

И. А. Семёнов

К вопросу об эффективности поиска конкретики¹ в Интернете

Человек постоянно находится в поиске, это свойство его неумейной натуры. Современные технологии предоставляют пользователю огромные возможности для поиска информации, однако из-за несовершенства существующих поисковых инструментов (в первую очередь полнотекстовых поисковых машин) найти что-то подходящее среди сотен тысяч выдаваемых ссылок бывает непросто. Это знает каждый, кто хоть раз искал что-нибудь в Яндексе, Рамблере или Google.

Всемирная паутина (World Wide Web) содержит огромный объем информации. Согласно исследованиям Berkeley [1], по состоянию на 2003 год он оценивался в 258,85 терабайта, и это только общедоступные данные. По данным [2], рост аудитории Интернета с 2003 по 2007 год составил 117% (на декабрь 2007 года аудитория оценивалась в 1,3 млрд человек). Если предположить, что объем данных в Интернете линейно зависит от числа пользователей (на самом деле эта зависимость скорее экспоненциальная), то в 2007 году суммарный объем информации в Интернете составлял не менее 561 терабайта. Оценить полный объем информации в Интернете в настоящее время крайне сложно, но можно с уверенностью сказать: он огромен. Кроме того, вся эта информация крайне хаотична, не предусмотрено практически никаких стандартов представления структурированной информации².

Таким образом, наиболее эффективным методом поиска является полнотекстовый поиск. Полнотекстовые поисковые машины — такие, как Google, Яндекс, Рамблер — собирают информацию в индексные базы, откуда она извлекается по запросу пользо-

вателя. Последовательно обходя все известные ему адреса, поисковый робот скачивает веб-страницы и индексирует их. При этом сама структура информации почти не учитывается, для поисковой машины нет разницы между страницей онлайн-энциклопедии и домашней страничкой. Правда, большинство поисковых машин умеют различать заголовки и подзаголовки, а также учитывают «ссылочный вес» страницы (количество ссылок на нее с других сайтов), но такая оценка далеко не всегда бывает эффективной.

Имея минимальное представление об алгоритмах работы поисковых роботов, можно оптимизировать веб-страницу таким образом, что ссылка на нее будет всплывать на первых местах по запросам, к которым она не имеет никакого отношения. Эта технология называется SEO (Search Engine Optimization). Основная идея SEO состоит в том, чтобы обмануть поискового робота, подсунув ему определенную информацию, отличную от той, которую видят посетители страницы. Из-за большого количества таких «оптимизированных» сайтов поиск по-настоящему релевантной информации становится непростым. В настоящий момент SEO не считается противозаконной технологией (в отличие от

¹ В английском языке слово concrete обозначает как бетон, так и нечто конкретное, основанное исключительно на фактах, а не на верованиях или догадках. — Прим. ред.

² На сегодняшний день World Wide Web Consortium (или W3C) постоянно ведет разработки в области подобных технологий и стандартов. В качестве примеров разработок, связанных со структурированием информации в Web, можно привести Common Web Language (CWL) или SKOS Simple Knowledge Organization System Reference. Однако, как верно отмечает автор статьи, ни одна из них не получила широкого распространения. — Прим. ред.