состоящему из одной буквы или цифры.

2.6. Формулировка запроса на естественном языке

В последнее время все большее распространение получает еще один способ отображения запроса — **на естественном языке** (**natural language query**; **free-text search**). В этом случае связь между словами устанавливается не с помощью логических операторов, а по правилам грамматики, и поисковое задание имеет вид обычного предложения либо фрагмента предложения. Допустимы даже такие варианты как «*Где можно найти информацию о связующих, в состав которых входит фосфат хрома?*».

Поисковая программа на первом этапе обрабатывает такое задание достаточно формально: отбрасываются служебные слова (*stopwords*), распознаются известные данной программе устойчивые словосочетания, термины объединяются операторами *or*, после чего начинается процесс поиска, аналогичный *Boolean search* в режиме *stemming*. Результатом поиска является очень обширный перечень записей, которые ранжируются по степени соответствия запросу (эта стадия обсуждается в п. 2.8).

Настоящий смысловой анализ текста пока что возможен только при небольших объемах материала узкой тематической направленности; он требует огромной мощности компьютера, поэтому еще не скоро будет применен в реальной поисковой работе.

2.7. Поле как элемент поискового задания. Поисковый бланк

Ранее (п. 2.5, в) на примере термина *bell* уже отмечалось, что поиск по всему объему записей может дать совершенно разные по смыслу результаты (*Bell* — автор; *Bell* — фирма; *bell* — предмет и понятие). Этот недостаток устраняется, если задание содержит ссылку на то, в каком поле должен находиться термин, интересующий пользователя.

Многие поисковые программы способны проводить поиск и по всем записям в целом, и отдельно по каждому из полей, и по их комбинациям. Например, в патентной базе данных не составляет труда найти опубликованный в *A* году патент автора *B*, если известно, что патент принадлежит фирме *C*, в его названии или реферате упоминается вещество *D*, а по тематике его можно отнести к классу *E*.

Существуют две методики составления задания для целенаправленного **поиска по полям** (**fielded search**).

В первом варианте *в тексте запроса* вместе с поисковыми терминами указываются коды соответствующих полей. Каждая программа имеет

свой список обозначений полей и свои правила формулирования этой части задания; запоминать их имеет смысл, только если с базой данных приходится работать особенно часто.

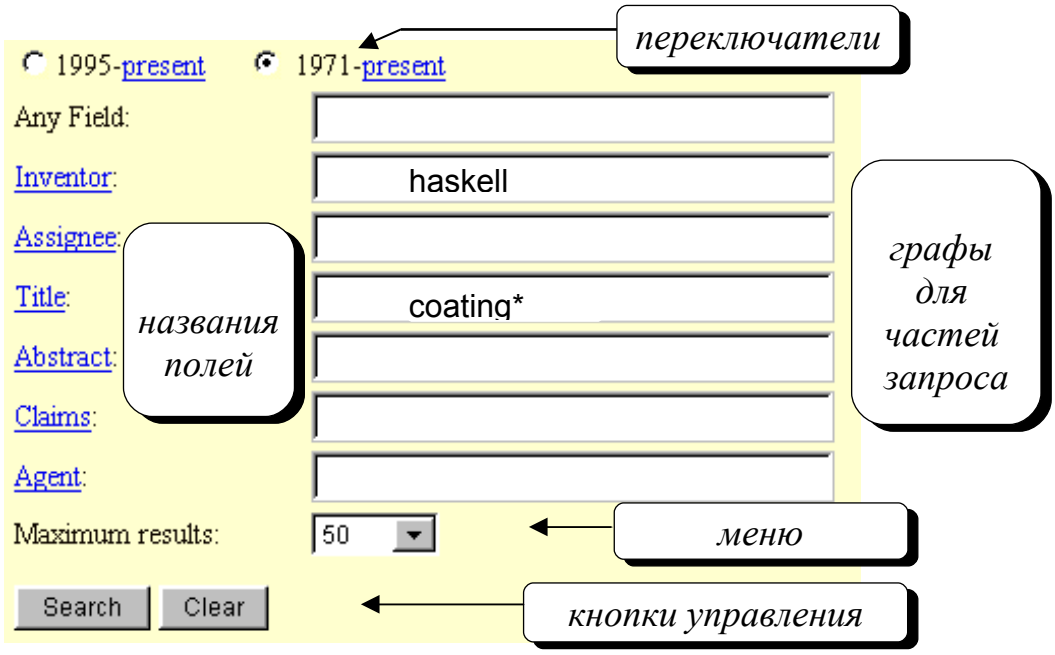
Примеры запросов для разных программ:

IN/haskell and TTL/coating*
haskell intitle:coating intitle:coatings

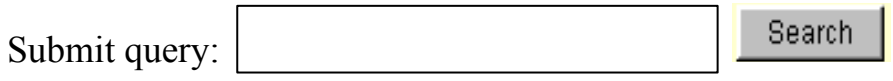
При большом количестве терминов текст может получиться достаточно громоздким, поэтому к такой методике прибегают в случаях тонкого информационного поиска.

В рутинной работе более удобен второй вариант: заполнение **поискового бланка**, *содержащего графы для разных полей*. На бланке могут присутствовать переключатели и меню, с помощью которых пользователь уточняет область поиска.

Пример бланка:



Отметим, что простейший поисковый бланк содержит только одну графу, предназначенную для записи поискового задания, и, по меньшей мере, одну кнопку, задающую команду начать поиск, например:



База данных может иметь несколько поисковых бланков.

Более простой по форме бланк обычно называется **основным** (**Basic**). Он предназначается для поиска в небольшом количестве полей либо во

всей записи в целом. Запрос в основном бланке формулируется с использованием логических операторов.

Бланк, содержащий большее количество граф, обычно называется **усложненным** (**Advanced**). Он предназначается для комбинированных поисковых заданий. Многие базы данных имеют усложненные поисковые бланки такой структуры, что запрос в них может быть сформулирован без логических операторов.

2.8. Список результатов поиска. Релевантность

Если программа в ходе поиска обнаруживает несколько записей, к пользователю поступает список результатов — оглавление, содержащее гиперсвязи к самим записям.

При работе с большими базами данных количество извлекаемых записей может достигать десятков, сотен и даже тысяч, и, конечно же, не стоит пытаться изучать все. Дело в том, что программа **ранжирует** документы — размещает их в порядке уменьшения **релевантности** (**relevancy**), т. е. *степени соответствия поисковому заданию*. Пользователь, продвигаясь от более соответствующего, стоящего во главе списка, к менее соответствующему, сам решает, на каком номере заканчиваются удовлетворяющие его документы.

При расчете релевантности, или *ранга*, принимаются во внимание следующие факторы:

- чем больше поисковых терминов присутствует в записи, чем ближе они расположены друг к другу, тем больше ее релевантность;
- ранг растет, если термин встречается в записи большее число раз, если он упоминается в названии документа;
- ранг зависит от плотности терминов: при прочих равных условиях короткие записи оцениваются выше, чем длинные;
- слова, редко встречающиеся в базе данных, имеют больший вес;
- релевантность повышается, если слова в записи расположены в том же порядке, в котором они расположены в запросе (этот фактор учитывается только частью программ);
- иногда учитывается расположение термина в документе: ранг увеличивается, если термин присутствует в начале записи.

Принципы поисковой работы, упомянутые в этой главе, стандартны для большинства научных баз данных, имеющихся в Интернете.

При работе с поисковыми системами следует учесть некоторые особенности применения тех же принципов, и об этом — следующая глава.

3. ПОИСКОВЫЕ СИСТЕМЫ

3.1. Структура поисковой системы

Поисковая система (**search engine**, **search service**) — это комплекс программ, баз данных и инструментов, предназначенных для сбора информации об имеющихся в Интернете документах и для извлечения нужных документов по признакам, указанным пользователем.

Основу любой поисковой системы составляет специальная программа, называющаяся **роботом** (*robot*), или **пауком** (*spider*), которая постоянно *сканирует веб-пространство*. Переходя по гиперсвязям с одной страницы на другую, с сайта на сайт, робот накапливает сведения о всех встреченных на его пути документах. На ранних этапах развития Интернета робот запоминал лишь адреса и названия страниц; в настоящее время роботы ведущих поисковых систем копируют веб-страницы целиком.

На основе полученной информации робот строит свои базы данных — **индексы** (*index*), в которых зафиксировано, на каких страницах встречается то или иное слово и в каком месте документа оно расположено.

С третьей частью поисковой системы — **поисковой программой** — каждый работающий в Интернете сталкивается непосредственно. *Поисковая программа* по запросу пользователя анализирует *индекс* и сообщает результат — какие именно веб-страницы удовлетворяют условиям, сформулированным в поисковом задании.

Работа по сканированию веб-пространства и построению индекса скрыта от публики, поэтому нередко именно поисковую *программу* называют поисковой *системой*.

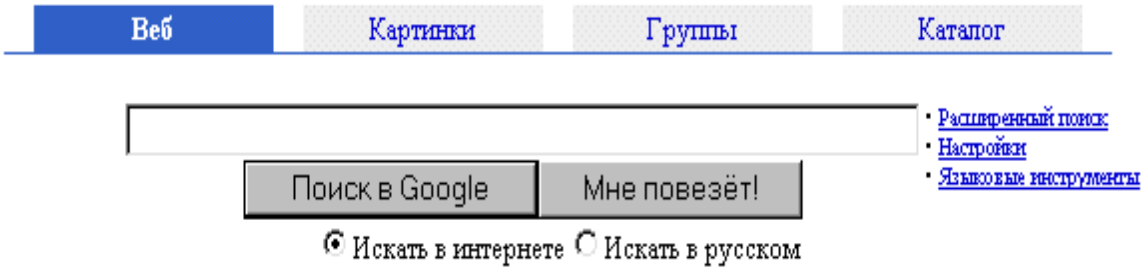
Индексы, сформированные роботами различных поисковых систем, хотя и перекрываются по своему содержанию, но не являются идентичными. Список страниц, попавших в индекс, зависит от многих факторов, в том числе от того, каким был исходный перечень ссылок, с которых робот начинал сканирование; сколько страниц с обнаруженного сайта робот включает в свой индекс (большая часть поисковых систем ограничивается главной и несколькими из тех, к которым ведут гиперсвязи с главной страницы); как часто робот посещает обнаруженный сайт (реально частота варьируется от ежеминутных для новостных сайтов до ежемесячных или даже более редких посещений); от других настроек робота (например, есть ли ограничения по доменам, по языку страницы и т. п.); от того, насколько простым является процесс добавления новых адресов по заявкам авторов сайтов.

В настоящее время Интернет сканируется достаточно большим коли-

чеством роботов, но самые большие базы данных создаются и постоянно обновляются роботами *Google*, *FAST*, *Teoma* и *Inktomi*. Индексы, сформированные ими, могут использоваться несколькими поисковыми системами — либо в чистом виде, либо в сочетании с другими источниками информации. Так, например, поисковая система *Google* использует индекс, созданный собственным роботом, и базу данных каталога *Open Directory*. В свою очередь, к индексу *Google* обращаются *Yahoo!*, *Netscape Search* и др. *Inktomi* вообще не работает с индивидуальными пользователями, а только готовит информационное и программное обеспечение для иных поисковых систем, например, для *MSN Search*.

Для доступа к поисковой системе создается сайт, на главной странице которого размещается интерфейс поисковой программы. Самый важный элемент такой страницы — это **основной поисковый бланк**, где пользователь формулирует *поисковое задание*, или *запрос*. На бланке, кроме редактируемого поля и кнопки, дающей команду начать поиск, могут быть размещены кнопки, выключатели и переключатели, конкретизирующие область поиска, гиперсвязи к другим элементам поисковой системы, в том числе, к усложненному бланку, к странице настройки интерфейса, к описанию правил работы с системой («*Help*»).

Пример *основного бланка*:



Усложненный бланк, в отличие от основного, содержит несколько редактируемых полей; бланк, как правило, насыщен переключателями, выключателями, выпадающими меню, которые позволяют создать сложное поисковое задание, не используя (или почти не используя) логические операторы в явной форме. Усложненный бланк обычно предназначается для проведения узконаправленного поиска.

3.2. Особенности информационного поиска

Поскольку поисковые системы предназначены для использования широкой публикой, то их создатели стремятся упростить процесс взаимодействия пользователя с программным обеспечением. На стадии формулирования запроса это проявляется в ненавязчивых рекомендациях избе-

гать использования операторов в задании, хотя система сама по себе способна обрабатывать сложные логические выражения. Изначально постулируемая нечеткость запроса предполагает и некоторую размытость критериев, по которым отбираются результаты поиска. *Избыточность извлекаемой информации* — вот с чем сталкивается пользователь, даже если он сформулировал узконаправленное задание с помощью операторов, разрешенных в конкретной системе. Ситуация чем-то напоминает ситуацию в магазине, когда, попросив черные носки, покупатель, конечно же, их получает, но заодно ему предлагают коричневые и белые носки, черные ботинки, темный костюм и, наконец, любую иную одежду.

Определенный смысл в таком подходе есть, поскольку ресурсы *WWW* огромны и нестандартизированы по содержанию и форме (здесь речь не идет о научных базах данных), а пользователь не всегда способен предугадать, какими словами может быть высказан ответ на его вопрос. Более того, при внешнем подобии запросов цели поиска могут существенно различаться — ведь задача *«Найти иголку в стогу сена»* (а именно ее решают с помощью поисковых систем) на самом деле означает многое:

Найти известную иголку в известном стогу сена;
Найти известную иголку в неизвестном стогу;
Найти неизвестную иголку в неизвестном стогу;
Любую иголку в любом стогу;
Самую острую иголку в стогу;
Большую часть самых острых иголок в стогу;
Все иголки в стогу;
Доказать, что в стогу иголок нет;
Что-нибудь похожее на иголки в любом стогу;
Сообщить, когда в стогу появятся новые иголки;
Иголки, стога — в общем, что-нибудь;
А где эти стога?
(Мэтью Колл)

Большой объем информации, извлеченной в результате поиска, устроит пользователя, только если поисковая программа способна высококачественно ранжировать обнаруженные документы по степени их соответствия запросу. Современные поисковые системы соперничают друг с другом не столько в области увеличения индексов, сколько в создании эффективных алгоритмов расчета релевантности. В настоящее время лидером на этом поле является *Google*, которая, кроме уже традиционных параметров (наличие терминов в документе, их частота, расположение, плотность), при определении релевантности веб-страницы учитывает ее

популярность в Интернете. Согласно подходу *Google*, чем больше ссылок направлено к данной веб-странице, тем она авторитетнее, тем выше ее ранг и тем она более ценная для пользователя; чем выше ранг страницы, тем, в свою очередь, больший вес имеют ссылки с нее. Как показывает практика, такой учет стихийного взаимного рецензирования весьма плодотворен, и в списках результатов поисковой системы *Google* самые подходящие документы действительно оказываются на первых местах.

3.3. Формулирование запроса на основном бланке

Большая часть поисковых систем рекомендует формулировать поисковое задание в виде набора терминов, разделенных пробелами; меньшая часть — в виде обычного вопроса на естественном языке, с соблюдением всех правил грамматического согласования между словами в предложении. Второй подход означает отнюдь не особо высокую интеллектуальность поисковой программы, а только то, что такая система имеет базу данных со стандартными ответами на стандартные вопросы. При поиске научной информации, т. е. не стандартной, лишние слова в запросе (*какой*, *почему* и т. п.) могут ухудшить ранг ценных документов в списке результатов, поэтому при работе с любой системой в поисковое задание целесообразно включать *только слова, несущие смысловую нагрузку* и действительно важные для данной поисковой задачи.

Многие программы при подсчете релевантности учитывают взаимное расположение терминов на странице — следовательно, в поисковом задании слова должны быть расположены в таком *порядке*, в котором пользователь хотел бы их увидеть в искомом документе. Если для данного поиска порядок слов не имеет значения, *на первое место* стоит ставить термин, самый важный с точки зрения пользователя — некоторые программы именно первому термину придают наибольший коэффициент при ранжировании.

Если запрос состоит только из слов, разделенных пробелами, программа, прежде чем начать поиск, сама размещает между терминами *операторы по умолчанию*, обычно — логическое «*и*».

Пример: запрос **aluminum phosphate coating** автоматически преобразуется в **aluminum and phosphate and coating**.

В идеальном случае список результатов такого поиска должен был бы состоять только из документов, содержащих *все* упомянутые в задании слова. Реально же список может быть расширен за счет документов, содержащих *не все* поисковые термины. Различные поисковые системы поразному относятся к такому расширительному толкованию запроса, но

замечена тенденция: чем длиннее поисковое задание, тем больше вероятность его нестрогого исполнения.

Небольшая часть поисковых программ использует логическое «*или*» в качестве оператора по умолчанию — понятно, что списки результатов поиска в таких системах должны быть особенно большими.

Все поисковые системы разрешают использовать логические операторы «*и*», «*или*», «*и не*» для построения узконаправленных поисковых заданий, однако формы отображения этих операторов бывают разные.

Логическое «*и*» может обозначаться словом ***and*** (большинство систем), символом ***&*** (некоторые системы), однако, почти все системы рекомендуют применять для этих целей знак «+» (*плюс*), записываемый слитно перед термином. Согласно описаниям поисковых программ, знак «*плюс*» показывает, что соответствующее слово обязательно должно присутствовать в искомом документе, поэтому одинаково правильны и одинаковый смысл имеют следующие варианты:

phosphate and silicate либо **phosphate & silicate**
phosphate +silicate
+phosphate +silicate

Если логическое «*и*» является оператором по умолчанию (а это почти всегда), то использование ***and*** или **+** в задании совершенно излишне. (Более того, в некоторых специфических случаях «*плюс*» неправильно интерпретируется поисковой программой.) В таких системах самый лучший вариант запроса — без операторов:

phosphate silicate

Только в одной ситуации применение «*плюса*» действительно оправданно — если поисковым термином является *стоп-слово*. Например, количество записей в англоязычной базе данных можно оценивать, проводя поиски по запросам **+a** и **+the**.

Логическое «*или*» может обозначаться словом ***OR*** (большинство систем) или, реже, символом, например, вертикальной чертой **|** . Для того, чтобы программа сумела отличить оператор ***OR*** от стоп-слова, его следует записывать заглавными буквами. *Примеры*:

phosphate OR silicate
phosphate | silicate

Логическое «*и не*» может обозначаться словами ***NOT***, ***AND NOT*** (часто — именно заглавными буквами) или символом, например, тильдой **~**, однако, почти все системы рекомендуют применять для этих целей знак «**-**» (*минус*), записываемый слитно перед термином. *Примеры*:

phosphate NOT silicate
phosphate ~silicate
phosphate -silicate

Для того, чтобы поисковая программа правильно интерпретировала запрос, термин, обозначенный оператором «*и не*», стоит располагать в самом конце задания.

Комбинированные запросы с несколькими разными операторами разрешены во всех поисковых системах. Абсолютное большинство поисковых систем придерживается стандартного порядка выполнения операций: сначала ***and*** и ***not***, затем ***or***.

Внимание! Поисковая система *Google* действует наоборот: сначала исполняется ***or***, затем ***and***.

Применение **круглых скобок** для указания очередности логических операций разрешено не во всех поисковых системах (например, *Google* круглые скобки не использует; *AlltheWeb* использует, но совершенно в иных целях).

Фразы и строки символов, ограниченные двойными **кавычками** слева и справа, правильно воспринимаются всеми поисковыми системами.

Операторы расстояния (аналоги ***near***) используются только несколькими поисковыми системами.

К **регистру** букв **нечувствительны** почти все поисковые системы. Только избранные программы (например, *Яндекс*) работают по более сложному алгоритму: если поисковый термин начинается с заглавной буквы, программа извлекает только те документы, в которых это слово начинается с заглавной буквы; если поисковый термин начинается со строчной буквы, то при извлечении материала регистр не учитывается.

Шаблоны не используются почти всеми поисковыми системами.

Поиск по словоформам терминов (*stemming*) проводится в одних системах (например, в *Рамблере*, *Яндексе*), не проводится в других (*Google*) и проводится только в части базы данных третьих (*AlltheWeb*, *MSN Search*).

Поиск в избранных полях веб-страниц возможен во всех поисковых системах. Проведение такого типа поиска имеет смысл только при решении некоторых узкоспецифических задач, поскольку абсолютное большинство информации *HTML*-документа размещено в одном обширном поле — его текстовой части.

Остальные поля помечены специальным образом в *HTML*-коде веб-страницы, и к ним, в частности, относятся:

- *название документа* — это тот текст, который отображается в строке заголовка (в верхней части окна) браузера;

- *URL ссылок* на иные страницы;
- *тексты ссылок* — те фрагменты страницы, от которых начинаются гиперсвязи; они обычно подчеркнуты либо выделены цветом;
- *URL графических элементов* страницы;
- *тексты подписей к графическим элементам*, появляющиеся во всплывающем окошке, когда курсор подведен к рисунку;
- *описание (Description)* — составленная автором аннотация;
- *ключевые слова (Keywords)* — составленный автором перечень терминов, характеризующих тематику страницы.

  Последние два поля не видны пользователю, если страница просматривается в браузере; с ними можно ознакомиться, открыв страницу в текстовом редакторе, например, в *Блокноте* (*Notepad*).

В большинстве систем для поиска по определенному полю задание записывается в формате **код_поля:термин**. Коды полей разнятся в разных системах.

Пример. Найти документы, в названиях которых присутствует слово ***journal***.

В *Google* запрос имеет вид: **intitle:journal**
В *AlltheWeb* запрос имеет вид: **title:journal**

Запрос по полю можно совмещать с запросом по всему документу.

Пример. В индексе *Google* требуется обнаружить документы, содержащие слово ***abc*** в адресе ссылки и слова ***free journals*** в тексте всей страницы. Задание имеет вид:

inurl:abc free journals

3.4. Формулирование запроса на усложненном бланке

Усложненный бланк (***Advanced Search Form***) содержит несколько редактируемых полей, каждое из которых предназначено для отображения определенной части поискового задания. В одних полях следует перечислять слова, которые должны присутствовать или отсутствовать в искомом документе, в других, уточняющих,— указывать иные признаки (время создания документа, формат и т. п.). В зависимости от своего содержания, запрос может быть либо рассредоточен по многим графам бланка, либо сконцентрирован в одной из них.

Если на то нет иных указаний на бланке, все графы объединены в единое целое логическим «*и*».

Операторы в явной форме редко используются на усложненном

бланке. Для того чтобы назначить *логическую взаимосвязь* между терминами, их просто нужно разместить по соответствующим графам, подписанным, например, так:

- **must include** (**contain** и т. п.) или **all of the words** — графа для слов, которые обязательно *должны быть* в искомом документе (на основном бланке перед такими поисковыми терминами стояли бы операторы ***and***);
- **may include** или **any of the words** — для слов, которые *могут быть* в искомом документе (на основном бланке перед такими терминами стояли бы операторы ***or***);
- **must not include** — для слов, которые *должны отсутствовать* в искомом документе (на основном бланке перед такими терминами стояли бы операторы ***not***);
- **the exact phrase** — для *фразы* или строки символов (на основном бланке эти термины были бы заключены в кавычки).

Если в такой графе записывают несколько терминов, их отделяют друг от друга пробелами.

Для поиска *по полям* на бланке могут присутствовать специальные графы, но чаще эти параметры поиска назначаются в выпадающем меню (содержащем, например, пункты **in page**, **in title**, **in URL** и др.).

Поиск по полям удобнее проводить именно с усложненного бланка, а не с основного, потому что в этом случае нет необходимости запоминать нестандартные обозначения кодов полей.

Достаточно часто на усложненных бланках присутствуют элементы (графы, меню и др.), с помощью которых можно указывать желаемый временной интервал создания искомого документа. К достоверности информации о том, когда была написана или изменена веб-страница, следует относиться скептически. По разным причинам каждый второй веб-сервер не сообщает эти сведения роботу в момент сканирования, поэтому в базу данных в таких случаях вместо даты создания документа вносится дата его последнего индексирования.

3.5. Результаты поиска

Записав задание, пользователь нажимает управляющую кнопку на бланке («**Поиск**», «**Search**», «**Find**») или клавишу **Enter** на клавиатуре (это допускается почти всеми поисковыми системами). Поисковая программа анализирует свою базу данных (или базы данных) и высылает список результатов поиска. Если список большой, на экран выводится только его первая часть (обычно 10—20 позиций); к остальным порциям

направлены гиперсвязи со страницы списка. Результаты поиска отсортированы по релевантности, поэтому страницы, наиболее удовлетворяющие заданию, обычно находятся в первой или второй порции списка.

В списке обычно приводятся следующие сведения: название веб-документа, его адрес и гиперсвязь к нему; сведения о документе, например, объем, дата создания или индексирования, формат файла; фрагмент текста, в котором присутствуют поисковые термины (эти термины могут быть выделены жирным шрифтом).

Пример фрагмента списка (поисковая система Google):

Analytical Chemistry Resources
Resources relating to **Analytical Chemistry** and Instrumentation, including: Journals, Research Groups, General **Resources**, Societies and Software. **...**
Description: Links to journals, societies, software, conferences and related areas, maintained for the Hamilton...
Category: Science > Chemistry > Analytical > Directories
www.netaccess.on.ca/~dbc/cic_hamilton/anal.html - 13k -
Cached - Similar pages

Пользователь анализирует текст и решает, стоит ли по гиперсвязи переходить к соответствующему оригиналу документа либо требуется вызывать следующую порцию списка.

Кроме перечня обнаруженных документов, страница результатов содержит дополнительные сведения и инструменты, которые могут быть полезны для уточнения поиска или изменения его направления.

а) Статистические данные

На странице указывается, сколько документов, удовлетворяющих заданию, обнаружено в базе данных. Если это число большое, а пользователь в первых порциях списка не находит ответа на свой вопрос, значит, следует изменить задание.

Некоторые поисковые системы сообщают статистические данные о каждом поисковом термине. Если в базе данных обнаружено мало страниц, содержащих конкретный термин, следует проверить, не допущена ли орфографическая ошибка при его написании в запросе.

б) Тематически близкий поиск (*Related Search*).

Поисковая система может подсказать, в виде каких словосочетаний термины данного запроса наиболее часто встречаются в заданиях других пользователей. Такая информация полезна, если возникает необходимость изменить направление поиска.

Перечень тематически близких запросов размещается на странице результатов в разделе, который озаглавлен, например, так:

У нас также ищут: ...
Related searches ...

в) Группирование страниц одного сайта

Если на сайте обнаружено несколько страниц, удовлетворяющих условию запроса, в список в явной форме включается один-два документа, характеризующихся наибольшими коэффициентами релевантности. Перечень остальных страниц этого сайта выводится на экран по ссылке:

Еще с этого сайта
See results from this site only.

г) Аналогичные страницы (*Find Similar*)

Программное обеспечение поисковой системы способно проанализировать веб-страницу и определить, какие слова и словосочетания особенно часто встречаются в документе и каково их взаимное расположение. Этих сведений может быть достаточно для того, чтобы обнаружить в базе данных иные страницы с такими же или близкими параметрами.

Для поиска страниц, аналогичных данной, в списке результатов может находиться гиперсвязь, например, с таким текстом:

Найти похожее
Find similar

д) Повторный поиск

Если в результате поиска обнаружено слишком много документов и в первых порциях списка нужная информация отсутствует, то стандартным действием является изменение запроса и начало нового поиска. Многие поисковые системы экономят время и усилия пользователя, позволяя проводить повторный поиск не по всей базе данных, а только по той ее части, которая была выделена в ходе предыдущего этапа работы.

Эта операция выполняется либо переходом по ссылке Search within these results, либо с поискового бланка при переключателе, включенном в положение «*искать в найденном*».

е) Копии, сохраненные роботом

Роботы поисковых систем, накапливая информацию для своих индексов, сохраняют копии страниц, и эти копии (полные либо только текстовые их части) могут быть доступны пользователю. Со страницы результатов поиска к ним ведут гиперсвязи «Cached», «Сохранено» и т. п.

Обращение к копии целесообразно в следующих ситуациях:

- если оригинал размещен на медленном или работающем с перебоями сервере (тогда именно копия более доступна, поскольку связь с

ведущими поисковыми системами обычно устойчива);

- если требуется быстро сориентироваться, насколько данная страница отвечает условиям запроса (*Google*, например, поисковые термины в копиях выделяет шрифтом разного цвета).

Следует учитывать, что автор может изменить или даже удалить страницу после ее индексирования, поэтому к моменту поиска копия не обязательно должна быть идентична оригиналу. В такой ситуации копия становится единственным источником уже исчезнувшей информации.

В базе данных поисковой системы копия остается неизменяемой до очередного индексирования соответствующего сайта.

3.6. Поисковые системы и кириллический текст

Поиск документов, написанных кириллицей, принципиально не отличается от поиска документов с латинским шрифтом. Русскоязычные ресурсы *WWW* наиболее полно отражены в российских базах данных *Рамблер* и *Яндекс*; для извлечения материалов белорусского происхождения (на белорусском и русском языках) предпочтительна *Open.BY*. Поисковые программы этих и некоторых иных систем мощные, учитывают словоформы терминов и прекрасно обрабатывают запросы с операторами.

Российские службы обычно индексируют следующие источники:

- сайты домена ***.ru*** — вне зависимости от языка веб-страниц;
- русскоязычные сайты стран СНГ;
- сайты иных доменов (например, ***.com***, ***.org***), если информация изложена на русском языке или ее содержание связано с Россией.

В базах данных превалируют документы на русском языке, заметную долю составляют англоязычные страницы, кроме того, встречаются документы и на иных языках (в том числе, на белорусском). Веб-страница может быть многоязычной (пример фрагмента такого текста: «*В журналах The Analyst и Zeitschrift für Naturforschung рассматривается ...*»).

Многоязычная среда базы данных не является препятствием при поиске, поскольку в задании допускается объединение терминов, записанных кириллицей и латиницей. Опции «*русский*» и «*английский*», встречающиеся на некоторых поисковых бланках, следует понимать в расширительном толковании как «*в основном, написанное кириллическим шрифтом*» и «*в основном, написанное латинским шрифтом*».

В заключение отметим, что кириллические документы (русскоязычные, в том числе) неплохо представлены в базах данных американских (а по сути, международных) поисковых систем *Google*, *All the Web*.

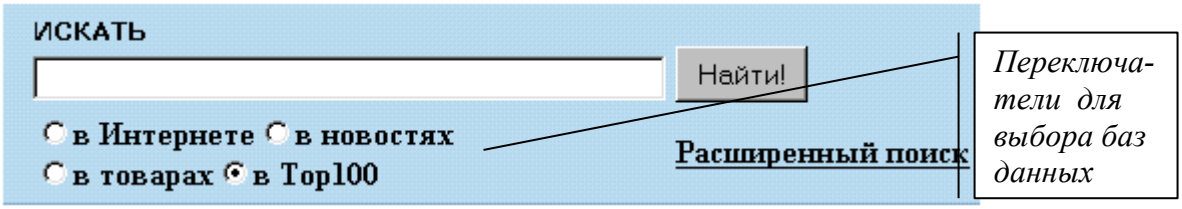
4. ПОИСКОВАЯ СИСТЕМА *РАМБЛЕР* http://www.rambler.ru/

4.1. Главная страница

На портале **Рамблер** (*http://www.rambler.ru/*) размещается одна из самых мощных поисковых систем русскоязычного Интернета. Поисковая программа *Рамблера* извлекает информацию из нескольких независимых источников, в том числе из тематического каталога *Rambler's Top100* и из базы данных, сформированной роботом-пауком.

Все тематические разделы каталога **Rambler's Top100** перечислены на главной странице портала. Структура каталога одноуровневая — подразделы в его разделах отсутствуют. Так, например, в категорию «Наука» собраны ссылки более чем на тысячу естественнонаучных, гуманитарных и даже астрологических сайтов, поэтому работа с каталогом в режиме *Browse* трудоемка и неэффективна.

Здесь же на главной странице портала находится **основной поисковый бланк** для работы в режиме *Search*:



Если переключатель установлен в положение «в Top100», поиск ведется только в каталоге *Rambler's Top100*. Каждая из записей каталога состоит из названия веб-страницы и краткой аннотации, и именно эти данные (но не содержимое веб-документа !) анализируются в процессе такого информационного поиска.

Если же переключатель установлен в положение «в Интернете», то поиск ведется и в каталоге *Rambler's Top100*, и в базе данных, сформированной роботом-пауком. А вот эта база данных содержит полные тексты HTML-документов.

Правила формулирования поискового задания одинаковы и для поиска в каталоге, и для поиска в базе данных робота.

4.2. Основной поисковый бланк: формулировка задания

- Простейшая форма запроса — набор поисковых терминов, разделенных пробелами.

Оператором по умолчанию является ***and***.
Примеры запроса:

редактор химических формул для Windows
окисление медь поверхность электрод

- Для проведения узконаправленного поиска используются логические операторы ***and*** (синоним: ***&***), ***or*** (синоним: **|**), ***not*** (синоним: **!**), а также круглые скобки.

  Пример запроса: **фосфат and (натрия or калия) not лития**
  Равноценный вариант: **фосфат & (натрия | калия) ! лития**

- В общем случае, программа **нечувствительна к регистру**. Однако если задание состоит из *двух, трех или четырех* терминов, каждый из которых написан с заглавной буквы, то извлекаются *только* те документы, в которых указанные термины, во-первых, записаны тоже с заглавной буквы, во-вторых, находятся очень близко друг к другу.
  Например, по заданию **Journal Chemistry** извлекаются документы с фрагментами:

  Journal of Analytical Chemistry
  "UKRAINIAN CHEMISTRY" JOURNAL IN 2001

- Программа работает в режиме **stemming**, причем и с русскими, и с английскими терминами.
- Фиксированная строка символов или фраза обозначается **двойными кавычками**, *например*:

  "1, 3-динитробензол"
  "температура кипения бензола"

- Внутри кавычек *stemming* отключен. Это можно использовать при поиске документа, содержащего термин в заданном падеже и числе. Например, по заданию

  правило "фаз"

  будут извлечены документы с фрагментами

  Правило фаз Гиббса
  в соответствии с правилом фаз

  но не будет извлечен документ, содержащий фрагмент

  как правило, эта фаза длится

- В запросе можно указать то максимальное расстояние, на котором термины имеют право находиться в искомой веб-странице (по-

дробности далее в п. 4.3).

4.3. Формулировка запроса — дополнительные сведения

- Оператор **NOT** в поисковом задании должен упоминаться не более одного раза. Если требуется исключить большее количество терминов, можно использовать конструкцию, например, такого типа:

  фосфат not (лития or рубидия or цезия)

- Кавычки не увеличивают чувствительность программы к регистру. Например, по запросу **"Белый"** будет извлечена информация и об Андрее Белом, и о белом фосфоре.
- Если внутри кавычек поставить точки между терминами, то каждая из точек будет восприниматься программой как шаблон, заменяющий собой один произвольный термин.
  Например, по заданию

  "журналы по . химии"

  будут извлечены документы с фрагментами

  журналы по органической химии
  Журналы по аналитической химии

  но не будут извлечены документы с фрагментами

  журналы по химии
  журналы по общей и неорганической химии

- В описании программы утверждается, что, если задание записано без операторов либо только с операторами ***and***, то извлекаются лишь те документы, в которых все поисковые термины присутствуют **на ограниченном участке длиной в 40 слов**. Как показывает практика, документы, в которых присутствуют не все поисковые термины либо все, но на большем расстоянии друг от друга, тоже обычно (но не всегда) извлекаются.
- Следующая конструкция (*внимание!* с одинарными кавычками) используется для указания **максимально допустимого расстояния** между терминами в документе:

  '(n, термин1 термин2)'

  где *n* — некое число, указывающее, что термин1 и термин2 в тексте должны находиться на участке длиной не более, чем в *n* слов. Порядок расположения терминов 1 и 2 допускается любой.
  Пример: По заданию

  '(4, фосфат натрий)'

извлекаются документы, содержащие фрагменты

фосфат натрия
натрий-фосфат
фосфата и глутамата натрия

Не извлекается документ с фрагментом

фосфат калия и силикат натрия

(здесь поисковые термины размещены слишком далеко друг от друга — на участке длиной в 5 слов).

4.4. Усложненный поисковый бланк («Расширенный поиск»)

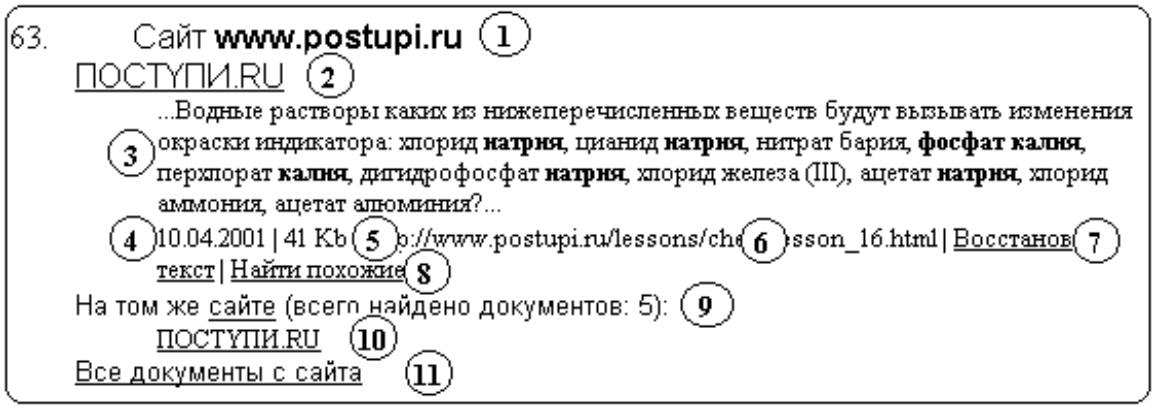
К **усложненному бланку** можно перейти по гиперсвязи «Расширенный поиск» с главной страницы портала (гиперсвязь находится справа от кнопки «**Найти!**» простейшего поискового бланка).

Особенностью *усложненного бланка* является то, что он содержит не одну графу, а несколько. С точки зрения эффективности поисковой работы, очевидных преимуществ у усложненного бланка нет. Иногда могут оказаться полезными следующие дополнительные возможности расширенного поиска: обнаружение документов, созданных в заданный период времени; извлечение страниц, находящихся на заданном сайте; поиск в избранных полях документа (в названиях веб-страниц или в заголовках внутри страниц). Поскольку информация о дате создания, о страницах сайта, о текстах заголовков по разным причинам далеко не всегда достоверно известна, то и к результатам такого вида поиска следует относиться лишь как к оценочным.

4.5. Результаты поиска

По умолчанию, результатом поиска является список обнаруженных сайтов и страниц, отсортированный по релевантности. Сайтам, зарегистрированным в каталоге *Rambler's Top100*, приписывается повышенная релевантность.

Каждый из пунктов списка содержит следующую *информацию* (см. рис. ниже): доменное имя сайта (1); название наиболее релевантной веб-страницы сайта (2); фрагмент этой страницы с выделенными жирным шрифтом поисковыми терминами (3); дату ее создания, а если эта дата неизвестна, то дату последнего индексирования роботом (4); объем страницы (5) и адрес (6); число страниц на данном сайте, удовлетворяющих поисковому заданию (9).



Фрагмент списка содержит следующие *гиперсвязи*:
(2) — к веб-странице, сведения о которой здесь приведены;
(7) Восстановить текст — вывод на экран копии данной веб-страницы, сохраненной роботом в момент индексирования (в копии поисковые термины выделяются цветом);
(8) Найти похожие — начать новый поиск, используя данную веб-страницу в качестве поискового задания;
(9) На том же сайте — к главной странице сайта, содержащего данную веб-страницу (поскольку главной странице условно приписывается адрес http://доменное_имя_сервера/ , то эта ссылка не всегда выполняется корректно);
(10) — к иным наиболее релевантным страницам данного сайта;
(11) Все документы с сайта — к списку всех известных роботу релевантных страниц данного сайта.

Кроме того, на странице результатов может быть фраза, предлагающая пользователю варианты уточнения поискового задания, например:

У нас также ищут: суперфосфат фосфаты фосфорит

В верхней части страницы находится простейший поисковый бланк с несколько измененным набором переключателей.

Установив переключатель в положение «в Интернете», пользователь может провести новый поиск по всей базе данных Рамблера.

Если установить переключатель в положение «в найденном», то новый поиск проводится только в тех документах, которые были обнаружены на предыдущей стадии работы.

5. ПОИСКОВАЯ СИСТЕМА *ЯНДЕКС*

http://www.yandex.ru/

5.1. Главная страница

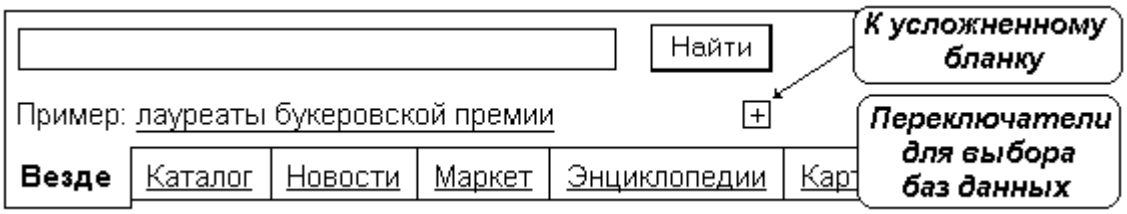
На портале **Яндекс** (*http://www.yandex.ru/*) размещается вторая по мощности российская поисковая система. Поисковая программа портала извлекает информацию из тематического каталога *Яндекс*, из базы данных, собранной роботом-пауком, а также из некоторых иных источников, например, энциклопедий.

Тематический **каталог** *Яндекс* имеет хорошо развитую структуру, удобную для ознакомления с ресурсами в режиме *Browse*. На главной странице портала размещены категории первого уровня и даже некоторые категории второго уровня; от них по гиперсвязям пользователь переходит к тематическим подразделам, например:

Наука и образование → Естественные науки → Химия

К сожалению, объем материала в подразделе «Химия» невелик (около сотни сайтов), и он лишь в незначительной степени отражает состояние химической части русскоязычного Интернета.

Здесь же на главной странице портала находится **основной поисковый бланк** для работы в режиме *Search*:



Устанавливая переключатель в одно из положений, пользователь указывает область поиска, например:

«**Везде**» — во всех базах данных портала Яндекс.

«**Каталог**» — только в тематическом каталоге.

«**Энциклопедии**» — в справочных изданиях (в том числе, в Большой Советской энциклопедии, Малом энциклопедическом словаре Брокгауза и Ефрона, Толковом словаре Даля, специализированных словарях по праву, экономике, истории, медицине, географии).

«**Картинки**» — в тех полях веб-страниц, которые содержат информацию о графических элементах (в адресе графического файла; в тексте гиперсвязи, ведущей к такому файлу; в тексте поясняющей надписи, появляющейся при подведении курсора к рисунку).

Особенностью поисковой системы Яндекс является то, что ее робот индексирует не только *HTML*-страницы, но и документы, созданные в других форматах (по состоянию на март 2003 г., в форматах *PDF* и *RTF*).

5.2. Основной поисковый бланк: формулировка задания

- Простейшая форма запроса — набор поисковых терминов, разделенных пробелами.

  Примеры запроса: **инсталлятор ChemSketch**
  тепловая труба капиллярная структура

  По запросу извлекаются документы, в которых все поисковые термины присутствуют в пределах *одного предложения* («строгое соответствие»). Кроме того, обычно извлекаются также документы, в которых поисковые термины находятся в разных предложениях, а также документы, в которых присутствуют не все поисковые термины («нестрогое соответствие»).
- Программа работает в режиме **stemming**, причем и с русскими, и с английскими терминами.

  Для отключения режима **stemming** перед термином следует поставить восклицательный знак (слитно со словом), например:

  правило !фаз
- Для проведения узконаправленного поиска используются **операторы**, обозначаемые следующими символами: **&** (логическое «*и*»), **|** (логическое «*или*»), **~** (логическое «*и не*»), а также круглые скобки.
- **Одинарные** операторы **&** и **~** указывают, что в искомом документе связанные ими оба термина (группы терминов) должны находиться/отсутствовать в пределах *одного предложения*.

  Пример запроса: фосфат & (натрия | Na) ~ калия
  или, что то же: фосфат (натрия | Na) ~ калия

  По этому заданию извлекаются документы, в которых есть предложение со словоформами *фосфат* и *натрий* (либо *Na*), но словоформа *калий* в этом же предложении должна отсутствовать.
- **Двойной** оператор **&&** означает, что связанные ими оба термина (группы терминов) должны присутствовать в пределах *всего* искомого документа.

  Пример: **(тепловая & труба) && (капиллярная & структура)**
  или, что то же: **(тепловая труба) && (капиллярная структура)**
  По этому заданию извлекаются документы, в которых есть предложение со словоформами *тепловая* и *труба* и предложение со словоформами *капиллярная* и *структура*.

- **Двойной** оператор ~~ означает, что следующий за ним термин (группа терминов) должен отсутствовать в пределах *всего* искомого документа. Вместо символов ~~ можно использовать знак - (минус), который записывают через пробел после предыдущего термина и слитно с последующим.

  Например, *по запросам:* **калий нитрат ~~ удобрение**

  или **калий нитрат -удобрение**

  извлекаются веб-страницы, в которых словоформы *калий* и *нитрат* присутствуют в одном и том же предложении, а словоформа *удобрение* отсутствует во всем документе.
- Если поисковый термин начинается с **заглавной** буквы, то извлекаются только документы, в которых это слово записано тоже с заглавной буквы.

  Если поисковый термин начинается со строчной буквы, то при извлечении материала регистр никак не учитывается.
- Фиксированная строка символов или фраза обозначается **двойными кавычками**.

  Внутри кавычек *stemming* отключен.

  Стоп-слова, находящиеся внутри кавычек, при поиске игнорируются.
- Для того, чтобы стоп-слово учитывалось при поиске, перед ним следует слитно поставить знак **+ (плюс)**.

  Например, по заданию **"фосфат и силикат"** извлекаются документы с фрагментами "**фосфат *и* силикат**", "**фосфат *или* силикат**" и т.п.

  По заданию **"фосфат +и силикат"** извлекаются документы только с фрагментами "**фосфат *и* силикат**".

5.3. Формулировка запроса — дополнительные сведения

- В запросе можно указать **допустимое расстояние** между поисковыми терминами в документе.
  - Если требуется, чтобы между терминами было *менее **n** иных слов*, используется конструкция **термин1 /n термин2** (такая запись разрешает *любой порядок* взаимного расположения этих терминов в веб-странице).

    Пример. По заданию **фосфат /1 калий** извлекаются только документы, в которых словоформы *фосфат* и *калий* находятся рядом, как в фрагментах *фосфатом калия* или *калий фосфату*.
  - Похожая конструкция, но со знаком «плюс» перед числом, **тер-**

мин1 /+n термин2, означает, что в тексте второй термин *обязательно* должен следовать *за* первым.

Пример. По заданию **система /+1 элемент** извлекается документ с фрагментом *система элементов*, но не извлекается с фрагментом *элемент системы*.

- Обратим внимание на общее и отличие в запросах

  система /+1 элементов
  "система элементов"

  Оба задания требуют, чтобы в документе поисковые термины находились рядом и именно в таком же порядке. Однако запрос **система /+1 элементов** разрешает любые словоформы терминов (например, *системами элементов*), в то время как запрос **"система элементов"** допускает только единственный набор слов, такой, какой записан внутри кавычек.
- Конструкция **термин1 && /n термин2** с удвоенным оператором **&&** означает, что допустимое расстояние между терминами измеряется не количеством слов, а количеством предложений. В частности, если поисковые термины должны находиться в искомом тексте либо в одном и том же предложении, либо в соседних, запрос принимает вид: **термин1 && /1 термин2**.
- Поисковая программа может работать и с более сложными алгоритмами ограничения расстояния, но в данном пособии они не рассматриваются.

- Запрос может содержать указание об области поиска (на конкретном сайте, в избранном поле веб-страницы). Синтаксис таких запросов мы здесь рассматривать не будем, поскольку для формулирования заданий этого типа более удобным является не простейший, а усложненный поисковый бланк (см. п. 5.4).
- На стадии формулирования запроса существует несколько механизмов регулирования длины списка извлекаемых документов и порядка расположения записей в этом списке:
  - В описании поисковой программы говорится, что в запросе желательно помечать знаком «плюс» те термины, которые обязательно должны присутствовать в документе. Как показывает практика, в большинстве случаев использование «плюса» не улучшает качество поиска, а иногда даже его ухудшает. И только в единственной ситуации «плюс» действительно необходим — если поисковым термином является стоп-слово (стоп-слово без «плюса» попросту игнорируется в ходе поиска).
  - В описании поисковой программы говорится, что в сложном

запросе пользователь может самостоятельно устанавливать значимость терминов, записывая их в формате **термин:n** , где ***n*** — произвольное число.

Пример:**медь:5 && (капиллярная структура)**

Значимость термина далее должна учитываться программой при подсчете релевантности документа и построении списка результатов. Как показывает практика, использование коэффициента ***n*** не оказывает существенного влияния на качество поиска.

5.4. Усложненный поисковый бланк («Расширенный поиск»)

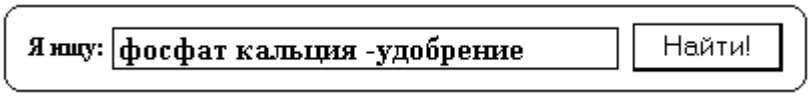
К **усложненному бланку** можно перейти от простейшего поискового бланка по гиперсвязи, расположенной под кнопкой «Найти» и обозначенной знаком «+».

Обращение к усложненному бланку целесообразно, в основном, в следующих ситуациях: поиск в избранных полях веб-страниц (в заголовке, описании, тексте ссылки, адресе); поиск на заданном сайте; извлечение документов с определенной датой создания (реально — датой индексирования роботом); поиск документов, ссылающихся на заданный; обнаружение страниц указанного формата (например, *PDF*) или содержащих в себе объекты указанного типа (например, апплеты).

Поскольку на бланке имеются поясняющие надписи и примеры заполнения редактируемых полей, здесь мы приведем только некоторые комментарии.

Усложненный бланк состоит из двух частей.

Верхняя часть (графа, помеченная словами «Я ищу») предназначена для формулирования запроса по правилам основного поискового бланка, например:



В нижней части бланка пользователь либо уточняет этот запрос, либо составляет вполне самостоятельное задание, записывая поисковые термины в соответствующих графах. Графы бланка, по умолчанию, связаны между собой операторами *and* (или операторами *not*, если возле графы написано «Исключить...»). Обе кнопки «Найти!» — верхняя и нижняя — равноценны и относятся ко *всему* запросу, который может быть размещен в одной или в обеих частях бланка.

Два редактируемых поля «Изображение» предназначены для поиска

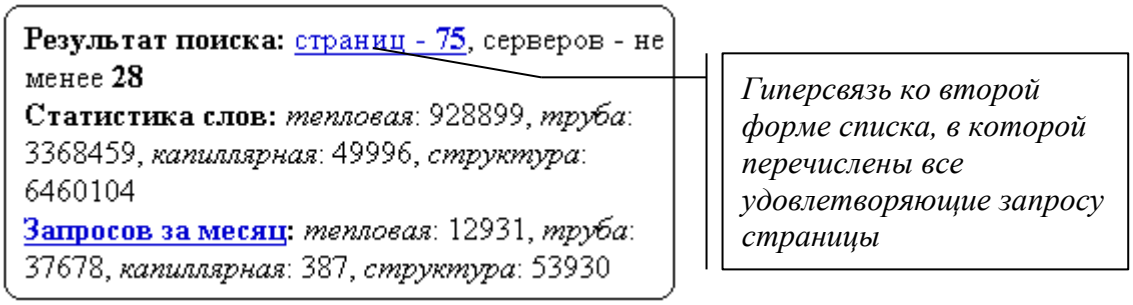
веб-страниц с рисунками и страниц с ссылками на рисунки:

- «Искать страницы, содержашие файл картинки» означает поиск в *адресах* графических файлов;
- «Искать страницы, содержащие картинку с подписью» означает поиск в текстах *поясняющих надписей*, появляющихся при подведении курсора к графическому элементу.

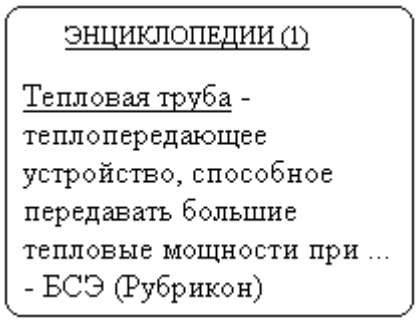
5.5. Результаты поиска

Результатом поиска является список обнаруженных сайтов и страниц, отсортированный по релевантности. Если несколько документов сайта удовлетворяет запросу, в список включается только та страница, которая характеризуется наибольшим коэффициентом релевантности.

В верхней части списка приводятся статистические данные о каждом из поисковых терминов, например:

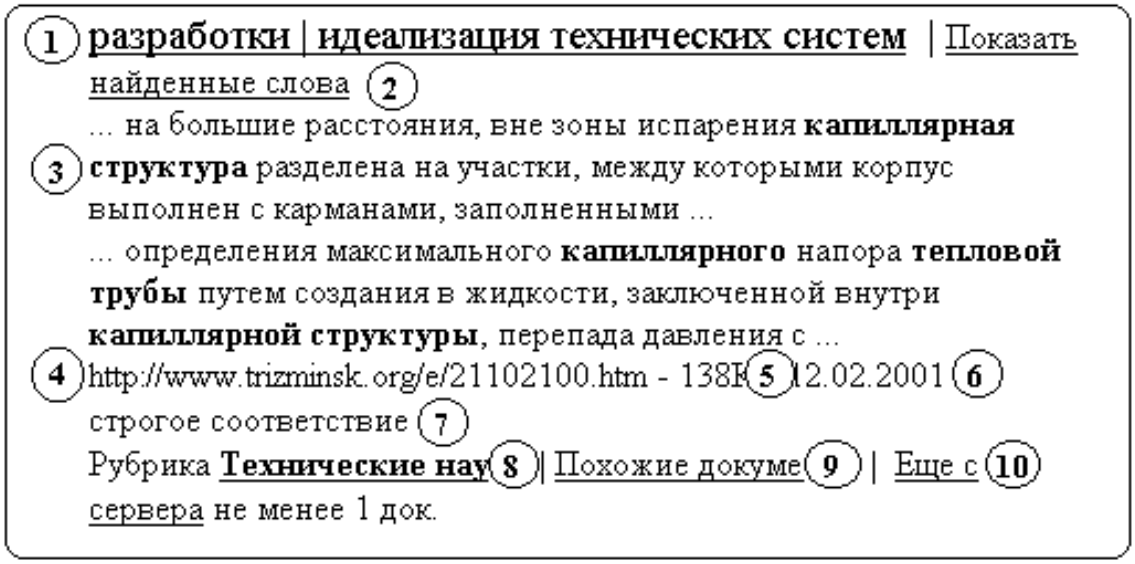


Здесь же приводятся результаты поиска в энциклопедиях и словарях с гиперссылками к соответствующим статьям, например:



Каждый из пунктов списка результатов веб-поиска содержит следующую *информацию* (см. пример ниже на рисунке): название наиболее релевантной веб-страницы сайта (1); фрагмент этой страницы с выделенными жирным шрифтом поисковыми терминами (3); адрес (4) и объем (5) страницы; дату ее создания, а если эта дата неизвестна, то дату последнего индексирования роботом (6); степень соответствия поисковому заданию — «совпадение фразы», «строгое соответствие» или «нестрогое

соответствие» (7); число страниц на данном сайте, удовлетворяющих поисковому заданию (10).



Фрагмент списка содержит следующие *гиперсвязи*:
(1) — к веб-странице, сведения о которой здесь приведены;
(2) Показать найденные слова — вывод на экран копии данной веб-страницы, сохраненной роботом в момент индексирования (в копии поисковые термины выделяются цветом);
(8) Рубрика ... — к соответствующему разделу тематического каталога Яндекс;
(9) Похожие документы — начать новый поиск, используя данную веб-страницу в качестве поискового задания;
(10) Еще с сервера — к списку всех известных роботу релевантных страниц данного сайта.

В верхней части страницы находится простейший поисковый бланк с переключателем «Искать в найденном». Если в этом переключателе установить флажок, то новый поиск проводится только в тех документах, которые были обнаружены на предыдущей стадии работы.

В нижней части страницы предлагаются варианты варьирования области поиска — по месту регистрации сайта (*например,* СНГ), по тематическому разделу каталога (*например,* в рубрике: Наука и образование).

6. ПОИСКОВАЯ СИСТЕМА *GOOGLE*

http://www.google.com/

6.1. Введение

Поисковая система **Google** (*http://www.google.com/*) в настоящее время является самой мощной в мире. По адресу разработчиков — это американская система, а по сути — интернациональная. В ее базе данных имеются сведения более чем о 3 миллиардах веб-страниц, созданных на разных языках, в том числе, на белорусском и русском. Поисковая программа *Google* способна извлекать информацию по запросам, написанным латиницей и кириллицей, а пользователь по своему желанию из сотни вариантов может выбрать наиболее удобный для него язык интерфейса программы.

Версия **Google** с белорусским интерфейсом находится по адресу

http://www.google.com/intl/be/

Версия **Google** с русским интерфейсом находится по адресу

http://www.google.com/intl/ru/

Еще один вариант руссифицированной поисковой системы находится по адресу **http://www.google.com.ru**/; база данных этого сайта не является абсолютным аналогом основной базы данных, поэтому результаты поиска в **http://www.google.com.ru**/ и в **http://www.google.com/intl/ru/** иногда могут несколько различаться.

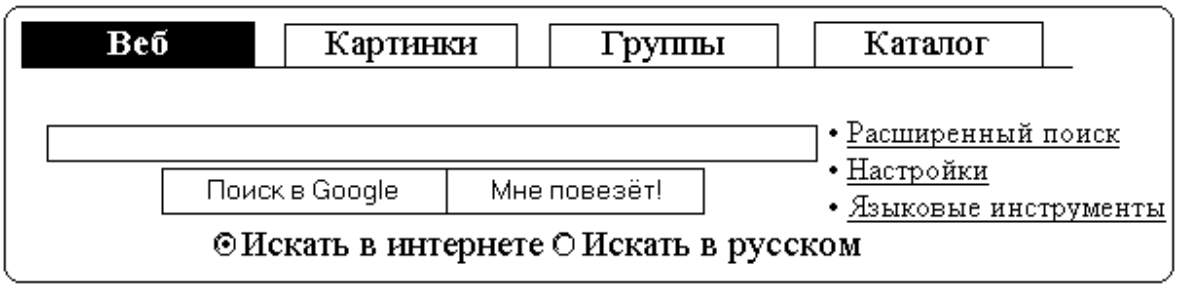
Google содержит информацию не только об *HTML*-документах, но и о файлах в других форматах (*PDF*, *PS*, *DOC*, *RTF*, *PPS*, *XLS*, в частности), а также о графических файлах.

6.2. Основной поисковый бланк

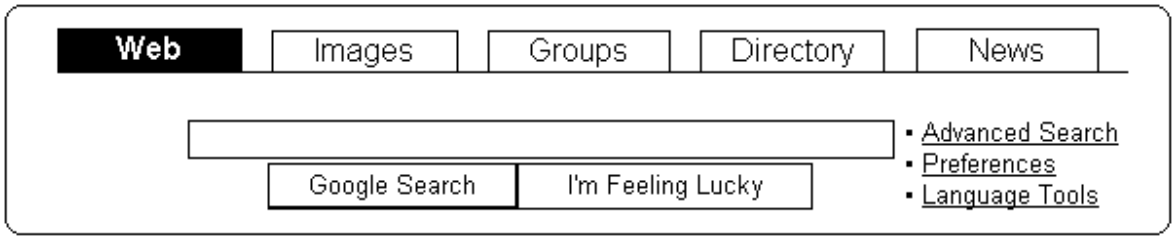
Главная страница поисковой системы *Google* построена исключительно функционально — на ней, кроме основного поискового бланка, размещено только несколько гиперсвязей. Структура бланка практически не зависит от языка интерфейса:

Белорусский (http://www.google.com/intl/be/)

Русский (http://www.google.com/intl/ru/ или **http://www.google.com.ru/)**



Английский (http://www.google.com/)



Ниже обсуждается строение русскоязычного бланка; в квадратных скобках приведены соответствующие надписи англоязычной страницы.

Кнопки, размещенные над редактируемым полем, служат для выбора информационного источника, в котором планируется вести поиск:

- **Веб [Web]** — поиск в базе данных, сформированной роботом.
- **Картинки [Images]** — поиск в тех полях веб-страниц, которые содержат информацию о графических элементах (в адресе графического файла; в тексте гиперсвязи, ведущей к такому файлу; в тексте поясняющей надписи и подписи к рисунку; в адресе страницы, содержащей графический файл).
- **Группы [Groups]** — в архиве телеконференций *Usenet Newsgroups*, содержащем более 700 млн сообщений за период с 1981 г.
- **Каталог [Directory]** — в тематическом каталоге, основой которого является каталог *Open Directory* (см. п. 17).
- **[News]** (на англоязычном бланке) — в сообщениях агентств новостей, в газетах, на страницах *News* некоторых веб-сайтов и т. п.

Справа от редактируемого поля находятся ***гиперсвязи***:

- Расширенный поиск [Advanced Search] — к усложненному поисковому бланку.
- Настройки [Preferences] — к странице, где можно изменить некоторые параметры поисковой работы (например, язык интерфейса).
- Языковые инструменты [Language Tools] — к странице, на которой находятся: (1) поисковой бланк, предназначенный для сужения области поиска по языку страницы, по стране регистрации сайта;

(2) гиперсвязи к поисковым системам *Google* в разных странах;
(3) *на англоязычном сайте* — бланки для перевода веб-страниц или произвольно взятого текста с некоторых европейских языков на английский и наоборот.

Под русскоязычным поисковым бланком есть гиперсвязь Google in English. Эта гиперсвязь может понадобиться в тех случаях, когда пользователь, желая работать с английским бланком, обращается по адресу *www.google.com*, но браузер автоматически переключается на русскоязычную страницу.

Под редактируемым полем размещены ***кнопки***:

- **Поиск в Google** [**Google Search**] — для вывода списка обнаруженных документов.
- **Мне повезёт!** [**I'm Feeling Lucky**] — для перехода к той странице, которая в списке результатов поиска стояла бы на первом месте.

На русскоязычном бланке находятся переключатели «Искать в интернете» и «Искать в русском». Рекомендуется проводить поиск с переключателем, установленным в положение «**Искать в интернете**».

Для поиска документов, написанных *латиницей*, пригоден любой бланк. Для поиска документов, написанных *кириллицей*, рекомендуется использовать бланки с русским или белорусским интерфейсами.

6.3. Основной поисковый бланк: формулировка задания

- Простейшая форма запроса — набор поисковых терминов, разделенных пробелами.
  В поисковое задание допускается включать слова на разных языках, записанные и латинским, и кириллическим шрифтами. Осложнения могут возникнуть, если в термине присутствуют специфические буквы, например, с диакритическим знаком (**ў** в белорусском слове, **ü** в немецком), поэтому следует быть готовым, что при использовании таких поисковых терминов программа не обнаружит все имеющиеся в ее базе данных подходящие документы.
- **Оператором по умолчанию** является ***and***.
  По запросу извлекаются документы, в которых *присутствуют все перечисленные* в задании поисковые термины.
- Слова в задании следует располагать в *том же порядке*, в котором они должны быть в искомом документе — это сильно увеличивает вероятность того, что нужная страница окажется в верхней части списка результатов поиска.
- Программа **не работает в режиме *stemming***.

По этой причине запросы по имени существительному в единственном числе и по тому же имени существительному, но во множественном числе, дадут разные результаты. При поиске в русскоязычной части базы данных разные результаты получатся при изменении падежа имени существительного, лица глагола и т. п.

Примеры: По запросу **journal** обнаруживается около 33 млн страниц, а по запросу **journal*s*** — около 9 млн; по запросу **фосфат калия** обнаруживается около 1,5 тыс. страниц, а по запросу **калий фосфат** — около 1 тыс.

- Программа **нечувствительна к регистру**.
  Примеры совершенно равноценных запросов: **менделеев**, **Менделеев, МЕНДЕЛЕЕВ**.
- Если запрос состоит из *одного* термина, то поиск по нему ведется в любом случае, даже если это слово размером в одну букву.
  Предлоги и другие часто встречающиеся слова считаются **стоп-словами**, если поисковое задание состоит из *нескольких* терминов.
  Пример: По запросу **in** можно узнать, что база данных содержит почти 2 млрд документов со словом *in*.
  Пример: По запросу **journals in chemistry** ведется поиск документов, содержащих слова *journals* и *chemistry*, а слово *in* при поиске игнорируется.
  Для того, чтобы стоп-слово учитывалось при поиске, перед ним следует слитно поставить знак **+ (плюс)**.
  Пример: **journals +in chemistry**
- Фиксированная строка символов или фраза обозначается **двойными кавычками**. Стоп-слова внутри кавычек *не игнорируются*.
- **Шаблон *** может использоваться для обозначения одного произвольного *слова* внутри кавычек.
  Пример: По заданию **"journal of * chemistry"** извлекаются документы с фрагментами *journal of analytical chemistry*, *journal of organic chemistry* и т. д.
- Запрос может содержать **операторы**.
  Поскольку логическое «*и*» является оператором по умолчанию, то в поисковом задании оно никакими символами не отображается.
- Логическое «*или*» обозначается оператором **OR**, записываемым именно **заглавными** буквами. Оператор **OR** особенно полезен, если требуется извлечь документы, содержащие словоформы поискового термина.
  Примеры запросов: **phosphate OR phosphates**
  фосфат OR фосфаты

Если бы последнее задание было записано в форме **фосфат or фосфаты**, то программа считала бы ***or*** стоп-словом (записано *строчными* буквами) и извлекала бы только те страницы, в которых одновременно присутствуют и слово *фосфат*, и слово *фосфаты*.

- Логическое «*и не*» обозначается знаком **-** (*минус*), который записывают через пробел после предыдущего термина и слитно с последующим.

  Пример: **phosphate -phosphates**

  Если бы задание было записано в форме **phosphate - phosphates** (минус отделен пробелами справа и слева), то велся бы поиск страниц, содержащих и слово *phosphate*, и слово *phosphates*.

  Если бы задание было записано в форме **phosphate-phosphates** (возле минуса пробелов нет), то велся бы поиск страниц, содержащих близко расположенные слова *phosphate* и *phosphates*.
- Программа использует совершенно **нестандартный порядок приоритетов** в выполнении логических операций: сначала выполняется операция ***OR***, а затем ***AND***.

  Например, по заданию **фосфат натрия OR перманганат калия** извлекаются страницы, в которых одновременно содержатся слова (1) *фосфат*, (2) *натрия* или *перманганат*, (3) *калия*.

  Использование **скобок** в запросе *никак не влияет* на порядок выполнения операций, т. е. запрос

  (фосфат натрия) OR (перманганат калия) равносилен запросу *фосфат натрия OR перманганат калия*, а по своей сути равен стандартному (но в данной ситуации бессмысленному) заданию *фосфат AND (натрия OR перманганат) AND калия***.**
- Специальные операторы используются для указания области поиска (в избранных полях, сайтах или доменах и т. д.). Поскольку почти все такие функции проще реализуются при поиске с усложненного бланка (см. п. 6.4), эти операторы здесь не рассматриваются.

6.4. Усложненный поисковый бланк (Advanced Search)

К **усложненному бланку** можно перейти с Главной страницы по гиперсвязи «Расширенный поиск» [Advanced Search], расположенной на основном поисковом бланке.

Обращение к усложненному бланку целесообразно, в основном, в следующих ситуациях: поиск в избранных полях веб-страниц; поиск на заданном сайте или в заданном домене; извлечение документов, созданных в определенный период; поиск документов, ссылающихся на задан-

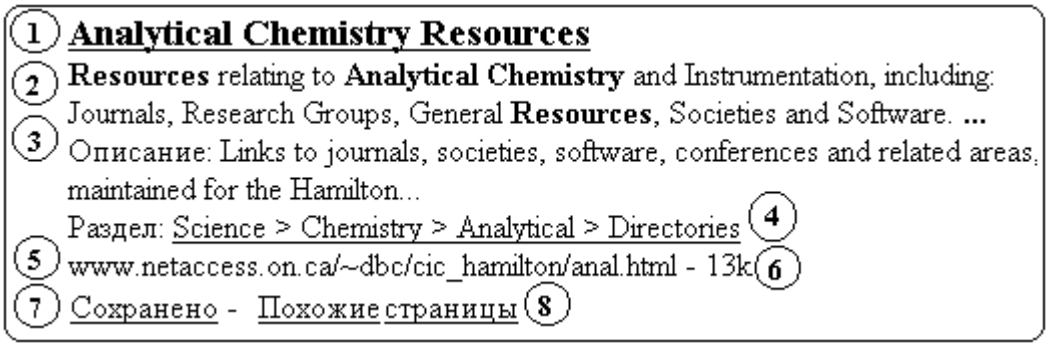
ный; извлечение страниц указанного формата (например, *PDF*).

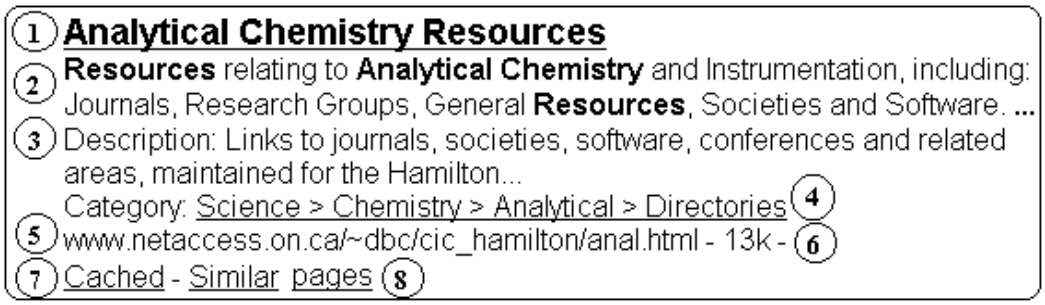
Кроме того, с этого бланка можно попробовать провести поиск веб-страниц, аналогичных заданной (качество такого поиска нередко разочаровывает).

6.5. Результаты поиска

Результатом поиска является список обнаруженных сайтов и страниц, отсортированный по релевантности.

Каждый из пунктов списка результатов содержит следующую *информацию* (см. примеры ниже на рисунке): название веб-страницы (1); фрагменты этой страницы с выделенными жирным шрифтом поисковыми терминами (2); ее адрес (5) и объем текстовой части (6). Если обнаруженная страница присутствует в тематическом каталоге *Google Web Directory*, приводится ее описание из каталога (3) и указывается, в каком тематическом разделе она находится (4).





Фрагмент списка результатов содержит следующие *гиперсвязи*:

(1) — к веб-странице, сведения о которой здесь приведены;

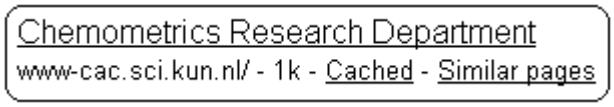
(4) — к соответствующему тематическому разделу в каталоге *Google Web Directory* (только для страниц, которые включены в этот каталог);

(7) Сохранено [Cashed] — вывод на экран копии данной веб-страницы, сохраненной роботом в момент индексирования (в копии поисковые термины выделяются цветом);

(8) Похожие страницы [Similar pages] — начать новый поиск, используя

данную веб-страницу в качестве поискового задания.

Если страница включена в базу данных *Google*, но еще не проиндексирована роботом, то в списке результатов приводится только ее название (если оно известно системе), адрес и (не всегда) размер, например:



Если несколько документов сайта удовлетворяет запросу, в список результатов включаются две страницы, которые характеризуются наибольшими коэффициентами релевантности, а для вызова остальных приводятся гиперсвязи, *например*:

[Дополнительные результаты с www.anachem.umu.se]

[More results from www.anachem.umu.se]

В таких случаях в самом конце списка помещается сообщение, *например*: «*Чтобы показать Вам наиболее значимые результаты, мы опустили некоторые, очень похожие на 34 уже показанных. Если Вы хотите, Вы можете повторить поиск, включив опущенные результаты*».

В списке отмечены документы иного, чем *HTML*, формата. В таких случаях по гиперсвязи В виде HTML [View as HTML] можно получить сохраненную роботом текстовую версию соответствующей страницы.

Пример:[PDF] Data Acquisition and Control

Формат файла: PDF/Adobe Acrobat — В виде HTML

Последняя ссылка особенно полезна тогда, когда пользователь хотел бы ознакомиться с содержанием страницы, но не имеет программмы-просмотрщика для файлов данного типа.

В нижней части страницы результатов находится гиперсвязь Поиск среди результатов [Search within results], по которой вызывается бланк для проведения поиска только в тех документах, которые были обнаружены на предыдущей стадии работы.

Другой вариант сужения области поиска — добавление терминов в исходное поисковое задание; в этом случае следует пользоваться стандартным поисковым бланком.

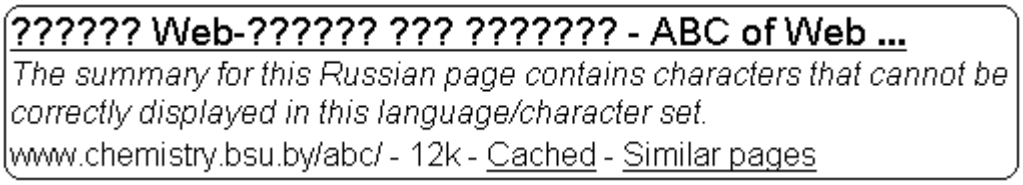
Создатели *Google* заслуженно гордятся тем алгоритмом, который они разработали и применяют для вычисления коэффициента релевантности. При расчете учитывается не только количество, плотность и расположение терминов на веб-странице, но и количество ссылок на этот документ с других сайтов. Странице приписывается ранг (*PageRank*), зависящий от количества гиперсвязей и ранга тех страниц, которые ссылаются на данную. В итоге, в верхней части списка результатов поиска оказываются такие страницы, которые удовлетворяют запросу и являются по данной

тематике самыми авторитетными.

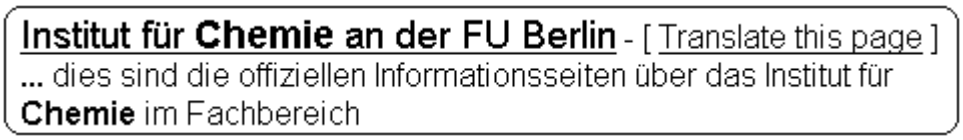
6.6. Англоязычный сайт: некоторые особенности

Проведение поиска с помощью англоязычного *Google* имеет свои минусы и плюсы.

Если в список результатов попадает русскоязычная, китайская и т. п. страница, то вместо кириллических букв или иероглифов мы видим вопросительные знаки и поясняющее сообщение о невозможности правильного отображения информации, например:



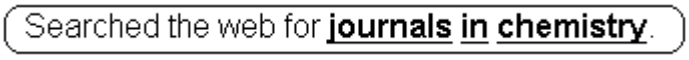
С другой стороны, если в списке присутствует немецкая, французская, итальянская, испанская, португальская страницы, можно воспользоваться автоматическим переводчиком на английский язык:



По ссылке Translate this page пользователь получает эту страницу с удовлетворительным английским текстом, графическими элементами и действующими гиперсвязями, которые ведут к переводам последующих страниц. Конечно же, текст на рисунках остается на языке оригинала.

Нелишне напомнить, что со страницы http://www.google.com/ по гиперсвязи Language Tools вызывается бланк для перевода тех веб-документов, адреса которых известны пользователю.

В верхней части страницы результатов цитируется запрос:



От подчеркнутого поискового термина гиперсвязь направлена к онлайновым словарям, в которых дается толкование данного слова.

Если в запросе допущена орфографическая ошибка, *Google* проводит поиск, но в то же время информирует пользователя и о других написаниях слов, *например*:

Searched the web for **jornals chemistry**.
Did you mean: *journals* chemistry

7. ПОИСКОВАЯ СИСТЕМА *ALLTHEWEB* http://www.alltheweb.com/

7.1. Введение

Поисковая система **AlltheWeb** (*http://www.alltheweb.com/*) по некоторым параметрам является ближайшим конкурентом системы *Google*, однако значительно уступает ей по известности среди пользователей. В базе данных *AlltheWeb* имеются сведения более чем о 2 млрд веб-страниц, созданных на разных языках, а ее поисковая программа способна извлекать информацию по запросам, написанным латиницей и кириллицей.

AlltheWeb способна анализировать не только *HTML*-документы, но и файлы в форматах *MS Word* и *Macromedia Flash*; она содержит сведения о графических, аудио-, видеофайлах и может вести поиск информации в *ftp*-архивах.

В отличие от других аналогичных систем, *AlltheWeb* не имеет тематического каталога.

7.2. Основной поисковый бланк

Основной поисковый бланк расположен на странице *http://www.alltheweb.com/*.

Кнопки, размещенные над редактируемым полем, служат для выбора бланка, предназначенного для конкретного вида поисковой работы:

- **Web** — поиск в основной базе данных, сформированной роботом;
- **News** — в сообщениях агентств новостей, в газетах и т. п.;
- **Pictures, Video, Audio** — поиск графики, видео- и аудиофайлов;
- **FTP files** — поиск в *ftp*-архивах.

Справа от редактируемого поля находятся гиперсвязи:

- **advanced search** — к *усложненному поисковому бланку*;
- **customize preferences** — к странице, где можно изменить некоторые параметры поисковой работы (в том числе таблицу кодировки символов).

Под редактируемым полем находятся переключатели «**Results in:**». Рекомендуется установить переключатель в положение «**Any Language**».

7.3. Основной поисковый бланк: формулировка задания

- Простейшая форма запроса — набор поисковых терминов, разделенных пробелами.
  В поисковое задание допускается включать слова на разных языках, записанные и латинским, и кириллическим шрифтами.
- **Оператором по умолчанию** является ***and***.
  По запросу извлекаются документы, в которых *присутствуют все перечисленные* в задании поисковые термины.
- S***temming*** не применяется к английским словам, но, как ни странно, применяется к русским словам.
- Программа **нечувствительна к регистру**.
- Фиксированная строка символов или фраза обозначается **двойными кавычками**.
- Запрос может содержать **операторы**.
  Поскольку логическое «*и*» является оператором по умолчанию, то в поисковом задании оно никакими символами не отображается.
- Логическое «***или***» обозначается совершенно нестандартно — для того, чтобы объединить термины оператором «*или*», их следует записать, *разделяя пробелами, внутри круглых скобок*.
  Пример: **sodium (phosphate phosphates)**
  По этому запросу извлекаются страницы, в которых присутствует слово *sodium*, а также либо слово *phosphate*, либо слово *phosphates*.
- Логическое «*и не*» обозначается знаком **-** (*минус*), который записывают через пробел после предыдущего термина и слитно с последующим.
  Пример: **phosphate -phosphates**
- *Булевы операторы* могут использоваться для формулирования задания, но только в одном из полей *усложненного* бланка.

7.4. Усложненный поисковый бланк (Advanced Search)

К **усложненному бланку** можно перейти с Главной страницы по гиперсвязи advanced search, расположенной на основном бланке.

Обращение к усложненному бланку целесообразно, в основном, в следующих ситуациях: поиск в избранных полях веб-страниц (в заголовке, адресе); поиск в заданном географическом регионе, в домене, в группе

серверов с заданными *IP*-адресами; извлечение документов, созданных в определенный период; поиск документов, ссылающихся на заданный; извлечение страниц определенного размера, указанного формата (например, *PDF*) и страниц, содержащих в себе заданные объекты (графические, аудио, видео, flash-анимации, апплеты, скрипты).

Усложненный бланк состоит из двух частей: верхняя, «*First select a type of search*», предназначена для указания логической связи между основными поисковыми терминами, и нижняя, «*Use the following to include and exclude additional criteria*»,— для назначения дополнительных условий поиска. Запрос может размещаться в одной либо в обеих частях.

В верхней части бланка задание формулируется в одном из двух разделов: «*Search for*» или «*Boolean*».

«**Search for**» — поисковые термины записывают в редактируемом поле, разделяя пробелами. Для установления логической связи между *всеми* записанными терминами в меню выбирают либо «*all of the words*» (равносильно операторам ***and***), либо «*any of the words*» (равносильно операторам ***or***), либо «*the exact phrase*» (равносильно кавычкам).

«**Boolean**» — в редактируемом поле записывают логическое выражение, в котором термины можно связывать друг с другом булевыми операторами ***and***, ***or***, ***andnot*** и круглыми скобками. Специфический оператор ***rank*** служит для повышения релевантности документов, в которых находится слово, отмеченное этим оператором.

Ниже, в разделе «**Word Filters**», пункты меню означают: «*Must include*» — должно содержаться обязательно; «*Should include*» — желательно, чтобы содержалось; «*Must not include*» — не должно содержаться. В каждое из редактируемых полей можно записывать только один термин. Для вызова дополнительных полей следует щелкнуть по гиперсвязи + Add a filter.

7.5. Результаты поиска

Результатом поиска является список обнаруженных сайтов и страниц, отсортированный по релевантности. Каждый из пунктов списка результатов содержит следующую *информацию*: название веб-страницы; фрагменты этой страницы с выделенными жирным шрифтом поисковыми терминами; ее адрес и объем; извлеченная из поля *Description* краткая аннотация (не для всех страниц).

Если несколько документов сайта удовлетворяет запросу, в список включается страница, характеризующаяся наибольшим коэффициентом релевантности; остальные вызываются по ссылке more hits from:

8. ПОИСКОВАЯ СИСТЕМА *MSN SEARCH* http://search.msn.com/

8.1. Введение

Поисковая система **MSN Search** извлекает результаты и из создаваемого вручную каталога *LookSmart* (см. п. 17), и из базы данных, собранной роботом фирмы *Inktomi* (*Inktomi* поставляет информацию для нескольких поисковых систем, но сама с индивидуальными пользователями не работает).

MSN Search заметно уступает системе *AlltheWeb* по объему индекса, но значительно опережает ее по популярности. Причина большей известности кроется не в каких-то уникальных и особо полезных качествах, а в том, что *MSN Search* (*MSN = Microsoft Network*) фактически является онлайновым продолжением браузера *MS Internet Explorer*.

Во-первых, адресное поле браузера *MS Internet Explorer* выполняет функции простейшего поискового бланка системы *MSN Search* (при соответствующих настройках браузера).

Во-вторых, в браузер интегрирован и более сложный бланк *MSN Search* — он открывается при нажатии кнопки «Поиск» [Search] в линейке кнопок *MS Internet Explorer*.

Основной поисковый бланк системы *MSN Search*, не привязанный к конкретному типу браузера, размещен по адресу *http://search.msn.com/*.

Поисковая программа *MSN Search* способна извлекать информацию по запросам, написанным латинским и кириллическим шрифтами.

8.2. Адресное поле браузера *MS Internet Explorer* как поисковый бланк

Адресное поле браузера ***MS Internet Explorer*** может использоваться в качестве поискового бланка системы *MSN Search*, если это разрешено настройками браузера (в *MSIE 5.0*, например, функцию поиска запрещают в *Tools → Internet Options → Advanced*, устанавливая переключатель в положение *Do not search from the Address bar*).

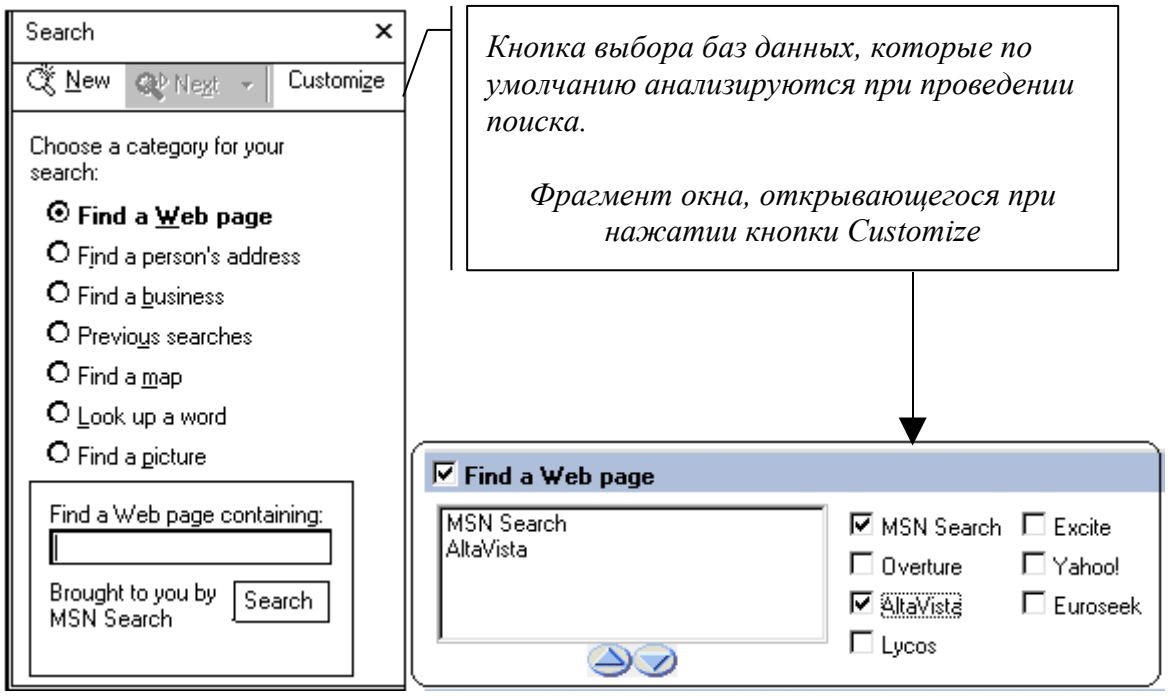
Для проведения поиска следует набрать поисковые термины (разделяя их пробелами) в адресном поле, затем нажать клавишу *Enter* на клавиатуре. Браузер отличит написанный текст от *URL*, пошлет запрос поисковой системе и выведет на экран страницу *MSN Search* с результатами поиска.

Правила формулирования запроса перечислены в п. 8.5.

8.3. Встроенный поисковый бланк браузера *MS Internet Explorer*

При нажатии кнопки «*Поиск*» [*Search*], находящейся в линейке кнопок браузера, в левой части окна открывается фрейм с поисковым бланком. Вид бланка зависит от положения переключателя «*Choose a category for your search*».

Если на переключателе выбран пункт «*Find a Web page*», бланк содержит одно редактируемое поле, предназначенное для написания запроса. Поиск проводится не только в индексе *MSN Search*, но и в индексах некоторых других поисковых систем. Пользователь сам может выбрать источники информации из списка доступных — для этого следует нажать кнопку Customize. (*Внимание!* Во время настройки бланка браузер должен быть подключен к Интернету).



Если на переключателе «*Choose a category...*» выбран пункт «*Look up a word*», поиск проводится в энциклопедии и толковых словарях. Следует быть готовым к тому, что значительная часть результатов такого поиска в полном объеме доступна только платным подписчикам соответствующих справочных изданий.

Остальные положения переключателя здесь не рассматриваются, так как они не имеют прямого отношения к поиску научной информации.

Правила формулирования запроса перечислены в п. 8.5.

8.4. Поисковый бланк на странице http://search.msn.com/

На странице *http://search.msn.com/* размещен простейший поисковый бланк системы *MSN Search*. Правила формулирования запроса изложены в п. 8.5.

Здесь же находятся категории тематического каталога, в основном идентичного каталогу LookSmart (см. п. 17). При проведении поиска содержимое этого каталога обязательно анализируется.

В верхней части страницы расположена гиперсвязь *Advanced Search*, ведущая к усложненному поисковому бланку.

8.5. Правила формулирования запроса в трех упомянутых выше простейших бланках

- **Единственная** форма запроса — набор поисковых терминов, разделенных пробелами.

  В поисковое задание допускается включать слова, записанные и латинским, и кириллическим шрифтами.

  Максимальная длина запроса — 150 символов.
- В простейшем бланке **операторы** *не используются*.
- **Оператором по умолчанию** является ***and***.
- Слова в задании следует располагать в *том же порядке*, в котором они должны быть в искомом документе — это увеличивает вероятность того, что нужная страница окажется в верхней части списка результатов поиска.
- S***temming*** применяется к английским словам и не применяется к русским словам.
- Программа **нечувствительна к регистру**.
- Фиксированная строка символов или фраза обозначается **двойными кавычками**.

  В список результатов, однако, попадают и страницы, в которых содержатся лишь фрагменты фразы, заключенной в кавычки.

8.6. Усложненный поисковый бланк (Advanced Search)

Браузер *MS Internet Explorer* не имеет встроенного *усложненного* бланка. К **усложненному бланку** можно перейти с Главной страницы системы *MSN Search* (http://search.msn.com/) по гиперсвязи Advanced Search.

В верхней части усложненного бланка находится поле для записи поисковых терминов. Логическая взаимосвязь между *всеми* терминами назначается в меню *Find*: пункт «*all of the words*» равносилен операторам ***and***, пункт «*any of the words*» — операторам ***or***, а «*the exact phrase*» — кавычкам.

Если в меню выбран пункт «*boolean phrase*», то в тексте поискового задания разрешено использование логических операторов ***OR***, ***NOT*** (записываются заглавными буквами), шаблона * для усечения правой части слова, кавычек для обозначения точной фразы. Оператор *AND* не записывается, поскольку он является оператором по умолчанию.

Среди других особенно полезных функций бланка следует упомянуть поиск только в названиях страниц (назначается пунктом «*words in title*» меню *Find*); включение/выключение режима *stemming* для английских слов; ограничение области поиска по региону, по языку страницы, по домену; выбор формата анализируемых документов (*HTML*, *PDF*, *DOC*, *XLS*, *PPS*); поиск страниц, содержащих объекты заданных типов (рисунки, аудио-, видеофайлы, анимации *Shockwave*, скрипты).

8.7. Результаты поиска

Поиск из простейшего бланка. Если поисковая программа находит достаточное количество подходящего материала в тематическом каталоге, то только такие страницы попадают в список результатов. Если поиск в каталоге малоуспешен, то основные сведения извлекаются из базы данных, созданной роботом *Inktomi*.

Поиск из усложненного бланка. Сведения извлекаются из базы данных, созданной роботом *Inktomi*.

В общем случае, список результатов может состоять из нескольких частей: *Web Directory Sites* (страницы из каталога), *Web Pages* (страницы из индекса *Inktomi*), *Featured Sites* (якобы самые лучшие сайты, но на деле главным образом те, которые включены в базу данных за плату).

Каждый из пунктов списка содержит следующую *информацию*: название веб-страницы, ее адрес, объем; аннотацию из каталога либо фрагмент самого документа.

Если поиск проводился из бланка «Поиск», встроенного в браузер *MSIE*, то список может состоять только из названий веб-страниц (конечно, с гиперсвязями к страницам-оригиналам).

9. ПРИМЕРЫ ДРУГИХ УНИВЕРСАЛЬНЫХ ПОИСКОВЫХ СИСТЕМ

Open.BY

http://poisk.open.by/

Поисковая система и тематический каталог белорусских ресурсов Интернета (базы данных содержат материалы домена *.by*, а также сайтов других доменов, если их содержание имеет отношение к Беларуси).

Поисковая программа анализирует весь текст веб-страницы. При формулировании комбинированного запроса разрешено использовать следующие операторы:

- логическое «*и*», которое может быть обозначено словом ***and*** или знаком ***&***;
- логическое «*или*», которое может быть обозначено словом ***or*** или знаком | (вертикальной чертой);
- логическое «*и не*», которое может быть обозначено словом ***not***, восклицательным знаком **!** или знаком **-** (минус), записываемым слитно перед термином.

Знаком + (*плюс*), записываемым слитно перед термином, можно отмечать слова, которые обязательно должны присутствовать в документе. Эта опция, в основном, дублирует функции оператора ***and***.

В сложном логическом выражении порядок выполнения логических операций может быть установлен с помощью круглых скобок, в том числе, вложенных.

Апорт

http://www.aport.ru/

Среди российских поисковых систем эта система является третьей по мощности. Робот *Апорта* достаточно часто обновляет базу данных, причем в нее включаются сведения не только о страницах первых уровней сайтов, но и тех, которые удалены от главной страницы.

AltaVista

http://www.altavista.com/

Эту поисковую систему иногда образно называют *Google* 90-х гг. Размер индекса *AltaVista* первым среди конкурентов перешагнул через отметку 100 млн записей; ее поисковая программа безусловно лидировала по своим возможностям и по качеству списков результатов. В этом году наблюдается стремление вернуть утраченные позиции.

10. СПЕЦИАЛИЗИРОВАННАЯ ПОИСКОВАЯ СИСТЕМА *SCIRUS*

http://www.scirus.com/

10.1. Общая характеристика

Scirus — это специализированная поисковая система, предназначенная для обнаружения научной, в том числе, химической информации. База данных *Scirus* содержит около 140 млн. записей (≈120 млн. веб-страниц и ≈20 млн. научных статей или рефератов) со сведениями не только о бесплатных ресурсах, но и о документах из платных баз данных, доступных лишь подписчикам.

Индекс поисковой системы формируется из источников, условно разделенных на два типа:

- **Web sources** — веб-страницы с сайтов университетов, конференций, промышленных компаний, научных обществ и отдельных ученых; страницы научных новостей, американские патенты из базы данных *USPTO*; препринты с сайтов *The Chemistry Preprint Server*, *E-Print ArXiv* и др.; технические отчеты *NASA*;
- **Journal sources** — полные тексты статей из журналов издательства *Elsevier Science*, размещенных на сайте *ScienceDirect*; рефераты статей из базы данных *Beilstein abstracts* (материалы, имеющие отношение к органической химии); полные тексты журналов онлайнового издательства *BioMed Central*; рефераты статей из базы данных *MEDLINE*.

Поисковая программа анализирует полный текст документов, которые могут быть в разных форматах (*HTML*, *PDF*, *PostScript* и др.).

В списке результатов поиска приводятся фрагменты обнаруженных документов. Степень доступности к полному документу может быть различной (подробности см. в п. 10.5).

Поиск с помощью *Scirus* имеет следующие **преимущества** по сравнению с поиском в индексах таких универсальных систем как Google:

- База данных *Scirus* содержит только научную информацию, поэтому полученный здесь список результатов содержит меньшую долю информационного мусора.
- *Scirus* индексирует не только те ресурсы, которые доступны роботам-паукам, но и документы из «скрытого» Интернета (статьи из реферируемых научных журналов, рефераты из библиографических баз данных, патенты).
- По утверждениям разработчиков, в индекс Scirus вносится инфор-

мация о всех научных документах, обнаруженных на сайте (большинство универсальных поисковых систем часто ограничиваются лишь двумя уровнями сайта, начиная с его главной страницы).

Недостатки поиска в *Scirus*:

- объем базы данных *Scirus* в десятки раз меньше, чем объем индекса *Google*, что, естественно, влияет на полноту обнаружения полезной информации;
- критерии деления веб-страниц на научные и не относящиеся к науке в значительной степени субъективны, что тоже влияет на полноту извлечения информации;
- *Scirus* индексирует, в основном, научные статьи издательства *Elsevier Science*.

10.2. Основной поисковый бланк: формулировка задания

Основной поисковый бланк (***Basic Search***) находится по адресу *http://www.scirus.com/*.

На бланке имеются два выключателя для назначения области поиска: *All journal sources* и *All Web sources* (перечень источников информации каждого типа приведен на предыдущей странице).

- Простейшая форма запроса — набор поисковых терминов, разделенных пробелами.
- **Оператором по умолчанию** является ***and***.
- Фиксированная строка символов или фраза обозначается **двойными кавычками.**

Все задание будет считаться фразой, если установлен флажок в выключателе *Exact phrase*, находящемся на поисковом бланке.

- Программа **нечувствительна к регистру**.
- Для формулирования задания можно использовать **операторы *or***, ***andnot***, а также круглые скобки. *Пример*:

  "aluminum phosphate" (binder or coating) andnot silicate

- Программа **не** работает в режиме ***stemming***
- Программа **не** использует **шаблоны**.
- Программа **может** проводить поиск в некоторых **полях** документов. Формат запроса:

  КодПоля:Термин, *например:* **au:shadyro**.

Веб-документы и записи в библиографических базах данных состоят из разных полей, что следует учитывать при планировании поиска (см. таблицу; имеющиеся поля отмечены знаком «*плюс*»):

Поле	Код	Web sources	Journal sources
Авторы	au:	-	+
Название документа	ti:	+	+
Название журнала	jo:	-	+
Ключевые слова	ke:	+	+
Сведения об авторе	af:	-	+
Адрес (URL)	url:	+	-

- В запросе можно объединять поиск по полю и поиск по всему документу.
  Пример: Найти статьи о *липидных мембранах* автора *Shadyro*.
  Запрос: **au:shadyro "lipid membranes"**
  (поиск проводится в области *All journal sources*).

10.3. Усложненный поисковый бланк (Advanced Search)

К **усложненному бланку** можно перейти с Главной страницы по ссылке Advanced Search, расположенной на основном поисковом бланке.

Обращение к усложненному бланку целесообразно в тех случаях, когда требуется отрегулировать область поиска по следующим параметрам: год публикации; тип информационного источника (реферат, статья, препринт и т.д.); база данных, в которой находится источник; область науки (химия и химтехнология, материаловедение и т. д.); формат документа (*HTML*, *PDF*).

Бланк имеет меню, с помощью которых можно назначать логическую взаимосвязь между поисковыми терминами и указывать, в каких полях документов следует проводить поиск.

10.4. Результаты поиска

Результатом поиска является список обнаруженных документов, отсортированный по релевантности. При расчете коэффициента релевантности учитывается степень соответствия поисковому заданию и количество ссылок на данный документ с других веб-страниц.

По желанию пользователя список может быть отсортирован по дате создания.

Scirus проводит частичное кластерирование извлекаемых документов: по типам источников (веб-страницы; статьи и рефераты), по основным опрошенным базам данных. Сведения о кластерах и гиперсвязи к кластерам приводятся на странице результатов:

Searched for:	All of the words "**aluminum phosphate**" (**binder** or **coating**) andnot **silicate**
Found:	**40 journal results** (**ScienceDirect** | **MEDLINE** | **Beilstein** | **BioMed Central** **132 Web results** (**All Preprints** | **NASA** | **US Patent Office**) **172 total**
Sort by:	**relevance** | **date**

Содержание записи в списке зависит от того, из какого источника получены сведения о документе.

Для журнальных статей в списке приводится следующая информация (см. пример ниже на рисунке): название статьи; сокращенное библиографическое описание; фрагмент документа. От названия статьи гиперсвязь направлена к самому документу в той базе данных, название которой указано в конце записи (в данном примере — это *ScienceDirect*).

1. Aluminum phosphate sealed alumina coating: characterization of microstructure
Vippola, M. / Ahmaniemi, S. / Keranen, J. / Vuoristo, P. / Lepisto, T. / Mantyla, T. / Olsson, E., *Materials Science and Engineering: A*, Jan 2002
The microstructure of aluminum phosphate sealed plasma-sprayed alumina coating was characterized by X-ray diffractometry, scanning electron microscopy, and analytical transmission electron microscopy. Microstructural characterization was...
Full text article available from SCIENCE DIRECT
similar results

Для веб-страниц в списке приводится следующая информация (см. пример ниже): название страницы; ее фрагмент; адрес; для некоторых (например, препринтов) — название соответствующей базы данных. От названия документа гиперсвязь направлена к самой веб-странице. Если на том же сайте присутствуют иные страницы, удовлетворяющие запросу, их перечень можно вызвать по ссылке more hits from.

3. Minnamari Vippola – Articles in international scientific journals with referee practice
Jun 2002
...of Aluminium **Phosphate Binder**. Journal of American Ceramic...MÃ¤ntylÃ¤, T., Olsson, E., **Aluminum phosphate** sealed alumina **coating**: characterization of microstructure...Sprayed Chromium Oxide **Coatings**: Effect of **Aluminum Phosphate** Sealing, In: Berndt, C...
more hits from
[http://www.tut.fi/units/ms/pin/personnel/vippola_p.htm...]

10.5. Доступность исходного документа

Степень доступности исходного документа зависит от его типа.

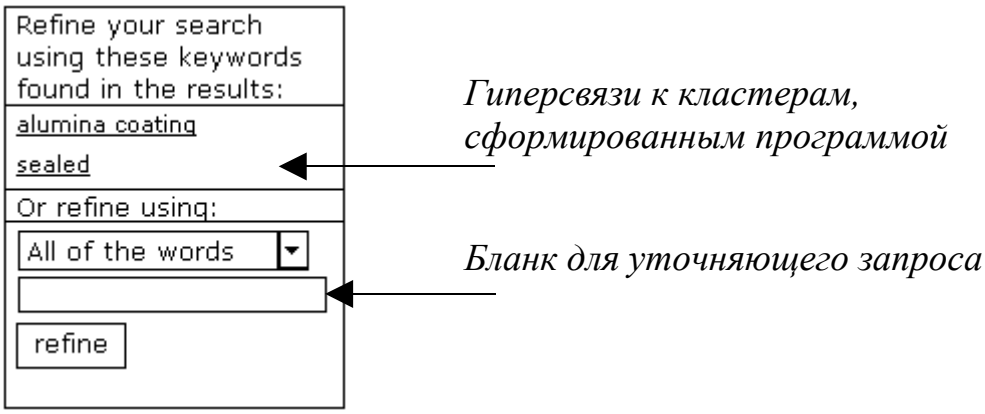
Документы группы **Web sources** (т. е. веб-страницы, патенты, препринты) полностью доступны и вызываются по гиперсвязи, начинающейся от названия документа в списке результатов поиска.

Доступность документов группы **Journal sources** частичная и зависит от того, в какой базе данных они обнаружены, в частности:

- ***ScienceDirect*** — любой пользователь может получить полное библиографическое описание, сведения об авторах, реферат статьи;
- ***Beilstein abstracts*** — полное библиографическое описание, рефераты статей доступны только зарегистрированным пользователям портала *ChemWeb* (регистрация бесплатна);
- ***MEDLINE*** — полное библиографическое описание, рефераты статей доступны только зарегистрированным пользователям портала *BioMed Net* (регистрация бесплатна).

10.6. Уточнение результатов поиска

Поисковая программа *Scirus* анализирует обнаруженные документы и находит в них некоторые общие ключевые слова, которые используются программой для еще одной стадии кластерирования результатов поиска. На странице результатов перечень кластеров приводится в бланке «Refine your search ...»:



Здесь же находится бланк «*Or refine using:*» для текста запроса. При нажатии кнопки *refine* поиск проводится только в тех документах, список которых был получен на предыдущей стадии работы.

Любая запись из списка результатов может использоваться в качестве нового поискового задания — для начала поиска такого типа требуется только щелкнуть по гиперсвязи similar results.

11. СПЕЦИАЛИЗИРОВАННАЯ ПОИСКОВАЯ СИСТЕМА *CHEMIE.DE*

http://www.chemie.de/?language=e (портал)
http://www.chemie.de/search/?language=e (поисковая система)

11.1. Общая характеристика

Chemie.DE — это двуязычный (немецкий и английский) информационный портал, на котором размещены специализированная поисковая система с тематическим каталогом, метапоисковая система для работы с библиографическими базами данных нескольких издательств, службы информирования о химических товарах, конференциях, обзоры научных новостей, а также ряд онлайновых справочников (словарь акронимов, пересчет единиц измерения физических величин и др.).

С главной страницы портала по гиперсвязи Chemistry Search Engine пользователь попадает на главную страницу поисковой системы и тематического каталога. База данных, общая для каталога и поисковой системы, содержит сведения о сотнях тысяч веб-документов в форматах *HTML*, *PDF*, *DOC*.

Работа в *Chemie.DE* возможна и в режиме *Browse*, и в режиме *Search*.

Каталог имеет развитую структуру. Особенностью его является то, что записи нижних тематических уровней одновременно принадлежат вышележащим уровням, поэтому некоторые категории оказываются излишне большими по объему и неудобны для просмотра.

Поисковые бланки размещены на всех страницах каталога и на страницах результатов поиска. Бланк, кроме редактируемого поля, содержит выключатели для указания формата искомых документов и меню для выбора языка веб-страниц (английского и/или немецкого).

При поиске анализируется весь текст документа. Поиск проводится в той категории, которая в данный момент выведена на экран. Для поиска по всей базе данных необходимо перейти на главную страницу каталога.

11.2. Поисковый бланк: формулировка задания

- Простейший запрос — набор терминов, разделенных пробелами.
- **Оператором по умолчанию** является ***and***. Для формулирования задания можно использовать **операторы *and*** (или знак + «плюс»), ***or***, ***not*** (или знак - «минус»), а также круглые скобки.
  Предостережение: Логические выражения, содержащие более трех терминов и нескольких операторов, не всегда правильно ин-

терпретируются программой.

- Программа **не** работает в режиме ***stemming***
- Программа **нечувствительна к регистру**.
- Фиксированная строка символов или фраза обозначается **двойными кавычками**.
- **Шаблон *** используется для замены произвольного количества символов в правой части слова (при этом остающаяся левая часть должна содержать не менее трех букв).
  Пример запроса: **(aluminium or aluminum) phosphate***

11.3. Результаты поиска

На странице результатов приводятся списки категорий и сайтов, обнаруженных в ходе поиска и отсортированных по релевантности. При расчете коэффициента релевантности учитывается количество ссылок на данный документ с других страниц, имеющихся в базе данных.

Каждая запись в списке результатов содержит следующую информацию: 1) название веб-страницы и гиперсвязь к ней; 2) фрагмент веб-страницы; 3) адрес; 4) объем; 5) дату создания; 6) количество гиперсвязей на обнаруженной странице (*например*, 13 links); 6) количество ссылок на данную веб-страницу с иных страниц, имеющихся в этой базе данных (*например*, 60 citations). *Пример*:

Internet Journal of Science
First issue (June 97) Abstracts **of** the 1997 Winter School in Biophysics Biophysical Aspects **of** Protein Folding. April 2 until April 6 1997 Storlien, Sweden First **Internet** Conference on Photochemistry and Photobiology November 17 - December 12 1997 Last iss
http://www.netsci-journal.com/ - 2k - last modified 2002-02-14 - 13 links - 60 citations

По гиперсвязи links можно перейти к перечню ссылок, имеющихся в документе, а по гиперсвязи citations — к перечню тех страниц, которые ссылаются на данную.

Со страницы результатов пользователь может сужать/расширять цели текущего поиска двумя способами: 1) изменяя формулировку запроса в поисковом бланке; 2) переходя в иную категорию — в поисковое задание в этом случае автоматически вносится ограничение области поиска. Категорию можно выбирать из перечня, присутствующего на странице, либо из меню *further restrictions*, имеющегося на поисковом бланке.

Прежде чем начать совершенно новый поиск, следует нажать на кнопку *stop Search*, которая появляется на поисковом бланке рядом с кнопкой *start Search* — иначе программа буде воспринимать новый запрос как уточнение прежнего задания.

12. МЕТАПОИСКОВЫЕ СИСТЕМЫ

Метапоисковые системы (*meta search engine*), в отличие от обычных поисковых систем, таких как *Google* или *AlltheWeb*, не имеют роботов-пауков и не формируют свои собственные базы данных. Вместо этого они обращаются к нескольким готовым индексам, выбирают нужную информацию и обрабатывают ее тем или иным способом. Обычно метапоисковая система посылает сформулированный пользователем запрос одновременно нескольким поисковым системам, причем нередко с «черного хода» — специально выделенным для них серверам.

В каких случаях метапоисковая система может быть *полезна*?

- Метапоисковая система позволяет существенно сэкономить время, если пользователь изначально планирует провести поиск в нескольких индексах (а базы данных, как мы знаем, имеют отличия у разных поисковых систем).
- По результатам, полученным одновременно от разных поисковых систем, пользователь может провести сравнительную оценку каждой из них с точки зрения качества решения поисковой задачи конкретного типа (при этом тестируются степень наполнения индекса материалом соответствующей тематики; реакция системы на данную конструкцию запроса и т.п.).
- Метапоисковая система может служить пробным инструментом при отработке методики поиска информации по новой или малознакомой для пользователя тематике — по полученным результатам можно судить о том, насколько удачно подобраны ключевые слова и как подобранные термины представлены в разных базах данных.

Какие *особенности* метапоисковых систем следует учитывать при работе с ними?

- Метапоисковая система не является инструментом, предназначенным для получения исчерпывающей информации о всех известных веб-ресурсах. Дело в том, что метапоисковая система, опрашивая поисковые системы, извлекает не все результаты, а только ту часть, которая характеризуется наибольшей релевантностью — чаще всего, 10-20 документов. Вполне вероятны случаи, когда прямой поиск в одном индексе обычной поисковой системы даст больше результатов, чем поиск в нескольких индексах с помощью метапоисковой системы.
- Если запрос состоит из нескольких логически связанных слов, то результаты поиска с привлечением метапоисковой системы могут

оказаться неудовлетворительными. Причина заключается в том, что разные программы пользуются разными правилами формулирования логических выражений. Правила, применяемые метапоисковой системой, могут не соответствовать правилам конкретной поисковой системы.

- Метапоисковая система не способна провести такой узконаправленный поиск, какой достигается с усложненного бланка обычной поисковой системы. Причина опять же кроется в невозможности учета в одном задании всех тонкостей синтаксиса, характерных для разных поисковых программ.
- При определении релевантности поисковая система анализирует весь документ, а метапоисковая — только ту его часть, которая попадает в список результатов поисковой системы. Для одной и той же веб-страницы коэффициенты релевантности, подсчитанные поисковой и метапоисковой системами, могут существенно различаться; соответственно, своим вмешательством метапоисковая система может ухудшить качество ранжирования, присущее исходным спискам.

В общем случае, поиск по распространенным, неспецифическим терминам может дать практически идентичные результаты и в отдельных поисковых системах, и в метапоисковых системах. Заметная разница наблюдается, если поисковое слово редкое, специфическое.

Структура списка результатов поиска у разных метапоисковых систем может быть различной.

Самый простой вариант — метапоисковая система только цитирует результаты, полученные от индивидуальных поисковых систем, не обрабатывая их. Пример такой системы— *Dogpile* (*http://www.dogpile.com/*).

Вторая группа метапоисковых систем обрабатывает полученные от поисковых систем результаты: объединяет их в один список, устраняет дубликаты (одни и те же страницы, обнаруженные несколькими системами), проводит ранжирование и только тогда передает окончательный список пользователю. Примеры: *Excite* (*http://www.excite.com/*), *MetaCrawler* (*http://www.metacrawler.com/*).

Третья группа поисковых систем проводит дальнейшую интеллектуальную обработку — группирует полученные результаты по неким общим свойствам в кластеры. Например, в части кластеров объединяются страницы, в которых есть сведения о конкретных веществах; в иных кластерах — страницы, содержащие информацию о неких общих свойствах разных веществ, или страницы, связанные с конкретным автором, с конкретным университетом и т. д. Понятно, что одна и та же веб-страница

по разным признакам может входить одновременно в несколько кластеров. Примеры систем этого типа: *Vivisimo* (см. п. 13), *Infonetware* (http://www.infonetware.com/).

13. МЕТАПОИСКОВАЯ СИСТЕМА *VIVISIMO*

http://www.vivisimo.com/

13.1. Общая характеристика

По оценкам многих экспертов, **Vivisimo** (*http://www.vivisimo.com/*) в настоящее время является самой лучшей метапоисковой системой. В ходе информационного поиска *Vivisimo* может анализировать базы данных универсальных поисковых систем и каталогов *MSN*, *Netcape Search*, *Lyco*s, *LookSmart, Gigablast, LII (Librarians' Index to the Internet)*, *BBC Search*, агентств новостей, а также таких источников информации, как энциклопедия *Britannica* и реферативная база данных *PubMed*. В явном виде в этом списке отсутствуют *Google* и *AlltheWeb*, но, поскольку их базы данных частично используются системами *Netcape Search* и *Lyco*s, то результаты поиска в *Vivisimo* в целом неплохо отражают содержание значительной части проиндексированного Интернета.

Поисковая программа *Vivisimo* способна извлекать информацию по запросам, написанным латинским и кириллическим шрифтами.

Главное достоинство *Vivisimo* — кластерирование результатов поиска.

13.2. Основной поисковый бланк

Основной поисковый бланк размещен на главной странице поисковой системы *Vivisimo* (*http://www.vivisimo.com/*) и содержит, кроме редактируемого поля, меню для выбора области поиска.

При загрузке страницы по умолчанию в меню устанавливается пункт *Search the Web*, предполагающий поиск в индексах ряда универсальных поисковых систем и каталогов. Группа пунктов *Search Top (Business, Tech, Sports) News* назначает поиск в ресурсах информационных агентств. Остальные пункты меню служат для выбора только одного из источников, но и в этом случае *Vivisimo* не просто заменяет собой поисковую программу оригинала, но и обрабатывает по своему алгоритму результаты поиска (см. п. 13.5).

С точки зрения работы с научной информацией интересны следующие упомянутые в меню ресурсы, не являющиеся поисковыми системами (от-

метим, что они доступны и непосредственно, не только через *Vivisimo*): *FirstGov* — сайт правительства США; содержит сведения о НИР, финансируемых правительством; *Britannica* — Энциклопедия Британника; *Delphion* — патенты США.

13.3. Основной поисковый бланк: формулировка задания

При формулировании задания используются правила, которые приняты в качестве стандарта большинством поисковых систем. Обращаясь к индексу за информацией, *Vivisimo* трансформирует запрос таким образом, чтобы он был понятен конкретной системе; если это не удается, то *Vivisimo* к этой системе с данным запросом не обращается.

- Простейшая форма запроса — *набор поисковых терминов, разделенных пробелами*.
  В поисковое задание допускается включать слова, записанные и латинским, и кириллическим шрифтами.
- **Оператором по умолчанию** является ***and***.
- Программа **нечувствительна к регистру**.
- Фиксированная строка символов или фраза обозначается **двойными кавычками**.
- Для проведения узконаправленного поиска используются **операторы**.
  Поскольку логическое «*и*» является оператором по умолчанию, то в поисковом задании его отображать не стоит.
- Логическое «***или***» обозначается оператором ***OR*** или ***or***.
- Логическое «***и не***» обозначается оператором ***AND NOT*** или ***NOT*** (либо знаком **-** (*минус*), записываемым слитно перед термином).
- Группа операторов используется для указания *области поиска*.
  Формат записи — **оператор:термин**, *например*,
  title:термин — поиск только в названии страницы;
  url:термин — только в тексте URL;
  domain:термин — в веб-страницах данного домена;
  linktext:термин — в тексте гиперсвязи.
  Пример 1: По запросу **linktext:atom OR url:atom** извлекаются страницы, в которых слово *atom* встречается либо в тексте гиперсвязи, либо в адресе ссылки.
  Пример 2: По запросу **"impedance spectroscopy" domain:edu** страницы, содержащие словосочетание *impedance spectroscopy*, извлекаются только из домена *.edu*.

13.4. Усложненный поисковый бланк (Advanced Search)

К **усложненному бланку** можно перейти с Главной страницы по гиперсвязи Advanced Search.

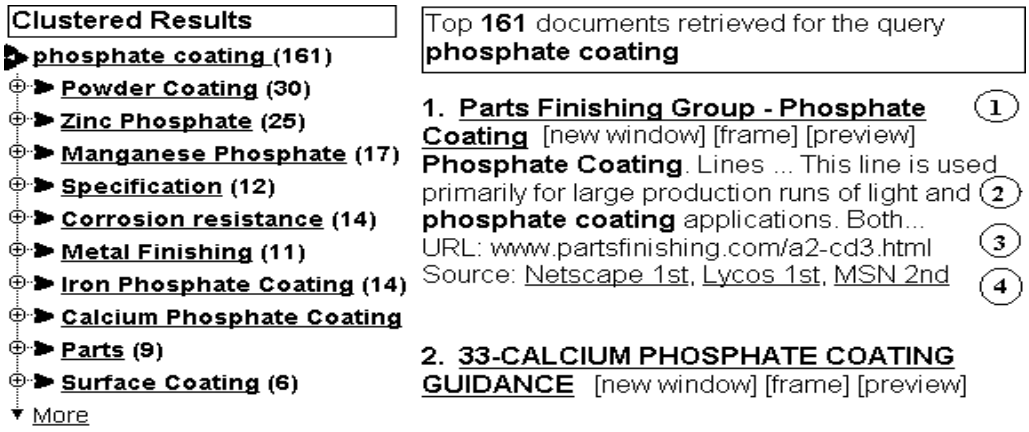
На бланке находится редактируемое поле *Search for*; задание в этой графе записывается в таком же виде, в котором оно формулируется в основном бланке.

Главное отличие усложненного бланка состоит в том, что пользователь может назначать любое сочетание источников, из которых планируется извлекать информацию — для этого достаточно расставить флажки в соответствующих выключателях. Меню «*Fast Selection*» раздела «*Search Engines*» служит для ускорения процесса настройки: при выборе любого пункта меню соответствующие флажки выставляются автоматически.

13.5. Результаты поиска

Результатом поиска является, во-первых, список обнаруженных сайтов и страниц, отсортированный по релевантности (учитываются места в списках результатов опрошенных поисковых систем); во-вторых — и это в *Vivisimo* главное — список кластеров, на которые разделены извлеченные страницы.

Пример страницы результатов (в левой части — кластеры, в правой части — список обнаруженных документов):

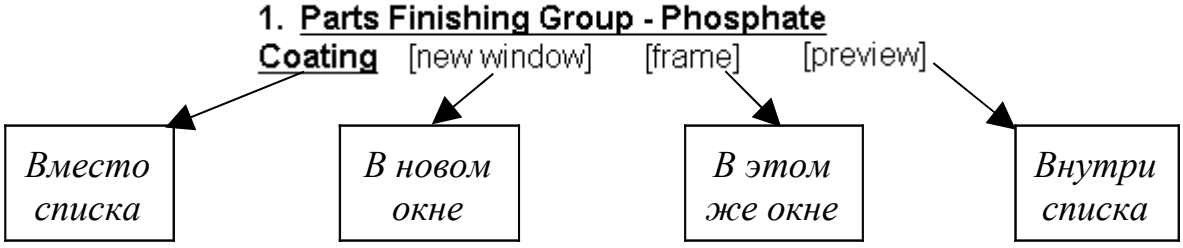


В левой части окна перечислены кластеры, по которым система может рассортировать обнаруженные документы, и число веб-страниц в каждом из кластеров. Как утверждают авторы программы, деление на кластеры проводится не просто по формальным признакам присутствия общих слов в документе, но и по результатам некоего смыслового анализа текста. Доля истины в этом утверждении имеется.

Если кластеров много, то на странице видна их первая порция; остальные можно вызывать по ссылке More .

Каждый из пунктов списка результатов содержит следующую *информацию* (см. пример на предыдущем рисунке, правая часть): название веб-страницы (1); фрагменты этой страницы с выделенными жирным шрифтом поисковыми терминами (2); ее адрес (3); порядковые номера этой страницы в списках результатов тех поисковых систем, где обнаружены сведения о ней (4).

Фрагмент списка результатов содержит следующие *гиперсвязи* (см. тот же расунок):
(1) — группа гиперсвязей к веб-странице, сведения о которой здесь приведены. Каждая из гиперсвязей по-своему выводит страницу на экран:



(4) — Гиперсвязи группы «*Source*» направлены к спискам результатов каждой из поисковых систем. При щелчке по такой гиперсвязи *Vivisimo* обращается к соответствующей поисковой системе с тем же запросом, после чего браузер в новом окне выводит результаты поиска — обычно ту часть, где упоминается именно данная страница.

После загрузки страницы результатов в памяти компьютера находятся сведения, не только выведенные на экран, но и невидимые пользователю. При необходимости эти сведения быстро обрабатываются без дополнительного обращения к серверу.

Так, например, перечень страниц, отнесенных к любому из кластеров, открывается немедленно при нажатии на гиперсвязь — название кластера. Для возвращения к полному списку документов служит самая верхняя строка в списке кластеров — та, где цитируется запрос.

В нижней части левого фрейма находится дополнительный поисковый бланк «Find in clusters» — его можно узнать по словам «Enter Keywords» в редактируемом поле. Бланк предназначен для анализа материала, полученного по последнему запросу и уже хранящегося в памяти компьютера. Если в графу внести некое слово и нажать кнопку «*Go*», программа проверяет названия, *URL* и отрывки страниц или аннотации на присутствие этого слова и выводит результат — список подходящих страниц — в правом фрейме.

14. ВЕБ-СТРАНИЦЫ ТИПА «ВСЕ В ОДНОМ» (*ALL-IN-ONE*)

Иногда к метапоисковым системам ошибочно причисляют веб-страницы, на которых размещены поисковые бланки нескольких поисковых систем. Такие бланки, как правило, простейшие, функционируют независимо друг от друга, и результаты поиска никак здесь не обрабатываются. Веб-страницы подобного типа называются «*Все в одном*» (*All-in-One*); они удобны на стадии ознакомления с разными поисковыми средствами. Пользователь, однако, должен учитывать, что настройки бланка на странице не всегда совпадают с настройками такого же бланка, но расположенного на сайте соответствующей поисковой системы (в частности, область поиска по умолчанию может быть сужена).

Пример: **All-in-One Search Page**
http://www.allonesearch.com/all1srch.html

Сайт содержит более 800 бланков для работы с универсальными и специализированными поисковыми системами, тематическими каталогогами и иными базами данных.

15. ТЕМАТИЧЕСКИЕ КАТАЛОГИ

Тематический каталог веб-ресурсов (**directory**) по своим функциям и структуре является электронным аналогом обычного каталога, используемого в библиотеке для систематизации сведений о печатной продукции. И в одном, и в другом случае весь информационный массив рассортирован по тематическим разделам и подразделам, только в веб-каталоге вместо шифров книг содержатся гиперсвязи к веб-документам. Единственное принципиальное различие заключается в том, что структура библиотечного каталога строго стандартизирована (например, в соответствии с УДК — Универсальной десятичной классификацией), а вот строение абсолютного большинства веб-каталогов зависит только от субъективной воли их создателей.

В отличие от базы данных поисковой системы, тематический каталог формируется не автоматически, а является продуктом ручного труда. Редакторы каталога посещают сайты, изучают их, составляют аннотации и распределяют материалы по соответствующим тематическим разделам и подразделам. Такая работа требует больших затрат времени, поэтому объем каталога, конечно же, меньше объема индекса поисковой системы.

Проигрывая в количестве, каталог выигрывает в качестве — ведь его материал был собран не случайным образом, а проходил стадию первичного рецензирования, в ходе которого отсеивались заведомо ложные или очевидно бессодержательные документы.

Как правило, каждая запись типичного каталога состоит из следующих элементов: 1) названия сайта или веб-страницы; 2) адреса и гиперсвязи к первоисточнику; 3) краткой аннотации. Лишь в редких каталогах вместо аннотаций можно обнаружить фрагменты веб-документов.

Тематические разделы и подразделы каталогов обычно называются **категориями** (***category***).

В большинстве каталогов поиск информации может проводиться и в режиме *Browse*, и в режиме *Search*.

В режиме *Browse* пользователь выбирает нужный ему раздел и, погружаясь по гиперсвязям иерархической лестницы категорий, достигает той части каталога, где сосредоточены списки сайтов интересующей его тематики, например:

Наука и образование → Естественные науки → Химия

Science → Chemistry → Organic Chemistry → Fullerenes → Research

В режиме *Search* пользователь формулирует в поисковом бланке текст запроса и получает список страниц, удовлетворяющих заданию. Совершенно не обязательно, что эти страницы извлекаются из одной и той же *категории*, поскольку для их обнаружения поисковая программа анализирует не тематическую направленность, а только факт наличия/отсутствия поисковых терминов в записях каталога.

Режим *Browse* целесообразен при ознакомлении с ресурсами по некоей достаточно широкой тематике, а режим *Search* более подходит для узконаправленного поиска.

Многие каталоги позволяют объединить достоинства обоих режимов работы: в них предусмотрена опция поиска не по всему каталогу, а только по его части — по категории того уровня, где в данный момент находится пользователь. В некоторых случаях это позволяет заметно уменьшить долю информационного мусора в списке результатов.

Пример. По запросу *"фосфат калия"* при поиске во всем объеме каталога большую долю списка результатов составляют ссылки на производителей и продавцов этого вещества. Сведения о сайтах такого типа практически не извлекаются, если такой же поиск проводится только в пределах категории «Естественные науки».

Большие информационные порталы нередко содержат и тематический каталог, и поисковую систему. Поисковая программа такого портала обычно позволяет проводить поиск как отдельно в тематическом катало-

ге, так и одновременно в базах данных обоих источников. С точки зрения пользователя, каталог в таких случаях выглядит лишь элементом поисковой системы. Мы же должны понимать, что каталог и индекс поисковой системы — это две независимые базы данных, которые формируются по разным алгоритмам, содержат информацию разной структуры (общее в них — только названия веб-страниц) и, как следствие, предполагают разные подходы при работе в режиме *Search*. ***Запрос для поисковой системы*** целесообразно строить из специфических терминов — тех характерных слов, которые присутствуют в тексте искомого веб-документа. ***В запрос для каталога*** следует включать термины, широкие по смыслу, т. е. слова, которые используются при написании аннотаций.

16. ТЕМАТИЧЕСКИЙ КАТАЛОГ *YAHOO!*
http://www.yahoo.com/

16.1. Общая характеристика

Портал *Yahoo!* относится к числу наиболее посещаемых мест в *WWW* и по праву многие годы является законодателем моды и неофициальным стандартом в организации структуры многоцелевых сайтов и тематических каталогов. Посетителям портала предоставляются средства поиска информации в универсальных и специализированных базах данных, а также услуги электронной почты, хостинга веб-сайтов, организации форумов и др. Весь информационный и сервисный массив портала формально разделен на две части: *Yahoo! Directory* (тематический каталог веб-ресурсов) и *Yahoo! network* (специализированные указатели ресурсов, предназначенные для широкой публики — новости, справочники, онлайновые магазины и т. д.). Поскольку научная информация сосредоточена главным образом в тематическом каталоге, в дальнейшем ограничимся рассмотрением только этого сектора портала.

Перечень категорий первого (частично и второго) уровня каталога размещен на главной странице портала (*http://www.yahoo.com/*), а также на отдельной странице (*http://dir.yahoo.com/*).

Поисковая работа в *Yahoo!* возможна в режимах *Browse* и *Search*.

Основной поисковый бланк присутствует на главной странице портала и на всех страницах каталога.

Поисковая программа *Yahoo!* извлекает информацию ***из собственного каталога и из базы данных поисковой системы*** партнера (в 2003 г. — из индекса *Google*).

16.2. Тематический каталог — работа в режиме Browse

Категория *Chemistry*, в которой, в основном, и находится вся научная химическая информация, размещена в категории первого уровня *Science* и, в свою очередь, состоит из тематических подразделов. Деление на подразделы проведено по различным параметрам (отрасли химии, методы исследования, типы информационных источников и т. д.).

Структура каталога стандартна на всех уровнях: открыв страницу категории, пользователь видит перечень соответствующих подразделов, а также список сайтов, которые по своему содержанию отнесены редакторами именно к данному тематическому уровню.

На странице используются следующие обозначения:

- символом @ (например, Geochemistry@) помечены ссылки на иные части каталога;
- словом *New* выделены поступления последней недели;
- «*солнцезащитные очки*» указывают на особенно интересные, по мнению редакторов, сайты.

Каждая запись состоит из названия сайта и краткой аннотации.

В верхней части страницы расположен бланк с переключателями:

- *the Web* — для поиска во всем каталоге и в индексе поисковой системы *Google*;
- *just this category* — для поиска только в данной категории каталога.

16.3. Основной поисковый бланк: формулировка задания

- Простейшая форма запроса — набор поисковых терминов, разделенных пробелами. **Оператором по умолчанию** является ***and***.
- При поиске **в каталоге** программа работает в режиме ***stemming***
  При поиске **в индексе** *Google* программа **не** работает в режиме ***stemming***.
- Программа **нечувствительна к регистру**.
- Фиксированная строка символов или фраза обозначается **двойными кавычками**.
- Логическое «*или*» обозначается оператором **OR**, записываемым именно **заглавными** буквами.
- Логическое «*и не*» обозначается знаком **-** (*минус*), который записывают слитно перед термином.

16.4. Усложненный поисковый бланк (Advanced Search)

К усложненному бланку можно перейти по гиперсвязи Advanced Search, находящейся на основном поисковом бланке.

Запрос формулируется в графах *include all of the words*; *include this exact phrase*; *include at least one of these words*; *exclude these words*.

По гиперсвязи *More options* на бланк выводятся дополнительные графы для конкретизации области поиска: по языку страницы и периоду ее создания; по стране, домену или сайту; по отдельным полям страницы (название, текстовая часть, адрес, а также ссылки на страницу).

В нижней части бланка в разделе *Search by URL (Web Address)* можно проводить поиск страниц, аналогичных заданной (*Find web pages similar to*), и страниц, ссылающихся на заданную (*Find web pages that link to*).

16.5. Результаты поиска

Результаты поиска поступают отсортированными по релевантности, причем сайты из каталога и из индекса Google находятся в общем списке.

Фрагмент списка результатов:

Directory Category Matches 1 - 3 of 3

- **Chemistry > Journals**
- Spectroscopy > **Journals**
- Biochemistry > **Journals**

Web Matches 1 - 20 of 943,000 | **Next 20**

1. **Internet Journal of Chemistry** - aims to promote the use of the Internet and development of network resources to enable chemists to better communicate. Features a customizable interface.
   www.ijc.com/ search within this site
   ➥More sites about: **Chemistry > Journals**

2. **Journal - Chemistry of Heterocyclic Compounds** - ...
   The English translation of the **journal** "**Chemistry** of Heterocyclic Compounds" is published in New York by "Consultants Bureau". For **...**
   www.osi.lanet.lv/hgs/hgs.html search within this site

Результаты поиска разделены на четыре основные группы:

- ***Inside Yahoo!*** — ресурсы из части *Yahoo! network*;
- ***Directory Category Matches*** — перечень категорий каталога *Yahoo! Directory*, названия которых соответствуют запросу;
- ***Sponsor Matches*** — список сайтов, помещенных в каталог за плату;
- ***Web Matches*** — информация о сайтах и страницах, извлеченная из каталога и из индекса поисковой системы партнера *Yahoo!*.

На рисунке на позиции 1 в списке результатов находится сайт, обнаруженный в тематическом каталоге. Приводится его название и гиперсвязь к первоисточнику, аннотация из каталога, адрес. В пункте *More sites about* помещена гиперсвязь к тому разделу каталога, в котором обнаружен данный сайт.

На позиции 2 в списке результатов (см. тот же рисунок) находится сайт, обнаруженный в индексе поисковой системы. Приводится название и гиперсвязь к первоисточнику, фрагмент веб-страницы, ее адрес.

Если, кроме упомянутой страницы, на соответствующем сайте есть иные удовлетворяющие условию документы, то их список можно извлечь, направляясь по гиперсвязи search within this site.

По ссылке Directory, находящейся под поисковым бланком, из списка результатов можно выбрать сайты, обнаруженные в каталоге.

17. ПРИМЕРЫ ДРУГИХ ТЕМАТИЧЕСКИХ КАТАЛОГОВ

17.1. Универсальные каталоги с разделами «Химия»

Многие *российские* предметные каталоги содержат раздел «Химия», однако объем научной химической информации, представленной в них, скуден. Тем не менее, такие источники тоже могут оказаться полезными:

Яндекс (*http://www.yandex.ru/*)
List.ru (*http://www.list.ru/*)
Mail.ru (*http://www.mail.ru/*)
KM (*http://www.km.ru/*)
AllBest (*http://www.allbest.ru/*)

На сайтах практически всех основных универсальных поисковых систем есть тематические каталоги. В м*еждународных* каталогах прорефе-рированы, в основном, англоязычные веб-ресурсы. Их объемы значительно больше, чем у российских; разделы «Химия» в таких каталогах

обычно состоят из подразделов, что еще больше повышает их ценность как источников информации.

Среди прочих каталогов своим подходом к созданию базы данных выделяется **Open Directory** (*http://dmoz.org/*). Этот многоуровневый каталог создается на общественных началах энтузиастами, которые отбирают материал и проверяют его качество. Каталог *Open Directory* получил признание у профессионалов: многие ведущие поисковые системы обращаются к нему в ходе информационного поиска. В частности, основу тематического каталога **Google Directory**, доступного с сайта **Google** (*http://www.google.com/*), составляет именно материал *Open Directory*.

Подобный подход используется при создании тематического каталога **LookSmart** (*http://www.looksmart.com/*), тоже высоко оцененного профессионалами и тоже используемого некоторыми поисковыми системами при информационном поиске. В отличие от *Open Directory*, *LookSmart* содержит в себе и информацию, включенную за плату.

17.2. Специализированные химические предметные каталоги

ChemDex (*http://www.chemdex.org/*)

Каталог химических веб-ресурсов *The Sheffield Chemdex* (или просто *Chemdex*) содержит около 10 тыс. записей, распределенных по одному-двум уровням тематических разделов. Большинство записей состоят из названий сайтов (страниц) и адресов, и только в меньшей части имеются, кроме того, очень краткие аннотации. Несмотря на такую внешнюю простоту, *Chemdex* является ценным поисковым инструментом, поскольку основу этого каталога составляют ссылки на проверенные временем и авторитетные информационные первоисточники.

На портале *Chemweb* (*http://www.chemweb.com/*) размещен второй вариант каталога *Chemdex*, отличающийся иным интерфейсом и, в некоторых разделах, чаще обновляемой информацией.

The Information Retrieval in Chemistry

http://macedonia.nrcps.ariadne-t.gr/

Каталог состоит из двух частей: *Chemistry General Index* и *Chemistry Related Fields*. Явно химическая информация находится в части *Chemistry General Index*; на первом уровне она разделена по отраслям химии, а на втором — по иным признакам (журналы, базы данных, конференции и т. д.). Часть *Chemistry Related Fields* содержит сведения по геохимии, биотехнологии, медицине, физике и другим смежным наукам.

Weblinks (*http://www.chemsoc.org/links/links.htm*)

Тематический каталог, размещенный на информационном портале *chemsoc* Королевского химического общества (*RSC*), имеет одноуровневую структуру. Объем его не очень большой, но достоинство каталога заключается в имеющихся здесь подробных аннотациях сайтов.

Chemistry Gateway

http://www.psigate.ac.uk/newsite/chemistry-gateway.html

Каталог примечателен тем, что он находится на стадии развития, поэтому новые веб-ресурсы в него попадают быстрее, чем в давно сформировавшиеся и оттого консервативные, такие как *Chemdex*. Тематическая структура каталога удобна для пользователя, прореферированные источники здесь снабжены аннотациями. *Chemistry Gateway* является частью большого каталога естественнонаучных ресурсов *PSIgate* (*http://www.psigate.ac.uk/newsite/*).

The Virtual Chemistry Center

http://www-sci.lib.uci.edu/HSG/GradChemistry.html

Каталог содержит разнообразную информацию, постоянно обновляется, однако из-за больших размеров файлов не очень удобен в работе.

Тематические каталоги могут иметь и более узкую специализацию. Например, ресурсы по аналитической химии собраны в **The Analytical Chemistry Springboard** (*http://www.anachem.umu.se/jumpstation.htm*), а ресурсы по химии для средней школы — в **Relevant (High School) Chemistry Resources on the Web** (*http://www.chemistrycoach.-com/high.htm*).

18. МЕТАСАЙТЫ

Метасайты (**metasite**) — небольшие сайты (нередко размером в одну веб-страницу), содержащие списки адресов других сайтов. Информация на метасайте может быть тематически упорядочена, поэтому не всегда возможно провести четкую границу раздела между маленьким каталогом и большим метасайтом.

В информационном поиске особую ценность представляют узкоспециализированные метасайты — они обычно создаются профессионалами в данной области и содержат сведения о действительно полезных и до-

стоверных ресурсах.

Примеры метасайтов:

Organic Chemistry Resources Worldwide
http://www.organicworldwide.net/
(органический синтез).

The Surfactants Virtual Library
http://www.surfactants.net/
(поверхностно-активные вещества).

Data and Property Calculation Sites on the Web
http://www.uic.edu:80/~mansoori/Thermodynamic.Data.and.Property_html
(термодинамика).

Electrochemistry and related subjects on the Internet
http://electrochem.cwru.edu/estir/inet.htm
(электрохимия).

Списки научных химических журналов
http://rzblx1.uni-regensburg.de/ezeit/fl.phtml
http://www.ch.cam.ac.uk/ChemJournals.html

Списки бесплатных научных химических журналов
http://www.chemistry.bsu.by/abc/current/fulltextjourn.html
http://www.chemistry.bsu.by/abc/current/trialjournal.html

Списки поисковых систем разных стран
http://www.searchenginecolossus.com/

ОГЛАВЛЕНИЕ

Учебное издание

Рагойша Александр Антонович

Поиск химической информации в Интернете

ч.I. Поисковые системы и тематические каталоги

Учебное пособие для студентов

химического факультета

Ответственный за выпуск *А.А.Рагойша*

Подписано в печать . Формат 60x84/16. Бумага тип. №1.

Усл. печ. л. . Уч.-изд.л. . Тираж 100 экз. Заказ

Белгосуниверситет. Лицензия ЛВ № 414 от 11.03.93.

220050, Минск, пр. Ф.Скорины, 4.