Синкевич Е.А. Процесс обработки информации поисковыми программами в сети Internet // Информационно-экономические аспекты стандартизации и технического регулирования. 2018. № 6. (46)

УДК 004.04

ПРОЦЕСС ОБРАБОТКИ ИНФОРМАЦИИ ПОИСКОВЫМИ ПРОГРАММАМИ В СЕТИ INTERNET

Синкевич Е.А., ФГБУН ВИНИТИ РАН

Аннотация. *Проведен анализ существующих способов современного поиска информации в сети Internet. Показана структура поисковых программ. Описан процесс индексации информации. Приведены способы выборки индексированной информации. Показаны особенности поиска научно-технической информации.*

Ключевые слова: поиск информации, поисковые машины, индексация гипертекстов, алгоритм поиска научно-технической информации

UDC 004.04

PROCESS OF PROCESSING INFORMATION BY SEARCH PROGRAMS ON THE INTERNET NETWORK

Sinkevich E.A. FSINI VINITI RAS

Annotation. *The analysis of existing methods of modern information search on the Internet has been carried out. The structure of search programs is shown, their comparative analysis is given. The process of indexing information is described. The methods of sampling the indexed information. The features of the search for scientific and technical information are shown.*

Keywords: information search, search engines, hypertext indexing, search algorithm, scientific and technical information

Поиск необходимой научно-технической информации, в настоящее время, становится все более актуальной задачей в условиях ежедневного роста ее объема. Для осуществления поиска на сегодняшний день уже сформировались методики и средства.

Однако, одним из немаловажных критериев быстрого поиска принято считать умение человека грамотно сформулировать свой запрос к поисковой

машине, работа с информационным массивом и отбор из этого искомой информации, так как в результате поиска выдается на много больший объем информации, чем был ему необходим изначально, при этом часть ее может вообще не иметь отношение к сформированному запросу [1].

К решению задачи поиска в основном подходят при помощи общедоступных и известных поисковых программ, таких как Яндекс, Google, Рамблер. Несмотря на то, что они снабжены дополнительными функциями уточнения и формулировки запроса обладают рядом общих недостатков: большое количество рекламы; результат поиска, без уточнений, не соответствует запросу пользователя.

Другими существенными недостатками (проблемами) являются поиск взаимосвязанной информации, расположенной в разных областях знаний, то есть разбросанной по различным рубрикаторам и индексам научно-технической информации, и восстановление хронологии становления научно-технической мысли, проектных решений и результатов их использования, например в промышленном производстве [2].

Для облегчения решения задачи по научно-техническому поиску информации и избавления от части рекламного контента были разработаны специальные программы такие как: Академия Google, Scholar, платформа Flexum, Scirus, ScienceResearch, BASE. Но и они, имея свои уникальные алгоритмы, используют все те же методы поиска информации [2].

Развитие поисковых программ в основном не меняет их структуры, а затрагивает лишь отдельные элементы, такие как [3]:

- «Паук» (spider) – перемещаясь по сети скачивает веб-страницы;

- «Червяк» (crawler) – анализирует найденные веб-страницы и извлекает находящиеся на них ссылки;

- Индексатор (indexer) – систематизирует информацию, найденную пауками;

- База данных (database) – хранилище всех обработанных данных, накопленных поисковой системой;

- Механизм выдачи результатов (search engine) – интерфейс пользователя для работы с базой данных.

Взаимодействие элементов поисковой системы схематично изображено на рисунке 1.

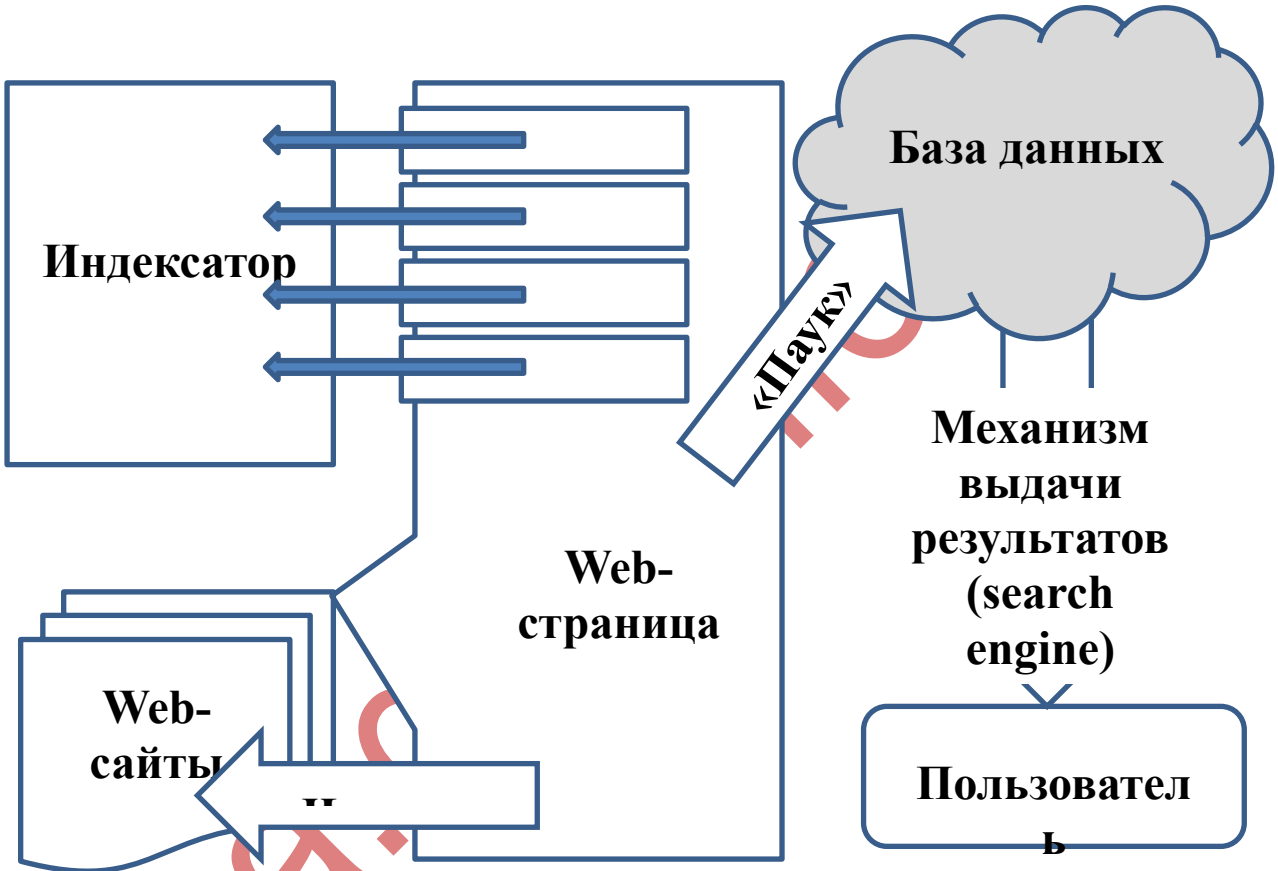

Рис.1 Взаимодействие элементов поисковой системы

Как можно увидеть из рисунка 1, основным элементом данной системы является индексатор, и от его работы зависит качество и скорость выдаемого результата поиска. Индексы в базе данных являются аналогом оглавления (содержания) обычной книги.

В настоящее время существует множество методов, решающих проблему индексирования XML-данных. В соответствии с их подходами мы можем классифицировать их следующим образом:

Методы, основанные на графах, строят структурную ветвь пути, которая может использоваться для повышения эффективности запросов, особенно для запросов с одним путем. В данной категории можно классифицировать, например, следующие методы: DataGuides, 1-индекс, иерархическая индексация, подход к поддержке XPath запросов (PP-индекс), следующий и предыдущий (F&B)-индекс, в XML XPath графический индекс (MTree) или компактное дерево (CTree).

Основанные на последовательностях методы преобразуют как исходные данные, так и запрос в последовательности. Таким образом, запрос XML-данных эквивалентен поиску соответствий подпоследовательности. В этой категории можно классифицировать, например, следующие методы: виртуального дерева (ViST), последовательности для индексации XML (PRIX).

Методы кодирования узлов применяют определенные стратегии кодирования для разработки кодов для каждого узла, для того, чтобы связь между узлами оценивалась вычислением. В эту категорию, мы можем классифицировать, например, систему индексирования и хранения XML (XISS), XR-дерево (см.рисунок 2) [4].

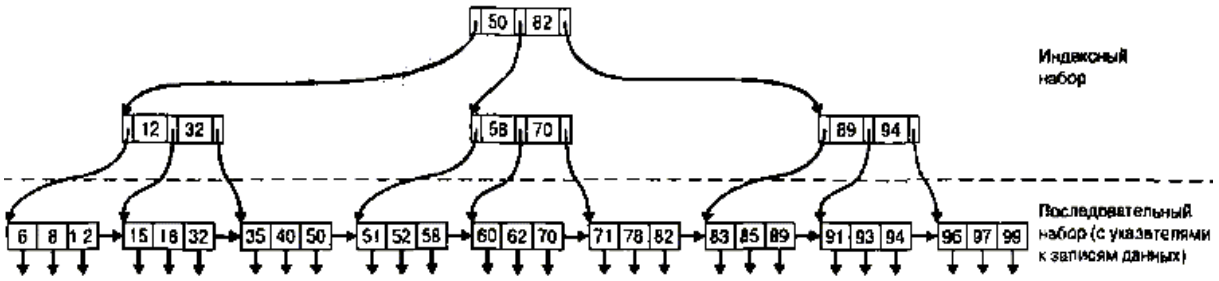

Рис. 2 Усовершенствованные сбалансированные древовидные индексы

Адаптивные методы могут осуществлять изменение структуры индекса в соответствии с рабочей нагрузкой запроса и применяются только для часто используемые запросы. В данной категории можно классифицировать,

например, следующие методы: адаптивный путь индекса для данных XML (APEX) и адаптивный индекс ветвления XML-запросов (AB-Index) [5].

Каждый метод имеет свои преимущества и недостатки, но один из приведенных выше методов не позволяет осуществлять поиск взаимосвязанной научно-технической информации. Например, нам нужно найти материал для изготовления изделия, технологические процессы для его обработки, станки, которые позволят реализовать технологию и т.п.

В частном случае, хронологию научного исследования и соответствующую ей информацию можно проследить по ранжированию выдаваемого результата поиска [6]. Само же обнаружение научно-технической информации в общем объеме индексируемых данных осуществляется измерением весов каждого употребляемого слова с целью выборки из массива специфических терминов, так как терминология отличается от ключевых слов. Ключевые слова, как правило, появляются в документах много раз, тогда как терминология может появиться в тексте документа всего лишь один раз, оказывая при этом существенное влияние на понимание контекста [7].

Таким образом, постоянно совершенствующие современные поисковые машины требуют создания различных специализированных надстроек, которые бы позволяли решать указанные выше проблемы и формировать корректный информативный результат в соответствии с запросами пользователя.

Список использованных источников и литературы

1. Алексеев А.В. Проблема поиска и обработки информации в современной информационной среде. – Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева 2017.
2. Кутовенко А. Профессиональный поиск в Интернете. – СПб.: Питер, 2011. – 256 с.: ил.
3. Александров Е. Интернет – легко и просто! – СПб.: Питер, 2011. – 129 с.

4. http://bourabai.ru / – Индексирование в базах данных
5. Eliška Šestáková Automata Approach to XML Data Indexing – Faculty of Information Technology, Czech Technical University in Prague, 2018
6. Ющук Е.Л. Интернет-разведка [Руководство к действию] – Екатеринбург, 2007. – 269 с.
7. Zhao, C, Dong, C and Zhang, X. 2018. ROCP: A Rapid Ontology Construction Platform from Unstructured Data. Data Science Journal, 17