УПРАВЛЕНИЕ ПОИСКОВЫМИ РОБОТАМИ НА САЙТЕ

магистр В.В. Стрельцов, канд. физ.-мат. наук, доц. Е.П. Черных,
Национальный технический университет "Харьковский политехнический институт", г. Харьков

Найти информацию о любом товаре с помощью компьютера и Интернета – привычное дело в наше время. Многие с удовольствием пользуются услугами поисковых систем и даже не задумываются над тем, как они работают.

Поисковая система сканирует сайт и этим же индексирует контент страниц сайта для записи в базу. Вследствие чего эти страницы сайта будут отображаться обычным пользователям при вводе нужного запроса. Для сканирования и индексации сайта, поисковые системы используют поисковых роботов. Поисковый робот – это важнейший элемент поисковой системы, в задачу которого входит сбор новых данных о сайтах и их обновлениях. Поисковая система может располагать не одним, а несколькими поисковыми роботами. Каждый бот представляет собой автоматический скрипт, имеющий свой алгоритм работы, свое конкретное задание для определенного сайта. Система обладает большим отрядом разных роботов, которые выполняют разные задачи: одни ищут новые страницы, другие отвечают за нахождение "мертвых" сайтов и чистку поисковых данных, третьи индексируют картинки, четвертые находят видео. Одно из важнейших значений для робота – корневой файл robots.txt, расположенный на подконтрольном сервере. Этот файл – инструкция для робота. Во-первых, robots.txt может вообще не допустить бота на сайт и сайт останется не проиндексированным. Во-вторых, может закрыть боту доступ к определенным страницам и файлам, появляется сложность в создании самого файла robots.txt.

Для решения данной проблемы был предложен подход – разработка программного модуля (скрипта), с помощью которого можно отследить визиты поисковых роботов. Скрипт будет содержать необходимую информацию: дату посещения, имя бота, IP-адресов бота и страницы, которые он посетил. Это позволит определить, какие роботы посещали данный сайт и какую страницу. Такой отчет поможет определить страницы, на которые нужно запретить доступ, чтобы при последующем посещении робота доступ на эти страницы был закрыт.

Предложенный подход позволит управлять поведением робота, правильно оформить файл robots.txt и повысить безопасность сайта.