ПОИСКОВАЯ СИСТЕМА

Авторы: Д. В. Барашев, Н. С. Васильева, Б. А. Новиков

ПОИСКОВАЯ СИСТЕМА, алгоритмы и реализующий их комплекс программ, предоставляющий возможность быстрого доступа к необходимой пользователю информации посредством поиска в большой коллекции доступных данных. П. с. необходимы для навигации (перенаправление пользователя по логически связанным данным к конечным сайтам) в огромном, продолжающем расти объёме информации; используются во всех отраслях деятельности человека, где необходимо обеспечить своеврем. доступ к данным. Обмен информацией в обществе осуществляется гл. обр. в текстовой форме, и не случайно, что весьма значит. долю информац. ресурсов современных П. с. составляет текстовая информация. Наибольшее распространение получили П. с., которые ищут информацию на сайтах *Всемирной паутины*; поисковые веб-серверы исполняют миллиарды запросов в день, способны с высоким качеством выполнять запросы пользователей за доли секунды. Существуют также специализир. П. с., которые предоставляют возможность работы с разл. коллекциями документов и данных. Составная часть П. с. – *поисковый сервер*.

Историческая справка

К первым П. с. можно отнести разработки (1930–1940-е гг.) оптико-механич. устройств для работы с массивами данных, хранящихся на микрофильмах; они позволили автоматизировать поиск документов при помощи предварит. разметки. «Статистическая машина» Э. Гольдберга (Германия) размечала документы на микрофильме при помощи перфорации плёнки и предоставляла возможность поиска документов посредством сравнения запроса и перфорированной разметки. В 1945 В. *Буш* описал машину Memex, которая хранила микрофильмированные книги и корреспонденцию, показывала их на экране и, помимо присваивания каждому документу определённого кода, предполагала также построение связей между

документами и последующую возможность навигации от документа к документу. В 1948 К. Муэрс (США) впервые ввёл термин «информационный поиск» («information retrieval») – процесс поиска неструктурированной документальной информации, были проведены первые эксперименты по компьютеризиров. поиску документов. Разработкам эффективных технологий системы текстового поиска уделяли большое внимание уже на ранних стадиях развития *информационных систем* (их назвали информационно-поисковыми системами, ИПС). В 1970-е гг. предложены метрики (характеристики) и механизмы оценки качества работы П. с., реализована одна из первых П. с. – MEDLARS (MEDical Literature Analysis and Retrieval System), позволявшая исполнять поисковые запросы к базе биомедицинской лит-ры; появились первые П. с., поддерживающие интерактивное исполнение запросов (online systems), напр. амер. справочно-правовая система LEXIS (ныне LEXIS Nexis). В 1980-е гг. большое внимание уделялось разработке механизмов взаимодействия пользователя с П. с. – пользовательских интерфейсов. Появление *Интернета* привело к созданию в кон. 1990-х гг. П. с. для Всемирной паутины, в основу которой легли результаты исследований и эксперим. разработок в области информац. поиска. Одним из способов организации доступа к информац. ресурсам Интернета стало создание каталогов сайтов, в которых ссылки на ресурсы группировались согласно тематике. Первым таким проектом стал сайт Yahoo (апр. 1994); в 1995 появились П. с. Lycos и AltaVista (последняя долгие годы была лидером в области поиска информации в Интернете). В 1997 С. Брин и Л. Пейдж создали П. с. Google в рамках исследовательского проекта в Станфордском ун-те, которая (на 2014) является мировым лидером по числу обрабатываемых запросов (более 122 млрд. запросов в месяц); первая в мире П. с., создавшая более 100 региональных версий. Назв. «Google» (Гугл) произошло от намеренно искажённого С. Брином слова «Googol», т. е. десять в сотой степени – 10^{100}. В 1997 официально анонсирована русскоязычная П. с. Яндекс (Yandex), которая обрабатывает св. 3 млрд. запросов в месяц (на 2013).

Особенности технологий и тенденции развития

Науч. основой П. с. являются математич. модели информац. поиска, индексные структуры, методы анализа данных (data mining), машинного обучения (см.

Распознавание образов), методы *искусственного интеллекта*, *математической статистики*, *компьютерной лингвистики*, обработки и анализа цифровых изображений и др. Стандартный подход к оценке П. с. подразумевает использование репрезентативных размеченных коллекций – набор документов, набор запросов и информация о релевантности (адекватности) документов коллекции каждому из запросов – и обычно составляется вручную авторами документов (экспертами в данной предметной области). Для получения объективной оценки релевантность документа запросу обычно оценивается несколькими людьми. Стандартные метрики, применяемые в совр. оценке текстового поиска, основываются на отношении релевантности документа запросу (рубрике). Для П. с. по очень большим коллекциям документов (напр., для поиска в Интернете) большую значимость обычно имеет точность результатов выдачи, поскольку пользователю важнее, чтобы все документы, выданные П. с. в ответ на его запрос, были релевантными, чем получить от П. с. абсолютно все документы из коллекции, соответствующие запросу (их может оказаться слишком много, и пользователь всё равно не сможет просмотреть все из них).

Технологии, используемые для реализации П. с., зависят от формата (спецификация структуры данных, записанных в компьютерном файле), объёма и способа размещения данных, по которым производится поиск. По способу размещения данных и их объёму можно выделить системы: веб-поиска, когда поиск осуществляется среди миллиардов документов, доступных в Интернете; корпоративного поиска, работающие со множеством внутр. данных корпорации; локального или персонального поиска, производящие поиск на жёстком диске пользователя; проблемно ориентированного поиска, создаваемые для поиска данных в определённой предметной области (напр., мед. лит-ра, заявки на патент).

Большинство П. с. осуществляют поиск по слабоструктуриров. или неструктуриров. данным, таким как текстовые или мультимедиадокументы, веб-страницы, электронные сообщения, *файлы*. Такие данные не имеют чётко определённой структуры в виде явно заданного набора атрибутов в противоположность структурированным данным, типичным для реляционных *баз данных*. Ключевым понятием, характеризующим технологию поиска в той или иной конкретной П. с., является модель поиска (включает

способ формирования представлений документов; способ формирования представлений поисковых запросов; вида критерия релевантности документов) и тип поддерживаемого поискового запроса. Модель булева поиска широко используется в системах текстового поиска, поддерживает обработку запроса, имеющего вид булева выражения, т. е. выражения, в котором ключевые слова используются в сочетании с операциями AND (И), OR (ИЛИ) и NOT (НЕ). В рамках данной модели текстовые документы обычно рассматриваются как множество слов. Такая модель позволяет определить наличие или отсутствие термина в документе и поддерживает точный поиск: документ либо удовлетворяет запросу, либо не удовлетворяет.

Альтернативой модели булева поиска являются модели поиска с ранжированием, позволяющие задавать поисковый запрос в произвольной форме (напр., фраза на естеств. языке при текстовом поиске или картинка-образец при поиске по изображениям), которая не предполагает использование строгих конструкций. Модели поиска с ранжированием поддерживают неточный поиск: П. с. решает, какие документы удовлетворяют запросу наилучшим образом, и сортирует список результатов по степени их соответствия запросу. Пример такой модели – векторная модель, в которой каждый документ представляется в виде вектора, отражающего содержание документа. Набор всех возможных векторов образует векторное пространство. Типичным векторным представлением для текстового документа является вектор из коэффициентов относительной важности слова из словаря коллекции для данного документа (вычисляется на основе взвешенной частоты слова в документе). Использование таких моделей требует значительно больших вычислит. ресурсов по сравнению с др. моделями, однако они обеспечивают существенно более высокое качество поиска. Пример векторного представления изображений – вектор частот цветов отд. *пикселов*. Моделью поиска с ранжированием является также вероятностная модель, основанная на использовании математич. аппарата теории вероятности для оценки вероятности релевантности документа запросу пользователя. На практике часто используются комбинации разл. моделей поиска.

Осн. направления развития П. с. включают поиск по разнородным источникам информации (комбиниров. поиск по текстовым, структурированным и мультимедийным данным), фактографич. поиск, разработку вопросно-ответных систем и др.

Литература

Лит.: Черняк Л. Статистическая машина Э. Гольдберга // Открытые системы. 2004. № 3; Vannevar B. As we may think (*http://www.ps.uni-saarland.de/~duchier/pub/vbush/vbush-all.shtml*); Колисниченко Д. Н. Поисковые системы и продвижение сайтов в Интернете. М., 2007; Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск. М., 2011.