ИНСТИТУТ «ОТКРЫТОЕ ОБЩЕСТВО»
(ФОНД СОРОСА)
Санкт-Петербургское отделение

ИНТЕРНЕТ-ЦЕНТР

Основы поиска информации в Интернете

В.А.Капустин

Методическое пособие

Базовый курс:

Основы профессиональной работы с информационными ресурсами Интернет

1998
Санкт-Петербург

Отказ от ответственности

Несмотря на то, что были предприняты все усилия для того, чтобы данный документ был свободен от опечаток, ошибочных сведений и устаревших ссылок на ресурсы Интернета, Санкт-Петербургское отделение Института "Открытое Общество" не несет никакой ответственности за убытки, как прямые, так и косвенные, которые могут вызваны использованием этого документа.

Авторские права

Содержание

Введение

Поиск информации – задача, которую человечество решает уже многие столетия. По мере роста объема информационных ресурсов, потенциально доступных одному человеку (например, посетителю библиотеки), были выработаны все более изощренные и совершенные поисковые средства и приемы, позволяющие найти необходимый документ.

Сначала эти средства совершенствовались в каталогах и информационных отделах крупных библиотек. В 70-е годы XX века появились базы данных, доступ к которым сначала обеспечивался через модемное подключение, а затем по протоколу telnet через Интернет. Стоимость работы с такими базами данных весьма велика. Например, одна минута работы с базой данных DIALOG (**www.dialog.com**) может стоить доллар, а вывод на экран одного элемента найденной записи (из, например, 70) – 20 центов. Такая высокая стоимость поиска информации потребовала создания эффективных приемов поиска.

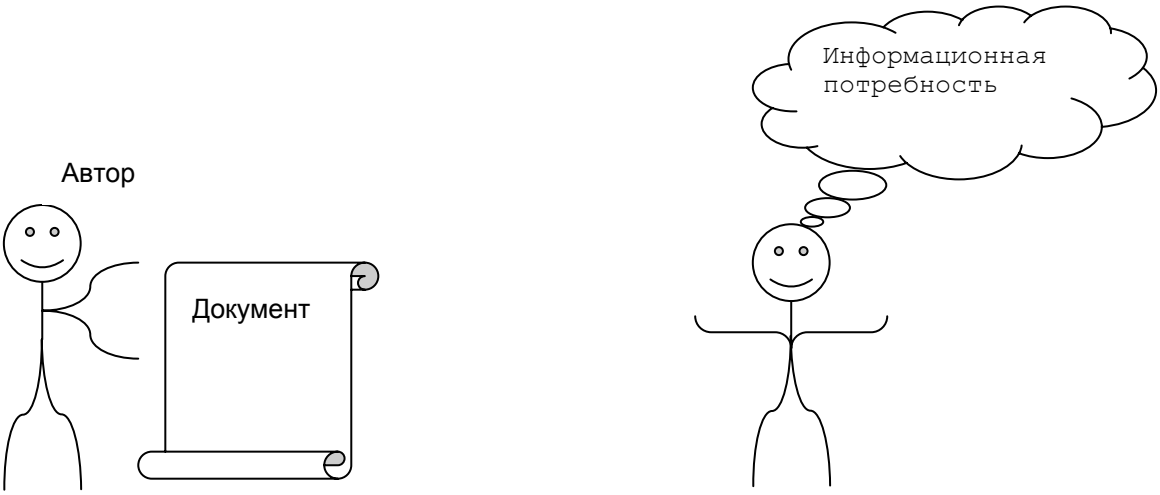
Исследования по методам поиска информации публикуются в научных журналах. В нашей стране – в журнале "Научная и техническая информация" (НТИ), в США – в Journal of American Society of Information Systems (JASIS).

Все найденные за много лет средства и приемы поиска информации доступны и эффективны и при поиске информации в Интернет.

Библиотеки используют, в основном, три вида каталогов: алфавитные, систематические и предметные. Информационно-поисковые системы (ИПС) Интернет, при всем их внешнем разнообразии, также попадают в один из этих классов. Поэтому, прежде чем знакомиться с этими ИПС, посмотрим, как устроены абстрактные алфавитные (словарные), систематические и предметные ИПС. А для этого придется познакомиться еще и с некоторыми терминами из теории информационного поиска. Наша экскурсия в теорию окажется полезной при встрече с очередной ИПС (а в Интернет их несколько сотен) – в этих ИПС вы станете узнавать знакомые черты.

Экскурсия в теорию информационно-поисковых систем

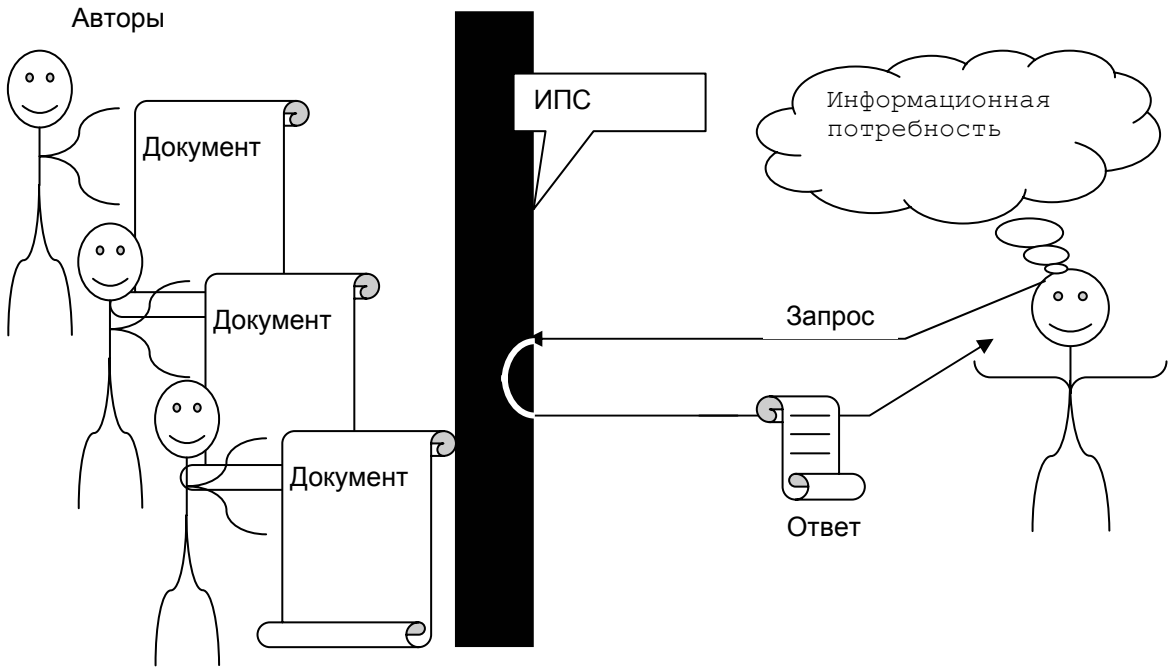
Итак, АВТОР создает ДОКУМЕНТ. А у нас (у вас) возникает ИНФОРМАЦИОННАЯ ПОТРЕБНОСТЬ:



Эта информационная потребность часто (как правило) даже не может быть точно выражена словами, и выражается только в оценке просматриваемых документов – подходит или не подходит. В теории информационного поиска вместо слова "подходит" используют термин "ПЕРТИНЕНТНЫЙ ДОКУМЕНТ", а вместо "не подходит" – "не пертинентный". Слово "пертинентный" происходит от английского "pertinent", что значит "относящийся к делу, подходящий по сути". Субъективно понимаемая цель информационного поиска – найти все пертинентные и только пертинентные документы (мы хотим найти "только то, что хотим, и ничего больше").

Эта цель – недостижима. Мы часто в состоянии оценить пертинентность документа только в сравнении с другими документами (конечно, если цель нашего поиска – редактор для Quake, а попался документ с кулинарным рецептом, то он явно непертинентен, но принять решение о пертинентности документа так просто удается далеко не всегда). Для того, чтобы было с чем сравнивать, необходимо некоторое количество непертинентных документов. Эти документы называются – "ШУМ". Слишком большой шум затрудняет выделение пертинентных документов, слишком малый – не дает уверенности в том, что найдено достаточное количество пертинентных документов (раз мы видим только пертинентные документы, нет никакой уверенности в том, что и среди тех документов, которые не попались нам на глаза, тоже не окажутся пертинентные). Практика показывает, что когда количество непертинентных документов лежит в интервале от 10% до 30%, ищущий чувствует себя комфортно, не теряясь в море шума и считая, что количество найденных документов – удовлетворительно.

Когда документов много, используется информационно-поисковая система (ИПС). В этом случае информационная потребность должна быть выражена средствами, которые "понимает" ИПС – должен быть сформулирован ЗАПРОС:



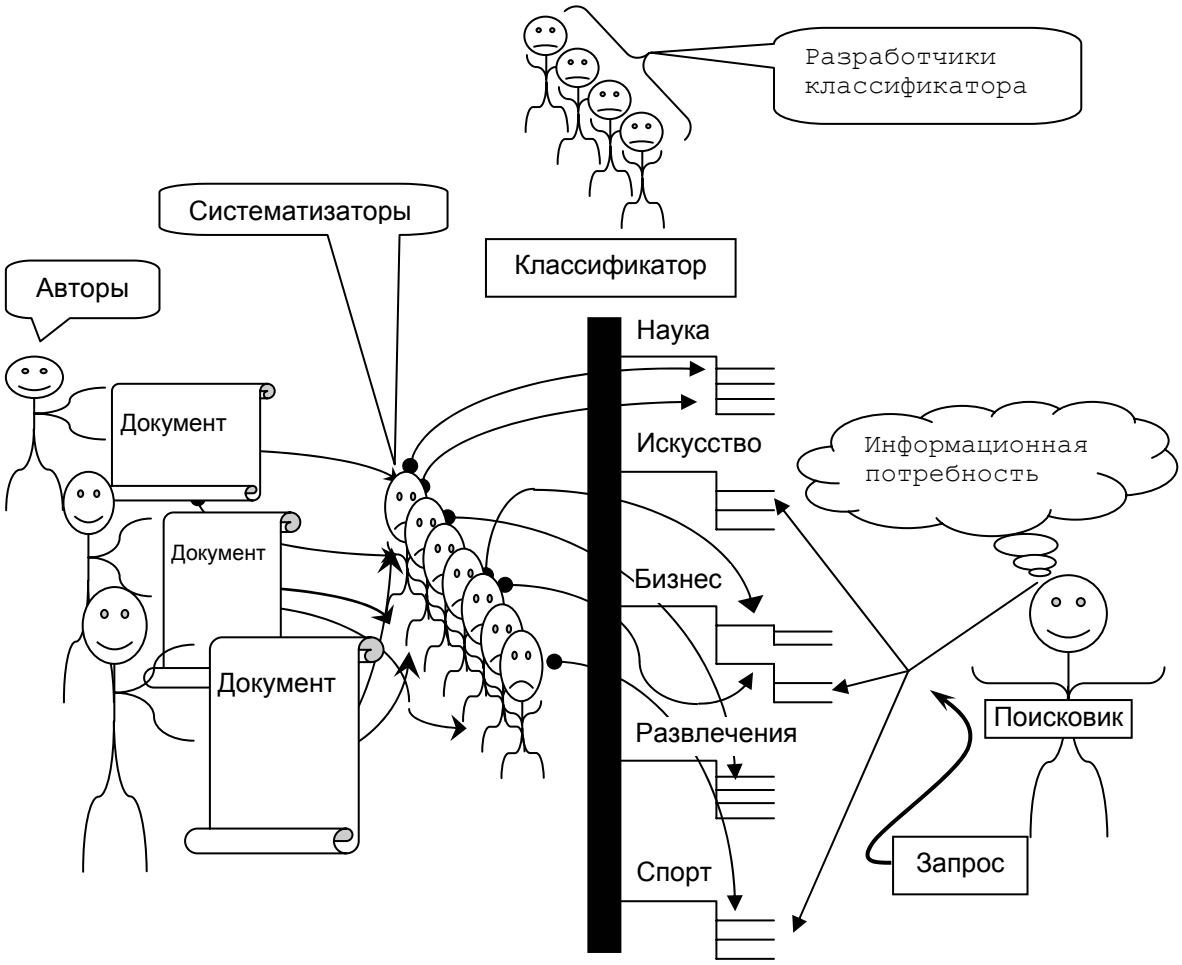
Запрос редко может точно выразить информационную потребность. Однако многие ИПС по разным причинам не могут определить, соответствует ли тот или иной документ запросу. Степень соответствия документа запросу называется РЕЛЕВАНТНОСТЬЮ. Релевантный документ может оказаться непертинентным и наоборот. Известна (американская) ИПС, которая на запрос, состоящий из единственного слова "Russia" (Россия), выдает список документов, в первом из которых этого слова нет вообще, но зато есть слово "Gagarin". Этот документ нерелевантен, но пертинентен для массовой американской аудитории. В случае, когда ищется информация о шлюпочных якорях (кошках), запрос, состоящий из слова "кошка", почти в любой ИПС даст массу релевантных, но непертинентных документов.

Классификационные информационно-поисковые системы

В классификационных ИПС используется иерархическая (древовидная) организация информации, которая называется КЛАССИФИКАТОРОМ. Разделы классификатора называются РУБРИКАМИ. Библиотечный аналог классификационной ИПС – систематический каталог. Классификатор разрабатывается и совершенствуется коллективом авторов. Затем его использует другой коллектив специалистов, называемых СИСТЕМАТИЗАТОРАМИ. Систематизаторы, зная классификатор, читают документы и приписывают им классификационные индексы, указывающие, каким разделам классификатора эти документы соответствуют.

Классический пример классификационной ИПС – **Yahoo** (**www.yahoo.com**). Едва появившись, Yahoo быстро завоевала признание качественной проработкой классификатора. Сейчас в Yahoo работают более 100 систематизаторов.

Классификационные ИПС обладают рядом специфических недостатков. Уже разработка классификатора связана с оценкой относительной важности различных областей человеческой деятельности. Например, сравнивая классификаторы многих ИПС Интернет (таких, как **Yahoo**, **Excite**, **Look Smart**), замечаем, что во многих из них нет раздела "Наука". Любая оценка является социальным действием; она связана с обществом, культурой, социальной группой, к которым принадлежит человек, выносящий оценку. Поэтому уже классификаторы, созданные разными коллективами в разных странах, могут иметь весьма различную степень полезности при поиске информации – все зависит от того, кто и что ищет. Но в создании классификационных ИПС участвуют еще и коллективы систематизаторов, также выносящих свои оценки о соответствии документов разделам классификатора.



Таким образом, при поиске информации с помощью классификационных ИПС возникает необходимость взаимодействия с другими культурами – культурами авторов, создателей классификаторов и систематизаторов.

Это непростая задача. Существует профессия, решающая эту задачу – переводчики. Хороший переводчик переводит не только слова, но и то, что называется "культурные реалии". В случае информационного поиска соответствующий профессионал называется "ИНФОРМАЦИОННЫЙ БРОКЕР". Он владеет когнитологическими методиками, знает, как устроены классификаторы и как их интерпретируют систематизаторы. Эти знания позволяют информационному брокеру в беседе с вами изучить вашу информационную потребность и превратить ее в запрос. В библиотеках такие "информационные брокеры" работают в информационных и библиографических отделах. Информационные брокеры Интернет у нас в стране уже встречаются, хотя пока еще редко.

Библиографы, понимая, что читатели не всегда глубоко изучают классификации, положенные в основу систематических каталогов, выработали два приема, облегчающие жизнь читателям. Эти приемы носят название "ОТСЫЛКА" и "ССЫЛКА", и оба они применяются создателями классификационных ИПС Интернет.

Эти приемы используются в ситуации, когда документ может быть отнесен к одному из нескольких разделов классификатора, а лицо, осуществляющее поиск (поисковик), может не знать, к какому именно разделу.

Отсылка используется тогда, когда создатели классификатора и систематизаторы в состоянии принять четкое решение об отнесении документа к одному из разделов классификатора, а поисковик с определенной вероятностью в поисках этого документа придет в другой раздел. Тогда в этом другом разделе помещается отсылка ("См.") в тот раздел классификатора, в котором действительно размещена информация о документах данного типа.

Например, информация о картах стран может быть размещена в разделах "Наука • География • Страна", "Экономика • География • Страна" или "Справочники • Карты • Страна". Принимается решение, что карты стран помещаются во второй раздел "Экономика • География • Страна"; тогда в остальные два раздела помещаются отсылки в него. Этот прием активно используется в ИПС **Yahoo** (отсылка обозначается в ней знаком @).

Ссылка ("См. также") используется в менее однозначной ситуации, когда даже создатели классификатора и систематизаторы не в состоянии принять четкого решения об отнесении документов к определенному разделу классификатора. В ИПС Интернет ссылка принимает разнообразные формы ("Relevant servers", "Похожие документы" и т.п.).

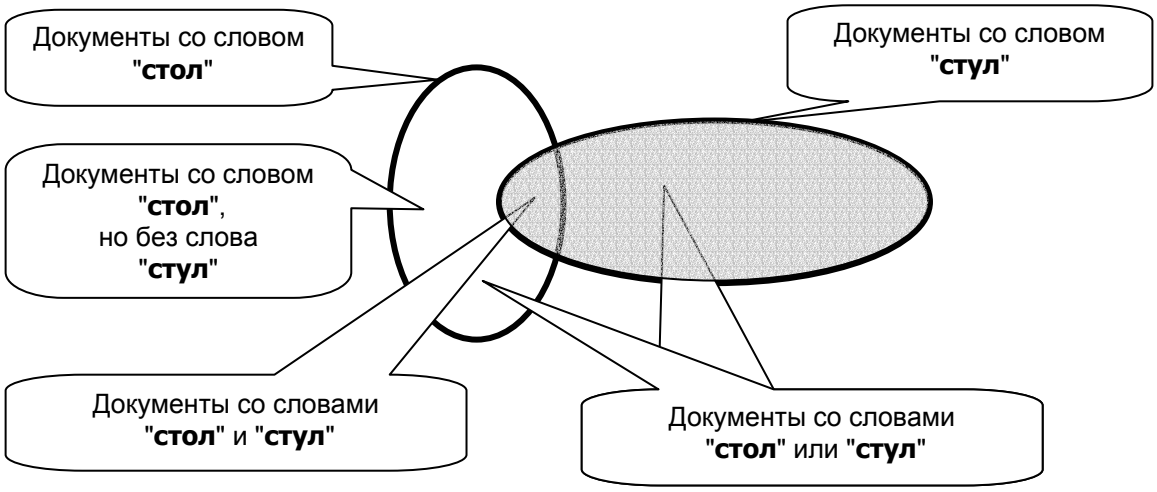
Классификационных ИПС в Интернет много (некоторые упомянуты в сводке ИПС в конце статьи). Большие классификационные ИПС (американская **Yahoo**, европейская **Yellow Web**, российские **Созвездие Интернет** и **Ау**) используют вспомогательные словарные ИПС по собственным рубрикам (аналоги библиотечных алфавитных указателей). Другие классификационные ИПС просто существуют совместно с ИПС словарного типа (**Excite**, **Lycos**, **Infoseek**).

Словарные ИПС

Культурные проблемы, связанные с использованием классификационных ИПС, привели к созданию ИПС словарного типа, с обобщенным англоязычным названием search engines. Основная идея словарной ИПС – создать словарь из слов, встречающихся в документах Интернет, в котором при каждом слове будет храниться список документов, из которых взято данное слово. Если поиск слов в таком словаре выполняется быстро, то можно отказаться от услуг разработчиков классификаторов и от услуг систематизаторов, оставаясь один на один с авторами документов.

К счастью, несмотря на обилие слов (и словоформ) в естественных языках, большинство из них употребляются нечасто, что было замечено ученым лингвистом Ципфом еще в конце 40-х годов нашего века. К тому же наиболее употребительные слова – это союзы, предлоги и артикли, т.е. слова, совершенно бесполезные при поиске информации. В результате словарь самой крупной словарной ИПС Интернет – **Alta Vista** – имеет объем всего лишь несколько Гбайт.

Поскольку слова в словаре упорядочены, поиск нужного слова может выполняться достаточно быстро – без последовательного просмотра. А наличие списков документов, в которых встречается искомое слово, позволяет ИПС выполнять операции с этими списками – их слияние, пересечение или вычитание (для наглядности списки документов изображены в виде овалов):



Вместо того, чтобы говорить "Список документов содержащих слово '**стол**' или документов, содержащих слово '**стул**'", употребляются сокращенные выражения, приведенные на рисунке. Дальнейшее сокращение эти выражения находят в языке запросов словарных ИПС: вместо "Найти список документов содержащих слово '**стол**' или документов, содержащих слово '**стул**'", большинству словарных ИПС достаточно написать что-то вроде

стол ИЛИ стул

Союз ИЛИ в запросе к словарной ИПС выступает в роли ЛОГИЧЕСКОГО ОПЕРАТОРА, связывающего множества искомых документов. Словарные ИПС используют три логических оператора: ИЛИ, И и И-НЕ ("но без"); как правило, эти операторы обозначаются одним из следующих способов:

Оператор	Полное обозначение	Сокращенное обозначение	Обозначение при простом поиске (кроме российской ИПС **Rambler**)
ИЛИ	**OR**	**|**	**пробел**
И	**AND**	**&**	**+**
И-НЕ	**AND NOT**	**&!**	**-**

Эти операторы имеют приоритет (прежде всего выполняется И-НЕ, затем – И, и лишь потом – ИЛИ), поэтому для составления сложных запросов могут использоваться скобки (исключение составляет лишь ИПС **Infoseek**, которая вместо скобок применяет другие обозначения). Как правило, словарные ИПС Интернет предоставляют пользователям два интерфейса – режим "сложного запроса" (advanced search"), в котором доступны все логические операторы, и режим простого поиска, в котором, как правило, невозможно использование скобок, и, следовательно, можно использовать не все сочетания операторов.

Давайте рассмотрим гипотетический пример поиска информации о столах. С учетом падежей слова "стол" и наших знаний о логических операторах, запрос к словарной ИПС мог бы выглядеть так:

стол ИЛИ стола ИЛИ столу ИЛИ столе ИЛИ столом

Хорошо, что это только одно слово, но писать такое уже довольно тоскливо.

Западные ИПС, ориентированные на английский язык, предлагают простое решение: вместо слова можно написать его начало, заменив изменяемую часть звездочкой:

стол*

Формально говоря, звездочка заменяет любое количество символов, поэтому говорят, что она обозначает правое усечение. Называть словом обозначение "**стол***" язык не поворачивается, поэтому для таких частей логических выражений запросов используется название ТЕРМИН. Звездочка для указанной цели (правого усечения) применяется всеми известными словарными ИПС Интернет.

Однако такой запрос отыщет и документы со словами "столовая", "столешница", "столоначальник" и даже "столб". Такое явление – искусственная синонимия – может сильно мешать при поиске, однако его проявление зачастую невозможно предусмотреть заранее.

Две российские ИПС (**Яндекс** и **Апорт**) "знают" русскую грамматику и в словаре хранят только так называемую "нормальную форму" слова (для существительного – именительный падеж единственного числа). Эти системы допускают написание запроса на естественном языке, нормализуя термины запроса, тем самым существенно упрощая поиск в русском Интернет.

Слова далекие и близкие

Описанные возможности словарных ИПС, хотя и достаточно мощные, зачастую оказываются совершенно недостаточными для поиска даже очень простой информации. Попробуем решить следующую задачу: отыскать сведения о продаже металлических стульев:

металлическ* И стул*

Но этому запросу отвечают прейскурант торговой фирмы, продающей плетеный деревянный стул (вторая строка прейскуранта) и металлический шкаф (178 строка прейскуранта). Оператор И отыскивает документы, в которых искомые слова встречаются в любом месте!

Для устранения этого недостатка некоторые ИПС хранят не просто список документов, в которых встречается слово, но и номер этого слова в конкретном документе. Это позволяет в языке запросов такой ИПС использовать оператор РЯДОМ, что решает поставленную задачу:

металлическ* РЯДОМ стул*

Многие ИПС не позволяют написать такой запрос – они не разрешают использовать термины с правым усечением совместно с оператором РЯДОМ, (только слова), но это ограничение постепенно снимается, – следите за информацией на конкретных ИПС.

Оператор РЯДОМ в различных ИПС обозначается по-разному (он имеется в **Alta Vista**, **Lycos**, **Апорт** и **Яндекс**, а также в ИПС телеконференций **DejaNews**, и во всех этих ИПС используются разные обозначения). Более того, в разных ИПС он может иметь и несколько различный смысл. Так, **Alta Vista** считает, что РЯДОМ – это не более чем через 15 слов в любом порядке, в то время как другие ИПС позволяют указывать требуемое расстояние между словами (ровно столько-то или не более чем столько-то). **Lycos** позволяет указывать расстояние и требуемый порядок слов. **Апорт** позволяет указывать расстояние между словами в словах и предложениях; **Яндекс** – в словах и абзацах (с возможностью указать порядок следования слов).

Ранжирование результатов поиска

Словарные ИПС способны выдавать списки документов, содержащие миллионы ссылок. Даже просто просмотреть такие списки совершенно невозможно. Было бы удобно иметь возможность задать формальные критерии (хотя бы относительной) важности (с точки зрения пертинентности) документов с тем, чтобы наиболее важные документы попадали бы в начало списка. Многие ИПС предоставляют такую возможность ранжирования результатов поиска. Методы ранжирования в разных ИПС различны. Так, **Alta Vista** позволяет (в режиме сложного поиска) указать перечень терминов, которые повышают ранг найденного документа (т.е. перемещают его в начало списка), что для **Alta Vista** особенно актуально, так как **Alta Vista** показывает только первые 200 найденных документов. **Rambler** и **Яндекс** позволяют указать вес каждого из терминов, участвующих в запросе, что позволяет весьма точно настраивать порядок следования найденных документов.

Английский тезаурус Alta Vista

Американский сервер ИПС **Alta Vista** (**www.altavista.digital.com**) предоставляет уникальный способ уточнения результатов поиска. Этот способ действует, только если в запросе использованы лишь англоязычные термины.

При нажатии на кнопку Refine возникает список понятий, встречающихся в только что найденных документах. С каждым понятием **Alta Vista** связывает список слов, которые видны тут же. Каждое понятие можно включить в новый запрос, исключить из него или игнорировать. Уже одно это позволяет резко повысить эффективность поиска за счет

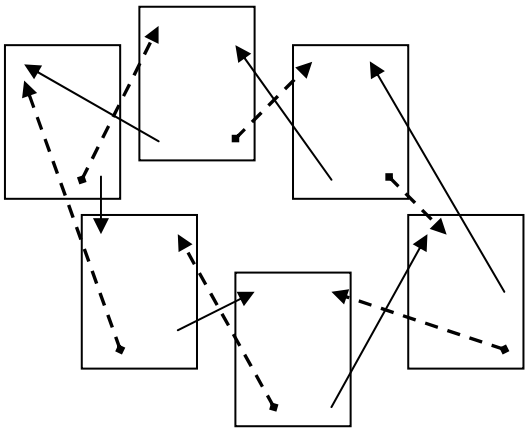
исключения понятий, не входящих в требуемую предметную область, и о сосуществовании которых с использованными вами терминами часто трудно догадаться.

Если ваш броузер поддерживает Java, то, нажав кнопку Graph, вы увидите схему связей между понятиями, и, вдобавок, сможете включать в запрос и исключать из него не только понятия целиком, но и отдельные слова, с ними связанные.

Web-кольца — предметная ИПС

Предметная ИПС с точки зрения пользователя устроена наиболее просто. Ищи название нужного предмета своего интереса (предметом может быть и нечто невещественное, например, индийская музыка), а с названием связаны списки соответствующих ресурсов Интернет. Это было бы особенно удобно, если полный перечень предметов невелик.

Так оно и было некоторое время назад. Web-мастера, занимающиеся одним предметом, начали ставить на своих серверах ссылки на серверы коллег, создавая кольцевые ссылочные структуры.



В июне 1995 г. появился сервер **www.webring.org**, объединивший несколько колец. В настоящее время на этом сервере "присутствуют" более 46 тыс. колец, которые в общей сложности включают более полумиллиона серверов, т.е. средний размер кольца – около 12 серверов. Есть, однако, и кольца-гиганты, содержащие тысячи серверов. Участники таких колец используют не только двусторонние ссылки (как показано на рисунке), но и ссылки "через сервер" и случайные ссылки, генерируемые программным образом.

Понятно, что найти нужный предмет интереса теперь непросто. **www.webring.org** обзавелся собственными вспомогательными ИПС – классификационной и словарной, помогающими найти название предмета.

Стратегия поиска

Дать общий рецепт эффективной стратегии поиска информации в Интернет, пожалуй, невозможно. Есть лишь некоторые принципы, позволяющие тратить меньше времени. Попробую их изложить.

Начну с примера. Если вам необходимо узнать, где растет древовидная черника, то вряд ли вы пойдете в алфавитный каталог библиотеки. Может быть, вы найдете нужную литературу с помощью систематического каталога. С несколько большей вероятностью – с помощью предметного. Но, скорее всего, ни один из библиотечных каталогов вам не поможет. Но зайдите в информационно-библиографический отдел крупной библиотеки, и дежурный библиограф достанет библиографический указатель по кустарничкам или какую-то похожую книжицу, из которой вы и найдете ответ на свой вопрос.

Подобную стратегию можно с успехом применять и в Интернет. В ИПС общего назначения можно утонуть в тысячах ссылок, выданных вам на простой запрос. ***Целью использования универсальной ИПС общего назначения может быть поиск специализированной ИПС, посвященной тематике вашего поиска***. Такая ИПС может быть распознана по наличию слов "информация (information)", "документ (document)" и т.п. в найденных в универсальной ИПС документах. Но часто специализированная ИПС может скрываться на сервере общественной, профессиональной или специализированной организации, издательства.

Иногда приходится разыскивать несколько информационных систем со все более узкой тематикой. Однажды ко мне обратились с просьбой срочно найти информацию о продаже судов-сухогрузов (по-английски – bulker). Запрос в **Alta Vista** (простой поиск)

+bulker* +sale*

дал нулевой результат; запрос

+ship* +sale*

тысячи ссылок на страницы, посвященные продажам катеров и яхт (впрочем, попалась и одна баржа). Внимательное изучение нескольких первых страниц списка результатов поиска показало, что в найденных текстах часто присутствует слово "marine (морской)". И тут я вспомнил, что есть в английском языке слово "maritime", означающее "все морское". Запрос

+maritime +information*

уже среди первых десяти ссылок содержал ссылку на расположенную на **www.GeoCities.com** информационную систему по морской тематике. Но и в ней информации о продаже сухогрузов не было. Зато была информация об отправке сухогрузов из портов мира, включающая сведения о владельцах судов. Многие из фирм – владельцев судов имели в своем названии слова "ship brokers (торговцы судами)". Этого английского выражения я не знал. Однако запрос в **Alta Vista**

+ship* +broker*

дал мне огромный список страниц, среди которых была одна с уже знакомым адресом – **www.GeoCities.com**. Оказывается, существует специализированная ИПС по торговцам судами! Второй найденный с помощью такой ИПС торговец содержал Web-сервер, на котором нашелся подходящий сухогруз.

Этот пример иллюстрирует еще один элемент стратегии: ***читайте найденные документы в поисках наиболее точных терминов и связей между терминами***. Возможно, вы мыслите в совершенно не тех терминах, которые используют авторы искомых документов (вспомним о культурных различиях!).

Третий элемент стратегии: ***используйте несколько ИПС***. Если вы регулярно занимаетесь поиском информации по какой-то тематике, ***отметьте те ИПС. которые для вас наиболее эффективны***.

Некоторые информационно-поисковые системы Всемирной Паутины

Название ИПС	URL	Местоположение	Вид ИПС	
			Классификационная	Словарная
Yahoo	www.yahoo.com	США	+	
Infoseek	www.infoseek.com	США	+	+
Lycos	www.lycos.com	США	+	+
Excite	www.excite.com	США	+	+
Look Smart	www.looksmart.com	США	+	
Euroseek	www.euroseek.net	Европа	+	+
Yellow Web	www.yweb.com	Европа	+	
Alta Vista	www.altavista.com	США	+ *	+
Ау	www.au.ru	Россия	+	
Созвездие Интернет	www.stars.ru	Россия	+	
Rambler	www.rambler.ru	Россия	+	+
Апорт	www.aport.ru	Россия		+
Яндекс	www.yandex.ru	Россия		+
WebRing	www.webring.org	США	Предметная	

*) Использует классификационную ИПС **Look smart**.