1.2. Поиск информации с помощью поисковых машин: понятие, принцип работы, преимущества и недостатки

Нapяду c кaтaлoгaми иcпoльзуютcя поискoвыe мaшины. Это уже более современный и удобный способ навигации и поиска в Сети. В отличие от каталогов, поисковая система – это полностью автоматизированная структура.

Задача поисковых машин – обеспечивать детальное разыскание информации в электронной вселенной, что может быть достигнуто только за счет учета (индексирования) всего содержания максимально возможного числа web-страниц. В отличие от каталогов, все они функционируют в автоматизированном режиме и имеют одинаковый принцип действия.

Поисковые машины более распространены чем каталоги, и число их, составляющее сегодня нескольких десятков, продолжает неуклонно увеличиваться. К глобальным поисковым машинам относятся:

**Google** (<http://www.google.com/>). Данная поисковая машина запущена в 1998 году. В настоящий момент эта система по всем значимым параметрам является единоличным лидером среди глобальных поисковых систем. Объем индексного файла Gооglе на сегодня составляет более 4,2 миллиapдoв wеb-cтpaниц и cтaтeй из гpупп нoвocтeй пo интeрeсaм. Дocтoинствoм Gооglе являeтся тo, чтo oн cпoсoбeн индeкcиpoвaть дoкумeнты нe тoлькo в видe НТML-фaйлoв, нo тaкжe дoкумeнты в фopмaтaх РDF, RТF, РS, DОC, ХLS, РРТ, WР5 и pядe дpугих. Gооglе пoзвoляeт дaжe пpocмoтpeть пpoиндeкcиpoвaнную cтpaницу, кoтoрaя былa удaлeнa или cepвep, на котором она расположена недоступен. Несмотря на тo, что Gооglе – это глобальная поисковая система, пользователи из неанглоязычных стран автоматически переадресовываются на интерфейс на их родном языке. На сегодня Gооglе обладает лучшими возможностями поиска иллюстраций c помощью режима "Поиск изображений" (Imаgеs, "Kapтинки").

**AlltheWeb** (<http://www.alltheweb.com/>). Была основана в Норвегии в 1997 году. Имеет базу данных более 600 млн. URL и поэтому считается одной из крупнейших в Сети. Данная поисковая машина знаменита скоростью поиска информации, но частенько появляется абсолютный спам в результатах. Поисковый робот индексирует каждое слово на странице, кроме меты тегов. Поисковая система Fast отдает предпочтение следующим факторам при ранжировании: titles, расположение ключевых слов. Одно из наиболее заметных нововведений в поисковике AllTheWeb - это так называемый "универсальный поиск", когда поисковая машина автоматически выдает информацию из разных коллекций. Так, в дополнение к каталогу web-страниц AllTheWeb имеет базу картинок, видеоклипов, MP3 и FTP-файлов из разных ресурсов Сети. При осуществлении поиска результаты выдаются из всех этих источников.

**Аltа Vistа** (<http://www.altavista.com/>). В переводе данная система означает «вид сверху». Была введена в эксплуатацию в 1995 году. На протяжении трех лет была лидером среди поисковых систем по объему индексного файла и сервисным функциям. Именно в ней был впервые опробован язык запросов: знаки "+" и "-", усечение с помощью знака "\*" и кавычки для поиска по точной фразе. В форме углубленного запроса были впервые эффективно использованы булевые операторы и оператор расстояния - NEАR.

Помимо перечисленных глобальных поисковых систем, в некоторых случаях, скорее по инерции, продолжают использоваться устаревшие поисковые сервисы, среди которых наиболее заметны HotBot (hotbot/) и Excite (excite/). Малый объем их индексных файлов на сегодня не позволяет полагаться на предоставляемые ими сведения. "Молодая" поисковая система как Ask (ask/ ) несмотря на внушительный объем проиндексированных документов, пока не представляет особого интереса. Она, к примеру, не способна осуществлять поиск документов на русском языке.

Глобальные поисковые машины уделяют свое внимание на зарубежные ресурсы Сети. Поиск информации на серверах в пределах отдельной страны является задачей локальных машин, которые специально приспособлены к особенностям конкретного языка. В русскоязычных странах к лидирующим поисковым системам на сегодня можно отнести Яндекс, Rambler и Апорт.

Яндекс (<http://www.yandex.ru/>). Запущен в сентябре 1997 года. На сегодня является лидером среди других российских поисковых систем. По запросу данная поисковая машина способна найти в русскоязычной части интернета необходимые вам веб - страницы, картинки, новости, статьи энциклопедий или товары и многое другое. При поиске учитывается морфология русского языка, машина сама исключает стоп - слова, анализирует расстояние слов друг от друга. Также позволяет искать документы на белорусском, украинском, английском, румынском, немецком и французском языках. Яндекс индексирует документы в форматах РDF,RТF,DОС,ХLS, РРТ, SWF, RSS и это отличает его от других поисковых систем. Для более точного запроса лучше всего использовать «Расширенный поиск». Также можно использовать и язык запросов Яндекс, который включает множество специальных символов: ~,&,,/,"",,(),|,$, #. Их употребление подробно описано в файле "Синтаксис языка запросов" (yandex/info/syntax.html). Кроме стандартной сортировки результатов — по релевантности (то есть по степени соответствия запросу), можно отсортировать документы по дате обновления (чтобы получить самую свежую информацию). Появилась новая функция — группировка документов по серверам. Эта возможность, дополняющая список серверов, позволяет использовать широкий набор критериев сортировки.

Rambler (<http://www.rambler.ru/>). Запущен в октябре 1996 года. Данная поисковая система содержит информацию о более чем 12 миллионах документов, которые расположены на серверах России и стран СНГ. К началу 2000 года Rambler устарел и утратил свои лидирующие позиции. Но в 2002 году была проведена модернизация всей программно-аппаратной части. Это позволило вернуть поисковой системе былой авторитет. Rambler поддерживает все кодировки русского языка. Данная поисковая машина обладает обычным и расширенным поиском. При поиске также можно пользоваться различными операторами такими как &, ||, \*, ?, or, and, NOT, кавычки, скобки (так называемый язык запроса). Сгруппировать найденную информацию можно либо по сайтам, либо по документам. Найденные документы по умолчанию сортируются по релевантности, но можно также их отсортировать и по дате. Каждая найденная ссылка снабжена функциями "Восстановить текст", "Все документы с сайта" и "Найти похожие". Все функции работают очень надежно.

Апорт (aport). Поисковая машина запущена в феврале 1996 года. Ее особенностью является то, что документы индексируются не только на серверах России и стран СНГ, но и на зарубежных серверах. Именно Апорт один из первых в России применил язык запроса, позволяющий улучшить результаты поиска. Более подробно о языке запроса, о используемых операторах можно узнать на самом сайте поисковой машины Апорт (aport/help.htm). Апорт позволяет также искать MP3-файлы, аудиофайлы. Найденные документы содержат сведения об адресе, дате опубликования и последней проверке документа [11].

Кроме перечисленных, в российском сегменте Сети существует еще несколько поисковых машин, которые не отличаются высокими показателями работы, но также могут быть применены в случаях, когда лидеры не дают результатов. К их числу относятся: КМ-поиск (go.km), Лупа (lupa/) и Tela Textorum (tela.dux).

Практически все всемирно известные каталоги и поисковые машины в настоящее время превратились во внушительные информационные корпорации с многомиллионными доходами. Заработав авторитет наиболее посещаемых мест в Сети, они предоставляют свои страницы для размещения рекламной информации, доходы от которой и составляют основу их бюджета. Постепенно поисковые сервера превращаются в многофункциональные порталы, в которых поисковый сервис остается главной приманкой для пользователей, но далеко не единственной и даже не основной из предоставляемых услуг. Помимо разыскания информации, такие сервера обычно предоставляют пользователям бесплатную электронную почту, возможность бесплатно размещать собственные страницы, сведения о погоде, текущих новостях, биржевые котировки, карты местности и т.д.

Для детального поиска документов используются специализированные поисковые системы – поисковые машины. Действие поисковых машин заключается в постоянном последовательном исследовании всех узлов Интернет, доступных данной системе поиска, со всеми их связями и ответвлениями. В связи с постоянным обновлением информации машина поиска регулярно возвращается через определенный срок (порядка месяца) к уже изученным узлам, чтобы обнаружить и зарегистрировать изменения. Вся прочитанная информация индексируется, то есть создается специализированная база данных, в которой закодированы все исследованные системой страницы Интернет [11].

При поступлении запроса от пользователя машина поиска рассматривает всю индексированную информацию и выдает список документов, соответствующих задаче поиска. Найденные документы ранжируются в зависимости от местоположения ключевых слов (в заголовке, в начале текста, в первых параграфах) и частоты их появления в тексте.

Несмотря на схожий принцип работы, машины поиска различаются по языкам запроса, зонам поиска, глубине поиска внутри документа, методам ранжирования и приоритетов, поэтому применение разных поисковых машин дает различные результаты.

В настоящее время в отечественных ресурсах действует довольно много поисковых машин. Есть фирменные и даже индивидуальные сайты, которые включают те или иные поисковые машины. Поисковые элементы существуют практически на всех Интернет-газетах и учебных порталах [10].

Наиболее распространенные из поисковых машин: Яndex ([www.yandex.ru](http://www.yandex.ru)); Апорт ([www.aport.ru](http://www.aport.ru)); МЕТА ([www.meta-ukraine.com](http://www.meta-ukraine.com)); Rambler ([www.rambler.ru](http://www.rambler.ru)); UANET ([www.uanet.com.ua](http://www.uanet.com.ua)); EIVisti ([www.el.visti.net](http://www.el.visti.net)); SEARCH ([www.search.kiev.ua](http://www.search.kiev.ua)); Pathfinder ([www.pups.kiev.ua](http://www.pups.kiev.ua)); Tela-поиск ([www.tela.dux.ru](http://www.tela.dux.ru)); Openweb ([www.openweb.ru](http://www.openweb.ru)); АУ (www.au.ru).

Следующее - это специализация поисковых машин. Так, например, top100 относится к поисковым машинам специализированного типа и больше предназначена для поиска технических элементов компьютеров и справок по ним. Поисковые машины km.ru, yandex.ru, rambler.ru, относятся к группе общего назначения. Именно с них следует начинать поиск в Интернете. Причем, поисковая система «Кирилл и Мефодий» на km.ru все же чаще используется для розыска материалов в разного рода энциклопедиях и учебных пособиях, словарях, которые вышли после 1990 г. На AltaVista – пожалуй наиболее полно представлены материалы, связанные с коммерческой деятельностью, Google – более специализируется на оперативных новостях. На машинах Fast, Инфоарт, Русский интернет и Aпорт часто можно бывает найти информацию, связанную с куплей-продажей. Следует также учитывать, что полностью адрес поисковых машин обычно включает префикс и выглядит примерно так: http://www.dig.ru. Некоторые поисковые системы используют чужие поисковые системы. Так система поиска mail.ru на сегодняшний день использует поисковую машину rambler.ru, но выводит информацию в собственном интерфейсе [11].

При наличии первичных сведений по теме поиска, документы можно разыскивать с помощью поисковых машин. При этом следует различать приемы *простого*, *расширенного*, *контекстного* и *специального поиска*.

Под *простым поиском* понимается поиск Web-ресурсов по одному или нескольким ключевым словам. Недостаток простого поиска заключается в том, что обычно он выдает слишком много документов, среди которых трудно выбрать наиболее подходящие.

При использовании *расширенного поиска* ключевые слова связывают между собой операторами логических отношений. Расширенный поиск применяют в тех случаях, когда приемы простого поиска дают слишком много результатов. С помощью логических отношений поисковое задание формируют так, чтобы более точно детализировать задание и ограничить область отбора, например, по дате публикации или по типу данных.

*Контекстный поиск -* это поиск по точной фразе. Он удобен для реферативного поиска информации, но доступен далеко не во всех поисковых системах.

*Специальный поиск* применяют при розыске Web-страниц, содержащих ссылки на заданные адреса *URL*, а также содержащих заданные данные в служебных полях, например, в поле заголовка [11].

Поисковый запрос может состоять из одного или нескольких слов, в нем могут присутствовать различные знаки препинания. Составлять самые простые запросы можно и не вдаваясь в тонкости языка запросов. Так, если ввести в поисковую строку несколько слов без знаков препинания и логических операторов, будут найдены документы, содержащие все эти слова, да еще и, если они находятся на строго ограниченном расстоянии друг от друга. Такой поиск с довольно большой вероятностью закончится отрицательным результатом. И здесь о понятии успешности не может быть и речи. Знание и правильное применение языка запросов поисковой машины поможет сделать поиск быстрым и эффективным.

Современные поисковые машины используют операторы для поиска в компьютерных сетях. Под оператором здесь понимается инструкция поиска. По этой причине запрос, включающий несколько слов, может содержать операторы. Поиск указанных операторов в документе не производится. Они служат лишь инструкцией поисковой машине. По этой причине они имеют свою орфографию и синтаксис.

Базовый алгоритм поиска с помощью поисковых машин:

- выбор поисковой машины, учитывая ее специфику;

- выявление основных слов (или словосочетаний), ключевых слов;

- если известна точная фраза из искомых материалов, используйте ее как цитату;

- после проведения предварительного поиска воспользуйтесь элементом уточнения: «Искать в избранном», задание поиска следует обязательно уточнить, введя хотя бы одно новое слово или ограничение;

- фиксируйте все интересные для вас адреса документов в Интернете (например, с помощью программы Блокнот), затем выберите среди них нужные для выполнения конкретной работы [11].

К преимуществам поисковых машин следует отнести следующие: малое количество в результатах поиска устаревших ссылок; намного большее количество Web-узлов, по которым производится поиск; более высокая скорость поиска; высокая релевантность поиска; наличие дополнительных сервисных функций, облегчающих работу пользователя, например, возможность перевода текста документа на иностранный язык, способность выделять все документы с определенного сайта, сужение критериев в ходе поиска, нахождение документов "по образцу" и т.д.

Однако у поисковых машин существуют некоторые недостатки: ограниченная область поиска. Если какой – либо сайт не был внесен в базу данных пoиcкoвoй мaшины, oн для нее не «существует», и его документы в результаты поиска попасть не могут; относительная сложность использования. Для того чтобы составленный запрос на поиск точно соответствовал тому, что именно требуется найти, нужно хотя бы немного представлять, как работает поисковая машинам, и уметь использовать простейшие логические операторы. Поисковые каталоги в этом смысле проще и привычнее; менее наглядная форма представления результатов запроса. Каталог выдает название сайта с его краткой аннотацией и другой полезной информацией. Результаты работы пoиcкoвoй мaшины менее наглядны; поскольку базу данных пoиcкoвoй мaшины пополняют программы – роботы, нечестные владельцы рекламных сайтов могут их «обмануть», из-за чего релевантность поиска может быть значительно снижена.