DOI: 10.24411/1993-8314-2019-10027

П. П. Кейно, *канд. техн. наук, доцент кафедры «Системное моделирование и автоматизированное проектирование» ФГБОУ ВО «Московский авиационный институт (национальный исследовательский университет)», science@blockset.ru*

О. В. Павлов, *студент кафедры «Системное моделирование и автоматизированное проектирование» ФГБОУ ВО «Московский авиационный институт (национальный исследовательский университет)», pavlov@smiap.ru*

К вопросу индексации файловых хранилищ на базе протокола FTP

Рассматривается система индексации серверов на базе протокола FTP. В статье описан набор функционала для индексации файлов и поисковой системы. Подробно рассмотрены основные проблемы, встречающиеся при обработке ответов на запросы по протоколу FTP и методы их решения. Несмотря на эволюцию файловых хранилищ, прошедшую за 34 года после создания протокола FTP, проблема индексации остается актуальной благодаря огромному массиву данных, хранящихся на существующих серверах. Разработанная система индексации состоит из трех основных частей: робота-индексатора, базы данных и Web-сервиса. Робот-индексатор способен работать с большинством типов FTP серверов и обрабатывать возвращаемые ими данные, избегая ссылочные ловушки и сохраняя результаты в базу данных. Web-сервис принимает запросы от конечного пользователя и возвращает результат поиска по базе данных. Пользователь может указать полное или частичное имя файла, тип файла и его размер. Ключевым отличием от существующих систем является отслеживание истории изменений файлов и серверов.

Ключевые слова: FTP, индексация, поисковая система, файловый протокол, поисковый робот

Введение

Индексация и последующее агрегирование информации, несмотря на развитие различных способов передачи и хранения данных, остается актуальной задачей на сегодняшний день. Также в связи с развитием технологий возникает проблема индексации данных, хранящихся на базе устаревающих технологий. Одной из таких проблем является поиск файлов на FTP-серверах. Несмотря на то, что протоколу FTP уже почти полвека, он до сих пор используется, а серверы на его основе хранят в себе колоссальный объем уникальных данных, причем эти данные представляют собой не только практический, но и исторический интерес [1]. По-прежнему актуальна задача балансировки нагрузки даже на ресурсах, основанных на базе протокола FTP [2]. Исходя из исследований предметной области, по-прежнему актуальны задачи «тонкой» настройки собственных FTP-серверов [3,4].

За долгую историю протокола было предпринято большое количество попыток создать систему индексации файлов на FTP-серверах, однако шло время, появлялись новые протоколы передачи данных, начиная от WebDAV [5] и заканчивая классическими облачными хранилищами, доступными через Web. Так, протокол FTP отдалился от широкого пользователя и про него стали забывать. Несмотря на то, что его поддержка присутствует на уровне различных операцион-