Содержание:

Информационные технологии поиска информации

Поиск информации: основные понятия, виды и формы организации

Поиск информации или информационный поиск представляет один из основных информационных процессов. Человечество издревле занималось им. Цели, возможности и характер поиска всегда зависели от наличия, информации, её важности и доступности, а также средств организации поиска.

Поиск – процесс, в ходе которого в той или иной последовательности производится соотнесение отыскиваемого с каждым объектом, хранящимся в массиве. Цель любого поиска заключается в потребности, необходимости или желании находить различные виды информации, способствующие получению лицом, осуществляющим поиск, нужных ему сведений, знаний и т.д. для повышения собственного профессионального, культурного и любого иного уровня; создания новой информации и формирования новых знаний; принятия управленческих решений и т.п.

Предполагается, что в дальнейшем будут созданы ИПС, способные автоматически адаптироваться с учетом уровня знаний и запросов конкретных пользователей, воспринимать запросы на естественном языке и, используя искусственный интеллект, выдавать им релевантную и пертинентную информацию. Для создания таких ИПС потребуются интеллект и знания конкретных пользователей ИПС или их посредников. Пока же от широкого круга пользователей поисковых систем требуется достаточно хорошо владеть данной предметной областью.

Существуют различные толкования термина "поиск информации" или "информационный поиск".

Термин "информационный поиск ввёл американский математик К. Муэрс. Он заметил, что побудительной причиной такого поиска является информационная потребность, выраженная в форме информационного запроса. К объектам информационного поиска К. Муэрс отнес документы, сведения об их наличии и

(или) местонахождении, фактографическую информацию.

С точки зрения использования компьютерной техники "информационный поиск" – совокупность логических и технических операций, имеющих конечной целью нахождение документов, сведений о них, фактов, данных, релевантных запросу потребителя.

"Релевантность" – устанавливаемое при информационном поиске соответствие содержания документа информационному запросу или поискового образа документа поисковому предписанию.

Существуют и другие определения. В любом случае, информационный поиск вызван потребностью удовлетворения информационных запросов пользователей, ожидающих с помощью поисковых систем оперативно получить необходимые им данные или сведения. Он является методом нацеленного поиска и извлечения релевантных документов и (или) фактов из различных источников информации, например, банков данных или запоминающих устройств. В качестве таковых выступают живые и неживые объекты, представляющие различные источники и носители информации.

Системы, обеспечивающие реализацию подобного поиска информации, называются поисковыми системами (ПС). В традиционных технологиях ПС представляют картотеки и каталоги, адресные и иные справочники, указатели, энциклопедии, справочный аппарат к изданиям и другие материалы.

В 1945 годы американский ученый и инженер В. Буш в статье "Возможный механизм нашего мышления" впервые широко поставил вопрос о необходимости механизации информационного поиска.

Начиная с 1960 годов, появляются автоматизированные поисковые системы, работающие с информацией. С этого периода ведутся интенсивные работы в области формирования и реализации принципов и методов информационного поиска.

"Поисковые системы" осуществляют поиск среди документов базы или иных массивов машиночитаемых данных, содержащих заданные слова.

Электронные ПС с помощью обычных или интеллектуальных терминалов (ПЭВМ) дают возможность пользователям производить поисковые запросы при помощи формальных и описывающих содержание элементов и с применением специальных

логических операторов; осуществляют поиск среди документов базы или иных массивов машиночитаемых данных, содержащих заданные слова. Поисковые системы позволяют осуществлять только поисковые процедуры и связанные с ними процессы.

Информационно-поисковые системы

ПС с большим набором функций и возможностей обычно входят в состав СУБД и именуются информационно-поисковыми системами. Они также создаются и используются для эффективного нахождения пользователями необходимых им данных, в том числе в Интернете.

Терминологически "информационно-поисковая система" (англ. "information retrieval system", IRS) – представляет систему, предназначенную для поиска и хранения информации; пакет программного обеспечения, реализующий процессы создания, актуализации, хранения и поиска в информационных базах и банках данных.

Информационно-поисковая система трактуется и как система, обеспечивающая поиск и отбор необходимых данных на основе информационно-поискового языка и соответствующих правил поиска, а база данных – как совокупность средств и методов описания, хранения и манипулирования данными, облегчающих сбор, накопление и обработку больших информационных массивов. Организация различных БД отличается видом объектов данных и отношений между ними.

Функционирование современных ИПС основано на двух предположениях:

1. документы, необходимые пользователю, объединены наличием некоторого признака или комбинации признаков;
2. пользователь способен указать этот признак.

Оба эти предположения на практике не выполняются, и можно говорить только о вероятности их выполнения. Поэтому, процесс поиска информации обычно представляет собой последовательность шагов, приводящих при посредстве системы к некоторому результату, и позволяющих оценить его полноту. При этом поведение пользователя, как организующее начало управления процессом поиска, мотивируется не только информационной потребностью, но и разнообразием стратегий, технологий и средств, предоставляемых системой.

Пользователь обычно не имеет исчерпывающих знаний об информационном содержании ресурса, в котором проводит поиск. Оценить адекватность выражения запроса, как и полноту получаемого результата, он может, отыскав дополнительные сведения, или так организовав процесс, чтобы часть результатов поиска могла использоваться для подтверждения или отрицания адекватности другой части. В то же время, для пользователей-профессионалов характерна устойчивость тематического профиля. Когда они являются "информационно-ориентированными", то им свойственно желание и способность организовать информационное пространство проблемы. Это означает, что пользователь создаёт по существу новый, "самостоятельный" проблемно-ориентированный, индивидуально обновляемый и пополняемый ИР, включающий помимо подборок документов также и метаинформацию, например, словари специальной терминологии, классификаторы предметных областей, описания ресурсов и т.д.

Особенность работы пользователя в режиме "самообслуживания", в контексте задачи автоматизации совокупной деятельности, означает, что система должна представлять среду, обеспечивающую поддержку функций потребителя по обработке найденной информации, а также традиционно относящихся к функциям информационного посредника (интерпретация запроса, его перевод на информационно-поисковый язык, выбор ИР, автоматизированный поиск и ручной отбор материалов), но также и такие "обеспечивающие" функции, как: структурирование информационной потребности, лексическая адаптация запроса, оценка, систематизация и обработка результатов поиска, причём на уровне как отдельного документа, так и информационных ресурсов в целом. Технические возможности, которыми располагает пользователь, позволяют ему создавать информационный ресурс – формировать массивы, систематизировать и создавать внешние представления их содержания для собственного или внешнего использования.

ИПС делятся на: традиционные (ручные, механические, электромеханические) и автоматизированные (электронные).

Автоматизированные ИПС (АИПС), используют компьютерные программно-технические средства и технологии и предназначаются для нахождения и выдачи пользователям информации по заданным критериям. Определяющими для понимания методов автоматизации поиска являются два следующих фактора:

1. сравниваются не сами объекты, а описания – так называемые "поисковые образы";

2. сам процесс является сложным (составным и не одноактным) и обычно реализуется последовательностью операций.

Данные в АИПС вводятся на основе специально разрабатываемых форматов ввода. Все сведения об одном объекте в ИПС представляются в виде систематизированных данных, образующих одну строку таблицы и называются записью. При этом, если ИПС представляет электронный каталог библиотеки, то любое библиографическое описание (БО) документа в нём – это одна запись, состоящая из полей, равных количеству элементов БО. Совокупность записей образует БД, которая, как правило, хранится в одном файле. Совокупность БД, объединенных одной СУБД, образует банк данных.

Поскольку АИПС инструмент, используемый человеком при поиске (а не интеллектуальным автомат для поиска информации – готовых решений задач основной деятельности), эффективность её использования зависит от того, насколько хорошо человек знает природу операционных объектов и свойства инструмента, посредством которого он работает с этими объектами.

Информационный поиск подразумевает использование определённых стратегий, методов, механизмов и средств. Поведение пользователя, осуществляющего управление процессом поиска, определяется не только информационной потребностью, но и инструментальным разнообразием системы – технологиями и средствами, предоставляемыми системой.

Стратегия поиска – общий план (концепция, предпочтение, установка) поведения системы или пользователя для выражения и удовлетворения информационной потребности пользователя, обусловленный как характером цели и видом поиска, так и системными "стратегическими" решениями – архитектурой БД, методами и средствами поиска в конкретной АИПС.

Выбор стратегии в общем случае является оптимизационной задачей. На практике в значительной степени он определяется искусством достижения компромисса между практическими потребностями и возможностями имеющихся средств.

Метод поиска – совокупность моделей и алгоритмов реализации отдельных технологических этапов: построения поискового образа запроса (ПОЗ), отбора документов (сопоставление поисковых образов запросов и документов), расширения и реформулирования запроса, локализации и оценки выдачи.

Поисковый образ запроса – записанный на ИПЯ текст, выражающий смысловое содержание информационного запроса и содержащий указания, необходимые для наиболее эффективного осуществления информационного поиска.

Методы поиска, т.е. выделение подмножества документов, потенциально содержащих описание решения задачи отбора документов (ОД), являются отражением процесса нахождения решения и зависят от характера задачи и предметной области.

Рассматривая поиск как итеративный процесс, методы сокращения пространства перебора (просматриваемого подмножества) образуют по существу методологическую основу стратегии поиска и могут быть разделены на следующие классы – методы поиска в:

1. одном пространстве (обычно, тематическом);
2. иерархически упорядоченном пространстве;
3. альтернативных пространствах;
4. динамическом (изменяющемся в процессе поиска) пространстве.

Реализуемый метод построения ПОЗа должен обеспечивать эффективные способы построения запроса для достижения целей различного типа.

Механизмы поиска – совокупность реализованных в системе моделей и алгоритмов процесса формирования выдачи документов в ответ на поисковый запрос.

Средства поиска, с одной стороны, – взаимозависимый комплекс информационно-поисковых языков (ИПЯ) и языков определения/управления данными, обеспечивающий структурные и семантические преобразования объектов обработки (документов, словарей, совокупностей результатов поиска), а с другой, – объекты пользовательского интерфейса, обеспечивающие управление последовательностью выбора операционных объектов конкретной АИПС.

Поисковые технологии – унифицированные (оптимизированные в рамках конкретной АИПС) последовательности эффективного использования отдельных средств поиска в процессе взаимодействия пользователя с системой для устойчивого получения конечного и промежуточных результатов.

Навигация как реализация процесса поиска по запросу в выбранной БД – целенаправленная, определяемая стратегией, последовательность использования методов, средств и технологий конкретной АИПС для получения и оценки

результата.

Средства навигации позволяют пользователю осуществлять управление процессом поиска. Они предоставляются пользователю в виде интерфейса, позволяющего организовать более или менее эффективный процесс взаимодействия с БД. При этом "дружественность" интерфейса характеризуется не только эргономичностью и понятностью, но и вариантностью выбора операционных объектов.

Процесс поиска информации представляет последовательность шагов, приводящих при посредстве системы к некоторому результату, и позволяющих оценить его полноту. Так как пользователь обычно не имеет исчерпывающих знаний об информационном содержании ресурса, в котором проводит поиск, то оценить адекватность выражения запроса, равно как и полноту получаемого результата, он может, основываясь лишь на внешних оценках или на промежуточных результатах и обобщениях, сопоставляя их, например, с предыдущими.

Процесс поиска можно представить в виде следующих основных компонент:

1. формулирование запроса на естественном языке, выбор поисковых системы и сервисов, формализация запроса на соответствующем ИПЯ;
2. проведение поиска в одной или нескольких поисковых системах;
3. обзор полученных результатов (ссылок);
4. предварительная обработка полученных результатов: просмотр содержания ссылок, извлечение и сохранение релевантных и пертинентных данных;
5. при необходимости, модификация запроса и проведение повторного (уточняющего) поиска с последующей обработкой полученных результатов.

Для уменьшения объёма отобранных материалов осуществляют фильтрацию результатов поиска по типу источников (сайтов, порталов), тематике и другим основаниям.

По используемым поисковым технологиям ИС можно разбить на 4 категории:

1. Тематические каталоги;
2. Специализированные каталоги (онлайновые справочники);
3. Поисковые машины (полнотекстовый поиск);
4. Средства метапоиска.

В Интернете ИПС размещается на одном или нескольких серверах. В ИПС собирается, индексируется и регистрируется информация о документах,

имеющихся в обслуживаемой системой группе веб-серверов. В документах индексируются все значащие слова или только слова из заголовков.

Тематические каталоги предусматривают обработку документов и отнесение их к одной из нескольких категорий, перечень которых заранее задан. Фактически это индексирование на основе классификации. Индексирование может проводиться автоматически или вручную с помощью специалистов, просматривающих популярные веб-узлы и составляющих краткое описание документов-резюме (ключевые слова, аннотация, реферат).

Специализированные каталоги или справочники создаются по отдельным отраслям и темам, по новостям, по городам, по адресам электронной почты и т. п.

Поисковые машины (самое развитое средство поиска в Интернете) реализуют технологию полнотекстового поиска. Индексируются тексты, расположенные на опрашиваемых серверах. Индекс может содержать информацию о нескольких миллионах документов. Например, в индексе популярной ИПС "AltaVista" более 56 млн. URL-адресов.

При использовании средств метапоиска запрос осуществляется одновременно несколькими поисковыми системами. Результат поиска объединяется в общий, упорядоченный по степени релевантности список. Каждая система обрабатывает только часть узлов сети, что позволяет расширить базу поиска. К подобному классу можно отнести и "персональные программы поиска", позволяющие формировать свои собственные инструменты метапоиска (например, автоматически опрашивать часто посещаемые узлы).

Базы информационных данных могут содержать практически любые виды информации, в том числе в любой комбинации. Информационный поиск осуществляется как по существующим в полнотекстовых ЭИР терминам, так и по специальным элементам, входящим в состав ИПЯ. Для формирования запросов используются специальные информационно-поисковые языки.

ИПС внутри найденной выборки обычно пытаются расположить документы в порядке их "релевантности", то есть близости к введенному пользователем запросу. Критериев такой близости много и выявление близких "по смыслу" к запросу документов не решает проблемы получения информации при отсутствии релевантного документа. Подобная ситуация достаточно тривиальна, в том числе и потому, что пользователь зачастую ищет документ, который сам собирается написать. Следует отметить, что в результате проведенного поиска пользователь

может получить как релевантные, пертинентные, так и нерелевантные и непертинентные подмассивы данных.

ИПС фактически являются системами информационного обеспечения и представляют собой базы и банки данных. В качестве объекта в них выступает индивид, организация, отрасль, регион и т.п. Субъектом информационного обеспечения является специалист-информатик, любой потребитель информации.

Организация поиска

Предлагается процедуру поиска необходимой информации разделить на девять основных этапов:

- Определение области знаний;
- Выбор типа и источников данных;
- Сбор материалов необходимых для наполнения информационной модели;
- Отбор наиболее полезной информации;
- Выбор метода обработки информации (классификация, кластеризация, регрессионный анализ и т.д.);
- Выбор алгоритма поиска закономерностей;
- Поиск закономерностей, формальных правил и структурных связей в собранной информации;
- Творческая интерпретация полученных результатов;
- Интеграция извлеченных "знаний".

Для проведения поиска первоначально на компьютере пользователя загружается интерфейс работы с соответствующей БД. Это может быть локальная или удалённая БД. Первоначально следует определиться с видом поиска (простой, расширенный и т.д.). Затем с набором предлагаемых для поиска полей. ИПС могут предложить для ввода одно или несколько полей. В последнем случае это обычно поля: автора, заглавия (названия), временного периода, вида документа, ключевых слов, рубрик и др. При формировании запроса практически все системы позволяют использовать логические элементы "И", "ИЛИ", "НЕТ".

Технологии поиска информации

Поисковые средства и технологии, используемые для реализации информационных потребностей, определяются типом и состоянием решаемой пользователем задачи основной деятельности: соотношением его знания и незнания об исследуемом объекте. Кроме того, процесс взаимодействия пользователя с системой определяется уровнем знания пользователем содержания ресурса (полноты представления, достоверности источника и т.д.) и функциональных возможностей системы как инструмента. В целом эти факторы обычно сводятся к понятию "профессионализма" – информационного (подготовленный/неподготовленный пользователь) и предметного (профессионал/непрофессионал) "профессионализма".

Процесс поиска информации обычно носит эмпирический характер. Он представляет последовательность шагов, приводящих при посредстве системы к некоторому результату, позволяющих оценить его полноту. При этом поведение пользователя, как организующее начало управления процессом поиска, мотивируется не только информационной потребностью, но и разнообразием стратегий, технологий и средств, предоставляемых системой.

Обычно пользователь не имеет исчерпывающих знаний об информационном содержании ресурса, в котором проводит поиск, поэтому оценить адекватность выражения запроса, как и полноту получаемого результата, он может, отыскав дополнительные сведения, или организовав процесс так, чтобы часть результатов поиска могла использоваться для подтверждения или отрицания адекватности другой части.

Операционными объектами, непосредственно участвующими во взаимодействии пользователей с поисковой системой являются поисковый образ документа (ПОД) и ПОЗ, соответствие которых устанавливается поисковым механизмом АИПС на формальном уровне. Адекватность образа действительному содержанию документа определяется качеством процесса свертки информации и уровнем знания субъектом средств отражения – концептуальной схемы предметной области и возможностей ИПЯ.

Поисковый образ документа – описание документа, выраженное средствами ИПЯ и характеризующее основное смысловое содержание или какие-либо другие признаки этого документа, необходимые для его поиска по запросу.

Большинство ПС изначально предлагают пользователям либо БО, либо ссылки на полные или частичные документы, их описание и другое, хранящиеся в различных

АИПС. Современные ПС позволяют определиться и указать какой и в каком виде источник информации интересует пользователя.

Методы обработки результатов поиска

По характеру преобразований (в контексте дальнейшего использования результатов обработки) методы обработки результатов поиска можно условно разделить на две группы:

1. Структурно-форматные преобразования;
2. Структурно-семантические преобразования (информационно-аналитические, логико-семантические).

Реализация поиска

Что обычно ищут в Интернете: персональные данные об индивидуумах и организациях; различные адресные данные; конкретные материалы (статьи, книги, фотографии, справочные данные, программное обеспечение и др.) в том числе место их хранения; где и сколько стоят те или иные материалы, услуги, продукты и т.п.; информационные сайты и порталы и др.

Общепринята организация поиска по начальным фрагментам слова (поиск с усечением справа), например, вместо слова "библиотечный" можно ввести его фрагмент "библиоте*". При этом будут найдены документы, в которых содержится не только слово "библиотечный", но и "библиотека", "библиотекарь", "бибилотековедение" и др. В каждом случае пользователь должен представлять, что именно он хочет найти, так как в предложенном ему варианте будет найдено гораздо большее количество документов, чем при задании данного слова полностью (без усечения). В подобном случае возможно в полученном массиве информации провести уточняющий поиск и в результате получить более релевантные и пертинентные данные.

Оформление результатов

С точки зрения ИПС результат поиска в ней есть совокупность (подмножество) найденных документов или ссылок на них. Обычно он представляется пользователю в виде списка. То есть простейшей выходной формой в данном случае будет список ссылок в виде полных или частичных БО, найденных ИР. Такой список может быть тут же распечатан или послан на какой-либо адрес

электронной почты, если такая возможность предоставляется ИПС и пользователь подключен к Интернету.

Графические и полнотекстовые ЭИР могут предлагаться пользователю только для просмотра, для копирования в различных форматах и масштабах, причём полностью или частично. Графические ИР обычно существуют в общепринятых форматах типа: JPG, GIFF, TIFF, BMP и др., а для текстовых материалов обычно используют текстовые форматы TXT, DOC и др., HTML и PDF – фактически графический формат, в котором могут сохраняться как текстовые, так и графические данные.

Полученные в результате поиска документы сохраняют.

Критерии оценки поиска

Критерием результата поиска является получение пользователем списка документов, одного документа или их частей, максимально удовлетворяющего его потребностям, сформулированным в поисковом запросе. В ИПС принято формировать список полученных в результате поиска документов по их релевантности. Различают критерии смыслового и формального соответствия между поисковым предписанием и выдаваемым документом.

Полнота и точность поиска являются взаимосвязанными показателями. Увеличение одного из них ведёт к снижению другого. В современных ИПС при сбалансированном поиске их значения составляет примерно 70%. Следует учитывать ситуацию, при которой список выданных поисковой системой ссылок содержит несколько, а порой и десятки разных адресов с одним и тем же текстом. Подобные ссылки характеризуются как дубликаты. Из них, при подсчёте коэффициентов учитывается только один документ.

Оценка и обработка результатов поиска

Учитывая, что идеальный результат поиска должен удовлетворять требованиям единственности, полноты и непротиворечивости, получаем, что различные виды поиска определяют различные требования к функциональным возможностям системы в части оценивания результата. Однако, для случая предметного поиска доказательство полноты является тривиальным: непустой результат поиска

подтверждает факт существования (или отсутствия) объекта, обладающего искомыми свойствами. При этом результат тематического поиска множественен и требует последующей систематизации – ещё одного процедурного шага для упорядочения полученного множества объектов по значениям не определённого явно основания. В свою очередь, проблемный поиск предполагает уже двухуровневую систематизацию.

Развитие процесса поиска осуществляется путём модификации выражения ПОЗ, путем реформулирования запроса и проведения повторного поиска в том же массиве данных или в подмассиве, полученном в результате осуществления первоначального поиска.

Интерфейсные средства обработки результата и развития поиска используют два типа операционных объектов – отдельные документы или коллекции документов.

Интернет-поисковые системы

Для получения информации в среде Интернета создаются специальные поисковые системы. Как правило, они общедоступны и обслуживают пользователей в любой точке планеты, где имеется возможность работы с Интернетом. Непосредственно для поиска используются поисковые машины, число которых в мире исчисляется несколькими сотнями. Они ориентируются на определенные типы запросов или их сочетание (библиографический, адресный, фактографический, тематический и др.). Кроме того, бывают полнотекстовые, смешанные и другие поисковые машины.

Для проведения поиска в Интернете (в WWW) функционирует множество сайтов и поисковых систем, поэтому необходимо не только ориентироваться в таких системах, но и уметь осуществлять в них эффективный поиск, то есть использовать соответствующие технологии.

"Технология поиска (англ. "Search Technology") означает совокупность правил и процедур, в результате выполнения которых пользователь получает ИР.

При поиске в Интернете рекомендуется обращать внимание на две составляющие: полноту (ничего не потеряно) и точность (не найдено ничего лишнего). Обычно соответствие найденных материалам этим критериям называют релевантностью, то есть соответствием ответа вопросу (запросу).

Поисковые системы характеризуются также временем выполнения поиска, интерфейсом, предоставляемым пользователю и видом отображаемых результатов. При выборе поисковых систем обращают внимание на такие их параметры, как охват и глубина.

Под охватом понимается объём базы поисковой машины, измеряемый тремя показателями: общим объёмом проиндексированной информации, количеством уникальных серверов и количеством уникальных документов. Под глубиной понимается – существует ли ограничение на количество страниц или на глубину вложенности директорий на одном сервере.

Каждая поисковая машина имеет свои алгоритмы сортировки результатов поиска. Чем ближе к началу списка, полученного в результате проведения поиска, оказывается нужный документ, тем выше релевантность и лучше работает поисковая машина.

Поисковые машины используют общие принципы работы, ориентированные на выполнение двух основных функций.

Первая функция реализуется программой-роботом, автоматически просматривающей различные сервера в Интернете. Находя новые или изменившиеся документы, она осуществляет их индексацию и передаёт на базовый компьютер поисковой машины. "Робот" – автоматизированный браузер, загружающий веб-страницу, изучающий её и, при необходимости, переходящим к одной из её гиперсвязей. Когда ему попадается страница, не содержащая связей, робот возвращается на одну-две ступени назад и переходит по адресу, указанному в одной из обнаруженных ранее связей. Запущенный робот проходит огромные расстояния в среде Интернета (киберпространстве), ориентируясь на развитие веб-сети и изменяя в соответствии с этим свои маршруты. Индексирующие роботы обрабатывают лишь HTML-файлы, игнорируя изображения и другие мультимедийные файлы. Они могут: обнаруживать связи с уже несуществующими страницами; устанавливать связь с наиболее популярными узлами, подсчитывая количество ссылок на них в других веб-страницах; регистрировать веб-страницы для оценки роста системы и др. Чаще всего роботы просматривают сервера самостоятельно, находя новые внешние ссылки в уже обследованных документах.

Вторая функция заключается в обработке выявленных документов. При этом учитывается все содержание страниц (не только полный текст, но и наличие иллюстраций, аудио и видео файлов, Java-приложений). Индексации подвергаются

все слова в документе, что дает возможность использовать поисковые системы для детального поиска по самой узкой тематике. Образуемые гигантские индексные файлы, хранящие информацию о том, какое слово, сколько раз, в каком документе и на каком сервере употребляется, составляют БД, к которой собственно и обращаются пользователи, вводя в поисковую строку ПОЗ (сочетания ключевых слов). Выдача результатов осуществляется с помощью специальной подсистемы, производящей интеллектуальное ранжирование результатов. В своих расчетах она опирается на местоположение термина, частоту его повторения в тексте, процентное соотношение данного термина с остальным текстом на данной странице и другие параметры, характеризующие возможности конкретной поисковой машины.

"Роботы" имеют ряд разновидностей, одной из которых является "паук" (англ. "spider"). Он непрерывно "ползает по сети", переходя с одной веб-страницы к другой с целью сбора статистических данных о самой "паутине" (Web) и (или) формирования некоторой БД с индексами содержимого веб.

Автоматизированные агенты "спайдеры" регулярно сканируют веб-страницы и актуализируют БД адресов (гиперссылки), средства индексирования информации, расположенные по указанным адресам. Полученные индексы используются для быстрого и эффективного поиска по набору терминов, задаваемых пользователем.

В разных системах эта цель достигается различным образом. Одни посылают "агентов" на каждую попадающуюся веб-страницу, индексируя все встречающиеся слова. Другие сначала анализируют БД адресов, определяя наиболее популярные (обычно подсчитывается число имеющихся ссылок на них). Именно эти веб-страницы в различной степени индексируются (только заголовки веб-страниц и ссылки, включая автоматическое аннотирование документов или весь текст).

Все чаще применяются "интеллектуальные агенты" – небольшие программы, обладающие способностью самообучаться, и действовать самостоятельно от имени своего владельца. Имея связь с компьютером пользователя, они выступают в роли персональных помощников, выполняющих ряд задач с применением знаний о потребностях и интересах пользователя. Интеллектуальные роботы-агенты ведут самостоятельный поиск в сети по собственным уникальным алгоритмам. Некоторые из них не только просматривают ключевые слова, но и осуществляют в Интернете семантический анализ информации, выявляя степень ее смыслового соответствия поставленной задаче.

Эффективный доступ к информации в Интернете обеспечивают такие зарубежные поисковые системы (машины), как Альта-Виста (AltaVista), "Lycos", "Yahoo", "Google", "OpenText", "Wais", "WebCrawler" и др. Их адреса в Интернете: www.altavista.com, www.yahoo.com, www.gogle.com, www.opentext.com,

К отечественным поисковым машинам относятся: Апорт ("Aport" АО Агама), Rambler (фирма Stack Ltd.), Яндех ("Yandex" фирма CompTek Int), "Русская машина поиска", "Новый русский поиск", и др. Их адреса в Интернете: www.aport.ru, www.rambler.ru, www.yandex.ru, search.interrussia.com, www.openweb.ru соответственно) и др.

Все эти поисковые машины позволяют по ключевым словам, тематическим рубрикам и даже отдельным буквам оперативно находить в сети, например, все или почти все тексты, где эти слова присутствуют. При этом пользователю сообщаются адреса сайтов, где найденные ИР постоянно присутствуют. Однако ни одна из них не имеет подавляющих преимуществ перед другими. Для проведения надежного поиска по сложным запросам специалисты рекомендуют использовать последовательно или параллельно (одновременно) различные ИПС.

Полнотекстовая поисковая машина индексирует все слова видимого пользователю текста. Наличие морфологии дает возможность находить искомые слова во всех склонениях или спряжениях. Кроме этого, в языке HTML существуют тэги, которые также могут обрабатываться поисковой машиной (заголовки, ссылки, подписи к картинкам и т.д.). Некоторые машины умеют искать словосочетания или слова на заданном расстоянии, что часто бывает важно для получения разумного результата.

Несмотря на общие принципы построения, поисковые системы отличаются тематикой, ее объемом, классификацией и интерфейсами. Для удобства перемещения (навигации) по имеющимся на поисковых машинах разделам некоторые из них используют специальный раздел "Карта".

Зачастую пользователю требуется текстовая и картографическая информация одновременно. В 80-е годы XX века эксперименты по решению этой проблемы начали проводить в Канаде, так появились первые географические информационные системы (ГИС) – компьютерные системы, позволяющие эффективно работать с пространственно-распределенной картографической информацией. ГИС – закономерное расширение концепции БД, дополняющее их наглядностью представления и возможностью решать задачи пространственного анализа. Они применяются для землеустройства, контроля ресурсов, экологии,

муниципального управления, транспорта, экономики, решения социальных задач и др. До 80-90% всей информации, с которой обычно имеют дело пользователи, может быть представлено в ГИС. ГИС – этап перехода к безбумажной технологии обработки информации.

При проведении поиска поисковые серверы обычно используют данные, хранящиеся в веб-страницах в тегах метаданных: (title), (meta name=”keywords”) и (meta name=”description”). Формируя свои страницы, следует отражать в этих тегах сведения о назначении сайта и его тематике.

При этом необходимо знать, что чем меньше количество ключевых слов включено в эти теги, тем с большей частотой они могут встречаться в текстах страниц сайта и, следовательно, тем выше их релевантность. Оптимальным считается частота таких слов не более 5%. Ключевых слов должно быть не очень много, они в большей степени должны состоять из одного или двух слов, образуя наиболее употребляемые термины. Чем большую релевантность имеют ключевые слова, тем большую конкурентоспособность они придают документу с точки зрения поисковых машин.

Полноту и точность ответа пользователь получает в зависимости от точности сформулированного им запроса. В результате поиска ему обычно предоставляется гораздо больше информации, чем ему необходимо, часть которой может вообще не иметь отношение к сформированному запросу. Легко заметить, что многое зависит не только от грамотно сформулированного запроса, но и от возможностей поисковых систем, которые весьма различны. При этом достаточно ярко проявляется "лесной синдром" (из-за леса не видно дров), заключающийся в том, что в полученных данных можно пропустить главные, необходимые сведения. Очевидно, никакие меры не являются исчерпывающими в условиях постоянного расширения среды и появления новых разнообразных ИР, что подтверждает трудности поиска в WWW.

Простые запросы в виде отдельных достаточно распространенных терминов приводят к извлечению тысяч (сотен тысяч) документов, абсолютное большинство которых пользователю не требуется (информационный шум).

Важным аспектом также является возможность таких систем поддерживать многоязычность, то есть способность обрабатывать запросы на различных языках. Пользователям предлагаются двуязычные словари, электронный переводчик и др. Кроме того, появились системы, осуществляющие мгновенный ("на лету") перевод

информационных ресурсов, найденных пользователем в Интернет и копируемых на его компьютер.

Актуальным является использование машиночитаемых тезаурусов. Электронный тезаурус – словарь, предназначенный для анализа текста и информационного поиска, включающий широкий набор семантических отношений между составляющими его терминами.

Создаются системы, позволяющие эффективно вести поиск в полнотекстовых БД. Они базируются на использовании технологий синтаксического и морфологического анализа текста (разбивка на элементы, распознаваемые программой) и оперативной обработки текстов на естественных языках.

Разработчики поисковых систем пытаются адаптировать их под начинающих и "средних" пользователей Интернета, количество которых неуклонно растет. В канадской системе (www.web-help.com), пользователям предлагается набор ссылок, подготовленных сотрудниками интернет-компании. На запрос пользователя сотрудник в реальном режиме времени находит и подключает на экран пользователя соответствующий (по его мнению) сайт. Метод удобен для нахождения конкретных фактов, статистики и т.п., которые другими способами непросто найти.

При организации одинакового запроса на разных поисковых машинах возможно получение различных по содержанию и широте охвата материалов. Искусство построения запроса требует знаний особенностей каждой конкретной поисковой системы и наличия опыта работы с Интернетом вообще. Некоторые поисковые машины предлагают квазиинтеллектуальные средства, позволяющие менее опытному пользователю, традиционно задавая вопросы на естественном языке, получать достаточно релевантные данные.

Обычно поиск в полнотекстовых БД осуществляется с использованием морфологических анализаторов (как правило, русских и английских), позволяющих автоматически находить существующие словоформы по фрагменту слова, слову, фразе, даже если в словах запроса присутствуют некоторые опечатки.

Используются метапоисковые системы, обеспечивающие в результате поиска получение суммарных данных с десятка поисковых систем, но при этом объем информации может быть весьма значительным. Частично данная проблема решается предоставлением ими общего списка, в начале которого будут данные, наиболее релевантные запросу. Другим способом удовлетворения потребностей

пользователей явилось создание тематически узконаправленных поисковых систем на веб-сайтах – порталов.

Важность проблемы информационного поиска в Интернете породила целую отрасль, задача которой заключается в том, чтобы помочь пользователю в его навигации в киберпространстве. Составляют эту отрасль специальные поисковые инструменты. Условно их можно разделить на поисковые средства справочного типа или просто справочники (directories) и поисковые системы в чистом виде (search engines).

Метапоисковые системы

Увеличение числа поисковых систем в Интернете обусловило появление "метапоисковых систем". Они дают возможность пользователю одновременно в едином пользовательском интерфейсе, используя индексы обычных поисковых систем, работать с несколькими БД. Пока еще "метапоисковые системы" не позволяют реализовать все возможности отдельных поисковых систем, но в большинстве своем он обладают существенными быстродействием и степенью охвата Web-пространства, что определяет их все более возрастающие значение и популярность.

Источники.

1. https://studfiles.net/preview/5566097/page:22/
2. https://studme.org/54387/informatika/tehnologii_obrabotki_informatsii
3. https://studme.org/97263/informatika/tehnologii_obrabotki_graficheskoy_informatsii
4. https://works.doklad.ru/view/CqAn6Z0cMcg.html
5. http://info-tehnologii.ru/vid_inf/aft_ofisa/tabl_inf/index.html